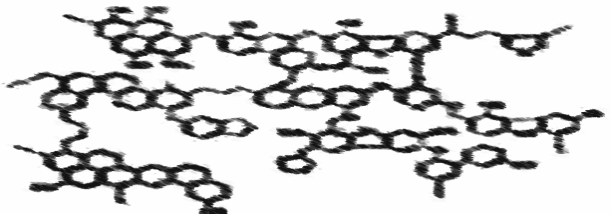
PROCEEDINGS OF THE 3-rd MOSCOW CONFERENCE ON COMPUTATIONAL MOLECULAR BIOLOGY MCCMB'07

MCCMB'07



PROCEEDINGS OF THE 3-rd MOSCOW CONFERENCE ON COMPUTATIONAL MOLECULAR BIOLOGY

⁂

Moscow, Russia,
July 27–31 2007

Moscow State University
INRIA, France
(the French National Institute for Research in Computer Science and Control)
Institute of Information Transmission Problems, Russian Academy of Sciences
Scientific Council on Biophysics, Russian Academy of Sciences
National Research Centre GosNIIGenetika
International Foundation of Technology and Investment
with financial support of
Russian Academy of Sciences
Ministry of Education and Science of the Russian Federation
Russian Fund of Basic Research

PROCEEDINGS

OF THE 3-rd MOSCOW CONFERENCE ON COMPUTATIONAL MOLECULAR BIOLOGY

MCCMB'07
Moscow, Russia,
July 27–31 2007

LOCAL ORGANIZING COMMITTEE

V.A. Sadovnichy (Moscow State University), chair

V.P. Skulachev (Faculty of Bioengineering and Bioinformatics of the MSU), deputy chair

M.S. Gelfand (Kharkevich Institute for Information Transmission Problems, RAS, Moscow), deputy chair, chair of the program committee

M. Regnier (INRIA, France), deputy chair

V.G. Tumanyan (Biophysics Council of RAS, Moscow), deputy chair

V.J. Makeev (GosNIIGenetika, Moscow), deputy chair

S.A. Spirin (Belozersky Institute of Physical and Chemical Biology, MSU), scientific secretary

A.V. Alekseevsky (Belozersky Institute of Physical and Chemical Biology, MSU)

N.K. Yankovsky (Institute of General Genetics, RAS, Moscow)

PROGRAM COMMITTEE

Inna Dubchak (Lawrence Berkeley Lab., USA)

Natalia G. Esipova (Engelhardt Institute of Molecular Biology, RAS, Russia)

Alexei V. Finkelstein (Institute of Protein Research, RAS, Pushchino, Russia)

Dmitry Frishman (Technical University of Munich, Munich, Germany)

Mikhail S. Gelfand (Kharkevich Institute for Information Transmission Problems, RAS, Moscow, Russia), Chair

Nikolay A. Kolchanov (Institute of Cytology and Genetics, Novosibirsk, Russia)

Alexei S. Kondrashov (University of Michigan, Ann-Arbor, USA)

Eugene Koonin (NCBI, Bethesda, USA)

Andrei M. Leontovich (Belozersky Institute of Phisical and Chemical Biology, Moscow State University, Russia)

Leonid Mirny (MIT, Cambridge MA, USA)

Andrei A. Mironov (Faculty of Bioengineering and Bioinformatics, Moscow State University, Russia)

Vladimir V. Poroikov (Orekhovich Institute for Bio-Medical Chemistry, RAMS, Moscow, Russia)

Mikhail A. Roytberg (Institute for Mathematical Problems of Biology, RAS, Pushchino, Russia)

Shamil R. Sunyaev (Brigham & Women's Hospital, Harvard University Medical School, USA)

CONTENTS

FINDING FUNCTIONAL REGULATORY SNPs

IRINA ABNIZOVA[1], LUISA FOCO[2], FEDOR NAUMENKO[3], TATIANA SUBKHANKULOVA[3], RENE TE BOEKHORST[4], LUISA BERNARDINELLI[1]

The research presented here combines the strengths of both genetics and genomics by investigating genetic variants, Single Nucleotide Polymorphisms (SNPs) in regulatory regions instead of genes. By bringing together the computational search and characterisation of regions in DNA that regulate gene expression on the one hand and information about individual variation in the structure of human DNA on the other hand, it aims to identify likely regulatory regions, the individual variation in their molecular make up and the effect this may have in the phenotypic expression of genes.

There is strong recent interest in **regulatory SNPs** [1-8]. There have been also demonstrated by combining experimental evidence and computation that the promoter regions of human genes provide a rich source of functional single nucleotide polymorphisms [4-8]. As many as 35% of promoter SNPs may be of functional significance [4]. There are, however, currently no computational tools, except of [8] for promoters, which can be used to assess directly from regulatory DNA sequence whether or not a given variant is likely to alter gene expression and hence be of functional significance.

Here, we present the approach that can allow *in silico* estimation of the likely functional consequences of single nucleotide changes in putative regulatory DNA. This approach is based on the integration of at least 16 sources of supervised sequence information about a given DNA stretch, with unsupervised methods [9,10]. We have also incorporated the novel method, which analyse a SNP functionality due to sensitivity of a mathematical model with respect to the SNP variant.

Essentially, the method consists of identifying regions in the human genome that are likely important in the regulation of gene expression and contain motifs that identity them as TFBSs. We then establish whether the motifs contain SNPs and if so, in how far these mutations destroy the signal by which regulatory proteins recognize the motifs as binding sites. Especially these SNPs could be strong candidates for further experimental verification to establish their possible role in the genesis of and susceptibility for particular diseases.

[1] MRC-BSU, Robinson Way, Cambridge, UK, irina.abnizova@mrc-bsu.cam.ac.uk, luisa.bernardinelli@mrc-bsu.cam.ac.uk
[2] University of Pavia, Italy, luisa.foco@unipv.it
[3] Queen Mary University, London, UK, f.naumenko@qmary.ac.uk, subkhankul@hotmail.com
[4] University of Hertfordshire, UK, r.teboekhorst@herts.ac.uk

Results. To test the method, we collected several known from literature disease-associated regulatory SNPs [1-3]. We checked if the disease-associated regulatory SNP is within one of the feature-predictions, and thus has a high score. We found that the scores of the disease-associated regulatory SNPs were among the highest scores for all SNPs for all our training sets. Furthermore, these SNPs appeared to be variant sensitive, namely some particular SNP variant changed the results of motif predictions. Interestingly, we found out that known disease-causal SNP variants formed significantly underrepresented motifs within local context.

1. Monsuur AJ, de Bakker PI, Alizadeh BZ, Zhernakova A, Bevova MR, Strengman E, Franke L, van't Slot R, van Belzen MJ, Lavrijsen IC, et al. (2005) Nat Genet. 37:1341-4.
2. Ueda H, Howson JM, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow DB, Hunter KM, Smith AN, Di Genova G, et al. (2003) Nature 423:506-11.
3. Morahan G, Huang D, Ymer SI, Cancilla MR, Stephen K, Dabadghao P, Werther G, Tait BD, Harrison LC, Colman PG (2001) Nat Genet. 27:218-21.
4. Hoogendoorn, B., Coleman, S. L., Guy, C. A., Smith, S. K., O'Donovan, M. C. and Buckland, P. R. (2004). Functional analysis of polymorphisms in the promoter regions of genes on 22q11. Hum. Mutat. 24, 35-42.
5. Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief. Bioinform. 6, 44-56
6. Pastinen, T. and Hudson, T. J. (2004). Cis-acting regulatory variation in the human genome. Science 306, 647-650
7. Hudson, T. J. (2003). Wanted: regulatory SNPs. Nat. Genet. 33, 439-440
8. tools
9. Paul R. Buckland , Bastiaan Hoogendoorn, Sharon L. Coleman, Carol A. Guy, S. Kaye Smith, Michael C. O'Donovan (2005) Strong bias in the location of functional promoter polymorphisms,
10. Khan I, et al. and Chuzhanova N. (2006) In silico discrimination of single nucleotide polymorphisms and pathological mutations in human gene promoter regions by means of local DNA sequence context and regularity, In Silico Biology 6, 0003
11. Irina Abnizova, Alistair G. Rust, Mark Robinson, Rene te Boekhorst and Walter R. Gilks, (2006) Prediction of TFBS using Markov models, J. of Bioinformatics and Comp. Biology, v4, n2, pp 425-441
12. Irina Abnizova, Rene te Boekhorst, Klaudia Walter and Walter R. Gilks, (2005), Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in eukaryotic genomes: the fluffy-tail test. BMC Bioinformatics, 6:109

2D MODELLING AND ANALYSIS OF SPATIALLY DISTRIBUTED CELLS TYPES OF PRIMARY SHOOT APICAL MERISTEM (SAM) OF ARABIDOPSIS THALIANA

ILYA R. AKBERDIN, EVGENY A. OZONOV, VICTORYA V. MIRONOVA, DMITRY N. GORPINCHENKO, NADEZDA A. OMELYANCHUK, VITALY A. LIKHOSHVAI, DENIS S. MIGINSKY, NIKOLAI A. KOLCHANOV

Development of organisms is a very complex process for understanding of that there are used methods of system computer biology along with experimental methods. It is well known that postembryonic development of the aboveground part of higher plants depends on the expression of apical shoot meristems, a dynamic structure which forms leafage, flowers and scape. The apical shoot meristem (SAM) is stem cells reservoir of plants and it regulates processes of growth and development in response to both incoming external signals (light, temperature) and internal signals (phytohormone, signal molecules). Therefore development rules of plant above ground level depend on mechanisms of meristem development in many respects. Object of our research is the apical shoot meristem of *Arabidopsis thaliana* during embryonic vegetative of developmental stages. Choice of the object as model object is determined by *Arabidopsis thaliana* is one of the most strongly studied of higher plant. There are strongly accumulated data both about molecular genetic processes and about spatial structures rules of the plant on the different stages his life cycle. In particular, there were revealed numerous genetic mutations which responsible for phenotypic anomalies in plant development. The cumulative experimental data allow starting construction of spatial distributed hierarchical model that will describe both molecular genetic processes and processes on the level of cell-cell interactions simultaneously. Development of this model allows to ascertain cause-and-effect relations between intracellular processes which are regulated gene networks and morphological characteristics of the plant and his separate parts (tissues, cell groups, individual cells). The cellular automaton was developed to model the development of shoot meristems of the Arabidopsis thaliana in embryogenesis on basis of experimental data from AGNS database (Arabidopsis Genenet Supplementary Database) (http://wwwmgs.bionet. nsc.ru/agns). Modeling covers the initiation of SAM, the formation of the SAM complex structure and its further functioning (Akberdin et al., 2007). Here the embryo is described as a two-dimensional array of cells, the rates of division of which depend on the cellular environment. The cells in the model may receive and, depending on the cell type, produce signals that should

Institute of Cytology and Genetics SB RAS, Novosibirsk State University, Lavrenteva ave, Novosibirsk, Russia, akberdin@bionet.nsc.ru, evgo1@gorodok.net

be received by other cells in the model. The biological meaning of signals is the concentration of certain diffusing substances, or morphogenes, which provide a specific influence on the cell.

Creation of a cellular automaton that imitates morphodynamics of embryo development by means of regulation of signals produced by different embryonic cells is a first step in modelling the process of development in general and in modelling the gene network for morphogenesis in particular. The formation of plant meristems in embryogenesis is characterized by a combination of a violent development of differentiating tissue and a stable development of its stem cells. Both processes were modeled in the cellular automaton being reported. Not only is this automaton a tool for predicting the dynamics of the division process and the cell differentiation process which underway in the systems being considered, but also for the examination of how real mutations influence the system.

1. Akberdin I.R., Ozonov E.A., Mironova V.V., Gorpinchenko D.N., Omelyanchuk N.A., Likhoshvai V.A., Kolchanov N.A. (2007). "A cellular automaton to model the development of shoot meristems of Arabidopsis thaliana", *Journal of Bioinformatics and Computational Biology* (in publication).

INTERLOCKS IS A CHARACTERISTIC FEATURE OF SANDWICH-LIKE DOMAINS

EVGENIY AKSIANOV[1], A.V. ALEXEEVSKI[1], A. KISTER[2], I. GELFAND[3]

Sandwich-like domains form a large group of protein domains with a similar architecture – two beta-sheets packed against each other – but rather different topologies. For their characterization and classification it is important to identify characteristic elements of their topology, i.e. elements contained in almost all sandwich-like domains and rarely contained in other classes of domains.

It was shown [1] that interlocks – two pairs of neighboring strands from two beta-sheets with special "interlocked" topology of the strands – are typical structural elements of sandwich-like domains. There are no publications on interlock occurrences in domains of other architectures. To investigate interlock spread in all solved protein structures, we have designed a computer aided procedure to classify all families in the SCOP 1.69 database into 3 groups: IL+ (all domains in the family contain an interlock), IL- (all domains are interlock-free) and IL+/-

[1] Moscow State University, A.N. Belozersky Institute of Physico-Chemical Biology, Vorobiovy gory, Moscow, Russia evaksianov@belozersky.msu.ru, aba@belozersky.msu.ru
[2] University of Medicine and Dentistry of New Jersey, USA
[3] Rutgers University, USA

(only a part of domains contain an interlock). First, we have developed an algorithm and computer program for detecting interlocks in 3D structures of protein domains. Briefly, the algorithm detects 4-tuples of strands with the interlock topology and checks their spatial arrangement by a set of local criteria. The details are described at http://monkey.belozersky.msu.ru/~evgeniy/i-locks/index.html. Testing showed that interlocks detected by the algorithm are confirmed by expert in more than 95% of cases. In the same time, about 10% of expertly confirmed interlocks were not detected by the algorithm.

All available protein structures were investigated by the following procedure.

1. Screening all domains of SCOP 1.69 database to check by the algorithm either a domain contains interlock or not. The results were presented in a table, which include also all levels of SCOP classification of protein domains. In the table SCOP families belonging to folds annotated as "sandwich" were considered as families of sandwich-like architecture.

2. Automatic expanding interlock detection by high sequence similarity of domains. We hypotysized, that domains with more than 60% identical amino acids in pairwise global Needleman-Wunsch alignment have very similar 3D structures and therefore, contain or not contain interlocks simultaneously. To do this, all domains in a family were divided into similarity groups and all domains of a group containing a representative with detected interlock were marked as IL+.

3. Representatives of many families and groups of domains were examined manually to correct possible algorithm mistakes. Among them (i) all families of sandwich-like domains having no interlock hits; (ii) all families and superfamilies out of sandwich-like folds; (iii) all families of IL+/- type.

By the described above procedure interlocks were detected in 9841 domains (14% of all domains in SCOP classification). In agreement with earlier results [1] the majority (93,5%) of domains annotated as sandwiches contains at least one interlock. There are 277 families of sandwich-like domains in SCOP 1.69, 224 of them were detected as IL+ by our approach. Additionally, 14 sandwich-like families were detected as IL+/-. Only 12 IL+ families and one IL+/- family were detected among 2614 not sandwich-like families. We conclude, that (i) interlock is characteristic sign of sandwich-like domains; (ii) there exist relatively small number of families, annotated in SCOP 1.69 as belonging to sandwich-like fold, such that all their representatives are interlock-free domains. Functional and evolutionary relations between IL+ and IL- sandwich-like families remains unclear.

The work was partially supported by Russian Foundation for Basic Research (grants No. 06-04-49558 and 06-07-89143) and INTAS grant 05-1000008-8028.

1. A.E. Kister et al. (2002) Common features in structures and sequences of sandwich-like proteins, *PNAS*, **99**: 14137–14141.

WATER-MEDIATED INTERACTIONS BETWEEN MACROMOLECULES

EVGENIY AKSIANOV[1], ANDREI ALEXEEVSKI[1], SERGEI SPIRIN[1],
OLGA ZANEGYNA[2], ANNA KARYAGINA[3]

The protein-nucleic acids (NA) interaction is usually characterized by hydrogen bonds (H-bonds) and hydrophobic interactions. Additionally, it was shown in a number of examples [1] that water-mediated bonds (WMBs) are also observed in X-ray complexes. In those cases a water molecule forms at least one H-bond with a protein donor or acceptor atom and at least one H-bond with a DNA atom. The aim of our work is to investigate the spread of WMBs among different families of NA-protein complexes.

WMBs are easily detectable in a single 3D NA-protein complex. Unfortunately, the reliability of water molecules in X-ray solved structures is less than for protein atoms. Thus a WMB observed in only one structure can be an experimental mistake or a specific feature of a particular crystal structure, which does not reflect in vivo and in vitro complex formation. So, only conserved water mediated bonds (CWMBs), i.e., the WMBs that were observed in a lot of complexes, should be used for an analysis.

It was shown that conserved water molecules correspond to immobilized water molecules detected by other methods (NMR, molecular dynamics) [2]. It is reasonable that conserved water bridges between proteins and NAs correspond to the most stable water-mediated links in an NA-protein complex.

A special procedure to inspect all available structural information and find all CWMBs between NAs and protein domains were developed. First, WMBs are detected in all available NA-protein complexes. Second, sequences of all proteins from the same SCOP family are pairwise aligned and regions of reliable alignment are detected. WMBs H-bonded with the correspondent atoms of aligned amino acid residues are considered as hypothetically aligned water

[1] Moscow State University, A.N. Belozersky Institute of Physico-Chemical Biology, Vorobiovy gory, Moscow, Russia evaksianov@belozersky.msu.ru
[2] Moscow State University, Bioengineering and Bioinformatics faculty, Vorobiovy gory, Moscow, Russia zanolya@yandex.ru
[3] N.F. Gamaleya Research Institute of Epidemiology and Microbiology, Institute of Agricultural Biotechnology, Moscow, Russia akaryagina@gmail.com

bridges. Additional verification is applied if WMBs are connected with amino acid residues not from the reliable regions. The pairs of hypothetically aligned water molecules for each family are used to find the hypothetical CWMBs (hCWMBs, a set of molecules aligned to each other) by an exhaustive search. To verify the found hCWMBs, structures from the family are superimposed using SSM program, and the wLake program [3] is used to detect clusters of aligned water molecules from different structures. After that, the clusters from protein-NA interfaces can be compared with the detected hCWMBs.

Totally 167 SCOP families of NA-binding domains were analyzed. 87 of them are represented by 10 or more structures from PDB files containing NA as well as water molecules. In 68 of 167 families (and in 35 of 87 families that contain over than 10 structures) hCWMBs presented in 10 or more structures were detected. Those results are used as a "guide" for a detailed analysis of the selected families. The complete list of observed hCWMBs is available at http://monkey.belozersky.msu.ru/~evgeniy/hcwmbs/index.html.

For example, in the family of Z-DNA binding domains 6 of 8 known structures contain both water and DNA molecules. We have detect 4 hCWMB presented in 4-5 structures each. Verification using a structure superimposition and wLake program showed that the 1st and the 2nd hCWMBs are overlapped and correspond to the same CWMB on the DNA-protein interface detected by wLake program. Similarly, the 3rd and the 4th hCWMBs form the second CWMB. Thus, hCWMBs detected by our method are very relative to the real CWMB given from the analysis of wLake results.

We conclude that some other detected hCWMBs can correspond to real conserved bonds. Thus CWMBs are rather wide-spread through protein-NA complexes. The results of the automatic analysis will be used for a manual annotation of protein-NA complexes.

This work was supported by RFBR grants 06-04-49558 and 06-07-89143 and INTAS grant 05-1000008-8028.

1. John WR Schwabe (2002) The role of water in protein—DNA interactions, *Curr Opin Struct Biol.*, **7**(1): 126-134.
2. A. Karyagina et al. (2005) The role of water in homeodomain-DNA interaction. In *Bioinformatics of Genome Regulation and Structure II*, N. Kolchanov and R. Hofestaedt (eds), 247-257 (Springer Science+Business Media).
3. B.P. Schoenborn et al. (1995) Hydration in protein crystallography, *Prog Biophys Mol Biol.*, **64**: 105-119.
4. E. Aksianov et al. (2006) A tool for comparative analysis of solvent molecules in PDB structures, In *Proceedings of the BGRS-2006*, 223-226.

PREDICTION OF PROTEIN FUNCTION BASED ON LOCAL SEQUENCE PROJECTION ALGORITHM

KIRILL ALEKSANDROV, B.N. SOBOLEV, A.E. FOMENKO, D.A. FILIMONOV, A.A. LAGUNIN, V.V. POROIKOV

Recently was obtained data about protein sequences of different organisms, including *Homo sapiens*; however functions of these proteins are unknown. Therefore, the functional annotation of amino acid sequences is one of the most important problems of bioinformatics. Different programs were successfully applied for recognition of some functional classes, nevertheless many functional groups still not predicted with required accuracy.

It is obvious, that the best prediction results can be obtained when a particular sequence is presented by the set of ordered unique descriptors. The sequential descriptors are required that represent ordered conserved fragment of any length and can be quickly calculated. We propose a Local Similarity Projection (LSP) algorithm. Each sequence from the training set is compared with the query sequence: the similarity scores are calculated for all query sequence position. Positional scores are used as descriptors weights in the recognition procedure. The suggested algorithm has the significantly more performance than the alignment methods using in addition more detailed data on the local similarity.

The LSP method was tested vs. three evaluation sets. The first set presented the serine proteinases (EC 3.4.21.X). Both tetrapeptide vocabularies and LSP method showed practically 100% recognition at the highest enzyme specificity level. The second set presented the superfamily of cytochromes P450. In this case one protein can interacted with many ligands and functional classes defined by substrate, inductor or inhibitor specificity are intersected. Phylogenetic clusters not always correspond to functional groups [2]. Substrates and inducers are better recognized for larger groups: the clear trend was shown for peptide vocabulary and LSP. Prediction for inhibitors was less accurate. The third set contains sequences from so-called “golden standard” [1] — the set of amino acid sequences with experimentally established functions.

Suggested method revealed the effective predictions with different sequence descriptions. Encouraging results were obtained for different types of functional classes.

1. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC. A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol.* 2006;7(1):R8.

Institute of Biomedical Chemistry of Rus. Acad. Med. Sci, Pogodinskaya Street, 10, Moscow, Russia 119121, dzimmu@yandex.ru

2. Yu.V.Borodina et al. (2003) If there exists correspondence between similarity of substrates and protein sequences in cytochrome P450 superfamily? *Nova Acta Leopoldina*., **87:** 47-55.

NETWORKS OF FUNCTIONAL COUPLING IN EUKARYOTES

ANDREY ALEXEYENKO

FunCoup (http://www.sbc.su.se/~andale/funcoup.html) is a statistical framework of data integration for finding functional coupling (FC) between proteins. It is capable of transferring information from model organisms (M. musculus, D. melanogaster, C. elegans, S. cerevisiae etc.) via orthologs found by InParanoid program (Remm et al., 2001). Data of different sources and various natures (contacts of whole proteins and individual domains, mRNA co-expression, protein co-occurrence in tissues and cellular compartments, similar phylogenetic profiles etc.) are collected and probabilistically evaluated in a Bayesian network (BN), trained on sets of known FC cases (e.g. KEGG, HPRD, GRID resources) vs. sets of randomly picked protein pairs as background reference. As a result of the integration, the confidence of individual links is drastically increased compared to single source based networks. To address known drawbacks of Bayesian estimators and genomic data integration, FunCoup has optimized several aspects:

Automatic discretization of continuous features as input for "data->likelihood" mapping;

Built-in confidence check while estimating likelihoods;

Choosing among alternative values from multiple pairs of co-orthologs;

Metrics for comparing mRNA expression profiles, sub-cellular co-localization, phylogenetic profiles across eukaryotic organisms;

Handling mutually redundant evidence with multivariate analysis;

Differential BN training on FC sets of different types (e.g. physical interactions, metabolic pathway links, signaling links) and then specific finding respective FC links (the multinet configuration, Friedman et al., 1997).

Compared to previous framework configurations of this sort (Suthram et al., 2006), the net gain in performance is tens of percentage points in either sensitivity or specificity. The number of simultaneously used model organisms (5-8) and individual datasets (30-50) has been estimated as maximal for practical purposes. It means that no significant further gain is expected given the current state of high-throughput data. However, novel approaches and high-

Stockholm Bioinformatics Center, Sweden, andale@sbc.su.se

throughput technologies in genomics and proteomics may well deliver new orthogonal datasets that will boost the performance of FunCoup. A comparable effect is expected of statistical post-processing of the collected evidence.

For a user of FunCoup, it is essential to know how likely a predicted functional link is to be true. Such confidence estimates as a calibrated form of the positive predictive value is to be provided. Despite the obstacles common for confidence estimates of integrated data, the new metric possesses many wanted features and is believed to be less biased. Each link in FunCoup thus contains information about underlying evidences and a confidence value.

FunCoup is a self-consistent framework that easily incorporates nearly any kind of data (continuous values of any distribution shape, binary data, character labels etc.) from any data source without human curation. It has thus been possible to generate networks for several organisms in respect of different types of functional coupling. A network for Ciona intestinalis, which had neither training sets nor sources of its own data, was created as well. In Ciona, the training set needed for supervised learning was obtained by extrapolating KEGG pathway members of organisms characterized in KEGG via MultiParanoid (Alexeyenko et al., 2006; http://www.sbc.su.se/~andale/multiparanoid/html/index.html) clusters of orthologs {human, Ciona, D. melanogaster, C. elegans}. Then evidence of FC was transferred (similarly to the other species' networks) from the better characterized model organisms. Understandably, it was limited to genes with orthologs to at least one eukaryote (7600 out of 10500 ENSEMBL 'high quality' gene models). As an independent validation, a significant part of FC links described in the comprehensive review of the Ciona embryonic development circuit (Imai et al., 2006) was successfully recapitulated by FunCoup.

The new networks for human, mouse, rat, worm, fly, Ciona, Arabidopsis, and yeast have been made available on the FunCoup website, which has also acquired a spectrum of visual (due to Medusa network applet – Hooper and Bork, 2005) and download functionalities: http://www.sbc.su.se/~andale/funcoup.html

1. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics, 2006, 22: e9-e15.
2. Friedman N., Geiger D., Goldszmidt M. Bayesian network classifiers. Machine Learning, 29, 131–163 (1997).
3. Hooper SD, Bork P. Medusa: a simple tool for interaction graph analysis. Bioinformatics. 2005 Dec 15;21(24):4432-3. Epub 2005 Sep 27.

4. Imai KS, Levine M, Satoh N, Satou Y. Regulatory blueprint for a chordate embryo. Science. 2006 May 26;312(5777):1183-7.
5. Remm M., Storm C.E., and Sonnhammer E.L.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol., 2001. 314: 1041-1052.
6. Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T. A direct comparison of protein interaction confidence assignment schemes. BMC Bioinformatics. 2006 Jul 26;7:360.

ASSOCIATIVE NETWORK DISCOVERY (AND) – SOFTWARE PACKAGE FOR AUTOMATED RECONSTRUCTION OF MOLECULAR-GENETIC ASSOCIATION NETWORKS

EWGENIA AMAN[1], PAVEL DEMENKOV[2], ARTEM NEMIATOV[3], VLADIMIR IVANISENKO[4]

The number of publications concerning biology, medicine and biotechnology grows dramatically over the years therefore it becomes virtually impossible to analyze available information for research and application purposes without automated analysis based on computer technologies. Development of information-computer software for automated operating and data extraction from text and factographic databases in the area of molecular biology, biotechnology and medicine is one of the most promising trends in systems biology.

To solve the problem of extracting data about molecular-genetic object interactions from texts the Associative Network Discovery (AND) software was developed. The AND allows to automatically reconstruct the networks of molecular-genetic interactions based on text- and data-mining methods [1].

AND consists of linguistic analysis module, association knowledgebase and visualization tool for reconstruction of associative networks. The linguistic module uses synonym dictionaries of molecular-genetic object names for extracting information about associations between proteins, genes, microRNA, substances and diseases from PubMed abstracts. The obtained data is stored in AND knowledgebase.

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia aman@bionet.nsc.ru
[2] Sobolev Institute of Mathematics SB RAS, Institute of Cytology and Genetics SB RAS
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Novosibirsk State University, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia salix@bionet.nsc.ru

We parsed 8 114 444 abstracts from PubMed database for the period from 1990 to 2006. Based on this texts 2497567 associations were extracted.

To estimate the accuracy of information extraction from text we compared the expert built gene network of NF-kB activation with NF-kB associative network. 89% common objects and 59% common interactions were identified among these networks.

AND knowledge base integrates information about associations extracted from texts with data about molecular-genetic interactions extracted from factographic databases like KEGG[2], IntAct[3], TRRD [4] among others.

The AND system can be applied for solving the wide range of problems concerned with systems biology, biomedicine and biotechnology: expanding of expert built gene networks, search of associations between gene networks and diseases, discovery of molecular mechanisms of pathology associations, search of candidate genes for genotyping assay and interpretation of microarray analysis results.

Work was supported in part by Russian Foundation for Basic Research № 05-04-49283-a, the CRDF Rup2-2629-NO-04 and № RUX0-008-NO-06, Interdisciplinary integrative project for basic research of the SB RAS № 49, and Grant for support of Leading Science Schools SS-4413.2006.1

1. S. Ananiadou, D.B. Kell, J. Tsujii (2006) Text mining and its potential applications in systems biology, *Trends Biotechnol.,* **24:** 571-579.
2. M. Kanehisa et al. (2006) From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res.* **34**, D354-357.
3. S. Kerrien et al. (2006) IntAct – Open Source Resource for Molecular Interaction Data, *Nucleic Acids Research*; **35**(Database issue): D561-565
4. N.A. Kolchanov, et al. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002, *Nucleic Acids Res.* **30:** 312-317

CONFORMATIONAL PECULIARITIES OF THE HIV-1 GP120 V3 LOOP IN THE HIV-RF AND HIV-THAILAND STRAINS

A. M. ANDRIANOV

The purpose of this work was to determine the local structure of the V3 loop of the virus strain HIV-RF and to compare its conformational characteristics with geometrical parameters of the homologous fragment of the HIV-Thailand gp120 protein computed earlier [1] using NMR spectroscopy data.

Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus, Kuprevich St., 5/2, 220141 Minsk, Republic of Belarus, andrianov@iboch.bas-net.by

The local structure of the HIV-RF V3 loop was determined by the NMR-based approach realized in a CONFNMR-2 computer program [1] and using a probability model of the protein conformation and a direct calculation of weighted average values of the molecule dihedral angles.

On analyzing the dihedral angles of the HIV-RF V3 loop, two overlapping β-turns III have been identified on its N-terminus (residues 3–7) converting to an elongated segment 9–14. According to the data obtained, certain conformers with the folded peptide backbone can be most probably realized in the central stretch of the loop (residues 15–20) belonging to the immunogenic crown of HIV-1. This stretch of the HIV-RF V3 loop has features of a metastable peptide producing an ensemble of structures, which, in addition to the dominating conformation (a combination of the inverse γ-turn with the β-turn IV), also contains minor conformations. Values of the internal rotation angles of amino acid residues in the C-terminal region of the HIV-RF loop V3 (residues 29–33) indicate that in aqueous solution this fragment forms a convoluted structure.

Comparative analysis of the secondary structures of the V3 loop in the HIV-RF and HIV-Thailand strains shows that variability of the amino acid composition of the gp120 protein does not cause essential reorganization of the loop structure. The data derived make it clear that the secondary structure elements found in the N-terminus and in the central stretch of the HIV-RF V3 loop are virtually preserved in the analyzed region of the HIV-Thailand gp120 protein. Minor changes observed in this region and associated with transformation of the overlapping β-turns III-III (HIV-RF) into a sequence of two β-turns (HIV-Thailand) and a slight enlargement of the elongated segment do not affect the structure of the virus 15–20 immunogenic crest. Close spatial folds of the main chain are also observed in segment 21–23, adjacent to the crown from the C-terminus; according to calculations, in the HIV-Thailand strain this segment forms a coil of helix 3_{10}, whereas a structure of the β-turn III presenting a fragment of the distorted α-helix is realized in this stretch of the HIV-RF V3 loop.

Our observations are inconsistent with the literature data on the conformational hyper-variability of the V3 loop in aqueous solution and suggest a possibility of conservation of some elements of its secondary structure in different HIV-1 isolates. Among the secondary structure elements common for the virus strains HIV-RF and HIV-Thailand, one needs to point out the β-turn III located in stretch 4–7 of the V3 loop and including a potential site of the gp120 protein N-linked glycosylation, which is used by the virus to strengthen the infectivity and defense against neutralizing antibodies. It should be noted that the β-turn III in the gp120 stretch under analysis was detected in work [2] during studies of conformational features of the V3 loop in the virus strain HIV-MN.

In that way, the calculations based on NMR data resulted in determination of structural elements of the V3 loop common for the two isolates of HIV-1, and these elements seem to be promising targets for realization of protein engineering projects designed for creation of drugs for prevention and therapy of AIDS.

The author appreciates the Belarusian Republican Foundation for Basic Research for financial support (project No X06-020).

1. A.M.Andrianov (2002) Local structural properties of the V3 loop of Thailand HIV-1 isolate, *J. Biomol. Struct. Dynam.,* **19:** 973-990.
2. A.M.Andrianov (1999) Global and local structural properties of the principal neutralizing determinant of the HIV-1 envelope protein gp120, *J. Biomol. Struct. Dynam.,* **16:** 931-953.

STRUCTURAL ANALYSIS OF THE HIV-1 GP120 V3 LOOP: APPLICATION TO THE HIV-HAITI ISOLATES

A. M. ANDRIANOV

The high-resolution 3D structure model of the HIV-Haiti V3 loop in water was generated in terms of NMR spectroscopy data by computer modeling method based on a “bottom-up” strategy for protein structure determination [1]. To reveal a common structural motifs occurring within V3 regardless of its environment variability, the simulated structure was collated with the one calculated previously [1] for the HIV-Haiti V3 loop in a water/trifluoroethanol (TFE) mixed solvent.

Comprehensive analysis of the dihedrals for the HIV-Haiti V3 loop in aqueous solution allows one to identify three extended β-segments (residues 2-4, 12-14, and 32-34), two stretches of distorted α-helix as well as three β-turns one of which is located in site with residues 4-7, whereas the rest of the two β-turns take up positions in central region 15-20. The values of dihedrals for the HIV-Haiti V3 loop amino acids located in segments 10-12 and 21-25 show that in water they adopt an unordered structure.

Examining the local structure of the HIV-Haiti V3 loop in a water/TFE mixed solvent reveals that altering the fragment medium results in its considerable structural conversion. Region 7-14, constituting in water the combination of helical, unordered, and extended segments, develops into the lengthy β-stretch. While, replacing the solvent stimulates the forming of the right-handed α-helix in segment 31-34, which conforms to our earlier findings [2] indicating that this

Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus, Kuprevich St., 5/2, 220141 Minsk, Republic of Belarus, andrianov@iboch.bas-net.by

V3 loop stretch longs for the coiled structures. Addition of TFE affects also the central region of the HIV-Haiti V3 loop (hexapeptide Gly-Pro-Gly-Lys-Ala-Phe) that determines the specificity of the virus binding with neutralizing antibodies: as follows from the data obtained, in a mixed water/TFE solvent this stretch makes up more compact, as compared to aqueous solution, spatial fold but its structure corresponds to the non-typical triple β-turn.

Among amino acids contributing to cell tropism, Arg-3, Pro-13, Gly-24, and Asp-25 retain their local structures, whereas for Ser-11, Ala-19, Thr-23, and Gln-32 the perceptible changes in dihedrals come to pass. In the list of the residues inclined to structural conservation, special attention must be paid to Asp-25 that is critical for the virus binding with primary cell receptor CD4 as well as to Arg-3 that is critical for utilization of CCR5 co-receptor and heparan sulfate proteoglycans. Along with Arg-3 and Asp-25, it is essential to mark out the (ϕ, ψ)-restrained residue located in position 4 of the Haiti-V3 loop that is highly conserved among CCR5-using viruses. Changes of environment do not affect the local structure of the amino acid in position 29 either, which stabilizes the V3 loop conformation and influences the intensity of the CD-4-activated gp120 protein binding with the co-receptor CCR5. Among structurally conservative amino acids, the residues in positions 10, 12, and 14 of the HIV-Haiti V3 loop should be also noted as those that significantly contribute to the interaction of the virus with the monoclonal antibody 447-52D possessing a wide spectrum of neutralizing activity. The conformationally stable amino acids of the Haiti-V3 loop should be also supplemented with segment 5-7 which includes one of the possible sites of the gp120 N-linked glycosylation.

The 3D structure model of the HIV-Haiti V3 loop built here may serve as a structural frame for computer-aided screening of the low-molecular ligands to be used as drugs against AIDS. In this case, the structurally conservative stretches of V3 may present the most suitable landing-places for molecular docking of the V3 loop and ligand structures followed by selecting the well-deserved applicants for the role of therapeutic agents.

The author appreciates the Belarusian Republican Foundation for Basic Research for financial support (project No X06-020).

1. A.M.Andrianov, V.G.Veresov (2006) Determination of structurally conservative amino acids of the HIV-1 protein gp120 V3 loop as promising targets for drug design by protein engineering approaches, Biochemistry (Moscow), 71: 906-914.
2. A.M.Andrianov (2002) Local structural properties of the V3 loop of Thailand HIV-1 isolate, J. Biomol. Struct. Dynam., 19: 973-990.

COMPARATIVE EVALUATION OF A NEW ALGORITHM OF GENERATING GAP-CONTAINING BLOCKS FROM MULTIPLE PROTEIN ALIGNMENTS

IVAN V. ANTONOV[1], ANDREY M. LEONTOVICH[1],
ALEXANDER E. GORBALENYA[2]

Depending on structural and functional roles played, different regions of proteins may evolve with considerably different rates. This link between evolution, structure and function is evident in multiple sequence alignments of proteins, with functionally and structurally important regions being relatively well conserved. In multiple sequence alignment, conserved regions are recognized as "blocks". Since 1990, when blocks were introduced [1], they have been used in constructing amino acids substitution matrices (BLOSUM) [2], alignment refining [3] and for homologs identification by scanning sequence databases. Several algorithms for deriving gap-free blocks from alignments have been proposed [1, 3, 4, 5]. These blocks may account for a sizable part of an alignment of relatively closely related proteins. However, in alignments containing numerous and distant homologs, number and size of blocks fall dramatically because of the gaps accumulation.

We have recently developed a new algorithm for generating blocks that may contain gaps. This algorithm was implemented in a program, dubbed Blocks Accepting Gaps Generator (BAGG). In this study we have compared blocks generated by two procedures: BAGG and Blocks Multiple Alignment Processor (BMAP) [5] program that is available from the Blocks server (http://blocks.fhcrc.org/blocks/process_blocks.html) and considered to be a standard for generating gap-free blocks from multiple protein alignments.

We have designed an original protocol for blocks evaluation through assessing the efficiency with which blocks identify homologous proteins in the Swissprot database. Manually curated seed alignments of protein families available from the Pfam database [6] were used as input to the BAGG and BMAP to generate blocks. Two sets of generated blocks were converted into the full HMM profiles using HMMER and these profiles were used to search SwissProt for homologs. A list of all protein hits above a threshold was compiled for each set (Hits list). It was compared, family by family, with the protein list in full PFAM

[1] Moscow State University, Lab. Bldg A, Vorobiovy Gory 1-73, Moscow 119992, Russia, pechkin@belozersky.msu.ru
[2] Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands, A.E.Gorbalenya@lumc.nl

alignments that contains proteins forming seed alignments plus homologs identified by the PFAM automatic procedure (6) (Full list). If a Swissprot hit was in a cognate protein family of the Full protein list, this hit was considered to be a true positive; otherwise it was treated as a false positive. Proteins that were in the Full, but not Hits list were considered false negative. According to this procedure, BAGG and BMAP blocks produced from protein alignments with relatively few gaps were comparable in retrieving homologous proteins from the Swissprot. However when highly diverged protein alignments were used, BAGG blocks significantly outperformed BMAP blocks. These results show that BAGG may be an efficient automatic procedure for identifying conserved regions in a wide range of protein families using protein alignments.

1. Smith, H., Annau, T., and S.Chandrasegaran (1990) Finding sequence motifs in groups of functionally related proteins, *Proc. Natl. Acad. Sci.,* **87:**826–830.
2. Henikoff, S., and J. Henikoff (1993) Performance evaluation of amino acid substitution matrices, *Proteins*, **17:**49-61.
3. Chakrabarti, S. et al. (2006) Refining multiple sequence alignments with conserved core regions, *Nucleic Acids Res.*, **34:**2598-606.
4. Henikoff, S., and J. Henikoff (1991) Automated assembly of protein blocks for database searching, *Nucleic Acids Res.,* **19:**6565-6572.
5. Henikoff, J., Henikoff, S., and S. Pietrokovski (1999) New features of the Blocks Database servers, *Nucleic Acids Res.,* **27:**226-228.
6. Robert, F. et al. (2006) Pfam: clans, web tools and services, *Nucleic Acids Res.,* **34:**D247-D251.

IMPROVING AUTOMATIC ANNOTATION OF PROTEINS BY THE NEGATIVE ASSOCIATION RULE MINING

IRENA I. ARTAMONOVA[1], GOAR FRISHMAN[2], DMITRIJ FRISHMAN[3]

The continuing reduction of sequencing costs and the success of metagenomics projects lead to exponential increase of the number of sequenced genes. The experimental studies and even manual expert annotation are much slower and thus the gap between the number of experimentally studied and all known gene products is permanently increasing. For that reason the only hope to get any information about most proteins is automatic annotation.

[1] Group of Bioinformatics, Vavilov Institute of General Genetics RAS, irenart@gmail.com
[2] Institute for Bioinformatics, GSF - National Research Center for Environment and Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany, astva@gsf.de
[3] Department of Genome Oriented Bioinformatics, Technische Universität Munchen, Wissenschaftzentrum Weihenstephan, 85350 Freising, Germany, d.frishman@wzw.tum.de

The automatic annotation based on the bioinformatics analysis is very efficient and sufficiently fast, but extremely error-prone. The most obvious and direct approach towards improving the reliability and coverage of unsupervised protein annotation entails the development of better bioinformatics tools. A complementary tactic is to improve the quality of protein sequence databases by retrospective search for errors in the total corpus of already available annotation. For such search we applied a negative association rule mining technique in addition to the previously developed positive association rule method [1].

A negative association rule is usually formulated in the form "Left-Hand-Side (LHS) implies not Right-Hand-Side (RHS)" and may be interpreted as "database entries that satisfy the LHS conditions are unlikely to satisfy the RHS condition". In the application of the association rules mining to the genome annotation we believe that if a rule has a high support (i.e., is applicable to many entries) and high strength (is satisfied by most entries), it reflects some biological regularity or maybe a peculiarity of the annotation process. Thus the exceptions to such rules may be annotation errors. Indeed, in the case of positive association rules, careful manual analysis demonstrated that about one half of exceptions to high-strength rules in the Swiss-Prot and PEDANT databases are actual annotation errors, which is significantly higher than the average several percent [1].

We applied the negative association rule technique to the analysis of the Swiss-Prot and PEDANT data. By design, negative feature combinations allow detecting only the over-annotation problems. Such problems are very rare in the case of manually curated databases, the main problem of which is under-annotation. Indeed, this approach is not effective for the curated Swiss-Prot database, and we reduced our efforts to the case of PEDANT annotation.

A large fraction (33%) of all negative association rules for the PEDANT annotation set included a taxon-specific FunCat label (e.g., fc75.03 – "animal tissue") on one side of the implication and the highest-level taxon of protein origin contradicting this specificity (in the given case, Bacteria or Archaea or Viruses for fc75.03) on the other. In theory, taxon-specific FunCat labels should only be present in the annotation of the genes belonging to the corresponding taxa. However, the homology-based transfer of such annotation attributes makes them prone to error. So if a taxonomically specific FunCat label is incompatible with the known gene taxon, it is the FunCat assignment which is guaranteed to be erroneous, since the protein origin is doubtlessly known. This simple test resulted in automatic correction of almost 50% of all exceptions in our set of strong negative rules.

To estimate the prevalence of errors among exceptions not corrected by the taxonomy procedure we selected randomly a sample of 100 rules and manually analyzed their exceptions. In 96% of the examined exceptions, at least one of the features constituting the rule was assigned wrongly to the given protein. The overall specificity of the approach was estimated to be as high as 98%: practically all feature combinations associated with exceptions included at least one annotation error. Thus the specificity of the negative rules is much higher than that in the case of positive rules, which has been estimated to be around 68% [1].

At the same time, the approach based on exceptions from strong negative rules yields much smaller coverage than positive rule mining: it allows identifying eleven-fold less annotation features that participate in incompatible feature combinations (0.6% for negative rules *versus* 6.7% for positive rules). On the other hand, this is still useful, since more than two thirds of these features do not get detected by positive rule mining.

We conclude that applying a combination of the positive rule mining and the negative rule mining represents a powerful way to enhance the fidelity of genome annotation.

This work was conducted in the framework of the BioSapiens project funded by the European Commission FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2003-503265.

1. I.I. Artamonova et al. (2005) Mining sequence annotation databanks for association patterns. *Bioinformatics*, **21:** iii49-iii57.

ANALYSIS OF SEQUENCE CONSERVATION AT THE NUCLEOTIDE RESOLUTION

SAURABH ASTHANA[1], WILLIAM S. NOBLE[2], JOHN A. STAMATOYANNOPOULOS[3], SHAMIL R. SUNYAEV[4]

It is widely assumed that human non-coding sequences comprise a substantial reservoir for functional variants impacting gene regulation and other chromosomal processes. Evolutionarily conserved non-coding sequences (CNSs) in the human genome have attracted considerable attention for their potential to simplify the search for functional elements and phenotypically important human alleles. A major outstanding question is whether functionally significant human non-coding variation is concentrated in CNSs or distributed more broadly across the genome. Here we combine whole-genome sequence data from four non-human species (chimp, dog, mouse, and rat) with recently available comprehensive human polymorphism data to analyze selection at single nucleotide resolution. We show that a substantial fraction of active purifying selection in non-coding sequences occurs outside of CNSs and is diffusely distributed across the genome. This suggests the existence of a large complement of human non-coding variants that may impact gene expression and phenotypic traits, the majority of which will escape detection using current approaches to genome analysis.

We further introduce a new computational method - SCONE (Sequence CONservation Evaluation) - for scoring evolutionary conservation at individual base pair resolution as well as at the level of sequence regions. SCONE estimates the rate at which each nucleotide position is evolving and computes a p-value for neutrality for the given rate estimate. We apply SCONE to multiple sequence alignment of 23 mammalian genomes available for 1% of genomic sequence. We find a clear relationship at the nucleotide level between SCONE scores and the allele spectrum of human polymorphisms in non-coding regions. We also examined the distribution of conservation scores for experimentally

[1] Biological and Biomedical Sciences Program, Harvard Medical School, HMS NRB, 77 Avenue Louis Pasteur, Boston MA 02115, USA, faplap@gmail.com
[2] Dept. of Genome Sciences, Univ. of Washington, 1705 NE Pacific Street, Seattle, WA 98195, USA, wnoble@u.washington.edu
[3] Harvard Medical School, HMS NRB, 77 Avenue Louis Pasteur, Boston MA 02115, USA, jstam@u.washington.edu
[4] Genetics Division, Brigham & Women's Hospital, Harvard Medical School, HMS NRB, 77 Avenue Louis Pasteur, Boston MA 02115, USA
ssunyaev@rics.bwh.harvard.edu

identified functional elements. Functional elements display an excess of conserved positions not embedded in long conserved regions. These positions are non-randomly distributed along the sequence.

The analysis of human polymorphism and functional features suggests that the majority of functionally important non-coding conserved positions reside outside of long conserved regions.

COMPARATIVE GENOMIC HYBRIDIZATION ANALYSIS OF DIVERSITY IN *LACTOCOCCUS LACTIS* STRAINS

J. BAYJANOV*, D. MOLENAAR‡, J. VAN HYLCKAMA VLIEG†‡, R.J. SIEZEN*†‡

High-throughput techniques like Comparative Genomic Hybridization (CGH) arrays help to understand genetic changes in closely related bacterial strains. Such experiments elucidate how strains are genetically related. The correlation of experimental results from CGH arrays and phenotypic information about strains helps to identify roles of genes in the generation of phenotypic traits (1). Trait-genotype correlations open more insights into how strains are diversified to survive and compete in the environment in which they grow (2).

We designed arrays containing 3.8×10^5 probes targeting proprietary and publicly available complete and incomplete *L. lactis* genome sequences. DNA from 40 different *Lactococcus lactis* strains was hybridized with these arrays. The signal variation in these CGH data is much higher than in usual CGH experiments due to the diversity of strains and the extremely high flexibility of bacterial genomes. Therefore, the interpretation of these CGH data requires novel tools and analyses. The CGH data were stored in a database together with sequence annotation data. The raw CGH data needs normalization to reduce systematic error caused by spatial features on the array. This was achieved by kernel smoothing. Strains were compared using their log ratio of normalized signal intensity values to the signal intensity values of well-studied strains. Signal intensity ratios are plotted against genome position of well-studied strains to screen DNA deletions and copy number changes, which is helpful to find the correlation between changes at genomic level and phenotypic traits of strains. Comparing strains for the presence/absence of genes will help to understand roles of genes in diversification of strains. A threshold value is determined us-

* Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Center, The Netherlands, J.Bayjanov@cmbi.ru.nl
† TI Food and Nutrition, Wageningen, The Netherlands.
‡ NIZO Food Research BV, Ede, The Netherlands.

ing information about presence/absence of genes in sequenced strains, and a gene is considered to be present in a strain if it has higher signal intensity than the threshold, otherwise absent. This method gave highly accurate results, which have been verified using genome annotation of sequenced strains.

1. Pretzer G, Snel J, Molenaar D, Wiersma A, Bron PA, Lambert J, de Vos WM, van der Meer R, Smits MA, Kleerebezem M. Biodiversity-Based Identification and Functional Characterization of the Mannose-Specific Adhesin of Lactobacillus plantarum. *J. Bacteriol. 2005 Sep;* **187(17)**: 6128-36.
2. Molenaar, D., F. Bringel, F. H. Schuren, W. M. de Vos, R. J. Siezen, and M. Kleerebezem. Exploring Lactobacillus plantarum genome diversity by using microarrays. *J. Bacteriol. 2005 Sep;* **187(17)**: 6119-27.

EXTENSIVE PARALLELISM IN PROTEIN EVOLUTION

GEORGII A. BAZYKIN[1], FYODOR A. KONDRASHOV[2], MICHAEL BRUDNO[3], ALEXANDER POLIAKOV[4], INNA DUBCHAK[5], ALEXEY S. KONDRASHOV[6]

Independently evolving lineages mostly accumulate different changes, which leads to their gradual divergence. However, parallel accumulation of identical changes is also common, especially in traits with only a small number of possible states. We describe parallelism in evolution of coding sequences in three four-species sets of genomes of mammals, *Drosophila*, and yeasts. Each such set contains two independent evolutionary paths, I and II. An amino acid replacement which occurred along path I also occurs along path II with the probability 50-80% of that expected under selective neutrality. Thus, the per site rate of parallel evolution in proteins is several times greater than their average rate of evolution, but still lower than the rate of evolution of neutral sequences. This deficit may be caused by changes in the fitness landscape, leading to a replacement being possible along path I but not along path II. However,

[1] Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA gbazykin@princeton.edu
[2] Section on Ecology, Behavior and Evolution, University of California at San Diego, La Jolla, CA 92093 USA
[3] Department of Computer Science and Banting & Best Department of Medical Research, University of Toronto, Toronto ON M5S 3J4 Canada
[4] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720 USA
[5] Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598 USA
[7] Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109-2216, USA kondrash@umich.edu

invariant weak selection assumed by the nearly neutral model of evolution appears to be a more likely explanation. Then, the average coefficient of selection associated with an amino acid replacement, in the units of the effective population size, must exceed ~0.4, and the fraction of effectively neutral replacements must be below ~30%. At a majority of evolvable amino acid sites, only a relatively small number of different amino acids is permitted.

MOLECULAR ASPECT OF THERMOPHILIC ADAPTATION

IGOR N. BEREZOVSKY

Exhaustive evaluation of all combinations of amino acids reveals a universalproteomic predictor of Optimal Growth Temperature in prokaryotes [1]. Whatmechanism does Nature use in her quest for thermophilic proteins?Positive and negative design [2] broaden the energy gap between native andmisfolded conformations in proteins, the main determinant of protein stability.The components of design are responsible for "from both ends of hydrophobicityscale" trend observed in thermophilic adaptation, whereby proteomes ofthermophilic proteins are enriched in hydrophobic and charged residues at theexpense of polar ones. Hydrophobic residues contribute mostly to the positivedesign, while repulsion between charged residues in non-native conformations ofproteins contributes to negative design [2].The frequency with which A and G nucleotides appear as nearest neighbors ingenome sequences is strongly and independently correlated with Optimal GrowthTemperature and points to the stacking as the major contributor to thethermostabilization of genomic DNA [1].

1. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. (2007) PLoSComput Biol 3, e52.
2. Berezovsky, I. N., Zeldovich, K. B. & Shakhnovich, E. I. (2007) PLoSComput Biol 3, e52.

Department of Chemistry and Chemical Biology, Harvard University,
12 Oxford Street Cambridge MA 02138, USA inberez@fas.harvard.edu

MATHEMATICAL MODELING OF THE HCV DRUGS COMBINATIONS EFFECT

K.D. BEZMATERNYKH, E.L. MISHCHENKO, V.A. IVANISENKO, V.A. LIKHOSHVAI

At the start of the 21st century, hepatitis C virus (HCV) remains a serious global health concern. HCV infection is the most common blood-born infection and a major cause of chronic liver desease in developed countries. According to worldwide estimates, up to about 2-3% of the human population is infected with HCV. The infection does not usually resolve, and about 80% of acute infections persist. Chronic HCV infection can cause progressive fibrosis of the liver, leading to cirrhosis and liver carcinoma [1, 2]

To date there are only two antiviral agents licensed for the treatment of HCV infection, interferon-alpha and ribavirin. The use of these agents in combination, and the modification of interferon-alpha with polyethylene glycol, has lead to the clearance of HCV in many patients. However, these agents are associated with significant side effects and are far from universally efficacious [3].

Nowadays there are presented many HCV NS3 protease and RNA-dependent RNA polymerase NS5B inhibitors as the new drugs against HCV. The estimation of the drug effect, the prediction of inhibitor's action in case of virus mutations and finding optimal cure strategy is possible with the help of mathematical modeling of the HCV inhibitors action.

The kinetics of the HCV RNA concentration in the presence of HCV NS3 protease inhibitors, HCV NS5B polymerase inhibitors and cellular factor inhibitor were calculated using the model of subgenomic HCV RNA replication in cell culture [10]. The kinetics of the HCV RNA concentration in the presence of 2 drugs combinations were obtained, combinations with synergetic effect were revealed. The dependences of the minimal cure time needed for full virus clearance from the value of inhibitory constant were obtained for each inhibitor. Based on these results we can recommend new treatment strategy against HCV for verification using experimental cells or animal model.

1. J.H.Hoofnagle (2002) Course and outcome of hepatitis C, *Hepatology*. **36:** S21-S29.
2. G.Dusheiko et al. (2000) The science, economics, and effectiveness of combination therapy for hepatitis C, *Gut*. **47:** 159-161.
3. J.C.McHutchison, K. Patel (2002) Future therapy of hepatitis C, *Hepatology* **36:** S245-S252.

Institute of Cytology and Genetics SB RAS, Novosibirks, Russia,
bezmate@bionet.nsc.ru

4. A.Pause et al. (2003) An NS3 serine protease inhibitor abrogates replication of subgenomic hepatitis C virus RNA, *J. Biol. Chem.* **278:** 20374-20380.
5. C.Steinkuhler et al. (2001) Hepatitis C virus serine protease inhibitors: current progress and future, *Curr. Med. Chem.* **8:** 919-932.
6. L.J.Stuyver et al. (2003) Dynamics of subgenomic hepatitis C virus replicon RNA levels in Huh-7 cells after exposure to nucleoside antimetabolites, *J. Virol.* **77:** 10689-10694.
7. Y.H.Koh et al. (2005) Design, synthesis, and antiviral activity of adenosine 5'-phosphonate analogues as chain terminators against hepatitis C virus, *J. Med. Chem.* **48:** 2867-2875.
8. V.K.Johnston et al. (2003) Kinetic profile of a heterocyclic HCV replicon RNA synthesis inhibitor, *Biochem. Biophys. Res. Commun.* **311:** 672-677.
9. L.Tomei et al. (2004) Characterization of the inhibition of hepatitis C virus RNA replication by nonnucleosides, *J. Virol.* **78:** 938-946.
10. E.L. Mishchenko et al. (in press) Mathematical model for suppression of subgenomic hepatitis C virus RNA replication in cell culture. *J Bioinform Comput Biol.*

NETWORK ALIGNMENT TOOLS FOR NOVEL INSIGHT IN CELLULAR MACHINERY

ANUP BHATKAR[1], GAUTAM LIHALA[1], MAHESH GUPTA

Abstract. Molecular networks represent the backbone of cellular activity within the cell. Research has revealed that protein–protein interaction (PPI) networks evolve at a modular level having scale free topology. As the amount of available data on these networks increases, discovery of conserved patterns in these networks becomes an important problem. Recent studies have taken a comparative approach toward interpreting these networks, contrasting networks of different species and molecular types, and under varying conditions. Many of the methodological and conceptual advances that were important for sequence comparison will likely also be important at the network level, including improved search algorithms, techniques for multiple alignment and better integration with public databases. In this review, we survey the field of comparative biological network analysis and describe its applications to elucidate cellular machinery and to predict protein function and interaction.

1. Kelley, B.P. et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. USA 100, 11394–11399

[1] Maulana Azad National Institute of Technology,Bhopal, India
anup801@gmail.com, gautamlihala@yahoo.com, mahesh_nitb@yahoo.com

2. Rhodes, D.R. et al. (2005) Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. 23, 951–959
3. Kelley, R. & Ideker, (2005) T. Systematic interpretation of genetic interactions using protein networks. Nat. Biotechnol. 23, 561–566
4. Zhang, L.V. et al. Motifs (2005) themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J. Biol. 4, 6
5. Thsato, Y., Matsuda, H. & Hashimoto, A. (2000) in Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB) 376–383
6. Barabási A., and Albert, R. .(1999) Emergence of scaling in random networks. Science 286, 509–512.
7. Qin, H., Lu, H.H.S., Wu, W.B., and Li, W. (2003) Evolution of the yeast protein interaction network. PNAS 100(22),12820–12824.
8. Koyuturk M., Kim Y., Topkara U., Subramaniam S., Szpankowski W., Grama(2006) A. Pairwise Alignment of Protien Interaction Networks. Journal of Computational Biology13, 182-199
9. Vázquez, A., Flammini, A., Maritan, A., and Vespignani, A.(2003) Modeling of protein interaction networks. ComPlexUs 1, 38–44.
10. Flannick J., Novak A., Srinivasan B.S., Harley H. McAdams and Batzoglou S. Græmlin(2006).: General and robust alignment of multiple large interaction networks, Genome Res. 16, 1169-1181
11. Altschul, S.F., Carroll, R.J., and Lipman, D.J.(1989) Weights for data related by a tree. J. Mol. Biol. 207: 647–653.
12. Tatusov, R.L., Koonin, E.V., and Lipman, D.J.(1997) A genomic perspective on protein families. Science 278: 631–637.
13. Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., and Ideker, T.(2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. 100: 11394–11399.
14. Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., and Ideker, T. (005) Conserved patterns of protein interaction in multiple species. Proc. Natl. Acad. Sci. 102: 1974–1979.
15. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249–255
16. Tohsato, Y., Matsuda, H. & Hashimoto, A. (2000). in Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB) 376–383
17. Gunsalus, K.C. et al. (2005).Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature 436, 861–865
18. Kemmeren, P. et al. (2002). Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol. Cell 9, 1133–1143
19. Rhodes, D.R. et al. (2005). Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. 23, 951–959

20. Jansen, R. et al. (2003).A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302, 449–453
21. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. Science 306, 1555–1558
22. Lu, L.J., Xia, Y., Paccanaro, A., Yu, H. & Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. Genome Res. 15, 945–953
23. Wong, S.L. et al. (2004). Combining biological networks to predict genetic interactions. Proc. Natl. Acad. Sci. USA 101, 15682–15687
24. Yeger-Lotem, E. et al. (2004).Network motifs in integrated cellular networks of transcription regulation and protein-protein interaction. Proc. Natl. Acad. Sci. USA 101, 5934–5939
25. Balaji S. Srinivasan, Antal F. Novak, Jason A. Flannick, Serafim Batzoglou, and Harley H.McAdams Integrated Protein Interaction Networks for 11 Microbes.
26. Pinter, R.Y., Rokhlenko, O., Yeger-Lotem, E. & Ziv-Ukelson, M. (2005).Alignment of metabolic pathways. Bioinformatics 21, 3401–3408

P-VALUE CALCULATION FOR HETEROTYPIC CLUSTERS AND ITS USE IN COMPUTATIONAL ANNOTATION OF REGULATORY SITES

VALENTINA BOEVA[1], J. CLEMENT[2], M. REGNIER[3],
VSEVOLOD J. MAKEEV[1,4]

Assessing statistical significance of multiple motif occurrences in the text is a common problem in computational biology, e.g. in finding of *cis*-regulatory modules (CRM) in genomes [1]. Here, the main difficulty comes from overlapping occurrences. So far, no tools have been developed allowing computing P-values for simultaneous occurrences of different motifs with overlaps. Here we present an algorithm, that computes the P-value to find $n_1,\ldots,n_k$ possibly overlapping occurrences of k different motifs in a random text. Motifs can be represented with a majority of popular motifs models without indels. In our implementation we included such motif models as a list of allowed words (the putative binding sites), Position Weight Matrix (PWM), IUPAC consensus and word with k mismatches. Zero or first order Markov chain can be adopted for the text. Our algorithm is inspired by Aho-Corasick automaton [2] and employs a prefix tree with suffix links. The algorithm runs with $O(N|S|)$ time complexity,

[1] GosNIIgenetika, Moscow, Russia, valeyo@imb.ac.ru
[2] GREYC, CNRS UMR 6072, Caen, France, Julien.Clement@info.unicaen.fr
[3] INRIA , Rocquencourt, France, mireille.regnier@inria.fr
[4] Engelgardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, makeev@genetika.ru

where N is the length of the text and $|S|$ is the number of the states of our automaton. The latter, in turn, is upper bound by the total number of possible words allowed by any of the motifs multiplied by the length of the longest word. The primary objective of the program is to assess the likelihood that a given genome segment is a CRM regulated with the known set of regulatory factors. The program can also be used to select the cut-off for PWM scanning and to assess similarity of different motifs. **Example:** In Fig. 1 the 3D-surface is shown for –log(P-values) calculated for various cutoff values in real biological sequence of the *even-skipped stripe 2* enhancer (Fig. 1a) and in a random sequence of the same length and with the same dinucleotide distribution (Fig. 1b). We took PWMs for transcription factors *bicoid* and *kruppel* that were reported to regulate the *even-skipped stripe 2* enhancer [1]. One can see that, first, p-values in the random sequence are much greater than in the enhancer sequence; and second, the shape of P-value distributions is different. We believe that cut-off values giving the minimal P-value (the biggest peak on the surface in Fig. 1) correspond to the best candidates for TF binding sites. As we expected, it was impossible to choose the appropriate cutoff for PWMs of factors from the random sequence data Fig. 1b.

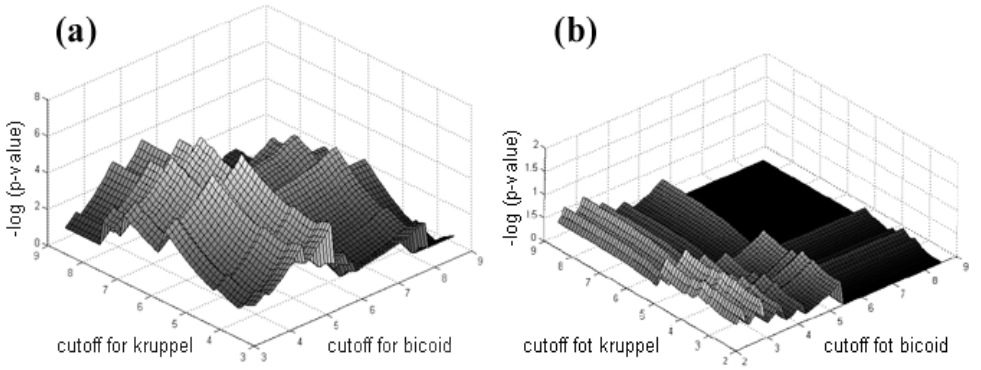


Figure 1. Distribution of $\log_{10}$P-value calculated for Markov(1) model as a function of cutoff values for PWMs for *bicoid* and *kruppel* in the *even-skipped stripe2* enhancer (a) and in a random sequence (b).

The web-page is available at http://favorov.imb.ac.ru/ahokocc/.

This work has been supported by a project EcoNet-12635WG, INTAS 04-83-3994 and INTAS 05-1000008-8028. The authors are pleased to thank Mikhail Roytberg and Andrey Mironov for helpful discussions.

1. D.A. Papatsenko, et al. (2002) Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers, *Genome Res.,* **12**(3): 470-81.
2. A.V. Aho, M.J. Corasick (1975) Efficient string matching: An aid to bibliographic search, *Communications of the ACM,* **18**(6): 333–340.

OPTIMAL WAY OF CONSIDERING INTRA-PROTEIN CONTACTS

NATALIA S. BOGATYREVA, DMITRY N. IVANKOV

A globular protein during folding goes from non-compact unfolded state to the most compact native state. Thus, the protein compactness increases, and one can consider compactness as a reaction coordinate on the route of folding. Quantitatively the increase in compactness is commonly expressed as either the decrease in accessible surface area (ASA)[1], or increase in the number of intra-protein interactions (contacts)[2], and, obviously, decrease in ASA must correlate with increase in the number of intra-protein contacts. The way of calculation ASA is rather common: it is calculated as the area of surface formed by a center of water molecule (represented by a ball of radius 1.4A), which is rolled over a protein. On the contrary, there are a number of different ways of considering intra-protein contacts. First, one can consider either atom-atom[2], or residue-residue, or C_α-atom[3] contacts. Second, the value of cutoff distance used for determining if a contact is formed or not, has wide range from 4A to 15A [ref.4].

In this work we use the idea that the change in ASA must be accompanied with the change in the number of intra-protein contacts to establish the best way(s) and parameters of considering intra-protein contacts.

For our analysis we took the set of protein domains based on 1.65 SCOP[5] release with pair-wise homology not higher than 25%. Then for each protein domain there were calculated:

1) the difference in ASA between completely extended and native protein conformations (YASARA [www.yasara.org] was used for generating completely extended conformation, ASA calculations and for addition hydrogen atoms to protein's native structure, when necessary);

2) the difference in the number of intra-protein contacts between the native and completely extended conformations. Here there were considered (i) atom-atom contacts, (ii) residue-residue contacts (i.e. two residue are in contact if they have two atoms in contact), (iii) C_α-atom contacts, and (iv) atom-atom

Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow region, Russia, bogat@phys.protres.ru, ivankov@phys.protres.ru

contacts with atom-specific cutoff value (i.e. two atoms are in contact if the distance between their van der Waals spheres is not higher than cutoff value). All contacts were calculated with and without hydrogen atoms.

We have shown that the best ways (i.e. best correlated with ASA calculations) for considering intra-protein contacts are the following (all these ways give correlation coefficients higher than 99.8% between change in ASA and change in the number of contacts):

atom-atom contacts with atom-specific cutoff value of 5.25A with hydrogen atoms taken into account (cutoff values from 4.5A to 5.5A are also good);

atom-atom contacts with cutoff value of 8A with hydrogen atoms taken into account (cutoff values from 7.25A to 8.75A are also good);

residue-residue contacts with cutoff value of 4A with hydrogen atoms taken into account (cutoff values of 4.25A is also good).

residue-residue contacts with cutoff value of 4.75A without hydrogen atoms.

We are grateful to Alexei V. Finkelstein and Oxana V. Galzitskaya for helpful discussions. This work was supported by the Russian Foundation for Basic Research (grant 07-04-01539).

1. E. Alm, D. Baker (1999) *Proc. Natl. Acad. Sci. USA*, **96**:11305-11310.
2. O.V. Galzitskaya, A.V. Finkelstein (1999) *Proc. Natl. Acad. Sci. USA*, **96**:11299-11304.
3. M.M. Gromiha, S.Selvaraj (2001) *J. Mol. Biol.*, **310**:27-32.
4. H.Zhou, Y.Zhou (2002) *Biophys. J.*, **82**:458-463.
5. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia (1995) *J. Mol. Biol.* **247**:536-540.

LIFE HISTORY OF THE SODIUM NEUROTRANSMITTER SYMPORTER FAMILY, SNF/SLC6

DMITRI Y. BOUDKO[1], ELLA A. MELESHKEVITCH[1], MELISSA M. MILLER[1], LYUDMILA B. POPOVA[1,2], BERNARD A. OKECH[1], DMITRY A. VORONOV[1,3], WILLIAM R. HARVEY[1]

To maintain homeostasis, biological systems recruit a network of ATPase pumps, ion channels, and secondary transporters. The role of the first two

[1] Whitney Laboratory for Marine Bioscience, University of Florida, 9505 Ocean Shore Blvd., St Augustine, FL, 32080, USA; boudko@whitney.ufl.edu.
[2] A.N. Belozersky Institute, Moscow State University, Moscow 119899, Russia.
[3] Institute for Information Transmission Problems Russian Academy of Sciences, Moscow 127994, Russia.

groups is to generate and regulate electrochemical membrane gradients. In contrast, the secondary transporters evolved a diversity of electrochemical gradient-coupled molecular mechanisms to balance intracellular concentrations of substrates and metabolites (43 SoLute Carrier families, SLCs; HUGO). The Sodium neurotransmitter Symporter Family (SNF a.k.a. SLC6) is one of the largest and most ancient families of secondary transporters which currently has been identified in all taxa except for plants as well as some protozoan and bacterial lineages. SNF encompasses a great diversity of transport phenotypes including sodium-dependent transporters for monoamine neurotransmitters, GABA, and some metabolic amino acids. Ongoing molecular, phylogenetic and structural study of "orphan" SLC6 members in our laboratory using comparative genomic model organisms revealed a large expansion of paralogous genes with a functional consensus in the accumulation of essential amino acids (Nutrient Amino acid Transporters, NAT subfamily of SLC6). Our data lead to several insights regarding the life history and biological role of the SNF. They suggest that metazoan NAT-SNF members evolved and acted in synergy as a key mechanism supplying essential amino acids utilizing pathways e.g. protein, neurotransmitter and hormone synthesis. The origin and set of expansions of NAT-SNF domain had dramatic impacts in the evolution. It generalized the acquisition of exogenous nitrogen/carbon rich substrates which liberate selective pressure on major metabolic pathways and led to massive loss of nitrogen fixation in prokaryotes and the extinction of essential amino acid synthesis cascades in organisms. On the other hand it facilities integration of metazoan organisms via redistribution of essential metabolites and enforces the evolution of sensory, motor and central neuronal functions that became critical to secure access of essential amino acids. Neuronal NATs provide genetic templates in the evolution of synaptic neurotransmitter transporters for monoamines, glycine and GABA neurotransmitters. The analysis of NAT functions is essential for understanding basis of somatic and symbiotic integration and genesis of multiple metabolic and neuronal disorders. It also leads to new approaches for effective suppression of disease vector, pathogen and pest organisms.

Supported by NIH R01-AI030464 (DB).

1. Boudko D.Y., Stevens B.R., Donly B.C., and Harvey W.R. (2005) Nutrient amino acid and neurotransmitter transporters. In *Comprehensive Molecular Insect Science*, vol. 4 (ed. K. I. a. S. S. G. Lawrence I. Gilbert), 255-309. Amsterdam: Elsevier.
2. Boudko, DY; Meleshkevitch, EA; Harvey, WR. (2005) Novel transport phenotypes in the sodium neurotransmitter symporter family. *FASEB J.* **19 (4)**: A748.

3. Boudko D.Y., Kohn A.B., Meleshkevitch, E.A., Dasher, M.K., Seron, T.J., Stevens, B.R. and Harvey, W.R. (2005) Ancestry and progeny of nutrient amino acid transporters. *Proc. Natl. Acad. Sci. U S A* **102**, 1360-1365.
4. Boudko, D.Y. (2006) Molecular basis of the essential amino acid absorption in vector mosquitoes. *Am. J. Trop. Med. Hygiene,* **75 (5)**: 170.
5. Okech, B.A., Harvey, W.R. and Boudko, D.Y. (2006) Distribution of two essential amino acid transporters in the larval alimentary canal of the African malaria mosquito *An. gambiae* (Diptera: Culicidae). *Am. J. Trop. Med. Hygiene* **75 (5)**: 4-5
6. Meleshkevitch, E. A., Assis-Nascimento, P., Popova, L. B., Miller, M. M., Kohn, A. B., Phung, L., Mandal, A., Harvey, W. R. and Boudko, D. Y. (2006) Molecular characterization of the first aromatic nutrient transporter from the sodium neurotransmitter symporter family *J. Exp. Biol.* **209**: 3183-3198.

THE INFLUENCE OF TANDEM REPEATS ON LD AND RECOMBINATION: CREATION AND DESTRUCTION

GEROME BREEN

There are >1 million candidate polymorphic TRs in the human genome (Breen et al., submitted), and many occur in gene regulatory regions. This is comparable to the number of SNPs (6-8 million). There is a wealth of evidence to support the view that TRs are often functional and numerous reports in the literature have shown their association with both monogenic and complex polygenic disorders. My group have recently identified a novel intron 8 VNTR in the dopamine transporter gene, associated with cocaine addiction (Guindalini et al., 2006) and attention deficit hyperactivity disorder (Brookes et al., 2006). This intron 8 VNTR also appears to modulate gene expression in reporter gene constructs and appears to be drug responsive in its regulation of expression, with the risk allele up and down-regulating its effects when the transfected cell lines are exposed to different compounds, such as cocaine and amphetamine. Thus, multiple strands of evidence that polymorphic tandem repeats are (a) useful for genetic fine mapping of complex disease loci, (b) that they are also often functional, and that (c) they may contribute to disease themselves.

One interesting aspect of these association studies (Breen et al., and Brookes et al. 2006) was the inability to map the effect of the VNTRs with SNPs in either study suggesting that, whatever their functional consequences, tandem repeats have properties that distinguish from SNPs in LD terms. Several studies have shown that a class of tandem repeats, microsatellites, are highly polymorphic, and have LD lengths in the 100 kb range when compared with the shorter, ~30 kb,

Institute of Psychiatry, London, UK

range for SNPs, probably due to the older age of SNPs. However, there is scant information on the linkage disequilibrium relationships between TRs and SNPs, especially with respect to haplotypes and LD blocks. The evidence that does exist is tantalizing: Oka et al., 1999 estimated microsatellite-microsatellite LD at 100kb while others (Kendler et al., 1999) found even more (up to 2 mb). Koch et al. (2000) found strong TRP-SNP LD at ADH4 (400kb). This is approximately 3-10 times more than SNPs (for the measure they used). Overall, it appears that pairwise LD is stronger for TR-TR combinations than TR-SNP and is weakest for SNP-SNP combinations. However, no large scale quantification, along the lines of the HapMap, of the role of TRs in linkage disequilibrium changes in the human genome has been carried out. This makes it difficult to reconcile this information about TRs longer LD with their known properties, such as being recombinogenic and possessing a higher mutation rate than SNPs.

As a pilot study for this project, we carried out a preliminary analysis looking at TR density and length between SNPs vs the change in LD between those SNPs for human chromosome 19. For this we used to LDU metric maps from Southampton derived from the HapMap phase I data on 744,000 SNPs (Maniatis et al., 2002). Human Chromosome 19 has 63,811,651 bases (2% or so the total) yet has 1590 genes (6.3% of the total RefSeq genes). It is similarly rich is tandem repeats with 18,940 tandem repeats (3% of the total) in the TRF UCSC Genome Browser March 2006 assembly of the human genome. Another 1042 perfect microsatellites not included in the TRF annotation from our database at www.microsatellites.org. We found that TR density and properties were associated with LD changes with genomewide correlate of ~0.1 with $p<10^{-6}$ for TR density, repeat unit size and number of repeats. Upon binning the genome into 100kb bins, the correlations increased to >0.5 for density and number of repeats. We now need to expand this analysis to the entire genome and to use the HapMap phase II data (>3,000,000 SNPs) and I will present progress towards this.

MODELING OF GENETIC FLOWS IN A STRUCTURED SINGLE-DIMENSIONAL POPULATION

YU.S. BUKIN

The use of DNA sequences for studies of genetic diversity allows one to increase dramatically potential resolving power of analysis and therefore the possibility find minor discontinuities in natural populations or to define more pre-

Limnological Institute SB RAS, Ulanbatorskaya 3 664033 IRKUTSK Russia,
bukinyura@mail.ru

cisely geographic borders of the populations. There are several thoroughly studied models of the migration in natural populations. These models differ by spatial setup: the island model stepping stone model by Kimura and the model with isolation by distance. All models describe the impact of migration on genetic diversity of the populations.

Often data sets consisting of large numbers of homologous DNA sequences are used in population genetic studies. Accordingly, accumulation of DNA diversity in populations accomplished with isolation/migration had been treated in several papers (Strobeck, 1987; Wilkins, et. al. 2002). In order to quantify the degree of isolation between two subsets of individuals most commonly F_{st} criterion is used (Wright, 1951). This value varies continuously between 0 and 1. The F_{st} approaches when two populations become more isolated and therefore gene flow between them approaches 0. For DNA sequences F_{st} may be calculated using a distance measure (Slatkin, 1991; Hudson, et. al. 1992).

There is special and very interesting case of single-dimensional populations where the width of areal is negligible if compared to the length of it. Examples of such populations are easily found among benthic invertebrates inhabiting in a littoral zone of a deep water body. At Lake Baikal there are many species which inhabit narrow zone at 10-100 m on a steep slope. This area is very important since it contains major part of species diversity in Lake Baikal or Lake Tanganyika. In this case one may usually presume that the major disrupting impact may due to geographic barriers interrupting this narrow zone. Sufficiently long stretches of bottom with unfavorable conditions may become the barriers. Genetic diversity in population of this zone studies currently is mostly estimated by comparing nucleotide sequences of mitochondrial genes.

Here we simulate gene flows in single dimensional populations using individual based approach of population dynamics. The methodology of individual-based simulations is well developed and is used widely to study different evolutionary scenaria (Dieckmann, et. al. 2004). Spatially subdivided populations were treated with this approach successfully (Doebeli, et. al. 2003). In our previous studies we used this approach to estimate the evolutionary consequences of different individual mobility (Semovski, et. al. 2002; Semovski, et. al. 2003).

We describe individual-oriented computer simulation of population dynamics processes including birth, death and migration in a single-dimensional population. Each individual has neutral randomly evolving and maternally inherited «DNA sequence», which follows the pattern of mtDNA. Probability of mutation is set to constant. This allowed us to study possible changes in sequence diversity patterns due to partial geographic isolation in natural lacustrine populations.

Accordingly, the general model had been modified by addition of a "geographic barrier" of different isolating power and length of existence. Using this model we simulated the process of genetic differentiation of groups in this organism taking into account isolation by distance and geographical barriers. Fst criterion was used in order to estimate of genetic flow.

With the help of this model we calculated different scenarios of migration and interaction of organism and determined stationary state of neutral DNA polymorphism with the help of Fst criterion. If DNA polymorphism in model correspond with real date it allow as make assumption that causes of genetic polymorphism in the model and real population are equal.

1. Dieckmann, U., Doebeli, M., Johan, A. J. Metz, Tautz D. (2004). Adaptive Speciation. Cambridge University Press.
2. Hudson, R.R., Slatkin, M., Maddison, W.P., (1992). Estimation of levels of gene flow from DNA sequence data. Genetics (US) 132: 583-589.
3. Semovski, S.V., Bukin, Y.S., Sherbakov, D.Y. (2002). Speciation in one-dimensional population: adaptive dynamics and neutral molecular evolution. Internet magazine "Investigated in Russia", 1397-1402, http://zhurnal.gpi.ru/articles/2002/125e.pdf.
4. Semovski, S.V., Bukin, Y.S., Sherbakov, D.Y. (2003). Speciation and neutral molecular evolution in one-dimensional closed population. International journal of modern physics, 14: 973-983.
5. Slatkin, M. (1991). Inbreeding coefficients and coalescence times. Genet. Res. 58: 167-175.
6. Strobeck, C. (1987) Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics 117: 149-153.
7. Wilkins, J.F., Wakeley, J. (2002). The Coalescent in a continuous, finite, linear population. Genetics 161, 873-888.
8. Wright, S. (1951). The genetical structure of population. Ann. Eugenics 15: 323-354.

TOWARDS ABSOLUTE TARGET CONCENTRATIONS FROM OLIGONUCLEOTIDE MICROARRAYS

C.J. BURDEN[1], Y. PITTELKOW[2], S.R. WILSON[2]

There is as yet no practical procedure for inferring absolute target concentrations from the fluorescence intensity data produced by gene expression arrays. Existing expression measures commonly used by experimental biologists are semi-quantitative in the sense that they only detect a ranking of target concentrations between distinct biological samples, and even then, only for those genes for which there has been a substantial change. At best, each currently available expression measure could be described as a gene-dependent, unknown increasing function of target molecule concentration, modulo statistical noise. The ultimate aim of the research presented here is to find an efficient and accurate algorithm for inferring absolute target concentrations from microarray fluorescence intensity data.

While the algorithms behind expression measures are often statistically sophisticated, very little attention is paid to the complex problem of understanding the physical processes involved in going from target concentration to observed fluorescence intensities. We have developed a mathematical model of this process which uses established principles of physical chemistry and statistical mechanics to describe the hybridisation of target molecules onto probes to form duplexes, and the partial dissociation of duplexes during the post-hybridisation washing step.

Any such model needs to consider a number of possible factors including, but not restriced to, probe-specific binding affinities, competitive hybridisation from non-specific targets, non-equilibrium hybridisation, bulk target-target hybridisation and probe-probe and probe-self interactions. In deciding which aspects are important we have, as a general principle, insisted that our modelling be consistent with the Affymetrix Latin Square spike-in experiments by using a statistically rigorous process of balancing parsimony with accuracy of fit. One important discovery we have made is the importance of including the post-hybridisation washing step to explain the differing asymptotic fluores-

[1] Centre for Bioinformation Science, Mathematical Sciences Institute and John Curtin School of Medical Research, Australian National University, Canberra, A.C.T.0200 Australia, Conrad.Burden@anu.edu.au
[2] Centre for Bioinformation Science, Mathematical Sciences Institute, Australian National University, Canberra, A.C.T.0200 Australia, Sue.Wilson@anu.edu.au, Yvonne Pittelkow@anu.edu.au

cence intensities between perfect- and mismatched probes at high spike-in concentrations.

To be useful, the model must be predictive as well as descriptive. Using a bootstrap analysis, we have tested the ability of our model to reproduce absolute spike-in target concentrations. We find that in general the method performs at least as well as MAS5, RMA or PLIER for the Affymetrix U95A Latin Square data set, particularly at higher concentrations where saturation effects are important.

1. C. J. Burden, Y. Pittelkow and S.R. Wilson (2004) Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays, *Stat. Appl. Gen. Mol. Biol.,* **3:** Article 35.
2. C. J. Burden, Y. Pittelkow and S.R. Wilson (2006) Adsorption Models of Hybridization Behaviour on Oligonucleotide Microarrays, *J. Phys.: Condens. Matter,* **18:** 5545-5565.

IDENTIFICATION OF FUNCTIONALLY LINKED GENES BY COMBINING POSITIONAL COUPLING IN BACTERIA AND CORRELATION OF EXPRESSION PROFILES IN EUKARYOTES

NADEZHDA A. BYKOVA[1], ROMAN A. SUTORMIN[1], PAVEL S. NOVICHKOV[2]

It is known that positional coupling of genes in bacteria may indicate functional correlation [1]. For the eukaryotes, the functional coupling manifests in similarity of gene expression profiles [2]. Here we combine positional coupling of genes in bacterial genomes with the correlation of gene expression in eukaryotic genomes using the relations between the bacterial and eukaryotic orthology groups. We believe that this may increase the reliability of the functional coupling prediction.

We used the bacterial and eukaryotic clusters of orthologous genes (COGs and KOGs, respectively) [3]. At the first step, the positional coupling for each pair of COGs was computed as the number of bacteria where representatives of these COGs were near each other in the chromosome. For each pair of KOGs we computed the correlation of gene expression based on microarray data [4, 5]. Then, pairs of two kinds were linked by relations of orthology between bacterial and eukaryotic ortholog clusters [3]. As a result we obtained quadruples of

[1] Moscow State University, Faculty of Bioengineering and Bioinformatics, GSP-2, building 73, Vorobiovy Gory, Moscow, 119992, Russia, nadelle4@mail.ru
[2] National Center for Biotechnology Information USA, pnovichkov@yandex.ru

ortholog clusters, characterized by the value of positional coupling and the expression correlation.

To estimate the reliability of functional annotation predictions based on quadruples, and compare this method with other published approaches, we compared our positional coupling data with the distances on the KEGG database of metabolic maps [6]. For that analysis, only genes encoding enzymes were considered. For each pair of COGs we computed the distance on metabolic map defined as the smallest number of the intermediate compounds between the catalyzed reactions. Distances larger than four were then merged and the corresponding genes were considered to be functionally uncoupled. These distances were compared with the list of positionally coupled pairs of COGs and, separately, with the quadruples of clusters. It turned out that for the pairs of the COGs, participating in quadruples, the fraction of functionally uncoupled pairs was twice lower (7%) than for all coupled COG pairs (14%). This confirms our suggestion that pairs of COGs found in quadruples are functionally coupled with a higher probability.

In each case we also determined the set of threshold values for varying probability of coupling. Orthologs with coupling score exceeding a threshold correspond to enzymes close on the metabolic map with the given probability (the thresholds P100% and P95% were defined for the probabilities 1 and 0.95, respectively). The P100% values differs about two-fold for COGs coupled only positionally (traditional approach, P100%=20) and for the pairs found in quadruples (P100%=9,5). Obviously, the lower is the threshold, the higher is the reliability of predictions. Taken together, these results demonstrate the usefulness of considering expression correlation in eukaryotic genes orthologous to bacterial ones for determination of the functional coupling.

To evaluate the quality of quadruples, we defined the "mixed score", taking into account positional coupling and expression correlation simultaneously. The mixed score is the product of the coupling score and expression correlation. For the mixed score, the threshold values (P100% and P95%) were also calculated, allowing us to estimate the reliability of functional coupling predictions.

The list of quadruples with the scores is available on the web at http://www.bioinf.fbb.msu.ru/cklink. It includes a search system allowing one to sort data by keywords in descriptions and names of ortholog clusters and a system of filters that may be used to intersect conditions to select interesting quadruples. The database contains 963 quadruples. As expected, the largest scores were assigned to ribosomal genes. They constitute 33% of the sample. At the same time, many quadruples contains uncharacterized (23) or poorly characterized (185) orthology groups. For some of them the values observed for the

mixed score are rather high. Therefore, the developed resource may serve for annotation of uncharacterized genes by predicting their functions based on the function of coupled genes, and also may further be used by biologists to fill in white spots on the metabolic map.

1. T. Dandekar et al. (1998) Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem Sci.*, **23**:324-328.
2. M. Gerstein , R. Jansen (2000) The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function?, *Curr Opin Struct Biol.,* **10**:574-584.
3. E. Koonin et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes, *Genome Biol.,* **5**:R7.
4. M. Eisen et al. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A,* **95**:14863-14868.
5. J. Stuart et al. (2003) A gene-coexpression network for global discovery of conserved genetic modules, *Science,* **302**:249-255.
6. M. Kanehisa et al. (2006) From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res.,* **34**:D354-357.

HYDRODYNAMIC VIEW OF PROTEIN FOLDING

S. F. CHEKMAREV[1], A. YU. PALYANOV[1], M. KARPLUS[2]

Free energy surfaces (FESs) are widely used to gain insight into protein folding (see, e.g. [1]). To construct the FES, the multidimensional conformation space of a protein is reduced to a pair of variables, which are intended to represent the folding process in a proper way, such as the radius of gyration and the fraction of native contacts. The FESs have played an important role in the justification of the "new view" of protein folding [2], according to which the FES of a protein is biased toward the native state [3,4], thus providing a guided search for the unique functional structure. At the same time, the FESs leave a large degree of uncertainty about the folding kinetics, because the same values of the free energy can correspond to different states which the protein visits when it folds and unfolds. It is thus of interest to see how the flows of representative points of the protein from the unfolded state to the native state are distributed over the conformation space and to determine their relation.

To address this issue, we introduce a hydrodynamic interpretation of protein folding. Two model systems are considered: a lattice α-helical hairpin and

[1] Institute of Thermophysics, SB RAS, and Novosibirsk State University, 630090 Novosibirsk, Russia, chekmarev@itp.nsc.ru
[2] ISIS, Université Louis Pasteur, 67000 Strasbourg, France, and Harvard University, Cambridge, MA 02138, USA , marci@tammy.harvard.edu

an off-lattice three-helix bundle protein. To simulate the folding process, a Metropolis Monte Carlo method and discrete molecular dynamics are used. We show that the average flow of transitions from the unfolded to the native state may concentrate in a very narrow region of the conformation space (Fig.1); i.e., only a small fraction of the low free energy portions of the FES is visited by trajectories that result in folding. A considerable portion of the conformation space is occupied by a flow "vortex", which is not evident from the FES (Fig.1); it presents a “repulsive” dead-end, hidden in the FES.

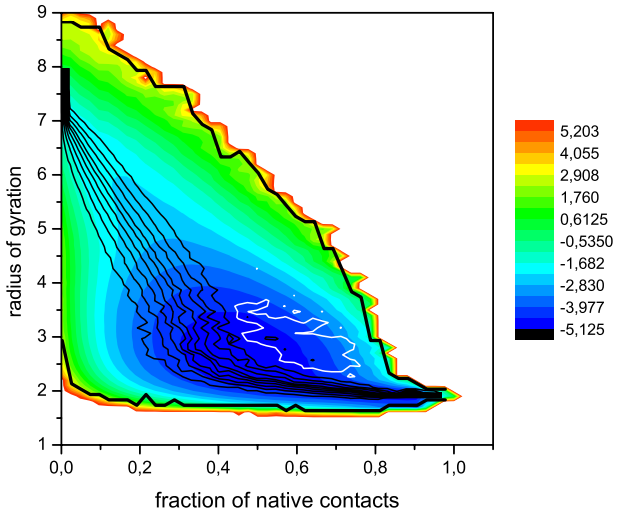


Fig. 1. Free energy surface and streamlines for the model α-helical hairpin, T = 0.575. The lower and upper black thick lines correspond to the fractions of the total flow equal to 0 and 1, respectively, and the thin lines between them to the 0.1, 0.2,...,0.9 fractions. The white line shows the flow "vortex".

The physical origins of the “vortex” regions and the “hydrodynamic” picture are discussed with reference to the two model systems.

This work was supported in part by the grant from the CRDF (RUP2-2629-NO-04). S.Ch. and A.P. also acknowledge support from the RFBR (#06-04-48587). M.K. acknowledges support from the National Sciences Foundation.

1. A.R.Dinner et al. (2000) *Trends in Biochem. Sci.,* **25:** 331-339.
2. R.L.Baldwin (1994) *Nature*, **369:** 183-184.
3. J.D.Bryngelson, P.G.Wolynes (1989) *J. Phys. Chem.* **93:** 6902-6915.
4. T.Lazaridis, M.Karplus (1997) *Science*, **278 :** 1928-1931.

AMPER: A DATABASE AND AN AUTOMATED DISCOVERY TOOL FOR GENE-CODED ANTIMICROBIAL PEPTIDES

ARTEM CHERKASOV

Increasing antibiotics resistance in human pathogens represents a pressing public health issue worldwide for which novel antibiotic therapies based on antimicrobial peptides (AMPs) may offer one possible solution. In the current study we utilized publicly available data on AMPs to construct hidden Markov models (HMMs) that enable recognition of individual classes of antimicrobials peptides (such as defensins, cathelicidins, cecropins, etc) with up to 99% accuracy and can be used for discovering novel AMP candidates.

HMM models for both mature peptides and propeptides were constructed. A total of 146 models for mature peptides and 40 for propeptides have been developed for individual AMP classes. These were created by clustering and analyzing AMP sequences available in the public sources and by consequent iterative scanning of the Swiss-Prot database for previously unknown gene-coded AMPs.

As a result, an additional 229 addtional AMPs have been identified from Swiss-Prot, and all but 34 could be associated with known antimicrobial activities according to the literature. The final set of 1045 mature peptides and 253 propeptides have been organized into the open-source AMPer database. The developed HMM-based tools and AMP sequences can be accessed through the AMPer resource at http://www.cnbi2.com/cgi-bin/amp.pl.

1. C.D. Fjell, R.W. Hancock, A. Cherkasov (2007) AMPer: A Database and an Automated Discovery Tool for Antimicrobial Peptides. *Bioinformatics,* **23**, *Epub ahead of print, PMID: 17341497*

Division of Infectious Diseases, Department of Medicine, Faculty of Medicine, University of British Columbia, 2733 Heather street, Vancouver, BC, Canada, V5Z 3J5, Canada, artc@interchange.ubc.ca

COMPUTING SEARCHING FOR NUCLEOTIDE SEQUENCES LIKE AGROBACTERIAL T-DNA FRAGMENTS IN PLANT GENOMES

M.I. CHUMAKOV, S.I. MAZILOV

Members of the genus *Agrobacterium* (family *Rhizobeaceae*) are natural soilborne plant-root-system residents that can transfer a portion of their Ti-plasmid DNA (T-DNA) into host-plant nucleus under condition of virulence-gene activation. *A. tumefaciens* transfers the ssT-DNA-VirD2 complex to the plant nucleus, where it becomes integrated in the plant chromosome, by using VirD2 and the plant repair system proteins in a sequence-independent manner [1]. We assumed that T-DNA might serve as a mutation factor to change plant adaptation to the environmental conditions. The aim of this work was to search for nucleotide sequences similar to agrobacterial T-DNA fragments in plant-genome data banks and to evaluate the role of naturally associated soilborne agrobacteria in plant evolution.

For computer searching for nucleotide sequences (GGCAGGA-TATT(CA/GG)G(T/G) TCTAA(AT/TC)) from agrobacterial T-DNA right border, the genes *npt*II, *rol*C described in [2] in plant-genome sequence databases (Gen-Bank, DDBJ - DNA Data Bank of Japan) we used the BLAST program 2.2.14, and 2.2.12 versions) at http://www.ncbi.nlm.nih.gov and http://www.ddbj.nig.ac.jp/search /blast-e.html, respectively, and Clustal X 1.81 program [3] for alignment of the sequences. All the checked variants of T-DNA right borders are listed :

1) ggcaggatattcagttctaaat; 2) ggcaggatattggggtctaatc; 3) ggcaggatatt-cagttctaatc; 4) ggcaggatattcaggtctaaat; 5) ggcaggatattcaggtctaatc; 6) ggcagga-tattggggtctaaat; 7) ggcaggatattgggttctaaat; 8) ggcaggatattgggttctaatc

We found from 2 to 115 nucleotide sequences similar to the T-DNA right border-like fragments (TRBLF) in different plant genomes, depending on the variant and length of the TRBLF (Table 1). Most of the TRBLFs were found in the corn genome. The length of the TRBLF fragments found in the corn and *Arabidobsis* genome ranged from 10 to 17 bp.

Institute of Biochemistry and Physiology Plants and Microorganisms, Russian Academy of Sciences, 13 Prospekt Entuziastov, Saratov 410049, Russia; Corresponding author: chumakov@ibppm.sgu.ru

Table 1. The total number of T-DNA right border fragments** observed in the plant genomes

T-DNA right border-like fragments *	*Zea mays* ****	*Petunia* sp. E<=214***	*Nicotiana tabacum* E<=437***	*Trifolium repens* E<=7.5***
1	115	21	17	9
2	83	16	20	4
3	96	21	20	10
4	62	19	27	3
5	80	20	29	3
6	95	16	15	2
7	93	9	22	2
8	95	9	20	2

* According to Table 1; ** The sum of 10-15 nucleotide fragments *** - the Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. For more details please see the calculations in the BLAST Course (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html). The default value (10) means that 10 such matches are expected to be found merely by chance. **** - 15 nucleotides (E=1.8), 14 nucleotides – (E=7.0), a 13 nucleotides (E=28).

We tried to search for the gene *npt*II in different plant genomes and found it (with an identity of 99.9% and E = 0) only in the *eIF-4A1* gene for the translation initiation factor eIF-4A1 (exons 1-5 from the *Arabidopsis thaliana* plant) and within the *gus* gene for β-glucuronidase protein, as a result of transfer of binary vector pBI121 to the *Arabidobsis* genome [2]. The 18-nucleotide-long *npt*II fragment (cgacggcgatgatctcgt) was also found in the *Arabidobsis* genome.

Within the *Arabidobsis* genome, we found a set of 18-nucleotide-long fragments (E=1.8) from the gene *rol*C. Within the *Zea mays* genome, we found a set of 21-nucleotide-long fragments (E=0.042); a set of 16-nucleotide-long fragments (E=11) was found within the *Tobacco* genome originating from the gene *rol*C.

Thus, by using BLAST program, we found a set of short and presented by: 10-15 nucleotides or 40-60% of full-length T-DNA right border-like fragments, and *npt*II-like (0.01- 0.02 % of full-length *npt*II) fragments within different plant ge-

nomes and data banks; and 18-21 nucleotides fragments from the gene *rol*C (0.03 - 0.04 % of full-length gene) in the *Zea*, *Arabidobsis* and *Tobacco* genomes.

We hypothesize that the full-length agrobacterial T-DNA insertion into the plant genome possible involved in plant evolution was eliminated during plant evolution, since T-DNA fragments are presented as 40-60% of full-length TRBLF. All fragments similar with *npt*II and *rol*C genes found in plant genomes were very short (10-21 nucleotides) and possibly were not represented as full-length transfer sequences.

This work was supported by a grant №02.512.11.2072 from the Russian Federal Agency for Science and Innovation.

1. Gelvin S.B. (2003) *Agrobacterium*-mediated plant transformation: the biology behind the "gene-jockeying" tool. *Microbiology and Molecular Biology Reviews* **67**, 16-37
2. Chen P.Y., et al. (2003). Complete sequence of the binary vector pBI121 and its application in cloning T-DNA insertion from transgenic plants. *Mol. Breed.* **11**, 287-293.
3. Higgins D.G., et al. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.,* **266**, 383-402.

REGTRANSBASE (RTB) - A DATABASE OF REGULATORY SEQUENCES AND INTERACTIONS IN PROKARYOTIC GENOMES

MICHAEL J. CIPRIANO[1], ALEXEI E. KAZAKOV[2], DMITRY RAVCHEEV[2], ADAM ARKIN[1], MIKHAIL S. GELFAND[2], INNA DUBCHAK[1]

RegTransBase, a database describing regulatory interactions in prokaryotes, is manually curated and based on published scientific literature. Although a number of databases describing interactions between regulatory proteins and their binding sites are currently being maintained, they focus mostly on the model organisms E.coli and B.subtilis, or are entirely computationally derived. RegTransBase describes a large number of regulatory interactions and contains experimental data which investigates regulation with known elements and contains the following types of experimental data: investigating the activation or repression of a gene's (or operon's) transcription by an identified direct regula-

[1] Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley CA 94720 USA, mjcipriano@lbl.gov
[2] Institute for Information Transmission Problems, RAS. Bolshoi Karetny pereulok 19 Moscow, 127994, Russia

tor; regulation of the gene's (or operon's) expression on the post-transcriptional level; mapping of a promoter or terminator; characterization of an operons' structure (co-transcription, complementation etc.); determining the transcriptional regulatory function of a protein (or RNA) directly binding to DNA (RNA); mapping of the binding site of a regulatory protein; characterization of a regulatory mutation; prediction of the binding sites of a regulatory protein, and others.

Currently, RTB content is derived from ~3500 relevant articles describing over 8000 experiments in relation to 155 microbes. It contains data on the regulation of ~7500 genes and evidence for ~6500 interactions with ~650 regulators.

RegTransBase additionally provides an expertly curated library of alignments of known transcription factor binding sites covering a wide range of bacterial species. Each alignment contains information as to the transcription factor which binds the DNA sequence, the exact location of the binding site on a published genome, and links to published articles.

RegTransBase builds upon these alignments by containing a set of computational modules for the comparative analysis of regulons among related organisms. These modules guide a user through the appropriate steps of transferring known or high confidence regulatory binding site results to other microbial organisms, allowing them to study many organisms at one time, while warning of analysis possibly producing low confidence results, and providing them with sound default parameters. We have used this tool to replicate the analysis from published articles, and expand those predictions to newly sequenced organisms. We present here the findings from this analysis

An intuitive, interactive user-friendly interface makes this knowledge also accessible to the larger microbiological research community at http://regtransbase.lbl.gov.

MODELING IN SYSTEMS BIOLOGY: PROGRESS, PROBLEMS AND APPLICATIONS TO BIOTECHNOLOGY AND BIOMEDICINE

OLEG V. DEMIN

Two approaches of computational systems biology are presented: pathway reconstruction and kinetic modeling. Pathway reconstruction collects all in-

A.N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, Russia, Institute for Systems Biology SPb, Sankt-Peterburgh, Russia, demin@genebee.msu.su

formation about players of interest, processes interconnecting them and their stoichiometry and can be considered as a powerful tool to search for drug targets, discover possible biomarkers and attribute them to particular cell state or phenomenon. In framework of kinetic modeling approach we mine, collect and integrate quantitative *in vitro* and *in vivo* experimental data produced by classical biochemistry, genomics, proteomics and metabolomics and use them to build and verify kinetic models [1,2]. These kinetic models when considered as a repository of all information about the system of interest can be applied to different problems of drug discovery and production such as screening optimization [3], investigation/prediction of drug safety [4] and optimization/maximization of yield of drug precursors. Several examples illustrating these approaches have been presented.

Kinetic modeling was applied to analysis of cell response to mutations leading to dramatic changing in gene expression and cell metabolism and cell colonies growth. We focused at a consideration of the following types of mutations: 1) growth arrest mutations in two genes leading to growth arrest when combined and 2) restitution mutations, in which one is growth arrest but when combined with other cell growth is restored. We used experimental data on doubling time for 9 strains of *E.coli* with mutations in purine nucleotides biosynthesis pathway and growing on medium with guanosine as the sole source of purines [5]. For analysis these effects we applied kinetic model of purine nucleotides biosynthesis which accounts for gene regulation of *de novo* pathway by purine repressor PurR and induction of enzymes in salvage pathway. On the basis of kinetic modeling mechanisms of genes interaction leading to arrest and restoration of cell growth were studied.

A kinetic model for a four-enzyme section of the shikimate pathway (AroB, D, E and K-catalyzed reactions) from Streptococcus pneumoniae

has been constructed and validated. The foundations of the model are sets of kinetic data collected for reactions of the four individual enzymes and for two-, three-, and four-enzyme linked reactions. It describes the dynamic behavior of the four-enzyme system with any concentration of enzymes or substrates. This model has been employed to design and optimize an inhibition-sensitive reconstituted pathway for a high-throughput screening effort on the shikimate pathway [3].

Kinetic models of metabolic pathways involved in synthesis and degradation of arachidonic acid and signaling networks initiated by prostaglandins in platelet and endothelium cells have been developed to understand/predict the mechanism of NSAID-stimulated adverse cardiovascular effects (clot formation). These models quantitatively describe the changes in dynamics and regu-

lations of the pathways caused by the following NSAIDs: aspirin, celecoxib, diclofenac, naproxen, indomethacin, ibuprofen.

1. E. Metelkin, I. Goryanin, O. Demin (2006) *Biophys. J*, **90:** 423-432
2. I. Goryanin, G. Lebedeva, E. Mogilevskaya, E. Metelkin, O. Demin (2006) *Methods Biochem Anal*, **49:** 437-488
3. M. Noble, Y. Sinha, A. Kolupaev, O. Demin, D. Earnshaw, F. Tobin, J. West, J. Martin, C. Qiu, W. Liu, W. DeWolf Jr., D. Tew, I.Goryanin (2006) *Biotechnology and Bioengineering*, **95:** 560-571.
4. E. Mogilevskaya, O. Demin, I. Goryanin (2006) *Journal of Biological Physics*, **32:** 245-271.
5. B. Hove-Jensen, P. Nygaard (1999) *Gen. Microbiology*, **135:** 1263-1273.

RASDB – REGULATION OF ALTERNATIVE SPLICING DATABASE

STEPAN DENISOV[1], RAMIL NURTDINOV[1], DMITRIY VINOGRADOV[2], ALEXEY KAZAKOV[2], GALINA KOVALEVA[2], MIKHAIL GELFAND[2]

The regulation of alternative splicing (AS) is an actual problem of modern molecular biology. Many cases of regulated splicing events are known, but the global picture is still obscure. The aim of our work is to collect all known cases of regulation of alternative splicing from published papers and to organize these data in formal way by creating a database.

From each paper we extracted the following information: all observed genes and mRNA isoforms (partial or full-length), cis-elements and trans-factors, which regulate expression of these isoforms. The information about the current size of RASDB is shown in Table 1.

At the second step, we aligned genes and isoforms with the corresponding genome (we considered the following genomes: *H. sapiens*, *M. musculus*, *R. norvegicus*, *G. gallus* and *D. melanogaster*). This allows us to consider all information about one gene from different papers simultaneously. Currently we are at the third stage: development of software tools for manual curation of these data and curation itself. We plan to create a web-interface to make the RASDB data publicly available.

RASDB has the following structure. There are five main entry types: article, gene, isoform, isoform alignment, regulator (i.e. cis-element or trans-factor) and links between entries. The description of an isoform includes not only its

[1] Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University. Vorbyevy Gory 1-73, Moscow, 119992, Russia, stepan@bioinf.fbb.msu.ru
[2] Institute for Information Transmission Problems, Russian Academy of Sciences. Bolshoi Karetnyi pereulok 19, Moscow, 127994, Russia, gelfand@iitp.ru

exon-intron structure, but also the data about the expression of this isoform on different developmental stages and in different tissues and organs (if it has been presented in corresponding article). The names of organs, tissues and developmental stages for human, mouse and rat were organized in a tree using eVOC Ontologies ([1], http://www.evocontology.org/). We plan to perform analogous work for *Drosophila*. We also store the type of an experiment from a curated vocabulary.

Table 1. Current size of RASDB. Number of genes, cis-elements, trans-factors are indicated

	Homo sapiens	*Mus musculus*	*Rattus norvegicus*	*Gallus gallus*	*Drosophila melanogaster*	Other organisms	Total
Genes	248	55	50	21	59	26	459
Cis-elements	247	47	66	49	56	168	633
Trans-factors	292	53	19	0	78	27	469

There are two types of regulators: cis-elements and trans-factors. Trans-factors are molecules (proteins in most cases) which bind to some regions of a transcript (cis-elements), and then promote or block inclusion of , splicing out of introns, and the use of alternative sites . In our database, cis-elements have the following properties: DNA position, type of regulated event (cassette exon, alternative donor or acceptor site, retained intron etc.), the regulated isoform and the type of experiment. Trans-factors are described by the molecule type (protein or RNA) and GenBank Accession.

The RASDB database can be used for sorting the complexity of regulatory motifs and trans-factors or for understanding the trends and characteristic properties of the evolution of regulated AS events.

We are grateful to members of our scientific group and external annotators for data input.

1. J.Kelso et al. (2003) eVOC: a controlled vocabulary for unifying gene expression data, *Genome Res.,* 13(6A):1222–1230.

PHYLOGENETIC ANALYSIS OF BIOLUMINESCENCE ORGANISM

DILIPAN ELANGOVAN, GEETHA PRIYA GURUSAMY, RAJADURAI MARUTHAMUTHU, RAMYA MOHANDASS, ANUSHA BASKAR

Luciferin is a chemical substance, which emits light in the presence of luciferase enzyme. The bioluminescent cells produce some form of luminescence within the electromagnetic spectrum, which may or may not be visualized through naked eye. This kind of property is unique in wavelength, duration, timing, regulatory of flashes. For this uniqueness luciferase from various bioluminescence organisms were subjected to analysis upon its evolutionary existence by using ProtParam, ClustalX, NJ plot and FM method.

Bioluminescence is the emission of light by a living organism as the result of a chemical reaction during which chemical energy is converted to light energy. The chemical reaction can occur either within or outside of the cell (1). The luciferase enzyme present in the cells of bioluminescent organisms that catalyzes the oxidation of Luciferin along with ATP and Mg^{2+} ions to produce light, oxyluciferin, CO_2 and AMP (3). Bioluminescence is a form of luminescence, or "cold light" emission where less than 20% of the light generates thermal radiation. A variety of species regulate their light production using luciferase, although it exists in organisms as different as in many creatures. The bioluminescence activity is due to presence of Luciferin and Luciferase. In a recently proposed mechanism, the enzyme luciferase modulates emission of color by controlling the resonance-based charge delocalization of the anionic keto form of the oxyluciferin in its excited state (2). Differences in emission of bioluminescence color among different species are due to variations in structure of luciferase. In bacteria, the expression of genes related to bioluminescence is controlled by an operon called lux operon. Luciferin and luciferase are not specific molecules. They are generic terms for a substrate and its associated enzyme (or protein) that catalyze a light-producing reaction. Ninety percent of deep-sea marine life is estimated to produce bioluminescence in one form or another. Most marine light-emission belongs to blue and green light spectrum, the wavelengths that can transmit light most easily through water. However, certain loose jawed fish emit red and infrared light. Non-marine bioluminescence is less widely distributed, but a larger variety in colors is observed (4). The two best-known forms of land bioluminescence are fireflies and glow worms. Other insects such as insect

P.G.Department of Biotechnology, Bishop Heber College (Autonomous), Tiruchirappalli, Tamilnadu, India, e.mail: protaamics@yahoo.co.in

larvae, annelids, arachnids and even a few species of fungi have been noted to possess bioluminescent abilities(5). This luciferin having awesome applications, where light is emitted when luciferase is exposed to the appropriate luciferin substrate. Photon emission can be detected by light sensitive apparatus such as a luminometer or modified optical microscopes. This allows observation of biological processes and stages of infection. Luciferase can be used in blood banks to determine if red blood cells are starting to break down. Laboratories can employ luciferase to emit light in the presence of certain diseases (6). Luciferase is used as a reporter protein in molecular studies, for example to test the activity of transcription from specific promoters with luciferase-transfected cells, or to detect the level of cellular ATP (7). Luciferase is a very heat sensitive protein that is used in studies of protein denaturation, testing the protective capacities of heat shock proteins (8). The opportunities for using luciferase continue to expand.

In this, our present study reveals the physical and chemical properties of *Renilla muelleni, Pleuromama sp. Gaussia princeps,, Cratomophus distinctus, Diaphanes pectinealis, Lampyris noctiluca,, Pyrocoelia rufa, Hotaria unmunsana, Phrixothrix hirtus, Phrixothrix vivianii* by ProtParam, and it subjected to multiple sequence alignment through ClustalX to find the rate and pattern of alignment.

The various amounts of amino acids, their molecular weight, isoelectric focusing point and total number of atoms in luciferase from *Renilla muelleni, Pleuromama sp. Gaussia princeps,, Cratomophus distinctus, Diaphanes pectinealis, Lampyris noctiluca,, Pyrocoelia rufa, Hotaria unmunsana, Phrixothrix hirtus, Phrixothrix vivianii* were analysed by ProtParam tool (9). Then the luciferase sequence of chosen organisms were taken in FASTA format from ExPASY server and aligned in ClustalX (10) to know about their sequence variation, and then for revealing evolutionary aspects, the sequence were loaded in NJplot (neighbor-joining method) to get the phylogenetic tree (11). From the phylogenetic tree, the branch lengths between the organisms are taken into account for calculation of distance by Fitch-Margoliash method (FM method) (12).

The physico-chemical properties of bioluminescent species that were taken in FASTA format incorporated into ProtParam, a tool from ExPASY server. The results of molecular properties of luciferase protein are displayed (Table-I). The results fetched from ProtParam, which show numerous variations between *Renilla muelleni, Pleuromama sp.* and *Gaussia princeps* and rest of the species taken for study. Further the differences observed from Multiple sequence alignment were evolutionarily analyzed by NJplot (Figure-I) and their phylog-

enies were calculated by Fitch-Margoliash method (FM method) show *Pyrocoelia rufa, Diaphanes pectinealis, Lampyris noctiluca* were very closely related and *Renilla muelleni,* distantly related when compared with other species and from distance based method (Table-II), it is observed that *Gaussia princeps* is much weighted and *Diaphanes pectinealis* is the least one.

Organism	No. of Amino acids	Molecular wt	Theoretical pI	Formula	Total No.of atoms
H. unmunsana	548	60476.9	6.12	$C_{2730}H_{4317}N_{703}O_{797}S_{23}$	8570
P. rufa	548	60787.0	6.07	$C_{2737}H_{4283}N_{717}O_{793}S_{27}$	8557
C. distinctus					
D. pectinealis	547	60448.7	5.90	$C_{2729}H_{4285}N_{703}O_{794}S_{25}$	8536
L. noctiluca					
Pleuromama sp.,	527	58317.1	5.83	$C_{2634}H_{4114}N_{680}O_{765}S_{24}$	8217
G. princeps	527	58317.1	5.83	$C_{2634}H_{4114}N_{680}O_{765}S_{24}$	8217
R. muelleni	198	22591.6	6.58	$C_{965}H_{1526}N_{278}O_{306}S_{21}$	3096
P. hirtus					
P. vivianii	185	19900.2	6.71	$C_{872}H_{1423}N_{237}O_{262}S_{15}$	2809
	311	36113.3	5.98	$C_{1645}H_{2504}N_{430}O_{462}S_{13}$	5054
	546	60952.5	6.95	$C_{2755}H_{4346}N_{718}O_{793}S_{23}$	8635
	545	59755.9	6.37	$C_{2704}H_{4269}N_{699}O_{788}S_{18}$	8478

Table-I: Results for bioluminescence organism using ProtParam tool

Figure: 1 Phylogeny tree by NJ Plot

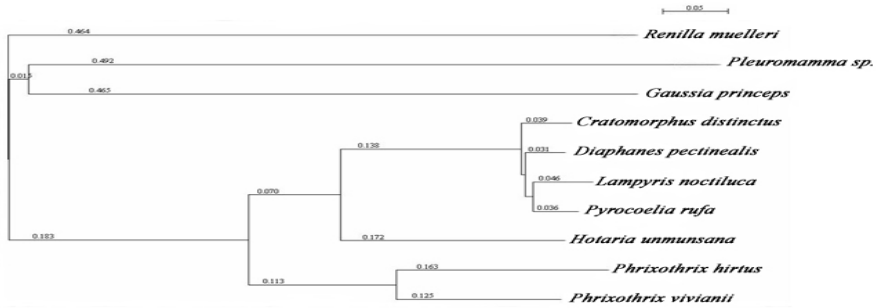


Table-II: Results of distance method and FM method

FM method	A	B	C	D	E	F	G	H	I	J
A	-	1.021	0.994	0.994	0.986	1.051	1.041	0.939	0.973	0.935
B	-	-	0.957	0.937	0.979	1.044	1.034	0.932	0.966	0.928
C	-	-	-	0.91	0.952	1.017	1.007	0.905	0.939	0.901
D	-	-	-	-	0.12	0.185	0.175	0.349	0.523	0.485
E	-	-	-	-	-	0.127	0.117	0.391	0.565	0.527
F	-	-	-	-	-	-	0.082	0.456	0.63	0.592
G	-	-	-	-	-	-	-	0.446	0.62	0.582
H	-	-	-	-	-	-	-	-	0.518	0.48
I	-	-	-	-	-	-	-	-	-	0.288
J	-	-	-	-	-	-	-	-	-	-
Distance based method	0.515	0.505	0.778	0.131	0.043	0.083	0.274	0.172	0.288	0.125

A-Renilla muelleni, B-Pleuromama sp.,C-Gaussia princeps,, D- Cratomophus distinctus, E-Diaphanes pectinealis, F-Lampyris noctiluca, ,G-Pyrocoelia rufa, H-Hotaria unmunsana, I-Phrixothrix hirtus, J-Phrixothrix vivianii and FM method-Fitch-Margoliash method.

The evolutionary analyses for bioluminescent organisms, which produce luciferase in its natural form, were done first time for its varied molecular content when observed by ProtParam. All species producing luciferase in its wild nature, belonging to Kingdom Eukaryota, Sub-Kingdom Metazoa, Phylum Arthropoda except *Renilla muelleni,* which comes under Phylum Cnidaria were put under scanner. On application of ProtParam, major fluctuations were obtained at its amino acid level as well in its molecular level. The NJplot shows *C. distinctus, D. pectinealis, L. noctiluca* and *P. rufa are closely related when compared with H. unmunsana which* come under the same Lampyridae family, upon their branch length. *Pleuromamma sp*. and *G. princeps* coming under Metridinidae family along with *R. mulleini,* a member of Renillidae family, all having different branch length, but falling under a common ancestor, metazoa. It's obvious that *R. mulleini* was very much diverse from the sub-kingdom, metazoa. Finally, *P.hirtus* and *P.vivianii*, belonging to same family, Phengodinae exhibit similar branch length and diverse from other members. When each branch lengths were scrutinized and compared with one another under distance based method, *Gaussia princeps* seemed to be much weighted and *Diaphanes pectinealis,* the least one. To sum up, all these evidences supports the present phylogenetic analysis as a reliable and strong foundation for comparative analyses of bioluminescent organisms.

1. http://en.wikipedia.org/wiki/luciferase
2. Branchini, B.R., Southworth, T.L, Murtiashaw, M. H., Magyar, R.A., Golzales, S.A., Ruggiero, M.C., Stroh, J.G. (2004), Biochemistry 43, 7255-7262.
3. White, E.H., Rapaport, E., Seliger, H.H., Hopkins, T.A. (1971) Bioorg. Chem. 1, 92-122.
4. http://en.wikipedia.org/wiki/luciferase
5. Greer LF 3rd, Szalay AA. Imaging of light emission from the expression of luciferases in living cells and organisms: a review Luminescence. 2002 Jan-Feb; 17(1):43-74.
6. http://www.lifesci.ucsb.edu/~biolum/
7. Jenkins DE, Hornig YS, Oei y, Dusich J, Purchio T. Breast Cancer Res. 2005; 7 (4): R444-54. Epub 2005 Apr 8.
8. Basha E, Lee GJ, Demeler B, Vierling E. Eur J Biochem. 2004 Apr; 271 (8): 1426-36.
9. http://www.expasy.ch/tools/protparam.html
10. Higgins D.G., Thompson J. D., and Gibson T. J. 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. 266: 383-402.
11. Saitou N. and Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees, Mol. Bio. Evol.4: 406-425.
12. Fitch W. M. and Margoliash E. 1987. Construction of phylogenetic trees, science. 155: 279-284.

RESTRICTION-MODIFICATION SYSTEMS AND BACTERIOPHAGE INVASION: WHO WINS?

FARIDA N. ENIKEEVA[1], MIKHAIL S. GELFAND[12], KONSTANTIN V. SEVERINOV[3]

We present a mathematical model of interaction between an invading phage and a restriction-modification system in a cell. This system can be described by a so-called pure-birth process with killing. Let $X(t)$ be a Markov chain with $N+1$ states $i=0,\ldots,N$ and a 'killing state' -1, where N is the number of restriction sites in the invading phage's genome. The system is in the state i if exactly i restriction sites of the DNA are methylated.

Let $\mu(t)$ and $\rho(t)$ be activities of methylase and restrictase in a cell, respectively. We suppose that at any state i the methylase and the restrictase choose a site to be methylated/restricted with the probability $1-i/N$. If all N sites are methylated, the phage survives and the cell dies. In this case the Markov chain hits the absorbing state N. If the restrictase manages to find an unmethylated site, the phage dies and the Markov chain hits the 'phage killing state' -1 meaning that the cell has survived the phage invasion. Let $\mu_i(t)=(1-i/N)\mu(t)$, $\rho_i(t)=(1-i/N)\rho(t)$.

Let $P_k(t)=P\{X(t)=k\}$ be the probability for k sites to be methylated in the system at the time t. Then the Kolmogorov first system of differential equations for this Markov process is given by

$$\frac{dP_0}{dt}=-(\mu_0(t)+\rho_0(t))P_0(t),$$

$$\frac{dP_k}{dt}=-(\mu_k(t)+\rho_k(t))P_k(t)+\mu_{k-1}(t)P_{k-1}(t),\quad k=1,\ldots N-1,$$

with the equations for the absorbing states

[1] Institute for Information Transmission Problems of RAS, Bolshoi Karetny pereulok, 19, GSP-4, Moscow, 127994 Russia, enikeeva@iitp.ru
[2] Faculty of Bioengineering and Bioinformatics, Moscow State University, Vorobyevy Gory 1-73, Moscow, 119992 Russia, gelfand@iitp.ru
[3] Waksman Institute, Department of Biochemistry and Molecular Biology, Rutgers, The State University of New Jersey, 190 Frelinghuysen Road, Piscataway, New Jersey 08854 USA; Institute of Molecular Genetics of RAS, 2 Kurchatov Sq., Moscow 123182, Russia, severik@waksman.rutgers.edu

$$\frac{dP_N}{dt} = \mu_{N-1}(t)P_{N-1}(t),$$
$$\frac{dP_k}{dt} = \sum_{i=0}^{N-1} \rho_i(t)P_i(t),$$

where the initial conditions are $P_0(0) = 1,\ P_k(0) = 0,\ k \neq 0.$

Solving the system of differential equations gives the probability of killing the phage at time t:

$$P_{-1}(t) = 1 - \left(\frac{1}{N} \int_0^t \mu(u)G(u)du + G(t) \right)^N$$

where

$$G(u) = \exp\left\{ -\frac{1}{N} \int_0^u \mu(v) + \rho(v)dv \right\}.$$

In particular, if $\mu(t) \equiv \mu$ and $\rho(t) \equiv \rho$ are constant, then $G(u) = \exp\left\{ -\frac{1}{N}(\mu + \rho)u \right\}$ and the stationary probabilities of a cell to be killed by the phage and to survive are, respectively,

$$P_N(\infty) = \left(\frac{\mu}{\mu + \rho} \right)^N \quad \text{and} \quad P_{-1}(\infty) = 1 - \left(\frac{\mu}{\mu + \rho} \right)^N.$$

This analysis may be easily generalized for the case when the cell is invaded by several identical phages (whose number satisfies the Poisson distribution) and when the methylase and restrictase activities in a cell also are random Poisson variables.

A MODEL OF EVOLUTION WITH CONSTANT SELECTIVE PRESSURE FOR REGULATORY DNA SITES

FARIDA N. ENIKEEVA[1], EKATERINA A. KOTELNIKOVA[23],
MIKHAIL S. GELFAND[14], VSEVOLOD J. MAKEEV[2]

Molecular evolution is usually described assuming a neutral or weakly non-neutral substitution model. New experimental [1–3] and computational [4] methods of identification of transcription factor binding sites (TFBS) produce an increasing amount of data about TFBS sequences, which creates a possibility to study evolutionary events in these regions. To reconstruct the evolutionary history of such sequences, one needs evolutionary models that take into account a substantial constant selective pressure.

The existing evolutionary models, which were successful in reconstruction of phylogenetic relations, can be applied to evolution of regulatory sequences only with a caution. Such models are historically related to the Jukes-Cantor and Kimura models of molecular evolution. Existing modifications of these models take into account various global characteristics like transition/transversion rate or local GC composition. They are not applicable to the case of strong selection for a specific nucleotide at a particular position. On the other hand, models developed specifically for the evolution of TFBS are needed to reconstruct the evolutionary origin of a particular TFBS and to evaluate the position-specific mutation rate and selective pressure.

Here we consider the simplest model of position-specific evolution with one preferred (consensus) nucleotide and three other (minor) nucleotides, the latter considered in a symmetric setting, without any selection or rate preferences [5]. Such a model can be deduced from physical requirements of the TF/TFBS interaction [6] and can explain the observed TFBS fuzziness. We build a rate matrix, which enhances the model of [5]. We calculate the substitution probability for each finite time and show that the nucleotide conservation in phylogenetic lineages can be non-trivial for some parameter values. In particular, our findings show that a nucleotide preferred at some position of a multiple

[1] Institute for Information Transmission Problems of RAS, Bolshoi Karetny pereulok, 19, GSP-4, Moscow, 127994 Russia, enikeeva@iitp.ru
[2] State Research Institute of Genetics and Selection of Industrial Microorganisms, 1st Dorozhnyj proezd, 1, Moscow, 113535 Russia, makeev@genetika.ru
[3] Ariadne Genomics Inc., 9700 Great Seneca Highway, Suite 113, Rockville, MD 20850 USA, ekotelnikova@gmail.com
[4] Faculty of Bioengineering and Bioinformatics, Moscow State University, Vorobyevy Gory 1-73, Moscow, 119992 Russia, gelfand@iitp.ru

alignment of binding sites for a TF in the same genome is not necessarily the most conserved nucleotide in an alignment of orthologous sites from different species. However, this effect can take place only in the case of a mutation matrix whose elements are not identical.

1. J. Wells, P.J. Farnham (2002) Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation, *Methods,* **26:**48–56.
2. A.S. Weinmann, , P.J. Farnham (2002) Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation, *Methods,* **26:**37–47.
3. A. Hube et al. (2006) The promoter competition assay (PCA): a new approach to identify motifs involved in the transcriptional activity of reporter genes, *Front Biosci,* **11:**1577–84.
4. E.A. Kotelnikova et al. (2005) Evolution of transcription factor DNA binding sites, *Gene,* **347:**255–63.
5. U. Gerland, T. Hwa (2002) On the selection and evolution of regulatory DNA motifs, *J Mol Evol,* **55:**386–400.
6. U. Gerland et al. (2002) Physical constraints and functional characteristics of transcription factor-DNA interaction, *Proc Natl Acad Sci U S A,* **99:**12015–20.

PREDICTION AND SIMULATION OF MOTION IN TRANSMEMBRANE PROTEINS

ANGELA ENOSH[1], NIR BEN-TAL[2] AND DAN HALPERIN[1]

Motivation: Motion in transmembrane (TM) proteins plays an essential role in a variety of biological phenomena. Thus, developing an automated method for predicting and simulating motion in this class of proteins should result in an increased level of understanding of crucial physiological mechanisms.

We have developed an algorithm for predicting and simulating motion in a pair of TM α-helices [1]. Our method employs probabilistic motion-planning techniques to suggest possible collision-free motion paths. The resulting paths were ranked according to the quality of the van-der-Waals interactions between the TM helices. The most energetically favorable motion pathways were selected to produce movies.

Our algorithm considers a wide range of degrees of freedom (dofs) involved in the motion, including external and internal moves. However, in order to handle the

[1] School of Computer Science
[2] Department of Biochemistry, Tel Aviv University, Ramat Aviv, 69978, Israel

vast dimensionality of the problem, we employed some relaxations on these dofs in a way that is unlikely to rule out the native motion of the protein.

Results: Overexpression of the RTK ErbB2 was implicated in causing a variety of human cancers. Recently, a molecular mechanism for rotation-coupled activation of the receptor was suggested [2]. We applied our algorithm to investigate the TM domain of this protein, and compared our results to this mechanism. A motion pathway that was similar to the proposed mechanism ranked first (http://www.cs.tau.ac.il/~angela/EGFR.html), and motions with partial overlap to this pathway followed in rank order [1].

More recently, we extended our method to handle more dofs, as in the case of the pore-forming domain of the potassium channel from *Streptomyces lividans* (KcsA). Potassium channels are the most common type of ion channels. They regulate cellular processes, such as neuronal signaling, secretion of hormones, and may also regulate cell volume and the flow of salt across epithelia. Diseases caused by mutations in ion channels, potentially impair cell-cell communication and lead to neurodegenerative disorders or muscular diseases.

Potassium channels use diverse mechanisms of gating (the processes by which the pore opens and closes), but they all exhibit very similar ion permeability characteristics and share high sequence similarity; particularly in the pore domain. The bacterial potassium channel from *Streptomyces lividans* (KcsA) was the first potassium channel whose structure was determined using x-ray crystallography [3]. Since then, several more potassium channel structures have been determined, and we now know the structure of the channel both in the open and closed conformations; the voltage-dependent potassium channel from *Aeropyrum pernix* (KvAP) [4] is considered to be in the open conformation, while KcsA is known in the closed conformation. Here we investigated the molecular details of the transition between these two states.

The pore-forming region in the potassium channel is composed of a bundle of eight helices from four identical monomers through which ions are conducted. Each monomer contributes a pair of TM helices connected by the P-loop region that contains the selectivity filter, which is tuned to select potassium over, e.g., sodium ions. All in all, the system includes 428 amino-acid, each of which adds two dofs (corresponding to its torsion angles) to the dimensionality of the motion configuration space. In addition, we consider side-chain flexibility during the gating phase, implying that the number of dofs in our problem is enormously high.

In order to deal with the vast number of dofs we used a slightly different technique.

First, we constructed the open conformation of KcsA by homology modeling, using the structure of KvAP as a template. Then, given the two conformations (the closed structure of KcsA and the open conformation of KcsA produced by KvAP) we employed probabilistic motion-planning techniques to find a motion pathway that connects the two conformations using a biased conformational exploration from the open toward the closed conformation and visa versa. Our algorithm simulates the motion, including all the dofs, and automatically produces a movie that demonstrates the motion.

Normal mode analysis revealed global corkscrew like counter-rotation of the extracellular and the cytoplasmic regions in the pore region of the potassium channel that leads to the opening of the pore [5]. We applied our algorithm to examine the mechanism of the pore region of KcsA in a more realistic context, considering movement of the helices, torsion-angle changes and side chain flexibility. Our algorithm suggests an energetically favorable pathway for the gating motion in KcsA that can be view at http://www.cs.tau.ac.il/~angela/ potassium.gif. This motion pathway is similar to the hinge-motion suggested by the normal mode analysis [5].

1. A. Enosh, S.J. Fleishman, N. Ben-Tal, D. Halperin (2007) Prediction and simulation of motion in pairs of transmembrane α-helices, *Bioinformatics* 23:e212-e218.
2. S.J. Fleishman, J. Schlessinger and N. Ben-Tal (2002) A putative molecular-activation switch in the transmembrane domain of erbb2, *Proc. Natl. Acad. Sci.* 99:15937-15940.
3. D.A. Doyle, J.M. Cabral, R.A. Pfuetzner, A. Kuo, J.M. Glubis, S.L. Cohen, B.T. Cahit and R. MacKinnon (1998) The structure of the potassium channel: molecular basis of K^+ conduction and selectivity, *Science* 280:69-76.
4. Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B.T. Chait and R. MacKinnon (2003) X-ray structure of a voltage-dependent K^+ channel, *Nature* 423:33-41.
5. I.H. Shrivastava and I. Bahar (2006) Common mechanism of pore opening shared by five different potassium channels, *Biophys. J.* 90:3929-394

ANALYSIS OF CORRELATIONS IN LOCATION OF HYDROPHOBIC AND HYDROPHILIC MONOMERS IN PROTEIN SEQUENCES

E.A. EROKHINA[1], L.V. GUSEV[2], V.V. VASILEVSKAYA[2], A.R. KHOKHLOV[2]

Characteristics of spatial structure of different classes of proteins (globular, membrane, and fibrillar) are largely determined by certain principles in location of hydrophobic and hydrophilic amino acid residues in the protein chain.

A method that allows effective identification of binary sequences of *AB* -copolymers with different statistics of monomer distribution was proposed in [1]. This method is based on calculating the introduced segmentation function $S(k)$, which characterizes the degree of correlations depending on distance k between them along the chain. It was shown that, for random copolymers, the segmentation function is a constant whose value depends on the composition of the *AB* -copolymer. For regular and Poisson sequences, the segmentation function is represented by a periodic function and an oscillating function coming to a plateau, respetively. The segmentation function of protein-like sequences grows at small k values and then comes to a constant value corresponding to the value of this function for a random sequence of the same composition (proportion of *A* and *B*-monomers).

The behavior of the segmentation function $S(k)$ was studied for different groups of proteins. Protein sequences were recoded from twenty-letter amino acid sequences to sequences of hydrophobic (*A*) and hydrophilic (*B*) monomers in accordance with amino acid classifications of refs. [2, 3] and in accordance with new two dimensional classification [4], in which amino acids are ascribed to groups on the basis of strict physicochemical measurements of relative solubility in aquatic and nonpolar phases and surface activity at the interface between these phases.

It was demonstrated that the segmentation function is sufficiently sensitive and allows easily revealing a number of characteristic features of protein sequences—the presence of periodicity in the primary structure and distribution of hydrophobic and hydrophilic groups, as well as the proportion of these groups [5].

Authors are grateful to the Russian Foundation for Basic Research (project 05-03-33077) and the Basic research program of the Division of Chemistry and Material Science of the Russian Academy of Sciences for financial support.

[1] Physics Department, Moscow State University, Leninskie gory, Moscow, Russia, 119991 erokhina@polly.phys.msu.ru

[2] Nesmeyanov Institute of Organoelemental Compounds (INEOS), Russian Academy of Sciences, Vavilova ul. 28, Moscow 119991, Russia

1. L.V. Gusev, V.V. Vasilevskaya, V.Ju. Makeev, P.G. Khalatur, A.R. Khokhlov (2003) Segmentation of heteropolymer sequences specifying subsequences with different composition and statistical properties, *Macromolecul. Theory Simul.*, **12**, 604-613.
2. D.L. Nelson, M.M. Cox (2000) *Lehninger Principles of Biochemistry*, Worth Publishers, New York, 1152 c.
3. B. Alberts, D. Bray, A. Johnson et al (1998) *Essential cell biology*, Garland Publishing, New York, 630 c.
4. I.M. Okhapkin, A.A. Askadskii, V.A. Markov, E.E. Makhaeva, A.R. Khokhlov (2006). Two-Dimensional Classification of Amphiphilic Monomers Based on Interfacial and Partitioning Properties. 2. Amino Acids and Amino Acid Residues. *Coll. Polym. Sci.*, **284**, 575-585.
5. A.Sh. Ziyatdinov, L.V. Gusev, V.V. Vasilevskaya, A.R. Khokhlov (2006) Analysis of correlations in location of hydrophobic and hydrophilic monomers in protein sequences. *Doklady Biochemistry and Biophysics*, **411**, 361-364.

RESTRICTION SITES AVOIDANCE IN BACTERIOPHAGE GENOMES AS A STRATEGY AGAINST RESTRICTION-MODIFICATION SYSTEMS: A WHOLE GENOME ANALYSIS

ANNA ERSHOVA[1], ANNA KARYAGINA[2], SERGEI SPIRIN[13], ANDREI ALEXEEVSKI[13]

Many bacterial and archaeal restriction-modification systems (RM-systems) prevent host cells from invasions of foreign DNA by cleavage the latter at specific sites. Typically, type II RM-system includes two proteins: the restriction endonuclease cleaves DNA at specific sites, the methyltransferase methylate the same sites preventing the cleavage. Bacteriophages posses several strategies to prevent DNA cleavage by host RM-systems [1]. One of them is the avoidance of host RM-systems cleavage sites [1, 2]. To reveal the spread of that strategy, we have computed the occurrences of 259 known type II RM-sites in 366 bacteriophage genomes.

If there are no site of a specific RM-system in a bacteriophage genome, then the bacteriophage can survive in the host carrying that RM-system. Moreover, a bacteriophage has a chance to survive in such host if the bacteriophage genome contains a small number **n** of restriction sites. Indeed, let **p** be a probability of a

[1] Moscow State University, Faculty of Bioengineering and Bioinformatics, Moscow, 119992, Russia, a_ershova@rambler.ru
[2] Gamaleya Institute of Epidemiology and Microbiology, Gamaleya st. 18, Moscow 123098, Russia, and Institute of Agricultural Biotechnology, Timiryazevskaya st. 12, Moscow 127550, Russia, akaryagina@gmail.com
[3] Moscow State University, Belozersky Institute, aba@belozersky.msu.ru

site methylation by methyltransferases before endonucleases cleave the DNA at this site. Assuming independence of site methylation, we may estimate the probability of methylation of all **n** sites in the phage genome by $\mathbf{p}^{\mathbf{n}}$. A methylation of a phage genome is inherited by its descendants in the same hosts (like it happens with host genomes). Thus, if **n** is sufficiently small, then there is a chance for phage infection to survive. A large number of sites makes that probability practically zero because of exponential dependence on **n**.

For each pair: (site sequence, bacteriophage genome) we have computed the number N_{obs} of observed sites in the genome, the expected number of sites N_{exp}, and the standard deviation of number of sites σ. The Markov model approach was used to compute N_{exp} and σ (like in [3]). If the condition $N_{obs} < (N_{exp} - 5.6\sigma)$ was fulfilled, then the site sequence was considered as significantly underrepresented in the genome. All 1,679 such pairs were selected for the analysis.

The histogram of the number of sites for significantly underrepresented site sequences in bacteriophage genomes shows an exponential fall from zero number of sites to 6–8 sites per genome and the uniform distribution from 10 to hundreds of sites. The exponential fall is in accordance with above reasoning. We suppose that cases of less than 8 sites in a genome manifest the site avoidance antirestriction strategy as a main survival tool in bacteriophages (671 cases in our data). That claim is in accordance with experimental data. A significant underrepresentation of dozens, hundreds or even thousands of sites (but expected numbers of sites are much more than observed!) could be accomplished with another bacteriophage antirestriction strategies or could be artifacts of the method, i.e., the reasons of underrepresentation of a word in a genome are different, or our assumption of independence of events of methylation before cleavage is wrong.

Our data allows to compare “site avoidance profiles” for bacteriophages and to predict bacteriophage–RM-system interactions.

This work was partially supported by Russian Foundation for Basic Research (grants 06-04-49558 and 06-07-89143) and INTAS grant 05-1000008-8028.

1. M.R. Tock, D.T.Dryden (2005), *Curr. Opin. Microbiol.* **8**:466-472.
2. P.M. Sharp (1986), *Mol. Biol. Evol.* **3**:75-83.
3. M.S. Gelfand, E.V. Koonin (1997), *Nucleic Acids Res.* **25**:2430-2439.

STRUCTURE OF LINE1 RETROTRANSPOSON PROMOTER REGIONS

A.V. FEDOROV, D.V. LUKYANOV

LINE1 retrotransposons makes up about 20% of mammalian genomes. It becomes clear that LINE1s are not just selfish mobile elements, since their involvement into various cellular processes have been shown. Obviously, LINE1 retrotransposons affect genome mostly via their expression followed by retrotransposition. Transcription of full-length RNA, the first step in L1 retrotransposition and propagation, provides the template for the synthesis of a DNA copy. The dissemination of LINE1 retrotransposons in the genome depends on their transcription, which can be detected only at definite stages of germ cell development, and in case of some cancers. The 5'-untranslated regions of LINE1 transposable elements are known to have promoter activity. These regions of rat, mouse and human LINE1s are not homologous, however similar pattern of LINE1 expression observed in different species. Factors that could be involved in the cell type-specific regulation of LINE1 transcription remain largely unknown.

In the present work we analyzed some structural features of rat, mouse and human LINE1 5'-untranslated regions that could influence their transcriptional potential. Promoters frequently co-localize with Matrix Associated Regions (MARs) that were shown to be involved in the regulation of their transcriptional activity. Computer analysis using MAR-Wiz 1.0 program (http://www.futuresoft.org/MAR-Wiz/) revealed that LINE1 promoters possesses no matrix association potential and do not contain such elements of secondary structure as hairpins and curved regions, that could bind similar structure-specific regulatory proteins. We proved that LINE1 promoters are non-curved DNA regions using DNA Curvature Analysis Software (http://www.lfd.uci.edu/~gohlke/curve/). Determination of the electrophoretic mobility of promoter DNA fragments under different conditions confirmed this result. Analysis of the LINE1s at http://genome.ucsc.edu/ using the ChrN_rmsk table allowed us to determine characteristics of genomic LINE1 5'-truncations and inverted structures, that could result in generation of double-stranded RNA. We also determined distribution of LINE1 retrotransposons in the assembled contigs of rat, mouse and human chromosomes, comparatively with the distribution of known heterochromatin regions, using Ensembl resource (http://www.ensembl.org/). We conclude that factors such as DNA bending and asso-

Institute of cytology RAS, Tikhoretsky pr. 4, Saint-Petersburg, Russia,
an_tn@mail.ru

ciation with nuclear matrix seem not to be involved in the modulation of the transcriptional potential of LINE1 5'-untranslated regions. In contrast, truncation of LINE1 promoter regions and their predicted partially inverted organization could play role in the in the regulation of LINE1 transcription.

A THREADING OF IMMUNOGLOBULIN-LIKE PROTEINS WITH SIMPLE ENERGY FUNCTION

SERGEY FERANCHUK[1], ALEXANDER TUZIKOV[1], VLADIMIR DULKO[1], TATSIANA KIRYS[2], JAIRO ROCHA[3]

The aim of our research is to investigate an effective first approximation for free energy function in protein folding problem. We test this approximation via the ability of threading algorithm to predict correctly the native conformation for given primary structure among several structures of the same folding type. We consider the protein folding type called "immunoglobuline-like beta sandwich" according to SCOP classification and take one representative protein from each superfamily. Proteins of this type consist of two beta-layers and differ in the order of beta-strands in these two layers. The result of threading algorithm here will be the assignment of beta-strands to particular segments in primary sequence. All the difference between structures is different distances between ends of coil segments (from the end of one beta-strand to the beginning of the next beta-strand). For this purpose we introduce to the free energy function the term for the coil entropy that depends on a distance between ends of the coil in the structure and a number of residues in the coil. When beta-layers connect together to form a sandwich, the side chains of even residues in each beta-strand become buried inside the protein. So another term in free energy function is the sum of hydrophobicity of even residues of each beta-strand. In first part of our research we consider various lengths of beta-strands and we should select the set of lengths that optimizes the energy function. This requires the term of free energy function that depends on number of hydrogen bonds between backbone atoms for the given configuration. Another problem is how to organize the enumeration of configurations in order to select the best. For this purpose we suppose that we can reconstruct the folding pathway of a

[1] UIIP NAS of Belarus, 6 Surganova st. 220012 Minsk, Belarus, feranchuk@gmail.com
[2] Belarusian State University, 4 Nezalezhnosty av. 220000 Minsk, Belarus, nushki@mail.ru
[3] University of Balearic Islands, Spain, jairo@uib.es

protein in a hierarchical form and follow this pathway in the enumeration, when the optimal solution is selected on each step. A folding pathway is reconstructed for each protein in the test set in a bottom-up direction, starting from beta-strands and connecting them into nodes of higher level according to the adjacency in primary sequence order. As for the result of enumeration, the positions of beta-strands in primary sequence are predicted well, but the free energy function is unable to select the native structure for given primary sequence. In the second part of our research we keep the lengths of beta-strands as they were in a native structure. In this case the enumeration is reduced to simple dynamic programming. The results are even better, and for several proteins the structure with best free energy for given primary sequence is the native structure. The discrepancies could be explained due to the fact that we disregard alpha-helices that occasionally meet in the structures, and we consider each fold as a separate protein instead of considering it to be a domain in more complex structure.

This work was partially supported by INTAS project 04-77-7178.

MULTI-ATOM VAN DER WAALS AND ELECTROSTATIC INTERACTIONS IN A CORPUSCULAR MEDIUM

ALEXEI V. FINKELSTEIN, D. N. IVANKOV,
N. V. DOVIDCHENKO, N. V. BOGATYREVA

Prediction of ligand binding or protein structure requires very accurate force field potentials: even small errors in potentials can make some “wrong” structure (from zillions possible) more stable than the single “correct” one. Despite huge efforts to optimize them, currently used force field potentials are still not able to bring an approximate, homology-based model of protein structure closer to its native conformation (1).

Van der Waals forces, which are very important for the structure and interactions of biological molecules, are usually treated as a simple sum of pairwise inter-atomic interactions even in dense systems like proteins (2-6). The same concerns electrostatic interactions, which are treated usually as pairwise interactions at the background of uniform, at least within protein body and bulk solvent, dielectric permittivity. The neglect of multi-body in Van der Waals interactions seems to follow the Axilrod-Teller theory (7) predicting a drastic decrease of three-atom interactions with distances; and indeed, detailed compu-

Institute of Protein Research, Russian Academy of Sciences, Pushchino, 142290, Moscow Region, Russia, afinkel@vega.protres.ru

tations of single-atom liquids (8) and solids (9, 10) show that multi-bogy effects amount to only about 5% of the total energy.

However, this work shows that multi-atom VdW interactions can become quite large in the presence of covalent bonds. A strict analysis of Van der Waals interactions in a medium, where each atom is considered as a three-dimensional harmonic quantum oscillator, shows that a specific coupling of multi-atom Van der Waals interactions with covalent bonding can, at extremes, by about 20-40% increase (or decrease) the interaction energy at certain angles between directions of interactions and covalent bonds. These significant energy effects are comparable to those caused by replacement of atoms (say, C by N) in conventional pairwise Van der Waals interactions.

Further, the same, in essence, "harmonic oscillator" model of a medium gives a possibility of a strict computation of electrostatic interactions in a non-uniform corpuscular medium. In conclusion, it is shown that, on the average, multi-body effects decrease the total Van der Waals energy in proportion to the square root of electronic part of that dielectric permittivity, which corresponds to small distances, where Van der Waals interactions take place.

These findings, which equally concern atomic interactions in biological molecules and solvents, imply a necessity to revise currently used all-atom force fields.

The authors are am grateful A.M. Dykhne, A.A. Vedenov, R.V. Polozov, M. Levitt and G. Vriend for seminal discussions. The work is supported in part by the RAS Program “Molecular & Cell Biology”, RFBR, INTAS and the Howard Hughes Medical Institute Award.

1. E.Krieger, T.Darden, S.B.Nabuurs, A.Finkelstein, G.Vriend Proteins 2004, 57: 678-683.
2. M.Levitt, M. Hirshberg, R.Sharon, V.Dagget Comp Phys Commun 1995, 91: 215-231.
3. A.MacKerell, Jr, J.Wiorkiewicz-Kuczera, M.Karplus J Am Chem Soc 1995, 117: 11946–11975.
4. W.L.Jorgensen, D.S.Maxwell, J.Tirado-Rives J Am Chem Soc 1996, 118: 11225-11236.
5. T.A.Halgren J Comp Chem 1995, 17: 490-519.
6. J.Wang, R.M.Wolf, J.W.Caldwell, P.A.Kollman, D.A.Case J Comp Chem 2004, 25: 1157-1174.
7. B.M.Axilrod, E. Teller J. Chem. Phys 1943, 11: 299-300.
8. R.J.Sadus Fluid Phase Equilibria 1998, 144: 351-360.
9. N.B.MacRury, B.Linder J Chem Phys 1971, 54: 2056-2966.
10. A.G.Donchev J Chem Phys 2006, 125: 074713.

A CONSTANT-TIME ALGORITHM FOR REGULAR BINARY MULTIGRID CELL INDEXATION

E. S. FOMIN

Multigrid methods belong to the most promising group of modern high-performance algorithms; because they allow (1) adjustment the local mesh size with the solution change scale by using many levels of discretization and (2) inartificial paralleling and vectorization of applications. The applications of multigrid methods are manifold: they range from structure mechanics and fluid dynamics to electromagnetism and molecular dynamics.

Use of any multigrid method requires a certain mapping of R^{dim} space vectors onto a variety of integer indexes associated with grid points. The mapping algorithm is determined by the grid type and structure, and the grid types used in practice are defined by the dilemma whether an efficient mapping algorithm can be constructed for the grid. The ideal solution for multigrids would be able to preprocess all relevant points of space R^{dim} within time O(n log n) into a data structure demanding O(n) memory and allowing the data extraction within time O(log n).

In this work, a constant-time algorithm O(1) for mapping a real vector $X \in R^{dim}$ onto a grid mesh index for a regular multilevel binary grid is presented. By a binary grid is meant such a grid whose mesh size at each i^{th} level is defined as $h = h_0 / 2^i$. The points of this binary grid are inartificially mapped onto a k-d tree data structure. The dimension "dim" of the space is implied to be limited, and it is not included into the computational complexity of the algorithm.

The idea of mapping algorithm can be demonstrated by example of a one-dimensional grid in the range [0, 1). In this range the index of grid mesh can be expressed as the bit sequence like 0010...1100. The bit at an i^{th} position of this sequence determines whether the point can be assigned to the left (0) or right (1) part of range $[a_i, b_i)$ at the i^{th} level of the grid. There is a direct bijection between such an index and the binary representation of a real $\in [0, 1)$: $c_{-1}*2^{-1} + c_{-2}*2^{-2} + .. + c_{-n} * 2^{-n}$, $c_i = \{0, 1\}$. Owing to such bijection, the index can be built by direct operation with bits of the real. The mapping algorithms implementing this idea have constant time behavior and are independent on the number of grid levels.

The implementation of the algorithm is machine-dependent because it uses information about the storage structure of reals. The upper limit of grid

Institute of Cytology and Genetics SB RAS, 10 Lavrentiev Ave, 630090, Novosibirsk, Russia, fomin@bionet.nsc.ru

levels is determined by the number of bits in the mantissa of the real; for example, it equals 53 for a double in the IEEE 754 standard, which allows up to $2^{53*\dim}$ grid meshes to be addressed.

In test calculations, the suggested algorithm showed a significant speed-up, especially for multilevel grids. For example, the calculation for a 3-level grid is accelerated by a factor of 7.5 and for a 50-level grid, by a factor of 125.

The algorithm has been included into MOLKERN library, developed for support of high performance calculations in molecular modeling. For example, it is used for building of neighbors list or extraction of values of potential functions at grid points.

To obtain the free trial version of MOLKERN, please contact the author by email: fomin@bionet.nsc.ru.

TEMPLATE LIBRARY MOLKERN AS A FRAMEWORK FOR BUILDING EFFECTIVE MOLECULAR MODELING PROGRAMS

E.S.FOMIN[1], N.A.ALEMASOV[2], Z.I.AKNAZAROV[2], A.S.CHIRTSOV[2], A.E.FOMIN[3]

MOLKERN is a template library for all software components required for the analysis, modeling and optimization of spatial macromolecular structures for proteins, cofactors, ligands and their complexes within the force field approximation. It provides a poweful framework for building of many common molecular modeling programs. Today's computer architecture - hierarchical structure of the memory subsystem, pipelining - is respected by the code to achieve efficient program execution. All algorithms implemented have a linear scaling O(N) both in memory and time requirements. The following operations can be performed:

support of multiple spatial molecular structure input/output file formats (pdb, hin, mol2, etc.);

modeling and edition of molecular structures: restoration of missing atoms, protonation of proteins at a given pH value, addition of disulfide bonds, and residue replacement;

optimization of molecular complexes with and without constraints;

search for hydrogen bonds, salt and water bridges, and surface atoms;

calculation of the molecule surface area by numerical and analytical methods;

[1] Institute of Cytology and Genetics SB RAS, 10 Lavrentiev Ave, 630090, Novosibirsk, Russia, fomin@bionet.nsc.ru
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Moscow Institute of Physics and Technology, Moscow, Russia

calculation of the contact interface between two molecules and its area;

energy calculation in vacuo and in water; general Born polarization energy calculation;

direct Gibbs free energy calculation;

search for active protein sites, rigid docking, flexible docking with regard to the mobility of protein atoms at the contact site;

virtual screening of chemical compound libraries;

The library codes, written in C++, use the well-known Standard Template Library (STL) and BOOST library. To avoid overhead associated with function calls, the MOLKERN library uses static polymorphism. The library runs under WINDOWS and LINUX. Some modules can run in parallel using LAM MPI.

Table. Run-time for energy minimization of proteins in water environment. All atoms of proteins are free. We used a Pentium IV computer (2.1 GHz) with the gcc 4.1.1 compiler.

PDB ID	# freedom	E [kCal/mol]	# iteration	time [s]	time / (#iteration * #freedom) [s]
1aie	1566	326.8	72	5.8	5.1e-5
2bin	8649	16279.4	6	5.0	9.6e-5
1f58	19821	14201.9	25	52.7	1.0e-4

To obtain the free trial version of MOLKERN, please, contact the author by email fomin@bionet.nsc.ru.

A FAST APPROXIMATE METHOD FOR CALCULATION OF HIGH DEGREE INTERSECTION AREAS OF ATOMIC SPHERES

E.S.FOMIN[1], A.S.CHIRTSOV[2]

Interaction of biomacromolecules in aqueous environment is a fundamental problem in physics, chemistry, and structural biology. Methods that explicitly take into account the effects of aqueous environment in calculations of energy and binding constants for protein-ligand complexes, prediction of protonization states are computationally consuming and can be hardly useful in problems of molecular docking and virtual screening of chemical compound libraries. Methods that implicitly take into account the aqueous environment effect are intensely developed. In these methods, the first order terms of interaction energy of macromolecules with aqueous environment are proportional to molecule surface area.

[1] Institute of Cytology and Genetics SB RAS, 10 Lavrentiev Ave, 630090, Novosibirsk, Russia, fomin@bionet.nsc.ru
[2] Novosibirsk State University, Novosibirsk, Russia

To date, a number of numerical and analytical methods for calculation of macromolecule surface are available [1]. Analytical methods are mainly based on summing up of the areas of the spheres surrounding atoms and on the direct inclusion-exclusion method for taking into account the contributions of two, three, four or more intersecting atomic spheres. The analytical expressions for high degree intersecting areas are so complex that one has to reduce these contributions to a signed sum of lower degree intersecting according to [2]. This work proposes a method for simple estimation of the contributions of many intersecting spheres.

The idea underlying the method is as follows. The relation between the segment area of the S_{seg} sphere and the area of its projection S_{plane} on the tangential plane $S_{seg} = \lambda(x)\, S_{plane}$ is nonlinear, where the $\lambda(x)$ coefficient is defined by the shape and area of the projected segment. In conditions of small relative areas and commensurability of the characteristic sizes of the segment, shape-dependence of the $\lambda(x)$ disappears. As a result, the $S_{seg}/\pi R^2$ value can be chosen as the x parameter. At small values of the parameter x, interpolation of the $\lambda(x)$ coefficient by low degree polynomials is becomes possible.

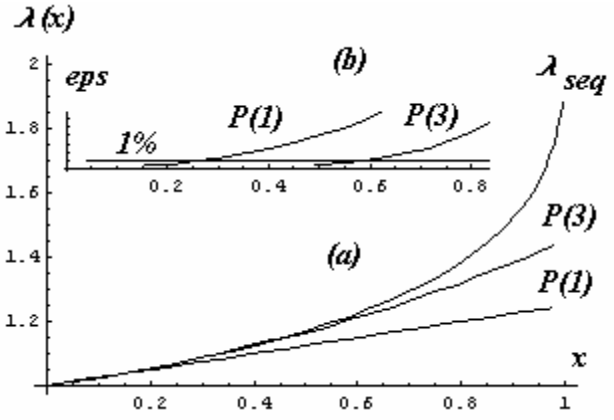


Fig. (a) Dependence of the $\lambda(x)$ coefficient for a spherical segment on the segment area and its interpolation by first and third degree polynomials. (b) Relative error of interpolation for the first and third degree polynomials for the $\lambda(x)$ coefficient.

The proposed method allows obtainment of the area value of high degree intersecting by calculating the area of its projection, followed by recovery of the result using the $\lambda(x)$ coefficient. In turn, it allows improvement of the calculation accuracy of a surface area by explicitly taking into account the contributions of high degree intersecting; there is no increase in computation complexity.

1. Jie Liang, Herbert Edelsbrunner, Ping Fu, Pamidighantam V. Sudhakar, Shankar Subramaniam (1998) Molecular Area and Volume Through Alpha Shape. *PROTEINS: Structure, Function, and Genetics* **33**:1–17.

2. Kratky, K. (1981) Intersecting disks (and spheres) and statistical mechanics. I. Mathematical basis. *J.Stat.Phys.* **25**:619-634.

MICROSATELLITES AND SHORT MINISATELLITES: GENERATION AND DEGENERATION

MARINA V. FRIDMAN[1], VALENTINA BOEVA[1], NINA OPARINA[2], VSEVOLOD J. MAKEEV[1,2]

Tandem repeats with different specific properties can be identified with instruments like TRF [http://tandem.bu.edu/trf/trf.html] and TandemSWAN [http://bioinform.genetika.ru/projects/swan/www/index.html] in genome sequences. Different types of filtration allow studying of evolution of repeats of a different length in genomes. In this presentation we limit ourselves with repeats with a repeat unit length varying from 3 to 25 bp.

Our analysis agrees with earlier observations on distribution of different type of tandem repeats, in particular 4-periodic repeats found in vertebrates genomes. Such patterns are characteristics for repeats identified with different repeat finders, both for precise and fuzzy repeats, which probably indicates that such bias is a result of biased repeat generation rather than with the subsequent degeneration. Moreover, TandemSwan computations indicate that most of the repeated units with length from 3 to 5 b.p. are equally frequent in all chromosomes, which probably also reflects biases in repeat generation.

For repeats with larger periods other biases are characteristic. We are going to discuss the possible reasons in our presentation.

During degeneration a part of microsatellite loci are likely to generate minisatellites with periods multiple to that of the initial microsatellite. In vertebrates this brings about frequent minisatellites with periods multiple to 4, whereas in *Drosophila* minisatellites with periods multiple to 6 are frequent. The less obvious consequence is that such mechanism imply that repeats with prime periods would be underrepresented in genomes, and this indeed can be observed.

Most of indels are found in very fuzzy repeats. Thus, mechanisms responsible for generation of indels, is likely to be also responsible for generation of multiple substitutions.

We failed to observe the reported prevalence of GC-rich minisatellites as compared to GC-rich microsatellites. In reality, all species analyzed has a very limited number of repeats with GC-rich period, and this fraction grows very slowly with the period. The only exception is 3-periodic repeats, which can be

[1] GosNIIgenetika, Moscow, Russia, valeyo@imb.ac.ru
[2] IMB, RAS, Moscow, Russia, makeev@genetika.ru

GC-rich much more often. Our analysis showed that 3-periodic repeats are found mostly in protein coding regions. Fraction of GC-rich repeats in mammals drops with the increased fuzziness, which can indicate the role of CpG methylation in this process.

GC-rich repeats contain less indels that average repeats identified with TRF. Thus, function for which repeats are recruited can require phase conservation.

STATISICAL APPROACH TO THE DESIGN OF SUBSET SEEDS FOR PROTEIN ALIGNMENT

E.FURLETOVA[1], G.KUCHEROV[2], L.NOE[2], M.ROYTBERG[1], I.TSITOVICH[3]

In [1] we have introduced the class of subset seeds and have demonstrated their advantages for the DNA comparison. In this work we study the features of subset seeds with respect to the comparison of amino acid sequences. In case of protein seeds the number of possible seed letters (i.e. subsets of the set of amino acid pairs AP) is $\sim 2^{210}$ thus one has to start with the choice of seed alphabet.

Let us fix a target set T of ungapped alignments of the given length L. We have two probability distributions on T: the *background* distribution, corresponding to the independent distribution of letters in the aligned sequences, and the *foreground* distribution, corresponding to the really interesting alignments. For the moment both distributions are considered as Bernoulli distributions, i.e. amino acid pairs in different alignment positions are considered as iid variables. Having learning sets of protein sequences and "true" alignments (e.g. corresponding substitution matrix BLOSUM62) one can for any amino acid pairs p estimate both its background probability $b(p)$, and its foreground probability $f(p)$. Thus the problem of similarity recognition can be treated as the hypothesis testing problem: one have to decide whether a given alignment was generated according to the background or to the foreground distribution. Let $D = \{p_1, \ldots, p_k\} \subseteq AP$ be a subset seed letter; $b(D) = \sum_{i=1,k} b(p_i)$ and $f(S) = \sum_{i=1,k} f(p_i)$ be its background and foreground probabilities. A letter D is *maximal* if there is no other letter D' such that $f(D') \geq f(D)$ and $b(D') \leq b(D)$. Our idea is

[1] Institute of Mathematical Problems of Biology, 4, Institutskaja str., 142290, Pushchino, Moscow Region, Russia, janny51@rambler.ru, mroytberg@impb.psn.ru
[2] LIFL/CNRS/INRIA, B^at. M3, Campus Scientifique, 59655 Villeneuve d'Ascq, C´edex, France, {Gregory.Kucherov,Laurent.Noe}@lifl.fr
[3] Institute for information transmission problems, 19, B.Karetnyi per., 127994, Moscow, Russia, cito@iitp.psn.ru

to restrict ourselves to maximal letters only. Moreover, following the Neumann-Pearson lemma, we will consider only "Neumann-Pearson" letters (NP-letters) that are defined as follows. Let t be a threshold, then NP[t] = {p $\in$AP| f(p)/b(p) $\geq$t}.

A letter D is *transitive* if D contains all pairs (a,a) and for all p_1, p_2, p_3 $\in$AP if (p_1, p_2), (p_2, p_3) $\in$D then (p_1, p_3) $\in$D. Transitive letters allows for a direct hashing scheme within the database search. Note, that in general NP-letters are not transitive. An embedded transitive alphabet (ETA) is a set of 20 transitive letters {R_1, …, R_{20}} such that $R_1 \subseteq R_2 \subseteq \ldots \subseteq R_{20}$. Transitive letters are in one-to-one correspondence with partitions of the amino acid alphabet A; thus an ETA can be represented as a list of embedded partitions; R_1 is a partition where all amino acids are separated (*Match)*. To find a good ETA we start with R_1 = *Match* and then recursively produce R_{k+1} from R_k according to the following algorithm. For all pairs of classes C_1, C_2 from the partition R_k we find Q(C_1, C_2) = f(Br(C_1, C_2)/ f(Br(C_1, C_2)), where Br(C_1, C_2) = {(a,b) $\in$AP | a $\in$ C_1, b $\in$ C_2}. Then we unite C_1, C_2 having maximal value of Q(C_1, C_2) into one class of R_{k+1}.

We have tested the Neumann-Pearson alphabet of non-transitive seeds and the ETA obtained by the above algorithm (the probability distributions corresponded to the matrix BLOSUM62; alignment lengths were 16 and 32.). For both alphabets we have constructed the seeds of span $\leq$ 5 giving the best selectivity for a given sensitivity. The selectivity of NP-seeds is better but the difference is small (~7%). E.g. for the sensitivity corresponding to the BLAST recommended vector seed we obtained the following selectivities: NP-seeds: 0.00097; transitive seeds: 0.001023; BLAST: 0.00065. The advantage of BLAST is due to a cumulative consideration of a contribution of several positions; the price for this is more complicated hashing scheme of the database search.

The work was supported by grants RFBR 06-04-49249, INTAS 05-1000008-8028

1. G. Kucherov, L. Noe, M. Roytberg. A unifying framework for seed sensitivity and its application to subset seeds. Journal of Bioinformatics and Computational Biology 4 (2006) 553–570

UBIQUITIN SYSTEM AS A MATTER OF SYSTEMS BIOLOGY

MURAT GAINULLIN, ALEJANDRO GARCIA

Ubiquitin system forms one of the most important pathways of cellular regulation. It tightly interplays and functionally overlaps with other cellular processes: signal transduction, control of gene expression, DNA repair, etc. Ubiquitin system consists of large number of dynamically interacting intracellular proteins and includes following constituents: (1) Enzymes combined into a multi-protein complex called ubiquitin-protein ligase (UPL) and catalyzing covalent attachment of ubiquitin or ubiquitin-like proteins (UBLs) to a target protein (i.e. ubiquitylation reaction). Each of three types of proteins (E1, E2 and E3), participating in UPL formation, as well as many optional adaptor proteins, are occur as different isoforms within a living cell. (2) Several enzymes cleaving a covalent bond between ubiquitin and target protein (deubiquitylating enzymes, DUBs). (3) Particular proteins serving as substrates for modification with ubiquitin or UBLs. The number of these proteins is extremely high and comprises a significant part (up to 1/6) of a proteome. (4) Different "ubiquitin receptors", recognizing attached ubiquitin or specifically assembled multiubiquitin chain as definite signal and in such a way bridging the ubiquitylation and subsequent ubiquitin-dependent cellular events (for latest review see [1, 2]). In the light of this complexity the ubiquitin system is ideally suited for a systems biology analysis [3].

At present, the amount of experimental data concerning ubiquitin system grows extensively. On the other hand, the unraveling of the respective functional consequences is not so fast, and many of the ubiquitin-dependent regulatory mechanisms still remain enigmatic. We try to reduce such divergence by systematic multilevel analysis of ubiquitylation using different computational methods. There are several tasks compounding this research. To fill a general need for collecting and systematizing experimental data concerning ubiquitylation we have developed a specialized resource, UbiProt Database, a knowledgebase of ubiquitylated proteins (http://ubiprot.org.ru) [4]. The database contains retrievable information about overall characteristics of a particular protein, ubiquitylation features, related ubiquitylation/de-ubiquitylation machinery and literature references reflecting experimental evidence of ubiquitylation. Currently we evolve a new resource, termed provisionally «Ubiquitin System Knowledge Database» using the BioUML workbench [5]. In general, this project aims to reconstruct the whole ubiquitin system, including protein-protein interaction, characterization of

Nizhny Novgorod State Medical Academy, 10/1 Minin Sq., Nizhny Novgorod, Russia, biochem@gma.nnov.ru

particular ubiquitin-dependent pathways, up- stream regulatory events and down-stream metabolic outcomes.

The dataset collected is subjected to computational analysis, including phylogenetic analysis and methods of structural bioinformatics. This allows us (1) to predict new mode of action of ubiquitylation on the target proteins. Suggested direct steric effects of ubiquitin attachment regulate activity of modified substrate through "loss-of-function" mechanism [6]. (2) First structural classification of HECT-domain UPL is elaborated. This provide new insights into the functional properties of some still not characterized proteins and offers a possibility to further investigation of specificity of this class of ubiquitylating enzymes. (3) Analysis of several biological features of identified target proteins, including their primary and 3D structures, domain architecture, topology and expression profiles, enables us to reveal new consequences concerning principles of target protein selection by particular ubiquitin ligases.

Obviously, the data obtained from these studies could be efficiently used in the future for studying pathogenesis mechanisms of various diseases associated with ubiquitin system defects as well as for development of novel therapeutic compounds.

1. M.Hochstrasser (2006) Lingering mysteries of ubiquitin-chain assembly. *Cell*, **124**, 27-34.
2. G.Nalepa et al. (2006) Drug discovery in the ubiquitin-proteasome system. *Nat.Rev.Drug Discov.*, **5**, 596-613.
3. M.C.Rechsteiner (2004) Ubiquitin-mediated proteolysis: an ideal pathway for systems biology analysis. *Adv.Exp.Med.Biol.*, **547**, 49-59.
4. A.L.Chernorudskiy et al. (2007) UbiProt: a database of ubiquitylated proteins, *BMC Bioinformatics* (in press)
5. F.Kolpakov et al. (2007) CYCLONET--an integrated database on cell cycle regulation and carcinogenesis. *Nucleic Acids Res.*, **35**, D550-D556.
6. A.L.Chernorudskiy et al. (2007) Evaluation of direct effects of protein ubiquitylation using computational analysis, *Biofizika* (in press)

DOES FOLDING NUCLEI COMPETE WITH AMYLOIDOGENIC REGIONS?

OXANA V. GALZITSKAYA, S. O. GARBUZYNSKIY

In addition to "normal," native protein structure, some proteins can also form alternative, misfolded structures. During the past years, it has been shown that some diseases are connected with protein misfolding and the formation of insolu-

Institute of Protein Research RAS, Pushchino, 142290, Russia,
ogalzit@vega.protres.ru

ble aggregates called amyloid plaques. For some proteins which are capable to form amyloid structures, those regions which are important for amyloid formation are already experimentally outlined (they are called amyloidogenic regions). In these proteins, we predicted those residues that are important for "normal" folding (that is, which are involved into the folding nucleus of the native structure) and compared them with amyloidogenic ones. The average of the predicted Φ-values (which reflect the degree of involvement of the amino acid residue into the folding nucleus) over 14 amyloidogenic regions (of 8 globular proteins in which amyloidogenic regions are now localized experimentally) is significantly greater (0.55±0.03) than the average Φ-value averaged over residues outside amyloidogenic regions (0.44±0.01). This demonstrates that amino acid residues in amyloidogenic regions in average are more included in the folding nucleus than amino acid residues from non-amyloidogenic regions. In total, 10 of 14 amyloidogenic fragments are located in the folding nuclei. This means that amyloidogenic regions are typically incorporated into the native structure early during its formation (not later than at the rate-limiting step of a "normal" folding process).

This work was supported by the program "Molecular and cellular biology," by the Russian Foundation of Basic Research (05-04-48750-a), by the INTAS grant (№ 05-1000004-7747) and by Howard Hughes Medical Institute (55005607).

EGOSAP: EVOLUTIONARY GENE ONTOLOGY-BASED SEMANTIC ALIGNMENT OF BIOLOGICAL PATHWAYS

JONAS GAMALIELSSON, BJOERN OLSSON

Given the large number of biological pathways that have been elucidated in later years, there is a great need for methods to analyze these pathways. One class of such methods compares pathways semantically, in order to discover parts that are evolutionarily conserved between species or to discover intra-species similarities. These algorithms should also return alignments between matching pathway fragments. Furthermore, these pathway alignment methods should rely on approximate, rather than exact, matching in biological pathways [1,2], since approximate matching can associate gene products that have different labels but are known to have similar function. Most previous work on such methods has focused on metabolic pathways, using the EC enzyme hierarchy to calculate match scores [3,1] and assuming that the topologies of compared pathways are known. However, sometimes only a set of gene products is avail-

University of Skövde, Box 408, 54128 Skövde, Sweden, jonas.gamalielsson@his.se

able. Some earlier methods [4-6] have tried to map sets of gene products onto known pathways, but merely for presentation purposes and not employing approximate matching using abstraction hierarchies or ontologies.

Here, we propose a method that uses similarity scores based on the Gene Ontology (GO) to find semantic alignments when comparing paths in biological pathways where the nodes are gene products. A known pathway graph is used as model, and an evolutionary algorithm (EA) is used to derive hypothetical paths that are semantically similar to paths in the model graph. The hypothetical paths are evolved from a set of gene products that can be selected by e.g. differential analysis of microarray gene expression data. The method takes advantage of all three sub-ontologies of GO, and uses semantic similarity measures to calculate match scores between gene products, to obtain a fitness measure. In the EA, each individual (hypothetical path) is represented as an initially random permutation from a query set of gene products. A hypothetical path is evolved for each of the longest, non-overlapping, paths in the model pathway. Gaps can also be part of the hypothetical path, creating a path alignment between the model path and the hypothetical path. Individuals are initialized by random permutation of gene products with the same length as the model path. Binary tournament selection is used to select individuals for the next generation and partially mapped crossover (PMX) is used as variation operator, since it guarantees feasible offspring. Further, a mutation operator is used to randomly switch gene products between two positions and to introduce new gene products from the query set. An elitist strategy is used to improve the EA's performance. A statistical test of significance shows whether an evolved alignment would be likely to appear by chance or not. Our evaluation of the method shows that given an artificial model pathway containing ~100 gene products, the EA is able to evolve a semantically perfect path that is statistically significant using the gene products of the model path as a query alphabet. This demonstrates that the EA is competent. We also show that the method is useful for studying real regulatory pathways for organisms such as *S. cerevisiae* and *M. musculus*. The proposed method is applicable to all types of biological pathways where nodes are gene products, e.g. regulatory pathways, signaling pathways and metabolic enzyme-to-enzyme pathways.

1. R. Y. Pinter et al. (2005) Alignment of Metabolic Pathways, *Bioinformatics*, **21:** 3401-3408
2. M. Koyutürk et al. (2004) An efficient algorithm for detecting frequent subgraphs in biological networks, *Bioinformatics*, **20:** i200-i207

3. Y. Tohsato et al. (2000) A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy, *ISMB 2000*, pp 376-383
4. K. D. Dahlquist et al. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, *Nature Genetics*, **31:** 19-20
5. P. D. Karp et al. (2002) The pathway tools software, *Bioinformatics*, **18:** S1-S8
6. H-J Chung et al. (2004) *Nucleic Acids Research*, **32:** W460-W464

VISUALIZATION AND FUNCTIONAL ANNOTATION OF COMPLETE GENOME SEQUENCES BY THE SEQWORD GENOME BROWSER

GANESAN H., RAKITIANSKAIA A.S., REVA O.N.

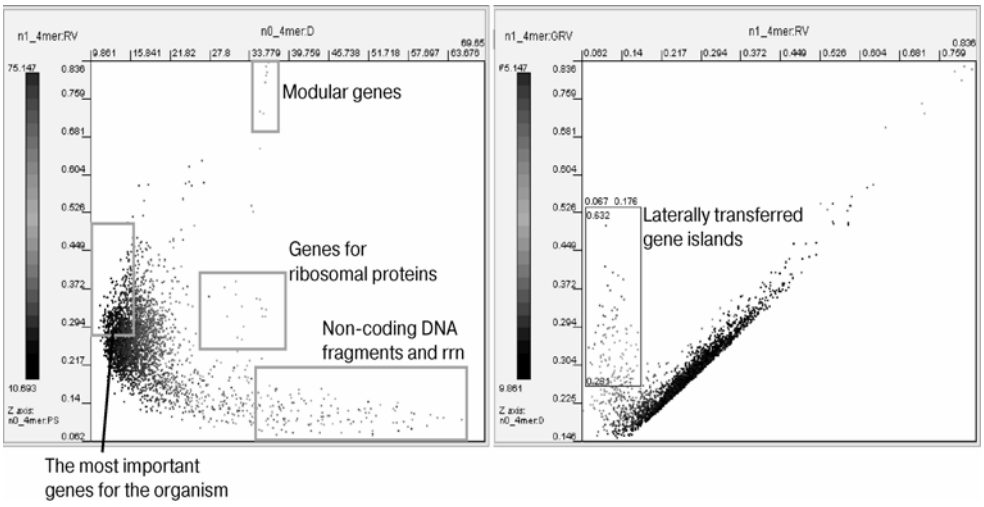
Complete sequencing of bacterial genomes has become a common technique of present day microbiology. Thereafter, data mining in the complete sequence is an essential step. New *in silico* methods are needed that rapidly identify the major features of genome organization.

We tested the usefulness of local oligonucleotide usage (OU) patterns to recognize and differentiate types of atypical oligonucleotide composition in DNA sequences of bacterial genomes. The order of nucleotides is governed not only by the encoded information, but also by physical and biological constraints. All sections of the genome should be exposed to the same constraints and consequently should have similar fingerprints of oligonucleotide frequencies. The frequency of each tetranucleotide is indeed approximately the same throughout the genome (Weinel *et al.*, 2002). However, there are always some regions that to occur which exhibit an atypical oligonucleotide composition indicating that this DNA has been exposed to particular constraints other than those seen in the bulk of the genome.Loci with alternative OU patterns termed as gene islands in fact belonged to different functional classes. We introduced into the literature some useful OU statistical parameters such as local pattern deviation (D), pattern skew (PS) and OU variance (OUV) to allow detection, visualization and distinguishing of gene islands (Reva and Tümmler, 2004, 2005).

Several procedures of identification of gene islands of different classes, namely: horizontally transferred elements, clusters of genes for ribosomal RNA and proteins, large multidomain genes, non-coding pseudogenes and the core genes which are the most important for a given organism to survive in its particular eco-niche; have been developed and published recently (Reva and

University of Pretoria, Biochemistry Dep., Bioinformatics and Computational Biology Unit, Lynnwood Road, Hillcrest, Pretoria 0002, South Africa,
oleg.reva@up.ac.za

Tümmler, 2005). All these algorithms were implemented in a Java-applet *SeqWord Genome Browser* linked to a database of OU patterns calculated for 168 bacterial genomes. The applet allows visualization of structural polymorphism of bacterial chromosomes, grouping genomic fragments by selected OU parameters (see figure below) and an automatic retrieving of loci of interest.



In this figure the dots on the plot correspond to overlapping genomic fragments of 8 kb stepping 2 kb along the chromosome. DNA fragments belonging to separate functional classes are outlined. The Web-based application of the program is freely available at the address: http://www.bi.up.ac.za/ Genome-Browser/.

Acknowledgements: this work was supported by the National Bioinformatics Network of South Africa (http://www.nbn.ac.za/).

1. C. Weinel *et al.* (2002). Global features of the *Pseudomonas putida* KT2440 genome sequence. *Envir. Microbiol.* **4:** 809-818.
2. O.N. Reva, B. Tümmler. (2004). Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* **5**: 90.
3. O.N. Reva, B. Tümmler. (2005). Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**: 251.

PREDICTION OF FOLDING RATES OF PROTEINS

SERGIY O. GARBUZYNSKIY, DMITRY N. IVANKOV, DANIELLE C. REIFSNYDER, NATALIA S. BOGATYREVA, ALEXEI V. FINKELSTEIN, OXANA V. GALZITSKAYA

Search and study of general principles that govern kinetics and thermodynamics of protein folding generate a new insight into the factors controlling this process. We demonstrate that it is possible to predict folding rate of a globular protein from its spatial structure [1, 2] and even from its sequence [3, 4, 5].

The folding rates obtained from dynamic programming method, method of kinetic equations and Monte Carlo simulations (all based on modeling of reversible unfolding of native spatial structure) correlate reasonably well with logarithms of experimentally measured folding rates at mid-transition [1, 2].

It has been shown that chain length is the main determinant of the folding rate for proteins with three-state folding kinetics [3]. The logarithms of folding rates of such proteins in water strongly anti-correlate with their chain length (the correlation coefficient being -0.80). However, chain length has no correlation with the folding rate (the correlation coefficient is -0.07 only) for two-state folding proteins [3]. On the other hand, a gross parameter reflecting protein topological complexity (contact order, *CO*) correlates with the folding rates of two-state folding proteins but not of three-state ones [6, 3, 7]. Absolute contact order parameter (equal to *CO*×length) allows successful predicting of folding rates of both two-state and three-state folding proteins [7].

Our analysis demonstrates that α/β proteins have both the greatest number of contacts and the slowest folding rates compared to proteins from the other structural classes [5, 8]. Since α/β proteins are thought to be the oldest proteins [9], it can be suggested that proteins have been evolved to pack more quickly and into looser structures [8].

This work was supported by the programs “Molecular and Cell Biology” and “Fundamental sciences – medicine” of Russian Academy of Sciences, by the Russian Foundation of Basic Research (05-04-48750-a), by the INTAS grant (№ 05-1000004-7747) and by Howard Hughes Medical Institute (55005607). D.C.R. was supported by a Fulbright Fellowship.

1. S.O.Garbuzynskiy, A.V.Finkelstein, O.V.Galzitskaya (2004) Outlining folding nuclei in globular proteins, *J. Mol. Biol.* **336:**509–525.
2. O.V.Galzitskaya, S.O.Garbuzynskiy, A.V.Finkelstein (2005) Theoretical study of protein folding: outlining folding nuclei and estimation of protein folding rates, *J. Phys. Condensed Matter*, **17:**S1539–S1551.

Institute of Protein Research RAS, Pushchino, 142290, Russia,
ogalzit@vega.protres.ru

3. O.V.Galzitskaya, S.O.Garbuzynskiy, D.N.Ivankov, A.V.Finkelstein (2003) Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics, *Proteins*, **51:**162–166.
4. D.N.Ivankov, A.V.Finkelstein (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure, *Proc. Natl Acad. Sci. USA*, **101:**8942–8944.
5. O.V.Galzitskaya, S.O.Garbuzynskiy (2006) Entropy capacity determines protein folding, *Proteins,* **63:**144–154.
6. K.W.Plaxco, K.T.Simons, D.Baker (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277:**985–994.
7. D.N.Ivankov, S.O.Garbuzynskiy, E.Alm, K.W.Plaxco, D.Baker, A.V.Finkelstein (2003) Contact order revisited: Influence of protein size on the folding rate, *Protein Sci.* **12:**2057–2062.
8. O.V.Galzitskaya, D.C.Reifsnyder, N.S.Bogatyreva, D.N.Ivankov, S.O.Garbuzynskiy (2007) More compact protein globules exhibit slower folding rates, *Proteins*, in press.
9. H.F.Winstanley, S.Abeln, C.M.Deane (2005) How old is your fold? *Bioinformatics*, **21:**i449–i458.

MUTABLE SITES ARE UNDER STRONGER NEGATIVE SELECTION

A. GERASIMOVA[1,2], F. KONDRASHOV[3], S. SUNYAEV [4], A. KONDRASHOV[2]

In this study we investigate correlation between mutation rate and the strength of negative selection within the same functional class of sites. We considered human nonsynonymous protein-coding sites and subdivided them into just two classes - those within and those outside CpG contexts, because in mammals this context excerpts by far the strongest influence on the rate of mutation at a site[1]. Then, we compare the rates of human-chimpanzee divergence and have found that the strength of negative selection within nonsynonymous coding sites of the human genome is substantially higher within hypermutable CpG contexts.

We used four sources of information on the impact of CpG on the transition and transversion rate in humans: direct data on Mendelian diseases[2], data on human-chimpanzee genome alignment[3], our data on ((human,chimpan-

[1] State Scientific Center GosNIIGenetika, Moscow, Russia,
a_gerasimova@yahoo.com
[2] Live Science Institute, University of Michigan, Ann Arbor, US
[3] Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, USA
[4] Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, USA

zee),orangutan) genome alignment, and data on evolution of a wider sample of mammals[1]. Below, we will assume that in humans CpG context increases the rate of transition mutations by the factor of 13.2, and the rate of transversion mutations by the factor of 3.3.

We produced ((human, chimpanzee), macaque) mRNA alignments to compare rates of a particular nucleotide substitutions that causes a particular amino acid replacement when this substitution occurs within vs. outside CpG context.

The genetic code provides 4, 5, and 8 kinds of amino acid replacements that are suitable for such analysis of transitions at 1st, 2nd, and 3rd positions of codons, respectively; for transversions the corresponding numbers are 2, 3, and 4, respectively (table not shown).

After that we calculated the average impacts of CpG context on the rate of non-synonymous evolution, together with the average impacts of CpG context on the probability of fixation of non-synonymous transitions and transversions. Accordingly, in coding regions CpG context increases the rate of transition mutations by the factor of 7.5, and the rate of transversion mutations by the factor of 2.2.

Thus, a non-synonymous transition that occurred inside CpG context will be fixed with the probability that is only ~60% of the probability of fixation of a non-synonymous transition outside CpG context. For transversions, the corresponding figure is ~70%.

1. Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA 101: 13994-14001.
2. Kondrashov AS (1995) Modifiers of reproduction under the mutation-selection balance: general approach and the evolution of mutability. Genetical Res 66: 53-69.
3. Ebersberger I, Metzler D, Schwarz C, Paabo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet. 70(6):1490-1497.

HIGH-THROUGHPUT IDENTIFICATION OF CATALYTIC REDOX-ACTIVE CYSTEINE RESIDUES AND SELENOPROTEIN GENES

VADIM N. GLADYSHEV, DMITRI E. FOMENKO,
GREGORY V. KRYUKOV, ALEXEY V. LOBANOV

Cysteine (Cys) residues often play critical roles in proteins, however, identification of their specific functions has been limited to case-by-case experimen-

Redox Biology Center and Department of Biochemistry, University of Nebraska, Lincoln, NE 68588 USA, E-mail: vgladyshev1@unl.edu

tal approaches. We describe a procedure for large-scale detection of catalytic redox-active Cys through homology to sporadic selenocysteine (Sec)-containing proteins. This method is not dependent on sequence motifs, structure and origin of the sequences and first identifies unique Cys/Sec pairs flanked by homologous sequences within the universe of translated nucleotide sequences; these pairs then serve as seeds for sequence analysis at the level of protein families and subfamilies. A variation of this method also allows identification of selenoprotein genes. Application of this method identified majority of known proteins containing catalytic redox-active Cys, while filtering out proteins in which conserved Cys are involved in other functions, such as non-redox catalysis, structural disulfides, posttranslational modifications and binding of certain metals. Moreover, for oxidoreductases containing multiple conserved Cys, the identity of the attacking catalytic Cys could be identified. We predicted redox-active Cys in several proteins, and directly verified the prediction in an S-adenosyl methionine-dependent methyltransferases family. Rapid accumulation of sequence information from genomic and metagenomic projects should allow detection of many additional oxidoreductase families as well as identification of redox-active Cys in these proteins. This method should also lead to a better understanding of composition, evolution and function of selenoproteomes.

1. Fomenko, D. E., Xing, W., Adair, B. M., Thomas, D. J., and Gladyshev, V. N. (2007) High-throughput identification of catalytic redox-active cysteine residues. *Science* **135**, 387-389.
2. Kryukov, G. V., Castellano, S., Novoselov, S. V., Lobanov, A. V., Zehtab, O., Guigo, R., and Gladyshev, V. N. (2003) Characterization of mammalian selenoproteomes. *Science* **300**, 1439-4313.

EVOLUTIONARY HISTORY OF BACTERIOPHAGES WITH DOUBLE-STRANDED DNA GENOMES

GALINA GLAZKO[1], JING LIU[12], VLADIMIR MAKARENKOV[3], ARCADY MUSHEGIAN[14]

Reconstruction of evolutionary history of bacteriophages is considered to be difficult, because of fast sequence drift and extensive gene shuffling in phage genomes, both of which must reduce phylogenetic signal. We compiled the pro-

[1] Stowers Institute for Medical Research, Kansas City, Missouri, USA, and University of Rochester, USA
[2] Interdisciplinary Graduate Program in Biomedical Sciences, Kansas University Medical Centre, Kansas City, Kansas, USA
[3] Departement d'informatique, Université du Québec à Montreal, QC, Canada
[4] Department of Microbiology, Kansas University Medical Centre, Kansas City, Kansas, USA , arm@stowers-institute.org

files of presence and absence of 803 orthologous genes in 158 completely sequenced phages with double-stranded DNA genomes and used these gene content vectors to infer the evolutionary history of phages. There were 18 well-supported clades, mostly corresponding to accepted genera, but in some cases appearing to define new taxonomic groups. Phylogenetic conflicts between this dendrogram and trees constructed from sequence alignments of phage proteins were exploited to infer 241 specific acts of intergenome recombination. Thus, a notoriously reticulate evolutionary history of fast-evolving phages can be reconstructed in substantial detail by quantitative comparative genomics.

IGLA-3D: A MODULAR ALGORITHM FOR PAIRWISE THREE-DIMENSIONAL PROTEIN STRUCTURE ALIGNMENT

IRINA V. GLOTOVA

We have developed a modular intermolecular algorithm for pairwise three-dimensional protein structure alignment, IGLA-3D. This algorithm combines four modules each of which solves a single problem (construction of an initial alignment, finding the best superposition of the geometric core of the corresponding structures, identification of the alignment geometric core and homology prediction, respectively). The advantage of IGLA-3D modularity is that it gives the opportunity to change independently any of the four modules and find the optimal combination of methods that may lead to a better solution of the protein structure alignment problem. In the current implementation of the algorithm comparison of the secondary structure element trees of the protein structures, rigid superposition of protein chain backbones and dynamic programming are applied.

To test the performance of the algorithm we used a set of 482 homologous and 458 non-homologous protein structures whose sequence identity is less than 25% [1]. IGLA-3D structural alignments were compared to the alignments obtained by two well-known structure alignment programs, DaliLite [2] and TM-Align [3]. Comparison of the test results showed that the number of false homology predictions for our algorithm is higher than for DaliLite and TM-Align. As for the geometric characteristics of the alignments, according to the evaluation of SI and MI values [4] IGLA-3D finds better alignments in more cases than DaliLite but in less cases than TM-Align. However, if the normalized root mean square deviation,

A.N.Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow119899, Russia, igbox@mail.ru

NRMSD is used as a measure for the geometric quality of the alignments, IGLA-3D gives better results than both DaliLite and TM-Align.

For the improvement of our algorithm performance some alterations are supposed to be made in each of IGLA-3D modules. In particular, these alterations may include finding a number of different initial alignments, identification of geometric cores through local alignments, as well as homology prediction based on similarity measures combining geometric and biological scores.

My special thanks to Vladimir K. Nikolaev and Vladimir V. Galatenko for their help and useful suggestions.

1. Liisa Holm and Chris Sander (1997) Decision support system for the evolutionary classification of protein structures, American Association for Artificial Intelligence.
2. Liisa Holm and Jong Park (2000) DaliLite workbench for protein structure comparison,
   Bioinformatics 16: 566-567.
3. Yang Zhang and Jeffrey Skolnick (2005), TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Research, 33(7): 2302-2309.
4. Rachel Kolodny et al. (2005) Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures, Journal of Molecular Biology, 346:1173-1188.

ATGC, SOFTWARE FOR NUCLEOTIDE SEQUENCE ANALYSIS

PAVEL K. GOLOVATENKO-ABRAMOV

In molecular biology research, scientist often faces necessity of analyzing nucleotide sequences in different ways. Today, there are a plenty of software products, web-based and local, which help researcher to solve his task in bio-informatics area. However, every such software has its disadvantages, from high system requirements (in case of desktop applications) or low processing speed due to connection problems (in case of online services) to complicated user interface and poorly tuned or even bugged operating algorithms. Particular problem can rise because of relatively high price for software. For all those reasons we developed new desktop application – ATGC – that solves few general tasks in sequence analysis.

ATGC is a pack of useful bio-informatics algorithms for exon alignment using mRNA and genomic DNA sequences as source data, docking multiple se-

Vavilov Institute of General Genetics, Gubkin St., 3, 119991 Moscow, Russia,
p.golovatenko@gmail.com

quences into contig, searching for forward, reversed and reverse-complemented repeats within source nucleotide sequence, searching for consensus sequences within source nucleotide sequence, and amino acid sequence prediction. Short description of core algorithms is given below.

Exon alignment algorithm is based on comparison of mRNA and genomic DNA fragments of fixed length by shifting comparison window along whole sequence. Length of compared fragments equals user-determined fixed value while running through intron, and changes to one while running through exon, up to next intron or end of DNA sequence. The extension of this algorithm implicates determining whether splice site consensus appears in point or presumable exon-intron switching. If not, the program shifts comparison window few positions back and forward to refine the border.

ATGC has two algorithms for docking sequences into contig. First (slow but precise) aligns every sequence with each other by calculating homology of overlapped regions while shifting one sequence towards another along whole length of chain. The solution appears in position of maximum homology. Homology is calculated as number of congruent nucleotides in corresponding positions in overlapped region. Congruency between corresponding nucleotides is calculated as number of conjunction results related to number of disjunction results between pair members in alphabet of polymorphic nucleotides. To substantiate this criterion, we used abstract model of nucleotide logical disjunction and conjunction. Second variant of docking algorithm is much faster, and it consists in shifting one sequence towards another until initially determined fixed-length non-ambiguous region on the edge of one sequence is not equal to corresponding region of another one.

Algorithm for searching for consensus subsequences within nucleotide sequence is based on regular expression expansion method. Program processes user-defined query with specified consensus structure including nucleotides, nucleotide blocks and gaps, and forms an array of possible exact matches which are then being sought by processing source nucleotide sequence. Method of searching for repeats within nucleotide sequence is also based on this algorithm, and implicates procedures of reversion and complement transitions.

In ATGC, the conception of graphical sequence representation is used, which makes manual manipulations with sequences (like manual shifting and alignment) much easier, as compared to plain text view mode. Application is performed with user-friendly, intuitively understood interface. We propose ATGC as free open-source software, which can be used for research, needs, as well as for educational purposes in bio-informatics.

A DATABASE SEARCH AND RETRIEVAL SYSTEM FOR THE ANALYSIS AND VIEWING OF BOUND LIGANDS, ACTIVE SITES, SEQUENCE MOTIFS AND 3D STRUCTURAL MOTIFS

ADEL GOLOVIN, KIM HENRICK

The three-dimensional environments of ligand binding sites have been derived from the parsing and loading of the PDB entries into a relational Macromolecular Structure Database (1). For each bound molecule the biological assembly of the quaternary structure has been used to determine all contact residues and a fast interactive search and retrieval system has been developed. The database was extended with small 3D structural motifs, PROSITE (2) patterns and profiles, Catalytic Sites. Novel algorithms for chemical substructure search, for / sequences search, for sequence patterns search, for super-secondary structure motifs matches and for small 3D structural motif groups searching are incorporated. The search engine is integrated with NCBI BLAST(3) sequence search where PSI-BLAST search is used. The interface provides functionality for visualization and creating a search criteria, and provides sequence and 3D multiple alignment options. It is a self integrated system where a results detail page is a search form itself and where search criteria can be refined. A large set of statistics is available for search. This set includes small 3D motifs binding statistics with respect to a ligand fragment library, distribution of small 3D structural motif sequences, occurrence of an amino-acid within a motif, correlation of amino-acids side-chain charges within a motif and Ramachandran plots for each residue in a motif. Statistics with respect to ligands include distributions of interactions against sets of amino-acids, nucleic-acids, motifs and catalytic sites. These statistics can be calculated on different sequence and structure families including PFAM(4), SCOP(5), CATH(6) and the Enzymes Catalogue(7) where a relative risk of a feature is highlighted for a particular family. Access to the data is also provided through the distributed Annotation System (DAS) protocol with the annotations available worldwide through DAS clients like SPICE(8), ENSEMBL(9) and DASTY. As a web service it is provided XML request with XML response where the output XML complies with the eFamily(38) scheme.

The database and its query interface for ligands search MSDsite(10) is available at http://www.ebi.ac.uk/msd-srv/msdsite , and integrated search engine with small 3d structural motifs MSDmotif is available at http://www.ebi.ac.uk/msd-srv/msmotif.

EMBL-EBI, Hinxton Hall, Genome Campus, Cambridge, UK, golovin@ebi.ac.uk, henrick@ebi.ac.uk

MSDsite project is funded by European Commission as the TEMBLOR, contract-no. QLRI-CT-2001-00015 under RTD program "Quality of Life and Management of Living Resources". MSDmotif is the BioSapiens project which is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health" contract number LHSG-CT-2003-503265.

1. Boutselakis, H., Copeland, J., Dimitropoulos, D., Fillon, J., Golovin, A., Henrick, K., Hussain, A., Ionides, J., John, M., Keller, P., Krissnel, E., McNeil, P., Naim, A., Newman, R., Oldfield, T., Pineda, J., Rachedi, A., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tate, J., Tromm, S., Velankar, S. and Vranken, W., (2003)
E-MSD: the European Bioinformatics Institute Macromolecular Structure Database
Nucleic Acids Research, 31, 458-462
2. Hulo N., Sigrist C.J.A., Le Saux V., Langendijk-Genevaux P.S., Bordoli L., Gattiker A., De Castro E., Bucher P., Bairoch A. Recent improvements to the PROSITE database. Nucleic Acids Res. 32:134-137(2004).
3. Altschul SF, HYPERLINK "http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Search&itool=pubmed_Citation&term=" Madden TL, HYPERLINK "http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Search&itool=pubmed_Citation&term=" Schaffer AA, HYPERLINK "http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Search&itool=pubmed_Citation&term=" Zhang J, HYPERLINK "http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Search&itool=pubmed_Citation&term=" Zhang Z, HYPERLINK "http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Search&itool=pubmed_Citation&term=" Miller W, HYPERLINK "http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Search&itool=pubmed_Citation&term=" Lipman DJ. GappedBLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25:3389-3402
4. Eric L.L. Sonnhammer, Sean R. Eddy, Ewan Birney, Alex Bateman, Richard Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research*
5. Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). *SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540*
6. Orengo CA, Thornton JM. Protein families and their evolution - A structural perspective. (2005) *Annual Review of Biochemistry*. Vol 74. p. 867-900.

7. Schomburg, D., Schomburg, I. Springer Handbook of Enzymes. *(2001) 2nd Ed. Springer, Heidelberg*
8. Andreas Prlic, Thomas A. Down and Tim J. P. Hubbard. Adding some SPICE to DAS. *Bioinformatics Volume 21, suppl_2 Pp. ii40-ii41*
9. T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark and E.Birney Ensembl 2005 *Nucleic Acids Research 2005 Jan 1;33 Database issue:D447-D453.*
10. Golovin A., Dimitropoulos D., Oldfield T., Rachedi A. and Henrick K. (2005) MSDsite: A Database Search and Retrieval System for the Analysis and Viewing of Bound Ligands and Active Sites. *PROTEINS: Structure, Function, and Bioinformatics 58(1): 190-9*

RECONSTRUCTION OF ANCESTRAL REGULATORY SIGNAL ALONG A PHYLOGENY

K. GORBUNOV, D. RADIONOV, O. LAIKOVA, M. GELFAND, V. LYUBETSKY

Well defined tasks are reconstruction of phylogenetic history of genes, species and inference of gene evolution events. A gene coding for a regulatory protein evolves together with its binding sites. We define a new task: reconstruction of ancestral regulatory signal (binding sites) on a known phylogenetic tree *G* of the regulatory protein (e.g., a transcription factor). Sequences of extant binding sites are assigned to leaves of tree *G* and aligned for each leaf separately. Ancestral sequences and their properties, e.g. base frequencies, are now to be reconstructed in nodes of tree *G*. Along some edges of tree *G* "abrupt" changes in the binding sites, for example because of regulatory protein changes, might have occurred; we aim at detecting such edges as well. Let us call such edges *evolutionary important*, and the their set *carrier of evolutionary scenario*. An evolutionary scenario also includes the distribution pattern of reconstructed ancestral binding sites or their properties (e.g., base frequency matrices) in nodes of tree *G*. The carrier and the pattern are defined via minimizing total change of the site or its property over all edges outside the carrier

Institute for information transmission problems RAS, Bolshoy Karetnyi lane, 19, 127994, Moscow, Russia, gorbunov@iitp.ru, lyubetsk@iitp.ru

according with maximum parsimony criterion. We developed an effective search algorithm that is robust against minor changes in topology of tree *G*, multiple alignment and properties of extant binding sites. The algorithm accounts for reliability of extant data and edges of tree *G*, as the lengths of edges. This approach to infer signal evolution along the phylogeny of its regulatory factor can be easily applied to study more complex systems, e.g., classic attenuation regulation.

The algorithm was used to make the following inferences: signal **NrdR**: an evolutionary important edge is at the origin of taxon *Thermus/Deinococcus*; signal **MntR**: an evolutionary important edge is at the origin of Corynobacteria; **LacI** family: evolutionary important edges are origins of two taxa - {EC_Laci, YE_Laci} and {YE_SerR, P37077, Q9F497, PM_SerR}+{EC_FruR, PA_FruR, VCA0519}. For individual sites of the signal, good quality scenarios appeared to precisely correspond to four palindrome pairs (5,16), (7,14), (8,13), (10,11) of the signal sites, i.e. the algorithm finds good quality scenarios that reflect palindrome structure of the signal, while the algorithm's objective function does not take it into account. Besides, for palindrome pair (9,12) four nodes can be marked in the tree of the protein factor: the ancestor is highly A-T rich, in its descendant A-T and G-C frequencies are similar, and the next generation descendants become largely different in A-T and G-C content: one of them remains A-T reach, the other becomes G-C reach. This result supports the idea about transition phase of low conservativity in nucleotide fixation during evolution; notably, it this case it is observed in the pair of complementary positions in accordance with palindrome structure of the signal. The next generation descendants mentioned above are taxa {BHI855, DF_ScrR, Q9ZHJ8, P74892, SAV_ScrR} and {VCA0654, P24508}; the ancestral taxon also includes {YE_TreR, EC_TreR, VC0909}.

For family **Fnr_crp**_azot, evolutionary important edges lead to taxon {*Clostridium difficile*, *Clostridium thermocellum*, *Treponema denticola*} and two subsequent edges to taxon {*Desul/fovibrio desulfuricans* G20, *Desulfovibrio vulgaris*}. For individual sites, good quality scenarios precisely correspond to four palindrome pairs of the signal: (6,13), (5,14), (4,15), (3,16).

For family **Irr**_alpha, evolutionary important two subsequent edges lead to taxon {*Rhodopseudomonas palustris*}, one edge – to taxon {*Brucella melitensis*}, and one edge – to the ancestor of two taxa, {*Nitrobacter winogradskyi* Nb-255} and {*Nitrobacter hamburgensis* X14}. Good quality scenarios precisely correspond to five palindrome pairs (4,18), (5,17), (7,15), (8,14), (9,13).

For family **Fur**_delta (subfamily of the Fur family), evolutionary important edges are at the root marking the family divergence into taxa {*Desulfovibrio*

vulgaris, *Desulfovibrio desulfuricans* G20} and the union of {*Desulfuromonas acetoxidans*} and {*Geobacter sulfurreducens* PCA, *Geobacter metallireducens*}. Similarly, scenarios were inferred for other signals.

CREATING A CRITICAL MASS OF DATA FOR GENOME ANNOTATION AND COMPARATIVE ANALYSIS

IGOR V GRIGORIEV

High-throughput sequencing a large number of genomes facilitates better genome annotation and comparative analysis. Sampling the tree of life both broadly and deeply allows analysis of differences between and within the clades on genomic level as well as identification of core sets of genes conserved between the genomes.

Over two dozens of eukaryotic genomes were sequenced and annotated at the DOE Joint Genome Institute (http://genome.jgi-psf.org) in the last few years. These broad and diverse set of genomes represents the major branches of the tree of life and hence enables efficient comparative analysis. Using same types of data and tools for assembly and annotation of different genomes minimize errors in comparative analysis. Analysis of multiple representatives of a clade improves quality of gene prediction and functional annotation [1-2]. Finished genomes can provide interesting insights into genome organization and evolution, such as comparative analysis of two *Ostreococcus* genomes [1]. Additional experimental data triggered by genome sequencing projects such as ESTs, proteomics and whole-genome microarrays assist with genome annotation and better understanding genome biology [3-5]. Using JGI Genome Portal over 500 biologists around the world manual curate predicted genes and functions in annotated genomes and improve their quality.

Creating a critical mass of genomes, supplementary experimental data, and users improves genome annotation and enables efficient comparative analysis to address a broad spectrum of biological questions and to find important industrial applications.

1. Palenik, B., et al. (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A, 104, 7705-7710.
2. Tyler, B.M., et al., (2006) Phytophthora Genome Sequences Uncover Evolutionary Origins and Mechanisms of Pathogenesis. Science, 313, 1261-1266.

US DoE Joint Genome Institute, Walnut Creek, CA 94598,USA, ivgrigoriev@lbl.gov

3. Jeffries, T.W., et al (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. Nature Biotech., 25, 319-326
4. Tuskan, G.A., et al (2006) The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science, 313, 1596-1604.
5. Vanden Wymelenberg, A., et al. (2006) Computational analysis of the Phanerochaete chrysosporium genome database and mass spectrometry identification of peptides in ligninolytic cultures reveal complex mixtures of secreted proteins. Fungal Genet Biol,. 43, 343-356.

THE HEDGEHOG SIGNALING CASCADE SYSTEM: EVOLUTION AND FUNCTIONAL DYNAMICS

K.V. GUNBIN*, D.A. AFONNIKOV, L.V. OMELYANCHUK N.A. KOLCHANOV

The theoretical studies of eukaryote morphogenesis currently tend to 1) build and analyze numerical models of morphogenesis, and 2) make feasible comparative and/or evolutionary analyses of genes controlling morphogenesis. When intricate molecular-genetic systems are simulated, insights are gained into how the systems functions so that, importantly, the system responses to changes in kinetic parameters can be estimated, and the classification of kinetic parameters can be made. Researchers are usually interested in those molecular-genetic system parameters whose small changes cause a strong response in the system's function (the hyper-responsive parameters), and, conversely, parameters whose significant changes cause no significant changes in the molecular-genetic system response (the inert parameters). Genes that define the hyper-responsive parameters are of biological interest. The morphogenetic process may prove to be very sensitive to impairment or weak mutational changes in the function of these genes. This makes their research timely and appropriate, particularly with reference to cancerogenesis and/or formation of developmental abnormalities. Here we study the Hh signaling cascade (Lum, Beachy, 2004) and compare the results we obtained using the model of the insect Hh-cascade (Gunbin *et al.*, 2007a) with those yielded by evolutionary analysis of the genes involved in the functioning of the cascade (Gunbin *et al.*, 2007b).

Taking into consideration the differences in the Hh-cascade for signal transduction between vertebrates and invertebrates, we compared the parametric robustness of the Hh-cascade model in invertebrates (Gunbin *et al.*, 2007a) and vertebrates (Lai *et al.*, 2004). As a result, we identified a parameter set whose

Institute of Cytology and Genetics SB RAS, Lavrentyev Ave. 10, Novosibirsk, 630090, Russia, * Corresponding author: genkvg@bionet.nsc.ru

small changes cause a strong response in the model's normal behavior and, in this way, we identified a set of hyper-responsive genes of the Hh-cascade (Table). The analysis of the molecular evolution of genes of the Hh-cascade was successful in that it allowed us to: 1) identify the genes that at the divergence step of the major Bilateria types (Ecdysozoa and Deuterostomia) evolved in the positive selection mode and 2) demonstrate that the genes are predominantly those we identified as developmental (Gunbin *et al.*, 2007b).

Table compares the evolutionary mode of the Hh-cascade genes (proteins), their functional load and response type. From the tabulated data it follows that, of the 9 genes under study, 5 can be assigned to the hyper-responsive genes and 2 of the 9 to the potentially hyper-responsive genes. It is of interest that all these genes belong to the developmental according to their functions. Another interesting feature, also shown in Table, is the consistency between the response type of a gene (hyper-response present) and the positive selection mode of the gene at the divergence time of the major Bilateria groups (Ecdysozoa and Deuterostomia). Of the 7 hyper-responsive genes, the evolutionary mode under positive selection was identified for 6, and the positive selection mode was not identified for 2 not hyper-responsive genes.

Table. Relation between gene evolution modes, divergence of Bilateria taxonomic types, and hyper-responsive kinetic parameters.

Protein (gene) name	Hyper-responsive kinetic parameters of the models corresponding to protein function	Type of the network response corresponding to change in kinetic parameters (+ - hyper-response; - -inertness; + - - intermediate effect)	Functional protein group	Events of positive selections related with divergence of taxonomic Bilateria types (Gunbin *et al.*, 2007b)
Hh	D_H **;K_{Shh} *	+ **,*	Developmental	+
Ptc	kd_1, M, k_5 **; K_{Ptc} *	+ **,*	Developmental	-
Smo	k_5 **; K_{Ptc} *	+ **,*	Developmental	+
Disp	D_H **; K_{Shh} *	+ **,*	Developmental	+
PKA	-	- **,*	Housekeeping	-
Slmb	-	- **,*	Housekeeping	-
Su(Fu)	kd_3**	+ - **	Developmental	+
Fu	kd_3**	+ - **	Developmental	+
Ci	k_0 **; $\upsilon_{max,G}$, K_2 *	+ **,*	Developmental	+

Designations: models for * - the Hh-cascade in vertebrates (Lai *et al.,* 2004); ** - the Hh-cascade in invertebrates (Gunbin *et al.*, 2007a).

Thus, genes of hyper-responsive type may be of particular importance for the compensatory changes during the structural reorganization of the Hh signaling cascade. Even small changes in their function might have resulted in great changes in the function of the entire network and, hence, these genes could serve as good candidate genes for "the within" sources of compensatory shift produced by mere point mutations.

The work is supported by Russian Foundation of the Basic Research (05-04-49141-a, 05-07-98012-p, 03-04-48506-a).

1. K.V. Gunbin, L.V. Omelyanchuk, V.V. Kogai, S.I. Fadeev, N.A. Kolchanov (2007a) Model of the reception of Hedgehog morphogen concentration gradient: comparison with an extended range of experimental data, *Journal of Bioinformatics and Computational Biology,* (In press).
2. K.V. Gunbin, D.A. Afonnikov, N.A. Kolchanov (2007b) The evolution of the Hh-signaling pathway genes: a computer-assisted study, *In Silico Biology,* (In press).
3. K. Lai, M.J. Robertson, D.V. Schaffer (2004) The sonic hedgehog signaling system as a bistable genetic switch. *Biophysical Journal,* **86:** 2748-2757.
4. L. Lum, P.A. Beachy (2004) The Hedgehog response network: sensors, switches, and routers. *Science,* **304**: 1755-1759.

CONSENSUS PREDICTION OF AMYLOIDOGENIC DETERMINANTS IN AMYLOID FIBRIL-FORMING PROTEINS

STAVROS J. HAMODRAKAS, VASSILIKI A. ICONOMIDOU

We combine the results of three prediction algorithms on a test set of twenty one amyloidogenic proteins to predict amyloidogenic determinants. Two prediction algorithms are recently developed prediction algorithms of amyloidogenic stretches in protein sequences, whereas the third is a secondary structure prediction algorithm capable of identifying 'conformational switches' (regions that have both the propensity for α-helix and β-sheet). Surprisingly, the results of prediction agree well and also agree with experimentally investigated amyloidogenic regions. Furthermore, they suggest several previously not identified amino acid stretches as potential amyloidogenic determinants. Most predicted

Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 157 01, Greece, shamodr@biol.uoa.gr, veconom@biol.uoa.gr

(and experimentally observed) amyloidogenic determinants reside on the protein surface of relevant solved crystal structures. It appears that a consensus prediction algorithm is more objective than individual prediction methods alone.

COMPUTATIONAL/EXPERIMENTAL APPROACHES FOR MICRORNA BIOGENESIS AND FUNCTION

A. HATZIGEORGIOU[1]

MicroRNAs (miRNAs) are ~22 nucleotide non-coding RNA, which regulate expression of protein-coding genes through translation repression and/or degradation of mRNA. They are known to regulate cell proliferation and death and it has been found that miRNA expression signatures can distinguish cancer subtypes or predict biological and clinical behavior within the same cancer type. The understanding of miRNA function is likely to lead to novel therapeutically treatment of cancer.

Combined computational/experimental approaches have played a significant role during recent years in the identification of novel microRNAs (miRNAs), as well as in the analysis of their function. We have developed several tools for analyzing the genomic organization and function of miRNAs (DIANA-microT, TarBase, and miRGen) and a microRNA gene finder. Recently we have supported research on edited miRNAs in brain and investigated the role of SNPs within miRNA targets.

DNA – „PROGRAMMING LANGUAGE OF LIFE“

RALF HOFESTAEDT[2]

During the last decades molecular genetics could identify and sequence different gene functional units (*DNA_Units*). Most of these units are analysed in syntax (sequence and genome) and semantic (metabolic function). This information is permanently growing and represented by different database systems (EMBL – sequences, PDB – structure and semantics etc.), which are distributed over the world (internet). Based on this knowledge it is possible to discuss the old and still open question if DNA can be interpreted as a programming language or must be interpreted as hardware. In this paper/talk we will show that the DNA can be interpreted as a programming language in the sense of computer science. Moreover, it is possible to describe the so called “Programming Language of Life” using classical methods of compiler systems.

[1] University of Pennsylvania, Center for Bioinformatics, Philadelphia, USA
[2] Bielefeld University, hofestae@techfak.uni-bielefeld.de

At the beginning we have to discuss if most of the relevant *DNA_units* are known today. Regarding to the sequenced and analyzed genomes including the actual molecular knowledge we can say:

Most of the *DNA_Units* are known.

For most of these units the function is known and seems to be universal.

To show that the *DNA_Units* can be interpreted as a programming language we have to:

a) Specify the functional units.

b) Show that the *DNA_Units* represent fundamental mechanisms of a programming language.

In this talk only b) will be discussed, because a) can be done using specific Chomsky-type-2 grammars [Brendel and Budde 84, Hofestädt 07]. The fundamental features of a programming language are:

F1. Data type (at least one is sufficient)
F2. Instruction (standard instructions or by definition)
F3. Control instructions
F4. Punctuation mark

In the case of control instructions we have to discuss different kinds:

***C1*: Composition S1; S2;...; Sn**

The semicolon denotes the following operator. The meaning of this operator is that the next instruction (S_{i+1}) will be executed if the pre-instruction (S_i) was already executed.

***C2*: If-Instruction (If B then S)**

Let S be a instruction and B a condition, which can be true/false. The meaning is that instruction S will be executed in the case that condidtion B is true.

***C3*: While-instructions (While B do S)**

Let S be a instruction and B a condition, which can represent the value true/false. The meaning is that the instruction S will be executed as long as B is true.

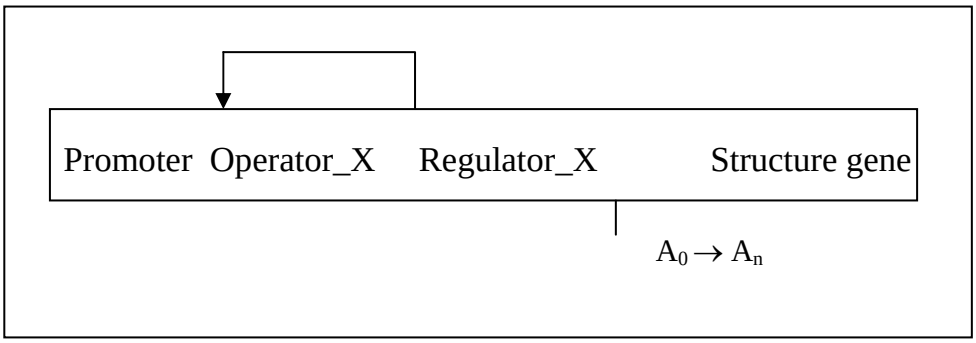
Interpretation

Let the DNA be the genetic program of a cell. In that case the cytoplasm can be interpreted as the data type representing metabolites. Therefore, we can say that the data type (metabolite) is available (F1). Instructions are chemical reactions caused by enzymes. Therefore, we can interpret enzymes as instructions which are presented by structured genes (F2). Structure genes are controlling the metabolism indirectly. Regarding specific cells we can see specific genes, which are active during specific time periods. This behaviour shows directly that DNA_Units can be interpreted as control instructions (F3). Finally the

DNA_Unit which is called spacer can be interpreted as the punctuation mark of this system (F4).

Defining and regarding specific operons it is able to show that C1 – C3 can be simulated by gene controlled regulatory networks. For one simple example showing that an operon can be interpreted as a If-instruction we focus to the operon L14 of E. coli, which regulates its own synthesis. The mechanism of this regulation process can be illustrated:

Example: Abstract representation of the Operon L14.



The boolean value of condition B will be specified by the state of the operator:

TRUE ::= if the Operator_X gene is free and
FALSE ::= if the Operator_X gene is blocked by the repressor.

Under this interpretation the instruction **If B then S** is simulated, because the operon will block itself after activation (protein synthesis). The synthesis of Regulator_X will be block the synthesis of Operon L14.

Regarding the illustrated Operon L14 example the theoretical extension to the **While instruction** is simple**.** Deleting the Regulator_X gene, which is inside of the OperonL14, will produce this effect**.** Therefore, we will discuss the structure and function of the Tryptophan-Operon, which will exactly show this effect.

However, based on these ideas this talk/paper will show that the *DNA_Units* represent the features of a programming language and that the *DNA_Units* can be interpreted as a programming language.

1. V. Brendel and H. Busse , "Genome structure described by formal languages," *Nucleic Acids Res.*, **12**, 2561 – 2568 (1984).
2. R. Hofestädt, "Extended Backus-Systems for the representation and specification of the genome," *Journal of Bioinformatics and Computational Biology*, in press.

RNA – PROTEIN INTERACTIONS AND THE SECONDARY STRUCTURE OF RNA

O.V. ILYICHOVA[1], P.K. VLASOV[2], M.A.ROYTBERG[3]

BACKGROUND: RNA-protein interactions play a key role in many fundamental biological processes due to their effects on RNA processing, turnover, transport, localization and translation. However despite of their functional importance, the specific mechanisms of protein – RNA interactions are still poorly understood. There have been written numerous works on the subject and they demonstrate that specificity of RNA-protein interactions is driven by a variety of hydrophobic, electrostatic, and hydrogen bond contacts between RNA and protein residues

INNOVATION: We have developed the database of RNA-protein interaction with respect to the RNA secondary structure. The database was created by analysis of PDB documents containing both protein and RNA chains. An element of the database corresponds to a contact between the RNA and protein atoms in a given PDB complex (a contact is a pair of atoms with a distance 4,5 À or less). Analogously to NPIDB database [1] for each contact we store its atoms, types of corresponding monomers, positions of the monomers within their chains, the total number of contacts for each monomer and the distance between the atoms. Besides this we store the information on the relation of RNA monomer (nucleotide) to the RNA secondary structure. Namely, we store the mark showing is the nucleotide, having a contact with the protein, paired or not; if it is paired – what is the position of the paired nucleotides, the length of the stem, etc; if not – what is the type of the corresponding loop or pseudoknot, etc.

We hope that the Data Base will give one better understanding of the role of RNA secondary structure in RNA-protein interaction (see e.g. [2, 3])

POTENTIAL APPLICATIONS: Identify and characterize preferable protein-binding substructures of secondary structure of RNA

Identification of inhibitors of RNA-protein interactions

[1] Moscow Institute of Physics and Technologies, 141700, 9, Institutskii per., Dolgoprudny, Moscow Region, Russia; ilyicheva@gmail.com
[2] Engelkardt Institute of Molecular Biology Russian Academy of Science, Vavilov str., 32 Moscow 119991, Russia, vlasov@imb.ac.ru
[3] Institute of Mathematical Problems of Biology, 4, Institutskaja str., 142290, Pushchino, Moscow Region, Russia, mroytberg@impb.psn.ru

The work was supported by grants RFBR 06-04-49249, INTAS 05-1000008-8028 and by the program of Russian Academy of Sciences (“ Molecular and cell biology”). The authors thank A.Alexeevsky, A. Karyagina and S.Spirin for advices and helpful discussions.

1. A. Alexeevski, S. Vasil’ev, A. Karyagina, R. Ledneva, S. Spirin, M. Titov (2006) Nucleic Acid-Protein Interaction DataBase (NPIDB) project. http://monkey.belozersky.msu.ru/NPIDB/
2. M. Terribilini, J. D. Sander (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins, *Nucleic Acid Research*.
3. N.Morozova, J.Allers, J.Myers, Y. Shamoo (2006) Protein-RNA Interactions: Exploring Binding Patterns with a Three-dimensional Superposition Analysis of High Resolution Structures, *Bioinformatics*, 22: 22, 2746–2752.

CHANGES IN ARGININE-RELATED TRANSCRIPTOME UNDER ACUTE MYOCARDIAL INFARCTION IN MOUSE: COMPUTATIONAL ANALYSIS OF MICROARRAY DATA

PAVEL S. IVANOV, ANASTASIA N. SVESHNIKOVA

The molecular processes in cardiomyocytes underlying ischemia and acute myocardial infarction (AMI) involve a complicated array of interconnected events at metabolic and genomic levels. High-throughput microarray technology provides an invaluable tool to study changes in expression of tens of thousands of genes under normal or pathological conditions. Recently, a genome-wide analysis of early transcriptional responses to AMI in mouse has been reported [1] with expression values measured at six time points for each gene. As a result of the preliminary statistical analysis of these microarray data, a dramatic increase in transcripts for arginase 1 (ARG1), the enzyme of polyamine biosynthesis and protein inhibitor of nitric oxide synthase (NOS), has been revealed. Since NO and polyamines have multiple and sometimes opposite effects on heart function [2,3] and the same is true for the role of polyamines in cell growth and apoptosis [4], we undertook a thorough statistical analysis of the arginine and polyamine related transcriptome from this microarray dataset (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4648).

First, we used the Significance Analysis of Microarrays (SAM) filtration [5] to select genes with most pronounced differential expression between mice with AMI and mice undergone sham surgeries as well as between different regions of heart tissue (infarcted tissue, surviving left ventricle free wall and interventricular septum). Second, we partitioned the genes with the most statistically significant changes in expression into groups by hierarchical agglomerative clustering with average linkage. We determined the threshold, Δ, of the SAM algorithm and the number of clusters in final partition, K, by a new Systematic Resampling (SyR) method proposed by the authors (unpublished results). Specifically, in SyR, the number of clusters, K, in statistically most reliable partition is estimated as $K = \mathrm{argmax}_k\ (p_{k\ \mathrm{adj}})$ where $p_{k\ \mathrm{adj}}$ is the bias-adjusted p-value for the measure of compactness of a k-cluster partition. We measured the compactness of partitions by Calinski-Harabash index, ch [6], and estimated p-values for this index by the second-order bootstrap. The threshold in the SAM filter was chosen as corresponding to $\max(p_{k\ \mathrm{adj}})$ in subsequent partitions when varying the number of clusters, k, from 2 to 29. This

Department of Biophysics, Faculty of Physics, M.V.Lomonosov Moscow State University, Vorobievy Gory, Moscow 119992, Russia, p-ivanov@mtu-net.ru

resulted in a filtered dataset with 397 genes which were subsequently partitioned into 11 clusters ($p_{11\ adj}$ = 0.96). These clusters correspond to different groups of co-expressed genes in myocytes under AMI.

Generally, arginine can be catabolized in multiple pathways why its origin in cell can be only of two kinds: the *de novo* synthesis from citrulline or the membrane transport by specific transporters [5]. We found that under AMI in infracted left ventricle, (i) the cellular uptake of arginine through y^+LAT1a transporter increases almost 2-fold, (ii) among four major pathways of arginine catabolism, only ornithine synthesis, that is the first step in the urea cycle, is AMI-induced, (iii) mitochondrial part of the urea cycle is presented at background expression levels, (iv) from three types of NOS only endothelial synthase is upregulated less than 2-fold, and (v) in polyamine pathway, ornithine decarboxylase increases about 3-fold and genes related to the spermidine synthesis are reliably upregulated. We used these results to correlate AMI-induced genes related to arginine-polyamines pathway with other parts of mouse transcriptome and as a basis of a quantitative model of their transcription regulation in cardiomyocytes under AMI.

1. M.H.Harpster et al. (2006) *Mamm. Genome,* **17:** 701-715.
2. C.E.Sears et al. (2004) *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **359:** 1021-1044.
3. Wallace et al. (2003) *Biochem. J.*, **376:** 1-14.
4. Y.-J.Zhao et al. (2007) *Eur. J. Pharmacol.*, doi: 10. 1016 /j. ejphar. 2007. 01. 096.
5. V.G.Tusher et al. (2001) *Proc. Natl. Acad. USA*, **98:** 5116-5121.
6. R.Calinski, J. Harabasz (1974) *Commun. Statistics*, **3:** 1-27.

NUCLEOTIDE CONTENT AND HYDROPATHY OF EXON, INTRON 5'- AND 3'-SITES IN THE LOWER FUNGI GENES

A.T.IVASHCHENKO, M.K.TAUSAROVA, V.A.KHAILENKO, S.A.ATAMBAEVA

Completely sequencing genomes of the lower fungi (Aspergillus fumigatus, Candida glabrata, Cryptococcus neoformans, Debaryomyces hansenii, Encephalitozoon cuniculi, Eremothecium gossypii, Kluyveromyces lactis, Magnaporthe grisea, Neurospora crassa, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Ustilago maydis, Yarrowia lipolytica) have essentially dif-

Kazakh National University named after al-Farabi, al-Farabi av., 71, Almaty, 050038, Kazakhstan, e-mail: a_ivashchenko@mail.ru

ferent content of genes with introns. For example, genome E. cuniculi has 0.7% of genes with introns, and genome C. neoformans contain 97.0% genes with introns. Properties of exons and introns have differences that depend on their number in genes. Average exon length decreases in 3-5 times at increase of intron number in genes of the lower fungi. The share of exons with lengths more than 400 nt decreases at increase of the number of exons with length 60-100 nt. In this case the average length of introns decreases, or does not change.

The genomes of C. glabrata, D. hansenii, E. gossypii, K. lactis, S. cerevisiae contain more one-intronic genes. The share of these genes was 1.5–5.0% from total genes number of genomes. Start nucleotides of the 5' sites (5'-S) of the introns of those genes are GUAUGU. Genomes of A. fumigatus, C. neoformans M. grise, N. crassa, S. pombe and U. maydis contain 37.8–97.0% genes with introns. Frequency of G, A, and U-nucleotides in the 5'-S of the introns in one-intronic genes of these organisms has been GUAAGU. Start nucleotides of the Y. lipolytica 5'-S introns are GUGAGU. The share of genes with introns in genome Y. lipolytica was 10.6% and 5'-S of the introns have other variant of nucleotide consensus.

The 3'-site (3'-S) of introns of studed organismes also distinguishes by nucleotides content. In all introns the 3'-S have consensus YAG, with more C-nucleotide in genomes of U. maydis, E. gossypii, Y. lipolytica, or T-nucleotide in genomes of D. hansenii, S. pombe. Consensus RCAG is present in genes of U. maydis, E. gossypii, A. fumigatus, C. neoformans.

Frequency of distribution of all four nucleotides are equiprobable at positions +2, +3 of the 5'-S of intermediate exons. Tendency of reduction of pyrimidines and increase of purines has been observed in all genes of all genomes at the positions -3 to -1 in the 3'-S of exons.

Introns have GU and AG at the flanking positions in the 5'- and 3'-S accordingly in almost all fungi genomes. Approximately 1.6% of C. neoformans one-intronic genes have GC dinucleotide at the positions +1, +2 in the 5'-S introns. About 0.2 and 2.8% genes with GC–AG-introns are present in one-intronic genes A. fumigatus and S. cerevisiae consequantly. Consensus of 5'-S GC–AG-introns of S. cerevisiae (GCAUGU), A. fumigatus (GCAAGU) and C. neoformans (GCAAGU) was equal with that for GU–AG-introns. All GC–AG-introns 3'-S have consensus YAG.

Hydrophobic-hydrophilic properties are critical for the interactions between snRNA, spliceosome proteins and pre-mRNA. The average hydropathy of nucleotides was calculated at the positions in the 5'- and 3'-splicing sites (5'-SS and 3'-SS accordingly) in exon/intron border. Nucleotide hydropathy is highly conserved in the interval -2 to +6 of the 5'-SS and in the interval -3 to +2 of the

3'-SS. The profile of hydropathy of the 5'-SS and 3'-SS remains close in genes of relative genomes. Hydrophobic-hydrophilic interactions provide recognition and interaction between snRNA, spliceosome proteins and pre-mRNA.

QUALITATIVE COMPARISON OF ORTHOLOGS DETECTION METHODS AND THEIR IMPLEMENTATION IN WEB-AVAILABLE DATABASES AND TOOLS BY THE EXAMPLE OF FABP FAMILY

A.E. IVLIEV[1], L.U. ANDREEVA[2], M.G. SERGEEVA[3]

Orthologs detection is one of the actively tackled problems in bioinformatics. It is important for experimental studies devoted to particular genes, elucidating the evolutionary origins of groups of organisms, systematizing the abundant sequence information in databanks and other objectives in biology. Although chronologically it was one of the first problems to be addressed, a broad range of new methods continues to appear now.

In the current work we compare different orthologs search methods by the example of applying them for characterization of eukaryotes-specific fatty-acid binding proteins (FABP) genes family. The family is actively studied now being involved in ω3/ω6 polyunsaturated fatty acids metabolism and associated with obesity, cancer and cardiovascular diseases.

The following methods were compared: Pfam [1], COG [2], HomoloGene [3], InParanoid [4], OrthoMCL [5], PhiGs [6], OrthologID [7], PHOG [8], Ensembl orthology detection pipeline [9]. Databases integrating distinct tools orthology predictions such as HCOP [10], YOGY [11] and PhyloPat [12] were also addressed. The resources were compared on the following parameters: search principles, ability to identify one-to-many or many-to-many relationships, ability to identify relationships between individual genes within the clusters, genomes sampling, redundancy caused by isoforms, database querying options and others. The opportunities of applying the methods for different research objectives are also discussed. The work was supported by RFBR (07-04-01160).

1. Finn RD et al (2006). *Nucleic Acids Res*, **34**(Database issue):D247-51.

[1] Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia, ivliev-alexa@yandex.ru
[2] Lyceum №1553, Moscow, Russia, luda_dg@mail.ru
[3] A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, 119992, Russia. sergeeva@genebee.msu.ru

2. Tatusov RL et al (2003). *BMC Bioinformatics*, **4**:41.
3. Wheeler DL et al (2007). *Nucleic Acids Res*, **35**(Database issue):D5-12.
4. O'Brien KP et al (2005). *Nucleic Acids Res*, 33(Database issue):D476-80.
5. Chen F et al (2006). *Nucleic Acids Res*, 34(Database issue):D363-8.
6. Dehal PS et al (2006). *BMC Bioinformatics*, **7**:201.
7. Chiu JC et al (2006). *Bioinformatics*, **22(6)**:699-707.
8. Merkeev IV et al (2006). *BMC Evol Biol*, **6**:52.
9. Hubbard TJ et al (2007). *Nucleic Acids Res*, **35**(Database issue):D610-7.
10. Eyre TA et al (2007). *Brief Bioinform*, **8**(1):2-5.
11. Penkett CJ et al (2006). *Nucleic Acids Res*, **34**(Web Server issue):W330-4.
12. Hulsen T et al (2006). *BMC Bioinformatics*, **7**:398.

GROUP BEST-BEST HITS METHOD: COMPROMISE BETWEE MANUAL AND AUTOMATIC ORTHOLOGS SEARCH. APPLICATION TO FAMILY-FOCUSED STUDIES

A.E. IVLIEV[1], M.G. SERGEEVA[2]

Searching genomes for orthologs of known proteins from a given species remains a widely occuring task for biologists studying evolution of their gene families of interest.

Web-available pre-computed results of genome-scale in-paralogs search tools, such as InParanoid [1], Phog [2], PhIGs [3] etc., are not always suitable for a particular research (for example because of genome sampling). Running the software locally either takes much time or requires significant computational powers. Thus, running InParanoid to compare 2 eukaryotic genomes results in hundreds of thousands Blast searches which takes days for an average PC. The majority of these searches can be avoided if smaller number of gene families is concerned in the research.

Here we propose using a "group Best-Best Hits" (gBBH) approach – based on the conventional BBH method – to identify in-paralogs and one-to-one orthologs of known genes. The exemplary evolutionary scenario scheme is depicted in Fig. 1a. The conventional BBH approach would miss several of the orthologs from SP2 in this typical case (Fig. 1b).

In gBBH approach the following algorithm is proposed (Fig. 1c). Blast search in SP2 is performed with A_1 as a query ("forward search"). For each

[1] Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia, ivliev-alexa@yandex.ru
[2] A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, 119992, Russia. sergeeva@genebee.msu.ru

SP2 gene from the forward hits' list starting from its top a blast search is performed in the reverse direction ("backward searches": *a_1* query – SP1 genome; *a_2* query – SP1 genome; etc). If the best hit of the current backward search belongs to the initial group of the SP1 paralogs, the query gene from SP2 is added to the pre-orthologs list. The backward searches are stopped as soon as a backward best hit outside of the initial group occurs for the first time while processing the forward hits' list. The described steps are repeated for all other SP1 genes as queries for forward searches to find relevant SP2 genes which haven't yet been added to the pre-orthologs list. In this case *a_1*, *a_2*, *a_3* and *b* will be found whereas out-paralogs of *A_1*, *A_2* and *B* will automatically be filtered out at the cost of approximately only 10 Blast searches as a whole. The final reconstruction of orthologous sets from the pre-orthologs list should further be done manually by building a phylogenetic tree.

The approach is thus a compromise between automatic and manual orthologs search and can be useful for family-focused evolutionary studies.

A Perl program based on the gBBH principle for searching orthologs in multiple genomes was written. The program was applied to study evolution of arachidonic acid cascade (75 genes from about 15 families in human). The Perl code is available at www.lipidomics.ru.

The work was supported by RFBR (07-04-01160).

1. Remm M et al (2001). *J Mol Biol*, **314(5)**:1041-52;
2. Merkeev IV et al (2006). *BMC Evol Biol*, **6**:52; [3] Dehal PS et al (2006). *BMC Bioinformatics*, **7**:201.

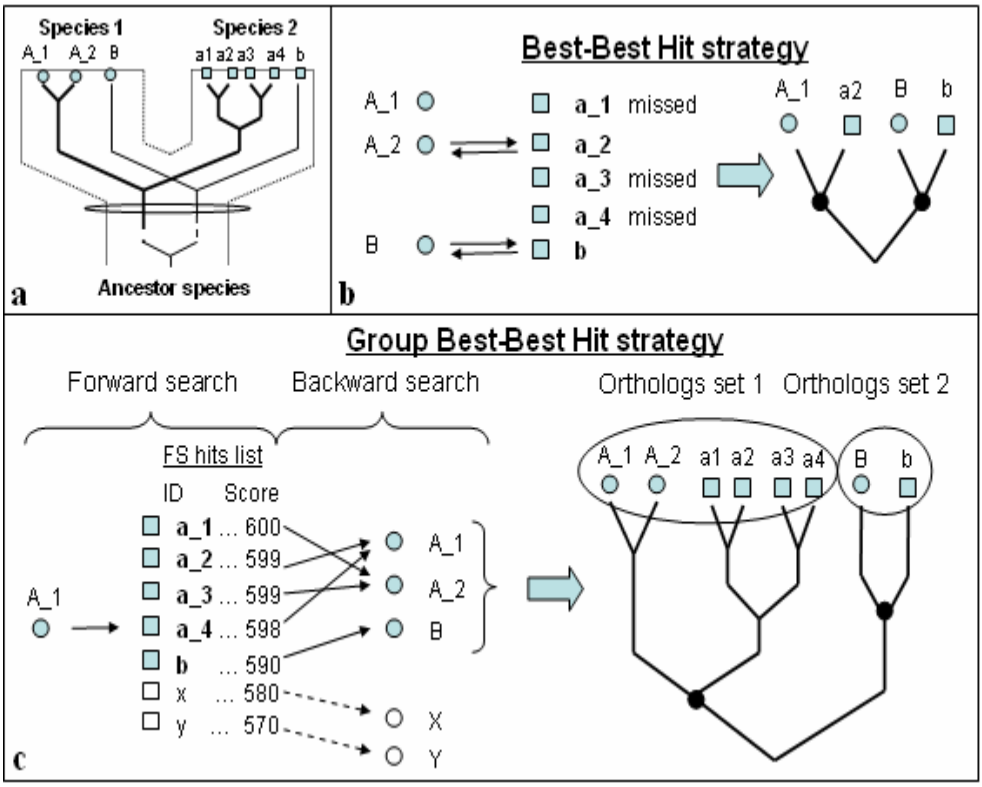


Fig. 1. (A) - evolutionary scenario scheme; (b) and (c) - BBH and gBBH approaches comparison.

VIRTUAL MACHINE FOR ANALYZING LIVING SYSTEMS

EKATERINA IZOTOVA, D.S. TARASOV

The possibility of using virtual machines for studying information processing capabilities of living cell arises from the notion that a living cell can be considered as computational device and its logical architecture can be studied independently from its physical organization. Virtual machine that resembles logical organization of living cell can be used for modeling regulation of biomolecular processes, can be target architecture for special biological modeling languages, enhance capabilities of genome annotation tools and provide further understanding of logical organization of living systems.

Kazan State University, 420075, Kremlevkaya 18, Kazan Russia, ied@mntech.ru

Virtual machines have been studied from 1960 primarily for the purpose of achieving compatibility between different hardware architectures. Since living cell can be considered as computational device [1] virtual machines becomes a relevant tool for molecular biology. In fact, systems based on kinetic models like E-CELL [2] are virtual machines designed to mimic cell behavior on digital computer. They provide virtualization by modeling physical (hardware) organization of living cell.

However, from the beginning of research in VM area, it was become clear that hardware architecture (machine from the point of view of engineer) and logical organization (machine from the point of view of programmer) are completely different [3]. Modeling exact physical organization of the system to simulate computational device behavior results in unnecessary complication and overload. The more is known about logical organization of cell, the less are the need for modeling exact physics and chemistry in order to understand its behavior.

Previously, we considered possible abstract machines for simulating logical organization of living cell and languages best suited for programming such machines [4, 5]. In this work we present implementation of living cell virtual machine (LCVM) and evaluate its functionality using a wide range of biochemical examples.

Our implementation has following features:

Rich programming language for describing biochemical models and making extensions for either for graphical environment or machine core architecture.

Support for modeling both well-mixed and crowded environments.

Parallel computations on SMP and cluster systems

Wide range of molecular biological processes that can be consistently and uniformly represented as LCVM programs, including enzymatic reactions and formation of non-covalent bonded complexes. Concept of metabolic pathway can be abstracted as LCVM operation and complete regulation subsystems can be represented as LCVM objects. Gene expression represented in LCVM programs as dynamic compilation of strings and differential splicing can be encoded as string processing system. This way, effects of many mutations can be represented in LCVM programs by modification of source strings in certain places. Possibility of mapping DNA mutations to source code strings suggests a possibility of creating DNA to LCVM translator.

1. S. Ji. (1999) The cell as the smallest DNA-based molecular computer, *Biosystems,* **52**:123-133.

2. M. Tomita, K. Hashimoto, K. Takahashi, et al. (1999) E-CELL: software environment for whole-cell simulation, *Bioinformatics* **15**:72-84.
3. G.M. Amdahl, G.A. Blaauw, P.P. Brooks (1964), Architecture of the IBM System-360, *IBM J. of Research and Development* **8**:87-101
4. D.S. Tarasov, N.I. Akberova, A.Yu. Leontiev (2002) Architecture of Cell Device, *Proceedings of BGRS'2002*, Novosibirsk, 216-218.

INFORMATION MEASURES FOR TRANSCRIPTION FACTOR BINDING SITES AND CONSERVED REGULATORY REGIONS

VIDHYA JAGANNATHAN, DOROTA RETELSKA,
EMMANUEL BEAUDOING, PHILIPP BUCHER

Information measures based on Shannon's entropy have been applied to nucleotide sequences as early as the first DNA and RNA molecules were deciphered (1). In my talk, I'm going to present two partly novel applications of information theory to living systems. The first one relates to the sequence specificity of a transcription factor which can be represented by an energy matrix according to Berg and von Hipplel (2). The elements ε_{ib} of such a matrix represent component interaction energies between a basepair b and a binding site position i. In practice, transcription factor binding sites are represented by base probability matrices p_{ib} derived from gap-free alignments of *in vitro* or *in vivo* selected binding sites. According to Schneider and coworkers (3), the information content of a base probability matrix is given by:

$$I = \sum_i \sum_b p_{ib} \log_2 \left(p_{ib} / q_b \right)$$

Here q_b is the background frequency of base b. However, this measure is not an intrinsic property of the DNA binding protein because the values p_{ib} depend on the stringeny of the binding conditions. Therefore, we propose as an alternative information measure:

$$I = \sum_i \sum_b e_{ib} \log_2 \frac{e_{ib}}{0.25}, \quad \text{where} \quad e_{ib} = \frac{\exp(-\varepsilon_{ib})}{\sum_b \exp(-\varepsilon_{ib})}$$

Here we assume that ε_{ib} are scaled in dimension-less RT units. According to Berg and von Hippel (2), $\varepsilon_{ib} = -(1/\lambda)\log(p_{ib} / q_b)$ with λ being an *a priori* unknown stringency parameter. We show that λ can be estimated by fitting

Swiss Institute for Experimental Cancer Research and Swiss Institute of Bioinformatics, Ch. des Boveresses 155, CH-1066 Eplainges s/Lausanne, Switzerland, Philipp.Bucher@isb-sib.ch

experimentally determined binding constants (K_b) for a few sequences to corresponding binding energies computed from the energy matrix. Using this approach we compute information measures for binding sites of four transcription factors characterized by high-throughput SELEX libraries. Base probability matrices were derived from more than 10'000 binding sites on average.

In the second application, we try to answer the question how much genetic information is conserved between two related genomes. To this end, we compute percent identity values for each genome position from the alignment in a surrounding window of 60 bp. The evolutionary rate $r^{obs}(i)$ at position i is then computed by a modified Jukes-Cantor model for gapped alignments. Taking into account the maximal, unconstrained mutational rate r^{max}, we define the amount of conserved sequence information at position i as $1-r^{obs}(\mathrm{i})/r^{max}$. By summing up these values over all genome positions, an estimate of the total amount of conserved genetic information is obtained. We applied this procedure to whole genome-alignments downloaded from the UCSC genome browser (4). As a major result we found that more than 50% of the conserved sequence information is non-coding in vertebrates. However, if the human genome is compared to increasingly distant other vertebrate genomes, one observes a faster decay of non-coding information as compared to coding information. Various evolutionary scenarios will be discussed to explain this observation.

1. L.L. Gatlin (1972) Information theory and the living system. Columbia University Press, New York
2. O.G. Berg, P.H. von Hippel (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters, *Mol. Biol.,* **193:**723–750.
3. T.D. Schneider et al. (1986) Information content of binding sites on nucleotide sequences, *J. Mol. Biol.,* **188:**415-431.
4. W.J. Kent . et al. (1992) The human genome browser at UCSC, *Genome. Res.,* **12:**51-54.

IDENTIFICATION OF FUNCTIONALLY IMPORTANT SITES IN POORLY CHARACTERIZED PROTEIN FAMILIES

OLGA V. KALININA[12], ROBERT B. RUSSELL[2], M.S. GELFAND[13]

Structural genomics projects generate many 3D structures of proteins of unknown function (Manjasetty et al., 2007). The comparative analysis of the sequences of related proteins may help to identify key functional positions and specificity determinants.

Recently, we developed SDPsite, a tool for prediction of protein functionally important sites (Kalinina et al., 2007). It is based on identification of SDPs (Specificity-Determining Positions, positions that are conserved within groups of proteins having the same specificity and differ between these groups) and conserved positions in an alignment of a protein family, and subsequent spatial clustering in a 3D structure of one of proteins from the family.

Testing SDPsite on a well-studied LacI family of bacterial transcription factors proves that the identified clusters are located in real functional sites of the protein. A benchmark using a set of protein families from the Conserved Domain Database with mapped functional residues demonstrates a good agreement of the prediction with the available experimental data.

We applied our method to a large dataset of uncharacterized protein structures, providing new, possibly valuable biological information.

O.V.K. was supported by INTAS Fellowship Grant for Young Scientists (04-83-3704).

1. B.A. Manjasetty, W. Shi, C. Zhan, A. Fiser, M.R. Chance. (2007) A high-throughput approach to protein structure analysis. *Genet Eng (N Y),* **28**:105-128
2. O.V. Kalinina, R.B. Russell, A.B. Rakhmaninova, M.S. Gelfand. (2007) Computational method for prediction of protein functional sites using specificity determinants. *Mol Biol (Mosk)* **41(1)**: 137-147

[1] Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia, 119992
[2] EMBL, Meyerhofstr., 1, Heidelberg, Germany, 69117
[3] Institute for Information Transmission Problems RAS, Bolshoi Karetny per. 19, Moscow 127994, Russia, ok81@yandex.ru

DISSECTING EVOLUTION OF IMMUNE SYSTEM: RAG1, TRANSIB AND CHAPAEV

VLADIMIR V. KAPITONOV

In vertebrates, numerous pathogens, including viruses and bacteria, can be neutralized due to enormous variability of surface receptors expressed by B and T immune cells. V(D)J recombination is a process responsible for the receptors variability. The RAG1 protein is a key element of the V(D)J recombination machinery that catalyzes DNA rearrangements involved in generation of the receptors variability. We have found recently that the ~600-aa RAG1-core and recombination signal sequences have been evolved from Transib DNA transposons more than 500 million years ago (1). Here I will show that the ~300-aa N-terminal portion of RAG1 was most likely derived from another DNA transposon belonging to a novel superfamily of "cut and paste" DNA transposons called Chapaev. Members of this superfamily are universally characterized by 4-bp target site duplications and the CAC terminal inverted repeats. Transposition of Chapaevs is catalyzed by the 500-1000 aa long Chapaev transposase encoded by the autonomous transposons. Excluding its ~300-aa N-terminus, which is similar to the RAG1 N-terminal portion, the transposase is not similar to other known proteins. Chapaev transposons populate genomes of diverse metazoans, including starlet sea anemones, hydras, mollusks, insects, worms, sea urchins, and lancelets. Based on results described here, it appears that the RAG1 protein has been evolved as an assembly of two different transposases, Chapaev and Transib. The reported results are important for understanding evolution of the immune system and computational identification of enzymes involved in DNA transpositions.

This study was supported by the NIH grant (5 P41 LM006252-08).

1. V.V.Kapitonov, J.Jurka (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons, *PLOS Biol.,* **3:** e181.

Genetic Information Research Institute, Mountain View, CA 95054, USA,
vladimir@girinst.org

A MODEL OF THE "MOLECULAR VECTOR MACHINE" FOR PROTEIN FOLDING

VLADIMIR. A. KARASEV[1], VICTOR V. LUCHININ[1], VASILY E. STEFANOV[2]

Earlier [1] we proposed a dodecahedral model of the canonical set of 20 amino acids based on antisymmetry of the side chains. Later [2] the model was refined. An additional anti-symmetry plane (III) was introduced (Fig.1, a). Side chains were arranged from the top to the bottom in the increasing order of their size. Shorter side chains are situated on the right whereas their heavier analogs on the left from the plane I. In this model four groups of chains can be distinguished: 1) chains symmetrical about plane I (e.g. Ser:Thr, etc.), 2) chains symmetrical about plane II (e.g. Ser:Cys, etc.), 3) chains symmetrical about plane III (e.g. Ser:His) and 4) chains symmetrical about the center of the dodecahedron (e.g. Ser:Trp). The aim of the present work was to search ways of practical application of the above model.

According to [3], the region of $N_iH...O_{i-4}=C_{i-4}$ hydrogen bond in the 4-link protein cycle is the target of physical operators (amino acid side chains) which reconstruct the triplet-encoded structure. Following the principles assumed in the model we introduced three planes in the above region and considered 20 vectors representing the action exerted on this bond [2]. The model of the amino acid structure was brought into this region (Fig.1, b), the NH-group being placed into the vertex Gly and atom O_{i-4} into the center of the dodecahedron, so that the above vectors were directed to the dodecahedron vertices corresponding to the names of the side chains.

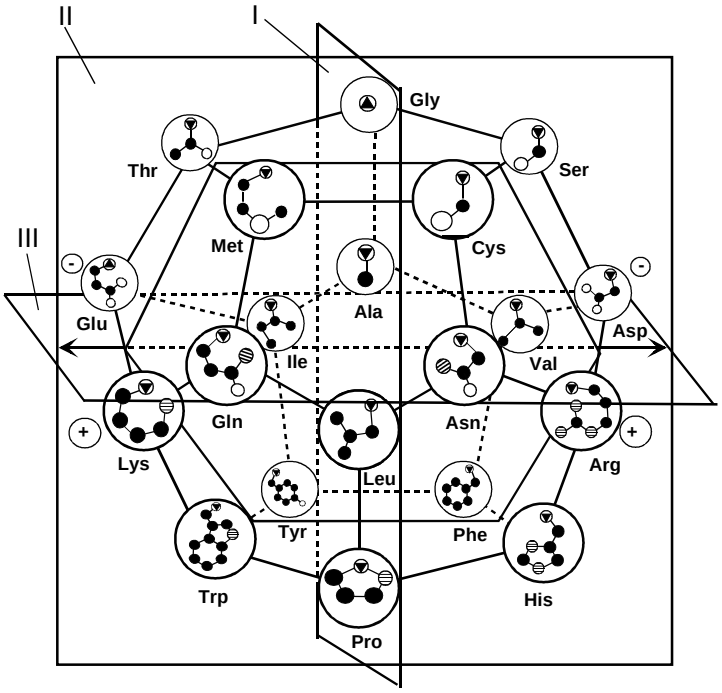
Thus, the amino acid side chains can be regarded as irreducible representations of the group composed by the vectors (dodecahedron diameters) and the structure itself as "molecular vector machine". The model accounts for certain features of the canonical set of amino acids, e.g. difference in the length of structurally similar Asp and Glu or oppositely charged Asp and Arg can be related to the effect of different orientation of vectors. Cyclic character of the side chains His, Trp, Phe, Tyr is likely to be associated with the rigid structure, which is required for the function of the vectors in the lower face of the dodecahedron could realize, etc.

The molecular vector machine (Fig.1, b) consists of two parts: dodecahedron and tetrahedron (block of C^{α}_i atom). Side chains connected with C^{α}_i (R_i)

[1] St.-Petersburg State Electrotechnical University, Prof. Popov str. 5, St.-Petersburg, 197376 Russia, e-mail:cmid@eltech.ru
[2] Department of Biochemistry, St.Petersburg State University, Universitetskaya nab. 7/9, St.-Petersburg, 199034 Russia e-mail: vastef@mail.ru

are directed towards the corresponding vertices of the dodecahedron and generate the encoded structure [3], whereas concomitant turn of the tetrahedron specifies the direction of the bond with the atom C^{α}_{i+1} of the chain [2]. We presume that operation of the molecular vector machine connected with the chain's growth, can provide co-translational protein folding. At present the model is being tested on different α-helical and β-sheeted fragments available from PDB.

(a)



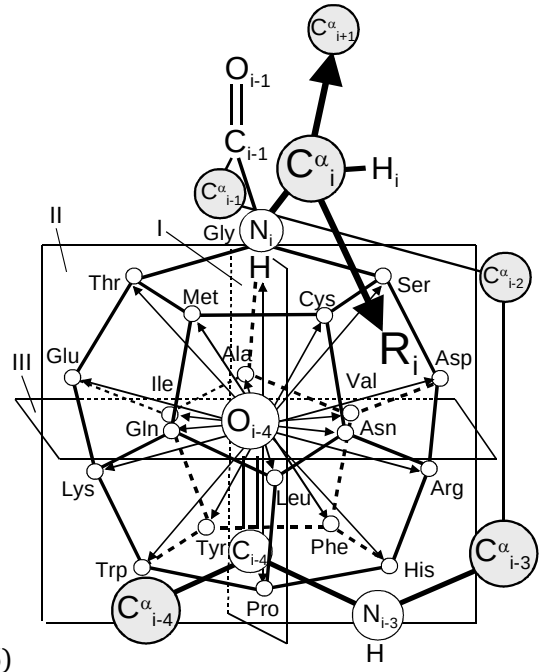


Fig. 1. Construction of the model of "molecular vector machine" for co-translational protein folding.

a – dodecahedral model of the canonical set of amino acids

I , II, III – symmetry planes. Side chains are oriented downwards; α-carbon atom is

symbolized by triangle in the circle

б – transfer of the structure of the amino acids' canonical set into the region of the of $N_iH...O_{i-4}=C_{i-4}$ hydrogen bond in the 4-link protein cycle.

1. V.A. Karasev et al. (2005). A dodecahedron-based model of spatial representation of the canonical set of amino acids. In: Mathematical Biology and Medicine, **Vol.8**. International Conference. "Advances in Bioinformatics and Its Applications", World Scientific Publ. Co., 482-493.
2. V.A. Karasev (2007). Amino acids of the canonical set as irreducible representation of the group of vectors – dodecahedron diameters. Dep. VINITI 25.04.07, № 461-B2007
3. V.A. Karasev, V.E. Stefanov (2001). Topological nature of the genetic code. J.Theor.Biol. **209:**303-317.

DISTRIBUTION OF MICROCIN J-LIKE AND MICROCIN C-LIKE ANTIBIOTIC SYSTEMS

ALEXEY KAZAKOV[1], M. S. GELFAND[1], KONSTANTIN SEVERINOV[2]

Microcins are a class of small ribosomally-synthesized antibiotics secreted by enterobacteria. Genes responsible for the synthesis, maturation and secretion of post-translationally modified microcins are often organized in operons. The small size of microcin precursors makes conventional similarity searches for homologous genes useless. However, the fact that the maturation and immunity genes tend to form clusters allowed us to distinguish candidate microcin loci genes from homologs with similar biochemical functions but unrelated cellular roles.

Taking into account the arrangement of genes encoding microcin precursor and microcin maturation enzymes, we were able to identify twenty microcin J-like antibiotic systems in fourteen bacterial species belonging to alpha-, beta- and gamma-Proteobacteria and Actinobacteria. In most of them, short open reading frames with sequence features characteristic for microcin J and for other threaded-lasso peptide precursors were found upstream of the first gene of the operon. The genomes of *Sphingopyxis alaskensis* RB2256, *Caulobacter* sp. K31, *Frankia* sp., and *Streptomyces avermitilis* contain several candidate *mccJ* loci and in several cases multiple microcin precursors are located upstream of candidate maturation genes. Thus, these bacteria may produce multiple microcin J-like lariat peptides.

Search for microcin C-like antibiotic systems using the same approach revealed fifteen such systems in twelwe species from diverse bacterial groups, such as beta-, gamma- and epsilon-Proteobacteria, Firmicutes and Cyanobacteria. The microcin C precursor is encoded by the 21-bp *mccA* gene, one of the shortest genes known. The candidate *mccA* genes are preceded by plausible Shine-Dalgarno sequences and in most cases, are separated from the *mccB* homolog by ~100 bp of noncoding DNA that, similarly to the case of the *E. coli* microcin C51 locus, contains putative transcription terminators. Heptapeptides structurally similar to the microcin C precursor were found in all identified gene clusters, except those from *Bartonella henselae* and two *Synechococcus* strains. In *Synechococcus*, two (strain CC9605) or three (strain RS9916) direct repeats each encoding 56 or 57 aminoacid polypeptides, respectively, were ob-

[1] Institute for Information Transmission Problems RAS, Moscow, Russia, kazakov@iitp.ru
[2] Institute of Molecular Genetics RAS, Moscow, Russia

served. The ORFs are sufficiently similar between the two *Synechococcus* strains and the C-terminal amino acids of all ORFs are asparagines, typical for microcin C-like precursors. So, the cyanobacterial microcin precursors may constitute a new subtype of C-related microcins, with potentially different properties. Our analysis demonstrates that peptides structurally similar to microcins J and C from *E. coli* may be produced by a variety of bacterial groups.

COMPUTATIONAL RECONSTRUCTION OF MICRORNA-MEDIATED GENE REGULATION FROM MICROARRAY DATA

RAYA KHANIN[1], VERONICA VINCIOTTI[2]

Over the past few years, it has emerged that gene expression in plants and animals is post-transcriptionally regulated by microRNAs (miRNAs). Mature miRNAs are small noncoding RNA molecules that regulate the target messenger RNAs (mRNAs) by influencing their stability, compartmentalization and translation. MiRNAs are estimated to comprise 1-5% of animal genes, making them one of the most abundant classes of regulators. Their widespread and important role in animals is highlighted by recent estimates that up to 30% of human protein-coding genes are subject to miRNA-mediated control [2] and is evidenced by their evolutionary conservation [4]. miRNAs play a central role in many biological processes, including developmental timing, cell proliferation, apoptosis, metabolism, cell differentiation and morphogenesis. Computational efforts to understand the post-transcriptional gene regulation by miRNAs have been focused on the target prediction tools (reviewed in [4]), while kinetic models of gene regulation by miRNAs are scarce.

In this work, we introduce a model of post-transcriptional gene regulation by miRNA focusing on the miRNA-mediated effect on the target mRNAs degradation. The expression μ of an mRNA transcript at time t is determined by its transcription and its degradation $d\mu / dt = p(t, TF) - \delta\mu$. It has been shown experimentally that miRNAs directly affect the levels of their target transcripts, presumably by accelerating their degradation rates ([1], [3]), and thus lowering their levels of expression. Thus, when mRNA is a target of a miRNA, its degradation rate, δ, depends on the level of this miRNA, m, that is $\delta = \delta_0(1 + \delta_m(m))$, where δ_0 is the basal degradation rate. A plausible model for the miRNA-mediated mRNA degradation is $\delta_m(m) = dm/(k + m)$, where

[1] Department of Statistics, University of Glasgow, UK, raya@stats.gla.ac.uk
[2] Department of Information Systems and Computing, Brunel University, UK, veronica@ida-research.net

d is the maximal miRNA-mediated fold change in the target mRNA degradation rate compared to δ_0 and k is the half-saturation constant. At the low levels of m (compared to k), the model reduces to a linear one: $\delta_m(m) = dm$. Recent estimates and experimental findings indicate that an average miRNA may regulate hundreds of genes ([2],[3]). The model therefore holds for each target mRNA i with gene specific kinetic parameters p_i, δ_{0i}, d_i, k_i.

We applied the above model to the microarray time-course (7 time-points and controls) measurements of human mRNAs in a miRNA124a transfection experiment [5]. The model is fitted to downregulated targets of miRNA-124a predicted by the pictar algorithm [2]. Using an inferential framework based on maximizing the likelihood, we quantify the miRNA-mediated downregulation of the target mRNAs. We also obtain the maximum likelihood estimate for miRNA124a half-life (prior to its incorporation into the RISC complex) to be $29h$ (with 95% confidence interval $(26h,49h)$). MiRNA half-life times are difficult to measure and the experimental estimates have not yet been obtained. Our model applied to microarray data makes a prediction for the miRNA124a half-life that can be experimentally verified. To the best of our knowledge, this is the first attempt to quantify miRNA-mediated effects on the target mRNAs.

Raya Khanin thanks Nikolaus Rajewsky for his encouragement of this work and for many helpful discussions on miRNAs and modeling in Systems Biology.

1. S.Bagga et al (2005). Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*. 122(4), 553-63.
2. A. Krek et al (2005) Combinatorial microRNA target predictions. *Nature Genetics*, 37(5), 495-500.
3. L.P. Lim et al (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027), 769-73.
4. N. Rajewsky (2006). MicroRNA target predictions in animals. *Nature*. 38,S8-13.
5. X. Wang and X. Wang (2006). Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Research*. 34(5): 1646-52.

CHANGES OF EXON AND INTRON LENGTHS IN HUMAN GENES

V.A. KHAILENKO, S.A. ATAMBAEVA, A.T. IVASHCHENKO

The length of introns in human genes varies from several tens up to tens thousand nucleotides and intron length is 20 times more than exon length on average. The changes of exon and intron lengths in genes of all human chromosomes depending on genes density of DNA and from number of introns in genes are studied. All chromosomes have been divided into three parts (1-6, 7-12, 13-Y) and average characteristics of genes were found in them. In these parts of chromosomes the quantity of genes compounded 9513, 6680 and 8319 genes accordingly. Each chromosome was divided into regions with length 1 Mbp and number of genes was found in them. Samples contained the regions with density 1-11, 12-20, 21 and more genes/Mbp. The genes of that samples have been arranged in groups with 1-2, 3-5, 6-9, 10-14, 15 and more introns in a gene. Length of exons (l_{ex}), introns (l_{in}), the sum of exon lengths (L_{ex}) in a gene, length of a gene (L_{gn}), number introns in a gene (N_{in}) and number of genes in a group (N_{gn}) were determined. Quantity introns and exons with a length in intervals 1-20, 21-40, 41-60 n and so on up to 400 n, and also with a length more than 400 n were analyzed.

In all chromosomes appeared a great heterogeneity of a genes density. The number of genes on 1 Mbp in chromosome 13 was equal to zero in 29 regions and in 7 regions it was in an interval of 12-20 genes/Mbp. There were not regions with density of genes 21 and above in chromosomes 13, 18 and Y. In chromosome 19, having the greatest density of genes in a genome of the human, regions without genes were absent, and in 41 regions genes density was above 21 genes/Mbp. The share of DNA, occupied protein-coding genes, was same in all three parts of regions. In regions with density 1-11 genes/Mbp share of protein-coding genes was 47, 47 and 40% in 1-6, 7-12 and 13-Y chromosomes accordingly. In regions with density 12-20 genes/Mbp the share of genes was equaled 53, 61 and 48%. In regions with 21 and more genes/Mbp genes occupied accordingly 46, 28 and 50%.

Obtained data show decrease of exon lengths in genes of all parts in process of augmentation in them numbers of introns. Decrease of exon lengths descended due to lowering a share of exons with length more than 400 n and maximum of exon lengths exhibited in an interval 100-140 n. In genes with 1-2 introns the share of exons with a length more than 400 n was several times

Kazakh National University named after al-Farabi, al-Farabi av., 71, Almaty, 050038, Kazakhstan, e-mail: a_ivashchenko@mail.ru

more, than in genes with 15 and more introns. Change of exon lengths in genes descended only by increase of intron number in a gene and did not depend on genes density in region of DNA.

The average intron lengths in group of genes in regions with density 1-11, 12-20, 21 and more genes/Mbp decreased in process of increasing of intron number in genes approximately in 1.5 times. The average intron lengths in genes of regions with density 1-11, 12-20, 21 and more genes/Mbp in chromosomes 1-6 were accordingly 11798, 4779 and 2191 n. For genes of chromosomes 7-12 in regions with density 1-11, 12-20, 21 and more genes/Mbp appropriate average values of intron lengths were 10565, 4165 and 1992 n. In genes of three samples in chromosomes 13-Y appropriate average intron lengths were 9361, 4294 and 2235 n. Thus the average intron lengths in genes depended on region in which these genes have been localized. Despite of lowering of average exon lengths in genes at augmentation in them number of introns, the common exon lengths increased several times. We shall note, that the sum of exon lengths in genes of samples with 1-2 introns was equal to length of genes without introns. Average sums of exon lengths in genes with 1-2 introns were relative in three samples of regions. Lengths of the sums of exons in groups of genes with 15 and more introns slightly differed among themselves as well.

In groups of genes with 15 and more introns the length of genes for appropriate samples in chromosomes 13-Y was 206629, 86848 and 46219 n. In genes with 1-2 introns their lengths in these parts were 18457, 8931 and 5330 n on the average.

Total exon lengths increased with rise of introns number in genes. This link in genes of samples with different of genes density for all chromosomes is described by dependence with high coefficients of correlation: $N_{in} = aL_{ex} + b$, where a and b - coefficients of linear regression. The total gene length depends on intron number. Multiple changes of genes length are well correlated with introns number in them and were described by the equation: $N_{in} = cL_{gn} + d$, where c and d - coefficients of linear regression.

Obtained data justify that variety of exon, intron and gene lengths is connected also with the intron numbers in genes and with the genes density in regions in all human chromosomes.

HIERARCHICAL ANALYSIS OF THE EUKARYOTIC TRANSCRIPTION REGULATORY REGIONS BASED ON THE DNA CODES OF TRANSCRIPTION

IRINA V. KHOMICHEVA[1], E.E. VITYAEV[1], E.A. ANANKO[2], V.G. LEVITSKY[2], T.I. SHIPILOV[2]

Eukaryotic regulatory regions are characterized by complex modular hierarchical structure and as the first level of organization possess the transcription factor binding sites (TFBSs). A pair of neighboring TFBSs organizes the so called composite element and in that case their joint action appears to be synergetic and different from if they act independently [1]. The up next level of organization consists of composite elements, promoters, silencers and enhancers. Block-hierarchical structure of eukaryotic regulatory regions provides flexible regulation of genes expression on transcription stage by switching separate elements. Thus each level of organization states its own task in front of investigators. First of all, there is the task of TFBSs prediction, the up next task is the TFBSs pattern discovery, and the highest hierarchical level corresponds to the system of integral regulation of transcription defined by the superposition of different DNA codes [2].

We developed ExpertDiscovery system [3] that finds the hierarchically complicating set of complex signals. The complex signal definition is introduced recursively:

the elementary signal (e.g. nucleotide, oligonucleotide) is the complex signal;

a pair of ordered complex signals located on some "distance" from each other is the complex signal. Distance varies in the user specified range;

the result of predicates "repetition", "orientation", "interval" applied to the complex signal is the complex signal.

The current signal becomes complicated, if the new complicated signal possesses the higher conditional probability value and the lower significance level (Fisher criterion).

ExpertDiscovery methodology provides the possibility of hierarchical analysis of regulatory regions. In the case when the elementary signals are nucleotides the system reveals the mutual interdependencies between the nucleotides rather distant from each other in the general case. We analyzed the DNA tar-

[1] Institute of Cytology and Genetics, 630090, Novosibirsk, Lavrentyev aven., 10, Russia khomicheva@bionet.nsc.ru
[2] Sobolev Institute of Mathematics, 630090, Novosibirsk, 4 Acad. Koptyug av., Russia vityaev@math.nsc.ru

gets of three protein families: steroidogenic factor-1 (SF1), sterol regulatory element binding protein (SREBP), early growth response factor 1 (EGR1). The comparison of the ExpertDiscovery system performance with the optimized positional weight matrix showed that the system doesn't yield to the matrix and even outperforms it.

At the next level of hierarchy as the elementary signals to construct the complex the system takes: putative functional sites; degenerate oligonucleotide motifs [4]; sites with conservative conformational or physical-chemical features [5]; the nucleosome formation sites [6].

The system was applied to regulatory regions analysis and prediction of interferon-stimulated genes, expressed in macrophage and genes of endocrine system. Complex signals discovered by the system correspond to the statistically and biologically significant combinations of potential TFBSs and degenerate motifs, among them biologically relevant cooperating TFBSs pairs, composite elements, verified in literature.

1. O. Kel-Margoulis et al. (2002) Transcompel. Nucl. Acids Res. 30: 332-334.
2. E.N. Trifonov (1997) Genetic level of DNA sequences is determined by superposition of many codes, Mol. Biol. (Mosk), 31: 759-767.
3. E.E. Vityaev, T.I. Shipilov (2006) Software for analysis of gene regulatory sequences by knowledge discovery methods. *Bioinformatics of Genome Regulation and Structure II*, Springer Science+Business Media, Inc., 491-498.
4. O.V. Vishnevsky, N.A. Kolchanov (2005) ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters, *Nucleic Acids Res.,* **33**: 417-422.
5. D.Y. Oshchepkov et al. (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition, *Nucleic Acids Res.*, **32**(Web Server issue): 208-12.
6. V.G. Levitsky et al. (2005) NPRD: Nucleosome Positioning Region Database, *Nucl. Acids. Res.*, **33:** 67-70.

MOLECULAR MODELING OF MRFP1 MUTANT STRUCTURES AND CORRELATIONS WITH THEIR PROPERTIES

EKATERINA E. KHRAMEEVA

The fluorescent proteins are genetically encoded, easily imaged reporters crucial in biology and biotechnology. They provide high sensitivity and great mobility while minimally perturbing the cell under investigation. At the present time new forms of red fluorescent proteins with improved properties, such as quantum yield, brightness, photo- and pH-stability, are created.

However, correlations between fluorescent protein properties and their structure, essential for such research development, have not been investigated yet. Methods of forecast of fluorescent protein spectral properties by their structure do not exist, i.e. the design algorithm for proteins with predefined properties has not been created. Analysis of mutant protein properties and subsequent molecular modeling of their structures *in silico* to find the correlations may solve the problem.

In this work molecular modeling simulations for mRFP1 and its 18 mutants containing point mutations in 66th position have been performed. Comparison of received structures has shown that the general form of β-barrel has remained the same and chromophores seem to be planar. Observed modifications in mutant structures are likely to cause variations of spectral properties, so they need investigating in details.

To obtain such more detailed information about conformational changes, all equilibrium bond, angle and dihedral values in chromophores and their 6Å environments have been calculated. However, variations of bond and angle values turned out too small to find valid correlations. Therefore, only dihedral values have been analyzed.

We calculated correlation coefficients between chromophore dihedral values and some known spectral properties defined experimentally, such as absorption maxima, extinction coefficients, excitation maxima, emission maxima and quantum yields. All dihedrals having the best correlation coefficients were found to be placed in the regions of connection between phenolic and imidazolidonic rings of chromophore and into the rings themselves. As chromophore is the system of conjugated π-bonds, planarity of its structure is the main characteristic ensuring high protein fluorescence and brightness. Thus, variations of such dihedrals point at deformation of chromophore planarity and hence changes of protein spectral properties.

Faculty of Bioengineering and Bioinformatics, Moscow State University, Russia,
khrameeva@yandex.ru

We also calculated correlation coefficients between dihedral values in 6Å chromophore environment and the same spectral properties. Residues containing dihedrals with the best correlation coefficients are located extremely close to the chromophore, so variations of their conformations are likely to relate to chromophore structure modifications caused by amino acid substitutions. The search of such residues is important for directed mutagenesis as these amino acid substitutions may affect chromophore structure and, hence, spectral properties, pretty much.

Correlations between mutant spectral properties, such as excitation and emission maxima, and volumes of substituting residues were also found, with excellent correlation coefficients for the selection containing only polar residues. Having such correlations, now we are able to predict some mutant properties by their structure and to plan mutagenesis by predefined protein properties.

1. D.V.Dmitrienko et al. (2006) Red Fluorescent Protein DsRed: Parametrization of Its Chromophore as an Amino Acid Residue for Computer Modeling in the OPLS-AA Force Field, *Biochemistry (Mosc.),* **71(10)**:1133-52.
2. N.C.Shaner et al. (2004) Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein, *Nat Biotechnol.,* **22**:1567–72.

ITERATIVE PROTEIN ALIGNMENT ALGORITHM (IPA)

TATSIANA KIRYS[1], SERGEJ FERANCHUK[1], ALEXANDER TUZIKOV[1], JAIRO ROCHA[2]

Many algorithms have been proposed to solve the problem of spatial comparison of proteins, such as the double dynamic programming method, the Monte Carlo method, the iterated dynamic programming method, the path extension method and the hierarchical alignment method. But in spite of the fact that this problem has been intensively studied during last decades none of the existing methods gives a satisfactory solution to the problem. Partially, this discrepancy is due to the absence of a commonly accepted measure for structural similarity.

[1] Belarusian State University, 4. Nezalezhnosty, 220000 Minsk, Belarus, nushki@mail.ru: UIIP NAS of Belarus, 6 Surganov, 220012 Minsk, Belarus, tuzikov@newman.bas-net.by
[2] University of the Balearic Islands, Spain, jairo@uib.es

In general, two problems must be solved to compare structure pairs automatically. The first one is to define similarity score functions between structure pairs. The second problem is to develop an efficient algorithm that finds the structural corresponding residues (alignment) with the largest similarity score.

Let two proteins be represented by coordinates of their C_α atoms in space and the secondary structures of proteins be known. The problem consists of finding Euclidean superposition of proteins as point sets in space, which minimizes rmsd, in other words it is necessary determine the largest common subsets, that are geometrically similar. Protein primary structure is not considered.

Our algorithm consists of two steps. At the first step initial matches are found, and at the second step they are refined using algorithm FICP [1].

Initial protein matching uses comparison of their secondary structure elements. Each element of secondary structure is described by a vector representing its principal axis of inertia. All possible pairs of vectors from both proteins are chosen to compute similarity function values. A number of matched vector pairs with the best values of similarity function are considered as initial matches. Then, for all initial matches FICP is applied to improve correspondence between proteins. The result depending on the goals may be given by several alignments of proteins.

FICP is a variation of the iterative closest point (ICP) algorithm for aligning two roughly pre-registered point sets under a set of transformations. This iterative algorithm has three basic steps:

1. pair each point of P to the closest point in M;

2. compute a motion that minimises rmsd between a fraction of paired points;

3. apply the motion to P and update rmsd.

The three steps are iterated; the iterations have been proved to converge in terms of the rmsd. The value of fraction of points in P used to compute the optimal motion is found as the minimum of a specific function.

The ICP algorithm was implemented in C++ using MATLAB. For closely-related proteins the results obtained by IPA are comparable with known data. At present SSM [2] is especially efficient compared to other algorithms of protein structure comparison in three dimensions. Therefore we compare our algorithm with SSM. In most cases our algorithm superimposes fewer residues than SSM, however the value of rmsd is also less. Our algorithm is quite fast, on average it converges in 20 iterations, each iteration takes several seconds. The results of the experiments are given in the report.

This work was partially supported by INTAS project 04-77-7178.

1. J. M. Phillips, R. Liu (2006) Outlier Robust ICP for Minimizing Fractional RMSD, *ArXiv Computer Science e-prints,* cs/0606098.
2. E. Krissinel, K. Henrick (2004) Secondary-structure matching ({SSM}), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallogr D Biol Crystallogr,* ***D60***: 2256-2268.

GRAPHICAL REPRESENTATION OF CELL/TISSUE TYPE RELATIONSHIPS

LARISA KISELEVA[1], RAYMOND WAN[2], PAUL HORTON[1]

Graphs are becoming more and more popular for representing various types of biological information, to the point that we can hardly imagine depicting schemes of protein interactions, signal transduction, metabolic pathways, food webs, or phylogenetic trees other than as a set of vertices and edges. However, although playing an increasingly important role in representing interactions and relations between molecules or organisms, graphs have not been extensively applied to visualization and analysis of relations between such important biological units as cells and tissues.

Recently we have proposed “Cell type relation networks”, undirected graphs in which nodes represent cell or tissue types, and edges represent similarity between given cell/tissue types. For a similarity measure, we use Pearson's or Spearman's correlation coefficient of global gene expression, which has been shown to be useful for automatic discrimination of cell tissue type (1).

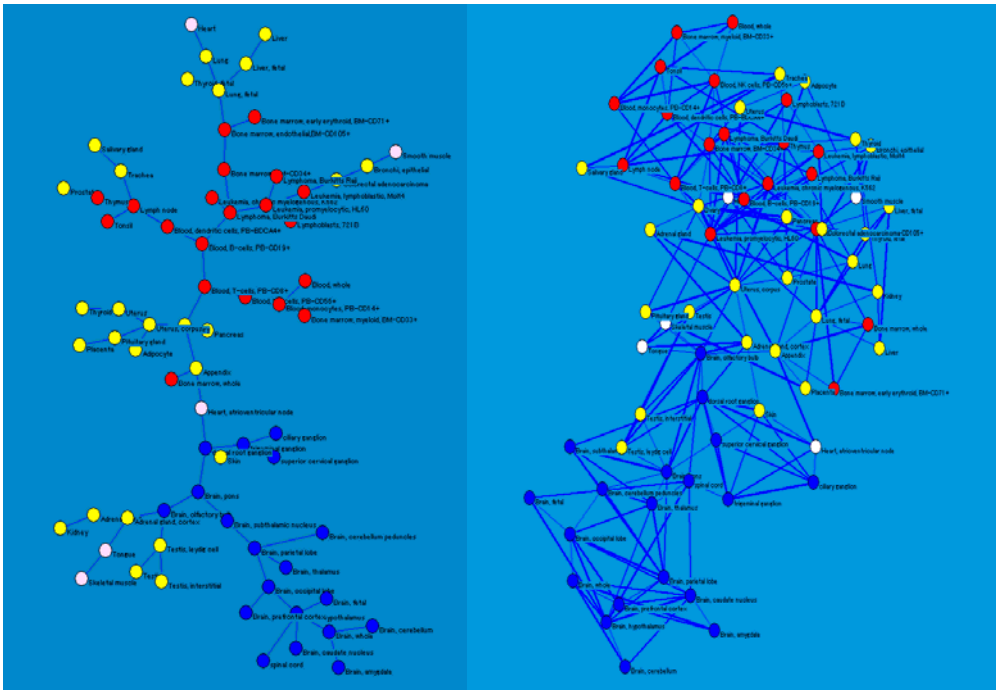
Expression level data of about 8000 genes in various human cell types were obtained from Gene Expression Omniubs, the largest publicly available repository of microarray data (2). We choose Spearman correlation as the similarity metric and calculated it between each pair of tissue types. To connect the nodes in this network we adopted the minimum spanning tree (MST), a concept of graph theory. As a result we obtained an undirected acyclic graph where cell types of similar origin and function appeared to be linked.

In this work we focus on issues of interpretation and reliability of cell/tissue relationship networks, as we have found noticeable differences between graphs produced using different similarity metrics or different experimental replicates. We propose several solutions to improve the graphical representation of rela-

[1] Computational Biology Research Center, 2-42 Aomi, Koto-ku, Tokyo, Japan, kiseleva-larisa@aist.go.jp
[2] Kyoto University, Institute forChemical Research, Bioinformatics Center, Japan rwan@kuicr.kyoto-u.ac.jp

tionships between various types of cells and tissues and expect applying the results of this work will be useful for studying the processes of development, cell differentiation, and disease progression.

1. P. Horton, L. Kiseleva, W. Fujibuchi (2006) RaPiDS:An algorithm for rapid expression profile database search, *Genome Informatics,* **17** (2): 67-76.
2. A.I. Su et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc. Natl. Acad. Sci. USA*, 101, **16**:6062-6067.



Cell/tissue type relationship networks

OPTIMIZATION OF RESOURCES DISTRIBUTION FOR HIGH-PERFORMANCE COMPUTATION

ALEXEY KOBETS[1], KIRILL VOTYAKOV[2], VASILY LUKOVNIKOV[1]

Modern scientific works in molecular biology demand the increasing computer capacities for data processing. This problem becomes sharper in conditions of resources shortage or distribution of resources in parallel calculations of complex computational-intensive tasks. The majority of existing computational models are based on parallel computing of a task, not caring about resources distribution problems. Therefore necessity for planning distribution of system recourses of the given computation is obvious - differently shortage of the resource can become a bottleneck in productivity of all system.

In this work we raise a question of optimization of resources consumption process for high-performance computation. We propose criteria for quality of planning by resources for groups of tasks at macrolevel. All resources of operation system can be divided on three groups: redistributable (f.e. CPU, network or disk throughput), non-redistributable (f.e. disk space) and partially-redistributable (f.e. physical memory).

We extended the model and terminology for redistributable resources from [1] for groups of computational tasks. Desirable function of consumption, on which it is possible to impose management by internal means of operational system schedulers, was offered for group of the tasks.

It was established, that for achievement of acceptable accuracy of distribution management it is necessary to impose certain restrictions (assumptions) on the behavior of function of group's consumption. With reference to algorithms of planning of redistributable resources used in modern operational systems, the disturbances caused by effects of environment, management and feedback should be limited on the time interval. Thus, imposing restrictions on function of consumption [2] and by changing consumption rate (satisfying criterion of quality of distribution, slowing down own time of group) and analyzing behavior of system in real time we can reach required accuracy of distribution of redistributable resources for group of computational tasks.

Using approach similar to Kalman filter [3] we created algorithm of imposed resources management for tasks in group to achieve required distribution of

[1] Moscow Institute of Physics and Technology, Chair of Computer Science, Dolgoprudny, Russia, kobets@sw.ru, vlukovnikov@sw.ru
[2] Institute of Biological Information Processing (IBI-2), Research Center Juelich, Germany, votyakov@fz-juelich.de

resources for the whole group, investigated a number of the edge effects connected with imposed distribution management and established, that the disturbances caused by such effects, are exponentially decreasing.

We are personally thankful for numerous discussions to A. G. Tormasov.

1. A. Tormasov et al. (2006), Basic problems of maintenance of accuracy of management at the imposed resource management in modern operation systems, *Processes and methods of processing of the information.*
2. I. Lukovnikov et al. (2006), Problems of management of distributed resources of OS, *Information technologies,* 10, 71-78.
3. R. E. Kalman (1960), New approach to linear filtering and prediction problems, *Transaction of the ASME – Journal of Basic Engineering*.
4. D. A. Solomon et al., Inside Microsoft® Windows® 2000, Third Edition, *Microsoft Press*, ISBDN 0-7356-1021-5

MODELLING AND ANALYSIS OF MOLECULAR PROCESSES IN DUCHENNE MUSCULAR DYSTROPHY USING PETRI NETS

I. KOCH[1], S. GRUNWALD[1], J. ACKERMANN[2], A. SPEER[1]

The classical approaches of systems biology, which are based on differential equations, can be used only, if a critical amount of parameters is known. In order to get nevertheless insights into the behaviour of the biological system, qualitative methods have been developed, which are all based on the incidence matrix of the underlying net graph. A more general approach represents the Petri net theory [1]. Petri nets provide the possibility to describe biochemical networks at different abstraction levels in a unique manner. There are techniques based on Petri net theory to compute structural as well as dynamic properties of networks without any knowledge of kinetic data. Since nearly 15 years Petri net modelling has been applied to different types of biochemical systems, to metabolic networks [2], signal transduction [3] and gene regulatory networks [4], and even to networks combining different abstraction levels [5].

Duchenne muscular dystrophy (DMD) is one of the most common inherited human neuromus-cular diseases. It usually affects boys before the age of six

[1] Technical University of Applied Sciences, Dept. Biotechnology/Bioinformatics, Seestr. 64, 13347 Berlin, Germany
[2] FluIT, Schloss Birlinghoven, 53757 S t. Augustin, Germany
ina.koch@tfh-berlin.de, stefanie.grunwald@web.de, joerg.ackermann@fluit-biosystems.de, astrid.speer@gmx.de

years and leads to death by respiratory and/or cardiac failure by the age of 20 years because no efficient therapy is available yet. The disorder is caused by mutations in the dystrophin gene followed by the absence or functional impairment of the subsarcolemmal cytoskeletal protein. Experimental research identified dystrophin as part of a complex interacting system of different proteins [6]. But, in contrast to the well analysed genetic aetiology, the network downstream of dystrophin, which represents the pathomechanisms of DMD, is neither registered nor understood completely. For this reason we have performed Real-Time PCR experiments to compare mRNA expression levels of relevant genes in tissues of affected patients and controls [7, 8].

In order to bring experimental data in context with the underlying pathway we have developed a Petri net model, which is based mainly on own experimental and literature data. We distinguish between up- and down-regulated states of gene expression. The analysis of the model comprises the computation of structural and dynamic properties with focus on a thorough t-invariant analysis including clustering techniques and the decomposition of the network into maximal common transition sets, which can be interpreted as functionally related building blocks [9]. We discuss all possible pathways, which reflect the complex net behaviour in dependence of different gene expression patterns. The resulted model serves as basis for a better understanding of pathological processes, and thereby for planning next experimental steps in searching for new therapeutic possibilities.

1. J.L. Peterson (1981) *Petri Net Theory and the Modeling of Systems*, Prentice-Hall, Inc. New Jersey
2. I. Koch, B.H. Junker, M. Heiner (2005) Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber Bioinformatics, *Bioinformatics,* **21**: 1219-1226.
3. M. Heiner, I. Koch, J. Will (2004) Model Validation of Biological Pathways Using Petri Nets - Demonstrated for Apoptosis, *BioSystems* **75**(1-3): 15-28.
4. H. Matsuno, Y. Tanaka, H. Aoshima, A. Doi, M. Matsui, S. Miyano (2003) Biopathways representation and simulation on hybrid functional Petri net, *In Silico* Biology, **3**(3): 389-404.
5. E. Simão, E. Remy, D. Thieffry, C. Chaouiya (2005) Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in *E.Coli*. *Bioinformatics,* **21**(Suppl 2): 190-196.
6. J.V. Chakkalakal, J. Thompson, R.J. Parks, B.J. Jasmin, (2005) Molecular, cellular, and pharmacological therapies for Duchenne/Becker muscular dystrophies, *FASEB J.,* **19**: 880-891.

7. M. Sifringer, B. Uhlenberg, S. Lammel, R. Hanke, B. Neumann, A. von Moers, I. Koch, A. Speer (2004) Identification of transcripts from a subtraction library which might be responsible for the mild phenotype in an intrafamilially variable course of Duchenne muscular dystrophy, *Human Genetics,* **114**:149-156.
8. S. Grunwald, A. von Moers, I. Koch, E. Hobbiebrunken, E. Wilichowski, A. Speer (2005) , *13th ESGT Prague 2005*, **69**:S59.
9. A. Sackmann, M. Heiner, I. Koch (2006) Application of Petri net based analysis techniques to signal transduction pathways, *BMC Bioinformatics*, **7**: 482.

SIGNALS INFLUENCING GENERAL TRANSLATION EFFICIENCY OF EUKARYOTIC MRNAS

ALEX V. KOCHETOV[1], VLADIMIR IVANISENKO[1], IGOR I. TITOV[1], NIKOLAY A. KOLCHANOV[1] AKINORI SARAI[2]

Sequence features of 5'-UTR (AUG triplets and open reading frames (ORFs), secondary structure) are emerging as important mediators of transcript-specific general translational control. However, these elements are commonly not taken into account in a prediction of mRNA translational characteristics and coding potential. Computational analysis of sequence organization of eukaryotic mRNAs provides a valuable tool to reveal significant features influencing translational activity.

First, we found significant correlation between the context of annotated translation start site (TSS) and the occurrence of in-frame downstream AUG codons [1]. This correlation probably reflects the complex organization of some eukaryotic translation initiation signals: usage of alternative TSSs and synthesis of several protein isoforms (potentially) possessing different functions [2]. In this case leaky scanning mechanism is likely to be used to synthesize additional protein forms frequently targeted to different subcellular compartments.

Second, we evaluated the role of small upstream ORFs in the control of alternative TSS selection. It was found that ca. 1000 human mRNAs contain small uORFs overlapping with annotated TSS and potentially delivering ribosomes to downstream alternative start sites by a reinitiation mechanism [3]. According to the results obtained, reinitiation of translation could also play an important role in the synthesis of new functional protein isoforms in eukaryotic cells.

[1] Institute of Cytology and Genetics, Novosibirsk, Russia, ak@bionet.nsc.ru
[2] Kyushu Institute of Technology, Iizuka, Japan, sarai@bse.kyutech.ac.jp

Third, the role of secondary structure in a selection of AUG codons in a weak context was investigated. It was found that a suboptimal start codon context significantly correlated with higher base pairing probabilities at positions 13 – 17 of CDS of mammalian mRNAs. It may be assumed that a stable hairpin in this region could delay the 40S ribosomal subunit just in the position providing an efficient interaction between the met-tRNAi-located anticodon and the start AUG codon. It is likely that this mechanism is used to enhance translation of some mammalian mRNAs *in vivo*. We developed a program (AUG_hairpin) aimed to prediction of "compensatory" hairpins within mRNAs and re-evaluation of translation efficiency of start AUG codons in suboptimal contexts [4,5].

This work was supported by the Programs of RAS (Dynamics of Gene Pools, Origin and Evolution of Biosphere), the RFBR (grant No. 05-04-48207), and JSPS. We thank SD RAS (grant No. 5.3) for partial support

1. A.V. Kochetov (2005) AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context, *Bioinformatics*, **21**: 837-840.
2. A.V. Kochetov et al. (2005) The role of alternative translation start sites in generation of human protein diversity, *Mol. Genet. Genomics*, **273**: 491-496.
3. M. Kozak (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**: 13-37.
4. http://wwwmgs.bionet.nsc.ru/mgs/programs/aug_hairpin/
5. http://gibk26.bse.kyutech.ac.jp/aug_hairpin/

APPLICATION OF COMPUTER SIMULATION FOR STUDY OF C-DOMAIN STRUCTURE OF M1 PROTEIN OF INFLUENZA VIRUS A BY TRITIUM PLANIGRAPHY METHOD

A.B. KOLOTILOVA, A.L. CHULICHKOV, E.N. BOGACHEVA, A.A. DOLGOV, A.V. SHISHKOV

The problem of the reconstitution of spatial structure of macromolecules from the primary amino acid sequence is in the center of attention of numerous researchers. In our experiments we have used an original method developed by us - the method of tritium planigraphy. The information obtained by tritium planigraphy brings the data about steric accessibility of hydrocarbon fragments of macromolecule, which by itself is directly connected with its spatial structure, reflects the «architecture» of the object [1].

N.N. Semenov Institute of Chemical Physics Russian Academy of Sciences, ul. Kosygina, 4, Moscow, 119991 Russia, e-mail: avs@chph.ras.ru

The introduction of tritium label in biological compounds is realized by the bombardment with the beam of hot tritium atoms of the target cooled down by fast freezing of the drops of protein solution sprayed over the wall of reactor at liquid nitrogen temperature. The introduction of tritium labels under such conditions occurs through single collisions of tritium atoms with the target, and the intramolecular distribution of the labels among the residues of amino acids is determined by their accessibility in the macromolecule. The range of hot tritium atoms obtained under such conditions is measured by few angstroms, and therefore the location of tagged species (e.g. of C-T bonds) is restricted by thin surface layer of investigated object (peculiar analog of suntan). Analysis of the distribution of labels in the investigated object is usually realized at the level of the separate amino acids, which is attained by the fragmentation of tagged proteins by various proteases into short peptides, where the amounts of identical amino acids are minimized. Such procedure allows determining the relative level of exposure of amino acid residues by tritium, gives detailed information on the structure of the surface and preliminary conclusions concerning the stacking of residues in macromolecule.

We propose semiempirical algorithm for the construction of spatial structure of proteins. The suggested algorithm of proteins spatial structure modeling consist in a combination of experimental data of tritium planigraphy method, computer simulation of the conditions of tritium labeling and prediction of spatial structure of protein by means of modified program Rosetta.

We have suggested semiempirical algorithm to construct the spatial structure of globular proteins which includes following stages:

Experimental determination of the profile of accessibility of amino acid residues by bombardment of the protein by tritium atomic beams.

Theoretical prediction of the elements of secondary structure of protein by traditional methods.

Determination of the profile of accessibility of amino acid residues in the isolated elements of secondary structure by the computer simulation of experiment.

Determination of contact regions between the secondary structure elements by the comparison of experimental and simulated profiles of accessibility.

Assembly of the secondary structure elements into the compact model (with the consideration of localized contact regions).

On the final stage we combine our data with the program Rosetta. The data of simulation algorithm of the tritium bombardment and the experimental data were compared with the theoretically predicted three-dimensional structure of the C-domain of M1 protein using the Rosetta program [2]. Basing on the re-

sults obtained this way the clusters with the best correlation between the methods were allocated. The application of the combined approach allowed reducing substantially the hypothetically possible of the C-domain spatial structures.

Updating of the program of construction of three-dimensional structure of proteins Rosetta in a mode *ab initio* is lead. Studying of algorithms of this program has shown an opportunity of the account of experimental data of tritium planigraphy for more correct construction 3D structures. Transition from approximation of side chains of amino acid the residues as centroids to fullatom description of protein is carried out. As a result it was possible to organize calculation of the accessible surface areas to solvent during construction of 3D structures. Functions are in addition entered into program Rosetta for calculation of surfaces of contact "active" (from the point of view of tritium planigraphy) atoms and procedure for reading experimental data and calculation of similarity factor. This factor is used as composed in the general Score function of program Rosetta. Trial calculations with the purpose to pick up parameters of algorithm of comparison and to define optimum weight of our new criterion in the general Score function of program Rosetta are lead.

This work was partially supported by the International Science and Technology Center (BTEP#82/ISTC#2816) and the Russian Foundation for Basic Research (06-03-32377).

1. E.N. Bogacheva, V.I. Goldanskii, A.V. Shishkov, A.V. Galkin, L.A. Baratova (1998) Proc. Natl. Acad. Sci. USA, 95:2790-2794.
2. K.M. Misura, D. Chivian, C.A. Rohl, D.E. Kim, D. Baker (2006) Proc. Natl. Acad. Sci. USA. **103**, 5361-536.

CHRUNTA – TANDEM REPEAT SEARCH AND CLASSIFICATION PROGRAM

KOMISSAROV A.S, PODGORNAYA O.I.

Centromeric (CEN) and pericentromeric (periCEN) regions remain «white spots» in chromosome maps obtained from sequencing of the human and mouse genomes. The periCEN heterochromatic regions of mammalian chromosomes usually consist of tandemly repeated satellite DNA (satDNA). The mouse major (MaSat, periCEN) and minor (MiSat, CEN) satellites are known for more than 20 years and well characterized at the molecular level. Mouse satDNA MS3 and MS4 fragments were recently isolated, characterized and mapped from a library of DNA fragments from isolated chromocenters of in-

Institute of Cytology RAS, 194064, Saint-Petersburg, Russia, ad3002@gmail.com

terphase nuclei [1]. Computer analysis of MS3 and MS4 sequences by alignment, fragment curvature, and search for matrix attachment region (MAR) motifs in comparison with MaSat and MiSat showed them to be new satDNA fragments. The CEN region consists of a small block of MiSat and MS3 followed by a periCEN block of MaSat combined with MS4. 2.2% of the total DNA consists of MS3 with a 160 bp long monomer. MS4 monomer is 300 bp long and accounts for 1.6% of the total DNA. MaSat and MiSat constitute significant portions of the mouse genome (~5% and ~1%, respectively) and both of them are present in mouse genome database, but we failed to find MS3 or MS4 in various databases by usual approaches. 23 sequences that are absent from the mouse genome database were cloned from the same chromocenters' fragments DNA library. We mapped 7 of them using fluorescent in situ hybridization (FISH) in interphase nuclei, which revealed them to compose more than 1% of the genome.

The aim of this work was to develop an approach of searching for a new class of tandem repeats based on few monomers cloned that are absent in databases due to satDNA variability. Chromosome Unknown database is expected to be enriched with various repetitive sequence types, including tandem repeats. Mouse Chromosome Unknown database consists of 54752 entries 2 Mb long in total [2]. Tandem Repeat Finder [3] was used in the initial search and found 35 thousands of unique consensus sequences. The CHRUNTA (**chr**omosome **un**known **t**andem **a**nalyzer) program was written to sort these sequences. GC content is the main parameter used by CHRUNTA and sequence homology is the secondary parameter. Thousands of consensus sequences were collected in a 100 groups according to GC content. Then these groups were sorted according to sequence homology by cluster analysis. Cluster volume depends on variability of specific tandem repeat types. Among obtained tandem repeat classes we identified all of rodent tandem repeats with the help of REPBASE [4].

The term "satellite DNA" arised from early experiments of ultrasonicated total DNA centrifugation in CsCl gradient where GC-rich DNA formed additional, "satellite" zone [5]. Among the 372 clusters obtained by CHRUNTA there were several of considerable size which contained satDNA consensus sequences. AT-rich MaSat and MiSat form the largest clusters (38% GC) and GC-rich MS3 and MS4 were found among consensus sequences in the second largest cluster (51% GC). 14 out of the 23 clones from chromocenters' library including those 7 mapped by FISH were found among minor clusters, probably due to the fact that they constitute a large fraction of the genome.

Consensus sequences of clusters differ from original cloned sequences by 5-15%, and now they are well recognized by BLAST [6]. Among revealed clusters we also found fragments homologous to SINE, LINE, repetitive elements of other species and about a 100 previously undescribed consensus sequences. Experimental verification of their belonging to the mouse genome and their repetitive nature is in progress. Nevertheless CHRUNTA has already proved its value for searching of unknown repetitive elements of the mouse genome.

1. I. Kuznetsova et al. (2006) High-resolution organization of mouse centromeric and pericentromeric DNA, Cytogenet Genome Res., 112:248-55.
2. ftp://ftp.ncbi.nih.gov/genomes/M_musculus/CHR_Un/
3. G. Benson (1999) Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res., 27(2):573-80.
4. O. Kohany et al. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor, BMC Bioinformatics, 7:474.
5. J.P. Thiery et al. (1976) An analysis of eukaryotic genomes by density gradient centrifugation, J Mol Biol., 108(1):219-35.
6. http://www.ncbi.nih.gov/BLAST.

A STOCHASTIC ADVANTAGE OF SEX?

ALEXEY S. KONDRASHOV[1] AND TIMOFEY A. KONDRASHOV[2]

Under epistatic selection, sex can reduce the mutation load even in an infinite population. However, if selection against deleterious mutations is multiplicative, sex can only be advantageous in a finite population. The only well-known process that leads to the advantage of sex of this kind is slowly-acting Muller's ratchet. Recently, Keightley and Otto (Nature 443: 89-92, 2006) described another mechanism, based on Hill-Robertson interference, for the advantage of sex under the same assumptions, and claimed that this mechanism could be of biological importance. We performed individual-based simulations of a population polymorphic for a modifier locus that can lead to either asex or sex and never observed a substantial advantage of sex under multiplicative selection against mutations. Thus, it does not seem that Hill-Robertson interference played a role in the evolution of sex, as long as only deleterious mutations

[1] Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA
[2] Dexter High School, Dexter MI 48130, USA

are considered. We discuss implications of our result for the general problem of the evolution of sex.

POSITION-SPECIFIC CORRELATIONS BETWEEN SEQUENCES OF LACI FAMILY DNA BINDING DOMAINS AND THEIR OPERATORS

Y. D. KOROSTELEV, O. N. LAIKOVA, A. B. RAKHMANINOVA

Specific DNA-protein interaction is involved in all genetic processes including replication, reparation, restriction and regulation of expression. While some regulating rules have been observed, the mechanisms providing specificity are not clear.

Our group developed a unique database of bacterial transcription regulators of the LACI family and their binding sites obtained from experimental data or predicted by comparative genomics methods (*O. Laikova*, MCCMB'05). The database size allowed for a detailed statistical analysis. We aligned 949 DNA binding domains and 2934 their respective operators and then examined correlations between positions in the domain alignment and in the operator alignment. As a measure of correlation between columns, a mutual information $I_{i,j}=^{4}\sum_{n=1}^{20}\sum_{a=1}[p_{i,j}(a,n)\cdot log(p_{i,j}(a,n))/(p_i(a)p_j(n))]$ was used, where $p_{i,j}(a,n)$ is the frequency of simultaneous occurrence of residue ***a*** in column ***i*** and nucleotide ***n*** in column ***j***. $p_i(a)$ and $p_j(n)$ are the frequencies of residue ***a*** occurrence in column ***i*** and nucleotide ***n*** occurrence in column ***j,*** respectively. A program has been created for calculating $I_{i,j}$ for each ***i-j*** pair. The statistical significance of obtained $I_{i,j}$ was estimated and corrected to phylogenetic trace by a previously published algorithm[1].

Only 3% (53 of 1660) significantly correlated pairs were found. These pairs correspond to nucleotides 701-706 and residues 3, 4, 13-16, 20, 26, 55, 57 in the 3D-structure of PurR_Ecoli (pdb: 1qpz). Residues 4, 15, 16, 20, 26 form close contacts with the DNA bases. All contacting pairs (8 of 8) of the residue side chain and the DNA base and all specificity determining positions (SDP) obtained in [1] were among correlated pairs.

The obtained correlations are position specific: in addition to standard arginine–guanine preferences we observed cases where, for an instance, pairs Arg55–G706 and Lys55-G706 are strictly forbidden, which is in agreement with the experimental data[2].

Moscow State University, Department of Bioengineerig and Bioinformatics, Moscow, Russia

Let [i-j] be the correlated pair of ***i***-column in the protein alignment and ***j***-column in the operator alignment. The list of 15 best correlated pairs contained [16-703], [16-704] and [16-711]. These pairs correspond to residue Thr16 that forms direct contacts with Cyt703, Ade704 and Thy711. The next pair from the list is [20-702] that in more than 2/3 of occurrences is represented by contacting pair Arg-G (in the PurR_Ecoli structure Arg20 is substituted by His20 which makes a water mediated H-bond with G702). Previously discussed residues His20 and Thr16 are located on the same side of the recognition helix and form direct contacts with DNA bases. They are the most crucial residues in the DNA recognition rules from Suzuki *et al.* [3]. Moreover, these residues seem to be anticorrelated: Arg20 avoids T702 (G702 is in 2/3 of occurrences) and, on the other hand, Thr16 prefers T702 (though they are not in direct contact). When both Thr16 and arginine702 are present, G702 is no longer preferred and T702 is no longer avoided (and not preferred).

To summarize, our approach based on statistical analysis of sequences allows to reveal residues and nucleotides that contribute to DNA recognition without using information of 3D-structures of complexes.

This is joint work with M. S. Gelfand.

We are grateful to A. A. Mironov for useful discussions.

1. Olga V. Kalinina, Andrey A. Mironov, Mikhail S. Gelfand and Aleksandra B. Rakhmaninova. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci 13(2): 443-56
2. Glasfeld A, Koehler AN, Schumacher MA, Brennan RG. The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions.
3. Masashi Suzuki, Steven E. Brenner, Mark Gerstein and Naoto Yagi. DNA recognition code of transcription factors. Protein Engineering vol.8 no.4 pp.319-328, 1995.

SIGNALING GLIA AND EVOLUTIONARY ORIGIN OF CIRCUMVENTRICULAR ORGANS IN VERTEBRATES

VLADIMIR KORZH

The circumventricular organs (CVOs) represent the semitransparent areas of the blood-brain barrier (BBB) used for hormonal communication between body and internal milieu of the brain. These include the organum vasculosum of the lamina terminals (OVLT), median eminence (ME) and adjacent neurohypophysis (NH), subfornical organ (SFO), area postrema (AP), pineal gland (PG), intermediate lobe of hypophysis (IH), choroid plexus (CP) and subcommissural organ (SCO). Several specialized neural structures that consist of signaling glia act as organizing centers during neurodevelopment in vertebrates. The roof and floor plate produce antagonistic dorsal and ventral signals. Despite different functions these organs share morphological characteristics especially at the level of the trunk, where they are organized initially in two continuous and parallel lines of columnar cells along almost all anterior-posterior (A-P) axis at the dorsal and ventral midline of the neural tube. This apparently simple organization of the trunk roof plate is fragmented at the brain level, where new signaling centers appear along the dorso-ventral (D-V) axis, including the midbrain-hindbrain boundary (MHB), zona limitans intrathalamica (ZLI), and rhombomeric boundaries, etc. Similarly, the expression patterns of molecular markers, which are expressed at the trunk level continuously, become discontinuous at the level of the brain. The position of these fragmented domains of expression correlates with position of CVO, which, similarly to the signaling glia of roof and floor plate, occupy dorsal and ventral midline of the neural tube. Using Tol2 transposon-mediated enhancer-trap (ET) we developed several transgenic zebrafish lines with GFP expression restricted in the trunk to the roof plate and/or floor plate (Parinov et al., 2004; Choo et al., 2006; reviewed in Korzh, 2007). Interestingly, in the brain the distribution of domains of GFP expression correlates with positions of CVOs of zebrafish. The mapping of transgene insertion sites provided means for identification of genes which expression has been recapitulated by GFP expression pattern and their regulatory elements. This work is in progress. The morphogenesis of CVOs has been studied during normal development and in several mutants of zebrafish affecting major developmental signaling pathways, including Wnt, Hedgehog and Notch, using confocal microscopy and time-lapse cinematography *in vivo*.

Institute of Molecular and Cell Biology, Singapore, vlad@imcb.a-star.edu.sg

The GFP-positive islets of columnar cells that give rise to CVOs are morphologically similar to the midline structures, occupy similar position and, depending on their position, express molecular markers of floor plate or roof plate. Taken together this evidence suggests common origin of CVOs and roof and floor plates. We speculate that emergence of BBB during evolution triggered specification of some regions of signaling glia into CVOs resulting in fragmentation of the continuous floor and roof plates. Alternatively, the regionalization of the brain created the "weak" points in BBB helping to transform adjacent signaling regions of glia into CVOs.

Author is thankful to M. Garcia-Lecea, I. Kondrichin, S. Parinov, A. Emelyanov, B. Choo and other members of VK's laboratory in IMCB whose efforts contributed into this project. VK's laboratory is supported by a grant from the Agency for Science, Technology and Research of Singapore.

1. S. Parinov et al. (2004) Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes in vivo. *Dev. Dynam,* **231**: 449-459.
2. B. Choo et al. (2006) Zebrafish transgenic Enhancer TRAP line database (ZETRAP). *BMC Dev Biol* **6**:5.
3. V. Korzh (2007) Transposons as tools for enhancer-trap screens in vertebrates. *Genome Biology* (in press).

VIRTUAL INFORMATION MODELING OF LIFE SYSTEMS

N.E. KOSYKH, S.Z. SAVIN, V.V. GOSTUYSHKIN

The work is dedicated to the original method of Virtual information modeling (VIM) allows describing life system with any degree of accuracy that is available at the moment. The information n-dimentional model of life system is being developed. It is a number of dotsin a closed space of the skull. Each dot the geometric model consists of the set of the numerical indexes. These numerical indexes reflect the definite anatomical, histological and physiological areas. An universal character of VIM allow using them as an instrument for developing new applications in mathematical morphology, bioinformatics and biomathematics. The cell VIM ideology of animate nature microobjects and based on game-theoretic approach to the tasks of image identification, axono-

Computer Center of FEBRAS, 68000 Khabarovsk, Kim Yu Chen str., 65, Russia
nilekosykh@mail.redcom.ru, savin@as.khb.ru, gostusvv@mail.ru

metric principles of creating specialized mega data bases for sequence analysis. There is a complex of connections between the points in a closed space, which cover the whole area or only its separate parts. VIM is represented by a basic set of constituents: information codes of points, the area of spatial location of points (informational spatial area - ISA), functional connections between the points forming ISA and other points of the system, quantitative characteristics of ISA. A multidimensional numerical model distinctive feature is its tendency to a maximum structural similarity to specific parts of life system as well models of molecular evolution, regulatory systems, energetic, information and metabolic pathways in cell. The scientific significance VIM is connected with the development of a universal approach to creating a unified information system for prospective research of bioinformatics, bioalgorithms, cytology, radiology and human biology and also in the theoretical field of neuroinformatics, neurocomputer development, neurophysiology and neurobionics. The application of VIM in medicine computer diagnostic and clinical radiology practice gives encouraging results. This can have a positive effect in medicine and population genetics, systems biology and human ecology and education in bioinformatics.

Resume. The actual scientific significance of VIM researches is connected with the development of a universal approach to creating a unified informational models of life systems for prospective research of molecules and genomes structure, pathophysiological mechanisms of genetic and immune illness, researches in stem cell therapy, ecological physiology of humans and animals, development biology, oncology, radiology, medicine computer tomography and also in the theoretical field of biomathematics, bioinformatics, bioalgorithms, synergetic and neurobionics.

REGULATION OF METHIONINE AND CYSTEINE BIOSYNTHESIS IN STREPTOCOCCI

GALINA YU. KOVALEVA

Methionine occupies a central position in cellular metabolism. Regulation of methionine biosynthesis in bacteria involves various regulatory systems including S- and T-boxes mechanisms for most Gram-positive bacteria except Streptococcaceae that bear neither S-boxes nor methionine-specific T-boxes upstream most methionine biosynthesis genes [1].

Research and Training Center on Bioinformatics, Institute for Information Transmission Problems, RAS, Moscow, Russia, kovaleva@iitp.ru

Our data suggests that RNA-level regulation (S- or T-boxes) of methionine biosynthesis is substituted by DNA-level system in Streptococci. This DNA-level system involves two similar transcriptional regulators that are present in each genome. Binding profiles were predicted for both transcriptional factors and motif distribution upstream structural genes allowed functional assignment of each regulator. Thus, we ascribe FhuR as a cysteine biosynthesis regulator with binding consensus sequence TGATA-N_9-TATCA-$N_{2\text{-}3}$-TGATA, and MtaR as a methionine biosynthesis regulator with regulon covering most methionine/cysteine metabolism genes and binding consensus sequence TATAGTT-N_3-AACTATA. Latter motif was initially found in our group as MET-box. In this study we improved initial prediction and more strictly defined the corresponding regulatory protein.

Observed transcriptional regulation in Streptococci seems to be both conserved and flexible, as strong candidate signals were found upstream non-orthologous analogous genes and *vice versa* some cases of potential regulator exchange were observed for orthologous proteins. Comparative genomic approach also allowed to predict two amino acid permeases as new candidate members of FhuR regulon and to uncover of non-orthologous replacement *in situ* for one of these candidates in *S. suis* genome.

This is joint work with Mikhail Gelfand. We are grateful to Dmitry Rodionov for useful discussions. This study was partially supported by grants from the Howard Hughes Medical Institute (55005610), the Russian Academy of Science (Program “Molecular and Cellular biology”), and INTAS (05-8028).

1. D.A. Rodionov, A.G. Vitreschak, A.A. Mironov, M.S. Gelfand. (2004) Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. *Nucleic Acids Res.*, **32**, 3340-3353.

DETECTION OF MACROMOLECULAR ASSEMBLIES IN CRYSTALLINE STATE

EUGENE KRISSINEL

Macromolecular assemblies are complexes of more than one polypeptide and/or nucleotide chain that are stable in native environment. The way, in which polypeptide chains assemble, represents the protein quaternary structure. Often, although not always, an assembly is the biological unit that performs a certain physiological function by facilitating a respective biochemical

European Bioinformatics Institute, Cambridge CB10 1SD, UK, keb@ebi.ac.uk

processes. Functionality of many, if not most, proteins is not independent of the context of a macromolecular assembly. The examples are numerous and include well-known complexes such as e.g. the two gene product haemoglobin, holoenzymes, ion channels, DNA polymerase, microtubules, nucleosomes, virons and many others [1]. The biological significance of macromolecular complexes is truly immense and cannot be underestimated.

Physiological function of macromolecular assemblies is known to be closely related to their 3D structure. Few experimental techniques are available to infer on the structure of macromolecular complexes. Certain conclusions about the size and outer shape of an assembly may be derived from the mobility and mass measurements [2] and small-angle scattering [3]. NMR [4] allows for atom-level data, however this technique is not applicable to medium-size and large structures. Electron microscopy is not applicable to all objects and offers only low-resolution images. In practice, very few (percents of) Protein Data Bank [5] depositions come with experimentally verified multimeric state.

More than 80% of PDB entries were obtained by means of X-ray diffraction on macromolecular crystals [6]. It is reasonable to expect that stable complexes do not change during crystallization and therefore they should be identifiable in crystal packing. However, by convention, a PDB entry contains only atomic coordinates for the asymmetric unit (ASU) of a crystal, which, generally, has nothing in common with a biologically-meaningful assembly. The lack of a direct relationship between ASU and macromolecular assembly poses considerable difficulties for the identification of the latter in crystal packing in a universal manner.

In this study, a technique for automatic detection of macromolecular assemblies in crystalline state is developed. In difference of previous approaches [7,8], based on scoring of individual macromolecular interfaces, the proposed technique stands on principles of chemical thermodynamics. The presentation will address physical-chemical principles of macromolecular complexation and outline a theoretical model developed to assess complex stability by calculating both binding and entropy terms of free energy of dissociation. The role of entropy and associated symmetry effects on complex stability will be revealed and discussed. The developed model of complex stability is employed for automatic identification of potentially stable macromolecular assemblies in crystal packing using a graph-theoretical approach. As will be shown, the technique gives between 80% and 90% correct answers, thus making protein crystallography a major source of data on macromolecular assemblies. The approach has other useful features, such as possibility to predict dissociation patterns. The tech-

nique is implemented in software available as a public web service from the European Bioinformatics Institute at http://www.ebi.ac.uk/msd-srv/prot_int /pistart.html.

1. J.M. Berg, J.L. Tymoczko and L.Stryer (2002) *Biochemistry*, W.H. Freeman and Co., New York.
2. T. Liu and B. Chu (2002) Light Scattering by Proteins, In: *Encyclopedia of Surface and Colloid Science*, A. Hubbard (ed), Marcel Dekker Inc., New York, 3023-3043.
3. D.I. Svergun and M.H.J. Koch (2002) Advances in structure analysis using small-angle scattering in solution, *Cur.Opin.Struct.Biol.*, **12**, 654-660.
4. J. Cavanagh, W.J. Fairbrother, A.G. Palmer III and N.J. Skelton (1996) *Protein NMR Spectroscopy*, Academic Press.
5. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne (2000) The Protein Data Bank, *Nucleic Acids Res.*, **28**, 235-242.
6. T.L. Blundell and L.N. Johnson (1976) *Protein Crystallography*, Academic Press Inc. London.
7. K. Henrick and J. Thornton (1998) PQS: a protein quaternary structure file server, *Trends in Biochem. Sci.*, **23**, 358-361.
8. H. Ponstingl, T. Kabir and J. Thornton (2003) Automatic inference of protein quaternary structure from crystals, *J. Appl. Cryst.*, **36**, 1116-1122.

RARE MISSENSE POLYMORPHISMS: THE GOOD, THE BAD AND THE UGLY

GRIGORIY KRYUKOV, SHAMIL SUNYAEV

Several recent reports showed that common complex phenotypes can be caused by multiple rare non-synonymous variants[1, 2, 3], and proposed association studies based on complete re-sequencing of candidate genes. In a study of this design, a cumulative frequency of rare deleterious mutations in a candidate gene, rather than individual SNPs frequencies, is compared between disease and control cohorts. The success of such approach critically depends on the proportion of deleterious mutations among all detected missense polymorphisms and on our ability to distinguish deleterious amino acid substitutions

Division of Genetics, Department of Medicine, Brigham and Women's Hospital
Harvard Medical School, Boston, MA 02125, USA,
gkryukov@rics.bwh.harvard.edu, ssunyaev@rics.bwh.harvard.edu

from neutral ones. If the majority of amino acid substituions detected in the study are neutral, then, the power of the method will be low because of the low signal to noise ratio.

It was not known what fraction of missense substitutions among *de novo* mutations and polymorphisms are strongly detrimental, mildly deleterious or effectively neutral. We estimated these values by comparing expected and observed numbers of nonsense, missense and synonymous changes among disease mutations, human SNPs identified by systematic re-sequencing projects and substitutions fixed in the human lineage after divergence from chimpanzee. As expected, fraction of deleterious mutations among common polymorphism was extremely low. However, despite commonly held belief that even among rare missense SNPs most are effectively neutral, our results indicate that the majority of human missense polymorphisms with detected frequency below 1% are, in fact, deleterious. This suggests that allele frequency alone can serve as a strong predictor of functional significance of missense polymorphic variants. We estimated that, on average, each human genome has approximately 600 moderately deleterious missense SNPs associated with selection coefficients in the range of 10^{-2}-10^{-3}.

Our results suggest that association studies aimed at detection of rare missense mutations enrichment might be efficient tool to study genetics of complex diseases and our work serves as a theoretical foundation for this approach[4].

1. Z. Wang et al. (2004) Mutational analysis of the tyrosine phosphatome in colorectal cancers, Science **304**: 1164-1166.
2. J.C. Cohen et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol, *Science* **305**: 869-872.
3. S. Romeo et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL, *Nat Genet.* **39**: 513-516.
4. G.V. Kryukov, L.A. Pennacchio, S. Sunyaev (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies, *Am J Hum Genet.* **80**: 727-739.

CONSTRUCTING PWM FROM UNALIGNED TFBS FOOTPRINTS

I.V. KULAKOVSKY[1], V.J. MAKEEV[2]

Positional Weight Matrix (PWM) is a widely used motif model for transcription factor binding sites (TFBS) at DNA. An aligned set of sequences containing a factor binding site is needed to construct the PWM representing binding motif for this factor. The problem is that sequence lengths of experimental footprints can vary dramatically and there may be sequences with lengths smaller than the actual motif length. To solve this problem one needs to consider genomic sequences containing the footprints; thus footprint sequences become extended with flanks of a selected length. In practice, algorithms that estimate motif length (e.g. [1]) tend to produce uncontrollably long motifs as the length of sequences in the input set increases. Here, we present a method producing multiple local alignment of footprint data which estimates the motif length that is stable when the length of the input sequences increases.

We used SP1 binding site for a model. Footprints for human Sp1 were obtained from TRANSFAC database (release September 2006) and mapped on UCSC build 35 of human genome obtaining 81 chromosome fragments. Chromosome regions made of Sp1 footprints with corresponding flanking chromosome sequences of a selected length were used as an input for

SeSiMCMC [2] Gibbs sampler run with –fp parameter. This allows one to get an alignment, with each sequence overlapping with the initial footprint (TRANSFAC) data. We varied flanking sequences in their length and found that the resulting motif length was slowly increasing with the flank length stabilizing at 15 for flank length 10. The final 15 bp motif was found in 77 of 81 sequences. Surprisingly, the motif found had a consensus of tandem repeat $(GCCCC)_3$ with many substitutions. The motif, the multiple local alignment made by SeSiMCMC, was used to construct a motif model, PWM using procedure described in [3]. PWMs were also constructed from Sp1 site alignments obtained from published data. To evaluate matrix selectivity, for each threshold we calculated the number of the scoring footprints and plotted it versus the probability to observe one or more sites with chosen threshold in a random sequence (the P-value). P-values we calculated with the AhoPro program (http://bioinform.genetika.ru/AhoPro). For a selected P-value and the sequence set PWM with a better selectivity finds

[1] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, ikulakovsky@inbox.ru
[2] Institute of Genetics and Selection of Industrial Microorganisms, FGUP GosNIIGenetika, Moscrow, Russia, makeev@genetika.ru

more motif-containing sequences (the PWM threshold for each matrix is determined by the P-value). We compared PWMs constructed with SeSiMCMC with those generated by classic Gibbs-sampler[4] and MEME[1] with a site length set to 15, as well as with the matrices published in [5] and the matrix obtained from SELEX-experiments [6]. As a control we also used dataset of 157 sequences from TRRD database (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/). These sequences contain Sp1 binding sites for different biological species and also were taken from TRANSFAC database.

All matrices constructed from the sequence set with correctly selected flank length showed a higher selectivity at our and the control datasets compared to matrix from SELEX-experiments and matrices taken from paper [4] over the whole P-value range. Thus, we believe that our method adequately constructs the PWM from the set of relatively short unaligned sequences obtained from footprinting experiments and optimally determines the motif length.

1. T.L.Bailey, C.Elkan (1995) // Machine Learning, 21(1-2):51-80
2. A.V.Favorov et. al. (2005) // Bioinformatics, 21(10):2240-2245.
3. A.P.Lifanov et. al. (2003) // Genome Res., 13-4: 579-588
4. W.Thompson, E.C.Rouchka, C.E. Lawrence (2003) // Nucl. Acids Res. 31:3580-3585
5. W.W.Wasserman, J.W.Fickett (1998) // J. Mol. Biol. 278(1):167-81
6. H.J.Thiesen, C.Bach (1990) // Nucl. Acids Res. 18(11):3203-3209

EXON SKIPPING AND ACTIVATION OF CRYPTIC SITES AS CONSEQUENCES OF SPLICING MUTATIONS

YERBOL Z. KURMAGALIYEV

A large fraction of mutations causing genetic disease disrupt correct splicing of genes. Such mutations may corrupt splice-sites and cis-acting elements, or create de novo splice sites. Disruption of a correct splicing pathway can lead to different changes in the splicing "phenotype": exon skipping, cryptic site activation, and intron retention (the rarest type). Here we compare exons, that demonstrate two most frequent types of the splicing pathway alterations caused by single nucleotide substitutions in the authentic splice sites: exon skipping (S-exons) and cryptic site activation (C-exons).

Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny per. 19, Moscow, 127994, Russia, kurmangali@mail.ru

The set of C-exons was obtained from previous works [1,2] and set of S-exons was collected by search in OMIM and PubMed.

An average S-exon is significantly shorter than a C-exon. When we compared strengths of partner splice sites (neighbors of authentic splice site affected by mutations), the intron partners did not show any difference in average site scores, whereas the exon partners were somewhat stronger in C-type exons, but this difference was not significant. Further, we measured the density of putative cis-acting elements: exonic splicing enhancers (ESE) and exonic splicing silencers (ESS). The density was defined as the number of candidate of ESE and ESS per base pare. Cis-acting elements were predicted using the published tools ESEfinder, RESCUE-ESE, and PESX. The average density of ESEs predicted by both ESEfinder and PESX was significantly higher in C-exons, and accordingly average density of ESSs predicted by PESX was significantly higher in S-exons. This is consistent with the theory that an exon containing many ESE and few ESS is less likely to be skipped, and is spliced using a cryptic site. The densities of ESEs predicted by Rescue-ESE did not show any difference between S- and C-exons.

Thus least two parameters of exons with corrupted authentic splice sites may a play role in the choice between skipping and cryptic site activation pathways: exon length and the density of cis-acting elements. Existence of a strong partner site via exon most likely also forces exon not to be skipped, as low statistical significance may be caused by the small sizes of datasets used in this analysis.

This is joint work with M.S.Gelfand.

1. X.Roca, R.Sachidanandam,A.R.Krainer (2003) Intrinsic differences between authentic and cryptic 5'-splice sites, *Nucleic Acids Res,* **31:**6321-6333.
2. I.Vorechovsky (2006) Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization, *Nucleic Acids Res,* **34-**4630-4641.

A SEARCH FOR THE GENE *FRUITLESS* IN ANTS

TATIANA KUZMENKO[1], MIKHAIL SKOBLOV[1],
SERGEY NUZHDIN[2], ANCHA BARANOVA[13]

The gene *fruitless* is a master-regulator that directs brain development in male- or female-specific ways by alternative splicing. Its function appears to be well-conserved among all insects, [1] but now noticeably characterized only in the fruit fly (*D. melanogaster*). [2] In this insect different isoforms of *fruitless* are expressed in groups of neurons responsible for certain patterns of **male**-specific **aggressive** behavior. Knocking *fruitless* out results in ablation of these behavior patterns without affecting other pathways. [3]

We aimed to determine whether *fruitless* performs the same function in cases where aggressive behavior not linked to sex. Social insects, specifically ants, are well suited as a model, as worker **females** demonstrate highly specific patterns of aggressive behavior undergoing well-described changes through the adult developmental stages. However, orthological *fruitless* gene has not been described in ants and full ant genomic sequences doesn't accessible yet. Therefore, we started this study from *fruitless* cloning.

We downloaded from databases all known partial or complete sequences of orthological *fruitless* mRNA among insects *(Apis melifera, Tribolium castaneum, Aedes aegypti, Anopheles gambiae, Bombyx mori, Lutzomyia longipalpis, Drosofila melanogaster)* and aligned them in ClustalX to determine the most conserved region, which turned out to be a BTB (Bric-a-brac, Tramtrack, Broad-complex) domain. However, designed for *fruitless* BTB domain primers had amplified *lola* – a *fruitless* homolog, which also contains conserved BTB domain. To increase specificity of PCR, we extended multiple alignments, using BTB domain's sequences of:

1) *fruitless* of the **honeybee** (*Apis mellifera*) – the closest relative of ants among insects with completed genomic sequence. This sequence was used as a basis for primer design;

2) *fruitless* of **flour beetle** (*Tribolium castaneum*), that is characterized by highest similarity to honeybee's *fruitless* according to distance trees composed in ClustalX. We used this sequence to determine the local conserved sequences inside BTB domain for making primers specific to these conserved sites;

[1] Research Center for Medical Genetics, RAMS, Moskvorechie Str., 1, Moscow, Russia, kuzmenkotv@gmail.com
[2] Department of Evolution and Ecology, University of California in Davis, UCD, Davis CA, USA svnuzhdin@ucdavis.edu
[3] Molecular Biology and Microbiology Department, George Mason University, Fairfax, Va., USA

3) *lola* of **ant** (*Camponotus maritimus*) – during primer design we excluded regions of high similarity between *fruitless* and *lola* to ensure *fruitless* amplification.

Using Oligo software we designed specific primers which must to amplify only *fruitless* region. We used genomic DNA and cDNA from the ant *C. maritimus* as templates for PCR. From both templates we obtained 261 bp amplicons and sequenced the genomic fragment. Its Blast 2-based alignment with honeybee genome and *lola* of ant (seed sequence) showed **89%** similarity to honeybee *fruitless* and very low similarity to *lola*. The next set of primers consisted of unique forward primer located within the amplified sequence and three reverse primers located within aligned sequences of *fruitless* of the honeybee and the flour beetle. These primers were spaced out from each other by ~150 bp in the 3'-direction of the mRNA transcript. All of them except the most distal reverse primer worked well on cDNA template. We sequenced largest of obtained amplicons (~470bp), aligned it with seed sequence described above and combined assembly sequence of 616 bp. Blast 2 alignment of the extended sequence with the honeybee's *fruitless* confirmed its high homology (86%). We suggested that our amplified fragment is likely the region of the *fruitless* gene of *C. maritimus*, based on the high similarity of the amplified fragment to other known *fruitless* genes.

We are planning to continue this study by extending the described fragment of *C. maritimus fruitless* gene in 5' and 3' direction through RACE-PCR. After getting full sequence of *fruitless* mRNA and genomic DNA, we will perform functional analysis of this gene in ants and investigate how expression of *fruitless* correlates with the aggressive behavior of such social insects like ants.

Authors acknowledge Kopp laboratory at the Department of Evolution and Ecology, UCD for the help in getting cDNA of *C. maritimus*.

1. D.A. Gailey, et al. (2005) Functional conservation of the fruitless male sex-determination gene across 250 Myr of insect evolution, *J Mol Biol Evol*, **23**(3):633-43.
2. D Yamamoto, Y Nakano, (1999) Sexual behavior mutants revisited: molecular and cellular basis of Drosophila mating, *J Cell Mol Life Sci.*, **56**(7-8):634-46
3. S.J. Certel, et al. (2007) Modulation of Drosophila male behavioral choice, *J Proc Natl Acad Sci USA*, **104**(11):4706-11.

FITNESS, CONSERVATION, AND TURNOVER OF TRANSCRIPTION FACTOR BINDING SITES

MICHAEL LAESSIG[1]

The evolution of regulation occurs under substantial selection.

For regulatory elements in yeast, we obtain strongly nonlinear fitness landscapes, which depend on their binding energy as molecular phenotype. We infer the evolutionary conservation and turnover rates for these elements.

Fitness itself is often time-dependent, as shown by a recent study of Drosophila genomes. These fitness ``seascapes'' vary at nearly the rate of neutral evolution and act as driving force of genome-wide adaptive sequence change. Consequences for the conceptual picture of molecular evolution are discussed.

STRUCTURE PREDICTION OF A-HELICAL MEMBRANE PROTEINS: THE NA^+/H^+ EXCHANGER 1 (NHE1) OF THE HEART AS AN EXAMPLE

MEYTAL LANDAU[2], KATIA HERZ,[3] ETANA PADAN,[3] AND NIR BEN-TAL[2]

Membrane proteins comprise 20-30% of the genome, but due to experimental difficulties, they represent less than 1% of the Protein Data Bank (PDB). The dearth of membrane protein structures makes computational prediction a potentially important means of obtaining novel structures. Recent advances in computational methods have been combined with experimental data to constrain the modeling of 3-dimensional structures. These advances and their application to predict the structure of NHE1 will be presented.

Eukaryotic Na^+/H^+ exchangers are membrane proteins that are vital for cellular homeostasis and play key roles in pathological conditions such as cancer and heart diseases. Using the crystal structure of the Na^+/H^+ antiporter from *Escherichia coli* (EcNhaA) as a template, we predicted the 3D structure of human NHE1. Modeling was particularly challenging because of the extremely low sequence identity between these proteins, but the model-structure is supported by evolutionary conservation analysis and empirical data. It also revealed the location of the binding site of NHE inhibitors; which we validated by conducting mutagenesis studies with EcNhaA and its specific inhibitor 2-

[1] Institut fuer Theoretische Physik, Universitaet zu Koeln
[2] Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel-Aviv University, 69978 Tel-Aviv, Israel
[3] Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

aminoperimidine. The model structure features a cluster of titratable residues that are evolutionarily conserved and are located in a conserved region in the center of the membrane; we suggest that they are involved in the cation binding and translocation. We also suggest a hypothetical alternating-access mechanism that involves conformational changes.

1. S.J. Fleishman, V.M..Unger and N. Ben-Tal. (2006) Transmembrane protein structures without x-rays. Trends in Biochem. Sci. 31: 106-113.
2. S.J. Fleishman and N. Ben-Tal. (2006). Progress in structure prediction of alpha-helical membrane proteins. Curr. Op. Struct. Biol. 16: 496-504.

VISUAL GENOMICS: GIGANTIC PALINDROME DISINTEGRATION AS A COMMON EVENT OF GENOMES EVOLUTION

S.A. LARIONOV[1], A.YU. LOSKUTOV[1], E.V. RYADCHENKO[1], M.S. POPTSOVA[2], I.A. ZAKHAROV[3]

We discovered that chromosomes of many species, from bacteria to human, have in their origins a gigantic palindrome (palindromes) that disintegrated during evolution process by inversion, duplications, mutaton drift and other kind of rearrangements [1,2] . This type of palindromes have length from megabases to several tens megabases (it is possible more long) and as a result of large scale rearrangements and drift have only context correlation nature [3,4]. This data obtained thanks to a highly visual character of 2D walk method, that allow us to see full chromosomes with several hundred megabases long as unique portret [5,6]. We analized this palindromes in detail in different scale. We use 2D walk method as a interface of genomes databases and annotations for detection and analisys functional-structural elements of chromosomes and sequences local properties. We considered this data in significant examples series and suppose that this results will be usefull for wide range of problems: from protein clusters prediction [2,7], and metabolical network organization [1] to a evolutionary modeling [4].

We would like thank S.Rybalko for his assistance in 2D DNA walk modeling. We also thanks A.Khokhlov, B.Dujon, H.Renauld , P.Schuster, A.Bairoch, L.Hurst, A.Valencia, J.Skolnik and I.Friedberg for useful comments and interest.

[1] Moscow State University, Physics Faculty, Moscow, Russia, e-mail: serglarionov@yandex.ru, loskutov@chaos.phys.msu.ru, evgeny@inbox.ru
[2] University of Connecticut, USA e-mail: mariapoptsova@mail.ru
[3] Vavilov Institute of General Genetics, Russan Academy of Sciences, Moscow, Russia, e-mail: iaz34@mail.ru

1. Larionov, S.A., Loskutov, A. & Ryadchenko E.V. (2007) (In Preparation).
2. Larionov, S.A., Loskutov, A. , Poptsova M.S., Ryadchenko E.V., Zakharov I.A. (2007) (In Preparation)
3. Larionov, S.A., Loskutov, A., Ryadchenko, E.V. (2005) Lessons of large scale comparisons from 2D DNA walk. Proc. of ESF Research Conf. : Comparative Genomics of Eukaryotic Microorganisms, Sant Feliu de Guixols, Spain, 12 - 17 November . (Oral presentation and poster session)
4. Larionov, S.A., Loskutov, A., Ryadchenko, E.V. & Rybalko, S. (2006) Gigantic palindrome diffusion and certain features of genomes evolution. Proc. of the Int. Symposium on Evolution of Biomolecular Structure, University of Vienna, Austria, 25-27 May . (Oral presentation and poster session)
5. Larionov, S.A., Loskutov, A. & Ryadchenko, E.V. (2005) Genome as a two-dimensional walk. Doklady Russian Academy of Sci., Physics, 14, 634-638,
6. Larionov, S.A., Loskutov, A. & Ryadchenko, E.V. (2005) What can we learn from 2D DNA walk?
7. ISMB-2005, Special Interest Group “Function prediction”. Detroit, USA, 24 June. (Poster session)
8. Larionov, S.A., Loskutov, A., Ryadchenko, E.V. & Rybalko, S. (2006) Visual genomics methods: gigantic palindromes and protein clusters prediction. Proc. of the Int Symp.: In-silico Analysis of Proteins: Celebrating the 20th Anniversary of Swiss-Prot, Fortaleza, Brazil, July 30 - Aug. 04 (Short oral presentation and poster session)

"EVOLUTIONARY CONSTRUCTOR" – METHODIC FOR SIMULATION OF COEVOLUTION IN COMMUNITY

S.A. LASHIN, V.V. SUSLOV, N.A. KOLCHANOV, YU.G. MATUSHKIN

An organism can not live beyond ecosystems, which function is of rehabilitation of comfortable conditions for community members on the base of information coded in organisms' genomes [Razumovsky, 1981]. Thus the process of taxa evolution can not be reduced to genome evolution. The realistic models of speciation should consider the interference of tempos and regimes of evolution at least at genomic and ecosystem levels. The close interaction between the both levels demonstrably appears in high-closed trophically linked communities of microorganisms: the change of one species (mutation, horizontal transfer, fluctuations of population size etc) may cause the whole system rearrangement [Zavarzin, 2003].

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, mat@bionet.nsc.ru

There are two approaches of evolution modeling. The flexible portrait modeling requires a lot of details about every object of a system which may lead to computational complexity growth taking into account trophic interactions between objects [Garey et al., 1982]. The generalized modeling does not allow simulation of system rearrangements. We suggest a novel method of evolutionary modeling and the "Evolutionary constructor" (EC) package, implementing this method. This method of stepwise simulation modeling handles both generalized and portrait models, allowing shift from "generalization" to "portraitness" and vice versa.

The life and evolution of networks of haploid unicellular organisms' populations in environment is modeled. The organisms consume substrates from and secrete products to environment. Products of some organisms may be consumed by the other ones as substrates (specific substrates). Also there are nonspecific substrates in environment, which concentration is provided by flow. Nonspecific substrates are necessary for the life of every organism in network. In order to use the substrates for reproduction and products synthesis the individual should "uptake" the substrates from environment. The concurrence for substrate between organisms (as inter- so intra- populational) is realized on the "uptake" stage. The efficiency of consumption and synthesis of substrates are considered as selective traits and regulated by the rate constant of corresponding process (which depend upon the corresponding "gene"). Then the "genotype" of individual is the three groups of constants: (c_i) – efficiency of specific substrate consumption, (d_i) – products synthesis rate, (r_i) – efficiency of nonspecific substrates consumption. The individuals having identical constants sets (c_i), (d_i) and (r_i) form **monomorphic population (MP).** In order to calculate the growth if it's size P the various equations were used (the example is shown below):

$$\Delta P = F_1(\vec{S}, \vec{C}, P) = \sqrt{r_0 n_0(P) \cdot \sum_{i=1}^{N} c_i s_i(P)} - k_{death} \cdot P^2 \quad ,$$

where $\vec{S}$ is the vector of the specific substrate amounts *uptaked* by the population proportional to the population size; n_o is the amount of unique nonspecific substrate; k_{death} is the mortality constant of the population. MP may consist of the only individual. The set of MPs may form polymorph population (PP). The PP is described by genetic spectrum – distribution of "gene" occurrence frequencies in population - in practice it is the distribution of a trait (efficiency of consumption/synthesis) occurrence frequencies in some limits. In these terms the PP is characterized by the set of genetic spectrums. The muta-

tion is the change of genetic spectrum profile. The notion of trait's threshold value (TTV) is also suggested: the trait assumed to be absent if its value less then TTV; partial transition of the spectrum through the TTV means the PP segregation on two subpopulations. TTV provides a convenient way to classify the network onto "species". Also this concept is used in modeling of horizontal transfer of gene material. The genetic spectrums arithmetic is developed to calculate population growth. It allows the "splitting" of PP onto several MPs (further the MP's growth is calculated) together with "merging" changed MPs in one PP again.

In spite of simplicity of EC prototype, some interesting biological results were obtained. The fitness growth of one population in the closed ring of MPs, inhibiting their neighbors (in ring) lead to disturbances of system functioning (oscillatory and chaotic-like regimes for rings of odd number of MPs and clusterization for rings of even number of MPs) which implied the system in dependence on environmental fluctuations. Modeling horizontal transfer of genes in symbiotic ring the population-acceptor and its nearest symbiont were found to get selective advantage (both in short and long period). The population-donor gets selective advantage only in long-term perspective.

The work is supported by the grants: RFBR 05-04-49068, 05-07-90274, 06-04-49556, 05-07-98011, SB RAS integrative project № 34, RAS Presidium program of molecular and cell biology "Biosphere origin and evolution".

1. S.M. Razumovsky (1981). On the Dynamics of Biogeocenoses. Moscow, Nauka (in Russ.).
2. G.A. Zavarzin (2003). Lectures on natural resource microbiology. Moscow, Nauka. (in Russ.).
3. Garey, M.R., Johnson D.S., Freeman, W.H. (1982) Computers and intrastability: A guide to the theory of NP-completeness. Moscow, Mir, (in Russ.).

COMPUTER SYSTEM FOR ANALYSIS AND MODELING 2D PLANT TISSUE

V.V. LAVREHA[1], S.V. NIKOLAEV[1], N.A. KOLCHANOV[1], A.V.PENENKO[2]

At present development of programs for analysis and modeling plant tissue is a very important problem in area of developmental biology. At the tissue level growth and development are determined by cellular biomechanics, molecule transportation, and cellular responses to their microenvironments. Biologically motivated models of such processes, when incorporated into models of tissue growth and development, would help to study the roles of these processes for growth and development tissue (Rudge, 2005; Shraiman, 2005). Obvious account of geometric characteristics of biological object allows to formulate quantitative tasks relatively mechanisms of process management of growth and development, and to check it in experiment with using methods of image getting and analysis.

Morphogenesis on the scales from cells - tissues - organs and up to the whole organism depends on differential cell growth and division. It is shown that deformation tensions that arise in different parts of growing biological tissue have dramatic effect on the all biological processes up to differential gene expression (Nelson, et al., 2005). The gene expression controls the cell functions. Thus biomechanics and biological functions form a regulatory loop. While developmental processes including morphogenesis in animals involve cell movement, it dose not occurs in plants, so the plant tissue development is, in a sense, easier to model and simulate. The growth of plant cells and tissues is shown to be dependent on water potential in the cells. The growing plant cells are enveloped with primary cell wall, which can yield under intracellular pressure (Cosgrove, 1986).

In result we developed the program system for simulation of “planar” plant tissue growth and development. The system includes graphical user interface to prepare an initial configuration of system to be modeled and to perform simulations.

The graphical user interface is an organizing component. It is designed to provide convenient interaction with the program. Its purpose is to join the stages of creation and development of cellular ensemble, and also to build and execute various scripts of cellular ensemble dynamics. Its functions:

interaction with user for generating geometry of cells ensemble,

[1] Institute of Cytology and Genetics SB RAS, pr. Lavrentieva 10, Novosibirsk, 630090, Russia, e-mail: vvl@bionet.nsc.ru
[2] Institute of Computational Mathematics and Mathematical Geophysics SB RAS, pr. Lavrentieva 6, Novosibirsk, 630090, Russia

redacting geometry of cells ensemble,
marking cells ensemble,
loading and saving files with cells ensembles,
selecting model of simulating cells dynamics and defining models parameters,
graphic presents results of modeling.

About views on the system. Biological aspects: orientation on plant tissue (no migration of cells); mechanic properties in plant cells are defined by cell walls. Functional aspects: graphic user interface; library of models

What we did before developing: (1) the components of the system was determed, and its structures were worked out; (2) data flows were described; (3) working scenarios were developed; (4) data structure was worked out.

Also during biological experiments on cells and tissues often images are the results. The images could be given as photos, different cuts and etc. These images contain a lot of information of different types: geometry and substances mixture (morphogenes, proteins, water) of cells. During even one experiment we can get huge number of images within a lot of information, that's why we need to use computer system to analyze it. And it is very important task, because after getting information from images we can say more directly about nature of processes going during experiment and accuracy of model constructing (Nelson, 2005).

1. *D.J. Cosgrove,* (1986) Biophysical control of plant cell growth, *Annu. Rev. Plant Physiol.,* **37**:377-405.
2. *C.M. Nelson, et al.* (2005) Emergent patterns of growth controlled by multicellular form and mechanics, *PNAS*, vol. **102**: 11594–11599
3. *T. Rudge, J. Haseloff* (2005) A computational model of cellular morphogenesis in plants. In: *ECAL*, M. Capcarrere et al. (Eds) LNAI 3630, pp. 78-87 (Springer-Verlag Berlin Heidelberg)
4. *B.L. Shraiman* (2005) Mechanical feedback as a possible regulator of tissue growth, *PNAS*, vol. **102**: 3318–3323

SELF-ORGANIZED BIOCHEMICAL DYNAMICS IN MIGRATING IMMUNE CELLS: A COMPUTATIONAL BIOLOGY APPROACH

D. LEBIEDZ

Cells react and adapt to changes in their environment by relaying information through signal transduction pathways. The spatiotemporal dynamics of signaling routes itself often encodes transduced information, and mathematical modeling and computer simulation techniques integrating quantitative in vivo experimental data are crucial on the way towards an elucidation of the enormous complexity of cellular signaling mechanisms.

Neutrophils are cells of the human immune system and exhibit central functions in the first line of defense against inflammation. Their major role is the detection, internalization (phagocytosis) and digestion of invading pathogens, e.g. microorganisms like bacteria.

Recently, biochemical NAD(P)H- [1] and Calcium-waves [2] have been detected in neutrophils by means of high-speed fluorescence imaging. These dissipative patterns seem to be related to polarization, direction finding and chemotactic migration of neutrophils as well as target-oriented secretion of reactive oxygen intermediates helping to destroy pathogens.

In a combined experimental / spatiotemporal modeling approach [3] we try to elucidate the underlying molecular biology and relate its dynamics to cell physiology.

1. H. R. Petty, A. L. Kindzelskii (2001) Dissipative metabolic patterns respond during neutrophil transmembrane signaling, *Proc. Natl. Acad. Sci. U S A.,* **98**: 3145-3149
2. A. L. Kindzelskii, H. R. Petty (2003) Intracellular calcium waves accompany neutrophil polarization, formylmethionylleucylphenylalanine stimulation, and phagocytosis: a high speed microscopy study, *J Immunol.* **170**: 64-72.
3. O. Slaby, S. Sager, O. S. Shaik, U. Kummer, D. Lebiedz (2007) Optimal control of self-organized dynamics in cellular signal transduction, *Math. Comput. Mod. Dyn. Sys.* (in press).

University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany), lebiedz@iwr.uni-heidelberg.de

A GRAPH-BASED APPROXIMATE STRING MATCHING METHOD FOR PREDICTING THE PLANTED (*L*,*D*)−MOTIF PROBLEM

LEE, CHAO-MING, WANG JUYING, LEE, HAHN-MING

The (l,d)-planted motif problem is a challenge problem for accessing the real transcription factor binding sites. Here we proposed a new computational scheme agrep, nr-grep and EMBM(Exact match by mutation) to assess the capability of predicting the planted motifs, eg.(10,2), (11,2), (12,3), (13,3), (14,4), (15,4), (17,5) challenge problems. We showed that the schemes can predict the planted motif through 100% percent comparing with the traditional algorithms like GibbsDNA, WINNOWER, SP-STAR, PROJECTION and Styczynski.

A GRAPH-BASED APPROXIMATE STRING MATCHING METHOD FOR PREDICTING TRANSCRIPTION FACTOR BINDING SITES

LEE, CHAO-MING, WANG JUYING, LEE, HAHN-MING

Here we proposed a new computational scheme using suffix array and approximate string matching techniques to assess the capability of predicting transcription factor binding sites. We showed that the proposed algorithm, which obtained better results than previously reported algorithms.

Keyword: motif, transcription factor binding sites, approximate string match, suffix array, scoring function.

Department of Computer Science and Information engineering, National Taiwan of Science and Technology, Taiwan

NOTCH SIGNALLING AND THE SOMITE SEGMENTATION CLOCK: MATHEMATICAL MODELLING AND EXPERIMENTAL VALIDATION

JULIAN LEWIS[1], FRANÇOIS GIUDICELLI[2], ERTUGRUL OZBUDAK[3]

The Notch signalling pathway is one of a handful of cell-cell communication mechanisms that are critical for almost every aspect of the development of multicellular animals: it provides a direct route by which a cell can influence gene expression in neighbouring cells with which it is in contact. Simple feedback loops in the Notch signalling pathway can give rise to many types of pattern in space and time, according to the logic and the quantitative parameters of the feedback circuitry. The inner ear, nervous system, vasculature, gut lining, and somitic mesoderm provide examples.

The importance of Notch signalling in temporal patterning is exemplified in the somite segmentation clock. This system offers an unusual opportunity to analyse one of the least-understood aspects of multicellular development: the control of developmental timing. The segmentation clock is a transcriptional oscillator that operates in the presomitic mesoderm (PSM) at the tail end of the vertebrate embryo and governs the spacing of somites - the embryonic segments of the vertebral column and musculature. In each oscillator cycle, one additional somite is delimited and emerges as a block of cells from the PSM. In the zebrafish, the cycle time is 30 minutes, and the oscillating genes include the Notch target genes *her1* and *her7* and the Notch ligand gene *deltaC*. The mechanism of oscillation can be plausibly explained in terms of a simple negative feedback loop through which the *her7* gene product directly inhibits its own expression, with cell-cell communication via the Notch pathway serving to keep the oscillations of neighbouring cells synchronized. Mathematical analysis of this simple model indicates that the oscillation period should be mainly determined by the transcriptional delay - the time from start to completion of the synthesis of a *her7* mRNA molecule. Failure of Notch signalling leads to an observed loss of synchrony that can be interpreted in terms of genetic noise arising from the association/dissocation kinetics of the interaction of Her7 protein with the regulatory DNA of the *her7* gene. Thanks to the special space-time ge-

[1] Vertebrate Development Laboratory, Cancer Research UK London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3PX, UK, julian.lewis@cancer.org.uk
[2] Present address: Laboratoire de Biologie du Développement, UMR CNRS 7622, Université Pierre et Marie Curie, 9 quai Saint Bernard, 75005 Paris, France
[3] Present address: Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA

ometry of the system, we can measure the delays, lifetimes, noise and other key parameters; and using heat-shock transgenics, we can also check the logic of the control circuitry. The system illustrates how even the simplest gene regulatory circuits can show rich and surprising behaviour, which can only be properly understood through a combination of mathematical modelling and quantitative experimentation.

1. Jiang, Y.-J., Smithers, L., Aerne, B., Haddon, C., Ish-Horowicz, D., and Lewis, J. (2000) Notch signalling and the synchronisation of the somite segmentation clock. Nature, 408:475-479.
2. Lewis, J. (2003). Autoinhibition with transcriptional delay. A simple mechanism for the zebrafish somitogenesis oscillator. Curr. Biol., 13:1398-1408.
3. Giudicelli, F. and Lewis, J. (2004) The vertebrate segmentation clock. Curr. Opin. Genet. Dev., 14:407-414.
4. Giudicelli, F., Ozbudak, E.M., Wright, G.J. and Lewis, J. (2007) Setting the tempo in development: an investigation of the zebrafish somite clock mechanism. PLoS Biology, in press.
5. Lewis, J. and Ozbudak, E.M. (2007) Deciphering the somite segmentation clock: Beyond mutants and morphants. Dev. Dyn., in press.

STATISTICS OF CLOSELY RELATED STRAIN PROTEOMES REVEALED STRIKING DIFFERENCES IN THEIR COMPOSITION

ELENA LITVINOVA, ALEKSANDRA B. RAKHMANINOVA

The number of completely sequenced prokaryotic genomes grows at a very fast rate. This set also contains genomes of closely related strains. Virtually all recent studies of closely related strains made an emphasis on tracing differences between pathogenic and non-pathogenic strains. Our aim was to study all possible strain diversities.

We have developed an integrated database PHOG-FUN. It merges data for phylogenetic orthologous rows and their functional annotation for a group of thirteen *Escherichia coli* and *Shigella* spp. complete genomes. The phylogenetic orthologous rows for this group were constructed by the PHOG-BLAST program developed in our group [1]. Functional annotations were obtained from two databases, GOA and GenProtEC. The former provides high-quality

Department of Bioengeneering and Bioinformatics, M.V.Lomonosov Moscow State University, Moscow, Russia, alenaaa@gmail.com

Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB) [2]. The latter is an annotation source for the laboratory strain *E.coli* K-12 [3]. We have used PHOG-FUN database to obtain biologically meaningful information about genome repertoires of closely related strains.

The total number of orthologous rows for the studied group of genomes is 8149. The average identity of every protein pair within one row common to all strains (the common core) is about 98%. Functional annotations were assigned to 60% of orthologous rows, with common core proteins being well annotated; 93% of them had an assigned function. We found that the distribution of row occupancy is nonuniform. At that 1526 (18%) orthologous rows were strain-specific and only 2121 (30%) orthologous rows contained genes from each strain.

Recent studies showed that *Shigella* might have been derived from multiple independent *Escherichia coli* strains[4]. We reconstructed NJ-trees based on two types of evolutionary distance estimates. The first one is pairwise difference in the number of common phylogenetic rows. The tree topology completely supports the current taxonomy tree obtained from NCBI, but it depends on the distribution of IS elements and phage-related insertions. The second tree based on multiple alignment concatenate of 2039 protein sequences belonging to the common core. The common core proteins were taken to eliminate the influence of lateral transfer events on the tree topology. It showed closer relationship between *S.flexneri* and enterohemorrhagic *E.coli,* but also contradicted the theory of independent *Shigella* origins. Thus the phylogenetic analysis showed that *E.coli* and *Shigella* are distinct taxa.

The performed analysis of several functional groups belonging to the top level of the MultiFun functional hierarchy showed that extensive genomic rearrangements perturbing metabolic pathways could be observed on pretty short evolutionary distances. Also we performed detailed analysis of orthologous rows containing genes encoding cellular components according to the GO classification. We detected striking differences between strains regarding the metabolism fatty acids and amines. The *Shigella* species were shown to lose various multisubunit enzyme complexes such as parts of the beta-galactosidase and citrate lyase complexes. The considered strain proteomes contained different ornithine carbamoyltransferase complex subunits. At the same time we found out that loss of some essential components could be explained by misannotations. In general our investigation revealed the fact that closely related strains could differ in essential components of the metabolism. Finally, the developed database was shown to be useful for detection of evolutionary events such as

recent gene duplications, domain shuffling, fusions and other genomic rearrangements.

We are grateful to I.V.Merkeev and A.A.Mironov for the provided PHOG-BLAST program.

This is joint work with M.S.Gelfand.

1. Merkeev IV, Novichkov PS, Mironov AA. (2006) PHOG: a database of supergenomes built from proteome complements.BMC Evol Biol.; 6:52.
2. Camon E, Barrell D, Lee V, Dimmer E, Apweiler R. (2004) The Gene Ontology Annotation (GOA) Database: an integrated resource of GO annotations to the UniProt Knowledgebase In Silico Biol;4(1):5-6.
3. Serres MH, Goswami S, Riley M. (2004) GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. Nucleic Acids Res.;32.
4. Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, Zhu Y, Yu J, Jin Q. (2007) Revisiting the molecular evolutionary history of Shigella spp. J Mol Evol.,64(1), 71-9.

DESIGN, DEVELOPMENT AND USE OF A DATA MANAGEMENT AND VISUALIZATION TOOL FOR OLIGONUCLEOTIDE PROBES

G. H. LÓPEZ-CAMPOS, F. MARTÍN-SÁNCHEZ

Nowadays there exists an increasing number of experimental techniques that are based in the use of oligonucleotide probes. An important aspect to deal with when using these techniques consists of avoiding or minimizing any cross reactivity of the probes with other sequences. Some of these techniques such as Microarrays or Reverse Line Blots play an important role and are currently applied in clinical microbiology and diagnosis. Therefore in this work a new tool for the management and annotation of oligonucleotide probes and primers in a clinical microbiology environment is presented. The system includes a module for probe information management and another one for the visualization of cross reactivity of the probes with a set of sequences of interest.

The system is based on a common database management system such as MS Access and embedded PERL code, integrated with other programs (BLAST [1] and BUSSUB [2]). With this structure, the system allows the user to annotate, manage and visualize probes and their interactions with the set of sequences of

Institute of Health “Carlos III”, Ctra. Majadahonda-Pozuelo Km2. Majadahonda, 28220. Spain, glopez@isciii.es

interest. The system has been designed to use the BLAST program to perform the searches of the probes and primers against one or several local databases defined by the user. These databases with the sequences of interest for the project can be downloaded from Internet in BLAST format or can be generated by the system from the files containing the sequences in FASTA format. The system also includes BUSSUB, an amplicon retrieval tool. The use of these tools is complementary to that one of BLAST. BUSSUB allows the user to include degenerate primers for the searches as well as to define stringency parameters such as number of possible mismatches in the recognition of the primers by the polymerase. The results from this tool also provides the amplicon sequence.

Once the searches have been performed the results are imported into the system for their storage and visualization. The system presents two different visualization modes depending on whether the visualization is from primers results or probes results. In the case of the primer visualization mode, the system offers the basic information of the primers as well as the information of the generated amplicon and from which sequences is generated.

In the case of oligonucleotide probes, the visualization tool graphically represents the probes and the organisms recognized by those probes. The system uses a transformation function to interpret BLAST results and gives a biological meaning to the differences between a gap or a mismatch in the recognition of a probe.

The system can be also integrated with AMANDA [3], an in-house built MIAME [4] compliant database for microarray experiment data management. This database includes LIMS capabilities for probe management.

Final result of this work is a probe and primer information management system for clinical microbiology laboratories and other similar environments. The system facilitates the annotation, analysis, visualization and management of oligonucleotide probes and primers. This tool can be applied to manage the information needed by different techniques such as RLBs, Real-time PCR assays or microarrays. In the latest case the tool can be integrated within a microarray experiment data management system giving a more powerful tool for the analysis of custom-made microarrays.

1. S.F.Altschul et al (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410
2. J.P. Sánchez et al (2004) BUSSUB: an virtual amplicon retrieval software. Poster ISMB 2004.
3. G.H. Lopez-Campos et al (2006) Analysis and management of HIV peptide microarray experiments. Methods Inf Med., 45(2):158-62

4. A.Brazma et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, Nat Genet., 29(4): 365-71

STRUCTURAL SIMILARITY ENHANCES INTERACTION PROPENSITY OF PROTEINS

DIMA LUKATSKY, BORIS SHAKHNOVICH, JULIAN MINTSERIS, KONSTANTIN ZELDOVICH, EUGENE I. SHAKHNOVICH

Several independent analyses of accumulating high-throughput and specific data on protein-protein interactions revealed a general statistical bias for homodimeric complexes. It was also shown experimentally that the sequence similarity is a major factor in enhancing the propensity of proteins to aggregate. The physical or evolutionary reasons for these striking observations remain unexplained.

We study statistical properties of interacting protein-like surfaces and predict two strong, related effects: (i) Statistically enhanced self-attraction of proteins; (ii) Statistically enhanced attraction of proteins with similar structures. The effects originate in the fact that the probability to find a pattern self-match between two *identical*, even randomly organized interacting protein surfaces is always higher compared with the probability for a pattern match between two *different*, promiscuous protein surfaces. This theoretical finding explains statistical prevalence of homodimers in protein-protein interaction networks reported earlier. Further, our findings are confirmed by the analysis of curated database of protein complexes that showed highly statistically significant overrepresentation of dimers formed by structurally similar proteins with highly divergent sequences ("superfamily heterodimers"). We predict that significant fraction of heterodimers evolved from homodimers with the negative design evolutionary pressure applied against promiscuous homodimer formation. This is achieved through the formation of highly specific contacts formed by charged residues as demonstrated both in model and real superfamily heterodimers.

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA

A GENOME-WIDE HUMAN-MOUSE EXPRESSION ALIGNMENT

MARTA ŁUKSZA[12], JOHANNES BERG[23], MICHAEL LAESSIG[23]

We present a method for genome-wide comparative cross-species analysis of expression data, which is based on an efficient gene alignment algorithm. A key component of our formalism is the correct normalization of (logarithmic) expression levels, which results in a well-defined statistics of the data sets. In particular, we construct a consistent overlap map between samples (e.g., experimental conditions or tissues) both within and across species, which accounts for statistical redundancies in the dataset. Given this map, we construct a gene alignment designed to display functional relationships between genes of two species.

The log-likelihood scoring function of this alignment is based on expression pattern overlap and sequence orthology, similarly to the statistical scoring introduced previously for graph alignment of co-expression networks [1]. This function especially accounts for tissue specific expression statistics and for mutual dependencies between tissues. It can be approximated by a bilinear form in the tissue-specific expression differences. This simple form of the scoring function allows the efficient analysis of the genome-wide datasets and leads to computational tractability of the problem.

The high-scoring significant gene mappings result in the alignment which is a many-to-many map and is thus capable of describing evolutionary changes such as gene duplications, recruitment of new genes into pathways, and non-orthologous gene displacements. It affords an efficient optimization algorithm involving the iterative update of both the sample overlap map and the gene alignment.

We apply the method to a genome-wide analysis of mouse and human expression data [2]. We discuss evolutionary conservation of functional gene clusters as well as examples of functional innovation.

1. J. Berg & M. Lassig (2006) Cross-species analysis of biological networks by Bayesian alignment, PNAS, 102: 10967-10972.
2. A.I.Su et al. (2004) A gene atlas of the human and mouse protein-encoding transcriptomes, PNAS, 101: 6062-6067.

[1] Institut für Biologie, Humboldt-Universität, Invalidenstraße 42, 10115 Berlin, Germany, mluksza@gmail.com
[2] Kavli Institute for Theoretical Physics, University of California, Santa Barbara, USA, berg@thp.uni-koeln.de
[3] Institut für Theoretische Physik, Universität zu Köln,Zülpicherstr. 77, 50937 Köln, Germany, lassig@thp.uni-koeln.de

BIBLIOMETRICS OF BIOINFORMATICS

A.V. LYUBETSKAYA

A bibliometric study of a bioinformatic database was performed in order to show how bioinformatics has developed since its emergence and what of its areas are currently the most popular.

The initial database consisted of papers published not later than January 1, 2007 and obtained from the PubMed database. These papers were selected using the following criteria. Firstly, a "bioinformatic" journal was defined as a journal that contains the word "bioinformatics" in its name. Secondly, a "bioinformatist" is an author who published at least one paper in a bioinformatic journal. Thirdly, the bioinformatic MeSH terms (medical subject headings) were manually selected from the list of all MeSH terms assigned to papers published in the bioinformatic journals, or in papers written by identified bioinformatists. Finally, a bioinformatic article is the one either published in a bioinformatic journal, or written by at least two bioinformatists; and in both cases assigned at least one bioinformatic MeSH term.

The data about the papers' authors, assigned MeSH terms, and publication dates were analyzed both separately and in combinations in order to identify co-authorship, thematic patterns and their temporal development.

Author groups containing 2 bioinformatists and 3 'other' authors (that are not bioinformatists by definition but still are among authors of bioinformatic articles) appeared to be the most successful. All authors may be roughly divided into two types: those who prefer to write papers with a certain group of co-authors, or those who have a big number of co-authors but a small amount of papers with each of them.

Correlations between groups of MeSH terms, publication dates and bioinformatists were studied closely. The main goal of our work was to understand how 'popularity' (relative usage of a term comparing to the previous year) of terms and their groups ("computers", "statistics", "linguistics", and "'proper' bioinformatics") changed through years and what authors appeared to follow 'fashion' in bioinformatics.

Almost the same fields were considered to be the most 'popular' for about 20 years: "sequence analysis", "statistical models", "single nucleotide polymorphisms", and "dinucleotide repeats". But the crucial shift of interest took place during last few years: fields concerning computers gained 'popularity', and at pre-

M.V. Lomonosov Moscow State University, Faculty of Bioengeneering and Bioinformatics, Moscow, Russia lyubetsky.anna@mail.ru

sent they are developing rapidly, as well as modeling. A period of time since 1998 till 2003 can be considered to be unambiguously successful for "'proper' bioinformatics", while "linguistics" in bioinformatics was popular since 1997 till 2001. "Statistics" was always an important but not overtly popular part of bioinformatics providing means of sequence analysis along with "computers".

Nowadays, the following bioinformatic fields may be called the most perspective ones: "synteny", "protein array analysis", "single nucleotide polymorphisms", "GC rich sequences", "computer simulation", "automated pattern recognition", "statistical models", "software" and "programming languages". The least 'popular' fields are ones concerning sequences, (but not "sequence analysis") such as "amino acid sequence", "base sequence", or "molecular sequence data". Thus, authors who wrote papers assigned a 'sequence' MeSH term turned out not to follow 'fashion' (in 85% of cases), and authors who wrote papers assigned "computational biology", "gene expression profiling", "algorithms", "oligonucleotide array sequence analysis", and "computer simulation" appeared to follow 'fashion' (in 45% of cases).

This is joint work with Mikhail Gelfand.

LONG HELICES IN MRNA PROCESSING

V. LYUBETSKY, A. SELIVERSTOV

An original algorithm was used to predict long mRNA helices; bacterial genomes were obtained from GenBank. Global search for long mRNA helices was performed; two examples are given below. Other cases of mRNA processing that involve long RNA helices will be also discussed.

Results: Brucella. Genes were found that are preceded by long RNA helices containing eight bases-long repeats in the shoulders that sometimes continue in the loop. The helices are found between genes on the same chain separated by a small region of up to 300 bases without a terminator, which suggests that the genes are part of one operon. The algorithm found two cases: (1) a long helix is preceded by hypothetical gene BMEI0570 and followed by gene mntH from Nramp family (E=2.5·10-156); (2) a long helix is preceded by gene gloA coding for Ni-dependent glyoxalase I and followed by hypothetic protein-coding gene BMEI1889. The latter is followed by gene BMEI1890 encoding a transporter containing domains DUF1775 (E=5.5·10-40) and DUF461 (E=2.4·10-45) (Table 1).

Institute for Information Transmission Problems RAS, Moscow, Russia
lyubetsk@iitp.ru

The same repeat AGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGU is present in 5'-shoulders of all mRNA helices found in B. melitensis 16M, B. suis as well as in genomes of Mus musculus, Rattus norvegicus and Ornithorhynchus anatinus. In the mammals, this sequence is not a part of RNA helix. The same repeat but of shorter length is found in B. abortus biovar 1 str. 9-941 and B. melitensis biovar Abortus 2308. 100-150 bases-long regions homologous to the found helices or to their loops occur in other parts of the Brucella genome but are not contained in any long RNA helix. Such homologs of any length were not found in other bacteria. Transcription of mntH gene in bacteria is often regulated by DNA binding protein MntR, however this protein is lacking in the Brucella genome. In studies (Rodionov D. et al., PLoS Comp. Biol., 2006, v. 2, p. 1568-1585) gene mntH is suggested to be part of the MUR regulon. Presence of a long mRNA helix in Brucella, unlike in other bacteria with MUR regulation, suggests an alternative regulation of mntH gene in Brucella.

Hypothesis: a helix containing the discussed repeat binds with a metalloprotein affecting mRNA stability. Stability of mRNA therefore depends on concentration of metal cations. A candidate metalloprotein in this case is ribonuclease H II, which binds Mn2+ ions and causes mRNA degradation within the long helix under high manganese concentrations. The found RNA helices with repeats might as well provide binding sites for a regulatory protein that protects mRNA under low metal concentrations.

Table 1. Orthologous genes in Brucella and E. coli with putative regulation.

B. melitensis 16M	*B. suis* 1330	*B. abortus* biovar 1 str. 9-941	*B. melitensis* biovar Abortus 2308	*E. coli* K-12	Gene annotation
BMEI0570	*BR1440*	*BruAb1_1435*	*BAB1_1459*		
BMEI0569	*BR1441*	*BruAb1_1436*	*BAB1_1460*	*mntH*	family Nramp (PF01566)
BMEI1888	*BR0056*	*BruAb1_0056*	*BAB1_0053*	*gloA*	Glycoxalase I (PF00903); EC 4.4.1.5
BMEI1889	*BR0055*	*BruAb1_0055*	*BAB1_0052*		Domain PF02451
BMEI1890	*BR0054*	*BruAb1_0054*	*BAB1_0051*		Domains DUF1775 (PF07987), DUF461 (PF04314)
BMEI1542	*BR0386*	*BruAb1_0411*	*BAB1_0415*	*rnhB*	Ribonuclease H II; EC 3.1.26.4

Results: chloroplasts. Genes *accD* (acetyl-CoA carboxylase beta subunit) and *atpH* (ATP-synthetase subunit) undergo RNA editing in chloroplasts of plants from genera *Anthoceros* and *Adiantum*. Minimal free energy (kcal/mole) was calculated for the RNA structure on a 40 bases-long mRNA region upstream genes *accD* and *atpH*, see Table 2. Long helices were found in this region in genera *Anthoceros* and *Adiantum* . The helices cover the putative ribosome binding site. Low energy accounts for high stability of the helices that, **hypothetically**, delay translation until the completion of editing. Table 2 shows free energies of the same mRNA region in representatives of all five genera. In *Huperzia lucidula* these two genes are without RNA editing, while possessing a low energy helix, and the ribosome binding site before gene *accD* is contained in the helix loop; other helices in this region have considerably higher energies (>-1.7 kcal/mol).

Table 2 (data partly shown). Presence of a pronounced helix before genes *accD* and *atpH*. Designations: 2nd and 4th columns show minimal free energies of helices in a 40nt mRNA region upsteam *accD* and *atpH* genes, respectively, kcal/mol; in 3rd and 5th columns "++" stands for presence, and "--" – for absence of gene editing.

Species	gene *accD*		gene *atpH*	
1	2	3	4	5
Anthoceros formosae	-7.0	++	-5.1	++
Adiantum capillus-veneris	-7.2	++	-5.2	++
Huperzia lucidula	-4.8	--	-2.9	--
Psilotum nudum	-0.8	--	-2.9	--
Pinus thunbergii	-3.6	--	-2.8	--

RNA STRUCTURES UPSTREAM *LEUA* GENES IN α-PROTEOBACTERIA

V.A. LYUBETSKY, A.V. SELIVERSTOV, O.A. ZVERKOV

Some α-proteobacteria (Rhizobiales: Agrobacterium tumefaciens, Aurantimonas sp. SI85-9A1, Brucella spp., Fulvimarina pelagi, Mesorhizobium spp., Rhizobium spp., Sinorhizobium spp.; Rhodospirillales: Magnetospirillum spp.; Rhodobacterales: Dinoroseobacter shibae, Jannaschia sp. CCS1, Loktanella vestfoldensis, Oceanicola spp., Rhodobacterales bacterium HTCC2654, Rhodobacter spp., Roseobacter denitrificans, Roseovarius spp., Sulfitobacter

IITP RAS, Moscow, Russia lyubetsk@iitp.ru

spp., Alpha proteobacterium HTCC2255) have conservative RNA structures in the 5'-untranslated region of the leuA mRNA (2-isopropylmalate synthase gene). In Rhizobiales these genes are the nearest genes for the leuA gene from Agrobacterium tumefaciens (/locus_tag="AGR_C_4114", /protein_id = "NP_355220.1"). In Rhodospirillales and Rhodobacterales these genes are the nearest genes for the leuA gene from Roseobacter denitrificans (/locus_tag = "RD1_1211", /protein_id="YP_681546.1"). Each considered RNA structure contains both a leader peptide gene with leucine codons and a conservative pseudoknot.

Contrariwise, no such structure is found upstream any leuA gene in Acidiphilium cryptum, Bartonella spp., Bradyrhizobium spp., Caulobacter cpp., Erythrobacter spp., Gluconobacter oxydans, Granulibacter bethesdensis, Hyphomonas neptunium, Maricaulis maris, Nitrobacter spp., Novosphingobium aromaticivorans, Paracoccus denitrificans, Parvularcula bermudensis, Rhodopseudomonas palustris, Rhodospirillum rubrum, Silicibacter spp., Sphingomonas spp., Sphingopyxis alaskensis, Stappia aggregata, Xanthobacter autotrophicus, Zymomonas mobilis.

Some of these bacteria have leader peptide genes upstream the leuA genes, see [1]. But they do not have the conserved pseudoknots. Bartonella spp. and Rickettsiales do not have the leuA gene.

Any conservative structure is a candidate for gene expression regulation mechanism. It is obvious, that coupling of the leader peptide translation and the kinetics of RNA secondary structure may be involved in the response on the variations of leucine concentration. Putative mechanism is based on degradation of mRNA. Considered structures do not contain obvious transcription terminators, i.e. long hairpins with poly-U tracts. On the other hand, pseudoknots are not closed to the initial codons in both Rhizobiales and Rhodobacterales. So we have no reason to propose a regulation on translation level.

The authors are grateful to A.G.Vitreschak for help and critical discussions.

The work was partially supported by grant ISTC 2766.

1. A.G.Vitreschak, E.V.Lyubetskaya, M.A.Shirshin, M.S.Gelfand, V.A.Lyubetsky (2004) Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis, FEMS Microbiology Letters, 234: 357-370.

EVOLUTION OF SPLICING IN INSECTS

D. B. MALKO[1], E. O. ERMAKOVA[2]

During the evolution, new mRNA and protein isoforms appear as a result of insertion, mutation and loss of functional elements of genes. Several recent studies analyzed the role of alternative splicing in evolution of gene structure was in mammalian genomes. The evolution of alternative splicing in insects is less well studied. One observation common to mammals and insects is lower conservation of alternative exons and splice sites compared to constitutive ones. The patterns of selection in constitutive and alternative regions in flies and in mammals seem to differ; however, the rates of nucleotide substitutions and selection patterns in alternatively spliced genes in the latter are subject of controversy.

Previous studies of splicing in insect genes were limited to the comparative analysis of the exon-intron structure in the genomes of two fruit flies and the malarial mosquito [1,2]. Here we consider patterns of selection and evolution of the exon-intron structure for 6794 clusters of orthologous genes of *A. gambiae* (Agam) and nine *Droshophila* species: *D. melanogaster* (Dmel), *D. simulans* (Dsim), *D. yakuba* (Dyak), *D. erecta* (Dere), *D. ananassae* (Dana), *D. pseudoobscura* (Dpse), *D. mojavensis* (Dmoj), *D. virilis* (Dvir), and *D. grimshawi* (Dgri).

The branches of the insect evolution tree were characterized by the rate of the following events for each cluster: intron insertion and loss, segment degradation (separately for constitutive and alternative segments), and splice site degradation (again separately for constitutive and alternative sites). Thus the obtained evolution tree describes separately evolution of the exon-intron structure and alternative splicing in the insect genomes.

region type	d_N/d_S
constitutive exons	*0.08*
constitutive parts of alternative exons	*0.08*
donor extensions	*0.10*
acceptor extensions	*0.12*
retained introns	*0.11*
cassette exons	*0.15*

[1] State Institute of Genetics and Selection of Industrial Microorganisms, 1st Dorozhny Proezd 1, Moscow, 117545, Russia, malko@genetika.ru
[2]Institute for Information Transmission Problems (the Kharkevich Institute), Russian Academy of Sciences, Bolshoi Karetny per. 19, Moscow, 127994, Russia, ermakova@iitp.ru

The patterns of intron gains and losses are different in the melanogaster subgroup (Dsim, Dyak, Dere), more "distant" flies of the Sophophora subgenus (Dana and Dpse), and the species of the Drosophila subgenus (Dmoj, Dvir, Dgri). Taking the Dmel structure as the reference, we observe that Dana and Dpse have similar counts of lost and gained introns, whereas in other species, intron gains prevail over losses (Dyak fourfold, Dsim sixfold, more distant ones, about fifty percent more). The nonsynonymous to synonymous substitution rate ratios (d_N/d_S) show that the distribution of positively selected, negatively selected and neutrally evolving sites differ not only between constitutive and alternative regions, but also between different types of alternative regions (see the table for average global rates).

This is joint work with M. S. Gelfand.

1. D.B. Malko *et al.* (2006) Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes, *Genome Res.,* **16**:505-509.
2. E.O. Ermakova *et al.* (2006) [Different patterns of evolution in alternative and constitutive coding regions of Drosophila alternatively spliced genes] Biofizika **51**:581-588 Russian

NETWORK ENTROPY AND CELLULAR ROBUSTNESS

T. MANKE, L. DEMETRIUS, M. VINGRON

Recent experimental efforts have highlighted the pervasiveness of molecular networks in biological sciences, as they control the information flow and regulation of many cellular signals.

For some model organisms, large-scale interaction data can already provide a glimpse of their global network architecture, but we have yet to understand the relation between structural and functional properties of biological networks. One important functional characterisation is the observed resilience of an organism against many perturbations, which have been studied systematically at a molecular level.

Here we present an analytical framework to characterise the macroscopic resilience of a network against microscopic perturbations. This work is based on a fluctuation theorem, which states that changes in network robustness are positively correlated with changes in network entropy [1]. Network entropy is a

Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany, manke@molgen.mpg.de, vingon@molgen.mpg.de

global quantity, which characterizes the diversity and uncertainty of microscopic pathways, and it is defined as the Kolmogorov-Sinai invariant of a Markov process, $P=(p_{ij})$, on a network

$$H = - \Sigma_{ij} \pi_i p_{ij} \log p_{ij} = \Sigma_i \pi_i H_i \qquad (1)$$

where p_i are the components of the stationary distribution of *P*.

Previously, we have utilized Eq. 1 and the fluctuation theorem to formulate an evolution model for biological networks based on entropic selection [2]. Now we refer to the second part of the equation, which leads to a natural ranking of individual proteins in an interaction network, according to their contribution to overall network entropy. H_i denotes the Shannon entropy.

In Fig. 1 we show that impairment of proteins with high entropic contribution ($\pi_i H_i$), results more frequently in a lethal phenotype, compared to random selection. This observation is robust against a number of known systematic errors, such as false positive and false negative interactions, as well as possible compartmental bias [3]. The observed enrichment is also more significant than for proteins which are ranked according to their degree. This illustrates that our analytical approach goes beyond the phenomenological studies based on node degree [4], and suggests a method to characterize proteins within their global network context. Moreover, the entropic framework is extendable to weighted networks, if more quantitative data is available. In a broader context, our work aims to bridge topological and dynamical (functional) properties of complex networks. To this end, we invoked a thermodynamic formalism, which can describe some properties of large systems with only a few macroscopic parameters. In particular the fluctuation theorem relates the return rate to steady state (robustness) to steady state properties (entropy). To the extent that the thermodynamic analogy holds, our approach should also be applicable to other complex networks.

1. L Demetrius, VM Gundlach, and G Ochs. (2004) Complexity and demographic stability in population models, Theor Popul Biol, 65 (3): 211-25.
2. L Demetrius and T Manke (2005) Robustness and network evolution -- an entropic principle, Physica A, 346(3-4): 682-96.
3. T Manke, L Demetrius, and M Vingron (2006) An Entropic Characterization of Protein Interaction Networks and Cellular Robustness, Royal Soc. Interface, 3(11): 843-50.
4. H Jeong et al. (2001) An Lethality and centrality in protein networks. Nature, 411 (6833): Figure 1. High-ranking proteins in the C.elegans network are more frequently essential (red), as compared to randomly chosen proteins (black). 41-2.

SNS-ALIGN: A TOOL TO ALIGN EVOLUTIONARILY DISTANT PROTEINS

GANIRAJU MANYAM[1], ANDREY MARAKHONOV[2], ANCHA BARANOVA[12], RAKESH MISHRA[3]

It has been shown before that the function of a protein may be retained even though many amino acids change as long as the critical sites are maintained through preservation of key residues that allow protein to retain the overall shape of the functional domains. This statement implies that the computerized recognition of the secondary structural features in the context of primary sequence may help to improve reliability of the protein alignments.

We developed structure and sequence alignment (SnS-Align) software that uses the combination of secondary structure prediction along with primary sequence alignment to score the degree of the protein similarity and reliably detect distantly related proteins. Taking a protein or DNA sequence as input, this tool identifies its homologue(s) in the evolutionarily distant organisms, which might not be found by comparing only the primary sequence alone. SnS-Align predicts secondary structures of all the input protein sequences or conceptually translated DNA sequences and replaces them by characters strings designating α-Helices and β-Sheets, respectively. Resulting sequences are represented both by these character strings and by amino acid sequences covering only the regions that can not be translated as α-Helices or β-Sheets (sandwiched sequence). These sandwiched sequences are used for local alignment using Smith-Waterman algorithm. Scoring matrices are calculated for each alignment in EMBOSS package. Scoring for α-Helices and β-Sheets is performed by taking the arithmetic mean of all amino acid scores in each of the scoring matrices. Amino acids that are conserved within the secondary structure deemed as functionally important and hence a bonus score is given to such amino acid positions.

We used human pannexin family for SnS-Align software testing. Three human pannexins represent vertebrate homologues of the innexins – invertebrate family of gap junction proteins. In vertebrates, gap junctions are mostly made of connexins that are specific only to Chordates. Although connexins and innexins have very different primary structures they nonetheless have some simi-

[1] Department of Molecular & Microbiology, George Mason University, Manassas, VA-20110, USA, gmanyam@gmu.edu
[2] Research Center for Medical Genetics, RAMS, Moskvorechie Str., 1, Moscow, Russia
[3] Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad-500007, INDIA, mishra@ccmb.res.in

lar features. Proteins of both unrelated families have similar topology with four transmembrane domains. SnS-assisted alignments of the connexins, innexins and pannexins are helpful for visualizing the structure similarities between these families and for deducing their evolutionary relationships. The standalone version of the tool can be freely downloaded from the following location: http://www.ccmb.res.in/rakeshmishra/tools.html.

1. Smith TF, Waterman MS (1981) Identification of common molecular subsequences, Journal of Molecular Biology, 147:195 –197.
2. Rice P et al. (2000) EMBOSS: The European Molecular Biology Open Software Suite, Trends in Genetics, 16:276-277
3. Piero Fariselli et al. (2006) The WWWH of remote homolog detection: The state of the art, Briefings in Bioinformatics, 8:78-87

ANTISENSE REGULATION OF HUMAN GENE *MAP3K13*: TRUE PHENOMENON OR ARTIFACT?

ANDREY MARAKHONOV[1], ANCHA BARANOVA[1], TATYANA KAZUBSKAYA[2], SERGEI SHIGEEV[3], MIKHAIL SKOBLOV[1]

Antisense regulation of gene expression is a widespread but not well understood mechanism of gene expression regulation [1]. Recently we have carried out a whole genome *in silico* search of *cis*-antisense clusters of transcripts in humans [2]. The developed database revealed a significant number of sense—antisense pairs consisting of one EST cluster expressed predominantly in normal tissues and another cluster with tumor-specific expression. The potential role of antisense transcripts in regulation of oncogenes and tumor suppressor genes is the most intriguing for the functional research. Here we describe and characterize an antisense mRNA *asLZK* overlapping human *MAP3K13/LZK* gene.

The protein coded by *MAP3K13/LZK* (mitogen-activated protein kinase kinase kinase 13 gene/leucine-zipper bearing kinase) is a member of serine-threonine protein kinases family involved in JNK/SAPK signal transduction pathway activated during mitogenesis. This pathway may be constitutively acti-

[1] Research Center for Medical Genetics, RAMS, Moskvorechie Str., 1, Moscow, Russian Federation, marakhonov@generesearch.ru, mskoblov@generesearch.ru
[2] Blokhin Cancer Research Center, Russian Academy of Medical Sciences, Kashirskoe Highway, 23, Moscow, Russian Federation
[3] Department of Forensic Medicine, Faculty of Medicine, People's Friendship University of Russia, Miklukho-Maklaya Str., 6, Moscow. Russian Federation

vated in tumors. Therefore, it was hypothesized that the transcription of *asLZK* antisense mRNA may lead to suppression of sense *MAP3K13* gene expression. Such event may lead to compensatory restrain added to one of the mitogenic signaling pathway, and unlikely to be supported by natural selection in the tumor cell population. To study the intriguing phenomenon of tumor-specific *asLZK* antisense we performed detailed *in silico* analysis of *asLZK*-like human sequences and their experimental evaluation in human tissues.

asLZK antisense transcript is represented by a cluster of 454 EST located within the first intron of MAP3K13 gene and is transcribed from the opposite DNA strand. We analyzed nucleotide sequence of antisense cluster *asLZK* by BLAST, and revealed 8 highly homologous loci spread between different chromosomes. The most interesting finding was an mRNA *RPL4* with 98 % homology to *asLZK*. This gene encodes ribosome large subunit protein L4. It is of interest that, according to *in silico* measurements, the level of *RPL4* expression is reflected by about 5500 ESTs stored in the database and is characterized by similar quantitative profiles in the normal and cancerous tissues. The level of *in silico* expression antisense *asLZK* is reflected by 454 ESTs, while 82 % of them were obtained from tumoral cDNA libraries. The rest of genomic loci similar to *asLZK* were less conservative (77—96 %), with no EST or mRNA were mapped to them. We believe that these loci represent silent pseudogenes of *RPL4* created in a result of retroposition.

We performed a multiple sequence alignment of *RPL4* mRNA, *asLZK* and its pseudogenes using ClustalW 1.83 program, then created PCR primers specific to *asLZK* antisense transcript. We also performed PCR based screening for *asLZK* expression in cDNA panel containing a number of normal human tissues (brain, muscle) and tumors (uterus, pancreas, kidney, ovary and rectum). No detectable expression of *asLZK* antisense locus has been revealed in any of the studied tissues, while both positive and negative control amplifications were read correctly. Therefore, we conclude that EST sequences that form *asLZK* cluster overlapping human *MAP3K13/LZK* gene are incorrectly mapped in dbEST on human genome. *asLZK* represents a silent pseudogene of *RPL4* created by its retroposition in the first intron of *MAP3K13* gene and does not participate in the regulation of *MAP3K13* expression. Therefore, an *asLZK* antisense to *MAP3K1* oncogene represents an artifact that needs to be corrected in the released human genome sequence.

The work in author's laboratory was supported by grant (07-04-00379-a) from RFBR.

1. M. Lapidot, Y. Pilpel (2006), EMBO reports 7:1216–1222.
2. D. Klimov et al. (2006), J Bioinform Comput Biol. 4(2):515—521.
3. A. Ikeda et al. (2001), J. Biochem. 130:773—781.

RNA POLYMERASE RESIDENT SITES IN BACTERIAL GENOMES: MULTIPLE OCCURRENCE AND PUTATIVE FUNCTION

I.S. MASULIS, M.N. TUTUKINA, K.S. SHAVKUNOV, V.I. LUKYANOV, O.N. OZOLINE

The traditional concept of promoter is based on its ability to provide RNA polymerase binding and transcription initiation. These functional properties should be supported by particular sequence and structural features of promoter DNA. Promoter-search software PlatProm summarizing inputs of different promoter-specific elements was used to depicture the whole-genome distribution of *E.coli* $E\sigma^{70}$ promoters (1). Besides known promoters, located in front of 471 genes, PlatProm pointed out 2251 potential transcription start sites for yet uncharacterized genes and nearly two thousand of promoter-like sites that can not be ascribed to any annotated gene. Part of them was found in intergenic regions and may indicate presence of novel genes. The most are those predicted within coding sequences, of which 1192 may provide antisense RNA products, while 709 coincide with the gene direction and may be required to produce new transcripts from intergenic loci or intensify the expression of properly oriented downstream genes. It should not also be excluded that some internal promoters anchoring RNA polymerase remain 'silent' and deficient in one of compulsory functions. Numerous intragenic RNA polymerase binding sites have been recently discovered using chromatin immunoprecipitation approach (ChiPChip assay) (2). Thus statistical assessment of intragenic promoter-like sites is coming to be relevant. In the present work we tried to discriminate transcriptionally active promoters from putative RNA polymerase residence sites in the frame of PlatProm facilities and present experimentally verified examples of potentially cryptic promoters.

Subsets, representing potential antisense (564 species) and co-directed promoters (267), were selected from the whole set of intragenic promoter-like signals. All of them are located more than 100 bp apart from the gene borders. Predicted intergenic promoters for novel genes were used for comparison. The control compilation possessing the whole assortment of specific features accounting by PlatProm was composed of 290 known promoters that were not

Institute of Cell Biophysics Russian Academy of Sciences, Pushchino, Moscow region, 142290, Russia, ozoline@icb.psn.ru

included in learning set. Promoters of four sets demonstrate similar patterns for consensus weight matrices, indicating that all of them may be recognized by $E\sigma^{70}$. Assuming that recognition *per se* is not always sufficient for productive RNA synthesis, we consecutively tested the presence of upstream A/T-rich sequences, deformable dinucteotides and periodicity in distribution of A/T tracts. These signs are proposed to reflect dynamic properties of promoter DNA, underlying conformational transitions of protein-DNA complex upon transcription initiation and early stages of elongation.

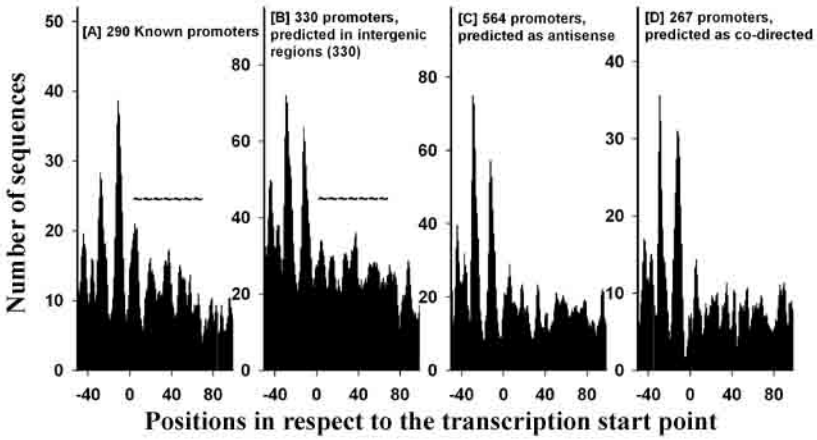


Fig.1. Distribution of paired A/T-tracts (www(n)13www) in the sequences of known [A] and predicted [B-D] promoters.

It was found that internal promoters demonstrate obvious depletion in the presence of periodically distributed A/T tracts in the early transcribed region (Fig.1C and D) as compared to promoters from control set (Fig.1A). Since predicted intergenic promoters were enriched by these elements (Fig.1B) it could not be explained by their underestimation in PlatProm weight matrices. Therefore, promoter population in terms of this feature appears to be non-homogeneous. Coupled A/T tracts represent specific signature proposed to assign the direction of RNA synthesis and regulate translocation of the transcriptional complex (3). Difference in the occurrence frequencies of A/T tracts in the structures of intergenic and internal promoters suggest that in fact a considerable fraction of the latter group is not suited for RNA polymerase escape and RNA synthesis. Seven internal promoter-like sites were tested in vitro for ability to bind RNA polymerase and initiate transcription. All promoter-containing fragments

interacted with RNA polymerase and formed transcription bubbles, but only two gave detectable products in the reaction of in vitro transcription. Thus, supporting the idea of "silent" promoters our date assume additional function for RNA polymerase-DNA interaction, yet to be elucidated. Broad spectrum of structural characteristics accounted by PlatProm opens a perspective of developing a tool for in silico functional dissection of unusual promoter classes.

Studies are supported by RFBR grant 07-04-01066a

1. Brok-Volchanski A., Masulis I., Shavkunov K., Lukyanov V., Purtov Yu., Kostyanicina E., Deev A., Ozoline O. (2005) in: Kolchanov N and Hofestaedt R (eds.), *Bioinformatics of Genome Regulation and Structure II,* Springer pp. 11-20;
2. Herring C., Raffaelle M, Allen T., Kanin E., Landick R, Ansari A., Palsson B. (2005) *J Bacteriol* 187, 6166-74;
3. Chasov V., Deev A., Masulis I., Ozoline O. (2002) *Molecular Biology* (Mosc.), 36, 682-688.

A STUDY OF GENES EXPRESSION EFFICIENCY ACCORDING TO ITS NUCLEOTIDE CONTENT BY BIOINFORMATICS METHODS

YURI MATUSHKIN, NIKITA VLADIMIROV, VITALI LIKHOSHVAI

It is well known that gene expression level correlates with genes' codon composition in many unicellular organisms. However there are some organisms where such correlation is missing. Previously we have shown the necessity of taking into account the inverted repeats in coding DNA sequences as causes of potential secondary structures, delaying ribosome move, for gene expression level estimation. The software package of Internet-available programs is developed. It allows estimation of potential expression level for coding sequences in unicellular organism genome sequenced. The index of elongation efficiency is calculated considering three factors: codon composition of coding sequence; frequency of inverted repeats; available energy of potential hairpins. The optimal individual combination of these factors is founded for every organism. The translation numerical characteristics are obtained for 384 unicellular organisms (351 bacteria's, 28 archeans, 5 eukaryotes). Presence of five evolutionary strategies of translation optimization for investigated organisms is proved basing on novel comprehensive sample. The significant difference of preferred

Institute of Cytology and Genetics SB RAS; Novosibirsk, 630090, Lavrent'ev ave. 10; Russia mat@bionet.nsc.ru

strategies frequencies is obtained for bacteria's and archeans. The reliable correlations between elongation efficiency index and expression level (obtained from microarray data) are shown for S.cerevisiae and H. pylori. The approach developed allows prediction of gene expression level also optimization of base ratio in transgenes for expression in predetermined organism-recipient not only in organisms in which used-codon frequency optimization is used but also in any other, for instance in H. Pylori or Mycoplasma gallisepticum. Web-version of program is available over Internet: http://wwwmgs.bionet.nsc.ru /mgs/ programs/eei-calculator/

The novel method of gene regulatory regions recognition based on oligonucleotide motifs is suggested. It is based on comparison of representation and characteristics of motifs distribution in investigated sequence and on gene regulatory regions sequences. The methods are implemented in software package ARGO_Proc. The method was tested on E.Coli sequences. The analysis of sample of E.Coli 5'- nontranslated region sequences in [-50;50] region has determined the set of significant oligonucleotide motifs. The most significant from them is NVAGGADN motif, which corresponds to Shine-Dalgarno consensus, which presence in this region is necessary for correct genes' translation.

The investigations were carried out about correspondence of M.gallisepticum genes elongation efficiency index (EEI) and two-dimensional electrophoresis data showing difference in proteins expression in different M.gallisepticum growth phases (cyanine straining of protein was used). Data about "count" of proteins is obtained in correspondence with growth phases (F1 - beginning of logarithmic phase, F3 - middle, F5 - stationary growth phase). The obtained results show the elongation efficiency index to be correlated with experimentally determined protein representation.

The elongation indexes for 1566 genes of Helicobacter pylori 26695 were calculated in various variants for the purpose of EEI adequacy estimation on correlation with experimental expression data. The necessity of local complementarity index (LCI) to be used was proved out. The primary structure optimization before start-codon was found to be very significant - when it is used, the correlation coefficient gets up on 30%.

The work was supported by Russian Foundation for Basic Research (No. 06-04-49556), Siberian Branch of the Russian Academy of Science (project № 10.4), Project "Evolution of molecular-genetic systems: computer analysis and modeling" of the RAS Presidium program "Biosphere origin and evolution".

SDPCLUST: A NEW TOOL FOR PREDICTION PROTEIN SPECIFICITY IN MPA

P.V. MAZIN, A.B.RAKHMANINOVA, O.V. KALININA

Introduction. The available data on protein sequences largely exceeds the experimental capabilities to annotate their function. So annotation *in silico*, i.e. using computational methods becomes of greater importance. To assign precise annotation automatically, it is crucial to be able to split a protein family into groups of protein of same specificity (specificity groups). Nowadays, there are some approaches to splitting families of protein sequences into specificity groups: Bete method [1] that uses Dirichlet mixtures and relative entropy to construct a phylogenetic tree from an alignment, and then the principle of the minimal cost of coding to divide it into subtrees; the giant component method [2] that allocates clusters based on sequence pairwise similarity; SPDsite method [3], developed in our group, that is based on Specificity Determining Positions (SDP). SDPs are alignment positions, in which distribution of amino acids correlates with splitting sequences into specificity groups.

In this study, we present a software package SDPclust, which includes methods to search SDPs and predict protein specificity. SDPclust is tested on simulated data constructed using standard sequence evolution model and on real data, namely, LacI family of bacterial transcription factors and OPA and MIP families of membrane transporters.

Materials and methods. SDPclust includes the following interconnected procedures: SDPlight to predict SDPs, SDPprofile to assign specificity to unannotated proteins, SDPgroup to split family into groups of specificity from a training sample (a small number of proteins from the considered family, for which specificity is known), SDPtree to construct a cluster tree of protein specificity.

SDPlight is a fast method to identify SDPs in a protein family alignment [4].

SDPprofile builds profiles over SDPs for all identifies specificity groups and calculates profile scores for given sequences [5].

SDPgroup is an iterative procedure for splitting family into groups of specificity using training sample:

Initiation: proteins from the training sample form initial specificity groups.

Department of Bioengineering and Bioinformatics Moscow State University, Vorob'evy Gory, 1-73, 119992, Moscow, Russia, iaa.aka@gmail.com

Step of iteration: SDPs are identified with SDPlight. For all sequences of the family, profile scores for all specificity groups are calculated using SDPprofile. Sequences are rearranged according to the maximal weight

End: The step of iteration does not result in rearramgement of sequences

SDPtree is a stochastic procedure to construct specificity cluster tree:

The family is randomly split into a large number of specificity groups

Starting from this splitting as a training sample, SDPgroup procedure is performed

Steps 1-2 were repeated 10000 times. A cluster tree was built based on how often two sequences fall into same specificity group.

Results and Conclusions. Previously we showed that results of SDPlight agree with structural and experimental data (Mazin and Kalinina, 2007). Testing SDPclust on generated data, for which specificity of proteins did not correlate with phylogenetic tree, shows that the method is capable to define groups of specificity even if they contradict phylogenetic data. When tested on well-studied LacI family, SDPclust performs better than other methods and agrees with data on protein specificity derived from genomic analysis [6].

We applied SDPclust to OPA family of transporters. The predicted SDPs agree well with structural and experimental data and allow to propose functional asymmetry of N- and C-terminal domains of proteins from this family. We also define a cluster tree of specificity and specificity groups for the OPA family and assign specifiicty to a number of unannotated family members.

1. K. Sjölander (1998), Proc Int Conf Intell Syst Mol Biol, **6**:165–174.
2. J. E. Donald, E. I. Shakhnovich (2005), Nucleic Acids Research, **33**(**14**): 4455–4465
3. O. V. Kalinina et al. (2007), Molecular biology, **41**(**1**): 137-147.
4. P. V. Mazin, O. V. Kalinina (2007), The XIV Int Scientific Conf for undergraduate and postgraduate students, and young scientists “Lomonosov”. 11 - 14 April 2007.
5. O.V. Kalinina et al. (2004), *Protein Science*, **13**: 443-456
6. O.N. Laikova (2003), Proc 1st Int Mosc Conf on Compl Mol Biol, 121-122.

IDENTIFICATION OF CPG ISLAND BOUNDARIES

JULIA MEDVEDEVA[1], IRINA ABNIZOVA[2], FEDOR NAUMENKO[3], MARINA FRIDMAN[1], NIKA OPARINA[4], VSEVOLOD MAKEEV[1]

CpG-islands are usually defined as CG-rich regions of genomic DNA, which are unmethylated and associated with 5' regions of different genes. CpG island methylation is thought to be the reason of gene suppression in normal case (e.g. gene imprinting, X-chromosome inactivation) and in disease case (e.g. cancer).

Yet, there are examples of CpG islands in which any of two their properties become violated. Significant part of CpG-islands is fully or partly methylated [1], unmethylated CpG-islands were found at the large distance of known genes [1]. Thus, function of CpG islands is not clear, and they deserve a detailed study.

CpG-island identified *in silico* are usually defined as regions with length > 200 bp, C+G content > 50 % and Obs[CpG]/Exp[CpG] > 0.6 [2], however this definition has only statistical functional justification. If island parameters vary, different sets of CpG islands come up; and their boundaries are especially variable. On the other hand, if CpG islands perform some clearly identified function, their boundaries detected *in silico* should be correctly identified to facilitate further experimental and computational studies of these genome regions.

Sp1 and CTCF proteins may play a role in CpG island boundary formation [3], [4]. Here we assume that (i) there are several types of CpG islands which differ in their structure and function; (ii) CpG islands distant from any transcription starts play functional role, possibly not related to gene suppression via methylation; (iii) all CpG islands have functionally justified boundaries.

We investigated sequence motifs located in the near-border regions of CpG islands using the method based on work [5]. This method allows detection of globally over-represented motifs as well as regions rich and poor with some motifs around CpG island borders. Our method performs automatic background correction for nucleotide bias. We searched for difference in significant motifs within following sets of CpG-islands:

1. CpG islands associated with different gene parts and CpG islands localized respectively far from known genes;

[1] GosNIIgenetika, Moscow, Russia, ju.medvedeva@gmail.com, marina-free@mail.ru, makeev@genetika.ru
[2] MRC-BSU, Robinson Way, Cambridge, UK, irina.abnizova@mrc-bsu.cam.ac.uk
[3] Queen Mary University, London, UK, fegir.naumenko@gmail.ru
[4] Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, oparina@gmail.com

2. CpG islands covering bi- and mono-directional promoters.

From this motif comparison, we studied different structural groups of CpG islands, and extracted protein binding sites and structural elements most likely responsible for CpG islands boundary formation. It appears that the role of Sp1 is substantially greater than the role of CTCF in CpG island formation.

1. Eckhardt, F. et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genetics*, 38: 1378–1385.
2. Gardiner-Garden, M., Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol.Biol.*, 196: 261–282.
3. Brandeis M, et al. (1994) Sp1 elements protect a CpG island from de novo methylation, *Nature*, 29: 435–438.
4. Yang Y., et al. (2003) Epigenetic regulation of Igf2/H19 imprinting at CTCF insulator binding sites. *J Cell Biochem.*, 5:1038-1055.
5. Irina Abnizova et al. (2007) Statistical and information characterization of Conserved Non-coding Elements in vertebrates (accepted into J. *of Bioinformatics and Comp. Biology*)

THE DATABASE OF PHYLOGENETIC ORTHOLOGOUS GROUPS (PHOG): THE ALGORITHM OF ITS CONSTRUCTION AND ITS APPLICATIONS IN COMPARATIVE PROTEOMICS

I. V. MERKEEV[1], A. A. MIRONOV[2]

In comparative genomics, it is frequently required to find genes that perform the same or a similar function in various organisms. Such genes are called orthologs and paralogs. It is very important to find such genes and put them to the appropriate clusters.

An algorithm was proposed that created clusters of orthologous groups at all nodes of the evolutionary tree [1, 2]. This algorithm starts from the leaves of the tree represented by organisms having completely sequenced genomes and runs to the root of the tree where the procedure is terminated. It starts comparing proteomes belonging to closely related organisms, and then it finds bidirectional best hits (BBHs) putting them into orthologous groups. Protein sequences belonging to one such group are multiply aligned and these multiple alignments are called phy-

[1] State Scientific Centre GosNIIGenetica, 1st Dorozhny pr., 1, Moscow, 113545, Russia, imerkeev@mail.ru
[2] Department of Bioengineering and Bioinformatics, Moscow State University, Vorob'evy gory, 1–73, Moscow, 119992, Russia, mironov@bioinf.fbb.msu.ru

logenetic orthologous groups (PHOGs), or supergenes, since the collection of PHOGs at a particular node forms a supergenome. The supergenome represents a proteome diversity of a particular node of the tree.

It is possible to find orhologous groups at all other nodes of the evolutionary tree by comparing child supergenomes, finding BBHs and aligning supergene multiple alignments into new multiple alignments, thus forming supergenomes corresponding to nodes lying higher in the tree. Existing algorithms for comparing multiple alignments use rigorous profile methods based on dynamic programming and they are very slow to perform this procedure for all nodes of the tree in reasonable time. Therefore, a new algorithm was proposed named PHOG-BLAST for fast profile search. This algorithm converts multiple alignments to sequences of letters belonging to a new alphabet. To find how this alphabet can look like, a special clustering procedure was applied to columns in multiple alignments belonging to well-known databases of multiple alignments. It was proven that there exist 20 clusters. Each cluster can be interpreted as a result of the evolution of just one ancestral amino acid.

This PHOG-procedure takes into account the possible different domain architectures of orthologous proteins, so it would be more appropriate call it a database of orthologous domains. It does so by running two times at each node. During the first time, it detects orthologous domains by finding BBHs and putting them into orthologous groups. During the second time, it cuts out N- and C- ends from orthologous proteins that are not aligned with other proteins in the same orthologous group and puts them to the second round of the comparison.

The PHOG database has applications in many areas of comparative genomics and proteomics:

1. Prediction of protein function.
2. Study of the genetic diversity of taxonomic groups.
3. Prediction of the domain structure of proteins.
4. Study of the rates and patterns of the molecular evolution.

1. I. Merkeev, P. Novichkov and A. Mironov (2006) PHOG: a database of supergenomes built from proteome complements. *BMC Evol. Biol.*, **6**:52
2. I. Merkeev and A. Mironov (2006) PHOG-BLAST – a new generation tool for fast similarity search of protein families complements. *BMC Evol. Biol.* 2006, 6:51

SIMULFOLD: SIMULTANEOUSLY INFERRING AN RNA STRUCTURE INCLUDING PSEUDO-KNOTS, A MULTIPLE SEQUENCE ALIGNMENT AND AN EVOLUTIONARY TREE USING A BAYESIAN MARKOV CHAIN MONTE CARLO FRAMEWORK

IRMTRAUD M. MEYER[1], ISTVÁN MIKLÓS[2]

Computational methods for predicting evolutionarily conserved rather than thermodynamic RNA structures have recently attracted increased interest. These methods are not only indispensable for elucidating the regulatory roles of known RNA transcripts, but also for predicting RNA genes[1]. Devising them has been notoriously difficult because a number of computational challenges have to be overcome. In order to get an accurate prediction of a conserved RNA structure, we have to have a high quality sequence alignment and an evolutionary tree relating several evolutionarily related sequences. These are two strong requirements which are typically difficult to fulfill unless the encoded RNA structure is already known. We present the first method, called SimulFold[2], that solves this chicken-and-egg problem by co-estimating all three quantities simultaneously. We show that our novel method can be successfully applied over a wide range of sequence similarities in order to detect conserved RNA structures, including those with pseudo-knots. We also show SimulFold's potential as an alignment and phylogeny prediction method. Our novel method overcomes several significant limitations of existing methods and has the potential to be used for a very diverse range of tasks.

SimulFold employs a novel theoretical framework for co-estimating an RNA structure including pseudo-knots, S, a multiple-sequence alignment, A, and an evolutionary tree, T, given several evolutionarily related RNA sequences, D, as input. We introduce a joint distribution of RNA structures, alignments and trees in a Bayesian framework. As it is not feasible to analytically calculate any interesting statistics in this model in reasonable computational time, we propose a Markov chain Monte Carlo (MCMC) method with which we can sample from the posterior distribution. Our novel theoretical framework allows us to sample (S, T, A) triples from the posterior distribution P(S, A, T|D) in a computationally very efficient way using a Bayesian Markov chain Monte Carlo. This

[1] UBC Bioinformatics Centre and Department of Computer Science, 2366 Main Mall, Vancouver, BC, Canada V6T 1Z4, irmtraud.meyer@cantab.net
[2] Hungarian Academy of Sciences, Reáltanoda utca 13-15, Budapest, Hungary 1053, MTA-SZTAKI, 1111 Budapest, Lágymányosi ut 11, Hungary and eScience Regional Knowledge Centre, ELTE, Pázmány Péter sétány 1/c, Budapest, Hungary 1117, miklosi@renyi.hu

is achieved by introducing a number of novel theoretical and computational tricks: we come up with a new expression for the prior P(S, A, T), propose a computationally very efficient way of jointly sampling structures and alignment and introduce a new type of MCMC sampler which we call a partial Metropolis importance sampler. The performance of SimulFold for predicting RNA secondary structures with and without pseudo-knots compares very well to the performance of RNAalifold, HXMATCH, Pfold and CARNAC across a wide range of average pairwise sequence identities and sequence lengths. We also present encouraging preliminary result that show SimulFold's potential as a alignment and phylogeny prediction program.

1. I.M.Meyer (2007), *A practical guide to the art of RNA gene prediction*, Briefings in Bioinformatics, in press.
2. I.M.Meyer and I.Mikós (2007), PLoS Computational Biology, under review.

DETERMINING THE POSITION OF RHIZARIA ON THE EUKARYOTIC TREE ON THE BASIS OF MULTIGENE ANALYSIS

K.V. MIKHAILOV[1], V.V. ALEOSHIN[2]

Reconstruction of the early stages of eukaryote evolution is a very active area of research that aims to decipher the phylogenetic message in the nucleotide sequences from the representatives of all major taxonomic lineages of eukaryotes. Our present knowledge of the first step in the radiation of eukaryotes may be better illustrated not by a tree but by a multifurcation that spawned around a dozen of "supergroups" [1], a condition that is referred to as the "Eukaryotic Big Bang". The taxon Rhizaria is one such supergroup. Organisms that belong to Rhizaria are numerous in the oceanic plankton and soil communities, a number of species lead a phytopathogenic or parasitic life style. Taxon Rhizaria embraces a large diversity of naturally established groups of protists: foraminiferans, radiolarians, desmothoracids, filosea amoebae, cercomonads and chlorarachniophytes. According to the fossil record Rhizaria is not only one of the largest protist supergroups but is also one of the most ancient: some fossil foraminiferans are thought to originate from the Cembrian period. The mono-

[1] Moscow State University, Division of Bioingeneering and Bioinformatics, Leninskie Gory,V.V. Lomonosov Moscow State University, Moscow 119992, Russia, Mikhailov_Kirill@mtu-net.ru
[2] Moscow State University, A.N.Belozersky Institution of Physico-Chemical Biology, Leninskie Gory,V.V. Lomonosov Moscow State University, Moscow 119992, Russian Federation, Aleshin@genebee.msu.su

phyly of Rhizaria was established on the basis of rRNA gene sequences and the sequences of the main components of the cytoskeleton, but the phylogenetic position of the taxon itself remained a mystery [2, 3]. Non-molecular approaches for tackling this problem have proved themselves futile since they were unable to infer the monophyly of Rhizaria, but recent availability of large amounts of sequence data begin to shed light on the phylogenetic position of this taxon. The sequencing effort has produced a large collection of expressed sequence tags for *Bigellowiella natans* – a member of chlorarachniophytes. Chlorarachniophytes are the only rhizarians that acquired an ability of photosynthesis – a virtue of a symbiosis with a photosynthetic eukaryote from the group of green algae. The plastids of chlorarachniophytes contain a remnant nucleus of the endosymbiont – the nucleomorph, the genome of which was sequenced [4]. However the genome sequence of the nucleomorph does not elucidate the phylogenetic relationship of the host cell, since the direction of gene transfer is only one way: from the endosymbiont genome to the host genome. In order to address this question we have searched the *Bigellowiella natans* database for the presence of ribosomal protein cDNA sequences of the host. A set of 61 (out of 80) full and partial ribosomal protein sequences was extracted from the database and aligned with the homologues sequences from over sixty representatives of other eukaryotic groups. The complete length of the alignment is approximately 10k aminoacid positions. The phylogenetic analysis was performed by neighbor-joining, maximum parsimony and maximum likelihood methods with aminoacid substitution rates and parameters optimised for each protein. The analysis yielded a sister group relationship between chlorarachniophyte *Bigellowiella natans* and another protist group, Heterokonta, with high statistical support. The discovered monophyly of a group uniting Rhizaria and Heterokonta allows us to reconstruct the last common ancestor of this group as a heterotrophic amoebaflagellate with a complex life cycle. Even further, this result may lead to a revision of our current views on the subject of plastid loss and gain as it prompts us to reconsider the central hypothesis of rhodophyte-derived plastid evolution, which assumes a single plastid gain event in the last common ancestor of Heterokonta, Dinophyta, Haptophyta and Cryptophyta [5, 6].

We thank the Russian Foundation for Basic Research for its financial support (grants 05-04-49705 and 06-04-49288).

1. S.L.Baldauf (2003) The deep roots of eukaryotes, *Science*, **300:** 1703-1706.
2. S.I. Nikolaev et al. (2004) The twilight of Heliozoa and rise of Rhizaria, a new supergroup of amoeboid eukaryotes. *Proc. Natl. Acad. Sci. USA*, **101:** 8066-8071.

3. F. Burki, J. Pawlowski. (2006) Monophyly of Rhizaria and multigene phylogeny of unicellular bikonts, *Mol Biol Evol*, **23:** 1922-1930.
4. P.R. Gilson et al. (2006) Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus, *Proc. Natl. Acad. Sci. USA*, **103:** 9566-9571.
5. T. Cavalier-Smith (2002) Chloroplast evolution: Secondary symbiogenesis and multiple losses, *Curr. Biol.,* **12:** R62-R64.
6. H.S. Yoon et al. (2002) The single, ancient origin of chromist plastids, *Proc. Natl. Acad. Sci. USA*, **99:** 15507-15512.

FOUR HELIX DESIGN USING AMINO ACID DOUBLETS

Z. MINUCHEHR[1], B. GOLIAEI[2]

Much more protein sequences are determined compared to protein structures in a daily manner (UniProtKB/Swiss-Prot Release 52.1 of 20-Mar-07 included 261513 entries**),** 3-D structure of proteins (PDB Holdings List: 27-Mar-07 included 42474 Structures in which 39000 are protein entries) is significantly much more difficult to determine. We should value the amino acid sequence of proteins in determining the overall fold of proteins as mentioned earlier but the relationship between sequence and structure is only partially yet understood (Baker and Sali 2001). Protein propensities remain excellent descriptors of amino acid tendencies for different secondary structures, alpha helices, beta strands, loops and turns. Examining the frequency of occurrence of different amino acids in protein secondary structures may give us an insight into the prediction of the three dimensional fold of the proteins or even de-novo design of a desired fold such as a four helix bundle. Studies have been conducted by different groups and us on preferences of amino acids in different secondary structures residues. There has been seen that alanine, glutamate and leucine tend to be present in alpha helices whereas valine and isoleucine tend to be present in strands, valine and isoleucine tend to destabilize alpha helices due to the steric clashes in the branching at beta carbon atom, but they are at the same time abundant in beta strands. Since studying the propensity of amino acids in different secondary structures is an important task to perform, as the protein data bank tends to grow, scientists have performed more specific studies on the propensity of amino acids in different positions in the secondary structures such as alpha helices(Aurora and Rose 1998; Kumar and Bansal

[1] National Institute for Genetic Engineering and Biotechnology, Tehran 14155-6343, Iran, (NIGEB) minuchhr@nrcgeb.ac.ir
[2] Bioinformatics Center, Laboratory of Biophysics and Molecular Biology, Institute of Biochemistry and Biophysics, University of Tehran, Iran goliaei@ibb.ut.ac.ir

1998; Goliaei and Minuchehr 2003) and less extensively on beta strands(Pal and Chakrabarti 2000; Ghamkhar, Minuchehr, and Goliaei 2005). Loops are also of functional importance in biology and may have key roles in recognition (Antibody hyper variable loops); ligand binding (e.g. Triosephosphate isomerase (Joseph, Petsko, and Karplus 1990)) or forming enzyme active sites (e.g. Serine protease, (Wlodawer et al. 1989)). Many studies has thus far tried to predict these particular structures (Burke and Deane 2001; Kuhn, Meiler, and Baker 2004; van Vlijmen and Karplus 1997) and many scientists have calculated the amino acid propensities in different secondary structures (Goliaei and Minuchehr 2003; Kumar and Bansal 1996; Penel, Hughes, and Doig 1999; Penel et al. 1999; Richardson and Richardson 1988), there is also some work done on loop regions for this purpose (Minuchehr and Goliaei 2005). Using our data on doublet propensities DLP's (Doublet Local Propensity) and singlet propensities SLP's (Single Local Propensity) and our data on loop regions, we designed two types of four helical bundle structures, the structures were then analyzed using different computational methods for the best predicted fold. The DLP oriented design showed a significantly better predicted fold compared to SLP, this shows the value of overlapping doublets used in calculating DLP's which is not present in the SLP values.

1. Aurora R and Rose GD (1998) Helix capping. *Protein Sci.* 7 (1):21-38.
2. Baker D and Sali A (2001) Protein structure prediction and structural genomics. *Science* 294 (5540):93-96.
3. Burke DF and Deane CM (2001) Improved protein loop prediction from sequence alone. *Protein Eng* 14 (7):473-478.
4. Ghamkhar, M, Minuchehr, Z, and Goliaei, B. Propensity calculation for amino acids in different positions in beta strands. Iranian Journal of Biochemistry and Molecular Biology 1(1), 93. 9-11-2005. Ref Type: Abstract
5. Goliaei B and Minuchehr Z (2003) Exceptional pairs of amino acid neighbors in alpha-helices. *FEBS Lett.* 537 (1-3):121-127.
6. Joseph D, Petsko GA, and Karplus M (1990) Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop. *Science* 249 (4975):1425-1428.
7. Kuhn M, Meiler J, and Baker D (2004) Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *Proteins* 54 (2):282-288.
8. Kumar S and Bansal M (1998) Dissecting alpha-helices: position-specific analysis of alpha-helices in globular proteins. *Proteins* 31 (4):460-476.

9. Kumar S and Bansal M (1996) Structural and sequence characteristics of long alpha helices in globular proteins. *Biophys.J.* 71 (3):1574-1586.
10. Minuchehr Z and Goliaei B (2005) Propensity of amino acids in loop regions connecting beta-strands. *Protein Pept.Lett.* 12 (4):379-382.
11. Pal D and Chakrabarti P (2000) beta-sheet propensity and its correlation with parameters based on conformation. *Acta Crystallogr.D.Biol.Crystallogr.* 56 (Pt 5):589-594.
12. Penel S, Hughes E, and Doig AJ (1999) Side-chain structures in the first turn of the alpha-helix. *J.Mol.Biol.* 287 (1):127-143.
13. Penel S et al (1999) Periodicity in alpha-helix lengths and C-capping preferences. *J.Mol.Biol.* 293 (5):1211-1219.
14. Richardson JS and Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science* 240 (4859):1648-1652.
15. van Vlijmen HW and Karplus M (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. *J.Mol.Biol.* 267 (4):975-1001.
16. Wlodawer A et al (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* 245 (4918):616-621.

HOW GENE ORDER IS INFLUENCED BY THE BIOPHYSICS OF TRANSCRIPTION REGULATION

LEONID MIRNY

What are the forces that shape the structure of prokaryotic genomes – the order of genes, their proximity and orientation? Co-regulation and coordinated horizontal gene transfer are believed to promote the proximity of functionally related genes and the formation of operons. However, forces that influence the structure of the genome beyond the level of a single operon remain unknown. Here we show that the biophysical mechanism by which regulatory proteins search for their sites on DNA can impose constraints on genome structure. Using simulations, we demonstrate that rapid and reliable gene regulation requires that the transcription factor (TF) gene be close to the site on DNA the TF has to bind, thus promoting the co-localization of TF genes and their targets on the genome. We use parameters that have been measured in recent experiments to estimate the relevant length and times scales of this process and demonstrate that the search for a cognate site may be prohibitively slow if a TF has a low copy- number and is not co-localized. We also analyze TFs and their sites

Massachusetts Institute of Technology, Cambridge, USA

in a number of bacterial genomes, confirm that they are co-localized significantly more often than expected, and show that this observation cannot be attributed to the pressure for co-regulation or formation of selfish gene clusters, thus supporting the role of the biophysical constraint in shaping the structure of prokaryotic genomes.

Our results demonstrate how spatial organization can influence timing and noise in gene expression.

MODELING OF THE PATTERN OF AUXIN DISTRIBUTION IN PLANT ROOTS

V.V. MIRONOVA[1], V.A. LIKHOSHVAY[1], N.A. OMELYANCHUK[1], S.I. FADEEV[2], E. MJOLSNESS[3]

Distribution of the hormone auxin in a plant root determines cell differentiation and direction of cell division, thereby forming the general root structure [1]. The highest auxin concentration is accumulated in the root meristem (Fig 1A), whereto auxin is delivered from the aerial parts of the plant through the vascular system. The main contribution to the auxin flow is made by PIN protein-facilitated auxin transport [2]. Experimental studies show that auxin induces its own transport at low concentrations (~100 nmol/l) and suppresses at higher ones (>10 μm) [3]. However, the mechanism controlling the formation of auxin distribution pattern in the root is still insufficiently understood.

[1] Institute of Cytology and Genetics, Lavrentyeva 10, Novosibirsk, Russia
kviki@bionet.nsc.ru
[2] Institute of Mathematics, av. Acad. Koptyug 4, Novosibirsk, Russia
fadeev@math.nsc.ru
[3] University of California, Irvine CA 92607, USA, emj@uci.edu

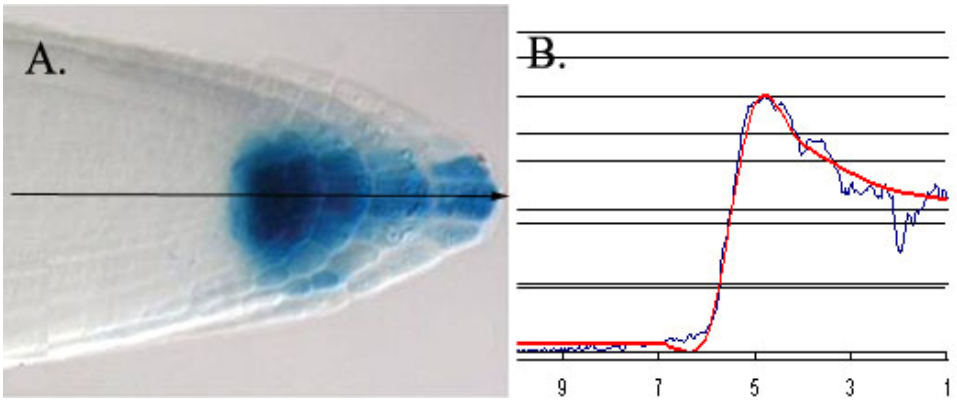


Fig. 1. A. Auxin distribution pattern in the root of *A .thaliana* and formation of the maximum in the root meristem [4]. B. Blue curve, qualitative auxin distribution curve obtained by scanning of Panel A; red curve, calculation of the model.

We present a simple one-dimensional model of auxin transport in the root:

$$\frac{da_N}{dt} = \alpha + P_t a_{N-1} - P_t a_N - K_d\, a_N - K_o\, a_N\, f(a_N)$$

$$\frac{da_i}{dt} = P_t(a_{i+1} + a_{i-1}) + K_o a_{i+1} f(a_{i+1}) - 2P_t a_i - K_d\, a_i - K_o\, a_i f(a_i),\ i = \overline{N-1,2}$$

$$\frac{da_1}{dt} = -P_t a_1 - K_d a_1 + P_t a_2 + K_o\, a_2 f(a_2).$$

(1)

$$f(a) = \frac{(\frac{a}{q_{11}})^{p_1}}{1+(\frac{a}{q_{12}})y^{p_1}} \cdot \frac{1}{1+(\frac{a}{q_2})^{p_2}}$$

Here N is the number of cells in the modeled area. The most proximal cell in the root (the last cell in the model array) is numbered N, and the most terminal cell at the end of the root is numbered 1. a_i is auxin concentration in cell i; K_d, coefficient of dissipation; P_t, rate coefficient of passive transport (diffusion), which is the same in both directions in this model; K_o>0, the rate constant of active transport, α , auxin influx into the last cell of the model. Hill's function *f(y)* is used to describe active transport with the following positive coefficients: $q_{11}, q_{12}, q_2, p_1, p_2$.

Parameter values were chosen for model (1) so that the steady-state auxin distribution should be in qualitative agreement with experimental data (Fig. 1B). Besides, the model reproduces experimental data on root tip regeneration and restoration of the auxin distribution pattern after tip laser ablation [5]. Also, the model allows *in silico* simulation of blurring of the terminal auxin distribution pattern with synchronous reduction of the auxin concentration peak and shift of the peak from the root tip to inner regions when active auxin transport from the cell is inhibited [3, 4].

Analysis of the model shows that it has numerous stable steady-state distributions with auxin concentration peaks at the root top, middle, and tip. In spite of this diversity, all the steady states have a feature in common: the presence of an auxin concentration distribution pattern in terminal cells of the root qualitatively corresponding to the experimental one (Fig. 1). Possible biological consequences of the modeling results are discussed.

Acknowledgements. We are grateful to V. Korolev, I. Gainova and A. Medvedev for providing the program package STEP+. This work was supported by the US National Science Foundation (FIBR EF-0330786 Development Modeling and Bioinformatics), Russian Federal Agency of Science and innovation (IT-CP.5/001), Russian Foundation for Basic Research (grant No. 05-07-98012), Russian Academy of Sciences (grant №10104-34/П-18/155-270/1105-06-001/28/2006 and the project on computer modeling and experimental design of gene networks), and the Siberian Branch of Russian Academy of Sciences (Integration Project No. 115)

1. Sabatini S. et al. (1999) *Cell.* **99**: 463–472
2. Friml J. et al. (2003) *Nature.* **426:** 147-153
3. Vieten A. et al. (2005) *Development.* **132:** 4521-4531
4. Wang J-W. et al. (2005) *The Plant Cell.* **17**: 2204–2216,
5. Xu J. et al. (2006) *Science.* **311:** 385-388

IN SILICO DESIGN AND IMPLEMENTATION OF A POLYKETIDE SYNTHESIS SYSTEM FOR PRODUCTION OF VIRTUAL LIBRARIES OF MACROLIDES

MEYSAM MOBASHERI , HOSSEIN ATTAR, SHARIAR SAIDI, AMIR HEIDARINASAB

The biochemical structure and biological activity of type I polyketide natural products are prescribed by modular multienzymic polyketide synthases(PKSs). The action of multifunctional PKS domains assemble a production-line pathway for biosynthesis of polyketide with a strong combinatorial potential, leading to an extraordinary diversity both in structure and function of these compounds. The astronomical number of these products make their experimental generation impractical. Thus developing knowledge based in silico approaches to produce virtual compounds is essential to discovery of novel polyketide structures and measuring their biological activities. Based on combinatorial chemistry and cheminformatics method, we developed a computational workbench capable of simulating the biosynthesis pathway of polyketide and producing virtual libraries of macrolides as a result of in silico manipulation of modules in PKS system. The produced libraries may further be filtered by the vHTS methods to pass the stages of hits and leads identification toward drug candidate discovery. Moreover, as a result of using this in silico approach, the selected virtual molecules provide the guideline for their genetic engineering regarding the genes encode the catalyzing enzymes of their biosynthesis reactions.

1. Heinz G. Floss, (2006) Combinatorial biosynthesis—Potential and problems, J. Biotech. , 124, 242–257
2. Gupta, S.; Bhattacharyya, B. (2003) Antimicrotubular drugs binding to vinca domain of tubulin. *Mol. Cell. Biochem.*, *253* (1-2), 41-47.
3. Gonzalez-Lergier, J.; Broadbelt, L. J.; Hatzimanikatis, V. (2005) Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways. *J. Am. Chem. Soc.*, *127*, 9930-9938.

Azad University, Science and Research Unit, Hesarak, Punak Square, Iran
m.mobasheri@seiau.ir

POLYMORPHISM OF ENZYMES CONTROLLING DRUG METABOLISM

I.M. MOKHOSOEV, A.A. TERENTIEV

Identification of enzymes participating in drug metabolism, both in Phase I and Phase II, has allowed studying in detail the individual differences in response to treatment with drugs. This depends on inter-individual variability in drug absorption, metabolism and excretion parameters that are determined by polymorphism of genes controlling these processes [1]. Experimental data indicate that individuals carrying certain alleles suffer from low sensitivity to drug treatment or adverse effects [2].

Human cytochromes P450 represent a large group of polymorphic enzymes that can alter drug metabolism [3]. They play an important role in Phase I of drug metabolism due to their monooxygenase activity and ability to incorporate one atom from molecular oxygen into drug substrate. In total, there are more than 2500 cytochrome P450 sequences known.

A polymorphism or difference in DNA sequence found at 1% or higher in a population is an important feature of enzymes catalyzing drug metabolism. As for cytochromes P-450s, the CYP2D and CYP3A are the best studied P450s regarding polymorphism. Each of these subfamilies of human cytochrome P450 is known to metabolize about 100 or more different drugs. The human cytochrome P450 polymorphism is an excellent subject for studying of interrelationships between amino acid substitutions and protein function.

Gene deletions, gene conversions with related pseudogenes and point mutations – all these types of mutations are revealed for genes encoding cytochrome P450s [4]. Point mutations leading to amino acid substitutions are revealed in a number of P450s and in some cases these mutations lead to the following consequences:

inactivation of the enzyme
reduction of the enzyme affinity to cytochrome P450-reductase
alteration of substrate specificity and affinity
destabilization of the enzyme

For example, L160H substitution in CYP2A6 causes inactivation of the enzyme, while R144C substitution in CYP2C9 reduced affinity to NADPH-cytochrome P450-reductase. However, there are no data regarding domains, modules or motifs in cytochromes P450 which may have functional importance. Available information, mainly, concerns variants of the enzyme. Nevertheless, it

Russian State Medical University, 117997 Ostrovityanova 1, Moscow, Russia
imokhosoev@mail.ru

will be of interest to reveal structural motifs in P450s, especially those involved in interactions with different substrates, ligands, and other proteins. For example, the well known cell adhesion RGD motif found in CYP2E1 and some short motifs participating in weak hydrogen bonding revealed in some cytochrome P450s.

Constructing of drugs with prolonged life in the body, diminished affinity to a certain enzyme catalyzing drug metabolism is a task for drug design. Crystal structures of such enzymes with drugs bound may be used for this purpose. To date, crystal structures of complexes with drugs are available in PDB database, at least, for 5 human P450s (cytochromes P450 3A4, 2A6, 2C8, 2C9 and 2D6). Modeling of 3D structures of the enzyme and their complexes with drugs on the basis of homology with other related enzyme may be used in case of lack of the enzyme crystal structure.

1. M. Ingelman-Sundberg (2001) *Drug Metabol. Disp.* **29**:570-573.
2. I.M. Mokhosoev, A. I.Archakov (1989) Biochemistry (Mosc.) **54**:179-186.
3. I.M. Mokhosoev, A.A. Terentiev (2005) *Adv. Nat. Sci.* (article in Russian) **12**: 67-68.
4. W.E. Evans, M.V. Relling (2004) *Nature*, **429**: 464-468.

DYNAMIC RESTRAINTS OF AMINO ACID SUBSTITUTIONS ARE POSSIBLE DURING PROTEIN EVOLUTION. MOLECULAR DYNAMICS SIMULATION STUDY OF ALPHA-FETOPROTEIN-DERIVED PEPTIDES

N.T. MOLDOGAZIEVA[1], A.A.TERENTIEV[1], K.V. SHAITAN[2]

It has been recognized that during protein evolution structural and functional restraints of amino acid substitutions exist [1]. These are determined by preservation of protein secondary structure, availability of amino acid residues for solvents, and hydrogen bonding, mainly, between the amino acid side chains and NH-group of polypeptide backbone. The preservation of such structural properties of proteins in local environment of functionally important sites is necessary for maintaining their biological activity. Thus, mutations resulting in amino acid substitutions leading to changes in protein conformation, disorder location of amino acid residues in functionally important regions and are

[1] Russian State Medical University, Ostrovityanova street, 1, Moscow, Russia, nmoldogazieva@mail.ru
[2] M.V. Lomonosov Moscow State University, Vorobyovy Gory,1 , Moscow, Russia, shaitan@moldyn.org

not secured in the population being eliminated during the evolution. Because dynamic features of proteins are also essential for their functioning, it was suggested that dynamic restraints of amino acid substitutions during the evolution should also exist. This means that there should be restraints preventing amino acid substitutions resulting in changes in conformational dynamics of amino acid residues surrounding functionally important regions.

In the present work conformational dynamics of a biologically active fragment of α-fetoprotein, the heptapeptide LDSYQCT (amino acid residues 14-20), and its analogs obtained by site-directed substitutions of amino acid residues was explored using equilibrium molecular dynamics simulation study. α-Fetoprotein (AFP) is the major mammalian oncofetal protein and structurally it is a mosaic, multi-modular one [2, 3]. Each module may function independently though binding to specific cell-surface receptor. The AFP segment LDSYQCT has been demonstrated to be one of its biologically active sites; however a role of individual amino acid in its functioning has not been yet studied.

The conformational dynamics of the peptide were conservative under the substitutions Y17F, Y17S, and D15E. Substitutions C19A and S16V resulted in only local changes in dynamic behavior of the peptide. Chemical modification of cysteine (C19) or dimerization of the peptide by producing a disulfide bond between cysteine residues of two parallel peptide chains, as well as substitutions C19G, C19S, Q18E, and D15N changed a set of probable conformations and dynamic behavior of all amino acid residues. The most significant changes occurred after substitutions of uncharged amino acid residues by charged ones, and *vice versa*.

Conformational and dynamic changes observed in the LDSYQCT peptide analogs depend on changes in intramolecular interactions (electrostatic or van der Waals interactions, hydrogen bonding, etc.) of amino acid side chains with each other or with NH- and CO-groups of polypeptide backbone [4]. Thus, substitutions resulting in disruption of intramolecular interactions enhance conformational flexibility of the peptide and destabilize its secondary structure. On the other hand, increasing of conformational mobilities of amino acid residues can promote changes in functional activity of the peptide. Amino acid substitutions similar to those, which occurred during the evolution of AFP in various biological species, do not markedly change conformational and dynamic behavior of all amino acid residues in the LDSYQCT peptide analogs. Substitutions Y17F and Y17S are permissible despite the differences in physicochemical properties of tyrosine, phenylalanine, and serine. Permissibility of substitutions of neutral hydrophilic amino acids by hydrophobic ones and also impermissibility

of substitutions of amino acids with similar sizes or hydrophobicities can be explained by specific features of their conformational dynamics. On the basis of these data the hypothesis about existence of not only structural but also dynamic restrictions of amino acid substitutions is proposed.

1. V. Chelliah, T.L. Blundell (2005) Quantifying structural and functional restraints on amino acid substitutions in evolution of proteins. *Biochemistry (Moscow)*, **70**: 835-840.
2. G.J. Mizejewski (2001) Alpha-fetoprotein structure and function: relevance to isoforms, epitopes, and conformational variants. *Exp. Biol. Med.*, **226:** 377-408.
3. A.A. Terentiev, N.T. Moldogazieva (2006) Structural and functional mapping of alpha-fetoprotein. *Biochemistry (Moscow)*, **71:** 120-132.
4. E.M. Popov (1997) *The Problem of Protein* [in Russian], Vol. 3, Nauka, Moscow.

DETECTING RECOMBINATIONS IN HIV WITH JUMPING PROFILE HIDDEN MARKOV MODELS (JPHMM)

BURKHARD MORGENSTERN

"Jumping profile Hidden-Markov-Models" (jpHMM) is a new approach tocomparative sequence analysis. It is a probabilistic generalization ofthe „jumping-alignment“ approach proposed by Spang et al. Given analigned set of input sequences together with a partitioning into knownsubtypes, each subtype is represented by a profile HMM. In addition tothe familiar transitions within these profile HMMs, our model allowstransitions between different subtypes. This way, different parts of aquery sequence can be aligned to different subtypes, depending on whichsubtype is locally most similar to the query. Jumps between subtypesoften indicate intersubtype recombinations. We applied our method to alarge set of genome sequences from human immunodeficiency virus (HIV)and hepatitis C virus (HCV) as well as to simulated recombined genomesequences.

University of Goettingen Institute of Microbiology and Genetics, Department of Bioinformatics, Goldschmidt str., 137077 Goettingen, Germany

LIMITATIONS OF ACQUISITION OF QUANTITATIVE DATA ON GENE EXPRESSION FROM THE CONFOCAL IMAGES OF DROSOPHILA EMBRYOS

EKATERINA MYASNIKOVA, SVETLANA SURKOVA, MARIA SAMSONOVA

In our experiments the data on gene expression is acquired by confocal scanning microscopy using fluorescence tagged antibodies [1]. The quantitative data are extracted from the confocal images at cellular resolution as described in [2]. The quantification is implemented using a pipeline of image acquisition, image processing, data acquisition and data processing methods. As a result the image files are transformed into the data presented in terms of nuclear locations and each nucleus is characterized by its coordinates, and the average fluorescence level. This approach has been applied to create a large data set of gene expression patterns in the fixed blastoderm stage embryos of *Drosophila melanogaster* [3]. However, the data are corrupted with the nuisance noise and distortions which arise in the course of data acquisition and processing. This kind of noise plays the key role among other sources of noise contributing into the total observed variability. The aim of this study is to estimate the errors at each step of image acquisition and quantification procedure using statistical methods and to discriminate them from the natural variability.

A considerable contribution into the total variability is made by errors occurring at the stage of scanning. The major source of errors is a shot noise from photomultiplier tube (PMT). The PMT gain (or voltage) roughly linearly increases the optical signal but exponentially increases electronic noise. The standard way to reduce this noise is averaging of multiple scans; however, some of residual noise is still detectable in the averaged image. We try to estimate the level of this portion of noise and to study how it is conditioned by microscope settings. We estimate the contribution of the residual shot noise into the averaged image and distinguish it from the intrinsic variability independent of the image acquisition procedure.

Obtaining of bright and high-contrast images is regulated by adjusting the gain and offset of the microscope PMT. We show that too high settings of gain and offset lead to the censored mean values not only at high (which is typical for any optical system) but also at low intensities. Using the statistical technique we propose a method for the correction of distortions in the quantitative data caused by the censoring.

St.Petersburg State Polytechnical University, St.Petersburg, 195251, Russia,
myasnikova@spbcas.ru, surkova@spbcas.ru, samson@spbcas.ru

One more source of errors lies in image segmentation procedure. The aim of segmentation is to detect exact edges of nuclei so as to read-off the data from intranuclear areas ignoring the internuclear space. It is shown that due to non-uniform distribution of intensities within nucleus even a very small error in edge detection leads to noticeable errors in mean intensities and nucleus-to-nucleus variation. We speculate that this effect can be explained both by protein localization and by smearing of nucleus edges in a confocal image of insufficiently high resolution.

Thus the statistical methods described in this work allow not only to estimate the contribution of experimental errors into the total data variability, but also to propose range of acceptable parameters providing the good accuracy of the data.

This work is supported by NIH grant RR07801, GAP award RBO-1286 and NOW-RFBR project 047.011.2004.013.

1. Kosman, D., Small, S., Reinitz, J. (1998). Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. *Dev, Genes, and Evol*, **208:** 290-294.
2. Janssens,H., Kosman,D, Vanario-Alonso,C., Jaeger,J., Samsonova,M. and Reinitz,J. (2005). A high-throughput method for quantifying gene expression data from early Drosophila embryo. *Development, Genes and Evolution*, **225(7):** 374-381.
3. Poustelnikova,E., Pisarev,A., Blagov,M., Samsonova,M., Reinitz,J. (2004) A database for management of gene expression data in situ. *Bioinformatics,* **20:** 2212-2221.

MALTASE-GLUCOAMYLASE GENE STRUCTURE AND EVOLUTION

DANIIL G. NAUMOFF

Maltase-glucoamylase (MGA) and sucrase-isomaltase (SUIS) are two mammalian glycosidases with complementary starch digestion activities [1-3]. Both enzymes are paralogues and consist of two homologous modules. Each module consists of a TIM-barrel type catalytic domain belonging to GH31 family of glycoside hydrolases [http://www.cazy.org/CAZY/] and several non-catalytic domains. GH31 family is widespread among almost all groups of living organisms, including Eukaryota, Bacteria, and Archaea. Several paralogous proteins, having GH31 domains, are encoded in many genomes.

State Institute for Genetics and Selection of Industrial Microorganisms, Moscow 117545, Russia, daniil_naumoff@yahoo.com

Sequence analysis of MGA gene and the flanking regions of the chromosome has allowed us to find from four to seven homologous fragments, encoding GH31 domains, in the mammalian genomes sequenced. Different copies from a single organism vary in sequence similarities. It suggests several recent tandem duplications in MGA gene. There is controversial information about expression mode of this gene cluster: at least in some species poly-modular mRNAs have been found (up to five GH31 domains in the case of rats).

We have performed a phylogenetic analysis of all known mammalian MGA sequences, using SUIS as an outgroup. It reveals a complicated history of MGA gene. Last common ancestor of MGA and SUIS genes appeared by GH31-encoding gene duplication followed by gene fusion (see figure). It happened at an early stage of Chordate evolution since bi-modular proteins are encoded in fish (*Tetraodon nigroviridis*) and ascidia (*Ciona intestinalis* and *C. savignyi*) genomes. Next duplication of MGA gene took place before divergence of placental mammals. Dog and cattle genomes encode four GH31 domains in the MGA loci. In primates, a new partial duplication happened, resulting in five-modular structure of chimpanzee, human, and macaque MGA loci. The most complicated MGA structure was found in rodents. In rats there are seven GH31 domains in this locus. Mouse genome encodes a six-modular structure. Only a series of line-specific partial duplications can explain the situation in rats and mice (see figure).. More detailed examination of the mammalian MGA loci allowed us to find several deletions and local rearrangements in different genomes. It allows to suggest that not all GH31 domains encoded by mammalian MGA loci are catalytically active.

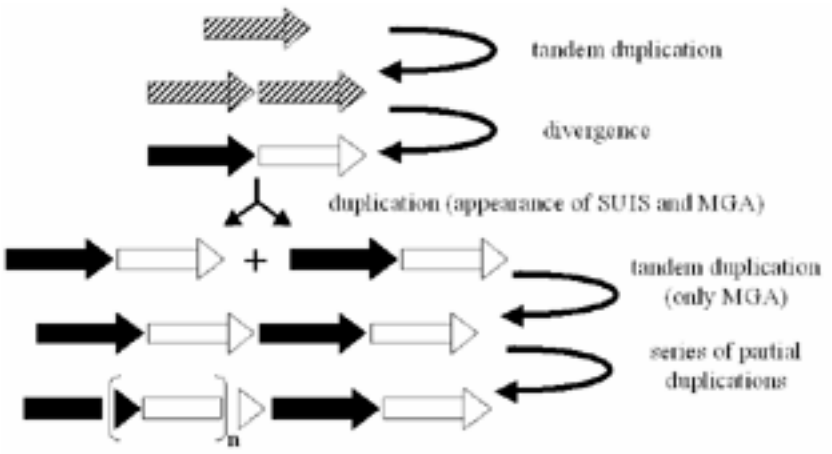


Figure. Evolution of maltase-glucoamylase (MGA) and sucrase-isomaltase (SUIS) genes. Each arrow corresponds to a gene fragment encoding only one GH31 domain. Copy number: $1 \le n \le 4$.

This work was supported by the Russian Foundation for Basic Research (grant 06-04-49079-a) and by grant of the Russian President for young scientists (MK-1461.2005.4).

1. W. Hunziker et al. (1986) The sucrase-isomaltase complex: primary structure, membrane-orientation, and evolution of a stalked, intrinsic brush border protein, *Cell*, **46:** 227-234.
2. B.L. Nichols et al. (1998) Human small intestinal maltase-glucoamylase cDNA cloning. Homology to sucrase-isomaltase, *J. Biol. Chem.*, **273:** 3076-3081.
3. B.L. Nichols et al. (2003) The maltase-glucoamylase gene: common ancestry to sucrase-isomaltase with complementary starch digestion activities, *Proc. Natl. Acad. Sci. USA*, **100:** 1432-1437.

INFORMATION STRUCTURE OF SHORT-CHAIN ALPHA-HELICAL CYTOKINES

A.N. NEKRASOV, L.E. PETROVSKAYA, V.A. TOPOROVA, E.A. KRYUKOVA, M.P. KIRPICHNIKOV

Short-chain alpha-helical cytokines (including GM-CSF, IL-2, IL-4, IL-13 and others) play an important role in the development and regulation of immune response processes. Despite the existence of the common fold, which includes the bundle of four alpha-helices, the level of amino-acid homology in the family is relatively low limiting the number of bioinformatics tools for the structural and functional predictions.

Our group has developed a new approach (ANIS method) to analyze protein sequences based on specific features of information entropy. This approach permits characterization of "information structure" (IS) of protein sequences. IS analysis of proteins is based on the identification of sites with Increased Degree Information Coordination (IDIC). IS has a hierarchic organization and can be presented as IDIC-trees with branches of lower order. **A**nomalously **D**istributed Density (ADD) of IDIC-branches in the lowest hierarchic level is likely to play an important role in functional properties of proteins.

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Miklukho-Maklaya 16/10, 117997, Moscow, Russia, an_nekrasov@mail.ru

Analysis of information structures of alpha-helical cytokines has revealed some common features. They consist of the limited number of elements (3 for GM-CSF, IL-2, IL-4 and IL-13). The highest level of IS similarity between IL-4 and IL-13 reflects the most noticeable homology in this protein family. Alpha-helices B and C predominantly contain ADD- sites; ADD+ sites is situated in helices A (except GM-CSF) and D.

The transduction of the regulatory signal requires interaction of a cytokine with cell-surface receptor. The study of interaction sites in the complex of IL-2 with its heterotrimeric receptor with the help of ANIS method revealed that inter-protein contacts most frequently include ADD-/ADD+ pairs. On the basis of these results assumptions were made about possible contact sites in other cytokine-receptor complexes.

It is known that natural antagonists of IL-2 and IL-4 represent the products of alternative splicing with the deletions of whole exons. We have shown that in the IS of these proteins the position of exon-encoded sequences coincides with the IS elements of highest order.

Therefore, ANIS method represents an effective tool for the studies of structural and functional organization of the protein molecules. With the help of this approach only on the basis of amino acid sequence data one can determine the localization of the sites participating in inter-protein contacts, including ligand-receptor interactions, predict the existence of deletion variants with different functional properties and design the deletion mutagenesis experiment leading to conformationally stable protein variants.

The work was supported by RAS 16-12315/2003 1002-251/П-10/145-143/010 403-087 and RFBR.

SIGNIFICANCE OF MOLECULAR MECHANISMS OF MORPHOGEN DETECTION FOR PATTERN FORMATION MODELING

S. NIKOLAEV[1], S. FADEEV[2], E. MJOLSNESS[3], N. KOLCHANOV[1]

While some models of spatial pattern formation take into account molecular mechanisms of morphogen detection, other ones often ignore explicit consideration such mechanisms ([1]). This ignorance is a way to simplify model.

[1] Institute of Cytology and Genetics, Novosibirsk, Russia, 630090, nikolaev@bionet.nsc.ru, kol@bionet.nsc.ru
[2] Institute of Mathematics, Novosibirsk, Russia, 630090, fadeev@math.nsc.ru
[3] School of Information and Computer Science, and Institute for Genomics and Bioinformatics, University of California, Irvine, CA, USA, emj@uci.edu

Before we considered a 1D model of shoot apical meristem structure homeostasis ([2]):

$$\frac{\partial Y}{\partial t} = \frac{1}{\tau_Y} g(h_Y + T_{YW} W) - d_Y Y + D_Y \frac{\partial^2 Y}{\partial r^2}$$

$$\frac{\partial C}{\partial t} = \frac{1}{\tau_C} g(h_C + T_{CY} Y) - d_C C, \quad 1 \le i \le n$$

$$\frac{\partial W}{\partial t} = \frac{1}{\tau_W} g(h_W + T_{WY} Y + T_{WC} C) - d_W W + D_W \frac{\partial^2 W}{\partial r^2}, \qquad 1 < i < n-1$$

The sigmoid function is in the form:

$$g(x) = \frac{1}{2}(1 + \frac{x}{\sqrt{1+x^2}}).$$

The model does not include a mechanism of morphogen detection. More over, the model structure implies that there is no morphogen consumption during its detection. So morphogen gradients are formed by it's diffusion and decaying.

In the present work we considered some mechanisms of morphogen detection, and studied its influence on the model structure and dynamics. For example, it is interesting, when we can ignore such mechanisms, and don't consider the influence of morphogen detection on morphogene dynamics in resulting model. Special interest was paid to question on significance of such mechanisms as additional restrictions for model parameter optimization.

1. A. Lander et al. (2002) Do Morphogen Gradients Arise by Diffusion?, *Developmental Cell,* **2:** 785–796.
2. S. Nikolaev et al. (2006) Analysis of a one-dimensional model for the regulation of the size of the renewable zone in biological tissue, *Computational technologies,* **11:** 65-79 (in Russian).

COMPUTATIONAL PREDICTION AND ANALYSIS OF TRANSCRIPTIONAL REGULATORY MODULES IN MAMMALS

A.A. NIKULOVA, A.A. MIRONOV

Regulation of gene expression is one of the basic elements of the genome life. In eukaryotes ranging from nematodes to mammals the number of coding genes is similar. A common explanation is that the organismal complexity may be attributed to phenomena such as alternative splicing, DNA rearrangement and transcriptional regulation. Thus, the identification of transcriptional regulatory elements and characterization of their interaction with the respective transcription factors (TFs) lie at the very heart understanding organismal complexity and development.

In higher eukaryotes, transcription factor binding sites (TFBSs) tend to be rather short (5-15 bp) and degenerate and often spread in extensive non-coding regions. So, because of the size of vertebrate genomes, thousands of potential transcription factor binding sites are expected to be found by chance. For more reliable prediction it is necessary to apply other criteria in addition to the sites' sequence.

TFBSs are known to be organized in groups confined to regions of a few hundred base pairs. These groups are referred to as modules or clusters. Various modules work together to provide the combinatorial regulation of gene transcription in response to various developmental and environmental stimuli. Moreover, recent work has confirmed that, in general, individual transcription factor binding sites are more conserved than is their surrounding DNA [1].

In this study, we used a site-clustering criterion for prediction of putative TFBSs in upstream regions of human muscle-specific genes. We also used site conservation in the human and mouse genomes to increase the precision of our method.

First, we used the CluSite algorithm [2] to search for clusters of sites that could be bound by muscle-specific TFs in 10 Kb upstream regions of 28 human and orthologous mouse genes reported as muscle-specific in a catalogue of regulatory elements [3]. Unlike most current predictive tools based on scanning a genome sequence with a sliding window of a fixed size and a fixed site score threshold, CluSite searches for the most statistically significant clusters of TFBSs. The background distributions of site scores for different TFs were calculated based on the genome sequence observations.

To estimate the quality of our predictions we compared the found TFBSs with known, experimentally characterized sites. This demonstrated high sensi-

Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University. Vorbyevy Gory 1-73, Moscow, 119992, Russia, nikanka@gmail.com

tivity, but rather low precision. To improve the precision, we applied a site-conservation criterion. We retained only those TFBSs that were found simultaneously in the human and mouse genomes. This filter resulted in 5-fold increase of precision with only slight reduction of sensitivity.

Next, we explored TFBSs arrangement in clusters. For this purpose we calculated the correlation coefficient for all pairs of TFBS's types (sites that bound different TFs or that occurred in different strands of DNA were considered to be of different types). The results of this analysis showed that TFBSs of some specific types were much more often placed next to each other than expected by chance. Moreover, the tendency of sites to occur in the same DNA strand was observed. Further, we examined distances between adjacent sites in clusters. Our analysis demonstrates that the distribution of distances between sites in clusters differs from the random (geometric) distribution. Moreover, there are preferred distances between particular types of sites. This fact may be caused by functional interactions between TFs binding to these sites.

A separate part of this project was to develop a method to identify genes regulated by muscle-specific TFs. In order to do that, we used the logistic regression analysis (LRA). Parameters for the LRA were chosen based on the assumption that if a gene was regulated by muscle-specific TFs, its regulatory regions would have more significant clusters with a large number of TFBSs. The positive test set consisted of 21 muscle-specific genes with unknown TFBSs positions. The negative test set contained 21 genes that were not expressed in the muscle tissue. As a result, our method could identify muscle-specific genes with 67% sensitivity and 71.4% specificity.

We are grateful to R.A. Sutormin and D.V. Vinogradov for useful discussions and to R.N. Nurtdinov for shared data.

1. Y. Liu, X.S. Liu, L. Wei, R.B. Altman, S. Batzoglou (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.,* **14:** 451–458.
2. A.A. Mironov (2005) An efficient algorithm for identification of clusters of transcription factor binding sites. *Proceedings of the International Moscow Conference on Computational Molecular Biology, MCCMB-2005,* 232–234.
3. W.W. Wasserman, J.W. Fickett (1998) Identification of Regulatory Regions which Confer Muscle-Specific Gene Expression. *J. Mol. Biol.,* **278**: 167-181.

INVESTIGATION OF THE AMINO ACID SEQUENCES OF BACILLUS SUBTILIS COMPLETE GENOME WITH PROTEIN FAMILY PATTERNS BANK PROF_PAT

L.P. NIZOLENKO[1], A.G. BACHINSKY[1], A.N. NAUMOCHKIN[1], A.A. YARIGYN[1], D. A. GRIGORIVICH[2]

Prof_Pat is a database of patterns, constructed for groups of related proteins from UniProt. [Bachinsky et al., 2000]. The version of Prof_Pat 1.19 contains more than 108,000 patterns and identifies more than 1 570 000 sequences from 1 939 634 full-length sequences in UniProt (rel. 8) with true or potentially true (putative) similarity. The bank is installed on http://wwwmgs.bionet.nsc.ru/mgs/programs/prof_pat/. Database searching program can examine large groups of protein sequences rather than just a few of them. Even as large as all amino acid sequences, translated from complete genomes of microorganisms.

We have investigated 4105 amino acid sequences of open reading frames of *Bacillus subtilis* presented in National Center for Biotechnology Information (NCBI) Internet site 23.01.2007 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ Bacillus_subtilis/). Only 5 sequences have no any similarity with patterns of Prof_Pat. For 817 of 4100 recognised sequences the similarity level was not high enough to be significant [Nizolenko et all. 2004]. And 3283 were confidently identified.

22 sequences, which similarity to any proteins except hypothetical ones with unknown function was not described up to now, recognised by Prof_Pat families of proteins with determined or predicted function. For 8 of them (Table 1) it is possible to derive some information about protein probable function when compare sequences with Interpro databank [Mulder, et al. 2007]. And for another 14 (Table 2) our prediction is unique and can be obtained at present when use Prof_Pat only.

Table 1. New similarity, confirmed by Interpro itself or some of it's member database.

Sequence	Prof_Pat	Interpro
GI:16077862	Radical SAM, Oxidoreductase	Radical SAM domain
GI:16078193	Hydrolase	Hydrolase BH2892 (ProDom, Superfamily)

[1] Theoretical Department, Research Institute of Molecular Biology, SRC VB "Vector", Koltsovo, Novosibirsk region, 630559, Russia, nizolenko@vector.nsc.ru
[2] Laboratory of Theoretical Genetics, Institute of Cytology and Genetics, Lavrentyev Ave., 10, Novosibirsk, 630090, Russia, odip@bionet.nsc.ru

GI:16078788	Hydrolase, Metallo-beta-lactamase	Metallo-hydrolase/oxidoreductase (Superfamily)
GI:16079244	(EC 2.1.1.-) Protein-L-isoD O-methyltransferase	S-adenosyl-L-methionine-dependent methyltransferase (SCOP fold)
GI:50812272	(EC 3.2.1.23) Zn-dependent metalloprotease	Protein DUF965. Family member is Zn-dependent metalloprotease
GI:16080037	Nucleoside 2-deoxyribosyl transferase YTOQ	N-deoxyribosyltransferase (Superfamily)
GI:16080230	Spore germination protein	Germination probable sporulation GERPF (ProDom)
GI:16080280	Spore coat protein	Spore coat protein YutH (Tigrfam)

Table 2. New similarity, predicted by Prof_Pat only.

Sequence	Prof_Pat	Interpro
GI:50812177	Multidrug ABC transporter, permease	Transmembrane regions
GI:16077362	Tellurium resistance protein	No hits reported
GI:16077388	Nucleotidiltransferase	No hits reported
GI:16077517	ABC transport system permease	Transmembrane regions
GI:16077694	Cold shock protein	No hits reported
GI:16081117	Responsible for oxetanocin A resistance	5 peptide_repeat.
GI:16078072	Transporter	Transmembrane regions
GI:16078137	Stage 0 sporulation regulator	No hits reported
GI:16078278	Zn dependent protease	No hits reported
GI:50812230	DNA binding transcription regulator	Transmembrane region
GI:16078875	Acetyltransferase	No hits reported
GI:16078959	Spore coat protein	No hits reported
GI:16079170	Prophage terminase, ATPase	No hits reported
GI:16079218	Spore coat protein	No hits reported

1. A.G. Bachinsky et al. (2000) *Bioinformatics,* **16:** 358-366.
2. L. Ph.Nizolenko, et al. (2004) *Molecular Biology,* **38:** 210–217.
3. N. J. Mulder, et al. (2007) *Nucleic Acids Research,* **35**(Database issue): 224-228.

RECONSTRUCTION AND ANALYSIS OF THE GENOME-SCALE METABOLIC NETWORK OF *LACTOCOCCUS LACTIS* MG1363

RICHARD A NOTEBAART[1], ROLAND J SIEZEN[123], BAS TEUSINK[123]

The genomic information of a species allows for the genome-scale reconstruction of its metabolic capacity. Such a metabolic reconstruction gives support to metabolic engineering, but also to integrative bioinformatics and visualization[1]. The core of genome-scale metabolic networks is formed by the set of relationships between genes, proteins and metabolic reactions (GPRs). As such the final metabolic network is the result of the integration between a genetic (i.e. genes/proteins) and a detailed metabolic reaction network (i.e. stoichiometry).

We have reconstructed the metabolic network of the industrially important model organism *Lactococcus lactis* MG1363. To reconstruct the network we applied a semi-automatic reconstruction method called AUTOGRAPH (AUtomatic Transfer by Orthology of Gene Reaction Associations for Pathway Heuristics)[2]. This method combines manually curated genome-scale metabolic networks of different organisms (e.g. *L. plantarum*[3]) with orthology to predict a network for a query organism (i.e. *L. lactis* MG1363). When genes from the query organism and genes from the curated genome-scale metabolic networks have orthologous relationships, metabolic reactions and protein complex information is transferred to the query organism. In this way we can accelerate the reconstruction process significantly.

The automatic reconstruction revealed ~500 genes to be associated to proteins (or protein complexes) and metabolic reactions. We have manually curated the predicted GPRs by using biochemical/physiological knowledge and where necessary genomic context based methods. Moreover, we have added *L. lactis* specific metabolic reactions based on physiological data, including the formation of cell components (lipids, cell wall, polysaccharides, etc) and the assembled overall biomass equation. By applying flux analysis using constraint-based modeling techniques we examined whether or not all cell components (and finally biomass) could be made. Thereby we closed existing gaps in the metabolic reaction network.

[1] Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, The Netherlands
[2] TI Food and Nutrition (WCFS)), Wageningen, The Netherlands
[3] NIZO food research BV, Ede, The Netherlands, R.notebaart@cmbi.ru.nl

We present a manually curated genome-scale metabolic network of *L. lactis* MG1363 consisting of 510 genes associated to proteins (or protein complexes) and 571 metabolic reactions of which ~90% has been associated to genes/proteins. The AUTOGRAPH method has shown to be very useful in accelerating the reconstruction process and the resulting metabolic network is currently being explored for both metabolic modeling (e.g. *in silico* growth experiments) and integrative bioinformatics.

1. Teusink B. et al. (2006), Modelling strategies for the industrial exploitation of lactic acid bacteria, *Nat Rev Microbiol.* **4(1)**:46-56
2. Notebaart RA. et al. (2006), Accelerating the reconstruction of genome-scale metabolic networks, *BMC Bioinformatics* **13**;7:296
3. Teusink B. et al. (2006), Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model, *J Bio. Chem.* **281(52)**:40041-8

SEARCH FOR STRUCTURAL FACTORS OPTIMIZING THE LIGHT-HARVESTING ANTENNA FUNCTIONING. THEORETICAL AND EXPERIMENTAL STUDIES

A.A. NOVIKOV[1], A.S. TAISOVA[2], N.V. FEDOROVA[2], L.A. BARATOVA[2], Z.G. FETISOVA[2]

This work deals with the problem of theoretical and experimental investigation of an optimal constitution of subantennae in photosynthetic light-harvesting antenna of the green filamentous bacterium *Oscillochloris trichoides*. At present, two subantennae were identified surely: chlorosomal BChl *c* subantenna B750 and membrane BChl *a* subantennae B805-860. Some indirect experiments indicated on the presence of minor amounts of BChl *a* in isolated chlorosomes, however, in absorption spectra of isolated chlorosomes, this BChl *a* subantenna was not visually identified (Fig.1). Using mathematical modeling of the functioning of the natural antenna, we showed that such intermediate-energy BChl *a* subantenna, connecting B750 and B805-860 ones, allows one to control the superantenna efficiency, i.e., to optimize excitation energy transfer from B750 to B805 by functional criterion and, hence, the exis-

[1] Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, 119992 Russia
[2] A.N. Belozersky Institute of Physico-Chemical Biology, M.V. Lomonosov Moscow State University, Moscow, 11999, Russia, zfetisova@genebee.msu.ru

tence of such intermediate-energy subantenna is biologically expedient. To prove experimentally the existence of the intermediate BChl *a* subantenna, *Osc. trichoides* chlorosomes were subjected to strong alkaline treatment. This led to selective degradation of BChl *a*, whereas BChl *c* was not affected. The alkaline treatment caused dramatic changes in the fluorescence spectra of chlorosomes measured under 77K (Fig.2).

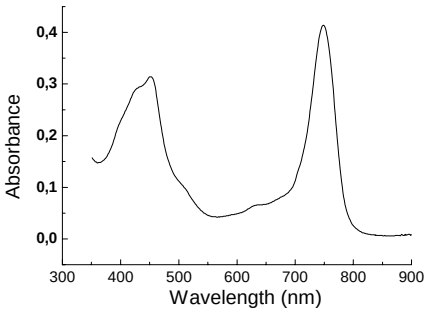


Fig. 1. Absorption spectra of *Osc. trichoides* chlorosomes

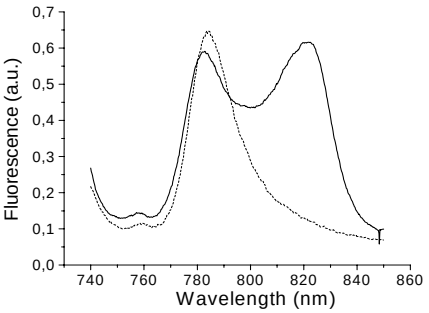


Fig. 2. Fluorescence emission spectra of *Osc. trichoides* chlorosomes under reducing conditions at 77K. Solid line - control chlorosomes; dashed line – alkaline-treated chlorosomes.

It should be noted that BChl *a* emission could slightly discerned in the fluorescence spectra of chlorosomes at room temperature (293 K). The effect of alkaline treatment on chlorosomal proteins was investigated by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). Three major chlorosomal proteins, CsmA, CsmN and CsmM, are found in isolated chlorosomes from *Osc. trichoides* with apparent molecular weights of 5.7, 11 and 18 kDa based on their relative mobilities in SDS-PAGE. Upon alkaline treatment, only the 5.7 kDa CsmA protein was removed from the chlorosomes concomitantly with the disappearance of BChl *a* fluorescence peaked at 821 nm. BChl *c* fluorescence remained unlike BChl *a* fluorescence. We suggested that the disappearance of BChl *a* emission is caused by the removal of BChl *a* in the baseplate. Selective BChl *a* disappearance should be expected only in case when

BChl *a* is associated with chlorosomal proteins, which are in close contact with the cytoplasmic membrane. Based on these results, we suggest that BChl *a* is associated with CsmA protein in *Osc. trichoides* chlorosomes. Thus, we proved that this theoretically predicted intermediate-energy BChl *a* subantenna in *Osc. trichoides* chlorosomes connecting B750 and B805-860 ones does exist.

This work was supported by the Russian Foundation for Basic Research.

STRUCTURAL PERTURBATIONS OF LONGITUDINAL AND LATERAL CONTACT SURFACES OF TUBULINS INDUCED BY INTERACTION WITH MICROTUBULE STABILIZING COMPOUNDS

A. Y. NYPORKO, Y. B. BLUME

Activity of many compounds, which are used as effective antitumor drugs, are conditioned by their ability to irreversibly stabilize the microtubules via high-affinity binding to tubulin molecules. Among these compounds, the taxane and epothilone derivates are the most studied and applicable in clinical practice. These substances share the same binding site that is located immediately near one of two lateral contact surfaces of β-tubulin molecule, but their behavior in binding pocket is characterized by individual features [1,2]. But one should mention, that structural basis of effectiveness of taxanes and epothilones, as well as other microtubule stabilizing compounds, are unstudied so far, in contrast to compounds with microtubule depolymerizing activity, which as rule bind to longitudinal contact interfaces and directly prevent the tubulin polymerization [3,4]. Thus, it seems reasonable to investigate the structural changes in tubulin functional surfaces due to interaction with microtubule stabilizing compounds.

In our work the normal heterodimer αβ-tubulin of *Eleusine indica* (its spatial structure was reconstructed by us earlier [5]) and its complexes with taxol and epothilone A have been studied. Spatial structure optimization and molecular dynamics of complexes were calculated using the *mdrun* module of GROMACS program package. Computation of molecular dynamics was carried out for 30 ns time interval. Structural changes were estimated by conformational energy dynamics (using *g_energy* module) and level of molecular oscillations (using *g_rms* module) of amino acid residues that formed longitudinal and lateral contact surfaces of tubulin subunits.

It was revealed that binding of both taxol and epothilone A results in stable decreasing of average level of conformation energy of contact surfaces at whole

Institute of Cell Biology and Genetic Engineering of National Academy of Science of Ukraine, acad. Zabolotnogo str., 148, Kiev, 03143, Ukraine, dfnalex@gmail.com

(~1000 kJ/mol for both complexes) as well as proteins at whole (400 kJ/mol for complex with taxol, 812 kJ/mol for complex with epothilone A). However, energy changes for individual surfaces are more complicated. So, energy reducing owing to binding of both ligands is observed for both β-subunit lateral contacts, for remote α-subunit lateral contact and for α-subunit longitudinal minus-end contact. An energy of near α-lateral contact surface reduces due to binding of epothilone only. The energy level of α-plus-end longitudinal contact surface slightly decreases in response to taxol binding, but essentially raises after epothilone A binding. The most unexpected changes occur in β-longitudinal contacts: taxol increases the energy of β-minus-end contact, and both compounds rise the energy of β-plus-end contact surface. It can be assumed that local energy increasing of β-plus-end surface may disturb structure of β-tubulin catalytic center and prevent hydrolysis GTP to GDP, that is obligatory condition of microtubule dissociation. But to finally answer to this question the development and analysis of more sophisticated multidimeric model is needed. The patterns of changes of molecular oscillations for each of contact surface in general correlate well with appropriate energy changes, with the exception of α-plus-end surface, where both ligands increase the oscillation level.

Thus, we can conclude that main effect of interaction of taxol and epothilone A with tubulin heterodimer shows itself in a number of distal structural changes of both tubulin subunits and an increase of general stability level of contact interfaces.

1. Snyder J.P., Nettles J. H., Cornett B., Downing K. H., Nogales E. (2001) The binding conformation of taxol in β-tubulin: a model based on electron crystallographic density *Proc Natl. Acad. Sci. USA*. **98**: 5312-5316
2. Nettles J.H., Li H., Cornett B., Krahn J.M., Snyder J.P., Downing K.H. (2004) The binding mode of epothilone A on alpha,beta-tubulin by electron crystallography. *Science.* **305**:866-9.
3. Blume Ya.B., Nyporko A.Yu., Yemets A.I., Baird W.V. (2003) Structural modelling of plant α-tubulin interaction with dinitroanilines and phosphoroamidates. *Cell Biol. Int.* **27:** 171-174.
4. Delye C., Menchari Y., Michel S., Darmency H. **(**2004) Molecular bases for sensitivity to tubulin-binding herbicides in green foxtail. *Plant Physiol* **136**: 3920–3932
5. Nyporko A. Yu., Blume Ya. B. (2005) Analysis of changes of tubulin functional surfaces that due to the interaction with GTP. *Proceeding of the International Moscow Conference of Computational Molecular Biology* (July 18-21, 2005, Moscow, Russia) – Moscow, 2005 – P. 262-263.

TRANSCRIPT DIVERSITY AT THE EXTREMES: ANALYSES OF ALTERNATIVE TRANSCRIPTION INITIATION AND TERMINATION

UWE OHLER

Similar to the arrival of large EST libraries which led to an explosion in the number of observed splice isoforms, recent high-throughput 5' end sequencing has seen a large increase in the number of genes with alternative start sites. As in splicing, the question is how much of this variation is due to (a) experimental noise; (b) biological noise; and (c) condition-specific functional variability. We have built a pipeline to infer alternative transcription start sites in /D. melanogaster,/ and will present results of our ongoing analysis. Among the questions we currently address are: Are alternative TSS utilized under different conditions? What are the sequence-specific determinants and characteristics of alternative TSS? What are the consequences on the isoform, i.e. do alternative TSS lead to changes in the protein and/or the 5' UTR sequence?

Regulation of transcript variation is also related to the usage of alternative cleavage/poly-adenylation sites, leading to longer or alternative terminal exons. The 3' untranslated portion of the terminal exon is the primary target area of post-transcriptional regulatory mechanisms, for instance of regulation by small RNAs: Animal miRNAs preferentially target complementary sites located in the 3'UTR of specific mRNAs. We have constructed a set of reliable alternative 3'UTR regions in the human genome, allowing us to study the distribution of miRNA target sites within (alternative) UTRs, with the result that more than 40% of target sites lie in alternative UTR segments which will not always be part of the mature transcript. This means that the miRNA-target gene relationship is dynamic in nature, and that it depends on the condition-specific presence of a particular isoform of its target gene.

Institute for Genome Sciences and Policy, Duke University, CIEMAS Bldg, 101 Science Dr, Box 3382, Durham, NC 27708

COMPARATIVE ANALYSIS OF TRINUCLEOTIDE REPEATS IN MAMMALIAN GENOMES

NINA OPARINA[1], MARINA FRIDMAN[2], VSEVOLOD MAKEEV[2]

Simple sequence repeats (SSRs) or microsatellites are frequent in genomes of species from all living kingdoms. Higher eukaryotes contain large fractions of SSRs in their genomes, differing in monomer type, length and copy number. Among all microsatellites, trinucleotide ones are of permanent interest due to their frequent occurrence in protein-coding genes and their role in disease-causing expansions. In mammals, microsatellites with trinucleotide monomers are found in genes associated with neurological disorders, such as fragile X syndrome, Huntington's disease and several forms of ataxia, and Myotonic dystrophy. DNA replication slippage and/or unequal recombination at these sites produce enlarged microsatellites causing the appearance of long monoamino-acid blocks in encoded proteins. Trinucleotide expansions, their mechanisms and consequences, are actively studied in several groups. In contrast to it, trinucleotide SSRs in non-coding regions of mammalian genomes were not analysed thoroughly.

Many of known disease-causing expansions are associated with polyglutamate-encoding (CAG)n repeats, relatively rare in genome. Whole genome frequencies of perfect trinucleotide SSRs with copy number >=12 shows that the most frequent repeats are AT-rich AAA, TTT, AAT, ATT and others. We have analysed human, mouse, rat and dog genomes and found certain differences in frequencies of trinucleotide SSRs with different monomer types. In most cases the frequency of triplet repeat positively depended on its composition and correlated with genome GC-content. In dog genome GC-rich trinucleotide repeats were often in contrast to human and rodent genomes. Surprisingly, CCC/GGG motifs were only relatively frequent GC-rich trinucleotide SSRs found in rodents and dog genomes but not in human genome. We have divided trinucleotide SSRs into subgoups: "new" ones, consisting of tandems with 100%ID and no indels and "elder ones" containing tandems with no indels but with 95-99%ID between monomers. Only repeats located in non-coding regions were analysed. The following findings were common for all studied species: 1) "new" SSRs occupy only short regions (mostly 8-10 copies, not more than 18 copies); 2) "elder" SSRs were longer with main fraction of tandems with 10-15 copies and a "long tail" of repeats with >18 copies; 3) there were profound differencies

[1] Engelhardt Institute of Molecular Biology, Moscow, Russia, oparina@gmail.com
[2] GOSNIIGenetika, Moscow, Russia, makeev@genetika.ru

in monomer types between "new" and "elder" SSRs: the fraction of AT-repeats decreased in "elder" SSRs in contrast to growing fraction of GC-repeats; 4) CpG-containing repeats were almost absent in all types of SSRs; 5) trinucleotide repeats known to be frequently connected to disease-causing expansions were rare both in "new" and "elder" SSRs. We also analysed subfractions of "new" SSRs consisting of short (<10 copies) and long (>11 copies) repeats. We proposed that short perfect trinucleotide SSRs could be the direct results of polymerase slippage or other mechanisms causing the tandem repeats origin in contrast to longer of imperfect ones containing traces of subsequent events. We have analysed syntheny regions and revealed that SSR monomer types were more similar between species for short "new" SSRs than for long "new" SSRs. Thus we concluded that trinucleotide tandem repeats originated due to highly similar mechanisms in all studied species.

INTEGRATED DATABASE OF HUMAN CIS-ANTISENSE GENE PAIRS

YURIY L. ORLOV, JIANGTAO ZHOU, VLADIMIR A. KUZNETSOV

Experimental studies and computational analysis of genome sequence reveals significant abundance of pairs of genes transcribed from opposite strand of the same locus and corresponding sense-antisense (SA) transcripts in several eukaryotic genomes including the human (Katayama et al., 2005), affecting up to 20% of genes in mammalian genomes (Chen et al, 2004). Recent works reported computer pipelines and databases collecting available information on SA transcripts (Engstrom et al., 2006; Zhang et al., 2006).

However, there is a little known about the gene expression patterns, regulation mechanisms and biological functions of these SA gene pairs in different cell types, in particular in cancer cells. To make up-to-date support for antisense gene expression analysis we created USAP (Unites Sense-Antisense Transcripts Pairs) database integrating genome data including manually curated compilations, published database sources on SA pairs, and clinical data on breast cancer based on Affymetrix U133A&B microarray platform. Our analysis revealed that SA pairs sharing nucleotides on the same chromosome loci can be grouped in about 8857 overlapping SA sequence clusters (based on GenBank mRNA annotation of February 2006). This number is at least two times greater than published before (Zhang et al., 2006). Due to the multiple occurrence of some sequences in more than one of the SA cluster overlaps, a

Genome Institute of Singapore, 60 Biopolis St. #02-01, Genome, 138672 Singapore, orlovy@gis.a-star.edu.sg, kuznetsov@gis.a-star.edu.sg

fraction of these SA cluster overlaps can be united into about 6301 joined SA chromosome territories including 329 "chains" with non-redundant sequences covering ~20% of whole chromosomes size. Up to 30% of human protein coding genes (defined by RefSeq IDs) have overlapped antisense transcript. Finally, we integrated into our USAP database SAGE (Serial Analysis of Gene Expression) data for the human genome, which contains 20316 short 17-mer tags (Ge et al, 2006). We found that 1271 SA cluster overlaps contain at least one SAGE tag thus proving SA transcription in antisense strand.

Data analysis of USAP DB supports the hypothesis that positive co-regulation of genes may be the dominant scenario for cis-antisense pairs. Using 249 Affymetrix microarrays from human breast cancer samples, Kuznetsov and co-authors (2006) have found that significant positive co-regulated pairs strongly dominate over negatively co-regulated pairs among different cell types including cancer cells. In particular, we identified over-expression and common positive co-expression SA transcripts in human breast cancer tissues. This pattern is reproducibly expressed in different breast cancer types. We also found specific positive co-expression SA transcript pairs which are uniquely associated with low- and in high-aggressive types of breast cancer cells. Thus, USAP database provides important resource of novel structural and functional elements in the human genome and could be used in evaluation of potential role of SA gene pairs and its transcripts in normal and pathological cells.

1. S. Katayama et al. (FANTOM Consortium) (2005) Antisense transcription in the mammalian transcriptome, *Science*, **309** (5740), 1564-1566.
2. J. Chen et al. (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res,* **32**: 4812-4820.
3. P.Engstrom et al. (2006) Complex Loci in human and mouse genomes, *PLoS Genet*. **2**(4): e47.
4. Y. Zhang et al. (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species, *Nucleic Acids Res*, **34**: 3465-75.
5. X. Ge et al. (2006). A large quantity of novel human antisense transcripts detected by LongSAGE, *Bioinformatics*, **22**(20): 2475-9.
6. V.A. Kuznetsov et al (2006) Genome-wide co-expression patterns of human cis-anti-sense gene pairs, In: *Proceeding of BGRS'2006, Novosibirsk*, **1**: 90-93.

DNA ELECTROSTATIC POTENTIAL DATABASE

ALEXANDER A. OSYPOV[1], PETR M. BESKARAVAINY[1],
SVETLANA G. KAMZOLOVA[1], ANATOLY A. SOROKIN[2]

The aim of the Database is to hold and provide all the available information about the electrostatic properties of model objects DNA together with comprehensive annotation of their sequences.

A large number of bacterial and bacteriophages promoters have been sequenced and some sequence-specific determinants involved in promoter functioning have been found. Despite this wealth of information about the sequence structure it is difficult from sequence data alone to pinpoint potential promoter sequences in genome DNA or to predict their strength and other functional characteristics. A large set of promoter search algorithms based on sequence specific recognition elements has failed in correct prediction of promoter sites in genome. It has recently been formulated that some additional information for promoter recognition can be coded in physical properties of DNA helix. Some physico-chemical characteristics of promoter DNA such as overall geometry, deformability, thermal instability and dynamical features have been shown to play an important role in modulating promoter activity. We have suggested a new approach to this problem based on analysis of electrostatic properties of promoter DNA (1).

We have developed a simplified method for calculation of electrostatic potential distribution for long DNA fragments as large as complete genomes. Using this method electrostatic properties of some genomes have been studied. Electrostatic interactions between promoter DNA and RNA polymerase have been shown to be of considerable importance in regulating promoter function (2). Electrostatic patterns of promoter DNA can be specified due to the presence of some distinctive motifs which differ for different promoter groups and may be involved as signal element in differential recognition of the corresponding promoters by the enzyme (1).

Given that we decided to develop the DNA Electrostatic Potential Database to hold and provide all the available information about the electrostatic properties of model objects DNA together with comprehensive annotation of their sequences. The Database is available for academic use at http://promodel.icb.psn.ru.

[1] Institute of Cell Biophysics of RAS, Pushchino, Moscow region, Russia, ao@icb.psn.ru
[2] The University of Edinburgh, Kings Buildings, Edinburgh, EH9 3JR, UK, asorokin@inf.ed.ac.uk

Genome sequences and their annotation are taken from EBI Genome Reviews (http://www.ebi.ac.uk/GenomeReviews) (3) or in some cases directly from literature. Also if available annotation data is taken from BioCyc (http://BioCyc.org) (4), RegulonDB (http://regulondb.ccg.unam.mx) (5) and some other DBs. Manually curated annotation is continuously added taken from different literature sources. The electrostatic potential around the double-helical DNA molecule is calculated by the original method (6) using the computer program of Sorokin, A. (7). All data provided is referenced and linked to its sources.

Further goals are the expansion of the Database to cover even more model objects, development of some tools for computational analysis of electrostatic properties of DNA and their integration into regulation prediction systems thus increasing their reliability.

The authors are grateful to Saveljeva E. G. for technical support.

1. A.A.Sorokin, A.A.Osypov, T.R.Dzhelyadin, P.M.Beskaravainy, S.G.Kamzolova. (2006) Electrostatic properties of promoter recognized by E. coli RNA polymerase Esigma70, *J Bioinform Comput Biol*, **4(2):** 455-467.
2. S.G.Kamzolova, V.S.Sivozhelezov, A.A.Sorokin, T.R.Dzhelyadin, N.N.Ivanova, R.V.Polozov (2000) RNA polymerase-promoter recognition. Specific features of electrostatic potential of "early" T4 phage DNA promoters, *J. Biomol. Struct. Dyn.*, **18(3):** 325-334.
3. P.Sterk, P.J.Kersey, R.Apweiler (2006) Genome Reviews: standardizing content and representation of information about complete genomes, *OMICS*, **10(2):** 114-118.
4. P.D.Karp et al, (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, *Nucleic Acids Res*, **33(19):** 6083-6089.
5. H Salgado et al. (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions, *Nucleic Acids Res*, **34:** D394-397.
6. R.V.Polozov, T.R.Dzhelyadin, A.A.Sorokin, N.N.Ivanova, V.S.Sivozhelezov, S.G.Kamzolova (1999) Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences, *J. Biomol. Struct. Dyn.*, **16(6):** 1135-1143.
7. A.A.Sorokin (2001) Functional analysis of E. coli promoter sequences. New promoter determinants, *Ph.D. Thesis. Pushchino, Institute of Theoretical and Experimental Biophysics RAS.*

COMPUTATIONAL APPROACH TO THE ANALYSIS OF THE PROPERTIES OF ELECTROSTATIC POTENTIAL PROFILE OF GENOME DNA

ALEXANDER A. OSYPOV[1], VALERY V. PANJUKOV[2]

Computational approach to the analysis of the properties of electrostatic potential profile of genome DNA is developing. Even the first attempt revealed considerable score in the formal metrics based on the frequency to amplitude ratio of the potential profile at the promoter regions compared to its distribution for the whole genome sequence, which opens the potential opportunity to formal computational annotation of the specific regions in the newly sequenced genomes.

A large number of bacterial and bacteriophages promoters have been sequenced and some sequence-specific determinants involved in promoter functioning have been found. Despite this wealth of information about the sequence structure it is difficult from sequence data alone to pinpoint potential promoter sequences in genome DNA or to predict their strength and other functional characteristics. A large set of promoter search algorithms based on sequence specific recognition elements has failed in correct prediction of promoter sites in genome. It has recently been formulated that some additional information for promoter recognition can be coded in physical properties of DNA helix. Some physico-chemical characteristics of promoter DNA such as overall geometry, deformability, thermal instability and others have been shown to play an important role in modulating promoter activity. An approach to this problem based on analysis of electrostatic properties of promoter DNA have been suggested (1).

Studies of electrostatic properties of some genomes using a simplified method for calculation of electrostatic potential distribution for long DNA fragments as large as complete genomes revealed the considerable importance of electrostatic interactions between promoter DNA and RNA polymerase in regulating promoter function (2). Electrostatic patterns of promoter DNA can be specified due to the presence of some distinctive motifs which differ for different promoter groups and may be involved as signal element in differential recognition of the corresponding promoters by the enzyme (1). Even more important is the discovered nonlinearity of the dependency of potential profile on the sequence, meaning that this property is vastly dependent on the whole se-

[1] Institute of Cell Biophysics of RAS, Pushchino, Moscow region, Russia, ao@icb.psn.ru
[2] Institute of Mathematical Problem of Biology of RAS, Pushchino, Moscow region, Russia, panjukov@impb.psn.ru

quence with flanking regions rather then the sequence text at the given point of consideration (1).

Although visual analysis of promoter regions revealed some characteristic properties of them, it is lacking in formality and is insufficient to the analysis of the large-scale data, available in the modern era of computational genomics. Especially one of the main properties recognized for the promoters, i.e. extreme peaks and valleys of potential compared to more or less even though somehow rough profile of the coding regions, attracts special attention and demands a formalization of some kind. To realize this intention we developed some algorithms and tools based on them and applied them to selected model objects. Even the first attempt revealed considerable score in the formal metrics based on the frequency to amplitude ratio of the electrostatic potential profile in the promoter regions compared to its distribution for the whole genome sequence. It's worth noting that this property can't be handled by common statistics considering only values of the potential especially after its smoothing, that facilitates its visual analysis, nor can it be extracted by the textual analysis of the sequence, given the indeterminancy of the electrostatic potential by its immediate literal composition.

The data obtained is held and provided in the DNA Electrostatic Potential Database together with all the available information about the electrostatic properties of model objects DNA together with comprehensive annotation of their sequences. The Database is available for academic use at http://promodel. icb.psn.ru.

Genome sequences are taken from EBI Genome Reviews (http://www.ebi. ac.uk/GenomeReviews). The electrostatic potential around the DNA molecule is calculated using the modified computer program of Sorokin, A. (3).

Further goals are the elaboration of the algorithms and tools for computational analysis of electrostatic properties of DNA and their integration into regulation prediction systems thus increasing their reliability.

1. A.A.Sorokin, A.A.Osypov, T.R.Dzhelyadin, P.M.Beskaravainy, S.G.Kamzolova. (2006) Electrostatic properties of promoter recognized by E. coli RNA polymerase Esigma70, *J Bioinform Comput Biol*, **4(2):** 455-467.
2. S.G.Kamzolova, V.S.Sivozhelezov, A.A.Sorokin, T.R.Dzhelyadin, N.N.Ivanova, R.V.Polozov (2000) RNA polymerase-promoter recognition. Specific features of electrostatic potential of "early" T4 phage DNA promoters, *J. Biomol. Struct. Dyn.*, **18(3):** 325-334.
3. A.A.Sorokin (2001) Functional analysis of E. coli promoter sequences. New promoter determinants, *Ph.D. Thesis. Pushchino, Institute of Theoretical and Experimental Biophysics RAS.*

RELIC TRANSPOSONS AND THE IMMUNOLOGICAL BIG BANG: THE IDENTIFICATION OF INVERTEBRATE MOBILE ELEMENTS SIMILAR TO HUMAN RAG1 GENE

YURI V. PANCHIN[1], LEONID L. MOROZ[2]

Animal genomes contain about 25000 genes. In addition millions of genes for antigen receptors are generated in cells of immune system by mechanism known as V(D)J somatic recombination from the sets of separate gene segments [1]. Although only jawed vertebrates possess this mechanism, another type of DNA rearrangement called transposition by which sequences of DNA (transposons) can move around to different positions within the genome of a single cell is common in different organisms. The components of V(D)J recombination system, Recombination-Activating Gene proteins (RAG1 and RAG2) and, recombination signal sequence (RSS), are thought to have "entered" the vertebrate genome by horizontal transfer as components of a transposable element (hypothetical "RAG transposon")[1-5]. The imaginary event of acquisition V(D)J recombination by jawed vertebrates via such transposon was dubbed the "Immunological Big Bang"[2,6]. Recently identified transposons, Transib [7], and NRAGTP (found in mollusk Aplysia and reported here) have terminal inverted repeats (TIRs) similar to RSS and putative transposase homologous to two different parts of RAG1 protein. Transib encode protein similar to C-terminal part of RAG1 and NRAGTP encode protein similar to N-terminal part of RAG1. These findings support and refine "RAG transposon" hypothesis and allow us to propose the scenario of V(D)J recombination machinery evolution. Although N- terminal region of the human RAG1 protein is dispensable for recombination it's Aplysia homolog appears to be a transposase. This disparity encourage the reassessment of RAG1 N- terminal domain role in V(D)J recombination.

Is NRAGTP a "selfish" DNA, genomic parasite, aimed only to its own reproduction or it has some function in Aplysia physiology and development? Transposons are not usually expressed in somatic tissues, whereas we show that NRAGTP RNA is expressed and NRAGTP transposition produces some

[1] Institute of Problems of Information Transmission, Russian Academy of Science, Moscow, & A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow 9, Russia, ypanchin@yahoo.com
[2] The Whitney Laboratory for Marine Bioscience, Evelyn F. & William McKnight Brain Institute of the University of Florida, Florida 32080, USA, moroz@ whitney.ufl.edu

genome rearrangements in somatic tissues. We speculate that similar to vertebrate adaptive immunity NRAGTP activity in mollusk may lead to GOD (Generation Of Diversity), creating diversity in genomic DNA and serving yet unknown function.

Acknowledgments: supported by RFBR grant 05-04-48401

1. H. Sakano H, et al.. (1979) Sequences at the somatic recombination sites of immunoglobulin light-chain genes.Nature.;280: 288-94.
2. D.G. Schatz (2004) Antigen receptor genes and the evolution of a recombinase. Semin Immunol. 4: 245-56.
3. M.A. Oettinger et al. (1990) RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science;248:1517–1523.
4. C.B.Thompson (1995) New insights into V(D)J recombination and its role in the evolution of the immune system. Immunity;3:531–3539.
5. S.D. Fugmann et al. (2006) An ancient evolutionary origin of the Rag1/2 gene locus. Proc Natl Acad Sci U S A. ;103:3728-33.
6. R.M. Bernstein et al. (1996) Proc Natl Acad Sci U S A. ;93:9454-9. Primordial emergence of the recombination activating gene 1 (RAG1): sequence of the complete shark gene indicates homology to microbial integrases.
7. V.V. Kapitonov, J. Jurka (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. 6:e181.

HUMAN "TRASH EST" STUDY

ALEXANDER Y. PANCHIN[1], SERGEY A. SPIRIN[2],
YURI V. PANCHIN[3], SERGEY A. LUKYANOV[4], YURI B. LEBEDEV[5]

Expression sequence tags (EST) are 500–1000 bp length sequences that represent RNA molecules derived from different sources (cell lines, tissues etc.). The human EST database contains over 8,000,000 sequences, with over 4,000,000,000 total letters. Normally RNA is synthesized from a genomic DNA matrix and therefore all EST's should contain sequences matching genomic sequences, with the exception of certain gaps, which result of mRNA splicing. Nevertheless, we found around 11000 EST in the human EST database, whose sequences have no match to sequences of the human genome database. The presence of these Trash EST's (TEST's) in the EST database could be a result of laboratory equipment, tissue, cell line or cloning contaminations (as E.coli or Bacteriophage lambda), but also TEST's could represent sequences from unidentified genes, as well as RNA sequences of unknown sources. Here we show the results of TEST analysis, as an attempt to identify the sources of human EST database contaminations.

Human TEST analysis shows that sequences from all different organism sources, i.e., viruses, bacteria, fungi, plants and mammals are presented in the EST database. Around 70% of those sequences derive from totally unknown sources. 12% of TEST's are bacterial sequences, a significant part of them derive from E.coli, thus representing laboratory contaminations. Yet some of those sequences are rather different from sequences of known bacteria, suggesting that they could belong to unknown species of prokaryotes, perhaps dwelling in human tissues. 4% of TEST's are sequences of bacteriophage genomes, laboratory contaminations as well. About 7% TEST's are very similar to mammalian sequences. The latter are the primary candidates for the human

[1] Faculty of Bioengineering and Bioinformatics of Moscow State University, Moscow, Russia, alexpanchin@yahoo.com
[2] A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, Russia, sas@belozersky.msu.ru
[3] Institute of Problems of Information Transmission, Russian Academy of Science, Moscow, & A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, Russia, ypanchin@yahoo.com
[4] Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia, Russian Academy of Sciences, luk@evrogen.ru
[5] Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia, lebedev_yb@ibch.ru

lost genes missing from the current genome assemblies, yet it appears that many of those sequences are actually sequences of other mammals, annotated as human by human mistake. 1% of TEST's seem to be sequences of some known and some yet unknown eukaryotic viruses.

We report interesting findings in the human EST database: some EST's, presented by different laboratories have sequences very similar to mRNA sequences of known cultural plants. These sequences encode plant-specific proteins such as chlorophyll a-b binding protein and ribulose bisphosphate carboxylase (RuBisCo) small subunit. Such EST's where found in EST databases of other mammals (mouse, dog, cow and rat). Some of those sequences were considered as murine bona fide mRNA and full length sequences deposited in mice NR and UniGene databases. We suggested that the occurrence of those sequences in the EST database may not be a simple contamination. The latter idea is supported by mouse expression array data, showing strong expression of RuBisCo mRNA in mouse tissues. Homological plant specific EST's have been annotated by different working groups, and the overall amount of the two listed types of EST's is rather high, comparing to average amount of EST's found for specific sequences. To explain those phenomena, we suggest a hypothesis that RNA from other species could migrate into human cells from food, and discuss this possibility within its probable similarity to double-stranded RNA (dsRNA) transport, described in C.elegans. We also support that hypothesis with the fact that mammals and worms have common channels designed for RNA transfer.

Acknowledgments: supported by RFBR grant 05-04-48401

EVOLUTIONARY ALGORITHM FOR PHYLOGENETIC TREE CONSTRUCTION

N.PERDIGÃO[1], D.MIGOTINA[1], A.ROSA[1]

Introduction. The similarity of molecular mechanisms of the organisms that here been studied strongly suggests that all organisms on Earth had a common ancestor. Thus any set of species is related and this relationship is called a phylogeny. Usually the relationship can be represented by a phylogenetic tree. Phylogeny is very important in Multiple Sequence Alignment (MSA) of sets of sequences, where the evolutionary relationship must be taken in account [1]. Example where this relationship is already taken in account is in progressive alignment algorithms [2,3,4].

[1] Instituto Superior Tecnico, Av. Rovisco Pais 1049-001 Lisboa, Portugal
p3rdigao@isr.ist.utl.pt, acrosa@isr.ist.utl.pt

Methods and DataSets. The intention of this work is the construction of phylogenetic binary trees using an Evolutionary Algorithm (EA) [5] as tool. The codification made in the chromosomes of the EA is:

A phylogenetic tree;

The α parameter that defines the Gamma distribution for models of replacement;

The K parameter that represents the Expected transition/transversion ratio;

The R parameter that represents the pyrimidine-transition/pruine-transition.

The EA have only one crossover operator that uses always two parents and return as offspring one or two children. The steps of the crossover are:

Selection of an interior branch of the father;

Creation of a temporary tree based on the previous excerpt;

Removal of all the elements present in the temporary tree of the mother;

Insertion of the temporary tree in a random branch of the mother.

The datasets used are 10, 20, 30, 40 and 55 sequences sets from rbcL[6].

Results and Conclusions: The results obtained by our program in terms of score are similar to phyML[7] and to TreePuzzle[8]. In terms of time phyML is the fastest.

1. D.Sankoff et al (1993) Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Addison Wesley.
2. D.G.Higgins et al (1992) ClustalV: improved software for multiple sequence alignment, CABIOS, 8: 189-191.
3. J.D.Thompson et al (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res., 22: 4673-4680.
4. J.D.Thompson et al (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res., 25: 4876-4882.
5. J.H.Holland (2003) Adaptation in Natural and Artificial Systems, MIT PRESS, Cambridge, Massachusetts.
6. P.O.Lewis (1998) A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data, Journal of Molecular Evolution, 15: 277–283.
7. S.Guidon et al. (2003) A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst Biol, 52(5): 696–704.
8. H.A.Schmidt (2002) Tree-Puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. BioInformatics, 18(3): 502–504.

EVOLUTION OF CPG ISLANDS IN MAMMALIAN GENOMES

I.M. PERTSOVSKAYA[1], A.A. MIRONOV[1]

Cytosine methylation is the only covalent DNA modification described in mammals. The cytosine methylation plays an important role in the regulation of gene expression and controls the genome stability in higher eukaryotes. Genetic studies have established that this epigenetic mark is required for embryonic development, genomic imprinting and X-chromosome inactivation. Alterations in DNA methylation are related to many human diseases, including cancer (1). DNA methylation is essential for the normal development of most multicellular organisms. The methylation status of promoter DNA sequences correlates with the transcriptional activity of genes (1). In mammals, methylation is restricted to CpG dinucleotides, which are largely depleted from the genome except at short genomic regions called CpG islands (2, 3). CpG dinucleotides are vastly underrepresented genome-wide compared to what would be expected by chance.

CpG islands are defined as genomic regions longer then 200 b.p. with an increased number of CpG dinucleotides and C+G ratio compared to the rest of the genome (4) . There are about 27000 annotated CpG islands in the human genome and 16000 annotated CpG islands in the mouse genome. In the human genome they cover about 0.7% of the DNA sequence and contain about 7% of all genomic CpG dinucleotides. Unmethylated CpG islands sometimes are found in the first exons and promoters of housekeeping and tissue-specific genes.

In this study we analyzed homology and genome position of CpG islands in human, mouse and dog. Using a compiled database, we found CpG islands associated with genes and their upstream regions. Only a small fraction of CpG islands are conserved in the upstream regions of ortologous genes of all three organisms. Non-linear correlation of the upstream region length and the number of CpG islands found in this region was detected. In the human genome, the largest number of these CpG islands were found in the 5000 b.p. upstream gene region. Despite the fact that the number of annotated CpG islands in the dog genome is about twice larger than in the human genome, the dog genome has tenfold less CpG islands in the 15000 b.p. upstream gene regions than the human genome.

Using the BLAT program we aligned all human CpG islands. We discovered about 100 CpG islands that are nearly identical to other human CpG islands

[1] Moscow State University, Vorobievi gori 1 – 73, Moscow, Russia,
inna.perts@gmail.com

and not located in annotated duplication regions. Some of them form homologous groups that contains several CpG islands. We discovered that about 10% homologous CpG islands are associated with genes. Four out of ten such CpG islands are associated with genes of the Frizzled family proteins and 80-90% conserved, although the coging regions of these genes are share between 20% and 40% sequence similarity by sequence. Also these CpG islands are similar not only with another associated with Frizzled gene CpG island, but also with not associated with genes CpG islands.

Another area of our study was investigation of the evolution of CpG dinucleotides in rodents (mouse and rat) using human as an outlier. Using triple alignments we examined positions immediately preceding guanine. We analyzed about 45 million positions outside CpG islands and 270 thousand positions in CpG islands. Our results demonstrate that mutations notCpG→CpG are rarer then mutations CpG→notCpG both in and outside CpG islands. In addition mutations of cytosine in CpG islands are rarer than mutations in other regions of the genome. These results are consistent with previous studies of the dependence of SNP and mutation rates on the local GC content (5, 6).

1. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet.; 39(4):457-66. (2007)
2. Gardiner-Garden M, Frommer M CpG islands in vertebrate genomes. J Mol Biol. 20 196(2):261-82 (1987).
3. Francisco Antequera, Adrian Bird Number of CpG islands and genes in human and mouse. Genetics Vol. 90, pp. 11995-11999 (1993)
4. Fazzari MJ, Greally JM Epigenomics: beyond CpG islands. Nat Rev Genet. 5(6):446 55 (2004)
5. Zhongming Zhao, Fengkai Zhang Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. Gene 366 316–324 (2006)
6. Gu J, Li WH. Are GC-rich isochores vanishing in mammals? Gene. 30; 385:50-6 (2006)

AN EVIDENCE FOR REGULATION OF SPLICING BY RNA SECONDARY STRUCTURES: CONSERVED COMPLEMENTARY MOTIFS IN DROSOPHILA INTRONS

DMITRI PERVOUCHINE[1], ANDREI MIRONOV[1]

The instances of RNA secondary structures that influence splicing, both constitutive and alternative, have been reported in various organisms. These include yeast, flies, vertebrates and even viruses [1-4]. Whether or not the secondary structure of pre-mRNA is, in general, an important factor for splicing efficiency has been debatable for years (see [5] for review). The current opinion is that it could be critical in several cases, whilst the majority of splicing pathways is weakly dependent on RNA secondary structures.

We used comparative genomics approaches to look for potential secondary structures in Drosophila introns. Eleven sequenced species of fruit fly were analyzed, yielding approximately 150 genes which have highly conserved complementary motifs in regions surrounding splice sites. Our findings suggest that these complementary motifs may mediate splicing by loop-out mechanism, which takes place, for instance, in the processing of DSCAM gene, where complementary interactions are responsible for mutually exclusive choice of exons [6].

Our list of predictions is enriched with alternatively spliced genes, suggesting a role of RNA secondary structures in regulation of alternative splicing. We find several cases when one motif is complementary to more than one target, which could be responsible for mutually exclusive choice of exons, as in the DSCAM case. A substantial part of our list consists of developmental genes and genes pertinent to the nervous system. All our predictions are highly significant with respect to various controls and background probabilistic models.

We suggest that regulation of splicing by RNA secondary structures is not limited to a couple of dozens of isolated cases that have been reported in literature. We believe that it is much more widespread than previously appreciated. In this work, we present evidence that RNA secondary structure could be a primary factor for determining splicing pattern of a gene, and a labile control element responsible for regulation of alternative splicing.

[1] Dept of Bioengineering and Bioinformatics, Moscow state University, Moscow, Russia, pervouchine@inbox.ru, mironov@bioinf.fbb.msu.ru

1. H. Halfter and D. Gallwitz. (1988) Impairment of yeast pre-mRNA splicing by potential secondary structure-forming sequences near the conserved branchpoint sequence. Nucleic Acids Res, 16(22):10413-10423.
2. Y. Chen and W. Stephan. (2003) Compensatory evolution of a precursor messenger RNA secondary structure in the Drosophila melanogaster Adh gene. Proc Natl Acad Sci USA, 100(20):11499-11504.
3. N. N. Singh, R. N. Singh, and E. J. Androphy (2006). Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. Nucleic Acids Research, doi:10.1093/nar/gkl1050
4. J. Kraunus, D. Zychlinski, T. Heise, M. Galla, J. Bohne, and C. Baum. (2006) Murine Leukemia Virus Regulates Alternative Splicing through Sequences Upstream of the 5' Splice Site. J Biol Chem, 281(49):37381-90.
5. E. Buratti and F. E. Baralle. (2004) Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. Molecular and Cellular Biology, 24(24):10505-10514.
6. B. R. Graveley. Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. Cell, 123(1):65{73, 2005.

TEXTURE ANALYSIS FOR IMAGING IN SYSTEMS BIOLOGY

LEONID PESHKIN

It is difficult to overestimate the importance of automated segmentation for microscopy images of cells. Practically every biological lab conducting research in cell biology or systems biology is using phase and fluorescent microscopy as a primary way to observe cell cycle events and to track localization and expression levels of individual genes and gene ensembles. A typical lab can collect hundreds or thousands of images daily. Currently available image processing software does not provide any functionality for segmenting and tracking individual cells. As a result researchers either have to average measurements over a population, or they are forced to perform mundane and extremely laborious manual segmentation which could not produce sufficient data for statistically significant analysis. A recently developed method known as "texture gradient" [1,2] has demonstrated quite promising results for segmenting natural images [2] such as still life, urban and rural landscapes, portraits and pictures of wildlife. The essence of the method is in reducing a variety of textures found in (a class of) images to a

Harvard University, Boston MA, USA, pesha@hms.harvard.edu

key set of "textons". Assuming that an edge is a geometrical place of points separating areas of distinct textures, the method compares the distribution of textons among areas surrounding a point. The more distinct these distributions are, the higher the probability that this point belongs to an edge.

In this work we investigate application of texture gradient methods to a particular domain---grayscale images of cell cultures resulting from microscopy. We are interested in both finding a cell outline and in attributing individual pixels to various parts of cell morphology. The latter part is the key contribution of this paper. Naturally, a boundary is thought of as a contour that represents a change in the pixel membership from one object or surface to another. This is the first step towards an automatic cell detection tool, which we are developing to register events and expression levels along the temporal dimension in movies. Our segmentation and classification results are applied by creating a mask which is superimposed onto fluorescence images that have been concurrently obtained in two channels. Via this method we are able to integrate fluorescent signals corresponding to the expression of two interacting genes (in our case these are p53 tumor suppressor protein and MDM2 ubiquitin-protein ligase) localized to the individual cells or nuclei. The problem of cell segmentation from phase images is so difficult that to the best of our knowledge it has not been addressed directly so far. Several research groups attempted to solve a somewhat simpler problem by reducing the task in various ways. A large body of work on segmentation is mostly concerned with segmenting nuclei rather than cells, since the main goal was to track the expression of genes localized in the nucleus (please see [3] and the references therein). Other researchers base segmentation on auxiliary fluorescence channels or combine fluorescence signal with clues from phase. In contrast, we are concerned with segmenting entire cells solely from phase images, in the interest of leaving other channels to track arbitrary unrelated signals.

1. 1. J. Malik, T.K. Leung, and J. Shi. (2001) Contour and texture analysis for image segmentation. International Journal of Computer Vision, 43(1):7–27.
2. 2. D.R. Martin, C. Fowlkes, and J. Malik. (2004) Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(5):530–549.
3. 3. C. Wahlby. (2003) Algorithms for Applied Digital Image Cytometry. PhD thesis, Uppsala University.

CLASSIFICATION OF MITOTIC ABNORMALITIES FOR AUTOMATED CYTOMETRY

LEONID PESHKIN[1], JOAQUIN GONI[1]

This paper is concerned with automated analysis of images resulting from a genome-wide depletion screening in drosophila cells for genes affecting metaphase mitosis. One of the possible effects of gene depletions is a change in the temporal pattern of mitosis, and consequently---in the distribution of mitotic cells within a well, in addition to phenotypical changes in the mitotic structure of the cell [1]. For example, early failures in mitosis result in the formation of abnormal metaphase spindles which can lead to mitotic delay, and potentially to chromosome missegregation during the ensuing anaphase. We present a computational approach to the automated classification of metaphase mitotic cells based on phenotypic differences. We automatically segment individual cells and detect phenotypic features in raw microscope images. Using such machine learning methods as Naive Bayes, Decision Trees and Support Vector Machines, we perform three-way classification of the mitotic cells into monopolar, bipolar and multipolar. We describe the extraction of basic geometric features from noisy input, the design of derrivative contextual features and the iterative process of supervised machine learning. Several classifiers which achieve human-level accuracy are presented and compared on an original benchmark dataset. This work provides an example of a reproducible computational protocol designed to augment a wet laboratory protocol in order to handle a wide range of cell lines and organisms.

1. G. Goshima, et al. (2007) Genes Required for Mitotic Spindle Assembly in Drosophila S2 Cells. Science, p 1141.

USING MACHINE LEARNING ALGORITHMS TO CLASSIFY DESIGNABLE AND NON-DESIGNABLE BINARY H/P PROTEIN SEQUENCES

MYRON PETO[2], ANDRZEJ KLOCZKOWSKI[12], ROBERT L. JERNIGAN[12]

By using standard Support Vector Machine, Naïve Bayes and other machine learning algorithms we were able to distinguish between two classes of protein sequences: those folding to highly-designable and non-designable protein con-

[1] Harvard University, Boston MA, USA, pesha@hms.harvard.edu
[2] Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011-3020, USA petom@iastate.edu, jernigan@iastate.edu

formations. We generated binary hydrophobic and polar (H/P) sequences that folded to a compact conformation on the 2-D triangular lattice using a specified energy function. Sequences threaded onto a specific lattice conformation with lowest energy were assumed to fold to that conformation. High-designable conformations had many H/P sequences folding to them for a given energy function and low-designable conformations had few H/P sequences folding to them under given energy function. We classified sequences as folding to either high- or low-designable conformations. Because of the necessary disparity between the total number of sequences for high designable and low designable conformations, we randomly selected a subset of the sequences belonging to high-designable conformations. By using several standard machine learning algorithms such as Support Vector Machine, Naïve Bayes, and Decision Tree, we were able to classify the two classes of sequences with high accuracy, over 95% in some cases.

1. Z. Begci, R. L. Jernigan, I. Bahar (2002) Residue coordination in proteins conforms to the closest packing of spheres, Polymer 43: 451-459
2. S. Sun, R. Brem, H. S. Chan, K. A. Dill (1995) Designing amino acid sequences to fold with good hydrophobic cores, Protein Engineering 12: 1205-1213
3. H. Li, R. Helling, C. Tang, N. Wingreen (1996) Emergence of preferred structures in a simple model of protein folding, Science 273: 666-669
4. R. Agarwala, S. Batzoglou, V. Dancik, S. E. Decatur, M Farach, S. Hannenhalli, S. Skiena (1997) Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model, Journal of Computational Biology 4: 275-296
5. A. Kloczkowski, R. L. Jernigan (1997) Computer generation and enumeration of compact self-avoiding walks within simple geometries on lattices, Computational and Theoretical Polymer Science 7: 163-173
6. I. H. Witten, E. Frank (2005) In: Data mining: practical machine learning tools and techniques 2nd Edition, Morgan Kaufmann, San Francisco, 2005

COMPARATIVE GENOMICS OF INTERGENIC SEQUENCES IN ENTEROBACTERIACEAE

MIKHAIL A. PYATNITSKIY

It is well known that noncoding DNA regions may demonstrate high levels of sequence similarity due to evolutionary conservation of an important biological function. While conserved intergenic sequences were extensively studied in eukaryotes, few works considered intergenic regions in prokaryotes. The published studies analyzed noncoding RNA, repeats and periodicity [1-3].

We developed several computational tools for the analysis of conservation patterns in prokaryotic intergenic DNA and applied them to Enterobacteriaceae as pilot case. Our approach includes finding all conserved gene pairs, i.e. pairs of adjacent genes which have corresponding adjacent orthologs in all analyzed species. In order to find orthologs we build a graph, where vertices are joined by an edge if the corresponding genes form a bi-directional best hit. Cliques in this graph represent sets of orthologs (similar to COGs) that are present in all selected genomes. For every pair of cliques we searched for adjacent orthologs that retained their relative orientation. Intergenic regions between selected conserved gene pairs were aligned using MUSCLE. At that the complete sets of orthologous intergenic regions were subject to multiple alignment.

We proposed a simple entropy-based statistics to assess conservation of a single position in a multiple alignment. By smoothing and thresholding this measure we obtain statistically significant high-conservative regions.

We also visualized differences in intergenic regions by plotting 2-D histograms of average identity vs number of columns with gaps in a sliding window. Windows where at least one sequence in the alignment contained a long deletion were excluded. This approach reveals some interesting patterns of evolution of intergenic regions.

The goals of this study are to estimate the fraction of intergenic regions subject to functional constraints and to develop a formal procedure for identification of regulatory sites by phylogenetic footprinting.

Software was implemented as set of R and Perl/Bioperl scripts and is freely available upon request.

This is joint work with Mikhail Gelfand.

Institute of Biomedical Chemistry, Moscow, Russia, mpyat@ibmh.msk.su

1. N.Rajewsky et al (2002) The Evolution of DNA Regulatory Regions for Proteo-Gamma Bacteria by Interspecies Comparisons, Genome Res., 12:298-308
2. L.Wilson, P.Sharp (2006) Enterobacterial repetitive intergenic consensus (ERIC) sequences in Escherichia coli: Evolution and implications for ERIC-PCR, Mol Biol Evol., 23(6):1156–68.
3. S.Hosid et al (2004) Sequence periodicity of E.coli is concentrated in intergenic regions, BMC Mol Biol., 5:14.

KNOWLEDGE-BASED POTENTIALS FOR PROTEIN ATOM INTERACTION BASED ON MONTE CARLO REFERENCE STATE

SERGEI V. RAHMANOV[1], VSEVOLOD J. MAKEEV[12]

Statistical potentials for interaction of protein atoms of different types are presented. The potentials are derived using a novel non-interacting virtual reference state for atom contacts, called Monte Carlo Reference State (MCRS). It models non-interacting elements in the macromolecular structure space as random 3D points. As a result, the expected contact density probabilities can be calculated to any desired precision for any individual structure, taking into account its size, shape, and molar fractions of contacting atoms, in a natural fashion. This allowed us to obtain atomic contact potentials for hydration of all types of protein atoms with explicit water molecules, and potentials for interaction of protein atoms with different ions, including calcium, zinc, etc. Due to the use of MCRS, the resulting potentials are very detailed, continuous, and extent over unlimited range of contact distances, including very short contacts in the sub-van-der-Waals distance range. The atomic hydration potentials (AHP) provide a detailed quantitative distance-dependent description of protein atom hydropathic properties. They can be used to predict protein-bound water molecules with a high accuracy, and a lower over-prediction error. This allows correct locating of individual water molecules in the context of protein and nucleic acid environment, and at macromolecular interfaces during docking and protein interaction modeling. Total macromolecular solvation energy estimates are made, as a sum of calculated hydration energy contributions from all nearby atoms over a first hydration shell, defined as space no closer than 2.5 angstrom to any structure atom, and not further away than 4.5 angstrom from at least one atom.

[1] GosNIIGenetika, Moscow, Russia, sergeira@genetika.ru
[2] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

Taking into account the fact that for the majority of soluble proteins, hydrophobic interactions account for about 90% of the total free energy providing stability of the native conformation [1], we applied AHP in fold recognition test. The improved Rosetta decoy set [2] was used, which includes on the average 1850 alternative misfolded decoy structures for each of the native protein structure. We demonstrate that in the majority of cases (27 of 41), the native fold was selected as that having the minimal calculated hydration energy over all of the decoys, and can be identified (often with a sizeable energy gap separating the native fold from the decoys), on the basis of hydration energy estimate alone, without any consideration of the internal protein energy. This result is on par with the best contemporary protein structure modeling potentials [3]. It points to the possible applications of AHP for dynamical protein structure modeling and analysis, for quantitative prediction of protein interactions, and for directed design of more thermally stable enzyme variants.

Empirical contact potentials for binding of various ions in proteins characterize interaction properties of all protein atoms in relation to metal ions, including calcium and zinc. They can be used to predict binding sites for different ions in proteins, to quantitatively characterize site specificities to different ions, and to carry out a design of novel metal binding sites.

1. Harano Y, Kinoshita M (2005) Translational-entropy gain of solvent upon protein folding. *Biophys J*, 89(4):2701-2710.
2. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* 2003, 53(1):76-87.
3. Dong Q, Wang X, Lin L (2006) Novel knowledge-based mean force potential at the profile level. *BMC Bioinformatics*, 7:324.

IDENTIFYING MICRORNAs AND THEIR TARGETS

NIKOLAUS RAJEWSKY

I will summarize what can be learned from predicting and analyzing microRNA targets. As an exaqmple, I will discuss the function of miR-150 in the immune system. Finally, I will present a new algorithm for the identification of microRNAs from deep sequencing data.

[1] Max-Delbrück-Centrum für Molekulare Medizin (MDC), Berlin, Germany, rajewsky@mdc-berlin.de

POSITIVE SELECTION AND ALTERNATIVE SPLICING IN HUMAN GENES

VASILY RAMENSKY[1], R.NURTDINOV[2], A.NEVEROV[3], A.MIRONOV[3], MIKHAIL GELFAND[4]

We have stidued the densities of single nucleotide polymorphisms and human-chimpanzee divergence in the coding regions of 6,671 alternatively spliced human genes. The alternatively spliced regions from minor isoforms experience lower selective pressure at the amino acid level and simultaneous selection acting against synonymous sequence variation. The results of the McDonald-Kreitman test demonstrate that, unlike the constitutive regions, minor alternatives are subject to the positive selection, with up to 24% of amino acids fixed by positive selection. This effect is observed both in conserved and non-conserved exons. The minor alternative exons being relatively young may be considered therefore as a natural substrate for positive selection.

A NOVEL APPROACH TO LOCAL SIMILARITY OF PROTEIN BINDING SITES AND ITS APPLICATION TO COMPUTATIONAL DRUG DESIGN

VASILY RAMENSKY[5], A.SOBOL[2], N.ZAITSEVA[2], A.RUBINOV[2], VICTOR ZOSIMOV[6]

Modeling of molecular interactions in protein-inhibitor complexes is the basis of modern computational drug design and on the other hand an extremely complicated and far from solved problem. Fortunately, modeling of binding can be supported by known structural data on protein-ligand complexes available, taking advantage of the similarity between the protein features responsible for binding.

[1] Engelhardt Inst. Of Molecular Biology, Vavilova, 32, Moscow, 119991, Russia, ramensky@imb.ac.ru
[2] Department of Bioengineering and Bioinformatics, Moscow State University, Vorob'evy gory, 1-73, Moscow, 119992, Russia
[3] State Scientific Center GosNIIGenetika, 1st Dorozhny proezd 1, Moscow, 117545, Russia
[4] Institute for Information Transmission Problems, RAS, Bolshoi Karetny pereulok 19, Moscow, 127994, Russia, gelfand@iitp.ru
[5] Engelhardt Inst. Of Molecular Biology, Vavilova, 32, Moscow, 119991, Russia Algodign LLC, Bolshaya Sadovaya 8, Moscow, 123379, Russia, ramensky@imb.ac.ru
[6] Applied Acoustics Research Institute, 9 May St. 7a, Dubna-1, 141981, Russia, victor.zosimov@niipa.ru

We introduce a novel notion of binding site local similarity based on the analysis of complete protein environments of ligand fragments. Comparison of a query protein binding site (target) against the spatial structure of another protein (analog) in complex with a ligand enables ligand fragments from the analog complex to be transferred to positions in the target site, so that the complete protein environments of the fragment and its image are similar. The revealed environments are called similarity regions and the fragments transferred to the target site are considered as binding patterns. The set of such binding patterns derived from a database of analog complexes forms a cloud-like structure (fragment cloud), which proves to be a powerful tool for computational drug design. It has been shown on independent test sets that the combined use of a traditional energy-based score together with the cloud-based score responsible for the quality of embedding of a ligand into the fragment cloud improves the self-docking and screening results dramatically. The cloud-based docking method has been implemented in the in-house software AlgoComb [1], which uses fragment clouds both for initial ligand anchoring and subsequent calculation of the mixed score.

AlgoComb performance with the mixed score was tested on the self-docking test with 100 protein-ligand complexes that have been used before for the comparative evaluation of eight available docking tools [2]. The performance in a self-docking test is measured by the percentage of cases in which the RMSD of the best scored position of the ligand does not exceed 2Å from the X-ray determined native position. The best performance attained by the reviewed tools is in the range 50-55% of all cases. AlgoComb has shown comparable 52% success rate when run with the energy-based scoring function without cloud usage and 82% when fragment clouds are used.

The performance of fragment cloud-assisted virtual ligand screening was tested on the HSV-1 thymidine kinase (TK, PDB ID 1kim) used as a target for the database of 990 drug-like molecules and 10 known TK inhibitors [3]. The success rate in virtual screening experiments is determined by a number of real inhibitors in a certain top fraction of the energy scores assigned to all tested compounds. The obtained ranks of the 10 true TK inhibitors are in the range 4-24, which is a clear improvement over the best results attained previously [2].

The usage of a fragment cloud as a source of positioned molecular fragments fitting the binding protein environment has been validated by reproduction of experimental ligand optimization results.

1. S.Nikitin et al. (2005) A very large diversity space of synthetically accessible compounds for use with drug design programs, J Comput Aided Mol Des, 19:47-63.
2. E.Kellenberger, et al. (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy, Proteins: Structure, Function, Bioinformatics, 57:225-242.
3. C.Bissantz, et al. (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem, 43:4759-4767.

COMPARATIVE GENOMIC ANALYSIS OF TRANSCRIPTIONAL REGULATORY NETWORKS IN SHEWANELLA SPECIES AND OTHER □-PROTEOBACTERIA

DMITRY A. RODIONOV[1,2]

Integrative comparative genomics approaches were used to infer transcriptional regulatory networks (TRNs) in 11 Shewanella species and a set of other □-proteobacteria with sequenced genomes. To accomplish this goal, we combined the identification of transcription factors (TFs), TF-binding sites (TFBSs) and cross-genome comparison of regulons with the analysis of the genomic and functional context inferred by metabolic reconstruction. The reconstructed TRNs for the key pathways involved in central metabolism, production of energy and biomass, metal ion homeostasis and stress response provide a framework for the interpretation of gene expression data. This analysis also helps to improve functional annotations and identify previously uncharacterized genes in metabolic pathways. Finally, we attempted to reconstruct possible evolutionary scenarios of these TRNs.

Using this approach, we identified candidate TFBSs for more then 20 TFs of known specificity, including global regulators (Crp, Fnr, ArcA, Fur, LexA) and specialized regulators of the metabolism of nitrogen (NarP, IscR, NsrR, DNR, NorR), amino acids (BirA, ArgR, MetJ, TrpR, TyrR, HutC), fatty acids (FadR, FabR), carbohydrates (PdhR, HexR, GntR), cofactors (BirA, IscR). Two novel highly conserved regulons, named IlvR (SO1898) and FadQ (SO2493), were tentatively predicted for the sets of genes involved in the degradation of branch

[1] Institute for Information Transmission Problems RAS, Moscow, Bolshoi Karetny per. 19, 127994, Russia, rodionov@iitp.ru
[2] Burnham Institute for Medical Research, La Jolla, 10901 N. Torrey Pines Rd., California 92037, USA

chain amino acids, and fatty acids, respectively. We also identified candidate TFBSs for previously uncharacterized sugar catabolic TFs, termed NagR, SdaR, ScrR, AraR, and BglR, tentatively implicated in the control of the utilization of N-acetylglucosamine, glycerate, sucrose, arabinose and □-glucosides, respectively. Finally, we have mapped the genes and operons controlled by five types of metabolite-binding riboswitches (B12, LYS, RFN, THI, GLY), and six translational attenuators of amino acid biosynthesis pathways (ilv, leu, his, thr, trp, phe operons).

Although some diversity of the predicted regulons is observed within the collection of Shewanella spp., the most striking difference in the overall regulatory strategy is revealed by comparison with E. coli and other □-proteobacteria. Multiple interesting trends in diversification and adaptive evolution of TRNs between lineages were detected including regulon "shrinking", "expansion", "mergers", and "split-ups", as well as multiple cases of using nonorthologous regulators to control equivalent pathways or orthologous regulators to control distinct pathways.

Within the Shewanella lineage, the two major diversification strategies are: constrained ("all or none"), when the regulon is either present or absent in its entirety with tightly conserved regulation of all genes (e.g. for local regulons), and permissive ("loose"), when most genes of a regulon are conserved between genomes, whereas the conservation of respective regulatory sites is much weaker and sometimes not mandatory (e.g. for global regulons). At that, the presence or absence of the constrained regulons is correlated with the pathway essentiality: the biosynthetic NrdR, BirA, and MetJ regulons are always present, whereas sugar catabolism regulons are either present or absent completely. Large regulons seem to be very flexible. Multiple gene and site gains/losses are observed in the LexA, □32, Fnr, ArcA, Crp, Fur and ArgR regulons, although in most cases a conserved core of the regulon can be defined.

Many aspects of metabolic regulation in Shewanella species are substantially different from TRN models that were largely derived from studies in E. coli. Among the most notable are the differences in TRNs for the central carbohydrate pathways. In enterobacteria the central carbon metabolism is controlled by catabolic regulators FruR and Crp, whereas Shewanella species use two other TFs, HexR and PdhR, for this control. The content and functional role of the Crp regulon is significantly different in these two lineages: the catabolism of carbohydrates and amino acids in enterobacteria, and the anaerobic respiration in Shewanella species.

This is joint work with Mikhail S. Gelfand, Andrei L. Osterman, Olga N. Laikova, Anna V. Gerasimova, Dmitry A. Ravcheev, Elizaveta A. Permina,

Alexey G. Vitreschak, and Alexey E. Kazakov. This study was partially supported by Howard Hughes Medical Institute (grant 55005610, "Comparative genomics and evolution of regulatory systems") and Russian Academy of Sciences (Program "Molecular and Cellular Biology").

AN ANALYSIS OF FREQUENCIES OF NUCLEOTIDE SUBSTITUTIONS IN TETRANUCLEOTIDE FRAGMENTS OF PROKARYOTIC GENOMES

SERGEY I. ROGOV, KUVAT T. MOMYNALIEV, VADIM M. GOVORUN

Results. We studied surroundings of nucleotide substitutions in prokaryotic genomes. 12 sets of three-way alignments of orthologous genes from triplets of closely related genomes were analyzed. In 9 sets, nucleotide substitutions of many types occur with maximum frequencies in the palindromic tetranucleotide CTAG. This effect is manifested strongly in genomes of Escherichia coli strains and Salmonella species. In both E. coli and Salmonella, the following substitutions occur with maximum frequencies in the CTAG tetranucleotide: A to C, A to G, A to T, C to A, C to T, G to C, and T to A. Also, in E. coli CTAG corresponds to the maximum frequencies of the substitutions T to C and T to G. The same tetranucleotide corresponds to the maximum frequencies of the substitutions C to G and G to T in Salmonella. Relationship between highest frequencies of different types of substitutions and CTAG is less evident in genomes of Bordetella, Buchnera, Chlamydia, Pseudomonas, Pyrococcus, Vibrio. Such a relationship is revealed for one to four types of substitutions in these genomes. In Helicobacter pylori, substitutions A to G, A to T, C to T, G to A, G to C, T to C occur with highest frequencies in palindromic tetranucleotide GTAC, whereas highest frequencies of the substitutions A to C, C to A, G to T, T to A associate with the palindromic tetranucleotide TCGA. In Streptococcus, palindromic tetranucleotides TGCA, CCGG and GATC are associated with maximum frequencies of the substitutions A to T, C to A and T to C, respectively. In Bacillus and Chlamydia, GGCC corresponds to the maximum frequency of the substitutions C to G. In Staphylococcus, no association between maximum frequencies of substitutions and palindromic tetranucleotides was found. Maximum frequencies of substitutions exceed the mean frequencies of the same type of substitutions 4 - 17 times.

High rate of mutations in CTAG may cause low abundance of this tetranucleotide. Underrepresentation of CTAG in many eubacterial genomes including E. coli was noticed earlier [1, 2]. However, in Buchnera, Chlamydia, Pyrococcus

Research Institute of Physico-Chemical Medicine, M.Pirogovskaya 1a, Moscow, Russia, sirogov@yahoo.com, bioinform@nm.ru, govorun@hotmail.ru

and Vibrio CTAG is not among rarest tetranucleotides. Influence of restriction-modification systems can be a possible cause of both underrepresentation of palindromic tetranucleotides and high mutation rates in them.

Description of method. Determination of frequencies of substitutions in tetranucleotides was performed in the following steps: 1) location of nucleotide substitution positions in gene sequences, 2) reconstruction of hypothetical ancestor gene sequences, 3) counting of tetranucleotide fragments covering substitution positions in ancestor gene sequences and 4) calculation of frequencies of substitutions in tetranucleotides. Method of location of nucleotide substitution positions is analogous to that of finding of amino acid substitutions in proteins [3]. Triplets of aligned nucleotide sequences of orthologous prokaryotic genes were analyzed. Sequence triplets were downloaded from ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/Cysteine [3]. Triplets consist of two sequences from close relative genomes (sister sequences) and a sequence from distant relative genome (outgroup sequence). If the outgroup and one of the sisters carries the same nucleotide X, but the second sister sequence carries nucleotide Y, the substitution X to Y was registered for the second sister sequence. According to that information, hypothetical ancestor sequences were reconstructed by replacing nucleotide Y with X. A pair of hypothetical ancestor sequences was reconstructed for every triplet of sequences and for each kind of nucleotide substitutions. Thus, 12 pairs of hypothetical ancestor sequences were produced from every triplet of sequences. Surroundings of the substitution positions in the reconstructed sequences were scanned with sliding frame of width four nucleotides. Tetranucleotides that include positions of substitutions were counted. These counts were interpreted as numbers of substitutions in particular tetranucleotides. If two tetranucleotides have the same sequence but differ in position of substitution they were considered different. Frequencies of substitutions in each tetranucleotide were computed as ratios of numbers of substitutions in tetranucleotide and total number of that tetranucleotide in reconstructed sequences.

1. Karlin S., Ladunga I., and BE Blaisdell B.E. (1994) Heterogeneity of Genomes: Measures and Values, Proc. Natl. Acad. Sci. USA, 91: 12837-12841.
2. Elhai J. (2001) Determination of bias in the relative abundance of oligonucleotides in DNA sequences, J. Comput Biol. 8:151-75.
3. I. King Jordan et al. (2005) A universal trend of amino acid gain and loss in protein evolution, Nature, 433: 633 -638.

PREDICTING TRANSCRIPTION FACTOR AFFINITIES TO DNA FROM A BIOPHYSICAL MODEL

H. ROIDER, A. KANHERE, T. MANKE, M. VINGRON

Theoretical efforts to understand the regulation of gene expression are traditionally centered around the identification of transcription factor binding sites at specific DNA positions. More recently these efforts have been supplemented by large-scale experimental data (ChIP-chip) for the relative binding strength of proteins to longer intergenic sequences. The question arises to what extent these two approaches converge. So far, a direct comparison has been made difficult by the presence of an arbitrary cutoff, which is commonly imposed on both in vivo data and on in silico binding site predictions.

Here we adopt the physical binding model of Berg and von Hippel to predict the fraction of bound transcription factors and the relative binding strengths of a given transcription factor to any sequence region. In contrast to the traditional search for binding sites, we do not impose any threshold, but integrate the contributions from strong and weak binding sites to calculate the overall binding strength of a transcription factor to a given region. This approach pertains directly to the experimental situation of ChIP-chip data, and we draw upon a large scale data set from S. cerevisae (Harbison 2004) to calibrate the parameters of the model. After calibration, our transcription factor affinity prediction (TRAP) tool can predict the relative binding strength of transcription factors even in the absence of large-scale experimental binding data.

We demonstrate that, within this probabilistic framework, a significant fraction of experimental low and high affinity binding data can be rationalized in terms of only two universal parameters.

Our method can assign high affinities to sequences where hit-based methods fail to report any "match", and it also accounts more accurately for differences in the binding strength of sites which are traditionally reported only as hits.

We compare our predictions to a number of traditional approaches and find that TRAP has a higher predictive power with respect to experimental binding ratios than any of the hit-based methods. While our approach is designed to compare different sequences with respect to their affinity for a given factor, I will also present a statistical approach to normalize different factors. With a proper normalization in place, we are able to compare different factors with each other, and rank them with respect to a given sequence of interest. Finally,

Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany, manke@molgen.mpg.de

I will illustrate the applicability of our approach to promoter regions of higher eukaryotes.

1. Harbison et al. (2004) Transcriptional regulatory code of a eukaryotic genome, Nature, 431 (7004): 99-14.
2. H. Roider, A. Kanhere, T. Manke and M. Vingron (2007) Prediciting Transcription Factor Affinities to DNA from a Biophysical Model, Bioinformatics, 23(2): 134-141.

PHYLOGENOMICS OF METAZOA: CONSTRUCTING THE GENE SET

LEONID RUSIN, V.A. LYUBETSKY

Modern biology has been revolutionized by the advent of comparative molecular sequence data and powerful methods of phylogenetic analysis that allow us greater insight into evolutionary history than ever before. Rapid development of bioinformatics and avalanche of genomic data from a great variety of organisms make it timely and feasible to ascertain evolutionary affinities within and among major branches of the tree of life using extensive molecular evidence. In this context, inferring deep phylogeny of Metazoa (largely represented by multicellular animals with true tissues and complex body plan) in many respects is considered a priority [1, 2]. The diversity of extant animal body plans far exceeds that accommodated in modern phylogenies, many are represented by aberrant, often neglected phyla with traditionally disputable evolutionary affinities. In this study, full advantage was taken of the bulk of existing, and continuously accumulating, genomic data to build phylogenetically informative datasets with extensive sampling of animal diversity.

A host of techniques, including original methods, was employed to mine for homologous genetic markers in databases containing complete metazoan genomes and EST libraries, assembling and translating ESTs, assessing markers' orthology and their informativity for establishing large-scale relationships. Original methods were used to control for quality of multiple sequence alignments of the found markers via reducing the amount of uninformative "noisy" columns [3] thus providing for the best possible resolution of individual-marker trees.

A constructed dataset contains 51 genes from 30 animal species of 11 phyla constituting a representative sample of extant metazoan diversity (Ctenophora,

Kharkevitch Institute for Information Transmission Problems, Bol'shoi Karetnyi lane 19, 127994 Moscow, Russia, rusin@iitp.ru, lyubetsk@iitp.ru

Cnidaria, Priapulida, Tardigrada, Platyhelminthes, Nematoda, Arthropoda, Annelida, Mollusca, Echinodermata, Chordata). Currently in phylogenomics, a multigene dataset for more than four phyla is obtained for the first time. It is used in sophisticated phylogenetic analyses to build a tree of the Metazoa based on total evidence from the 51 gene and on combining individual gene trees into a supertree. The outcome of this analyses provides important new knowledge and insights into a number of fundamental biological questions, e.g. ancestral nature of the metazoan body plan, origins of the secondary body cavity (coelom) among the bilaterally symmetric animals and its fate in groups currently considered acoelomic, such as nematodes and flatworms.

Research was supported by grant RFBR 05-04-49705.

1. G. Giribet (2002) Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion? Molecular Phylogenetics and Evolution, 24: 345–357.
2. A. Schmidt-Rhaesa (2003) Old trees, new trees – is there any progress? Zoology, 106: 291–301.
3. V.A. Lyubetsky et al. (2005) Removing noise in multiple protein alignment. Information processes, 5: 380-391.

BENCHMARKING OF INTERNET SERVERS FOR RECOGNITION OF TRANSMEMBRANE SEGMENTS IN BETA-BARREL PROTEINS FROM GRAM-NEGATIVE BACTERIA

NATALIYA S. SADOVSKAYA

Beta-barrel outer membrane proteins perform a variety of functions and play significant roles in cells life. Thus the analysis of such proteins is an important problem in bioinformatics. It is addressed by several servers that predict positions of transmembrane segments (TM-segments) in beta-barrel membrane proteins. We benchmarked seven available servers using the consistency criterion: predictions of TM-segments in homologous proteins should be similar.

Starting with beta-barrel porins families (class TC.1B) from the TCDB database http://www.tcdb.org/ 1, we identified homologous proteins from gram-negative bacteria in the COGs database 2. Total 5673 proteins pairs from 14 COGs were analyzed. The obtained sequences were aligned with ClustalW. TM-segments were predicted by seven servers:

State Scientific Center “GosNIIGenetika” Moscow, Russia, nathalie@iitp.ru

B2TMPRED (http://gpcr.biocomp.unibo.it/cgi/predictors/outer/pred_outercgi.cgi), B2TMR (http://gpcr.biocomp.unibo.it/predictors/), HMM-B2TMR (http://gpcr.biocomp.unibo.it/predictors/), PRED-TMBB (N-best method) (http://bioinformatics.biol.uoa.gr/PRED-TMBB), PRED-TMBB (Posterior decoding method) (http://bioinformatics.biol.uoa. gr/PRED-TMBB), PRED-TMBB (Viterbi method) (http://bioinformatics.biol.uoa.gr/PRED-TMBB), TMBETA-NET (http://psfs.cbrc.jp/tmbeta-net/).

The default settings were used. Overlapping and adjacent segments predicted in one protein were merged. All pairs of orthologs were used to calculate the index Q (Jackard's coefficient of community) and the segment consistency index C, measured as the fraction of common TM-segments in two proteins 3. The average values of the consistency indices Q and C and standard deviations (σ) are listed in Table 1.

The most consistent predictions was made by B2TMR followed by B2TMPRED and HMM-B2TMR (the latter assigned some proteins to the "non-transmembrane" class).

This is joint work with Mikhail Gelfand.

Table 1. The Q and C consistency indices and standard deviation (σ) in different identity intervals.

ID	0 - 50%		51 - 100%		all	
Server	Q ± σ	C ± σ	Q ± σ	C ± σ	Q ± σ	C ± σ
B2TMR	0,67 ± 0,15	0,84 ± 0,11	0,85 ± 0,18	0,93 ± 0,11	0,68 ± 0,15	0,84 ± 0,11
B2TMPRED	0,49 ± 0,15	0,68 ± 0,14	0,79 ± 0,23	0,88 ± 0,17	0,50 ± 0,16	0,68 ± 0,14
HMM-B2TMR	0,55 ± 0,26	0,71 ± 0,31	0,70 ± 0,35	0,78 ± 0,35	0,55 ± 0,27	0,71 ± 0,31
PRED-TMBB (N-best method)	0,37 ± 0,18	0,56 ± 0,24	0,67 ± 0,27	0,78 ± 0,25	0,38 ± 0,19	0,57 ± 0,25
PRED-TMBB (Viterbi method)	0,37 ± 0,18	0,56 ± 0,24	0,67 ± 0,27	0,78 ± 0,25	0,38 ± 0,19	0,57 ± 0,25
PRED-TMBB (Posterior decoding method)	0,37 ± 0,17	0,56 ± 0,23	0,66 ± 0,28	0,78 ± 0,24	0,37 ± 0,18	0,57 ± 0,24
TMBETA-NET	0,36 ± 0,08	0,54 ± 0,10	0,66 ± 0,24	0,79 ± 0,20	0,37 ± 0,10	0,54 ± 0,11

1. M.H. Jr. Saier (1999) A functional-phylogenetic system for the classification of transport proteins, J Cell Biochem Suppl, 32–33:84–94.
2. R.L. Tatusov et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes, Nucleic Acids Res., 29:22-28.

3. N.S. Sadovskaya et al. (2006) Recognition of transmembrane segments in proteins: review and consistency-based benchmarking of internet servers, J Bioinform Comput Biol., 4:1033-1056.

THE MODIFICATION OF MUSCLE MULTIPLE SEQUENCE ALIGNMENT ALGORITHM FOR MULTIPROCESSORS

ALEXEY N. SALNIKOV

The goal of this work is to reduce the time of creating a multiple sequence alignment by means of parallelizing an heuristic algorithm. We choose the algorithm MUSCLE [1] whose realization is available at http://www.drive5.com/muscle as a program called "muscle".

The MUSCLE algorithm has several stages. On the first stage a matrix of similarity of the input sequences is created. Using the matrix, a binary cluster tree is build with neighbor-joining or UPGMA algorithms. Using that tree as a guide tree, a progressive alignment is performed: a multiple alignment is created from pairwise alignments of sequences and profiles (see [1]). On the last stage, the created multiple alignment is improved. An improved guide tree is used to realign appropriate subalignments. The last stage can be repeated several times.

In our parallel version of the program, the stage of progressive alignment is modified. The creation of profiles for subalignments is done independently on different processors, if it is possible. We build a parallel program as a graph-program by means of PARUS [2]. PARUS is a parallel programming language that allows building parallel programs in data flow graph notation. To define a data flow graph, one must define source vertices, inner vertices, a drain vertex, and directed edges connecting them. A realization of a data flow graph as a parallel program is called graph-program.

In the parallel version of MUSCLE, each source vertex corresponds to one or several input sequences. In the latter case, it represents a node of the guide tree (the choice of that node depends on a user-defined parameter). Each inner vertex represents a node of the guide tree, and the edges are the branches of the guide tree directed from the leaves to the root. The drain vertex correspond to the root of the guide tree. In the source vertices, alignments of the correspondent sequences are created by the original algorithm of MUSCLE, and the profiles are generated from that alignments. Each inner vertex aligns the pair of

[1] Russia, 119992, Moscow, Leninskie Gory, MSU, 2-nd educational building, VMK Faculty, salnikov@cs.msu.su

profiles from the upstream vertices to create a new profile that can be sent to the next inner node or to the drain node.

So, the data flow is directed from the leaves to the root of the guide tree. To improve the efficiency of the graph-program on a multiprocessor computer, the bottom (closer to leaves) layers of the tree are compressed into one layer of the data flow graph. To do that, a new parameter that defines a number of layers of the tree to be compressed is introduced into the algorithm. As a result of the compression, a source vertex of the data flow graph can correspond to a group of sequences.

The time of program execution for modified MUSCLE algorithm is reduced due to the possibility to execute vertices of the graph-program on different processors. Aligning the LTR5 collection (all known LTR5 objects in human genome described in [3]) by the original MUSCLE program got 1h8m at the PrimePower 850 machine. Our parallel version using all 12 processors made that task in 28m (2.4 times faster). The improvement looks rather low; however note that the guide tree in this example is rather unbalanced, while the algorithm is most efficient for a balanced guide tree. It should be noted that the parallel version uses the memory in a more effective manner.

The work is partly supported by INTAS, grant 05–1000008–8028.

1. Robert C. Edgar (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics, 5:113.
2. Alexey N. Salnikov (2006) PARUS: a Parallel Programming Framework for Heterogeneous Multiprocessor Systems, Lecture Notes in Computer Science, 4192:408-409.
3. Alexeevski A.V., Lukina E.N., Salnikov A.N., Spirin S.A. (2004) Database of long terminal repeats in human genome: structure and synchronization with main genome archives, Proceedings of the fours international conference on bioinformatics of genome regulation and structure, 1:28-29.

PERIODIC PATTERN OF SECONDARY STRUCTURES IN PROKARYOTIC AND EUKARYOTIC mRNAs

S.A. SHABALINA[1], A.Y. OGURTSOV[1], N.A. SPIRIDONOV[2]

Several hypotheses have been proposed to explain the non-random use of synonymous codons and the relatively high GC content at the 3rd degenerate codon sites observed in many life forms, however, no firm conclusion has been reached1-5. The functional importance of mRNA secondary structure and the idea that the redundancy of the genetic code may allow conservation of mRNA folding6 has been supported by several lines of evidence. Conservation of secondary structure features was demonstrated for retroviral mRNA, where folding in RNA stem regions disrupted by silent mutations on one strand is restored by compensatory mutations on the other strand7. Significant biases in favor of local RNA structures have been found in several bacterial species and yeast8. Analysis of synonymous nucleotide polymorphysm in enteric bacteria and compensatory nucleotide substitutions in Drosophila suggested selective constraint on mRNA secondary structures9. Moreover, synonymous mutations affecting mRNA and pre-mRNA structure and stability can be highly deleterious and have implications in disease in humans10.

To study the relationship between the genetic code and mRNA folding, we evaluated sequence conservation, free Gibbs energy of secondary structure formation, and nucleotide involvement in secondary structure elements for 19,317 human and 20,892 mouse mRNA sequences folded in silico11. In the CDSs, we found pronounced periodic patterns of mRNA secondary structure stability and nucleotide base pairing. Notably, base pairing at the 3rd CG-rich codon sites and their contribution to mRNA secondary structure stability are significantly higher than contributions of the 1st G-rich sites or the 2nd AU-rich sites. We show that this pattern is created by the structure of the genetic code, and the dinucleotide relative abundances are important for the maintenance of mRNA secondary structure. Although synonymous codon usage contributes to this pattern, it is intrinsic to the structure of the genetic code, and manifests itself even in the absence of synonymous codon usage bias at the 4-fold degenerate sites. While all codon sites are important for the maintenance of mRNA secondary structure, degeneracy of the code allows regulation of stability and periodicity of mRNA secondary structure. We demonstrate that the 3rd degenerate codon sites contribute most strongly to mRNA stability in mammals.

[1] NCBI, NLM, NIH, Bethesda, MD 20894 USA
[2] Center for Drug Evaluation and Research, FDA, Bethesda, MD 20892 USA

These results convincingly support the hypothesis that redundancies in the genetic code allow transcripts to satisfy requirements for both protein structure and RNA structure. Our data evidence that selection may be operating on synonymous codons to maintain a more stable and ordered mRNA secondary structure, which is likely to be important for transcript stability and translation. Similar universal periodic pattern of nucleotide base pairing in mRNAs was observed in flies, worms and bacteria.

We also demonstrate that functional domains of the mRNA (5'UTR, CDS and 3'UTR) preferentially fold onto themselves, while the start codon and stop codon regions are characterized by relaxed secondary structures, which may facilitate initiation and termination of translation. Characteristic translation regulatory signals such as Shine-Dalgarno and Kozak sequences for prokaryotic and eukaryotic organisms possess specific secondary structures.

1. Karlin, S., and Mrazek, J. (1996) What drives codon choices in human genes? J. Mol. Biol., 262: 459-472.
2. Kanaya, S., et al. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol., 53: 290-298.
3. Duret, L. (2002) Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev., 12: 640-649.
4. Comeron, J.M. (2004) Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics, 167: 1293-1304.
5. Chamary, J.V., and Hurst, L.D. (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol. Biol. Evol., 21: 1014-1023.
6. White, H.B.r., et al. (1972) Messenger RNA structure: compatibility of hairpin loops with protein sequence. Science, 175: 1264-1266.
7. Konecny, J., et al. (2000) Concurrent neutral evolution of mRNA secondary structures and encoded proteins. J. Mol. Evol., 50: 238-242.
8. Katz, L., and Burge, C.B. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res., 13: 2042-2051.
9. Innan, H., and Stephan, W. (2001) Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics, 159: 389-399.

10. Duan, J., et al. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet., 12: 205-216
11. Shabalina, S.A, Ogurtsov, A.Y., Spiridonov, N.A. (2006) A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res., 34: 2428-2437.

REACTION OF HUMAN HELA CULTURED CELLS TO TOTAL PROTEIN SYNTHESIS INHIBITION

LEV I. SHAGAM[1], OLGA V. ZATSEPINA[2]

The nucleolus is the major nuclear structural domain that serves for ribosome production and is involved in regulation of cell death by apoptosis. Activity and morphology of the nucleolus are known to change following a variety of cell stimuli [1].

The present research was aimed at revealing traits of human cells' early response to total protein synthesis inhibition by an anti-cancer drug anisomycin. Human cervical HeLa cultured cells were used in the study.

A3 antigen (a nucleolar protein, which is likely to be a part of RNA polymerase I transcriptional complex) is a cytological marker of ongoing protein synthesis in human cell cultures [2]. It migrates from the nuclei to numerous discrete foci in the nucleoplasm in response to protein synthesis inhibition, which precedes apoptosis of the cultured cells. However, only some members of the population (around one third), "sensitive" cells, demonstrate the A3 topology alteration.

Using an original computer program aimed at revealing cellular clusters we have shown that "sensitive" cells are situated regularly, in clusters. Cells demonstrating A3 antigen migration to the nucleoplasm and the ones in the S-period of cell cycle have been shown to coincide (except 10-20% from both of the groups). These observations argue in favour of the idea that S-period cells are the most sensitive to the total protein synthesis inhibition as judged by localization of the RNA polymerase I machinery.

We are grateful to Dr. Prof. A.V. Alexeevski, Dr. D.A. Alexeevski and Dr. T.A. Alexeevski for their significant contribution to statistical processing of our data, and to Dr. A.A. Grigoriev for supplying cells.

[1] Moscow State University, Moscow, Russia, levshagam@mail.ru
[2] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia

1. F.-M. Boisvert et al. (2007) The multifunctional nucleolus, Nature Reviews Molecular Cell Biology, 1–12 (published online 23.05.07).
2. Т.И. Булычева, И.А. Калинина, А.А. Григорьев, О.В. Зацепина (2006) Штамм культивируемых клеток мышиной гибридомы А3, используемый для получения моноклональных антител к антигену ядрышек клеток человека. Патент на изобретение №2005115525/13(017785) от 19 сентября 2006 г.

IN SILICO SEARCH FOR NATURAL ANTISENSE TRANSCRIPTS IN HUMAN GENOME AND ANALYSIS OF THEIR EXPRESSION PATTERNS

MIKHAIL SKOBLOV[1], DMITRY KLIMOV[2],
TATIANA TYAZHELOVA[3], ANCHA BARANOVA[4]

Both mRNA expression in a eukaryotic cell and efficiency of its translation into proteins are controlled by many regulatory levels subsequent to transcription initiation. As mRNA is a single strand molecule, the expression of a complementary antisense strand may alter transcription, elongation, processing, stability, and translation of the template RNA. Functional antisense RNAs have been identified in bacteria, but later were shown to be involved in gene regulation and differentiation in several eukaryotic organisms, including mammals. Natural antisense transcripts (NAT) usually arise via separate transcription initiation on the opposite DNA strand at the same genomic locus as the sense strand. Computational analysis of data from large-scale sequencing projects has revealed a surprising abundance of antisense transcripts in several eukaryotic genomes. As some antisense transcripts have been shown to regulate gene expression, it is possible that antisense transcription might be a common mechanism of regulating gene expression in eukaryotic cells.

We created an algorithm that allows high-throughput mapping of NATs. We used exact coordinates of transcripts and their orientations on the plus/minus chains of the human genome archived at NCBI server (NCBI http://www.ncbi.nlm.nih.gov/). In-house software "Antisense Searcher" was written on

[1] Research Center for Medical Genetics, RAMS, Moskvorechie Str., 1, Moscow, Russia, mskoblov@generesearch.ru
[2] Vavilov Institute of General Genetics, RAS, Moscow, Russia
[3] National Hematology Research Centre, Moscow, Russia
[4] Molecular Biology and Microbiology Department and Center for Biomedical Genomics and Informatics, George Mason University, Fairfax, Va., USA, abaranov@gmu.edu

C++ and SQL. This program fulfills following tasks: 1) forming EST and mRNA transcripts in clusters on every chain of DNA; 2) retrieving all overlapping pairs of transcripts that are located on different DNA strands with more than 20 nucleotide overlaps; 3) retrieving an intersection of two previous sequence sets. EST clusters that contain only 1 or two ESTs were filtered at the subsequent stage of analysis. By this method we mapped approximately 13,500 NATs.

To study expression patterns of natural antisense pairs we created C++ - based software "Antisense Cluster Filter". This software allowed us to retrieve tissue expression field for all the transcripts from the lists of NATs. We used cDNA library descriptions available from CGAP website (CGAP http://cgap.nci.nih.gov/) and other sources. By that, our data describing NATs were updated by information of pattern expression of transcripts.

We sorted the NATs data by two criteria: 1) prevalence of expression in tumor or in normal cells 2) tissue specificity. In both cases we found about hundred NATs in which one of the pair expressed only in tumor cells or in specific tissue. These pairs will be experimentally study.

The work in author's laboratory was supported by grant (07-04-00379-a) from RFBR.

1. Klimov D, Skoblov M, Ryazantzev A, Tyazhelova T, Baranova A. In silico search for natural antisense transcripts reveals their differential expression in human tumors. J Bioinform Comput Biol. 2006 Apr;4(2):515-521.
2. Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD.., Over 20% of human transcripts might form sense-antisense pairs., Nucleic Acids Res. 2004;32(16):4812-20.
3. Reis EM et al., Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. Oncogene. 2004 Aug 26;23(39):6684-92.

INFLUENZA VIRUS MEMBRANE PROTEOME STRUCTURAL INVESTIGATION BASED ON ENZYME PROTEOLYSIS AND MALDI-TOF MASS SPECTROMETRY

JULIA SMIRNOVA[1], LARISA V. KORDYUKOVA[2], NATALYA V. FEDOROVA[2], LUDMILA A. BARATOVA[2], MARINA V. SEREBRYAKOVA[3], MICHAEL VEIT[4]

Influenza virus possesses lipoprotein envelope surrounding segmented RNA genome of negative polarity. The envelope is enriched in cholesterol and (glyco)sphingolipids aquired from the host cell plasma membrane "raft" domains. Three transmembrane proteins are incorporated into the envelope of Influenza A virus: glycoproteins hemagglutinin (HA) and neuraminidase (NA) and a minor M2 protein – an ion channel. Major structural component – matrix M1 protein – is membrane-associated and covers the lipid envelope from inside.

Membrane proteome molecules HA and M1 have crucial functions in the viral life cycle, specifically, they drive membrane fusion and budding reactions. Assembly of the progeny viral particles at the cell plasma membrane occurs via interaction of M1 protein both with HA, NA and M2 cytoplasmic tails and ribonucleoprotein complex. Chemical character of these interactions is unknown. In the present work we analyzed by MALDI-TOF mass spectrometry types of fatty acids attached via thio-ester bonds to conserved cysteine residues of the HA2 C-terminal anchoring segment of various virus strains. We also attempted to create a method for assessing possible role of these fatty acids in HA/M1 interactions.

Bromelain digested influenza virions lacking HA ectodomains (subviral particles) were subjected to chloroform/methanol [1] or octylglucoside/Igepal [2] extraction at room temperature to isolate HA2 C-terminal segment. Subsequent MALDI-TOF mass spectrometry analysis revealed that Influenza A virus HA possessed not only palmitate as was known earlier but also stearate (from 10 % for H1 subtype strains till 33 % for H7 subtype ones). H1 and H7 subtypes

[1] Department of Bioengineering and Bioinformatics, Moscow State University, Leninskie Gory, Laboratornii korpus "A", Moscow 119992, Russia, yulya_82@list.ru
[2] Belozersky Institute of Physico-Chemical Biology, Moscow State University, Leninskie Gory, Laboratornii korpus "A", Moscow 119992, Russia, kord@belozersky.msu.ru
[3] Institute of Physico-Chemical Medicine, ul. Malaya Pyrogovskaya, 1a, Moscow 119992, Russia, mserebr@mail.ru
[4] Institute of Immunology and Molecular Biology, Faculty of Veterinary Medicine, Berlin Free University, Philippstr. 13, 10115 Berlin, Germany, mveit@zedat.fu-berlin.de

belong to evolutionary distinct groups of Influenza A virus HA, so obviously differ markedly in their amino acid structure. There was no stearate detected in the mutant A/FPV HA (H7 subtype) as well as in the Influenza B virus HA which both lacked the conserved cysteine located in the transmembrane domain (TMD). Oppositely, the only conserved cysteine of hemagglutinin-esterase-fusion glycoprotein of Influenza C virus which is positioned in TMD was mostly stearoylated (90%). Apparently, this stearate is necessary for proper incorporation of HA into a raft domain before virus budding.

To assess HA2 C-termini fatty acylation role in HA/M1 interactions, bromelain-digested Influenza A and B subviral particles were prepared in the presence of various concentrations (0-200 mM) of 2-mercaptoethanol (ME). There was no (in the presence of 0-5 mM ME), partial (50 mM) or almost full (100; 200 mM) deacylation of the HA2 C-termini inside the subviral particles detected by MALDI-TOF mass spectrometry. Cold solubilization by combination of non-ionic detergents octylglucoside/Igepal has shown that the relative quantity of M1 protein in the HA/M1 co-solubilizate was only slightly reduced when HA2 C-termini were deacylated. It was suggested that HA/M1 interactions are based rather on protein/protein than acyl/protein contacts.

This work was supported by ISTC grant #2816p, RFBR grant #06-04-48728.

1. M..Serebryakova, L.Kordyukova, L.Baratova, S.Markushin (2006), Eur. J.. Mass Spectrom., 12:51–62..
2. V.Radyukhin, N.Fedorova, A.Ksenofontov et al. (2007) Biochem. J.., in preparation.

RECOGNITION OF PROTEIN FUNCTION USING THE LOCAL SIMILARITY

BORIS SOBOLEV, K.E. ALEKSANDROV, A.E. FOMENKO, D.A. FILIMONOV, A.A. LAGUNIN, V.V. POROIKOV

The functional annotation of amino acid sequences is one of the most important problems of bioinformatics. Different programs were successfully applied for recognition of some functional classes, nevertheless many functional groups still not predicted with required accuracy.

We adopted PASS (Prediction of Activity Spectra for Substances) algorithm [1] for the recognition of protein functional classes. Three different descriptions

[1] Institute of Biomedical Chemistry of Rus. Acad. Med. Sci, Pogodinskaya Street, 10, Moscow, Russia 119121, boris.sobolev@ibmc.msk.ru

of amino acid sequences were used, including peptide vocabularies and two original descriptor types. Using the MNA (Multilevel Neighborhoods of Atoms) descriptors [1] an amino acid sequence is presented as a single molecule: the polypeptide fragment is described by a set of atoms accounting their surroundings in a chemical structure. We performed the leave-one-out cross validation test using the peptide vocabulary and MNA descriptors for 77 enzyme classification (EC) groups, including 1267 sequences [2]. The peptide vocabularies gave the best result at the peptide length of 4 amino acid residues. In case of MNA descriptors the accuracy of prediction was higher than for peptides and achieved 0.98. However, the small increase of prediction accuracy required the significant increasing of computing resources.

It is obvious, that the best prediction results can be obtained when a particular sequence is presented by the set of ordered unique descriptors. The sequential descriptors are required that represent ordered conserved fragment of any length and can be quickly calculated. We propose a Local Similarity Projection (LSP) algorithm. Each sequence from the training set is compared with the query sequence: the similarity scores are calculated for all query sequence position. Positional scores are used as descriptors weights in the recognition procedure. The suggested algorithm has the significantly more performance than the alignment methods using in addition more detailed data on the local similarity.

The LSP method was tested vs. two evaluation sets. The first set presented the serine proteinases (EC 3.4.21.X). Both tetrapeptide vocabularies and LSP method showed practically 100% recognition at the highest enzyme specificity level. Another set presented the superfamily of cytochromes P450. In this case one protein can interacted with many ligands and functional classes defined by substrate, inductor or inhibitor specificity are intersected. Phylogenetic clusters not always correspond to functional groups [3]. Substrates and inducers are better recognized for larger groups: the clear trend was shown for peptide vocabulary and LSP. Prediction for inhibitors was less accurate.

Suggested method revealed the effective predictions with different sequence descriptions. Encouraging results were obtained for different types of functional classes.

1. V.V.Poroikov, D.A.Filimonov (2005) PASS: prediction of biological activity spectra for substances, In: Predictive Toxicology, C.Helma (Eds.), 459-468 (Marcel Dekker, New-York).
2. A.Fomenko et al. (2006) Prediction of protein functional specificity without an alignment, OMICS, 10: 56-65.

3. Yu.V.Borodina et al. (2003) If there exists correspondence between similarity of substrates and protein sequences in cytochrome P450 superfamily? Nova Acta Leopoldina., 87: 47-55.

CONFORMATIONAL CHANGES IN ACTIN-BINDING PROTEINS, REVEALED BY SINGLE PARTICLE ELECTRON MICROSCOPY

O.SOKOLOVA[1], S.MAITI[2], N.GRIGORIEFF[3], P.LAPPALAINEN[4], B.L.GOODE[5]

Cell locomotion, endocytosis, and intracellular motility of vesicles, organelles and pathogens all rely on rapid assembly of actin networks. At the heart of these processes are actin nucleators that, upon activation by nucleation promoting factors (NPFs) such as SCAR/WASp family proteins, stimulate actin assembly.. Formins are thought to processively cap the fast-growing ends of actin filaments, while Arp2/3 complex seeds actin polymerization by forming a pseudo-actin trimer of its two actin-related subunits, Arp2 and Arp3 bound to WASP (Rodal et al., 2005). Arp2/3 complex activation stimulates the formation of membrane protrusions downstream of the Rho-family GTPases. The recent studies demonstrated also that formation of membrane protrusions depends on controlled interplay between direct membrane deformation by IRSp53/MIM family proteins and the actin cytoskeleton (Mattila et al., 2007).

Most of the actin binding proteins are composed of multiple domains, performing both regulatory and signaling functions. Here we show using electron microscopy (EM) and single particle averaging that various actin binding proteins constantly undergo major conformational changes. To identify the domain arrangement within the EM reconstructions, we used docking of their known crystal structures into the reconstructed 3D volume. We demonstrated that wild-type Arp2/3 complex exists in solution in three distinct conformations having variable degrees of separation between Arp2 and Arp3 (open, intermediate and closed). Activation of the Arp2/3 complex upon binding the WASP closes the structure. On the other hand, autoinhibited mouse formin exists in closed conformation; the number of open molecules increases dramatically after its activation. Next, we studied the conformational changes during activation of actin binding proteins MIM and IPRs53. We created the difference

[1] Moscow University, Moscow, Russia, 119992, sokolova@moldyn.org
[2] Brandeis University, Waltham MA, 02454 USA
[3] HHMI and Brandeis University, Waltham MA, 02454 USA
[4] University of Helsinki, Helsinki, Finland, 00710
[5] Brandeis University, Waltham MA, 02454, USA goode@brandeis.edu

maps to compare the structures of wild-type, regulator-liganded and G-actin-bound proteins. The obtained results explained the observed ligand-induced conformational changes inside the MIM/IRSp53 proteins.

1. Rodal, A. et al. (2005), Conformational changes in the Arp2/3 complex leading to actin nucleation. Nat. Struct. Mol. Biol., 12(1), 26-31
2. Mattila, P.K. et al (2007), Missing-In-Metastasis (MIM) and IRSp53 deform PI(4,5)P2-rich membranes by an inverse BAR domain like mechanism J. Cell Biol. (in press).

NESTED ARC-ANNOTATED SEQUENCES AND STRONG FRAGMENTS.

T.A. STARIKOVSKAYA[1], M.A. ROYTBERG[2]

1. Definitions and statements. The nested arc-annotated sequences (NAAS) [1] represent RNA secondary structures. A NAAS is a word in the alphabet {A, U, G, C} and a set of nested arcs connecting its letters. The width of the NAAS is a minimal distance between positions connected with the arc. A d-optimal structure for a given word is a NAAS on the word having maximal possible number of arcs among the NAASs of width d or greater. A word is d-strong, if any of its d-optimal structures has an arc between the first and the last characters of the fragment.

The run-time bound of a dynamic programming algorithm finding an optimal structure for a given word w can be improved by replacement of the number of all fragments of a word w (~ n2, where n = n(w) is a length of the word w) by the number F(w) of 1-strong fragments of the word w [2]. An.A.Muchnik [personal communication] showed, that for all n there is a word of length 6n containing at least n2 1-strong fragments. We are interested in the behavior of F(n) where F(n) is an average number of strong fragments of a random word w of length n (all letters of w are iid variables).

Statement 1. Let G(n) be a number of strong words of length n; g(n) = G(n)/4n . Then for all n ≥ 0

$F(n+1) = n/4 + n\cdot g(1) + (n-1)\cdot g(2) + \ldots + 1\cdot g(n)$

2 Computer experiments.

[1] Lomonosov Moscow State University, Leninskie gory, GSP-1, Moscow 119991, Russia, tat.starikovskaya@gmail.com
[2] Institute of Mathematical Problems of Biology, 4, Institutskaja str., 142290, Pushchino, Moscow Region, Russia, mroytberg@impb.psn..ru

To investigate the average number of strong fragments of a random word we have performed computer experiments. For each n $\in$ {1,..., 1500}$\cup${4000, 5000, 6000, 9000} we have created a set R[n] consisting of 200 random sequences of length n. Besides this, we have considered 2000

eucariotic RNA sequences with lengths from 37 to 6393 bp . They were divided into 61 groups {G[n]}, G[n] contains sequences of lengths from 100(n-1)+1 to 100n. For each of groups R[n] and G[n] we have calculated the mean values of F(w) over the group, the mean values are called FR[n] and FG[nmean] respectively; nmean=50+100(n-1). The left graph of Fig.1 shows the plots for log-ratios log (X(n)))/log(n) for X(n)= FR[n] and FG[n] (note: if X(n) = nα, then log (X(n)))/log(n)=α). The graph demonstrates significant difference between the FR[n] and FG[n]: one can not see non-trivial upper limit for the value log(FR[n])/log(n) (obviously, log(F(n))/log(n) ≤2); while FG[n] shows the limit ~ 1.2. The right plot shows that the difference cannot follow from the fact that real RNA secondary structures are of width ≥ 3.

The work was supported by grants RFBR 06-04-49249, INTAS 05-1000008-8028.

The authors thank Andrey A. Muchnik and A.L. Semenov for many helpful discussions and O.Ilyicheva and P.Vlasov for the help with the preparation of data.

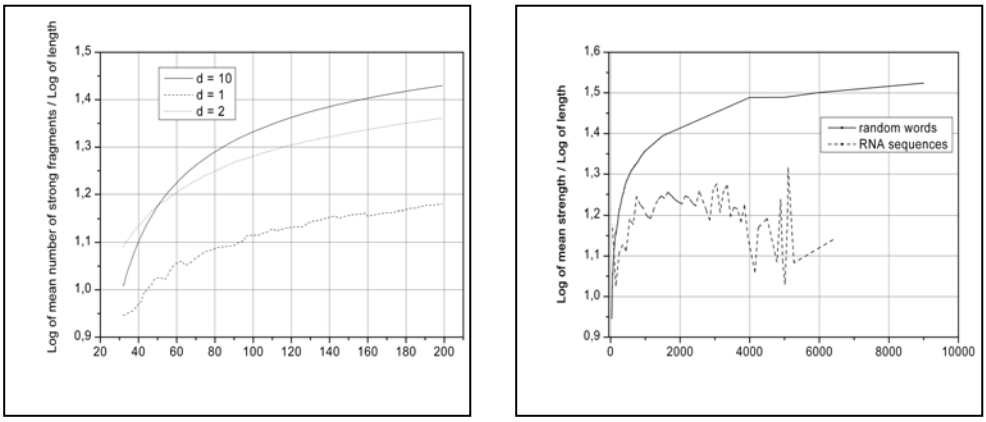


Fig.1. The log-ratios log(X(n))/log(n) for different functions X(n). (Left). The solid line corresponds to X(n) =FR[n]; the dotted line corresponds to X(n) = FG[n]. (Right) The lines correspond to X(n) = DR[d, n]; d=1, 2, 10. Here DR[d, n] is an average number of d-strong fragments over a set R[n].

1. G. Blin and H. Touzet (2006) How to compare arc-annotated sequences: The alignment hierarchy. In: 13th International Symposium on String Processing and Information Retrieval (SPIRE), volume 4209 of Lecture Notes in Computer Science, pages 291-303. Springer Verlag
2. Wexler Y., Zilberstein C., Ziv-Ukelson M. A Study of Accessible Motifs and RNA Folding Complexity. Proceedings of RECOMB 2006: 473-487

AUTOMATED SEARCH FOR REGULATORY MOTIFS IN UPSTREAM REGIONS OF GENES FROM THE FUNCTIONAL SUBSYSTEMS

ELENA STAVROVSKAYA[1], M. CIPRIANO[2], I.L. DUBCHAK[2],
A.A. MIRONOV[1], MIKHAIL S. GELFAND[3]

Introduction: The genes involved in one biological process are often co-regulated. A common motif found upstream of these genes may prove the fact of co-regulation. It is reasonable to search for a motif in a group of closely related genomes, because regulatory motifs are often moderately conserved during the evolution. We have developed a pipeline that searches for putative regulatory motifs in a functional subsystem for a given group of genomes.

Methods: Our pipeline searches for palindromic ungapped motifs of fixed length. As input data it needs a functional subsystem of interest, a group of genomes and the expected motif length.

The functional subsystems are taken from the SEED database [1]. A subsystem is a set of functional roles that together implement a specific biological process or a structural complex. A subsystem may be thought of as a generalization of the pathway. A functional role is an abstract function that a protein performs. A subsystem can be represented as a spreadsheet. Each column in the spreadsheet corresponds to a functional role from the subsystem, each row represents a genome, and each cell identifies the genes within the genome that encode proteins which implement the specific functional role within the designated genome.

To select a correct group of genomes is rather difficult. On the one hand, the genomes should be sufficiently close for the motif to be conserved. On the other hand, if the genomes are too close the upstream regions of the orthologous

[1] Department of Bioengineering and Bioinformatics, Moscow State University, Leninskiye Gory 1-73, Moscow, 119992, Russia, stavrovskaya@gmail.com
[2] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[3] Institute for Information Transmission Problems, Bol'shoi Karetnyi per. 19, Moscow, 127994, Russia, gelfand@iitp.ru

genes are similar and it is impossible to identify the motif. Our pipeline has a special procedure to filter too close genomes from the initial group.

The pipeline implements three approaches. The first one is to search through each genome of the group separately (through the row of the spreadsheet) and than cluster the found motifs. The second one is to search through each functional role of the subsystem (through the column of the spreadsheet) and then cluster the results. And the third one is to search in the upstream regions of all genes corresponding to the subsystem (through the whole spreadsheet).

To find motifs in a set of upstream regions we apply the algorithm SignalX [2].

To cluster the resulting motifs we use the ClusterTree-RS algorithm [3].

Results: We have tested the pipeline for 10 functional subsystems in alphaproteobacteria. The regulators and their regulatory motifs corresponding to the subsystems were known. The results are listed in the table.

SEED subsytem	regulator	motif is found
Hemin_transport_system	RirA	+
Transport_of_Iron	Irr	+
	Fur	+
Transport_of_Manganese	Mur	+
Denitrification	NnrR	+
Ribonucleotide_reduction	NrdR	-
NAD_regulation_experimental	NadQ	+
Nitrogen_fixation	NifA	+
Transport_of_Zinc	ZUR	+
Transport_of_Nickel_and_Cobalt	NikR	-
Pyrimidine_utilization	RutR	+

1. R. Overbeek et al. (2005) The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes, Nucleic Acids Research, 33(17): 5691-5702.
2. A.A. Mironov (2006) Threshold selection using the rank statistics, Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure, 1: 110-113.
3. E.D. Stavrovskaia et al. (2006) ClusterTree-RS: the binary tree algorithm for identification of co-regulated genes by clustering regulatory signals, Mol. Biol. (Mosk), 40(3): 524-532.

INTERACTION OF THE CELLULAR MEMBRANE WITH NO. SIMULATION OF THE PENETRATION OF NO INTO MODEL BIOMEMBRANE

VASILY E. STEFANOV, BORIS F. SHEGOLEV, ANDREY A. MAMONOV

We used computer simulation to investigate interaction of the biomembrane with NO – a small non-charged radical with one unpaired electron considered as one of the most important biochemical regulators. As free diffusion is believed to be a driving force in the process of NO penetration through lipid bilayer membranes [1], taken into account hydrophobicity of NO, we undertook numerical simulation of interaction between NO and biological membrane. The main method of investigation was Molecular Dynamics with the use of GROMACS programs [2]. Other methods and software were: HyperChem 7.0, MNDO, VMD.

At the first stage, using g_membrane program, we constructed a membrane model, consisting of 140 phosphatidyl choline and 60 phosphatidylethanolamine molecules, on the basis of the simplest parallelepiped cell with a uniform distribution of lipids in the bilayer. Then, shrinking of the cell was performed by means of the program SHRINK in order to exclude penetration of water between lipid molecules. After that solvation was carried out with 7000 water molecules added. Solvent molecules were inserted by means of GROMACS algorithm. The obtained membrane model was optimized using molecular dynamics (mdrun).

Construction and preoptimization procedures for NO molecule in the duplet state was similar to those done with phospholipids. Calculated values of the bond length, dipole momentum and, hence, charges of atoms in the molecule were very close to those determined in the experiment. Finally, the system "membrane – water – NO" was generated. 15 NO molecules were added to the equilibrated solvated membrane (program GENBOX; the number of water molecules was reduced to 6590) and energy minimization was performed. After that, the final simulation was undertaken at a constant pressure, over the time interval equal to 1200 ps. Parameters x, y, z of the cell were 5.5 nm, 10.0 nm and 9.0 nm, equilibrium being reached 500 ps after the start. The equilibrium thickness of the solvated membrane, estimated as the distance between the phosphorus atoms of the phosphate groups, was about 4.25 nm.

After 1.2 ns simulation with NO, the membrane thickness became equal to 4.5 nm. It was found that at the beginning of the simulation process the orien-

[1] Department of Biochemistry, St.Petersburg State University, Universitetskaya nab. 7/9, St.-Petersburg, 199034 Russia, vastef@mail.ru

tation of NO molecules is perpendicular with respect to the bilayer plane. However, as penetration of NO molecules into the lipid bilayer proceeds, deviation from their normal orientation may occur. To assess the depth of their penetration into the lipid bilayer we plotted the values of the normal (with respect to the bilayer plane) coordinate of the nitrogen atom of NO. The obtained pattern demonstrated a slow rise in the value of the normal coordinate, testifying to the diffusion of NO into the bilayer. A few sharp peaks, observed in the plot, may either be connected with transmembrane diffusion of NO or, more likely, originate from transition of NO through the cell boundary, which can be regarded as an artifact. This observation requires further analysis. Also plotted were functions of radial distribution of atoms with respect to the ammonia nitrogen of phospholipids' polar heads at the beginning and at the end of the simulation process. We estimated values of the area of bilayer surface per one molecule at the beginning and after 1.2 ns of simulation with NO at constant pressure. The obtained values were 0.244 nm2 and 0.255 nm2, respectively. These data are compatible with the presumption of the biomembrane permeability for NO [1].

The results of the undertaken simulation of interaction of the model biomembrane with NO suggest that penetration of NO into the phospholipid bilayer of the membrane takes place in the time interval 0.5 – 1.0 ns, preceded by the normal orientation of NO molecules with respect to the bilayer plane during 0.3 – 0.5 ns at a distance 0.5 – 1.0 nm from the membrane surface. Once penetrated through the membrane surface, NO molecules begin to change their orientation from normal to the tangential one, apparently due to the field effects. Two closely related important trends are distinctly traced. These are a decrease in the membrane thickness and increase in the surface area per phospholipid molecule. The simulation clearly demonstrated that initial effects of NO solving in the lipid phase occur, which supports the idea of NO diffusion through the biological membrane as a plausible mechanism accounting for the penetration of this physiologically important mediator into the cell.

1. W. Subczynski, M. Lomnicka, J. S. Hyde. (1996) Permeability of nitric oxide through lipid bilayer membranes. Free Radicals Research, 24:343-349
2. E. Lindahl, B. Hess, van der Spoel. (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. Journal of Molecular Modelling, 7:306-317

EXPRESSION PROFILING OF SINGLE NEURONAL PROGENITOR CELLS

TATIANA SUBKHANKULOVA, F.J LIVESEY

In the development cerebral neocortex, neurons are born on a predictable schedule. Studies in invertebrates have shown that individual progenitor cells undergo asymmetric divisions repeatedly and produce particular types of neuronal progeny at specific points in the lineage trees. In the cerebral neocortex, the first born neurons are reelin-positive Cajal-Retzius cells. Multipotent progenitor cells then produce other cortical neurons in inside-outer order: neurons destined for the deep layers 6 and 5 are born first, followed by those destined for the more superficial layers 4,3, and 2. It has been demonstrated that cortical progenitors preserve neurogenic timing in vitro and layer-specific neurons are born in appropriate order (1). Although the neural stem cells were shown to change neuropotency during development it still remains unclear if the population of early neuronal progenitor cells is heterogeneous consisting of a few different cell lineages, or it is homogeneous with high degree of developmental plasticity.

Here we have examined the diversity in gene expression profiles of early progenitor cells from mouse neocortex at day 11 of embryonic development (E11), when 98 % of cells are believed to be neural progenitors, possessing morphological and functional similarity. Previously it has been shown that microarray expression profiling based on global polyadenylated PCR-based amplification technique generates reliable data from picogram amounts of RNA (2), and that sampling effect (the random picking of the low abundant transcripts) is not significant for mentioned technique (paper in preparation).

Shortly, 12 single cells were extracted from mouse neocortex (E11), lysed, and cellular mRNA was undergone to PCR-based amplification following by hybridizations on expression microarray slides containing 23232 65-mer oligonucleotides (Sigma-Genosys). The statistical analysis reveled that neuronal progenitor cells demonstrate high heterogeneity which proven to be a result of the real differences in gene expression levels. Unsupervised clustering allowed dividing tested cells into two major groups according to their expression patterns. The most intriguing difference was found in transcription factor's expression levels. Notably, we discovered that that many transcription factors were up-regulated in first group and down-regulated in other cell's group, and vice versa. Among them

[1] Gurdon Institute and Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, CB2 1 QN, UK ts300@cam.ac.uk, rick@gurdon.cam.ac.uk

we identified regulatory genes, which play important roles in early neurogenesis and establishment of neuronal types such as Neurog2, Sox10, NeuroD1, Foxp4, Mash1, Math1, Eomes (Tbr2), Foxp1, Hoxa2 and number of other TFs which also may be significant in mammalian neurogenesis.

Recently, proneural genes NGN1/2 and Mash1 has been shown to be sequentially expressed in ventricular zone progenitors in distinct phases of the cell cycle (3). Notably, our microarray and real time PCR data also confirmed that NGN1/2 and Mash1 controversially up- and down-regulated in tested cells.

This study brings insight into the fundamental characteristics of mammalian neuron progenitor cells, demonstrating the high diversity in seemingly identical progenitors, and identifies the regulatory factors that may determine cell fates.

1. Shen Q et al., (2006) The timing of cortical neurogenesis is encoded within lineages of individual progenitor cells, Nat Neurosci. 9(6):743-51.
2. Subkhankulova T, Livesey FJ (2006) Comparative evaluation of linear and exponential amplification techniques for expression profiling at the single-cell level. Genome Biol, 7(3):18.
3. Britz et al., (2006) A role for proneural genes in the maturation of cortical progenitor cells. Cereb Cortex., 16: 138-151.

FUNCTIONAL ANNOTATION OF THE HUMAN GUT BACTERIAL METAGENOME

L.S. SYCHEVA[1], M. KAZANOV[2]

Introduction

Bacteria are organized in consortia, which may include few or many species.

In the past the research of bacterial consortia was hampered by the fact that most species could not be cultivated as pure cultures (e.g. 60 to 80 % of bacteria in human feces [1]). Some of these difficulties were resolved after emergence of new research technologies such as sequencing, 2D electrophoresis and mass-spectrometry.

One of the most interesting bacterial consortia is the human microbiote that forms a symbiotic organ covering the internal part of the gut, other mucous surfaces and skin.

The aim of this work was functional annotation of the bacterial metagenome from human gut.

Methods and Materials

DNA sequences of fosmids containing random clones of the human gut bacteria were kindly provided by M. Leclerc (INRA, France). There were 79 sequence fragments of the total size 3 Mb. The average size of a fragment was 35 kb.

DNA sequences were annotated using Artemis [2]. Gene functions were predicted by the analysis of best protein BLASTP [3] hits. to their translated sequences. The threshold for functional annotation was set to 30% of amino acid identitiy. Genes were assigned to clusters of orthologous genes (COGs) using Cognitor [4]. The taxonomy of a fosmid was assigned at the most detailed level, as which the taxonomies of the best hits to most genes from the fosmid were consistent. Genes encoding transport RNAs were predicted using tRNAScan [5].

Results

The human gut microbiote sample contained genome fragments of about 50 species. The bacterial consortium was rather heterogeneous and taxa were represented unevenly. The phyla most highly represented in the human gut microbiote were the Bacteroidetes (59 % of all annotated genes) and the Firmicutes (15 % of all annotated genes). The Enterobacteriales (e.g. Escherichia coli) were rather rare.

[1] Moscow State University, Department of Bioengineering and Bioinformatics, 119992, Moscow, Russia, lada.sychova@gmail.com
[2] Institute For Information Transmission Problems RAS, 127994, Moscow, Russia

One of the most popular gene functions in the Bacteroidetes were carbohydrate transport and metabolism. The annotated genes from this functional group encode a large spectrum of various carbohydrate cleavage enzymes. It is consistent with the theory that the human gut microbiote provides the essential part of carbohydrate metabolism processes in the human organism.

We are grateful to Marion Leclerc for sharing data prior to publication. This is joint work with Mikhail Gelfand.

1. L.J. Forney, X. Zhou, C.J. Brown (2004) Molecular microbial ecology: land of the one-eyed king. Microbiology, 7: 210-220.
2. K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M-A. Rajandream, B.Barrell (2000) Artemis: sequence visualisation and annotation. Bioinformatics, 16 (10): 944-945.
3. S.F. Altschul, T. L. Madden, A. A. Schlffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25: 3389-3402.
4. R.L. Tatusov, E.V. Koonin, D.J. Lipman (1997) A genomic perspective on protein families. Science, 278: 631-637.
5. T.M. Lowe, S.R. Eddy (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 25: 955-964.

OBJECT ORIENTATION AND BIOLOGICAL TAXONOMY: APPLYING PROGRAMMING CONCEPTS TO SPECIES CLASSIFICATION

DENIS TARASOV, E.D. IZOTOVA, N.I. AKBEROVA

Biological taxonomies are constructed with two purposes: practical (rapid identification of species and retrieving relevant information) and theoretical (revealing evolutionary history of species) [1]. Development of biological taxonomy has a long history starting perhaps from Aristotle and shaped to its present form by many works, including that of Linneus, de Candole, Hennig and many others. Modern taxonomy attempts to use data from diverse sources such as morphology, physiology, genomics, proteomics and metabolomics to obtain integrated picture of relationship between organisms.

[1] Kazan State University, 420008, Kremlevkaya 18, Kazan Russia,
dtarasov@mntech.ru

A vast variety of approaches to taxonomy can be (crudely) said to fall into 3 categories:

Intuitive taxonomy – species are classified by person with large experience in the area, based on intuitive view and informal reasoning. This approach allows to use data on different levels of complexity (from molecular to ecological) but lacks repeatability and objective measures of quality

Numerical taxonomy – species are classified by using a large numbers of equally weighted, noncorrelated characters. Procedure is repeatable and formally definable, but can not properly deal with complexity levels and semantics of different characters

Cladistics - arranges organisms by their order of branching in an evolutionary tree by using sets of shared derived characters. In current practice often applied in form of maximum parsimony analysis of short DNA sequence fragments that fails to take into account full complexity of living systems.

Contemporary biological nomenclature and structure of classification are governed by Nomenclature Codes evolved during last centuries and often criticized for containing to much "dead weight" of the past [2].

Our approach is based on the notion that ontological development of an organism can be viewed as computation determined by genetic program. Thus, in order to describe any given organism a computer program can be used instead of a simple list of characters.

It this work we propose to view species as object oriented programming classes, individual organisms as object instances and higher order taxa as abstract classes. This way, the whole taxonomical system can be viewed as object oriented framework (program) and the shortest possible program corresponds to the best classification. Although, such program is effectively unknowable [3], complexity of two (or more) classifications can be compared using standard software complexity metrics to determine their relative quality and suggest ways to further improvement.

To test our approach we applied it to taxonomy of microscopic fungi genus Trichoderma – a group of organisms with disputed taxonomical structure. We demonstrated how information about genetic structure, biochemical pathways and morphology of individual isolate can be described in form of ontogenetic development program written in object-oriented programming language and how such programs can be used for purposes of species classification and identification. The resulted object-oriented taxonomy is found to be more compact and less ambiguous than currently accepted system.

1. E. Mayr (1982). The growth of biological thought: Diversity, evolution and inheritance. Harvard University Press, Cambridge, Mass.
2. H.C. Godfray (2002) Challenges for taxonomy. The discipline will have to reinvent itself if it is to survive and flourish, Nature 417: 17-19
3. G. J. Chaitin (1987). Algorithmic Information Theory. Cambridge University Press.

KULLBACK-LEIBLER MARKOV CHAIN MONTE CARLO (KLMCMC) – AN ALGORITHM FOR FINITE MIXTURE ANALYSIS AND ITS APPLICATION TO GENE EXPRESSION DATA

TATIANA TATARINOVA[1], ALAN SCHUMITZKY[2]

In this presentation we describe Bayesian analysis of nonlinear hierarchical mixture models with a finite but unknown number of components. Our approach is based on Monte Carlo Markov Chain (MCMC) methods. One of the applications of our method is directed to the clustering problem in gene expression analysis. From a mathematical and statistical point of view, we will touch upon the following aspects:

1. Theoretical and practical convergence problems of the MCMC method;
2. Determination of the number of components in the mixture,
3. Computational problems associated with likelihood calculations.

In the existing literature, these problems mainly been addressed in the linear case. Developing a method for the nonlinear case is one of the main contributions of this work.

In the framework of mixture models the clustering problem is equivalent to the problem of determining which mixture component an observation is most likely to come from. Our approach is a combination of three previously developed methods:

1. Birth-Death MCMC (BDMCMC) approach outlined by Mathew Stephens (1997),
2. Random Permutation Sampler (RPS) by Sylvia Fruhwirth-Schnatter (2001),
3. Choosing the optimal number of components using the weighted Kullback-Leibler distance by Sahu and Cheng (2003), and
4. Relabelling strategy developed by Mathew Stephens (1997)

[1] Ceres, inc 1535 Rancho Conejo Road, Thousand Oaks, CA, 91320, USA
ttatarinova@ceres-inc.com
[2] Department of Mathematics, University of Southern California, Los Angeles, CA, USA 90007

We address both theoretical and practical aspects of convergence, label switching and model parameter selection.

We illustrate the KLMCMC using both simulated and real-life (yeast expression time series) datasets and compared the performance of the algorithm with existing methods.

1. Mather Stephens (1997). Bayesian Methods for Mixture of Normal Distributions. PhD thesis, University of Oxford.
2. Sylvia Fruhwirth-Schnatter (2001), Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models, JASA, 96 (453): 194-209.
3. S.K. Sahu, R. Cheng (2003). A Fast Distance Based Approach for Determining the Number of Components in Mixtures, Canadian Journal of Statistics, vol 31(1):3-22.

STRUCTURAL AND FUNCTIONAL MAPPING OF PROTEINS AS A BASIS FOR MODELING OF INTER- AND INTRACELLULAR PROCESSES

A.A. TERENTIEV, N.T. MOLDOGAZIEVA, A.N. KAZIMIRSKY

Modeling of living cell is the intriguing task which requires accounting a very complicated network of biochemical pathways with involvement of thousands of proteins differing in their physicochemical properties, structure and functions. The most of these proteins are mosaic, multi-modular and polyfunctional ones [1]. Each individual module of these proteins may function independently through binding to specific cell-surface receptor. At that, the same module may be a constituent of different non-homologous, unrelated proteins. For example, EGF-like modules and their repeats have been revealed in a whole number of extracellular matrix proteins, coagulation factors, membrane-associated proteins, etc. and may be responsible for their participation in regulation of cell proliferation and differentiation. Another example is tripeptide RGD which has been found in different cell adhesion proteins and provides their interaction with integrins. Cooperative and coordinated action of different multi-modular and polyfunctional proteins provides fine regulation of complexity of inter- and intracellular processes.

The method of structural and functional mapping of proteins presented here may be a useful tool to describe interrelationships between multi-modular and

[1] Russian State Medical University, 117497 Ostrovityanova street, 1, Moscow, Russia, nmoldogazieva@mail.ru; aaterent@mtu-net.ru

polyfunctional proteins [2, 3]. Structural and functional mapping of proteins implies searching of possible functionally active sites in their primary and/or spatial structure by comparison of structures of different physiologically active proteins available in databases Swiss-Prot/TrEMBL or Genbank. For comparison of primary structures, local alignment programs (BLAST, FASTA) may be used. The motifs revealed then may be chemically synthesized and tested for biological activity.

Alpha-fetoprotein (AFP) is the first protein primary structure of which is mapped. It is a multi-modular and polyfunctional protein containing to date more than twenty functionally important sites with proposed or experimentally confirmed biological activity. In its domain I (amino acid residues (aa) 2-187) the following sites were localized: cyclin-binding motif 1 RTLHR (aa 1-5), where H also binds Ni and Cu ions; EGF-like motif 1 LDSYQCT (aa 14-20), where C does not participate in disulfide bridging and contains a free SH-group; heavy metal-binding sites 1 (aa 19-39) and 2 (aa 51-71); fatty acid-binding site 1 (aa 42-62); apoptosis-regulating peptide 1 (aa 79-102); TGF-ß1-like motif 1 (aa 123-125); bilirubin-binding site 1 (aa 136-148); epitopic site SKAENAVE (aa 175-182). In domain II of AFP (aa 194-379) the following sites are located: cell-adhesion motif RGD (aa 262-264); fatty acid-binding site 2 (aa 209-228); apoptosis-regulating peptide 2 (aa 224-237); glycosylation site (N233); histidine-rich site VAHVHEHC (aa 244-251); bilirubin-binding site 2 (aa 261-277); EGF-like motif 2 IMSYICS (aa 266-272); cyclin-binding motif 2 (aa 312-329); hetero- and homodimerization motif 1 (aa 340-361). In domain III of AFP (aa 386-577) the following sites were identified: fatty acid-binding site 3 (aa 419-438); major estrogen-binding site (aa 428-449); growth inhibitory peptide (GIP) (aa 446-479); minor estrogen-binding site (aa 458-471); apoptosis-regulating peptide 3 (aa 463-478); hetero- and homodimerization motif 2 (aa 497-560); segment of major histocompatibility complex GVALQTMKQ (aa 524-532).

As a whole, more than 400 aa (about 70%) are in functionally important sites and part of the sites are overlapping. Peptides constructed on the basis of the biologically active segments of proteins may be used as leads for targeted biopharmaceutical agents.

1. N.T. Moldogazieva, A.A. Terentiev (2006) Alpha-fetoprotein and growth factors. Structural and functional relationships and analogies. [Adv. Biol. Chem.] (article in Russian), 46: 99-148.

2. G .J. Mizejewski (2001) Alpha-fetoprotein structure and function: relevance to isoforms, epitopes, and conformational variants, Exp. Biol. Med, 226: 377-408.
3. A.A. Terentiev, N.T. Moldogazieva (2006) Structural and functional mapping of alpha-fetoprotein. Biochemistry (Moscow), 71: 120-132.

NPIDB, A DATABASE OF STRUCTURES OF NUCLEIC ACID – PROTEIN COMPLEXES

M.L. TITOV[1], A.V. ALEXEEVSKI[2], S.A. SPIRIN[2], A.S. KARYAGINA[3]

The resource NPIDB (Nucleic acids – Protein Interaction DataBase) includes a collection of files in the PDB format containing structural information on DNA-protein and RNA-protein complexes, and a number of online tools for analysis of the complexes. Those tools are: an original program CluD [1] for analysis of hydrophobic clusters on interfaces, a program for detecting potential hydrogen bonds, visualization of structures with Jmol (http://jmol.sourceforge.net/). SCOP [2] and Pfam [3] domains presented in protein chains of structures are detected.

Structures of protein – nucleic acid complexes are extracted from PDB as files in the PDB format representing both asymmetric units (PDB entries "as is") and biological units. Structures are revised by our experts in order to correct possible mistakes (such as duplication of atoms) and inconvenience (such as two or more variants of a structure posed in one coordinate space, see, for example, PDB entry 1QPI, where two variants of each DNA chain are superimposed). The manually corrected entries are also included into the Database as "revised biological units". Thus, in some cases, the NPIDB content differs from the original PDB one; in particular, for some complexes (1FJL, 1QPI, etc.) there are some additional biological units in NPIDB compared with the PDB. All structural files of NPIDB are available for download.

Update of the content is done weekly by a special program module.

NPIDB is available via Internet: http://monkey.belozersky.msu.ru/NPIDB/. The main part of the web interface is the list of available structures. The list can be ordered according to PDB code, date of creation, type (DNA-protein or RNA-

[1] Institute of Agricultural Biotechnology, 42 Timiryazevskaya st., Moscow, 127550, Russia, mlt@iab.ac.ru
[2] Belozersky Institute of Physical and Chemical Biology, Moscow State University, Moscow, 119992, Russia, sas@belozersky.msu.ru
[3] Gamaleya Institute of Epidemiology and Microbiology, 18 Gamaleya st., Moscow, 123098, Russia, akaryagina@gmail.com

protein), or author's classification (HEADER row of the PDB file). Each NPIDB entry has its own web page, containing general information, links to other resources (such as PDBsum, NDB, etc.), a table describing biological units, tables describing Pfam and SCOP domains of the presented protein chains, and the list of available actions (including Jmol visualization of the complex).

The web interface contains also the lists of all presented Pfam and SCOP domains. Each domain type has its own web page containing the list of entries that include domains of that type. A collection of representatives of Pfam domain types is available for download. For each domain type, that collection contains a PDB-format file describing a fragment of a protein chain representing the domain together with the fragments of nucleic acid chains that are in contact with the protein domain.

The work is supported by the Russian Foundation of Basic Research, grants 06–07–89143 and 06–04–49558, and INTAS, grant 05–1000008–8028.

1. A.Alexeevski et al. (2003) CluD, a program for determination of hydrophobic clusters in 3D structures of protein and protein-nucleic acid complexes, Biophysics 48 suppl. 1, S146–S156.
2. A.G.Murzin et al. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540.
3. R.D.Finn et al. (2006) Pfam: clans, web tools and services, Nucleic Acids Research 34, Database issue, D247–D251.

JUDGMENT ALGORITHM FOR DETECTION OF PERIODICITY AND ITS APPLICATION

DAISUKE TOMINAGA, KATSUHISA HORIMOTO[1]

Judgment whether a time series of biological data has periodicity or do not is important and done widely to find circadian genes, check montly change of hormons, etc. For gene expression, DNA microarrays are very popular and its costs are becoming more reasonable today. Many sets of time series data are published on the web. Most of these published microarray data contain upto 20 or 30 sampling points for each gene. However, many sampling points are needed to analysis its time series generally, choice of reliable methods are restricted. For example, auto correlation analysis, auto regression model, Fourier

[1] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, AIST annex CBRC, 2-24 Aomi, Koto, Tokyo 135-0064, Japan, tominaga@cbrc.jp

transform analysis, and curve fitting are often used. All these methods need many sampling points to reduce influence of errors on reliability of analyses, models or fittings. These methods also need criteria to judge the data has periodicity or do not. Criteria are threshold of correlation coefficients, difference between models and given data, amplitude of Fourier spectra, or accuracy of curvefitting. These values are determined by persons analyze given data according to characteristics of data and his/her experiences.

On the other hand, microarray data often contain over ten thousands of time series of gene expression. Periodicity judgment criteria for each gene expression time series must be determined depending on its signal/noise ratio, average expression level, distribution of expression levels, etc, however, it is impossible actually for microarray data. Automatic criterion determination is strongly needed.

We developed a judgment method by combining Bayesian Information Criterion (BIC) and Discrete Fourier Transform (DFT), named 'piccolo', and proved its ability by applying to microarray data of mice to find circadian genes. Our methods shows higher sensitibity and specificity than simple DFT and a curve fitting method.

We applied the piccolo algorithm for detection of circadian genes to two kinds of data: one is a set of time series data generated by normal distributed random number, the other is a series of DNA micro-array data (GDS404 on the GEO database of NIH, USA, mice published on the web). Piccolo is compared with the cosinor method and simple DFT. The sensitivity (true positive divided by sum of true positive and false negative) for known seven circadian genes in the case of GDS404 of the cosinor method, simple DFT, and piccolo are 2/7=0.286, 5/7=0.714, 6/7=0.857, respectively. p-value analysis for GO terms of detected and all genes also shows piccolo's high sensitivity (p-value of Simple DFT and piccolo are 0.12 and 0.057). We present the piccolo algorithm for detection of circadian genes through the combination of DFT and BIC. From studies on both real and random data, we show evidence suggesting piccolo is more accurate than either DFT alone or the popular cosinor method.

1. Y. Sakamoto, et. al. (1986) Akaike Information Criterion Statistics (D. Reidel Publishing Company, Tokyo, Japan)
2. G. Schwarz (1978) Estimating the dimension of a model, Annals of Statistics, 6: 461-464.

BIOPHYSICAL METHODS IN BIOINFORMATICS: CLASSICAL MOLECULAR MECHANICS AND FUNCTIONAL RESIDUES IN PROTEINS

IVAN TORSHIN

Unified models for analysis of structure-function relationship in proteins and other macromolecules are essential for post-genomic biology, biochemistry and biophysics [1]. In particular, such models can, potentially, be used for in silico characterization of the novel gene products as well as for characterization of the protein structures established in the 'structural genomics' (structural proteomics) initiatives. A new physico-chemical model of biomacromolecular function called "molecular energetic profiles" [2,3] is presented here. The method is based on molecular mechanics calculations and with subsequent analysis of the relative contributions of the amino acid residues to the protein stability. Important functional residues in proteins are characterized by unusual energy values and the results of calculations allow prediction of the functional sites in proteins of different biochemistry (enzymes, proteins involved in protein-protein interactions, metal, DNA-, hem- and fatty-acid binding proteins) and structure (all-alpha, alpha/beta, all-beta and ion channels). The 1st test set included about 20 proteins with functional regions annotated in considerable detail on the base of previously published biophysical and biochemical data. On average for this test set, 72% of the topmost destabilizing residues and 64% of the topmost stabilizing residues of the proteins were either known functional residues or were spatially clustered around the known functional site regions.

One of the remarkable features of the present method is that it can be applied not only to the spatial (3D) protein structures but also to the lower levels of the structural organization of proteins. This feature of the method is extremely important since to obtain spatial structure requires protein purification, crystallization, Xray/NMR/synchrotron data collection, solving the three-dimensional structure of the protein and only then characterizing the solved structure. Here, we present application of the method of the molecular energetic profiles which does not require any data on the spatial organization of a protein except the secondary structure. Using the same test set, we show that, on average, 55% of the most destabilizing residues and 55% of the most stabilizing residues in calculations without spatial structure of proteins were either functional or were spatially clustered around the known functional sites. The results of calculations performed on 1st test set as well as on the 2nd set of 216

[1] Private consulting, Moscow, 125239, Russia, tiy135@yahoo.com

proteins were also compared with the results of the classical technique of bioinformatics: analysis of sequence identities. The results indicate that the residues calculated to have extremal energetic properties also tend to have lower values of sequence entropies in calculations performed on the most extended conformation ($P<0.001$ for the trend). Thus, the method of molecular energetic profiles can be used for prediction of the functional residues not only on the base of the 3D structures of proteins but also on the base of the amino acid sequence and the secondary structure prediction. It is important to notice that this result can be achieved without relying on any kind of sequence similarities with known proteins. Biophysical, biochemical and cellular aspects pertaining to the basics of the method are discussed. As the functional residues form the biophysical basis for the higher levels of the biological function of any protein, the method can have a wide range of potential applications including annotation of proteins, rational protein engineering and, in perspective, discovery of the genetic markers for the phenotype-genotype association studies.

1. I.Y. Torshin (2006). Bioinformatics in post-genomic era: the role of biophysics, pp1-10 (Nova Biomedical Books, NY, USA).
2. I.Y. Torshin (2004). Computed energetics of nucleotides in spatial ribozyme structures. The Scientific World JOURNAL; 4:228-247.
3. Torshin IY (2005). Computed energetics of macromolecules. Part I: identification of the functional residues in spatial structures of proteins and RNA. In: Bioinformatics: New Research, P. Yan (Ed), 101-126 (Nova Science Publishers, NY, USA).

STORIES ABOUT THE EVOLUTION OF REGULATORS: HOW FRUR BECAME CRA AND HOW RBSR BECAME PURR

OLGA TSOY[1], W. ZAKIRZIANOVA[1], DMITRY A. RAVCHEEV[2]

The increased number of completely sequenced bacterial genomes now allows one to perform detailed comparative analysis of regulatory interactions and reconstruct evolution scenarios for various regulatory systems.

We analyzed the evolution of the three regulatory systems, FruR, PurR and RbsR, in gamma-prteobacteria. All these systems were previously well-studied experimentaly in Escherichia coli 1, 2, and PurR and RbsR regulation was investigated in some representatives of gamma-proteobacteria by the comparative genomic approach 3, 4.

To consider these systems in more detail, we analyzed evolution of these regulons in the genomes of 19 gamma-proteobacteria from the Enterobacteriales (E. coli, Salmonella typhi, S. typhimurium, Yersinia pestis, Y. pseudotuberculosis, Erwinia carotovora, Photorhabdus luminescens), Pasteurellales (Pasteurella multocida, Haemophilus ducreyi, H. influenzae, H. somnus, Mannheimia succiniciproducens,), Vibrionales (Vibrio cholerae, V. fischeri, V. parahaemolyticus, V. vulnificus, Photobacterium profundum), and Pseudomonadales (Pseudomonas aeruginosa, P. putida, P. fluorescens, P. syringae).

In Esherichia coli, transcriptional factorFruR (Cra) regulates a number of genes encoding enzymes involved in sugar catabolism and anabolism. Our results demonstrate that such situation is typical only for the Enterobacteriales, where FruR controls expression of the most genes involved in the glycolitic and gluconeogenic pathways.

In the Pasteurellales genomes the FruR regulon is decaying. In the Vibrionales and the Pseudomonadales, FruR functions as a regulator of the single fruBKA operon. Most likely, FruR was a local regulator of the fru operon in the gamma-proteobacterial ancestor, but then the regulon expanded to glycolisis/gluconeogenesis and then to other carbon metabolic flux pathways and phosphotransferase systems.

A slightly different situation was observed for the pair of homologous regulators RbsR and PurR. Both regulators were found in the Enterobacteriales, Pasteurellales and Vibrionales.

[1] Lomonosov Moscow State University, Moscow, Russia, borh@bk.ru
[2] Institute for Information Transmission Problems, RAS, Moscow, Russia, ravcheyev@iitp.ru

In the Pseudomonadales, only one copy of the transcriptional factor was found. This protein was equally similar to both PurR and RbsR, and was co-localised with genes for ribose utilisation. Analysis of the predicted regulatory regions of this operon allowed us to detect a conserved motif that predicted to be a putative binding site of the Pseudomonadales transcription factor. Although the regulatory protein in Pseudomonadales resided in the ribose operon, as RbsR in E. coli, its binding site was more similar to the PurR binding motif.

Thus, we propose that PurR and RbsR originated from a common ancestor protein whose function was similar to the RbsR. After the duplication, one copy, RbsR, conserved the function but changed the ligand motif, while the other copy, PurR, conserved the motif but changed the ligand specificity and became the regulator of the purine biosynthesis and linked metabolic pathways.

This is joint work with Mikhail Gelfand. We are grateful to Andrey Mironov for kindly provided software. This study was supported by grants from the Howard Hughes Medical Institute, the Russian Academy of Science (under program “Molecular and Cellular Biology”), and INTAS.

1. M.H. Saier, T. M. Ramseier (1996) The catabolite repressor/activator (Cra) protein of enteric bacteria, Journal of Bacteriology, 178: 3411-3417.
2. H. Zalkin, P. Neidhardt (1996) Biosynthesis of Purine Nucleotides, In Escherichia coli and Salmonella. Cellular and Molecular Biology. F.C. Neidhart (Eds.), 1325-1333 (ASM Press).
3. D.A. Ravcheev et al. (2002) Purine regulon of gamma-proteobacteria: a detailed description, Genetika, 38:1203–1214.
4. O.N. Laikova et al. (2001) Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria, FEMS Microbiol Lett, 205:315-322.

HYDROPATHY OF HUMAN PRE-MRNA SPLICE SITES

A.S. TURMAGAMBETOVA[1], G.F. BOLDINA[1], A.T. IVASHCHENKO [1]

Revelation of conservative nucleotides sequences in the 5' (5'-S) and the 3' (3'-S) sites of exons and introns promoted understanding of mechanism of splicing. The properties of the donor splice site (5'-SS) formed by the 3' exon and the 5' intron sites and also the acceptor splice site (3'-SS) formed by the 3' intron and the 5' exon sites define splicing efficiency. There are different meth-

[1] Kazakh National University named after al-Farabi, al-Farabi av., 71, Almaty, 050038, Kazakhstan, a_ivashchenko@mail.ru

ods of representation of the 5'-SS and the 3'-SS by the way of alphabetic characters of nucleotides [1, 2, http://genes.mit.edu/pictogram.html]. Nucleotide content varies near canonical dinucleotides in the 5'-S and the 3'-S of introns, it confirms that several nucleotides make contribution on splicing process [3, 4]. One of the characteristics of recognition of the 3'-SS of human pre-mRNA is the availability of the polypyrimidine sequence (PPS) placed upstream of the AG dinucleotides in the 3'-S of introns.

Common disadvantage using of pictogram utility is absence of criteria reflecting quantitative change of the 5'-SS and the 3'-SS properties at a variation of nucleotide composition. As it has been shown in some researches so far hydropathy of intron sites plays an important role during the first step of splicing, however quantity assessment has not been done [3, 4]. We have used hydropathy coefficient of nucleotides to characterize splice sites.

Genes have been distributed into 15 groups and hydropathy profiles of the 5'-SS and the 3'-SS were determined. Genes containing only 1-3 introns have been taken from 1, 4, 13, 19, 21 and 22 chromosomes. Conservative hydropathy profile of the 5'-SS and the 3'-SS was noticed in all groups containing 9-59 genes. Hydropathy profile was provided by two canonical nucleotides of all the 3'-S of exons, except the terminal exons. All the internal exons also have similar hydropathy profile at +1, +2 positions in the 5'-S.

The 5'-S and the 3'-S of introns have the clearest consensus of nucleotides. The 5'-S and the 3'-S of introns are formed by six and five nucleotides accordingly. The hydrophobic site located upstream of the 3'-SS formed by approximately 10 nucleotides for the most part pyrimidine nucleotides and the 3'-SS form a conservative hydropathy profile which are recognizable by proteins and snRNAs [5, 6]. The expanded region between the 5'-SS and the PPS has weaker hydrophobic properties in comparison to the polypyrimidine sequence. Thus, heterogeneity of exon and intron hydropathy, including splice sites, creates conditions for primary recognition of the 5'-SS and the 3'-SS. The features of the 5'-SS and the 3'-SS containing dinucleotides (GU – AG, GC – AG or AU – AC) in the 5'-S and the 3'-S of U2-type and U12-type introns determine peculiarity of splicing.

It has been demonstrated that introns are more hydrophobic in comparison to exons, thus they mostly form a pre-mRNA core. Exons are mainly located on pre-mRNA surface as well as the 5'-SS and the 3'-SS that facilitate interaction and ligation of exons.

1. T.S.Schneider, R.M.Stephens (1990) Sequence logos: a new way to display consensus sequences, Nucl. Acids Res., 18: 6097-6100.
2. J.Gorodkin et al. (1997) Displaying the information contents of structural RNA alignments: the structure logos, Comput.Appl.Biosci., 13: 583-586.
3. P.L.Lim, C.B.Burge (2001) A computational analysis of sequence features involved in recognition of short introns, Proc. Nat. Acad. Sci. USA, 98: 11193-11198.
4. T.A.Thanaraj, F.Clark (2001) Human GC-AG alternative intron isophorms with weak donor sites show enhanced consensus at acceptor exon positions, Nucl. Acids Res., 29: 2581-2593.
5. J.Královičová, M.B.Christensen, I.Vořechovsky (2005) Biased exon/intron distribution of cryptic and de novo 3' splice sites, Nucl. Acids Res., 33: 4882-4898.
6. N.Sheth et al. (2006) Comprehensive splice-site analysis using comparative genomics, Nucl. Acids Res., 34: 3421-3433.

THEORETICAL STUDY OF THE EVOLUTION OF THE MOLECULAR-GENETIC SYSTEM CONTROLLING THE CELL CYCLE

I.I. TURNAEV, K.V. GUNBIN, L.V. OMELYANCHUK, V.A. LIKHOSHVAI

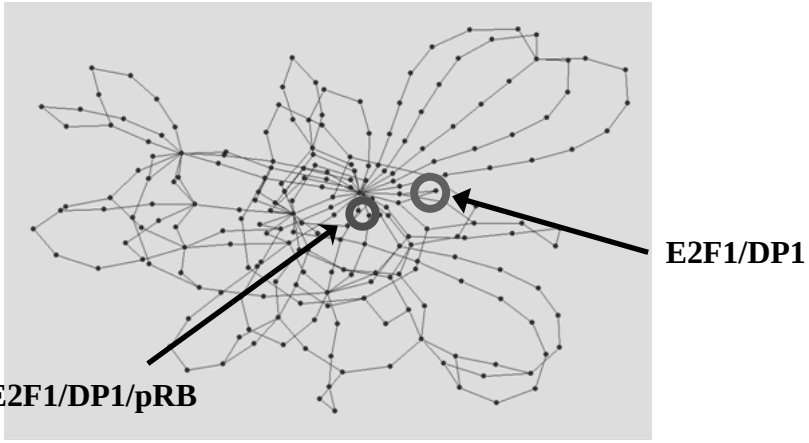
A comparative phylogenetic analysis of protein involved in cell division of 3 taxonomic groups (Eubacteria, Archaea, Eukaryota) was performed. Processes of the cell cycle were analyzed: temporal control of the cell cycle processes (cyclins), DNA replication, protein synthesis (transcription, translation), chromosome segregation, cytokinesis. As a result, a subsystem of the gene network of the cell cycle differing in the degree of evolutionary conservation was identified: 1) conserved processes common to pro- and eukaryotes (basal transcription, cytokinesis), 2) nonconserved processes that can be lost or acquired during evolution of pro- and eukaryotes (checkpoints mechanisms), 3) processes whose evolutionary changes are not dramatic (DNA replication end repair). The key factors in evolutionary conserved processes are proteins of the RNA-polymerase complex (36-48% similarity), tubulins in eukaryotes and proteins of Fts cluster in prokaryotes (44-54% similarity). Checkpoints proteins are characterized by weaker evolutionary conservation (no similarities between proteins or up to 30% similarity for individual protein kinases). An intermediate evolutionary conservation is a feature of DNA replication proteins (~30%

Institute of Cytology and Genetics, Novosibirsk, Lavrentyev aven., 10, Russia
turn@bionet.nsc.ru

similarity; between the DNA polymerase δ-subunits in Schizosaccharomyces pombe and Streptococcus agalactiae is 71%). Based on the protein structural similarities that reflect their relatedness [Robert et al., 2006], we hypothesize the existence of a common molecular-genetic mechanism of the cell cycle. It is independent of the control checkpoints which arose at the time eukaryotes appeared. For example, the gene network for the higher eukaryote cell cycle has a regulatory circuit, which includes homologous genes E2F1 and E2F6, the circuits are the checkpoints for the transition from G1 to the S-phase of the cell cycle (Figure). It is of importance that the E2F1 and E2F6 proteins now perform antagonistic functions.

In the context of the hypothesis, this element might have appeared during late evolution of eukaryotes as a superstructure over the basal system of cell cycle regulation. Previous studies on the dynamics of the function of molecular-genetics of the cell cycle in pro- and eukaryotes are consistent with the hypothesis. For example, the dynamics of the transcriptional activity of RNA polymerase II complex by the TFIID factor in eukaryotes is periodic [Yonaha et al., 1995]. The interaction of RNA polymerase with ribonuclease [Kim et al., 2006] and helicase RapA [Sukhodolets et al., 2000] may be involved in providing the periodic dynamics of the gene expression in prokaryotes.

Fig. Illustration of the role of the E2F protein family in the “Cell Cycle G0/G1-S” gene network [Ananko et al., 2005].



Within the framework of the proposed hypothesis, a minimum mathematical model for cell cycle dynamics was built and analyzed numerically. The study demonstrated the cyclic mode of gene expression functioning, and model it qualitatively reproduced the main parameters of the bacterial cell cycle [Allman

et al., 1991]. The model may be helpful in understanding the dynamics of cell cycle function and evolution.

Acknowledgements

Project "Evolution of molecular-genetic systems: computer analysis and modeling", Program of the RAS Presidium "Origin and evolution of the of biosphere". The authors are grateful to Dr. Ponomarenko M.P. for valuable discussions.

1. E.A. Ananko et al. (2005) GeneNet in 2005, Nucleic Acids Res., 33: D425-D427.
2. R. Allman et al. (1991) Cell cycle parameters of Escherichia coli K-12, J. Bacteriol., 173: 7970-7974.
3. J. Kim et al. (2006) Construction of an in vitro bistable circuit from synthetic transcriptional switches, Mol. Syst. Biol., 2: 68.
4. 4. F.D. Robert et al. (2006) Pfam: clans, web tools and services, Nucleic Acids Res., 34: D247-D251.5. M.V. Sukhodolets, D.J. Jin (2000) Interaction between RNA polymerase and RapA, a bacterial homolog of the SWI/SNF protein family, J. Biol. Chem., 275: 22090-22097.
5. M. Yonaha et al. (1995) Cell cycle-dependent regulation of RNA polymerase II basal transcription activity, Nucleic Acids Res., 23: 4050-4054.

MODELING OF PROTEIN-PROTEIN INTERACTIONS IN STRUCTURAL GENOMICS

ILYA VAKSER, ANDREY TOVCHIGRECHKO, ZHENGWEI ZHU, JAGTAR HUNJAN, ANATOLY RUVINSKY, YING GAO

Structural information is important for understanding protein interactions. Computational approaches are needed to generate protein structures, particularly for genome-scale studies, with experimental techniques providing representative templates for modeling, and are indispensable for understanding the molecular machinery of life.1,2

Protein docking. The computational approaches to structural prediction of protein-protein complexes (docking) have been rapidly developing. An important problem in structural genomics is recreation of the network of connections between proteins in a genome. The major aspects of this problem are: (1) the number of protein-protein interactions is very large, and (2) most protein

[1] Center for Bioinformatics, The University of Kansas, 2030 Becker Drive, Lawrence, KS 66047, USA, vakser@ku.edu

structures have to be models of limited accuracy. Thus, the structure-based methods for building this network have to be (a) fast, and (b) insensitive to the inaccuracies of modeled structures. Our docking program GRAMM has been shown to adequately address these issues.

Genome-wide modeling. Our GWIDD resource is a comprehensive environment for genome-wide structural modeling of protein-protein interactions. It contains interaction information for multiple organisms. The structures of the participating proteins are modeled or their crystallographic coordinates are retrieved and docked by GRAMM. The resource is not restricted to interactions in the GWIDD database - other sequences or structures may be entered at various stages. The system provides three different entry points: (1) Sequence: search for interacting proteins and retrieval of their sequences, (2) Structure: modeling of structures or retrieval of crystallographic coordinates for sequences selected in step one or for sequences entered directly, and (3) Docking: docking of protein structures obtained in the previous step or structures entered directly.

Environment for docking methodology development. A key element in designing better docking approaches is validation on experimentally determined structures (benchmark sets). Our Dockground project aims at development of a comprehensive public resource for design and validation of protein docking methodologies. The project includes four major integrated databases of protein-protein complexes: (a) Bound structures: comprehensive, regularly updated and curated database of co-crystallized complexes; (b) Unbound structures: the structures from the bound set in the unbound form, crystallized and simulated; (c) Modeled structures: the structures from the bound dataset refolded as models; and (d) Docking Decoys: predicted matches for protein pairs from (a), (b), and (c) datasets for validation of scoring techniques.

Principles of protein interactions: Energy landscapes. GRAMM docking procedures allows systematic sampling of intermolecular energy landscapes, revealing distribution of energy basins and their characteristics. The landscape analysis showed that, in general, the number of energy basins is small, they are well formed and correlated with actual binding modes, and the pattern of basins distribution depends on the type of the complex. The results indicated dependence of the funnel size on the type of the complex (smaller for antigen-antibody, medium for enzyme-inhibitor, and larger for the rest of the complexes) and the funnel size correlation with the size of the interface. Guidelines for the optimal sampling of docking coordinates, based on the basins distribution, basin size estimates, and the forcefield parameters were explored.

The described public resources are available at www.bioinformatics.ku.edu. The studies were supported by The University of Kansas and NIH grants R01 GM074255 and R01 GM061889.

1. G.R.Marshall, I.A.Vakser (2005) Protein-protein docking methods, In: Proteomics and Protein-Protein Interaction: Biology, Chemistry, Bioinformatics, and Drug Design, G.Waksman (Ed), 115-146 (Springer).
2. R.B.Russell et al. (2004) A structural perspective on protein–protein interactions. Curr. Opin. Struct. Biol., 14:313-324.

STRUCTURAL STUDIES OF PROKARYOTIC TRANSCRIPTION INTERMEDIATES

DMITRY G. VASSYLYEV

RNA polymerase elongation complex (EC) is both highly stable and processive, rapidly extending RNA chains for thousands of nucleotides. Understanding the mechanisms of elongation and its regulation requires detailed information concerning the structural organization of the EC. The 2.5Å resolution structure of the Thermus thermophilus EC revealed the post-translocated intermediate with the DNA template in the active site available for pairing with the substrate and shed significant light on the basic principles of transcription elongation.

We have also determined the 3.0Å resolution structures of the EC with a non-hydrolyzable substrate analog, AMPcPP, and with AMPcPP plus the inhibitor streptolydigin. In the EC/AMPcPP structure, the substrate binds to the active ("insertion") site closed through re-folding of the trigger loop (TL) into two □-helices. In contrast, the EC/AMPcPP/streptolydigin structure reveals an inactive ("pre-insertion") substrate configuration stabilized by streptolydigin-induced displacement of the TL. Our structural and biochemical data suggest that TL re-folding is vital for catalysis and have three major implications. First, despite differences in the details, the two-step, pre-insertion/insertion mechanism of substrate loading may be universal for all RNAPs. Second, freezing of the pre-insertion state is an attractive target for design of novel antibiotics. Finally, the TL emerges as a prominent target whose re-folding can be modulated by regulatory factors.

Department of Biochemistry and Molecular Genetics, University of Alabama at Birmingham, USA dmitry@uab.edu

DOCKING STUDIES ON ANTIVIRAL DRUGS FOR SARS

MR. VIRUPAKSHAIAH. DBM, MR. RACHANAGOUDA PATIL, MR. HEGDE PRASAD

The importance of Protein – Ligand binding in biological systems should not be underestimated. Protein – Ligand binding has an important role in the function of living organisms and is one method, which the cell uses to interact with the variety of molecules that comes in contact with it. Ligand is a molecule with a high level of specificity towards a particular type of proteins and is of prime importance in the cells functioning and survival.

Severe acute respiratory syndrome or SARS is a respiratory disease in humans which is caused by the SARS coronavirus. The treatment of coronavirus-associated SARS has been evolving and so far there is no consensus on an optimal regimen. The mainstream therapeutic interventions for SARS involve broad-spectrum antibiotics and supportive care, as well as antiviral agents and immunomodulatory therapy.

The Protein- Ligand interaction plays a significant role in structural based drug designing. Ligand is a molecule with a high level of specificity towards a particular type of proteins and is of prime importance in the cells functioning and survival.

In the present work we have taken the receptor Angiotensin converting enzyme 2 and identified the drugs that are commonly used against SARS (Severe Acute Respiratory Syndrome). They are Lopinavir, Ritonavir, Ribavirin, and Oseltamivir.

The receptor Angiotensin converting enzyme 2 (ACE2) was docked with above said drugs and the energy value obtained are as follows, Lopinavir (-292.3) , Ritonavir (-325.6), Oseltamivir (-229.1), Ribavirin (-208.8) .Depending on the least energy value we have chosen the best two drugs out of the four conventional drugs. We tried to improve the binding efficiency and steric compatibility of the two drugs namely Ritonavir and Lopinavir. Several modifications were made to the probable functional groups (phenylic, ketonic groups in case of Ritonavir and carboxylic groups in case of Lopinavir respectively) which were interacting with the receptor molecule. Analogs were prepared with using software Marvin Sketch and were docked using a docking software HEX against the same receptor the, energy value obtained are, Lopinavir analog 17 (-332.7), Ritonavir analog 12 (-330.8). From this we came to know that some of the modified drugs are better than the original drugs. Of these molecules Lopinavir

Depart of Biotechnology, Basaveshwar Engineering College, S.Nijalingappa, Vidhyanagar, Bagalkot-587 102, Karnataka, India, veerudbm1@rediffmail.com

analog (17), Ritonavir analog (12) are probable lead molecules than the rest of the drugs for SARS owing to their high-energy value.

Further work can be carried out to improve the steric compatibility of the drug based upon the work done above for a more energy efficient binding of the drugs to the receptor.

FUNCTION AND EVOLUTIONARY ANALYSIS OF THE T-BOX REGULON IN BACTERIA

A.G. VITRESCHAK[1], A.A MIRONOV[123], V.A. LYUBETSKY[1], M.S. GELFAND[12]

The bacteria use a wide range of regulatory mechanisms to control gene expression. While the most common regulatory mechanism seems to be regulation of transcription by DNA-binding proteins, there are other important mechanisms, in particular, regulation of transcription (by premature termination) and translation (by interference with intiation) via formation of alternative RNA structures in 3'-untranslated regions.

T-box antitermination is one of the main mechanisms for regulation of genes involved in the amino acid metabolism in Gram-positive bacteria. The T-box regulatory sites consist of conserved sequence and RNA secondary structure elements. Using a set of known T-box sites, we constructed the common pattern and used it to scan available bacterial genomes. The initial scanning using the RNA-pattern program identified about eight hundred T-boxes in ninety bacterial genomes. T-boxes were widely distributed in Gram-positive bacteria mainly in the Firmicutes, but also in the Actinobacteria. Moreover, T-boxes were found in some Gram-negative bacteria (δ-proteobacteria) and in other groups (Deinococcales/Thermales, Chloroflexi, Dictyoglomi). The majority of T-box-regulated genes encode aminoacyl-tRNA synthetases. Two other groups of T-box-regulated genes are amino acid biosynthetic genes and transporters, as well as genes with unknown function. Analysis of candidate T-box sites resulted in new functional annotations. We assigned the amino acid specificity to a large number of candidate amino acid transporters and a possible function to amino acid biosynthesis genes.

[1] Institute for Information Transmission Problems (the Kharkevich Institute), RAS. Bolshoj Karetny pereulok 19, Moscow, 127994, Russia
[2] Department of Bioengineering and Bioinformatics, M.V.Lomonosov Moscow State University. Vorobievy Gory 1-73, Moscow, 119992, Russia
[3] State Scientific Center GosNIIGenetika. Pervy Dorozhny proezd 1, Moscow 117545, Russia

We then studied the evolution of the T-boxes. Analysis of the constructed phylogenetic trees demonstrated that in addition to the normal evolution consistent with the evolution of regulated genes, T-boxes may be duplicated, transferred to other genes, and change specificity. We observed several cases of recent T-box regulon expansion following the loss of a previously existing regulatory system, in particular, the arginine regulon in Clostridium difficile and the methionine regulon in Lactobacillaceae. Finally, we describe a new structural class of T-boxes containing duplicated terminator-antiterminator elements.

COLLAGEN-LIKE PATTERNS IN THE HUMAN GENOME

VLASOV P.K., VLASOVA A.V., ESIPOVA N.G, TUMANYAN V.G.[1]

Collagen fibrillar proteins are essential protein family found in all animals that have mostly structural but also other functions [1-5]. Collagens are necessary for a consistent portion of the animal body and make up about 25% of the summary protein mass for mammalians [5]. Every collagen protein has the fibrillar (collagen) region with three helical chains coiled on each other, and a globular (non-collagen) region, but the exact size and the relative ratio of these regions vary in different collagens. In addition, collagen-like segments occur in other proteins and, apparently, empower these proteins with a variety of specific functions [6-7]. Collagen regular structure may be described by a simple template: (Gly-X-Y)n where X and Y are any amino or imino acid residue. This pattern corresponds to the specific periodicity of a nucleotide coding sequence where the glycine codon occurs every six nucleotides (two codons).The existence of this kind of sequence periodicity pattern leads to a unique sequence-structure interplay [8]. Given the importance of the collagen protein family and the high number and diversity of collagen genes, it is imperative to develop a method capable of searching and recognition of collagen-like segments in genome sequences.

Initial approaches of collagen gene scans in genomes were based on the standard BLAST algorithm [9]. The special “collagen sequence patterns” were applied to find collagen-like segments in bacterial genomes. However, any sequence-specific pattern restricts the sensitivity of alignments as it is difficult to construct a universal collagen sequence query and to take into account the high sequence variability of different collagen genes. Thus, there are clear limitations of simple BLAST-based searches.

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia, vlasov@imb.ac.ru

However, the specific nucleotide periodicity of collagen genes mentioned above can be used to recognize similar patterns in the genome sequence. Here, we realize a new method of a thorough search for collagen-like patterns (CLPs) in any nucleotide sequence. Our approach correctly identified all annotated exons in the fibrillar region of collagen gene. Our program, CollagenFinder, unlike many standard gene prediction programs, can scan any nucleotide sequence regardless of length and nucleotide content to annotate the CLP-regions.

We annotated CLPs in all human chromosomes. The results were compared with the GenBank collagen annotation as well as with the results of a popular gene prediction program Genscan. A high level of correspondence of CLP prediction and collagen genes annotation provides evidence for a high rate of accuracy of our approach. Indeed, our program has recognized 85% of all collagen exons, which is better than 60% for GeneScan level. The prediction results show that our proposed approach can indeed improve the collagen gene prediction accuracy, and it is better to combine standard GeneScan method and our approach.

Our method marks CLP in the coding regions of many non-collagen proteins. Some of these proteins have annotated collagen fragments (i.g. acetylholinesterase), but most of CLP were founded in gene (exon/intron) regions that are not annotated as collagen containing proteins. The annotation of these CLPs gives addition information about the structural and functional roles of human proteins. Interestingly, many more CLPs were founded in intergenic regions. The functional or evolutionary role of these regions retains unknown. The results obtained denote existence of the strong specific periodicity through the human genome. According to our results, the human genome has numerous regions with 9-nucleotide periodicity that correspond to CLPs. Being of high specificity in respect of sequence predicted CLP regions may serve as useful markers for identification of various genome regions with divergent biological functions.

We summarized all CLPs predicted in the entire human genome. Additionally, we include information comparing our predictions with the GenBank annotation. Database of predicted CLPs for all human chromosomes is available on (http://strand.imb.ac.ru/Collagen-cgi/clp.cgi).

Acknowledgments

The authors are grateful to F. Kondrashov for critical reading of the manuscript and valuable comments. This work was supported by grant from the Russian Foundation for Basic Research (No 05-04-49625) and grants on Molecular and Cellular Biology and Fundamental Sciences – Medicine of Presidium RAS.

1. S. Ricard-Blum, F. Ruggiero (2005) The collagen superfamily: from the extracellular matrix to the cell membrane, Pathol Bio., 53: 430-42.

2. J. Myllyharji, K.I. Kivirikko (2004) Collagens, modifying enzymes and their mutations in humans, flies and worms, Trends in genetics, 20: 33-43.
3. K. Gelse, E. Poschl, T. Aigner (2003) Collagens - structure, function, and biosynthesis, Adv Drug Deliv Rev., 55: 1531-46.
4. K.I. Kivirikko (1993) Collagens and their abnormalities in a wide spectrum of diseases, Ann Med., 25: 113-26.
5. E. Vuori (1990) The Family Of Collagen Genes, Annu.Rev.Biochem., 59: 837-72.
6. K. Hakansson, K.B. Reid (2000) Collectin structure: a review, Protein Sci., 9: 1607-17.
7. K. Sastry, R.A. Ezekowitz, (1993) Collectins: pattern recognition molecules involved in first line host defense, Curr Opin Immunol., 5: 59-66.
8. C.M. Stultz (2006) The folding mechanism of collagen-like model peptides explored through detailed molecular simulations, Protein Sci., 15: 2166-77.
9. M. Rasmussen, M. Jacobsson, L. Bjorck (2003) Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins, The journal of biological chemistry, 278: 32313–32316,

CONTEXTUAL ORGANIZATION OF 3`-END CONTEXT OF TRANSLATION START SITE IN EUKARYOTIC mRNAs

O.A. VOLKOVA, A.V. KOCHETOV

5'-terminal part of protein coding sequences (CDS) in eukaryotic mRNAs is characterized by specific contextual organization. It may result from several factors including the usage of a protein-specific amino acids, preferable usage of optimal codons to provide mRNA with a high translation elongation rate, and a specific organization of the translation initiation signal. It is known that nucleotide sequence flanking AUG codon modulates its ability to be recognized as a translation start site by the scanning 40S ribosomal subunits [1-3]. However, interrelationship between these features and mRNA translation initiation efficiency was not investigated in detail.

We analyzed statistical deviations in amino acid frequencies at N-terminal positions 2-4 of proteins in Arabidopsis thaliana, Homo sapiens and Saccharomyces cerevisiae genes. It was found that the most frequent amino acids in pos. 2 (Ala, Glu, Gly, Asp) are encoded by a guanine-started codons. This might result from the functional significance of G at position +4 of CDS. Despite the valine-encoding codons were also started from G (i.e., GUN), this amino acid was found to be un-

Institute of Cytology and Genetics, Novosibirsk, Russia, ov@bionet.nsc.ru, ak@bionet.nsc.ru

derrepresented. Probably, this underrepresentation resulted from the negative influence of U in pos. +5 on the effect of G in pos. +4 [1]. Our results suggest that AUGGU combination can be unfavorable in vivo not only in mammalian, but also in plant cells. It was also found that yeast mRNAs are characterized by another contextual organization of start codon 3'-end context.

Interestingly, there was a significant difference between the representations of different synonymous codons of overrepresented amino acids in pos. 2. All alanine encoding synonymous codons were overrepresented that demonstrated the significance of this amino acid itself as a component of translation initiation signal. However, GCG was found to be most strongly overrepresented. It should be noted that GCG is characterized by lowest average frequency in mammalian CDS in comparison with other alanine-encoding codons: probably, this nucleotide combination has some specific advantage over other GNN codons.

Serine was found to be also overrepresented at 2nd position of CDS and UCG codon is especially overrepresented in all analyzed organisms. It is possible that UCG also increases AUG recognition - probably because some specific features of this triplet may abrogate the negative influence of U at position +4. However, this bias may also resulted from the importance of Met - Ser at N-end for the functional activity of some proteins.

Acknowledgments

This work was supported by the Programs of RAS (Dynamics of Gene Pools) and RFBR (grant No. 05-04-48207). We thank SD RAS (grant No. 5.3) for partial support.

1. Kozak M. (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. EMBO J., 16: 2482–2492.
2. Niimura Y, Terabe M, Gojobori T., and Miura K. (2003) Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. Nucleic Acids Research, 31: 5195-5201.
3. Sawant S.V., Kiran K., Singh P.K., and Tuli R. (2001) Sequence Architecture Downstream of the Initiator Codon Enhances Gene Expression and Protein Stability in Plants. Plant. Physiol., 126:1630-1636.

SEQUENCE-STRUCTURAL CHARACTERISTICS OF HUMAN MIRNAS

PAVEL S. VOROZHEIKIN[1], ALEXANDER YU. IVANISENKO[1],
ALEXANDER I. KULIKOV[12], IGOR I. TITOV[13]

MicroRNAs (miRNAs) are small RNAs of about 20-23 nt and, presumably, regulate as much as 10 % of human genes [1]. MicroRNAs are processed from pre-miRNAs which form hairpin structure, typically shorter than a hundred nt. In the cytoplasm pre-miRNA is cleaved by Dicer, RNase III type enzyme, and produces RNA duplex. In most cases one duplex strand is selected as mature miRNA while the other one is degraded. Otherwise mature miRNAs are generated from both duplex strands in partially complementary pairs. It is still unclear when a single miRNA or a miRNA pair will be produced. In this report we present a study of sequence-structural determinants of miRNA processing from pre-miRNA and an investigation of the miRNA genomic distribution.

First, we built genome-wide map of human miRNA localization. We found that miRNAs scatter in all chromosomes and the distribution of miRNAs between chromosomes is highly inhomogeneous. Using null-hypothesis of random miRNA localization within a chromosome, we confirmed the earlier observations that miRNAs tend to cluster and found a typical cluster size of about 400 bp. Genomic locations of CpG islands and miRNAs were compared.

Second, we analyzed pre-miRNA sequences to find sequence-structural determinants of miRNA excision from miRNA genes. Following the Hidden-Markov-Model approach for the problem of miRNA prediction [2], we found that in most cases it predicts the mature miRNA together with its complementary partner rather that a single miRNA. Besides, known miRNA pairs (about 20% of miRNA set used in our work) were found to be usually shifted by few nt. This shift probably reflects consecutive cleavage of two (of four) duplex termini by Dicer.

Finally the role of secondary structure and nucleotide context in selection of miRNA was investigated. Calculating base-paring probabilities of pre-miRNAs by partition function approach we found that pre-miRNAs tend to form loops in the center and near the boundaries of further miRNA duplex. Then we calculated nucleotide frequencies near the duplex boundaries and evaluated the rules of miRNA boundary recognition. We bring together our results building a

[1] Novosibirsk State University, Novosibirsk, Russia, pacha_1@ngs.ru
[2] Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia
[3] Institute of Cytology and Genetics, Novosibirsk, Russia, titov@bionet.nsc.ru

model of miRNA excision, where miRNA is processed from pre-miRNA in three consecutive steps: Dicer recognition of miRNA ends, duplex excision and stochastic end cleavage.

This work was supported by the Program of RAS and SD RAS "Origin and Evolution of Biosphere" (Contract number 10002-251/II-25/155-270/200404-082).

1. B. John et al. (2004) Human microRNA targets. PLoS Biol 2 (11), e363.
2. J.-W. Nam et al. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. NAR, 33 (11): 3570-3581.

COMPUTATION OF ELECTROSTATIC EFFECTS FOR MEMBRANE PROTON PUMP - BACTRIORHODOPSIN

KIRILL VOTYAKOV[1], ALEX KOBETS[2]

The most extensively characterized proton pump is bactriorhodopsin (bR). bR uses the light energy to transfer a proton from the cytoplasm of Halobacterium salinarum to the extra cellular space. In spite of the high resolution structural bR models available up to date the molecular mechanism of the bR functioning is still under discussion [1]. Many functionally important titratable sites are buried within the protein core. Therefore the results of theoretical calculations of the electrostatic effects may considerably improve the understanding of the bR functioning.

This work concerns some methodological questions on the computation of the electrostatic effects for the membrane proteins. The continuum electrostatics model is used [2]. In this model the bR and the lipid molecules of the membrane are treated as a low dielectric region in a high dielectric medium which represents the solvent. On the bases of the experimental X-ray structures the protein atoms are modeled as spatially fixed partial atomic charges embedded in the low dielectric region. The lipid region is modeled atomicless. The electrostatic potential is obtained from the solution of the Poisson–Boltzmann equation. The titration of a single protonation site is determined by its pKa value.

We investigate the dependence of the protonation probability of the bR titratable sites on the value of the protein dielectric permittivity and the quality of the used structural models. Two methods of the accounting of the protein

[1] MIPT, Dolgoprudny, Russia, Center of Biophysics and Physical Chemistry of Supramolecular Structure, kvotyakov@aqt.ru
[2] MIPT, Dolgoprudny, Russia, Chair of Computer Science, kobets@sw.ru

interior water molecules are compared. The calculated pKa values are compared with the experimental data.

The bR structures 1m01 [1.47 A], 1c3w [1.55 A], 1qhj [1.90A] and 1cqw [2.25 A] are used to estimate the influence of the structure quality on the correctness of the calculated pK'a values. The calculated pK'a for R82, D85,96,115 for the all used structures show closely related values ($\Delta pK < 0.5$ pH unit). The considerable difference ($\Delta pK \sim 3$ pH unit) is seen for D212, E196, 204 between 1cqw and the other models. The reason for it is the difference in the "background" energy term induced by the difference in the water molecules network in the vicinity of this residue. The calculated pK'a values for the key titratable sites show good correlation between each other and the experiment for the quality structures (1m01 and 1c3w).

The water molecules in the protein cavities are treated explicitly and implicitly to understand the necessary structure quality. The model with explicit water molecules shows stable right results in the wide range of permittivity from 2 to 10. For the model with implicit water molecules right protonation states are obtained only with high protein permittivity. The screening of stabilizing sites interactions is supposed to be the reason for it. Therefore the availability of the interior water network is extremely important for the accurate electrostatic calculations in the bR.

All numerical calculations of the Poisson-Boltzmann equation are done with the MEAD package [3]. We are thankful for numerous discussions to V. I. Gordeliy and S. Grudinin.

1. Luecke, Schobert, Richter, Lanyi (1999), Structure of bacteriorhodopsin at 1.55 angstroun resolution, JMB, 291:899-903.
2. D. Bashford and M. Karplus (1990), pKa's of ionizable groups in proteins: atomic detail from electrostatic model. Biochemistry, 29:10219-10225.
3. Bashford, D. et al. (1997), An object-oriented programming suite for electrostatic effects in biological molecules. ISCOPE97, 233-240, Berlin, , Springer.

CMDB: A DATABASE FOR COORDINATED MUTATIONS

YU.V.VYATKIN, D.A. AFONNIKOV

The effect of mutations on protein structure-function is an important issue in biology. Comparative analysis of homologous protein sequences is now in-

Institute of Cytology and Genetics, Novosibirsk, Lavrentyev aven. 10, Russia,
ada@bionet.nsc.ru

tensely used to resolve the issue. Study of compensated substitutions in proteins, which allows estimation of the effect of pairwise residue interaction in protein structure on amino acid substitution, is a promising approach. With reference to the pairwise interactions, novel methods are been intensely developed. However, approaches that integrate analysis of evolutionary, structural and biochemical information about proteins are imperative for a better understanding of the biological mechanisms that provide the coordinated mode of amino acid evolution. For a more complete analysis of coordinated substitutions in proteins, we developed the Correlated Mutations Database (CMDB). The idea underlying CMDB is integration of available information about protein evolutionary and structural features with the results of search of coordinated substitution mode.

A computational pipeline implemented in multiprocessor version was developed for filling up CMDB. The advantage of the pipeline is that it enables to perform parallel computations with a linear speedup that depends on the number of processors used. As a result, computational time is reduced by many times as compared with the single processor version. This pipeline makes possible identification of homologous sequences for proteins with known spatial structure, making of multiple alignments, finding of phylogenetic relationships in a protein family, also detection of coordinated substitutions using a set of methods [1-3]. The results for original protein sequences are integrated with those for structural annotations. These overall results are formatted so that they are convenient for storage in CMDB and queries in search of information about protein sequences according to protein identifier, homology, and PDB ID can be easily done. CMDB outputs information about multiple alignments of protein families, phylogeny, and the identified coordinated substitutions in text and image format.

Additionally, programs, which allow evaluation of the structural functional features of protein positions at which substitutions are coordinate, were developed. CMDB-assisted analysis was performed for interaction between coordinated evolutionary mode, features of physicochemical interactions of residues in protein structure, and structural characteristics of the residues.

Thus, CMDB proved to be a valuable computational tool for studying the mode of coordinated substitution mode in amino acid residues.

The work is supported by the Ministry of Education of the Russian Federation grant “Development of the Higher School Scientific Potential” 2.1.1.4935, Russian Foundation of the Basic Research (05-04-49141-a, 05-07-98012-p), SB RAS integration projects 49, State Contract №10104-34/П-18/155-270/1105-

06-001/28/2006-1. The computation was performed in part at the High Performance Computing Center, SB RAS.

1. D.A. Afonnikov, D.Y. Oshchepkov, N.A. Kolchanov (2001) Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. Bioinformatics, 17:1035-1046.
2. U. Gobel, C. Sander, R. Schneider and A. Valencia (1994) Correlated Mutations and Residue Contacts in Proteins Proteins, 18:309-317.
3. W.R. Atchley, K.R. Wollenberg, W.M. Fitch, W. Terhalle, and A.W. Dress (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol. Biol. Evol., 17, 164-178.

QUESTIONING THE ASSUMPTIONS: A STRATEGY FOR GRADUATE EDUCATION IN STATISTICAL METHODS FOR BIOINFORMATICS

SUSAN R. WILSON[1]

Statistical methods are very widely used in bioinformatics research, both in the development of new methodology and in the design of studies and analyses of the resultant data. What is the general standard of these methods? In medical research, wherein there is widespread evidence of the extensive use of statistical methods, standards are generally low.1 Although no comparable survey has been done for bioinformatics research, the anecdotal evidence is that the standards are not as high as they could be.

One way to improve these standards is through appropriate education. A useful strategy at the graduate level is to take a set of relevant research papers and start questioning the assumptions underpinning the approach and methods used therein. In the presentation a range of suitable problems will be overviewed. For illustration, this abstract briefly considers two papers.

Popular sequence comparison methods, such as BLAST, rely on local alignment. These methods assume contiguity between homologous segments. Is this assumption appropriate? Often not, so an alignment-free sequence comparison approach that uses as a test statistic D2, the number of matches of words of a given length k between two given sequences has been proposed. A two-sample Kolmogorov-Smirnov (K-S) test was implemented2 to evaluate the asymptotic distributional results for D2. What assumptions does the K-S test make? Several, the most serious limitation being that the distribution to which the data

[1] Centre for Bioinformation Science, The Australian National University, Canberra, ACT 0200, Australia, Sue.Wilson@anu.edu.au

are being compared must be fully specified. So if the parameters are estimated from the data, as they were for these simulations, the critical region of the K-S test is no longer valid. Does the size of this error matter? Yes it does, and we have shown elsewhere that other tests, such as the Shapiro-Wilk test for evaluating normality, give far more accurate results.

Signatures, such as lists of genes or scores derived from a set of gene expression values, are being proposed, particularly in cancer studies. How stable are these signatures? Using 10-fold cross-validation (CV), we recently re-analysed a subset of data from a relatively large breast cancer study3, and found that the resultant ten signatures were very unstable, having very few genes in common either with each other or with the original published signature. There are a large number of questions that can be asked of studies such as this. What is the data source for these samples? Briefly, the samples were a subset of frozen tumour samples from patients with lymph-node negative breast cancer that had been submitted for steroid-hormone receptor measurement from an intake of 25 hospitals. How heterogeneous are they? The data show considerable variation in many known breast cancer prognostic factors, such as therapy type, age, menopausal status, tumour grade and stage for example. Data on these factors were not made available on the web.

'Because society depends on sound statistical practice, all practitioners of statistics, whatever their training and occupation, have social obligations to perform their work in a professional, competent, and ethical manner' (from Ethical Guidelines for Statistical Proactice, 1999, American Statistical Association). Courses such as the one being developed here can ensure that computational molecular biologists know how to ask, and find answers to, appropriate questions of the methods and data appearing in the literature, and hence incorporate high quality statistical standards into their own research.

Acknowledgements: Drs C. Burden, S. Foret and Y. Pittelkow are involved with these projects.

1. A.M. Strasak et al. (2007) The American Statistician, 61: 47-55.
2. R.A. Lippert et al (2002) PNAS, 99:13980-13989.
3. Y. Wang et al (2005) Lancet, 365: 671-679.

ELECTRON-TRANSFER PATHWAYS IN NATIVE AND MUTANT GM203L BACTERIAL REACTION CENTERS

ANDREY G. YAKOVLEV[1], MICHAEL R. JONES[2], JANE A. POTTER[2], PAUL K. FYFE[2], LYUDMILA G. VASILIEVA[3], ANATOLI YA. SHKUROPATOV[3], VLADIMIR A. SHUVALOV[3]

The photosynthetic bacterial reaction center (RC) is a specialized pigment-protein complex within which a series of fast electron-transfer (ET) reactions occurs to convert light energy into the chemical free energy of charge-separated states. With regard to the mechanism of charge separation, special attention is deserved by a crystallographically-defined water molecule HOH55 (file 1AIJ, Protein Data Bank) located in the structure of the Rb. sphaeroides R-26 RC between PB and BA [1,2]. An important feature is that water HOH55 is within hydrogen-bonding distance of both the oxygen of the 131-keto carbonyl group of BA and the nitrogen of the residue His M202 that provides the axial ligand to the magnesium of the PB BChl [1-3]. Thus a direct link between PB and BA through HOH55 and His M202 seems to emerge from the crystallographic data. The X-ray crystal structure of GM203L mutant RCs has shown that replacement of Gly M203 by Leu caused exclusion of water HOH55 without gross changes in the protein structure outside the immediate vicinity of the M203 site.

This work has addressed the influence of water molecule HOH55 on electron transfer between the primary reactant P and the acceptor BA in reaction centers of purple bacteria [1-3]. This was done by comparing femtosecond time constants and oscillations in the kinetics of charge separation in native and GM203L mutant RCs of Rb. sphaeroides. We have previously shown [4-6] that in native and Pheo-modified RCs a frequency at 32 cm-1 is clearly observed in the FT spectra of the oscillatory component of the kinetics of population of the product states P+BA- and P+HA- (measured by monitoring the time evolution of absorbance changes at 1020 and 760 nm, respectively). A remarkable result of the present measurements is that the frequency at 32 cm-1 is completely absent in the FT spectra of GM203L mutant RCs obtained for the kinetics at 1020

[1] Department of Photobiophysics, Belozersky Institute of Chemical and Physical Biology, Moscow State University, Moscow 119899, Russia, yakov@genebee.msu.su
[2] Department of Biochemistry, School of Medical Sciences, University of Bristol, University Walk, Bristol BS8 1TD, United Kingdom, m.r.jones@bristol.ac.uk
[3] Institute of Basic Biological Problems, Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russia, shkur@issp.serpukhov.su, shuvalov@issp.serpukhov.su

and 760 nm. As the 32 cm-1 frequency corresponds to one of the frequencies of water molecule rotation [6], and because the HOH55 water is absent in the GM203L mutant, it is reasonable to conclude that, in native RCs, HOH55 has an effect on electron transfer from P* to BA.

According to the kinetic analysis, the decay of the P*-stimulated emission at 940 nm in GM203L mutant RCs follows a single-exponential with a time constant of approximately 4.3 ps. In native RCs, the kinetics of stimulated emission decay can be described by two exponents with time constant of approximately 4.3 ps (20%) and 1.1 ps (80%), respectively. We have considered two possible explanations for this slowing of P* decay in the GM203L mutant.

The first is that the mutation causes a change in the energetics of the primary reaction. The X-ray crystal structure of the GM203L mutant rules out a significant change in the edge-to-edge distance of P and BA, or in their relative orientations, but the mutation could cause a change in a parameter that affects the driving force for the reaction, such as the redox potential of P/P+ or BA/BA-, or a change in the polarity of the environment of P and BA that could affect the reorganization energy for the primary reaction. A second possibility is that the lack of water HOH55 in the GM203L RC results in the disappearance of the 1.1 ps electron-transfer component due to the removal of a favorable route for electron tunneling. The slower component (4.3 ps) could reflect other, less effective electron-transport pathways, probably related to non-specific electron tunneling from P* to BA. The position of water HOH55 is such that it could complete an electron-transfer chain involving the following polar atoms: N-Mg(PB)-N-C-N(HisM202)-HOH55-O=(BA) [6-8].

According to quantum-mechanical calculations [9], the maximal electron π-spin density in the P* state is localized on the nitrogen atoms that are coordinated to the central Mg of the PB molecule . This implies that the liganding Mg nitrogen atom of His M202 is also very close (about 2 Å) to the maximal spin density on PB. This would be consistent with the proposal that the aforementioned chain of polar groups represents a major (80%) route for transfer of an electron with a high rate (1/1.1 ps-1). When this chain is broken due to the absence of HOH55, the ET rate is strongly decreased (1/4.3 ps-1) and electron transport probably proceeds along other ('non-specific') pathways for electron tunneling. Such a fraction does not exceed 20% in native RCs but it is a major one in GM203L mutant RCs.

In conclusion, the results presented here provide unambiguous evidence that water molecule HOH55 has a strong influence on primary charge separation in native reaction centers of Rb. sphaeroides, and give insights into changes in detailed protein structure that occur on the time scale of charge separation.

This work was supported by a PCB RAS grant from the Russian Academy of Sciences, grant № 05-04-48554 from Russian Fund for Basic Research, and funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom.

1. U. Ermler, G. Fritzsch, S.K. Buchanan, H. Michel (1994) Structure 2: 925.
2. J. Deisenhofer, O. Epp, I. Sinning, H. Michel (1995) Mol. Biol. 246: 429.
3. M.H.B. Stowell, T.M. McPhillips, D.C. Rees, S.M. Soltis, E. Abresch, G. Feher (1997) Science 276: 812.
4. A.M. Streltsov, S.I.E. Vulto, A.Ya. Shkuropatov, A.J. Hoff, T.J. Aartsma, V.A. Shuvalov (1998) J. Phys. Chem. B 102: 7293.
5. A.G. Yakovlev, A.Ya. Shkuropatov, V.A. Shuvalov (2002) Biochemistry 41: 2667.
6. A.G. Yakovlev, A.Ya. Shkuropatov, V.A. Shuvalov (2002) Biochemistry 41: 14019.
7. V.A. Shuvalov, A.G. Yakovlev (2003) FEBS Lett. 540: 26.
8. A.G. Yakovlev, L.G. Vasilieva, A.Ya. Shkuropatov, T.I. Bolgarina, V.A. Shkuropatova, V.A Shuvalov (2003) J. Phys. Chem. A 107: 8330.
9. M. Plato, F. Lendzian, W. Lubitz, E. Trankle, K. Mobius (1988), In: The Photosynthetic Bacterial Reaction Center: Structure and Dynamics, J. Breton, A. Vermeglio, (Eds.), p. 379 (Plenum Press, New York and London).

CONFORMATIONAL CHANGES IN POLYPEPTIDES / PHASE TRANSITION

ALEXANDER YAKUBOVICH, I. A. SOLOV'YOV, A. V. SOLOV'YOV, WALTER GREINER

The phase transitions in finite complex molecular systems, i.e. the transition from a stable 3D molecular structure to a random coil state or vice versa (also known as (un)folding process) occur or can be expected in many different complex molecular systems and in nano objects, such as polypeptides, proteins, polymers, DNA, fullerenes, nanotubes.

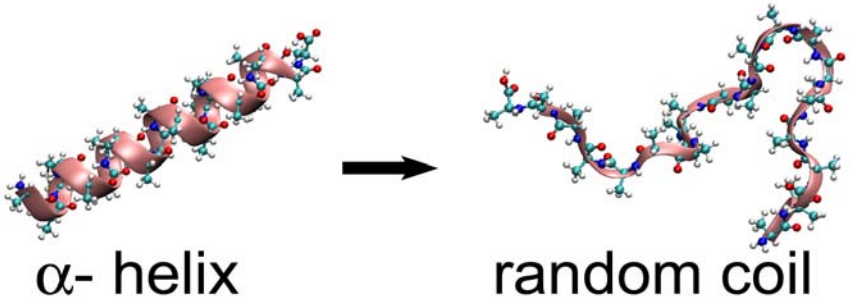


Fig. 1 The characteristic structural change of alanine polypeptide experiencing an α-helix↔random coil phase transition.

We suggest a novel ab initio theoretical method [1] for the description of phase transitions in the mentioned molecular systems. In particular, it was demonstrated that in polypeptides (chains of amino acids) one can identify specific, so-called twisting degrees of freedom, responsible for the folding dynamics of the amino acid chain, i.e. for the transition from a random coil state of the chain to its α-helix structure (see Fig. 1). The essential domain of the potential energy surface of polypeptides with respect to these twisting degrees of freedom can be calculated and thoroughly analysed on the basis of ab initio methods such as density functional theory (DFT) or Hartree-Fock method. It is shown [1] that this knowledge is sufficient for the construction of the partition function of a polypeptide chain and thus for the development of its complete thermodynamic description, which includes calculation of all essential thermodynamic variables and characteristics, e.g. heat capacity, phase transition tem-

Frankfurt Institute for Advanced Studies, Max-von-Laue str. 1 60438 Frankfurt am Main, Germany, ilia@fias.uni-frankfurt.de

perature, free energy etc. The method has been proved to be applicable for the description of the phase transition in polyalanine of different length by the comparison of the theory predictions with the results of several independent experiments and with the results of molecular dynamics simulations.

1. A. Yakubovich, I. Solov'yov, A. Solov'yov, and W. Greiner, (2006) Eur. Phys. J. D (Highlight paper), 40:363-367, Europhysics News, 38:10

ANALISIS OF GENETIC DIVERGENCE OF DIFFERENT VIPERA SPECIES (REPTILIA: VIPERIDAE, VEPERA) FROM GENE SEQUENTION OF CYTOCHROME OXIDASE SUBUNT III AND 12S RIBOSOMAL RNA

R.V. YEFIMOV[1], E.V. ZAVIALOV[1], V.G. TABACHISHAN[2]

In the present time the polymorphism of the nucleotide sequence of the mitochondrial genes is used for the study of evolution processes and phylogenetic reconstruction of different genus and species of animals. Such unique properties as the maternal inheritance of characters, absence of recombination and the high level of variability allow to use the mitochondrial DNA as the high-quality instrument of genetic analysis.

The method elaboration for the investigation of the nucleotide sequence of the mitochondrial genes for research of evolution processes and phylogeography by the example of different adder species is the purpose of this study.

The nucleotide sequence of the mitochondrial genome fragments, including the genes of 12S ribosomal RNA and cytochrome oxidase subunt III of some adder species: V. nikolskii, V. berus, V. renardi from different habitat of Chuvash and Mordovia Republic, Volgograd, Saratov, Samara, Penza regions was determined. The size of the sequences obtained was 669 and 674 nucleotide pairs, respectively. This makes up practically the full nucleotide sequence of the said genes except short end fragments.

All samples fall into two groups as a result of compare of nucleotide sequences. In one of them there were specimens of V. nikolskii from Saratov region and in other – V. berus and V. Nikolskii from Chuvash and Mordovia Republic, Volgograd, Samara, Penza regions

[1] Chernyshevsky Saratov State University, Russia, 410012, Saratov, Astrakhanskay str., 83, EfimovRV@Rambler.ru
[2] Saratov branch of A.N. Severtsov Institute of Ecology and Evolution RAS Russia, 410028, Saratov, Rabochaya str., 24

The results of this investigation revealed that populations of V. nikolskii from Saratov region are high-avidity by genetic features. This fact allows to draw the north boundary of specific natural habitat of V. nikolskii on the valley of Saratov reservoir. Adders of Middle Volga, which spatially confined to the territory V. berus, must be attribute to the last species.

HOW THE STRIPES ARE PAINTED: FEED-FORWARD MECHANISMS OF DEVELOPMENTAL PATTERN FORMATION IN DROSOPHILA

ROBERT ZINZEN, MICHAEL LEVINE, DMITRI PAPATSENKO

Development of an organism from a single cell is a progressive gain of spatial system complexity/spatial system information in time. Large fraction of the spatial information is encrypted in regulatory DNA sequences of developmental genes. How the discrete sequence information "unfolds" into the complex tissues and body parts is one of the central problems in systems biology.

I will describe quantitative models for spatial gene expression based on gene response to transcriptional signals – maternal morphogenes. On the example of genes involved into formation of Anterior-Posterior and Dorso-Ventral embryo polarity, I will demonstrate spatial outcome of antagonistic protein gradients in embryo. From the point of formal network analysis, the interactions between spatial gradients of developmental genes may be explained by feed-forward mechanisms[1,2].

Observed feed-forward interactions between Dorso-Ventral and Anterior-Posterior gradients in fly embryo explain formation of spatial gene expression patterns in many instances, even in seemingly unrelated systems, such as eye-spot patterns in butterfly wings.

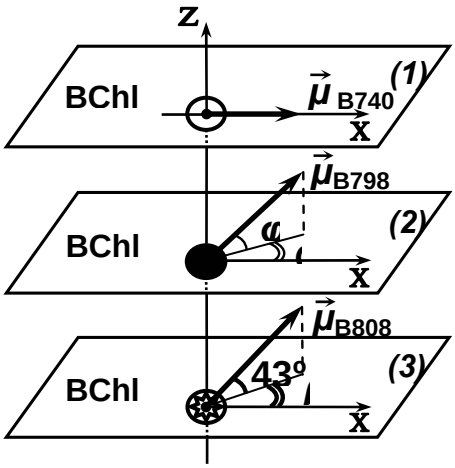
1. Zinzen RP, Senger K, Levine M, Papatsenko D. (2006) "Computational models for neurogenic gene expression in the Drosophila embryo" *Curr Biol.*, 16(13):1358-65.
2. Zinzen R, Papatsenko D. (2007) "Enhancer Responses to Similarly Distributed Antagonistic Gradients in Development" *PLoS Comput. Biol.*, e84.eor doi:10.1371/journal.pcbi.0030084.eor

University of California, Berkeley, USA dxp@berkeley.edu

OPTIMAL STRUCTURAL COORDINATION OF LIGHT-HARVESTING SUBANTENNAE AS AN EFFICIENT STRATEGY FOR LIGHT HARVESTING IN PHOTOSYNTHESIS. MODEL CALCULATIONS

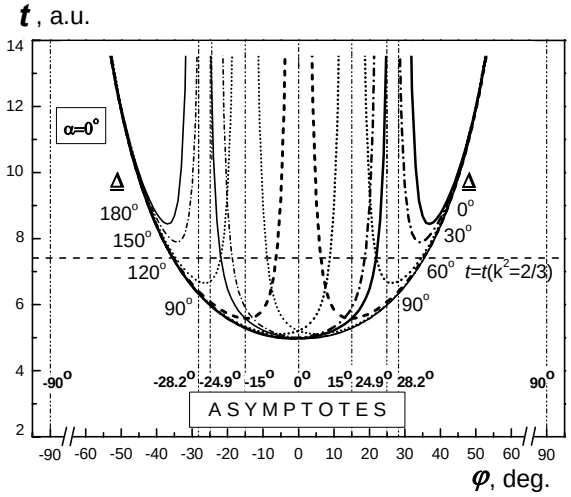
A.V. ZOBOVA[1], A.C. TAISOVA[2], Z.G. FETISOVA[2]

This work continues a series of our investigations on efficient strategies of functioning of natural light-harvesting antennae, initiated by our concept of rigorous optimization of photosynthetic apparatus structure by functional criterion [1]. Using computer modeling for the functioning of the natural antennae, we suggested some basic principles for designing optimal model systems. Targeted searches for these principles in vivo allowed us to recognize some of them in natural antennae (see [2] and the references herein). This work deals with the problem of finding the optimal orientation of Qy transition moments of light-harvesting bacteriochlorophyll (BChl) a molecules of a subantenna B798 (absorption maximum, at 798 nm) in the green bacterium Chloroflexus aurantiacus. We considered infinite three-dimensional antennae having translational symmetry along the X and Y axes, which provides its preparation from elementary fragments representing linear one-dimensional antenna units parallel to the Z axis and containing pigment molecules of three uniform subantennae, namely, B740 and B798 and B808 (Fig.1).



[1] Lomonosov Moscow State University, Department of Bioengineering and Bioinformatcis, Moscow, 119992, Russia
[2] A.N. Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, 119992, Russia, zfetisova@genebee.msu.ru

B798 is the acceptor for oligomeric BChl c B740 subantenna and the donor for monomeric BChl a B808 one. The spatial orientations of Qy transition dipoles of pigments are known only for two subantennae (B740 and B866) [3-5]. Using the probability matrix approach, we computed the time (t, a.u.) of excitation energy transfer from B740 to B808 as a function of φ (φ is the angle between the orientation of B798 Qy transition dipoles and the antenna plane). We varied the angles φ, α, β (see Fig.1). Fig.2 shows a typical set of parametric curves t(α,Δ,φ) for R12=R23. For any fixed values of α and Δ=α-β parameters, each t(φ) graph has two branches.



Each branch has two asymptotes, one variable (φa∈[+28,2°÷-28,2°]), and one constant (φa = ±90°) (see Fig.2). For R12=R23, the t(φ) dependences have deep stable minima near φ=0°, where t values are notably lower than those for randomly oriented systems. When R23 increases approaching to its possible upper limit (R23 ≈ 2R12), the deep stable minima of t(φ) dependences move continuously from φ=0° to φ<±25°. Thus, the model calculations have shown that polarization of the baseplate BChl a Qy transition dipoles B798 of the green bacterium Chloroflexus aurantiacus in the antenna plane (or "cy près") is biologically expedient being optimal for excitation energy transfer B740→B798→B808-866. Corresponding experiments are in progress.

The work was supported by the Russian Foundation for Basic Research (Grant 05-04-49494).

1. Z.G. Fetisova, M.V. Fok (1984) Molec. Biol. (Engl. transl.) 18: 1354-1359

2. A.A.Novikov, A.S. Taisova, Z.G. Fetisova (2006) J.Bioinform.Computat.Biol., 4: 887-909
3. V.I. Novoderezhkin, A.S. Taisova, Z.G. Fetisova, R.E. Blankenship, S. Savikhin, D. Buck, W.S. Struve (1998) Biophys. J., 74: 2069-2075
4. Z.G. Fetisova, A.M. Freiberg, K.E. Timpmann (1988) Nature, 334: 633-634
5. V.I. Novoderezhkin, Z.G. Fetisova (1999) Biophys. J., 77: 424-430

AN USING OF DL-SYSTEMS TO MODEL OF THE RENEWABLE ZONE SIZE CONTROL IN GROWING TISSUE

U.S. ZUBAIROVA, S.V. NIKOLAEV, N.A. KOLCHANOV[1]

The meristem structure remains constant during plant growth, but its resident cells changes. As a result of horizontal division of overlying cells of central zone shift down and transform into cells of organizing center. In turn the cells of organizing center also shift down and transform into the cells of rib-zone. There are some facts in recent articles about distribution of the length of cell cycle in shoot apical meristem, but in our work we considered only stochasticity of cells divisions.

In short, the model is the following. Cells of shoot apical meristem are not differentiated, but they are determined by expression of certain genes. 2-4 cells around vertical axis of meristem in 3-4 upper layers express CLV3. It is the central zone. The cells beneath the central zone express WUS. It is the organizing center (2-3 cell). Mechanism that provide such constant structure is the subject of intensive research.

The main concept of our model is the following. We have one dimensional array of n cells. Substances Y, Z, W can be synthesized in the cells with rates depending on the concentrations of these substances. It is assumed that Y and W are diffusible. The Y is synthesized in first cell and diffuses through the cells of the array. The rate of the Y synthesis depends on concentration of W in the cell. The C does not diffuse and only decay. The rate of the C synthesis depends on concentration of Y. Substance W diffuse through cell-array and regulate synthesis of Y in the cell 1. Its rate of synthesis depends on concentrations of Y and C. Thus we can formulate this model in terms of Cauchy problem.

In this work we studied influence of the cell division on dynamics of compartmental structure of the renewable zone. We also interested in relations between character time of the cell cycle and diffusion of morphogens and influence of this factors on the systems stability.

Institute of Cytology and Genetics SB RAS, pr. Lavrentieva 10, Novosibirsk, 630090, Russia, ulyanochka@gorodok.net

Another part of model is cell division. Production of substance W defines the renewable zone. Each cell is characterized with its length l. At initial time all cells are labeled with initial values of length according to normal distribution. Next time cells grow and value of l increase. When length l of a cell achieves critical value, the cell length divides in certain relation k, which is also stochastic variable. Concentration of substances in its child cells is the same as in the parent cell. Lengths of the cells outside of the renewable zone are unchanged.

Model is realized in stochastic parameterized dL-system [1,2]. It is a mathematical framework for modeling plants and simulating their development in a manner suitable for animation. The key concept is the integration of discrete and continuous aspects of model behavior into a single formalism, called dL-systems, where L-system-style productions express qualitative changes to the model (cell division, and cell type), and differential equations capture continuous processes (change of concentration Y, C and W). Differential L-systems extend parametric L-systems by introducing continuous time flow in place of a sequence of discrete derivation steps. Concentration of substances defines productions of L-system at each derivation step.

We used program package Mathematica for modelling, because it is a very convenient instrument for work with a rewriting systems.

Computer simulations with the model demonstrated movement of the zone boundaries as a result of occasionally simultaneous division of some cells. These movements was followed by recovering “normal” structure. This model behavior is in consistence with some experimental data. However, in some computer experiments we observed destruction of the system, when many cells divide occasionally in the same time.

1. Przemyslaw Prusinkiewicz, Mark Hammel, and Eric Mjolsness. Animation of Plant Development. Proceedings of SIGGRAPH 93, pp. 351-360.
2. P. Prusinkiewicz and A. Lindenmayer. The algorithmic beauty of plants. Springer-Verlag, NewYork, 1990. With J. S. Hanan, F. D. Fracchia, D. R. Fowler, M. J.M. de Boer, and L. Mercer