Proceedings of the International Moscow Conference on Computational Molecular Biology MCCMB 2015

Moscow, 16-19 July 2015

ISBN 978-5-901158-27-2

Conference materials are available at http://mccmb.belozersky.msu.ru/2015

Седьмая Московская международная конференция по вычислительной молекулярной биологии MCCMB'15
Москва, 16-19 июля 2015
Сборник трудов. ISBN 978-5-901158-27-2

Organizers

Department of Bioengineering and Bioinformatics of M.V.Lomonosov Moscow State University
Biological Department of M.V. Lomonosov Moscow State University
Institute for Information Transmission Problems, RAS
Engelhardt Institute of Molecular Biology, RAS
Vavilov Institute of General Genetics, RAS
Non-commersial Partnership "Bioinformatic seminar"
The Scientific Council on Biophysics of RAS
Kazan Federal University

Acknowledgements

We are grateful for generous support of the following organizations and companies:
Russian Foundation for Basic Research
Genotek

A method of detecting local gene synteny rearrangement

V. A. Lyubetsky, L. I. Rubanov, O. A. Zverkov, L. Yu. Rusin, A. V. Seliverstov

Institute for Information Transmission Problems, Russian Academy of Sciences
Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia, lyubetsk@iitp.ru

A. G. Zaraisky

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences
Moscow 117997 Russia, azaraisky@yahoo.com

Problem statement and data. Local rearrangement of gene synteny is a likely driving force in many aspects of species evolution. We describe a method and its computer implementation (http://lab6.iitp.ru/ru/lossgainrsl) for detecting genes with local synteny disruption. The underlying algorithm operates with narrow windows around protein-coding genes. The computer program computes 30 species for about 25 mins on an average PC, which demonstrates high performance.

Define two sets of species corresponding to *upper* and *deeper* levels of a phylogeny respective of a certain point; each set is structured according to taxonomy. The algorithm finds genes in a pre-defined species from one of the two sets that we call the *basic* species. The algorithm can be run in two modes: finding genes with undisrupted synteny in at least *one* of the taxa or, otherwise, in *each* of them. Interestingly, our algorithm outputs short gene lists even in the first mode. Henceforth, we refer to the first mode and imply "gene loss" if "gene gain" is not explicitly stated.

We exemplify the following four cases with listing upper and then deeper species on the tree; the basic species is *Xenopus tropicalis* (clawed frog). (**0**) The gene is missing in reptiles, birds, mammals but present in the frog and at least one gnathostomatous fish. (**1**) The gene is missing in birds and mammals but present in the frog and at least one representative of gnathostomatous fish and reptiles. (**2**) The gene is missing in mammals but present in the frog and at least one representative of gnathostomatous fish, reptiles, and birds. (**3**) The gene gain: its presence in the frog and at least one representative of reptiles, birds, and mammals but absence from all fishes and tunicates. These taxa are well represented in the Ensembl database. Fish: *Petromyzon marinus* (agnathan fish); *Astyanax mexicanus*, *Danio rerio*,

Gasterosteus aculeatus, *Latimeria chalumnae*, *Lepisosteus oculatus*, *Oreochromis niloticus*, *Oryzias latipes*, *Poecilia formosa*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Xiphophorus maculatus* (gnathostomatous fish); amphibians: *Xenopus tropicalis* (clawed frog); reptiles: *Anolis carolinensis* (lizard), *Pelodiscus sinensis* (turtle); birds: *Gallus gallus* (chick), *Meleagris gallopavo* (turkey), *Taeniopygia guttata* (zebra finch), *Anas platyrhynchos* (duck), *Ficedula albicollis* (flycatcher); mammals: *Ornithorhynchus anatinus* (platypus), *Sarcophilus harrisii* (Tasmanian devil), *Monodelphis domestica* (opossum), *Homo sapiens* (human), *Mus musculus* (mouse), *Cavia porcellus* (Guinea pig); tunicates: *Ciona intestinalis*, *Ciona savignyi.*

The method. All protein-coding genes from the basic species are tried. For gene *X* we verify a predicate in the 2- and 3-species modes. In the 2-*species mode*, in each deeper species homologue *X** is checked for its synteny. For gene *X* in the basic species two witness ("anchor") genes *Y* and *Z* are defined as different from *X* and each other, and co-located within a window of length *l*. Their homologues *X**, *Y**, and *Z** co-located within window of length l_1 are identified in the deeper species. The predicate is satisfied downwards if in a deeper species there exists such co-location of *X**, *Y**, and *Z**. The predicate is satisfied upwards if in each upper species there is no homologue *X** or its synteny is disrupted (no homologue of a witness gene within window *l* occurs within window l_2 in the upper species). The predicate is satisfied if it is satisfied both upwards and downwards. In the exemplified results the window length was set equal: $l=l_1=l_2$=2Mb. The method allows to set any window size and any number of witnesses. E.g., for a high quality genome assembly two witnesses can be called downwards, and one – for a poorer assembly. The method and its implementation allow for a more fine detection of synteny by checking for gene transfer to the antisense strand, etc.

The problem of effectively detecting true homology and orthology, and the choice of associated parameters are known to be far from completion. We employed several approaches and assumed the answer that does not depend on the choice of one of them. In the first approach, homologues of *X*, *Y*, and *Z* are defined with the BLAST algorithm using pre-set expect value thresholds *E* (for upper species) and E_1 (for deeper species). The second

approach is based on our original method of protein clustering, where each non-singleton cluster is considered a protein family (the same threshold E_1 is used). In the third approach orthology is assumed according to the Ensembl database. The three approaches produce three gene lists satisfying the 2-species predicate, and their intersection is the 2-*species list* in the basic species. We set $E=10^{-9}$ to detect genes in upper species that diverged in sequence but retained synteny and assume $E_1 = E$.

Our results suggest that each individual approach produces significant overprediction, and the intersection of the three lists lowers overprediction for a better confidence. The program output includes all intermediate gene lists for independent analyses. The program allows for high flexibility in its logic owing to the built-in interpreter for accepting user-defined search scenarios. We widely tested the clusterization approach on plastid genome data.

We discuss causes of under- and overpredictions of the method. Thus, for case (2) and the frog gene *foxa2*, all three approaches produce downward predictions: two witnesses in fish, reptiles, birds (BLAST), and two witnesses in birds (clusterization and Ensembl). Analysis of Ensembl contigs of the frog does not detect any *foxa2* witness upwards in mammals, which therefore places it in the 2-species list. However, in Danio *foxa2* is located at coordinate 42.4 (chromosome 17) with witnesses *pax1a* and *nkx2.2a* (42.85 and 42.92, respectively); in human – at coordinate 22.6 (chromosome 20) with the same witnesses (21.7 and 21.65, respectively). Nevertheless, their orthologues in the frog, *pax1* and *nkx2-2*, are located in a different contig with respect to *foxa2*, which makes the local synteny of *foxa2* in the frog and Danio doubtful to confirm with Ensembl data. To rule out such cases, the algorithm implements the 3-species predicate, which juxtaposes not only the basic and upper species but each deeper with each upper species as well. More specifically, the 3-*species predicate* operates downwards as the 2-species predicate, whereas the upward operation is more constrained: gene *X* in the basic species satisfies the 2-species predicate upwards, *and* none of the deeper species contains a homologue *U* (a "*substitute*" of *X*) with two witnesses, which has a homologue *U** with two homologous witnesses within the corresponding windows in an upper species. Call the gene list in the basic species obtained with the 3-species predicate the 3-*species list*. It does not include *foxa2* because there is another substitute of *foxa2*

downwards. The intersection of the 2- and 3-species lists is defined as the *gene list*.

Examples of gene lists. For results (**0**): ENSXETG00000006007 (only last ID digits are given hereafter, with gene names in brackets), 6008, 9881, 16048 (*foxo1*), 17278 (*ccdc150*), 20840 (*ndc80*), 26169, 31554, 31627, 33120, 33176, 33543. For results (**1**) the following genes are added: 13385, 27743, 33873. For results (**2**) the added gene are: 2741 (*c3orf17*), 3913, 4328 (*commd5*), 4350, 4496, 10491 (*pcbp2*), 10494 (*znf830*), 10495 (*bag1*), 12462 (*gtf2e1.2*), 13081, 13652, 17116, 20301 (*nog2*), 20838 (*taf13*), 21667 (*rpl18a*), 22934, 23300 (*tprkb*), 24517, 24648 (*mocs3*), 25525 (*pnhd*), 27268 (*nog4*), 30850, 32142, 32294, 33929 (*MFSD8*). For results (**3**) the list is: 1468 (*qrfp*), 1739 (*OR6B1*), 12323 (*hmgn2*), 16850 (*prnp*), 19153 (*arc*), 20235 (*stra8*), 20472 (*pacrgl*), 24783 (*lif*), 27360, 27419 (*a4galt*), 30940 (*IL22*), 31063 (*OR11L1*), 31314 (*SFTPC*), 32124 (*enam*), 33385, 33448 (*MGC145290*), 33476, 34067.

Discussion of examples. The gene *nog2* does not have any witness upwards in mammals, and thus is included in the 2-species list. With the Ensembl data the "proper" substitute is difficult to choose. A more sophisticated situation is observed with *nog4*. Distinct from the Ensembl data, the algorithm does not detect its homologues in mammals, and therefore places it in the gene list. Genes *foxo1*, *sfrpx*, *pnhd* are placed in the gene lists, and are experimentally predicted in our group to condition regeneration capacities that were lost in amniotes. The gene *sfrpx* has a formal substitute in *Gasterosteus aculeatus* but they are not orthologous, it is included in the gene list. The prothymosin-coding gene ENSXETG00000006008 has a mammal homologous domain (PF 03247) involved in determining cell division, but the protein sequence and local synteny underwent considerable changes (as inferred by the method). This may suggest its important role in mammal evolution. The presented results are used to discuss the impact of local synteny rearrangement on shaping evolutionary patterns.

Research was partly supported by the Russian Foundation for Basic Research (grant 13-04-40196-H).