SEARCH FOR REGULATORY SIGNALS IN GROUPS OF ORTHOLOGOUS GENES OF GAMMA - PROTEOBACTERIA

***Danilova L.V.* [1], *Gelfand M.S.* [2]**

[1] Institute of information transmission problems, Moscow, Russia, e-mail: dlv2k@mail.ru
[2] Integrated Genomics-Moscow, Russia, e-mail: gelgand@integratedgenomics.ru

Key words: *gamma - proteobacteria, Orthologous genes, regulatory signals*

Introduction

Recognition of common regulatory signals in sets of DNA sequence fragments is an old and still actual problem of computational molecular biology. There are different approaches to this problem. One of them is analysis of upstream regions of orthologous genes from related genomes [1, 2]. The underlying assumption is that there is a conserved signal for orthologous regulators. This is not always correct, but in a sufficient number of cases this assumption holds, making the comparative technique a promising approach. Here we apply a previously suggested algorithm to analysis of genomes from the *Escherichia coli* group in order to test its applicability to other, less studied taxonomic groups.

Materials and Methods

Complete genomes of gamma-proteobacteria Escherichia coli, Escherichia coli O157, Salmonella typhi, Salmonella typhimurium, Yersinia pestis, Vibrio cholerae, Haemophilus influenzae, Pasteurella multocida were considered.

A pair of genes from two genomes was considered to be orthologous if these two genes were the closest relatives of each other in these two genomes. Then, pairs of orthologs were merged into clusters using the single linkage algorithm. Transitivity was not required and small differences in the similarity level were ignored (thus one gene could have more than one ortholog in any given genome).

Upstream regions of length 200 bp were selected. No overlaps with other genes were allowed, so if the distance to the upstream gene was shorter than 200 bp, only the spacer was selected.

Closely similar fragments were filtered out, retaining *E. coli* fragments whenever possible. This allowed us to search for conserved regulatory signals without interference from insufficiently divergent sequences from closely related genomes (strains). The criterion of excessive similarity was matches in at least 35 out of any 40 consecutive positions.

After the filtration step, there were 1967 subsamples of at least three fragments. After that, 345 sequences of known regulatory sites of *E. coli* were taken from the dpinteract database [3] and matched to the samples. Both directions of DNA sequence fragments were considered. Total 311 sites were found in 239 sequences. Other sites were not found either because they were located outside of the selected fragments or because the gene had no orthologs. A known sites could be placed in several samples if it was located between divergently transcribed genes, since each known site was considered in both direct and complementary directions.

The program implementing the earlier proposed algorithm [4] was used. It accepts as input some sequences (here 3 through 40) of length from 40 up through 200 bp, and outputs a system of similar words in each sequence. The system quality is defined by optimization of the pairwise similarity of words. It also takes into account additional features of words, e.g. their palindromicity. Signals of length 15, 20, 22, and palindromic signals of length 15, 16 and 22 were considered.

Implementation and Results

The results are shown in the Table. Ninety nine out of 311 known sites were found (that is, coincided with predicted signals or were subwords of the signals). Other sites were not found either becase the signals were too weak to be identified or because orthologs lost the regulation.

Discussion

The known sites *E. coli* considered for Table only. Our samples consist of region of orthologous genes from related *E. coli* genomes. If we have found site some regulator in *E. coli* fragments that means that we have found sites orthologuos genes the same regulator.

This work showed that such approach is reasonable for study taxonomic groups. We plan to apply that approach and our algorithm for *Bacillus subtilts* group and alpha-proteobacteria

Table. Nubmer of known sites *E.coli* in samples and detected sites.

Regulator	Number of known sites	Number of known sites in samples	Direct site	Complementary site	Detected sites
arcA	14	9	2	7	9
argR	17	20	14	6	3

cpxR	12	6	4	2	2
crp	49	41	26	15	3
cspA	4	6	3	3	1
cynR	2	4	2	2	1
cytR	5	4	4	0	0
deoR	3	1	1	0	0
dnaA	8	4	3	1	1
fadR	7	9	7	2	1
farR	4	8	2	6	4
fnr	14	10	6	4	1
fruR	12	6	3	3	4
fur	9	9	6	3	4
galR	7	4	3	1	2
gcvA	4	1	1	0	1
glpR	13	14	9	5	4
hns	15	11	6	5	4
hu	3	1	1	0	0
iclR	2	1	0	1	1
lacI	3	0	0	0	0
lexA	19	17	13	4	9
malT	10	17	9	8	5
melR	2	4	2	2	0
metJ	15	20	13	7	11
metR	8	10	7	3	1
narL	11	7	4	3	1
narP	8	10	6	4	0
ntrC	5	4	3	1	1
ompR	9	7	5	2	2
pdhR	2	2	2	0	1
purR	22	15	12	3	8
rpoN	6	4	3	1	3
torR	4	12	8	4	6
tyrR	17	13	13	0	5
Total	345	311	203	108	99

Acknowledgements

This study was partially supported by grants from INTAS (99-1476), HHMI (55000309), LICR (CRDF RBO-1268) and RFBR (00-15-99363). We are grateful to P.Novichkov for the help with the data, and V.A.Lyubetsky, the scientific leader of L.V.Danilova, K.Yu.Gorbunov and A.A.Mironov for discussions.

References

1. McCue L.A., Thompson W., Carmack C.S., Ryan M.P., Liu J.S., Derbyshire V., Lawrence C.E. (2001) Phylogenetic footprinting of transcription factor binding site in proteobacterial genomes. Nucl. Acids Res. 29, 3, 774-782.
2. Terai G., Takagi T., Nakai K. (2001) Prediction of co-regulated genes in Bacillus subtilts on the basis of upstream elements conserved across three closely related species. Genome Biol. 2, 11.
3. Robison K., McGuire A.M., Church G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. J. Mol. Biol. 284, 241-254.
4. Danilova L.V., Gorbunov K.Yu., Gelfand M.S., Lyubetsky V.A. (2001) Algorithm of regulatory signal recognition in DNA sequences. Mol. Biol. 35, 6, 987-995.