Molecular Biology, Vol. 34, No. 2, 2000, pp. 222–231. Translated from Molekulyarnaya Biologiya, Vol. 34, No. 2, 2000, pp. 253–262.

GENOMICS. PROTEOMICS. BIOINFORMATICS

UDC 577.1

Software for Analysis of Bacterial Genomes

A. A. Mironov, N. P. Vinokurova, and M. S. Gelfand*

State Research Center GosNIIGenetika, Moscow, 113545 Russia;
** E-mail: misha@imb.imb.ac.ru*

Received November 20, 1999

Abstract—GenomeExplorer is a program for comparative analysis of regulation in prokaryotic genomes. The program has options for signal search, comparison of gene samples, search for paralogs and orthologs, iterative construction of signal profiles. The program has a convenient graphic interface, allowing for navigation in the annotation window, in the genome map, and in the table of gene similarities. The use of the system clipboard allows one to export the results of analysis into Word and Excel, and to call external programs via the Internet.

Key words: transcription, translation, regulation, operator, computer analysis

INTRODUCTION

Prediction of regulatory signals is an important component of genome annotation. Analysis of regulation not only is interesting *per se*, but allows one to make more precise functional annotations of genes made by protein analysis methods. There exist numerous methods of analysis of transcription factor binding sites and other signals (for review see [1–3]), but the reliability of the existing algorithms is not satisfactory [4–5].

Availability of numerous complete and almost complete genomes makes it possible to sharply increase the specificity of predictions. The comparative approach is based on the assumption that composition of regulons (sets of coregulated genes) is conserved in related genomes. Thus the true sites are located upstream of orthologous genes, whether false positives are scattered at random. Making predictions with a low threshold for a group of related genomes, and making total pairwise comparison of genes with upstream candidate sites, a user can check consistency of predictions and thus to increase their reliability. Moreover, the comparative approach makes it possible not only to describe counterparts of known regulons in little studied genomes, but also to discover new members of well studied regulons [6].

This approach allowed us to describe regulons of purine, arginine, and aromatic amino acids metabolism in *Escherichia coli* and *Haemophilus influenzae* [7, 8], SOS-regulons of *E. coli*, *H. influenzae*, and *Bacillus subtilis* [9], regulation of ribosomal protein operons in *E. coli* and *H. influenzae* [10], as well as the riboflavin biosynthesis regulon in a large group of bacteria [11]. Currently we are studying regulons of catabolite repression in enterobacteria; SOS-response in gamma-purple bacteria, the *Bacillus/Clostridium* group, and mycobacteria; attenuation of aromatic amino acid operons in gamma-purple bacteria and *Chlamydia trachomatis*, as well as several archaeal regulons.

The existing methods of computer analysis are not convenient for these studies. The differences in formats of input and output, lack of convenient interfaces and other similar problems make the research very time consuming. Thus we have developed *GenomeExplorer* software specifically intended for comparative analysis of genomes and, in particular, regulatory signals. Here we described the main features of the program. We start with the main features (main windows and menus, analysis tools), and then describe the technical points (installation, customization, etc.). The following notation is used throughout: **boldface, menus**; *italics, programs and databases*; ***boldface italics, file and folder names***; underlined typewriter font, keyboard; **boldface typewriter font, input/output.**

WINDOWS AND MENUS

GenomeExplorer runs under *Microsoft Windows* and retains the main features and defaults of this environment. A system of menus is used. Simultaneous opening of several copies of the program allows one to perform comparative analysis of different genomes.

Windows

The main program has two main windows: the annotation window that contains annotations of genes and other functional units downloaded from the genome features table (in the *GenBank* or *EMBL* for-

0026-8933/00/3402-0222$25.00 © 2000 MAIK "Nauka/Interperiodica"

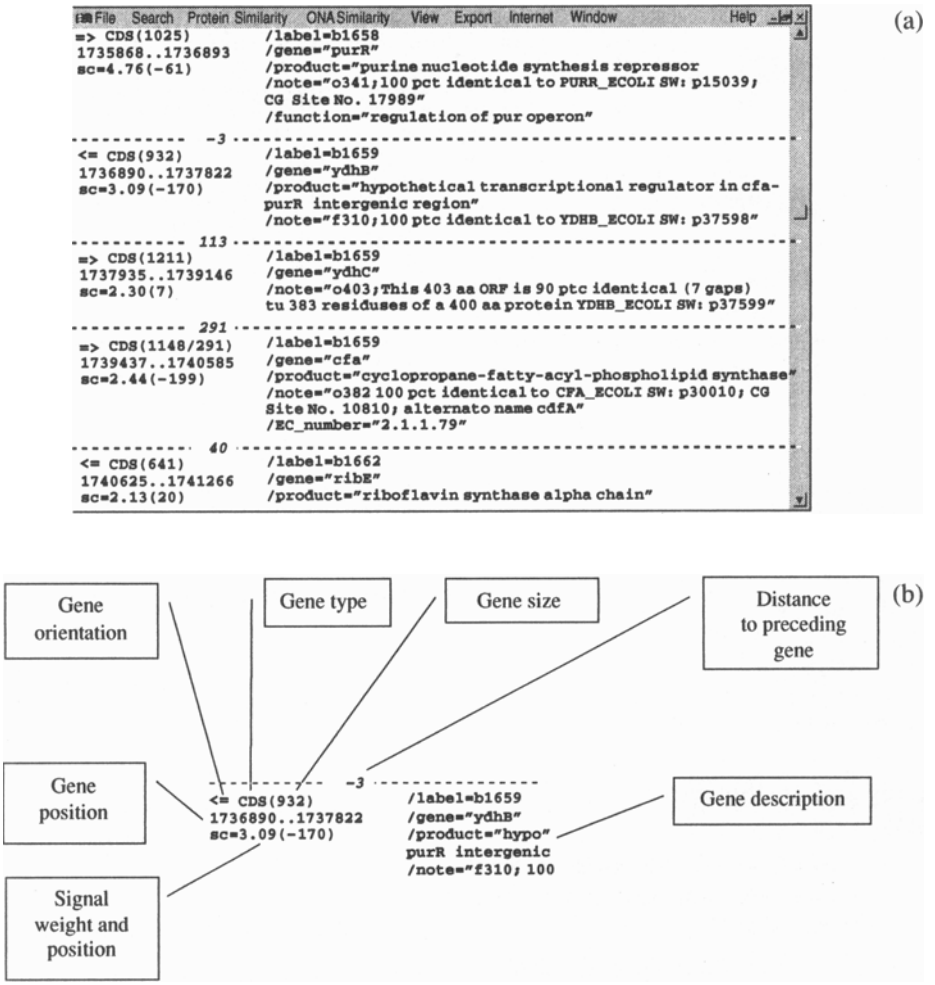

Fig. 1. (a) The annotation window. (b) Cell corresponding to one gene.

mat) and the genetic map window where the same units are presented in the graphical form.

The annotation window consists of individual cells corresponding to genes and other functional units (Fig. 1a). Each cell contains information about the unit type, strand, length, distance to the previous unit, and description taken from the features table (Fig. 1b). In this window the user can:

- navigate using arrows (↑, ↓, PgUp, PgDown, Home, End) and the vertical scroll bar;
- mark and unmark functional units (space; delete all marks—Shift + space);
- navigate over marked units (⟸, ⟹);
- perform text search, DNA and protein similarity search, analysis of signals, output, etc. (see below);
- work with the local menu (hit on the right mouse button, see below).

The map window is evoked from the main menu **View** ⟶ **Map**. It consists of three areas: a circular map of the complete genome in the upper left corner that shows the current chromosomal position, a linear map at the bottom showing the local neighborhood of the current gene, and the brief description of the genes from the current region in the upper right corner (Fig. 2). In this window the user can:

- navigate in the genome using arrows, clicking the left mouse button on the circular map, or using the horizontal scroll bar;
- perform text search, DNA and protein similarity search, signal analysis; output the analysis signals (see below);
- work with the local menu (hit on the right mouse button, see below).

Click on the left mouse button in the active window performs positioning in both windows. If a text file is open, click of the left mouse button on a word is interpreted as a gene name, on a number, as a genome position, with automatic positioning in the annotation window.

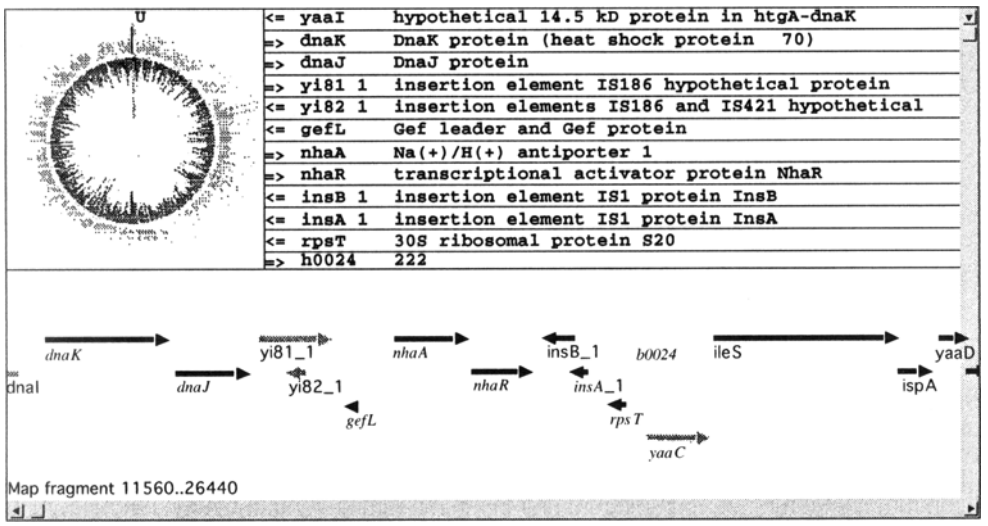

Fig. 2. The map window.

Menus

The main menu (top panel) has the following points:

File—loading and customization (see "Service programs").

Search—gene search given the name, a fragment, etc., marking selected genes (see "Service programs").

Protein similarity—analysis of protein similarities (see "Gene comparison")

DNA similarity—analysis of regulatory signals and nucleic acid similarities (see "Analysis of regulatory signals").

View—setting windows (see below).

Export—output (see *Service programs*).

The local menu of the annotation window (click on the right mouse button) has the following points:

Mark/Unmark—mark the current gene in the annotation window (synonym: space bar).

Copy—copy to the clipboard (see *Service programs*).

Export—output to file (see *Service programs*).

Mark operon—mark genes potentially cotranscribed with the already marked genes (located on the same strand); mutual positioning of genes (upstream/downstream) and maximum allowed intergenic distances are set in a special dialog window.

Unmark all—deselect all marked genes (synonym: Shift + space).

Get Distance—show the intergenic distance.

Annotate—add annotation to file ***ann** that will be kept in the same folder as the current genome and will be downloaded during all subsequent analyses together with the basic genome annotation.

The local menu of the map window has the following points:

Zoom in—increase resolution of the local map.

Zoom out—decrease resolution of the local map.

Zoom ...—set resolution of the local map.

Copy—copy to the clipboard (see below).

Auxiliary Windows

The main menu point **View** opens the following auxiliary windows:

Features 1—main annotation window (if it has been closed).

Features 2—additional annotation window.

Map—map window.

Details—results of the last site or similarity search for the current gene.

Histogram—histogram of site or similarity scores for the last search.

Info—genome data: length, GC-content, number of genes, number of marked genes.

Select Features—dialog window: select functional units to be shown in the annotation window.

GENE COMPARISON

Gene comparison is done using the standard Smith–Waterman algorithm [12] that constructs the optimal local alignment of two sequences. The alignment is characterized by score *sc* (using affine deletion penalties), length *len*, and normalized score $lsc = sc/\sqrt{len}$. The set of genes to be compared is set from the main menu, the alignment parameters is set in the dialog box. There exist two groups of the alignment parameters.

The first group is the alignment parameters *per se*:

IniGap—penalty for gap initialization,

Del—penalty for gap extension,

Cutoff—threshold for alignment significance (in *lsc* units; the default value is 9; this parameter should be decreased for analysis of short fragments; the threshold for highly significant alignments is 12),

Correct ends—use the termini correction procedure to maximize *lsc*,

amino acid mathing matrix (the default matrix is BLOSUM62 [13]).

The second group consists of filtration parameters using Roytberg's method [14] for fast similarity search:

Use Fast Search—use filtration (always used for analysis of nucleic acid similarities),

Diag. Dist.—maximum distance between pairs of matching *l*-grams,

Diag. Shift—maximum allowed number of deletions for construction of *l*-gram chains,

CutOff1—minimum total length of *l*-gram chains (stricter filtration corresponds to larger values),

Band—alignment width (maximum total number of deletions),

L-gram—*l*-gram size (strict filtration for $l = 5$, soft filtration for $l = 3$),

16-alphabet—use 16-symbol alphabet assuming equivalence of similar amino acids (the strictest filtration),

8-alphabet—use 8-symbol alphabet assuming equivalence of similar amino acids,

4-alphabet—use 4-symbol alphabet assuming equivalence of similar amino acids (the softest filtration; obligatory value for DNA comparisons).

If filtration is switched on (fast search), the alignment is constructed only if two genes have at least one chain of coinciding *l*-grams in the given alphabet, such that the total number of symbols in the chain exceeds **CutOff1**, the distance between adjacent *l*-grams in each sequence is less than **Diag. Dist.**, and the total number of deletions between *l*-grams is less than **Diag. Shift**.

Protein Alignment

The program performs comparison of various groups of genes or, more exactly, encoded proteins, dependent on the problem specifics. All protein alignments are evoked from the main menu. The results are output to the special window (see below "Comparison window").

Protein similarity—analysis of protein similarities:

Protein—search for homologs in the current genome for one or more proteins from a file (respectively, file ***.ppt*** in the *SwissProt* or *GeneProt* format, or file ***.ppc*** in the *FASTA* format; if file ***.ppc*** contains several sequences, they are separated by the star symbol *).

Compare selection—comparison of marked genes from two or more genomes. The dialog window for the choice of genomes appears.

Paralogs—search for homologs (paralogs) in the current genome for the current gene.

Ortologs—search for homologs of the marked genes from the current genome in another genome (genomes) and consequent search for homologs of the found genes in the current genome. The choice of genomes is done in the dialog window. The two-step procedure allows one to distinguish between orthologs (defined as symmetrical best hits) and paralogs.

Motif—search for protein motifs. The motifs are set in the *ProSite* format: each element of a motif is a symbol of an amino acid, group of amino acids in brackets, or X denoting an arbitrary amino acid. Capitals and lower case letters are synonymous, other symbols are ignored. Thus **G–L–[v, l, i]–x–c–f–[S, T]** is equivalent to **GL[VLI]XCF[ST]**. Besides, it is possible to set the maximum allowed number of mismatches with the given motif.

Parameters—setting the parameters of the protein alignment. These parameters will be accepted as default for consequent searches.

Clipboard—search for homologs in the current genome for one or more proteins from the clipboard (in the *FASTA* format).

XYZ	signal name
a c g t	column headers: order of nucleotide weights
0.23 -0.24 -0.24 0.25	positional nucleotide weights for the first
0.05 -0.17 0.08 0.05	box; the number of rows equals
-0.05 -0.05 -0.23 0.34	the profile lengths
...	
MinScore=2.5, MaxScore=4.8	minimum allowed and maximum observed scores of the first box
- - - - - - - - - - - - - - - - - - - -	separator
Dist 2..4	interval of allowed interbox distances
0 0 0	distance weights; the number of weights equals the interval length (no particular preferences in the shown example)
- - - - - - - - - - - - - - - - - - - -	separator
0.20 -0.14 -0.14 0.08	positional nucleotide weights for the second
0.11 -0.33 0.11 0.11	box; the number of rows equals
-0.05 -0.05 -0.23 0.34	the profile lengths
...	
MinScore=3.0, MaxScore=5.5	minimum allowed and maximum observed scores of the second box,
**********	end of description
useless stuff	comments

Fig. 3. Sample ***.pat*** file with a recognition for a two-box site.

Nucleotide Alignment

Variants of the nucleotide alignments are listed below. They are evoked from the main menu. The results are output to the specific window (see below "Similarity window").

DNA similarity—analysis of nucleotide similarities:

DNA—search for similar fragments in the current genome for a nucleotide sequence from a file (file ***.dne*** in the *FASTA*, *GenBank*, or *EMBL* format).

Site Search—see "Analysis of regulatory signals".

Psi-Site—see "Analysis of regulatory signals.

Protein/DNA—similarity search for an amino acid sequence (given in file ***.ppt***) in the current genome formally translated in six reading frames.

Parameters—setting parameters:

DNA-DNA—setting parameters for nucleotide alignments; these parameters will be then accepted as defaults.

Protein-DNA—setting parameters for protein-DNA alignments; these parameters will be then accepted as defaults.

Clipboard—search for homologs in the current genome for one or more sequences from the clipboard (in the *FASTA* format).

Comparison Window

The results of the gene similarity analysis are output to the comparison window. This window is super-

vised by program *cmpwnd* and a retained result (file ***.cmp***) can be evoked from this program or from the main menu (**File** → **Open CMP file**).

The similarity window consists of four areas:

• The list or table of alignments (top left). In the list format each pair of aligned genes is represented by a line containing the gene names and a bar representing the alignment in pseudocolors. In the table format each pair is represented by a dot in the rectangle table. The color of the dot corresponds to the alignment score. Click on the left mouse button at a line (in the list format) or at a dot (in the table format) displays the detailed information about the alignment (see below) and positions on the aligned genes in both genomes.

• The scale of pseudocolors corresponding to the similarity levels (top right).

• The detailed data about the current alignment (middle): similarity level (score *sc*, normalized score *lsc*, percent identity, percent of similar amino acids), alignment length, gene names, position of the aligned segment in the genes, graphical representation of the alignment in the pseudocolors.

• current alignment (bottom).

The similarity window menu has the following options:

File

New window—open the second window with the same set of alignments (convenient for comparative analysis).

Open—open a file with alignments (***.cmp***).

Save—output alignments to a file (***.cmp***).

Save As—save the file with alignments under a new name (***.cmp***).

Exit—close the similarity window.

Edit—edit the set of alignments.

Sort—sorting:

by Score—by the normalized alignment score (*lsc*),

by First Gene—by the name of the first gene,

by Second Gene—by the name of the second gene,

Unsorted—return to the initial order.

Copy—copy to the clipboard:

Current Alignment—the current alignment,

Table—the table of alignments. The list of parameters to be output is set in the dialog window. The output format allows one to easily paste the results in *Excel* and *Word*.

Delete—delete:

Current—the current alignment,

By Score—all alignments with scores below the threshold.

Symmetrize Matrix—symmetrize the list by inversion the order of genes in each pair, so as each gene appears in the table both in a line header and a column header.

Find—search:

Gene—by a gene name,

Motif—by a protein motif,

Next Search—repeat the search.

View

Gene Table—alignments in the table format,

Gene List—alignments in the list format,

Parameters—parameters used to construct the alignments (if inaccessible, all zeroes are shown).

ANALYSIS OF REGULATORY SIGNALS

The option **DNA similarity** → **Site Search** allows the user to search for functional sites in DNA using recognition rules in the form of sets of profiles (positional nucleotide weight matrices). The format of ***.pat*** files containing the rules is presented in Fig. 3. Such a file contains one-three profiles, minimal thresholds for individual profile scores, and description of allowed distances between the signal parts.

During the search a dialog window opens, that provides for setting the boundaries of potential regulatory regions (relative to gene starts) and the minimum threshold on the total site score.

After completion of the analysis, each gene in the annotation window gets additional characteristics, namely, the score and the position of the best site in the regulatory region of this gene. Genes having candidate sites with scores exceeding the threshold become marked, whereas all old marks are removed. The results can be output to a file ***.gls*** (the main menu **Export** → **Extended Gene List**) or the clipboard (the local menu, **Copy** → **Site Report**), see below ("Output").

Files ***.pat*** for several signals are distributed with the program. New signals can be created in two ways.

If a training sample of experimentally characterized aligned sites is available, the corresponding file ***.pat*** can be created using the auxiliary program *signal* that is contained in the distributed package. A file containing the training sample of aligned sites in the ASCII format is created using an arbitrary text editor. Each line in this file is of the form

<site_name> <blanks> <sequence>,

where the site name can contain only alphanumeric characters and underscores _ (underscore), the sequence can contain only upper and lower case let-

ters (**ACGTacgt**) and, in particular, cannot contain blanks. All fragments should have the same length (that would be the length of the generated profile) and be aligned (biologically equivalent residues should occupy same positions).

When the file with the learning sample is created, program *signal.exe* should be evoked. To input the learning sample and construct the profile, the user should chose the menu option **File → Open Aligned DNA**. The sample is output to the screen (more conserved positions are shown by brighter colors) and positional nucleotide weights are computed as

$$w[b,k] = \log(N[b,k]+0.5)$$

$$-0.25 \sum_{i=\mathrm{A,C,G,T}} \log(N[i,k]+0.5),$$

where $N[b,k]$ is the count of nucleotide b in alignment position k. The profile is output to a ***.pat*** file using option **File → Save Profile**.

If only one site or a motif generated by mutational analysis is known, profile can be generated using option **DNA similarity → Psi-Site**. This is a DNA analog of the well known protein analysis program *Psi-BLAST* [15]. The program works as follows. An initial site or motif is set. Candidate sites similar to the initial motif are determined and the corresponding genes are analyzed manually using functional or evolutionary (comparative) reasoning. The selected sites serve as the base for the new profile, etc. (only the best site for each gene is used, although the window evoked by **View → Details** shows all sites with scores exceeding the threshold). The site search is repeated, and the procedure is iterated the desired number of times. Intermediate results can be retained in files ***.pat*** (the current rule) and ***.gls*** (list of sites). The search parameters (boundaries of regulatory regions, thresholds, etc.) are set in the special dialog window that is open during all the time when the *Psi-site* procedure is active.

SERVICE PROGRAMS

Input

Initialization (downloading of files) and termination of a working seance is done from the main menu, point **File** with the following options:

File

Open sequence—donwload genome(s); open the annotation window.

Open CMP file—download file with alignments (see *Gene comparison*); open the similarity window.

Open Text File—download text file.

Options—set up colors and fonts; add menu points (see *Installation and parameter files*).

Exit—quit.

Output

Output of analysis results, protein and gene sequences, and genome fragments can be performed in three ways: using the main menu (to a file), the local menu of the annotation window (to a file or the clipboard), and the local menu of the map window (to the clipboard). At that, some objects can be output in any way, and some objects require a specific menu.

Output from the main menu is done from the point **Export** having the following options:

Export—output to a file:

Fragment—annotated genome fragment (file ***.dne***, format *EMBL*). Large fragments can be output from the map window using highlighting with pressed left mouse button.

Collection—set of genome fragments corresponding to marked genes (file ***.dnc*** in the *FASTA* format, retains strand, genome positions, corresponding gene).

Current Protein—amino acid sequence of the current gene (file ***.ppt***, format *SwissProt*).

Prot. collection—amino acid sequences of all marked genes (file ***.ppc***, format *EMBL*).

Extended Gene List—complete results of similarity or site search (file ***.gls***).

Brief Gene List—partial information about results of similarity or site search (file ***.gls***).

Histogram—histogram of similarity or site scores.

Output from local menu of the annotation window can be done to a file or the clipboard.

Copy—copy to the clipboard:

Marked text—marked text from the annotation window.

Current Prot.—amino acid sequence of the current gene (format *FASTA*).

Prot. collection—amino acid sequences of all marked genes (format *FASTA*).

Site—regulatory sites of the current gene (see "Analysis of regulatory signals").

Site report—regulatory sites of all marked genes (see "Analysis of regulatory signals").

Export—output to a file:

Current prot.—amino acid sequence of the current gene (file ***.ppt***, format *SwissProt*).

Gene—nucleotide sequence of the current gene (file ***.dne**, format *EMBL*).

Prot. collection—amino acid sequences of all marked genes (file ***.ppc**, format *FASTA*).

Output from the local menu of the map window allows for copying to the clipboard.

Copy—copy to the clipboard:

DNA sequence—genomic DNA fragment marked on the genomic map (format *FASTA*).

Linear map—marked fragment of the local map in the graphic representation that can then be printed in the standard way.

Search

Directly from the main menu or using shortcuts one can perform the following searches:

Search

Find Text—search for a text fragments in all fields (shortcut: F4).

Find Product—search for text fragment in the /product field (shortcut: Ctrl-P).

Find Gene—find a gene given the name shortcut: Ctrl-G).

Find Position—find a functional unit covering the given positions.

Sequence—find an exact appearance of a sequence fragment (shortcut: Ctrl-S).

Next Search—repeat the previous search (shortcut: F3).

Select by List—use an existing ***.gls** file to mark genes. Old marks are deleted.

Add Selection—all marks from an existing ***.gls** file. Old marks are retained.

Text (but not sequence) search patterns can include wildcards:

?—one arbitrary symbol,

*—several arbitrary symbols, so that **HI?31** will find, for example, **HI031**, **YHI131**, **ZZHIZ3199**, etc., whereas **HI?31** will additionally find **HI45631** and **YHI66631999**.

INSTALLATION AND PARAMETER FILES

Installation of the program is done as follows.

1. Create folders for programs, genomes, and recognition profiles.

2. Copy files ***genome.exe***, ***cmpwnd.exe***, ***genome.hlp***, ***blosum62.mtx*** to the program folder.

3. Copy files ***.pat** to the profiles folder.

4. Fill the genomes folder. Genomes are files in the *EMBL* or *GenBank* format. If genomes are downloaded from the Internet, they should be saved in the ASCII format (***.txt**), and not as HTML files.

5. In order to create the list of available genomes, call *genome* and choose **File > Open Sequence**. An empty window will appear. Pressing the **Add to List** button allows one to set the name of the genome file and to describe the taxonomy in the second window. To add the genome to the list automatically, press **Browse** and find the genome file.

Customization is done using the main menu option **File → Options**. In the dialog window one can change the colors and fonts of the description elements and the genetic map. These parameters as well as the default alignment parameters and the current folder name are retained in the file *genome.ini* in the ***Windows*** folder. It is a text file that can be edited using standard tools.

To create new menu points, use the option

File → Options

Tools—add calls to external programs.

Internet—add calls to Internet URLs in the *Netscape* browser.

SAMPLE SCENARIO OF REGULON ANALYSIS

A typical situation where *GenomeExplorer* can be used is described below. Let the learning sample contain several transcription factor binding sites from one ("old") genome. The aim is to find analogous files in a related ("new") genome and, possibly, to find new binding sites in both genomes. This section describes the basic scheme of such analysis, whereas the next section is dedicated to possible problems.

First, one should check whether the new genome contains an ortholog of the relevant transcription factor. Indeed, if there is no regulator, there is no reason to assume conservation of the regulon. The degree of similarity of the two factors should be sufficiently high and uniformly distributed along the global alignment. The domain similarity in this case is insufficient (but cf. [9]). However, we cannot provide any specific estimates for the "acceptable" similarity level.

At the next step one should identify in the new genome orthologs of the regulated genes from the old genome. This can be done using the **Ortologs** option. The working definition of orthologs is that each of them is the closest relative (in its genome) to the other one: formally, genes g^* from the old genome and h^* from the new genome are orthologous if $sc(g^*, h^*) > sc(g, h^*)$ for any other gene g from the old genome, and $sc(g^*, h^*) > sc(g^*, h)$ for any other gene h from the new genome. Thus, in the table format of the similarity window, a pair of genes can be identified as

orthologs if the dot corresponding to this pair is the brightest dot (in the pseudocolor scale) in its row and its column.

After that, site search in the upstream regions of the selected genes is performed using the **Site Search** option. This is done using the recognition rule generated using the learning sample from the old genome, but with lower thresholds. The identified candidate sites can be used to construct a new recognition rule: dependent on specifics of the problem, it could be reasonable both to create a new rule for each genome and to construct a single rule, but on a larger learning set of sites.

The final step involves analysis of new operon members. To do that, one should perform independent site search in the two genomes, and then compare genes having candidate sites in upstream regions (**Compare Selection**). This generates pairs of related genes with candidate sites in upstream regions. To validate the prediction, one has to make sure that these pairs indeed included orthologs and not paralogs (*Orthologs*).

The use of the clipboard allows one to analyze the generated gene and protein samples with other programs and Internet servers, as well as export preliminary and final results to test editors.

Regulation analysis can be performed even if only one representative of the signal is known and there are no data about the regulator. Of course, in this case existence of the regulator and conservation of the regulon cannot be verified directly. Nevertheless, the general scheme stays. Procedure **Psi-Site** is used to find genes whose upstream regions contain sites similar to the known one. At that, simultaneous application of **Psi-Site** to two genomes allows one to apply to gene selection not only functional, but evolutionary reasoning: at each step the next iteration uses sites from upstream regions of orthologous genes.

In this case simultaneous analysis of several related genomes increases reliability of predictions. One or more genomes can be retained for control: if the constructed recognition rule gives reasonable and consistent results on the control genomes not used to construct the rules, this is the critical evidence for the validity of the rule. However, even in the simpler case described in the beginning of this section, simultaneous analysis of several genomes with varying level of taxonomic relatedness increases reliability of predictions.

LIMITATIONS

The described technique cannot be applied in the case when the considered genomes are very close (e.g., they are strains of one species), since in this case the degree of conservation of noncoding regions is so large, that appearance of candidate sites in similar positions is not two independent events.

Special care should be exercised when the analyzed regulon contains members of large multigene families (e.g. transporters or transcription factors). In this case it might be difficult to resolve the orthology relationships. Sometimes the latter can be analyzed using positional analysis: if genes adjacent to two homologous genes are orthologous, the genes in question also are likely orthologous.

One should keep in mind that the majority of genes annotated in complete genomes have not been analyzed experimentally. In particular, annotated gene starts are often incorrect. A warning sign is loss of similarity at the left side of protein alignment: this is easily detectable on the graphic representation of alignment in the similarity window. A related problem is sequencing errors, and in particular frameshifts. In the latter case two adjacent "genes" align with two parts of one gene from another genome (although the same phenomenon can be caused by bona fide gene fusion events). Finally, some, especially short, annotated genes could be just results of overprediction by statistical gene recognition programs.

A more substantial problem is caused by changes of the operon structure that can occur even at short evolutionary distances. Such changes involve breaking of operons into two parts and integration of new, sometimes functionally unrelated, genes into operons. Thus comparison of genes having candidate upstream sites should involve possible cotranscribed genes (**Mark operon**). The analysis of results should involve not only candidate sites upstream of a particular gene, but also sites in regulatory regions of upstream genes if they are transcribed in the same direction.

GenomeExplorer is available for Russian researchers free of charge.

ACKNOWLEDGMENTS

This study was supported by Anchorgen, Inc. (http://www.anchorgen.com). The authors are grateful to Yu.V. Kozlov, E.V. Koonin, R. Overbeek, A.B. Rakhmaninova, D.A. Rodionov, M.A. Roytberg, and J. Fickett for advice and comment.

REFERENCES

1. Gelfand, M.S., *J. Comput. Biol.*, 1995, vol. 2, pp. 87–115.
2. Frech, K., Quandt, K., and Werner, T., *Comput. Appl. Biosci.*, 1997, vol. 13, pp. 89–97.
3. Gelfand, M.S., *Mol. Biol.*, 1998, vol. 32, pp. 103–120.
4. Fickett, J.W., and Hatzigeorgiou, A.G., *Genome Res.*, 1997, vol. 7, pp. 861–878.
5. Robison, K., McGuire, A.M., and Church, G.M., *J. Mol. Biol.*, 1998, vol. 284, pp. 241–254.

6. Gelfand, M.S., *Res. Microbiol.*, 1999, vol. 150 (in press).

7. Mironov, A.A., and Gelfand M.S., *Molecular Biology*, 1999, vol. 33, pp. 109–114.

8. Mironov, A.A., Koonin, E.V., Roytberg, M.A., and Gelfand, M.S., *Nucleic Acids Res.*, 1999, vol. 27, pp. 2981–2989.

9. Gelfand, M.S. and Mironov, A.A., *Mol. Biol.*, 1999, vol. 33, pp. 772–778.

10. Vitreschak, A., Bansal, A.K., Titov, I.I., and Gelfand, M.S., *Biofizika*, 1999, vol. 44, pp. 601–610.

11. Gelfand, M.S., Mironov, A.A., Jomantas, J., Kozlov, Y.I., and Perumov, D.A., *Trends Genet.*, 1999, vol. 15, pp. 439–442.

12. Smith, T.F. and Waterman, M.S., *J. Mol. Biol.*, 1981, vol. 147, pp. 195–197.

13. Henikoff, S. and Henikoff, J.G., *Proteins*, 1993, vol. 17, pp. 49–61.

14. Roytberg, M.A., *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 1992, vol. 2, pp. 113–126.

15. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., *Nucleic Acids Res.*, 1997, vol. 25, pp. 3339–3340.