GRESA DEVELOPMENT TOOLS: THE ENVIRONMENT OF DEVELOPMENT AND TESTING OF APPLICATIONS IN THE FIELD OF THE DNA REGULATION SEQUENCES ANALYSIS

[1]Lobanova M., [1,3]Cheremushkin E., [1]Konovalova T., [2]Tchekmenev D., [1]Beschastnov E., [3]Dunaev A., [3]Murzin F.

[1] **Institute of Cytology and Genetics SB RAS, Novosibirsk,**

[2] **BIOBASE GmbH, Wolfenbuettel**

[3] **Institute of Informatics Systems SB RAS, Novosibirsk**

We have developed stable and easy-to-use C++ based program interface designated for the use by DNA analysis application programmers. This environment considerably simplifies the creation of fast-working applications in the area of bioinformatics. The main concepts used by GRESA DT are those of transcription factor (TF) and TF-binding sites. The environment is hierarchically organized in 3 levels – core, commonly accepted tools and experimental research tools.

Introduction

GRESA DT provides an easy-to-use, stable, and consistent programming interface based on the C++ language for regulatory DNA sequences analysis application programmers. GRESA DT environment is used to reduce otherwise complex tasks to only a few lines of code, thus keeping speed and stability of the application. GRESA DT environment considerably simplifies development of fast applications in the field of bioinformatics and is simple enough in use. Profound knowledge of the C++ language is not necessary, base understanding of classes is enough for use of the GRESA DT. Extensive documentation makes its idle time in use for novice bioinformatitian programmers. Support of the majority of the standard sequence formats and of various widely used methods is realized in the environment.

Description

Both experimental, and theoretical researches are carried out intensively for disclosing mechanisms of gene expression regulation. One of the base concepts playing a key role during a transcription, are transcription factors which are represented by the regulatory proteins able to recognize a specific short DNA sites. Therefore big attention is paid for detailed studying and recognition of the corresponding nucleotide fragments, named cis-elements or transcription factors binding sites (TFBS or sites)[1]. Despite of a variety of approaches, the problem of construction of exact methods of recognition TFBS cannot be considered finally solved. The reason is the big variety of the contextual, physical, chemical and conformational features of TFBS; mechanisms of DNA-protein interactions between TFBS and transcription factors; Specificity of the context surrounding TFBS, rate of conservatism of the nucleotide context in evolution.

Structure

GRESA DT environment has hierarchical structure.

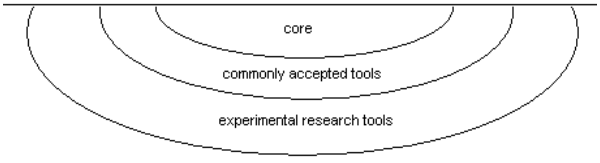

Fig.1 Hierarchical structure of GRESA package. Package consists of three tool levels, core tools, commonly accepted tools and experimental tools.

The package "core" consists of the classes representing the basic standard objects of bioinformatics science of regulatory DNA sequences:

Sequence – DNA sequence. Represents linear nucleotide sequence, designated by letters A, C, G, T. Also has the name, the description and binding to genome (chromosome number, start position on a chromosome and a direction "+" or "-").

Site – a subsequence of DNA sequence, usually 10-20 b.p., having position, length, and direction.

Factor – the object realizing properties of the transcription factor. Transcription factor is protein which binds a DNA site.

Alignment – set of the aligned sequences. In each of them gaps between nucleotides can be inserted. Alignment reflects evolutionary similarity of sequences.

Set of sequences, sites, factors – classes in which saving and loading from standard formats is realized.

And also there is a set of auxiliary classes. Above objects classical operations, such as, reception of the complimentary sequences, search etc. is realized.

The set of the standard tools consists of such applications as MATCH [2], COMATCH (search of composite modules), FOOTPRINT [3, 4], and CM SEARCH.

MATCH – a method of site search based on weight matrixes. The most widely used method.

COMATCH – a method, searching composite elements and sites with two domains.

FOOTPRINT – a method which is taking into account evolutionary similarity of sequences. In the beginning alignment of sequences and then search of sites which have met on both sequences in one and the same block of alignment are made.

CM SEARCH – a method of search of the composite modules regulating group of genes. For the given set of genes the common module, regulating these genes is searched.

The set of experimental tools will consist of yet not published applications, taking place in a stage of development. Among them context using sites search, estimations of quality of recognition methods.

Development and application

GRESA DT environment all time is in the development stage. Extreme Programming technology of environment development makes it possible to support the stable working version. Stability, at rather big and distributed group of developers, is supported by large number of the automated tests. Life cycle of the separate application consist of stages when the application is in a stage of experimental development, then passes in a stable stage.

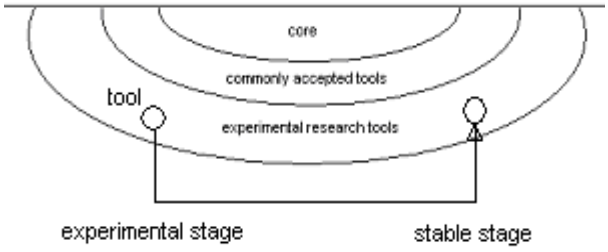

Fig.2. Life cycle of the separate application consist of stages when the application is in a stage of experimental development, then passes in a stable stage.

Further it can proceed into a set of the standard tools. Any team member can make changes to any class, the main thing – to keep successful performance of tests.

At present GRESA DT is used for regulatory DNA sequences processing. Application covers the wide range of site recognition tasks. Also some methods of regulation prediction are realized on the basis of the predicted sites. There is an opportunity to construct combinations of methods, for

example as a basis of footprint-search of sites it is possible to take either result of match-search, or results of any other method, supporting the necessary format. In GRESA DT comparative testing methods are implemented. The testing system estimates quality of site recognition. At present TRANSFAC [5] database is used for testing.

References

1. A biophysical approach to transcription factor binding site discovery. Djordjevic M, Sengupta AM, Shraiman BI.
2. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E.
3. Blanchette M., Schwikowski B., Tompa M. " Algorithms for phylogenetic footprinting. " *J Comput Biol.* 9, 211-223 (2002)
4. Whole genome human/mouse phylogenetic footprinting of potential transcription regulatory signals. Cheremushkin E, Kel A. Pac Symp Biocomput. 2003; 291-302.
5. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S. and Urbach, S. " The TRANSFAC system on gene expression regulation. " *Nucleic Acids Res.* 29, 281-283 (2001)