RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS

PROCEEDINGS OF THE THIRD INTERNATIONAL CONFERENCE ON BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE

Volume 1

BGRS' 2002
Novosibirsk, Russia
July 14 - 20, 2002

IC&G, Novosibirsk, 2002

International Program Committee

Nikolay Kolchanov, Institute of Cytology and Genetics, Novosibirsk, Russia (*Chairman of the Conference*)
Ralf Hofestadt, University of Bielefeld, Germany (*Co-Chairman of the Conference*)
Philip Bourne, SDSC, San-Diego, USA (*Co-Chairman of the Conference*)
Nickolai Alexandrov, Ceres Inc., Malibu, USA
Philipp Bucher, Swiss Institute for Experimental Cancer Research, Switzerland
Julio Collado-Vides, National University of Mexico, Mexico
Jim Fickett, AstraZeneca, Boston, USA
Paolo Frasconi, University of Florence, Firenze, Italy
Sergey Goncharov, Sobolev Institute of Mathematics, Novosibirsk, Russia
Igor Goryanin, GlaxoSmithKline, UK
Charlie Hodgman, GlaxoSmithKline, UK
Elza Khusnutdinova, Institute of Biochemistry and Genetics, Ufa Sci. Centre RAS (Ufa), Russia
Lev Kisselev, Engelhardt Institute of Molecular Biology, Moscow, Russia
Boris Kovalerchuk, Central Washington University (Ellensburg), USA
Luciano Milanesi, ITBA, Milan, Italy
John Reinitz, The University at Stony Brook, N.Y., USA
Akinori Sarai, RIKEN Tsukuba Life Science Center, Tsukuba, Japan
Ilya Shindyalov, San Diego Supercomputer Center, USA
Rustem Tchuraev, Institute of Biology, Ufa Sci. Centre RAS, Ufa, Russia
Masaru Tomita, Institute for Advanced Biosciences, Keio University, Japan
Edgar Wingender, GBF, Braunschweig, Germany
Nikolay Yankovsky, Institute of General Genetics, Moscow, Russia
Lev Zhivotovsky, Institute of General Genetics, Moscow, Russia

Local Organizing Committee

Dagmara Furman, Institute of Cytology and Genetics, Novosibirsk,
Nadya Omelianchuk, Institute of Cytology and Genetics, Novosibirsk,
Sergey Lavryushev, Institute of Cytology and Genetics, Novosibirsk,
Galina Kiseleva, Institute of Cytology and Genetics, Novosibirsk,
Elena Borovskikh , Institute of Cytology and Genetics, Novosibirsk,
Nikolay Shkel, Institute of Cytology and Genetics, Novosibirsk,
Andrey Kharkevich, Institute of Cytology and Genetics, Novosibirsk,

The information about the Conference BGRS' 2002 is presented at http://www.bionet.nsc.ru/meeting/bgrs2002/

Our sponsors

Organizers

Institute of Cytology and Genetics, SB RAS

Siberian Branch of the Russian Academy of Sciences

Grants

INTAS Conference Grant

Glaxo Wellcome Inc.

Russian Foundation for Basic Research

Ministry of Industry, Science and Technologies of the Russian Federation

Information sponsors

San Diego Supercomputer Center, United States

RIKEN Tsukuba Institute

Bielefeld University, Faculty of Technology

http://www.karger.com/

In Silico Biology

Others

KWESTA-group: computers, computer accessories, service

CONTENTS

REGULATORY GENOMIC SEQUENCES

GENOME STRUCTURE AND FUNCTION

INTRODUCTION

Four volumes of Proceedings of the Third International Conference on Bioinformatics of Genome Regulation and Structure – BGRS' 2002 (Akademgorodok, Novosibirsk, Russia, July 14-20, 2002) incorporate about 180 annotated extended abstracts (short papers) devoted to the actual problems in bioinformatics of genome regulation and structure.

The Conference BGRS' 2002 is organized by the Laboratory of Theoretical Genetics of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. BGRS' 2002 is the third in the series. It will continue the traditions of the previous conferences, BGRS' 98 and BGRS' 2000, which were held in Novosibirsk in August 1998 and 2000, respectively.

As the greatest scientific event within the period between the Conferences BGRS' 2000 and BGRS' 2002, could be undoubtedly viewed the completion of human genome draft sequencing. This event has initiated the beginning of the post-genome era in biology. This era is characterized by sharp increase in research scale in the fields of transcriptomics, proteomics, and systemic biology (gene interaction, gene network functioning, signal transduction pathways), without loosing the fundamental interest to studying structural genome organization.

The structure and regulation of genome are the counterparts of life at molecular level; that is why understanding of fundamental principles of regulatory genomic machinery is impossible unless their structural organization is known, and *vice versa*.

The huge volume of experimental data that has been acquired on genome structure, functioning and gene expression regulation demonstrate the blistering growth. Development of informational-computational technologies of novel generation is a challenging problem of bioinformatics. Bioinformatics has entered that very phase of development, when decisions of the challenging problems determine the realization of large-scale experimental research projects directed to studying genome structure, function, and evolution.

By analyzing the papers submitted for publication in the four-volume issues of the BGRS' 2002, the Program Committee came to a conclusion that participants of the Conference have concentrated their attention at consideration of the hottest items in bioinformatics listed below: (i) regulatory genomic sequences: databases, knowledge bases, computer analysis, modelling and recognition; (ii) large-scale genome analysis and functional annotation; (iii) gene structure finding and prediction; (iv) comparative and evolutionary genomics; (v) computer analysis of genome polymorphism and evolution; computer analysis and modelling of transcription, splicing and translation; structural computational biology - genomic DNA, RNA and protein structural and functional organization; (vi) gene networks, signal transduction pathways and genetically controlled metabolic pathways: databases, knowledge bases, computer analysis, and modelling; principles of organization, functioning, and evolution (vii) data warehousing, Knowledge Discovery and Data Mining; (viii) analysis of fundamental regularities in genome functioning, organization, and evolution.

The researchers working in the fields of experimental biology are also invited to participate in the work of BGRS' 2002 in order to develop a sort of interface between experimental and computer-assisted researches in the fields of genomics, transcriptomics, proteomics, structural and systemic biology, as well as for contributing to promotion of computational biology to experimental research. These results are highlighted in the fourth volume of BGRS' 2002 Proceedings.

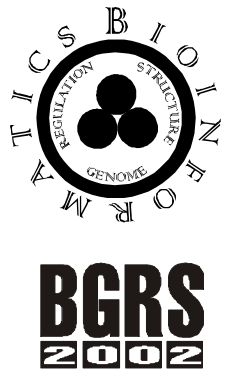
All the questions listed above will be suggested to consideration of participants of BGRS' 2002 at plenary lectures, oral communications, poster sessions, Internet computer demonstrations, and round-table discussions.

The Conference is sponsored by Siberian Branch of the Russian Academy of Sciences, by the Institute of Cytology and Genetics SB RAS, by Russian Foundation for Basic Research, by Russian Ministry of Industry, Science and Technologies, by the Company Glaxo Research and Development Limited, by independent International Association formed by the European Community INTAS. The Organizing Committee of the Conference tender thanks to all the sponsors for financial support.

Professor Nikolay Kolchanov
Head of Laboratory of Theoretical Genetics
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Chairman of the Conference

Professor Phil Bourne
SDSC, San-Diego, USA
Co-Chairman of the Conference

Professor Ralf Hofestaedt
Faculty of Technology
Bioinformatics Department
University of Bielefeld, Germany
Co-Chairman of the Conference



REGULATORY GENOMIC SEQUENCES

TRANSCRIPTION REGULATORY REGIONS DATABASE (TRRD): ITS STATUS IN 2002

Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G., * Kolchanov N.A.

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: kol@bionet.nsc.ru
*Corresponding author

Key words: *database, transcription regulation*

Motivation: The main purpose for development of TRRD was to provide a most comprehensive and adequate description of the structure–function organization of transcription regulatory regions in eukaryotic genes.

Results: The current release of TRRD comprises the data on 1405 genes, 2158 regulatory units, 7419 expression patterns, 6646 transcription factor binding sites, and 14 locus control regions. All the data have been inputted into TRRD from 4779 annotated publications and are distributed between the seven following databases: TRRDGENES, TRRDUNITS, TRRDEXP, TRRDSITES, TRRDFACTORS, TRRDLCR, and TRRDBIB. SRS is used as a main searching and navigation tool.

Availability: http://www.bionet.nsc.ru/trrd/

Introduction

Transcription is the first and key stage of an intricate multilevel process of gene expression. An avalanche of experimental data on the transcription level of gene expression regulation brings about an exigent need to develop computer databases for arrangement, storage, and use of the information obtained. The Transcription Regulatory Regions Database (TRRD), developed at the Institute of Cytology and Genetics SB RAS (Novosibirsk), provides an integrated description of transcription regulation of eukaryotic genes transcribed by RNA POL II. TRRD complies the information on structure–function organization of extended transcription regulatory regions, which may include several hierarchical levels. Transcription factor binding sites, described in TRRDSITES, belong to regulatory elements of the first level. Functionally related sets of sites compose regulatory units (promoters, enhancers, and silencers), representing the next level of regulation (described in TRRDUNITS). Note that regulatory units may be localized to different gene regions: 5'-flanking, 3'-flanking, exons, and introns. Locus control regions (LCR) form the next regulatory level. LCRs are responsible for concerted expression of several genes within one locus and may contain several regulatory units (enhancers, positive, or negative regulatory regions) and transcription factor binding sites. The database TRRDLCR contains the structure–function characteristics of LCRs. TRRDEXP comprises the data on qualitative distinctive features of gene expression in any organs, tissues, cell types, and cell lines in a form of expression patterns. The database TRRDFACTORS compiles the information on the transcription factors interacting with binding sites. An entry of TRRD corresponds to a particular gene. All the pieces of general information on genes together with hierarchically organized presentation of all the regulatory elements are accumulated in TRRDGENES. All the information is inputted into the database by experts in biology basing on analysis and annotation of papers reporting experimental data. Each type of experiment is designated with specific digital code, indicated in the fields ExperimentCodes (AG) of the databases TRRDGENES, TRRDSITES, and TRRDUNITS. All the bibliographic information is available in TRRDBIB.

Methods and Algorithms

The formats of TRRD releases 4.1 and 4.2 are described in (Kolchanov et al., 1999; Kolchanov et al., 2000); the formats of new databases TRRDLCR and TRRDUNITS as well as new fields supplemented to TRRDSITES (TRRD release 6.0), in (Kolchanov et al., 2002).

Syntactic and semantic analyses using original programs accompany the data input into TRRD. The former program—TRRD-INPUT, realized in Visual FoxPro 5.0 using OLE technology and ActiveX elements (Ananko et al., 1998)—checks the inputted terms for their compliance with the controlled vocabularies. The latter program—TRRD-Pars, realized in Visual C++ medium—verifies in a dynamic mode the information on nucleotide sequences and positions of transcription factor binding sites from TRRDSITES using the sequences from EMBL/GenBank referred to in TRRD. Another function of the program TRRD-Pars is automated generation of a block of fields in the database TRRDUNITS containing nucleotide sequences of regulatory units retrieving the necessary information from TRRD and EMBL/GenBank. These fields are generated through comparing the corresponding data in TRRDGENES, TRRDSITES, and EMBL/GenBank including several stages. At the first stage, the starting point used for describing a regulatory unit in question in TRRDGENES (transcription start, translation start, or beginning of the sequence) is considered. The data on positions of sites in EMBL/GenBank sequence and their distance from the starting point is used to determine the position of the starting point in EMBL/GenBank sequence (Fig. 1a). At the second stage, the regulatory unit in question is localized and the corresponding

region of EMBL/GenBank sequence is extracted using the position of the starting point found in combination with annotated data on the positions of the regulatory unit bounds relative to the starting point (Fig. 1b).

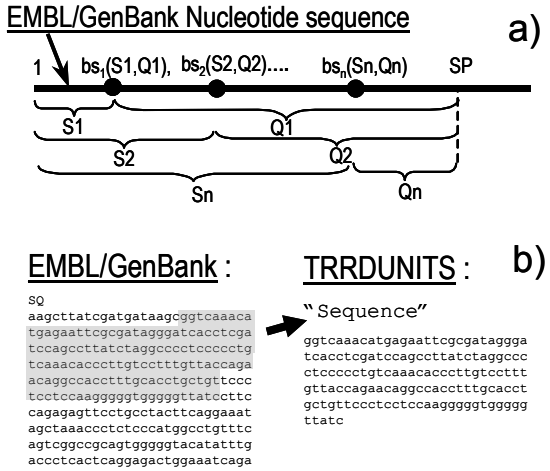


Fig. 1. Extraction of nucleotide sequences of regulator units from EMBL/GenBank: (**a**) determining the position of starting point (SP) in EMBL/GenBank entry from positions of sites relative to SP (S1, S2,…Sn) and the position of first nucleotide of the sites (Q1, Q2,…Qn) and (**b**) extracting the region of nucleotide sequence corresponding to the regulatory unit in EMBL/GenBank.

Sequence Retrieval System (SRS) v. 6 is used to integrate the seven databases of TRRD with one another and other available informational and software modules of GeneExpress-2 (Kolchanov et al., 1999).

A new version of the program TRRD-Viewer, presenting the data compiled in TRRD as maps of gene regulatory regions, was developed in Java using JDK 1.1.8 and tested in web browsers under MS, Windows, and Linux operational systems.

Implementation and Results

Informational content. TRRD is the largest informational module of GeneExpress-2 (Kolchanov et al., 1999), available at <http://wwwmgs.bionet.nsc.ru/mgs/gnw/>. TRRD is supplemented with new information monthly. The numbers of entries in TRRD release 4.2.5 (Kolchanov et al., 2000) and the current release 6.01 are listed in Table.

Table. TRRD informational contents.

Database	Number of entries in release 4.2.5*	Number of entries in release 6.01**	Number of indexed fields in release 6.01
TRRDGENES	760	1405	24
TRRDUNITS	–	2158	11
TRRDEXP	3403	7419	17
TRRDSITES	3604	6646	16
TRRDFACTORS	2862	5735	14
TRRDLCR	–	14	40
TRRDBIB	2537	4779	9

*As of March 01, 2002; **As of May 14, 2002.

Within TRRD, the following topic sections are developed, uniting genes according to their functional characteristics: Heat Shock-Induced Genes (HS-TRRD), comprising 106 entries; Interferon-Inducible Genes (IIG-TRRD), 114 entries; Erythroid-Specific Regulated Genes (ESRG-TRRD), 66 entries; Genes of Lipid Metabolism (LM-TRRD), 103 entries; Endocrine System Transcription Regulatory Regions Database (ES-TRRD), 124 entries; Glucocorticoid-Regulated Genes (GR-TRRD), 64 entries; Plant Genes (PLANT-TRRD), 136 entries; Cell Cycle-Dependent Genes (CYCLE-TRRD), 55 entries; and Redox-Sensitive Genes (ROS-TRRD), 81 entries.

The visualization system TRRD-Viewer was supplemented with additional options compared with the previous version (Kolchanov et al., 1999). The new version of TRRD-Viewer (Kolchanov et al., 2002) provides a quicker loading and increased processing power of graphical tools compared with the previous version. Operation of the TRRD-Viewer is shown by the example of human CYP7 gene regulatory regions (Fig. 2).

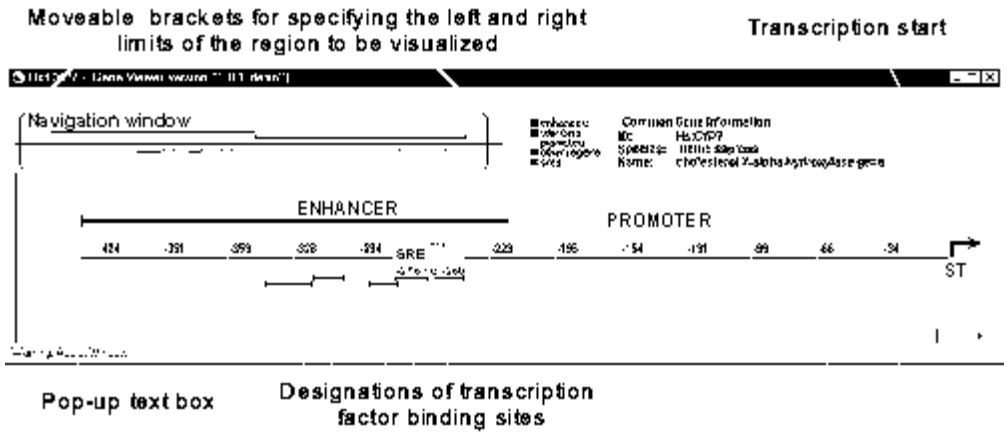


Fig. 2. The interface of TRRD-Viewer.

Data search in TRRD. SRS (Sequence Retrieval System) is the major tool for searching TRRD for the data of interest using key words. The total number of indexed fields available for the SRS-based search amounts to 131 (Table). The search for genes of interest according to their names using browsers is also provided. A quick access to the genes from the topic sections listed above is available via the corresponding TRRD sections. The program BLAST (http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/units_blast.html) allows the regulatory regions homologous to an analyzed DNA sequence to be searched for.

Analysis of DNA sequences is available using the program BinomSite. This program searches for the regions homologous to the transcription factor binding sites compiled in TRRD (http://wwwmgs.bionet.nsc.ru/mgs/programs/mmsite/) in a DNA sequence of interest.

Relational version of the TRRD database was developed in ORACLE8i medium. XML representation is used as an exchange format. A specialized loading program, transforming the data into XML format, is used for loading the data from flat file. By now, the relational TRRD version comprises 102 tables (52 informational and 50 linking tables); the TRRD data scheme is available at http://www.bionet.nsc.ru/trrd/RelScheme/.

Discussion

Several databases compiling the information on various aspects of eukaryotic gene transcription regulation—EPD, TRANSFAC, and COMPEL—are now available. However, neither of them provides an integrated description of transcription regulation. TRRD is a unique database, as it contains simultaneously the data obtained while studying extended regulatory regions, transcription factor binding sites, and specific expression patterns of various eukaryotic genes. The informational system TRRD comprises over 130 indexed fields (Table 1) distributed over 7 databases. TRRD accumulates only published experimental information upon its syntactic and semantic verification. TRRD contains the largest in the world collections of annotated regulatory units of eukaryotic genes (about 2000) and transcription factor binding sites (6431). The data on gene expression patterns and regulatory elements responsible for their realization provides for the first time the possibility to analyze molecular genetic systems of different organisms at the level of gene networks.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, and 00-04-49255); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors are grateful to I.V.Lokhova and L.V.Katokhina for bibliographic support; D.A.Grigorovich and E.V.Maksakov, for development of software; experts in biology O.E.Belova, T.V.Busygina, V.M.Merkulov, T.N.Goryachkovskaya, V.V.Suslov, T.M.Khlebodarova, S.A.Fedorova, S.S.Ibragimova, O.G.Smirnova, and A.L.Proskura, for annotating the literature.

References

1. Ananko E.A., Naumochkin A.N., Fokin O.N., Frolov A.S. (1998). Programs for data input to the Transcription Regulatory Regions Database. Proc. First International Conf. on Bioinformatics of Genome Regulation and Structure, (BGRS'98), ICG, Novosibirsk. 1:29-32.
2. Kolchanov N.A., Ananko E.A., Podkolodnaya O.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (1999). Transcription Regulatory Regions Database (TRRD): its status in 1999. Nucl. Acids Res. 27:303-306.

3. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30:312-317.
4. Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Korostishevskaya I.M., Romashchenko A.G., Overton G.C. (2000). Transcription Regulatory Regions Database (TRRD): its status in 2000. Nucl. Acids Res. 28:298-301.
5. Kolchanov N.A., Ponomarenko M.P., Frolov A.S., Ananko E.A., Kolpakov F.A., Ignatieva E.V., Podkolodnaya O.A., Goryachkovskaya T.N., Stepanenko I.L., Merkulova T.I., Babenko V.V., Ponomarenko Yu.V., Kochetov A.V., Podkolodny N.L., Vorobiev D.V., Lavryushev S.V., Grigorovich D.A., Kondrakhin Yu.V., Milanesi L., Wingender E., Solovyev V., Overton G.C. (1999). Integrated databases and computer systems for studying eukaryotic gene expression. Bioinformatics. 15:669-686.

THESAURUS AS A TOOL FOR SEARCHING TRRD DATABASE

**** Ananko E.A., Grigorovich D.A., Podkolodny N.L, Ignatieva E.V., Podkolodnaya O.A., Korostishevskaya I.M.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: eananko@bionet.nsc.ru
*Corresponding author

Key words: *transcription regulation, database, thesaurus*

Resume

Motivation: One of the problems of when searching for the information needed in databases is to take into account the presence of synonyms of numerous terms used and hierarchical organization of concepts. The majority of retrieval systems search only for the certain textual string (keyword), which is frequently inconvenient for the user.

Results: A retrieval system assisting the user in working with TRRD, which addresses particular biological problems, has been developed. While operating, the retrieval system uses hierarchically organized thesauruses of cells, tissues, and organs, permitting the queries to the SRS TRRD version not only by a particular keyword, but also retrieval of the data on all the connected words (daughter terms with reference to the initial word) or all its synonyms simultaneously. A specialized section of the TRRD database on tissues and organs, helping the user to obtain additional reference information, were created.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/thesaurus/.

Introduction

Development of informational technologies arises the problem of searching for the information needed. A large diversity of retrieval systems is presently available. However, the most critical problem encountered by users of these systems is synonymy of terms. Most frequently, it is impossible to retrieve the necessary information without knowing the synonyms. In addition, similarly to other specialized databases, it is necessary to find all the information connected with a word of interest when working with molecular biological databases. For this purpose, the user has to search not only for the main keyword, but also for all the words connected with it. This is a time-consuming and inconvenient procedure, giving no guarantee that the entire information of interest is retrieved.

Hierarchically ordered thesauruses and vocabularies of synonymic terms help solving such problems. We developed a specialized retrieval system using such thesauruses for the TRRD database (Kolchanov et al., 2002). This assists the TRRD users in searching for genes expressed in certain cells, tissues, and organs. The system is available at the specialized TRRD section (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/thesaurus/).

Methods and Algorithms

The following types of relations are used in the hierarchically organized thesauruses of cells, tissues, organs, and ontogenetic stages: scope note (SN), broader–narrower (BT/NT), whole–part (TT/NT), relationships (RT), and synonymy (USE/UF).

The retrieval program forms a set of daughter keywords from the keyword tree, then a subset of synonyms, and creates a query to the SRS table TRRDEXP4 (Kolchanov et al., 2002), retrieving the corresponding fields where at least one word from the keyword set is met. Then, the entries with a zero expression level are cut off, and two SRS tables—TRRDGENES4 and TRRDEXP4—are automatically linked. Upon linking, the user receives a list of the entries from the table TRRDGENES4.

Implementation and Results

The technology for developing and maintaining controlled glossaries and vocabularies for TRRD was developed and has been optimized (Ananko et al., 1998). Basing on these glossaries and vocabularies, hierarchically organized thesauruses of cells, tissues, organs, and ontogenetic stages, including synonyms, were developed. Thesauruses on mammalian tissues and organs provide the user with supplementary information on their cell compositions, localizations, tissue origins, and functions of entire organs and their parts. The thesaurus-based search for information in TRRD was worked out, and a specialized retrieval system maintaining the relations of the broader–narrower (BT/NT), whole–part (TT/NT), relationships (RT), synonymy (USE/UF), and other types while its operation was developed. This retrieval system is a component of the system GeneExpress (Kolchanov et al., 1998), is available via the Internet (Fig. 1), and links the thesauruses with the TRRD database, allowing the list of genes expressed in a tissue or organ in question to be retrieved (Fig. 2).

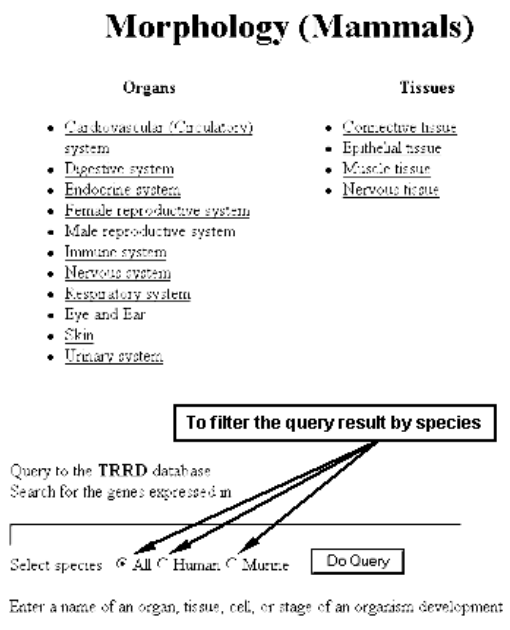


Fig. 1. The main page of the informational system on mammalian tissues and organs.

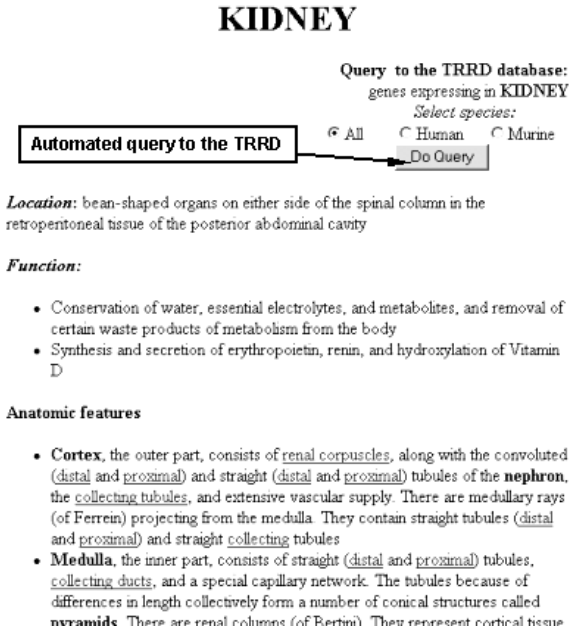


Fig. 2. The page containing reference information on kidneys.

The retrieval system developed addresses particular biological problems, namely, search for the genes expressed under conditions specified. This system allows queries to the TRRD SRS version to be made not only by a keyword inputted, but also by all the related words (daughter with respect to the initial word) within the corresponding glossary or vocabulary simultaneously. In this process, the two SRS tables of TRRD—TRRDEXP4 and TRRDGENES4 (Kolchanov et al., 2002)—are automatically linked. The work of user is essentially facilitated, as he(she) receives a list of genes with indication of all the synonyms and organism species (Fig. 3), not expression patterns, as is typical of a conventional SRS-based retrieval system. For example, when a query is generated using 'kidney' as a keyword, the system searches for the database entries containing other words, namely, parts of kidneys, such as 'kidney cortex'| 'tubules' |'glomerulus' |'proximal convoluted tubules'. The entries where the expression level in organs are marked as 'none|undetectable', that is, the expression in this organ was studied and found to be close to zero, are excluded. Overall, 120 of such genes were found in the TRRD database (Fig. 3).

Such retrieval system is convenient for the user and allows a set of tissue-specific genes to be formed easily.

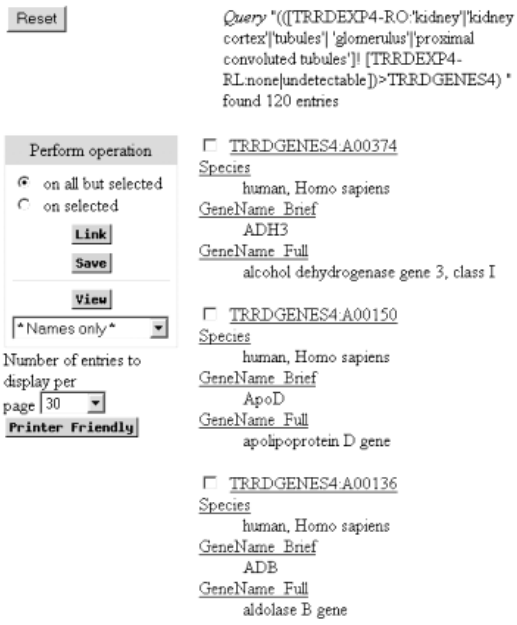


Fig. 3. A result of the query to database using 'kidney' as a keyword.

Discussion

The developed approach to organization of retrieval system offers great opportunities for using the TRRD database while solving particular biological problems. Further development of the retrieval system is in progress. It includes developing hierarchically ordered thesauruses for two additional fields: external stimuli and transcription factors. These thesauruses will be available at the TRRD www-site in the nearest future. This will allow users to search for the genes regulated by certain external stimuli and transcription factors.

In addition, we are developing a modified version of the retrieval system that would search for the genes whose expression meets simultaneously several conditions, namely tissue-specificity, ontogenetic stage, external stimulus, and transcription factor.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 00-07-90337, 00-04-49229, 00-04-49255, 01-07-90376, 01-07-90084, and 02-07-90359); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health USA (grant № 2 R01-HG-01539-04A2), and US Department of Energy (grant № 535228 CFDA 81.049). Authors are grateful to I.V.Lokhova and L.V.Katokhina for bibliographical support, and to G.B.Chirikova for translation of the paper into English.

References

1. Ananko E.A., Naumochkin A.N., Fokin O.N., Frolov A.S. (1998). Programs for data input to the Transcription Regulatory Regions Database. Proc. First Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'1998). ICG, Novosibirsk. 1:29-32.
2. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30:312-317.
3. Kolchanov N.A., Podkolodny N.L., Ponomarenko M.P., Ananko E.A., Ignatieva E.V., Kolpakov F.A., Levitsky V.G., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Vorobiev D.G., Lavryushev S.V., Grigorovich D.A., Ponomarenko J.V., Kochetov A.V., Orlova G.V., Kondrakhin Yu.V., Titov I.I., Vishnevsky O.V., Orlov Yu.L, Valuev V.P., Ivanisenko V.A, Oshchepkov D.Yu., Omel'yanchuk N.A., Pozdnyakov M.A., Kosarev P.S., Goryachkovskaya T.N., Fokin O.N., Kalinichenko L.A., Kotlyarov Yu.V. (2000). Integrated system on gene expression regulation GeneExpress-2000. Proc. Second Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2000). ICG, Novosibirsk. 1:12-18.

MATHEMATICAL TOOLS FOR REGULATORY SIGNALS EXTRACTION

Regnier M.

INRIA Rocquencourt - Domaine de Voluceau BP 105, Le Chesnay, 78153, France, e-mail: Mireille.Regnier@inria.fr

Key words*: regulatory signals, statistics, protein binding sites, Markov models*

Resume

Motivation: Statistics for the number of occurrences of a set of words has numerous applications and a great interest arose recently (van Helden, 1998; Hampson et al., 2002; Reinert et al., 2000; Régnier, 2000; Apostolico et al., 1999). The underlying assumption is as follows: a biological function that is enhanced or avoided is associated to a word, or a set of words that is overrepresented-or underrepresented-. This regulatory signal can be for example a protein binding site. We provide below a brief formalization of the main issues addressed in biological applications, that we classify into two main classes.

Exceptional words in long sequences One studies a long text-typically, a genome- and assumes a probability model on it. The goal is the extraction of exceptional words -either overrepresented or underrepresented-. To achieve this goal, one needs an algorithm that searches for candidate motifs and mathematical tools to assess statistical signicance of these candidates. A recent survey on algorithms can be found in (Lonardi, 2001; Marsan, 2002).

Set of independent identically distributed sequences A set of sequences is given. These sequences are relatively short, generated independently according to a common distribution, but may have different lengths (van Helden, 1998). One looks for exceptional signals, a word H, or a set of words H. One counts the number of sequences where H, or H, is actually observed and compares this number with its expected value. A typical example is the characterization of polyadenylation signals in human genes (Beaudoing et al., 2000). This scheme underlies the software *RSA-oligonucleotides* of van Helden (van Helden et al., 1998). An important related problem is the alignment of two (or more) sequences.

Results: In both cases, one needs a probability model on the input texts. Different background probability models are discussed in (Hampson et al., 2002). In this paper, we assume that the model is Markovian, or Bernoulli. We derive new tractable formulae and provide efficient algorithms that actually compute them. We also discuss widely used approximations. We state their validity domains and occasionally extend them. We discuss the critical domains -or phase transition phenomena- that are observed (Blanchette, Tompa, 2001; Robin, Schbath, 2001; van Helden et al., 1998_ and suggest a few solutions.

Large Sequences

In this section, we concentrate on the statistical signicance of word avoidance or overrepresentation in a large sequence, typically a genome. We assume that this sequence is randomly generated according to a Bernoulli or Markov model. The distribution of the number of occurrences of a given word H (or, possibly a given set of words $\mathbb{H}$) has been studied by various authors (Pevzner et al., 1989; Bender, Kochman, 1993; Régnier, Szpankowski, 1997; Régnier, 2000; Reinert et al., 2000). Notably, the expectation and the variance V (H) have been extensively studied. It is well known that the number of occurrences converge in distribution to a normal law. Still, this does not provide valuable information on the so-called P - value. Nevertheless, our combinatorial results on the distribution allow to prove (Denise et al., 2001; Denise, Régnier, 2002):

Theorem 1 Let H be a given pattern, and P(H) be its probability of occurrence. Let a be a real number such that a > P(H). Then:

log Prob(NH >= na) /n ~I(a) ; (1)

where

$$I(a) = a\log(\frac{D(z_a)}{D(z_a) + z_a - 1}) + \log z_a \qquad (2)$$

$$D(z) = (1-z)A(z) + P(H)\,z^m \qquad (3)$$

and z_a is the largest real positive root satisfying $0 < z_a < 1$ of

$D(z)^2\,(1 + (a-1)z)D(z) - az(1-z)D_0(z) = 0$ (4)

A similar result holds for avoided words (Vandenbogaert, 2002).

Approximation Quality

A more precise result is established in (Denise, Régnier, 2002). E.g.

$P rob(NH \geq na) \sim 1 / \sigma_a \sqrt{n}\ e^{-nI(a)+\delta a}$

where δ_a and σ_a are some functions of a and z_a. It turns out (Denise et al., 2001) that this expression is very close to the exact expression computed by the software *Excep* (Klaerr-Blanchard et al., 2000). Still, as the computation reduces to the numerical solution of a polynomial equation, this computation is much faster and numerically very stable. As a matter of fact, the possible range for the length n of the text and the probability P(H) is much larger. Applications to avoided words - restriction-modication systems for bacteria- can be found in (Vandenbogaert, 2002; Vandenbogaert, Makeev, 2002).

This can be compared to some common approximations for the p-values, e.g. the p-values computed for the normal law with mean P(H) and variance V (H) or for the (compound) Poisson distribution. First, we point out that similar results (Dembo, Zeitouni, 2002) are known for these laws, that allow to skip tedious computations. When a P(H) is not too large, a local development proves theoretically that the Poisson approximation (or the compound Poisson approximation) is much tighter than the normal approximation, a fact that was experimentally observed for some ranges of n and p in (Reinert, Schbath, 2001; Nicodème, 2001).

Small Sequences

In this context, one searches for a signal (a word H, possibly with errors, a structured motif) that occurs in a set of L small sequences more often than expected. Typically, these sequences are upstream regions of (possibly co-regulated) genes. One proceeds in two steps:

(i) compute for a given signal, say a word H, the probability p that it occurs at least once in the text;

(ii) compute P_L (k), the probability that k out of L sequences contain the motif at least once.

When the p-value P_L (k) is very small, one concludes that the extracted signal is relevant.

Poisson approximation A very common approximation (Buhler, Tompa, 2001) for p is:

$p = 1- (1- P (H))^{n-m+1}$

where n is the size of the sequence. It has been observed by many people (van Helden et al., 1998) that this approximation is bad when H is a self-overlapping word. We point out here that combinatorial results on words [Gonnet and Odlyzko 1981, Régnier2000] allow for an exact computation of p. E.g, 1-p is the n- th coefficient of z^n in the Taylor development of $A_H(z) /D(z)$. Practically, when n is small, this is easily computable by a symbolic computation system and it turns out that the Poisson approximation is rough. When n is larger, combinatorial analytics allow to show that $p\sim 1-e^{-n \log\rho}$ where ρ is the smallest real positive root of D(z) = 0. It is easily checked that $\rho\sim P(H) /A_H (1)$. The approximation

$p = 1-e^{-n \log(1+ P(H)/ AH (1))}\sim 1-e^{-n P(H)/AH (1)}$

is very tight.

Our main observation here is that PL (k) is the tail distribution of a Bernoulli process. Hence, it is known (Dembo, Zeitouni, 1992; Waterman, 1995) that $PL (k) = e^{-nI(a)}$ with A = k /L and

$$I(a) = a \log \frac{a}{LP(H)} \quad (6).$$

This approximation is very tight and was recently implemented in *RSA-tools*.

Conditional Expectation

The overrepresentation -or under representation- of a signal modifies the statistical properties of the sequence. It is valuable to eliminate artefacts of a strong signal in order to extract a weaker signal. It is also of interest to cluster similar motifs in a single degenerate signal.

Theorem 2. (Denise, Régnier, 2002) Let H and F be two patterns with probabilities of occurrence P(H) and P (F). Let A_{HF} (z) and A_{FH} (z) be their correlation polynomials. Assume that k occurrences of H are found in the text, with a = k/n, a > P(H). Then, the expected number of occurrences of F, knowing that H occurs k times is:

$E(F/NH = k) \sim n\nu(a)$

where

$\nu(a) = z_a [(1-z_a) A_{HF} (z_a) + P (F) z_a^m][(1- z_a) A_{FH} (z_a) + P(H) z_a^m] /(D(z_a)(D(z_a) + z_a -1))$:

We briefly present below two applications. Similar problems are studied in (Blanchette, Sinha, 2001).

Polyadenylation In (Beaudoing et al.), short (around 50 bp) upstream regions of EST of human genes are searched for polyadenylation signals. The largest Z-score is assigned to AAUAAA, that actually is the dominating signal. The computation of a Z-score using E(F=AAUAAA) as the new expected value for each pattern F drastically reduces the Z-

scores of artefacts AUAAAN and NAAUAA and assigns the 2-nd rank in Z-scores to AUUAAA, that actually is the second (weak) signal (Denise et al., 2001).

Arabidopsis thaliana RSA-tools was "blindly" used on upstream regions of *Arabidopsis thaliana*, by M. Lescot. In one test, the two highest Z-scores were assigned to H = ACGTGG and F = CACGTG. The number of occurrences were NH = 32 and N F = 52. The conditional expectations are: E(H/F) = 24 and E(F/H) = 20. Our conclusion is that H is better explained by F than F is explained by H. This leads to chose F = CACGTG as the significant signal, although its Z-score is smaller. As a matter of fact, the biological knowledge confirms that F is the regulatory signal.

Open problems

It is an interesting challenge to derive the rate function for a set of consensus words, or degenerated words (IUPAC code). The degree of the polynomial equation to be solved is proportional to the number of words. Still, it is likely that it can be kept down by a suitable use of combinatorial properties of consensus words (Clément et al., 2002). One can also extend the approximation (5). An other interesting issue is the derivation of the conditional expectation when a set of words is overrepresented. As above, one expects a simplification of the formulae for consensus words. Finally, it is worth extending (6) for sequences of different lengths.

Acknowledgements

I am grateful to magali lescot (marseille) for fruitful discussions and for providing data on arabidopsis thaliana. This research was partially supported by the future and emerging technologies program of the eu under contract number ist-1999-14186 (alcom-ft), the french-russian liapounov institute and intas 99-1476.

References

1. Apostolico A., Bock M.E., Lonardi S., Xu X. (1999) Efficient detection of unusual as words. J. of Computational Biology.
2. Beaudoing E., Freier S., Wyatt J., Claverie J., Gautheret D. (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. Genome Res. 10, 1001-1010.
3. Bender Edward A., Kochman F. (1993) The Distribution of Subwords Counts is Usually Normal. Eur. J. of Combinatorics. 14:265-275.
4. Blanchette M., Sinha S. (2001) Separating real motifs from their artifacts. Bioinformatics. 817, 30-38.
5. Buhler J., Tompa M. (2001) Finding Motifs Using Random Projections. In RECOMB'01, 6-76. ACM-, Proc.RECOMB'01, Montreal.
6. Clement J., Dutour I., Régnier M. Combinatorial algorithms on approximate words, 2002. submitted.
7. Denise A., Régnier M. (2002) Rare Events on Random Strings, in preparation;http://algo.inria.fr/regnier/index.html.
8. Denise A., Régnier M., Vandenbogaert M. (2001) Assessing statistical significance of overrepresented oligonucleotides. Proc. First Intern. Workshop on Algorithms in Bioinformatics, Aarhus, Denmark, August 2001;LNCS 2149, 85-97.
9. Dembo A., Zeitouni O. (1992) Large Deviations Techniques. Jones and Bartlett, Boston.
10. Guibas L., Odlyzko A.M. (1981) String Overlaps, Pattern Matching and Nontransitive Games. J. of Combinatorial Theory. Series A, 30:183-208.
11. Hamspon S., Kibler D., Baldi P. (2002) Distribution patterns of over-represented k-mers in non-coding yeast dna. Bioinformatics. In press.
12. Klaerr-Blanchard M., Chiapello H., Coward E. (2000) Detecting localized repeats in genomic sequences: A new strategy and its application to *B. subtilis* and *A. thaliana* sequences. Comput. Chem. 24(1):57-70.
13. Lonardi S. (2001) Global detectors of unusual words: design, implementation, and applications to pattern discovery in biosequences. Phdthesis, Purdue University, 45 pages, August, 2001.
14. Marsan L. (2002) Inférence de motifs structures: algorithmes et outils appliques à la détection de sites de fixation dans des séquences génomiques. Phdthesis, University of Marne-la-Vall□ee.
15. Nicodème P. (2001) Fast Approximate Motif Statistics. J. of Computational Biol. 8(3):235-248.
16. Pevzner P.A., Borodovski M., Mironov A. (1989) Linguistic of Nucleotide sequences:The Significance of Deviations from the Mean: Statistical Characteristics and Prediction of the Frequency of Occurrences of Words. J. Biomol. Struct. Dynam. 6:1013-1026.
17. Régnier M. (2000) A Unified Approach to Word Occurrences Probabilities. Discrete Applied Mathematics. 104(1):259-280. Special issue on Computational Biology;preliminary version at RECOMB'98.
18. Régnier M., Szpankowski W. (1997) On Pattern Frequency Occurrences in a Markovian Sequence. Algorithmica. 22(4):631-649. preliminary draft at ISIT'97.
19. Robin S., Schbath S. (2001)Numerical Comparison of Several Approximations on the Word Count Distribution in Random Sequences. J. of Computational Biol. 8(4):349-359.
20. Reinert G., Schbath S., Waterman M. (2000) Probabilistic and Statistical Properties of Words: An Overview. J. of Computational Biol. 7(1):1-46.
21. Vandenbogaert M. (2002) Statistical measurements applied to bacterial rm-systems through genome-scale analysis, and related taxonomic issues. presented at BBC 2002, Belgian Bioinformatics Conference.
22. van Helden J., Andre B., Collado-Vides J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol. 281:827-842.
23. Vandenbogaert M., Makeev V. (2002) Analysis of bacterial RM-systems through genome-scale analysis and related taxonomic issues. submitted to BGRS'02, Novossibirsk.

24. Waterman M. (1995) Introduction to Computational Biology. Chapman and Hall, London.

BIOINFORMATICS ANALYSIS OF PHOH FUNCTION AND REGULATION IN ACTINOBACTERIA

**** Kazakov A.E., Vassieva O., Gelfand M.S., Osterman A., Overbeek R.***

Integrated Genomics inc., Moscow, Russia, e-mail: kazakov@integratedgenomics.ru
*Corresponding author

Key words: *regulatory sites, phosphate regulon, lipid metabolism regulation*

Resume

Motivation: PhoH protein is a putative phosphatase belonging to a phosphate regulon in Escherichia coli.

Results: Positional coupling of this gene in different groups of bacteria indicates its involvement in phospholipid metabolism. The regulatory site discovered upstream of phoH paralog ylaK in Actinobacteria links this gene to a regulon, which includes enzymes of fatty acid beta-oxidation.

Introduction

The function of one of the members of the phosphate regulon in *E. coli*, *phoH,* is still undefined, although PhoH was shown to possess phosphatase activity (Kim et al., 1993). In *Bacillus subtilis*, this gene is located in one locus with the gene encoding diacylglycerol kinase (Kim et al., 1997).

Materials and Methods

Multiple sequence alignment was constructed using the CLUSTALX program (Thompson et al., 1997). Phylogenetic tree was constructed using the PHYLIP package program PROML (maximum likelihood method) (Felsenstein, 1996). Positional analysis of *phoH* orthologs in different groups of bacteria was made using ERGO database (Overbeek et al., 2000). The recognition profiles (positional weight matrices) were constructed using aligned samples of experimentally verified sites. The positional nucleotide weights in these profiles were defined as (Mironov et al., 1999):

$$W(b,k) = \log [N(b,k) + 0.5] - 0.25 \sum_{i=A,C,G,T} \log [N(i,k) + 0.5] \quad (1)$$

where $N(b,k)$ denoted the count of nucleotide b at position k. The score of a L-mers candidate site was calculated as the sum of the respective positional nucleotide weights:

$$Z(b_1 \ldots b_L) = \Sigma_{k=1 \ldots L} W(b_k ,k) \quad (2)$$

The comparative approach to the analysis of transcriptional regulation in bacterial genomes is based on the assumption that sets of genes regulated by orthologous transcription factors are conserved in related genomes. Thus the candidate sites occurring upstream of orthologous genes are true, whereas false positives are scattered at random. Unique members of regulons may be lost, however, use of additional genomes decreases the number of "orphan" regulon members.

Results

Sequences of proteins similar to BS-PhoH and BS-YlaK (close homolog of PhoH) were aligned and a phylogenetic tree was constructed to distinguish orthologs of these two proteins. PhoH orthologs were found in 76 organisms including many Gram-positive and Gram-negative bacteria. BS-YlaK orthologs were found in 41 organisms.

A conserved locus containing *phoH* as well as six another genes was identified. The list of genes found in the *phoH* loci includes:

1. miaB encoding protein involved in methyltiolation of isopentenylated A37 derivatives in the tRNA (in 30 Gram-negative organisms).

2. yqfF encoding possible metal-dependent phosphohydrolase (in 9 organisms, mainly Gram-positive bacteria).

3. yqfG encoding conserved protein with unknown function (in 55 organisms).

4. dgkA encoding diacylglycerol kinase (in 11 organisms, mainly Gram-positive bacteria).

5. ybeX encoding CBS domain-containing protein (in 33 organisms, including Gram-negative bacteria and Actinobacteria).

6. era encoding GTP-binding protein ERA (in 13 organisms, mainly Gram-positive bacteria).

7. lnt encoding apolypoprotein N-acyltransferase (in 24 Gram-negative organisms).

Search for possible regulatory sites upstream of *ylaK* orthologs revealed a conserved 18-bp pseudopalindrome in *Mycobacterium tuberculosis*, *Mycobacterium bovis* and *Thermomonospora fusca*. An iterative signal search procedure was applied to the genomic sequences of *M. tuberculosis* and *T. fusca* using PSI-SITE program. First, a recognition rule was

generated using all three initially identified sites. Second, ten best sites were selected in each genome and used for generation of new organism-specific recognition rules. Third, these recognition rules were applied to appropriate genomes and sites scoring at worst 10% below the highest possible value were selected (Table 1).

All genes identified in *M. tuberculosis* and *M. bovis* seem to be orthologous. Among genes identified in *T. fusca*, only *RTFU00810* and *RTFU00852* have orthologs in *M. tuberculosis* and *M. bovis*. *RMB00348* and *RMT05630* (orthologs of *RTFU00852*) probably are not co-regulated with *ylaK* orthologs.

Some of the identified genes seem to be co-transcribed with other genes. Thus, *RMB00929* and *RMT05592* are probably co-transcribed with *RMB00928* and *RMT04295* respectively, the latter encoding the alpha subunit of the fatty oxidation complex. *RTFU02009* probably forms an operon with *RTFU02062* that encodes short-chain precursor of Acyl-CoA dehydrogenase.

Then, potential regulatory sites were identified in *M. tuberculosis* and *T. fusca* using the same recognition rule with a lower cutoff. Genes having candidate sites upstream of orthologous genes were selected. Six pairs of genes (in addition to the *ylaK* orthologs) were identified (Table 2).

Table 1. Members of the *ylaK*-related regulon with strong candidate sites.

ERGO database name	Alias	Proposed function	Site position	Site score	Site sequence	Orthologs
Mycobacterium bovis						
RMB00929	None	3-ketoacyl-CoA thiolase (EC 2.3.1.16)	-74	4.75	GGTgCCGGTaCgGGaCCT	RMT05592
RMB02442	BS-ylaK	PhoH protein homolog	-87	5.05	aGgACCGGcCCCGGTCCT	RMT04260, RTFU00810
RMB04839	None	Unknown	-106	4.75	GGTAgCGGcaCCGGcCCT	RMT06543
Mycobacterium tuberculosis						
RMT04260	phoH2	PhoH protein homolog	-87	5.05	aGGACCGGCCCCGGTCCT	RMT04260, RTFU00810
RMT05592	fadA	3-ketoacyl-CoA thiolase (EC 2.3.1.16)	-76	4.75	GGTGCCGGTACGGGaCCT	RMB00929
RMT06543	None	Unknown	-104	4.75	GGTAgCGGCACCGGCCCT	RMB04839
Thermomonospora fusca						
RTFU00810	None	PhoH protein homolog	-68 -89	5.29 4.93	GGGGCTGGTCCCGGTCCT GGGGCCGGTCCCGGCCCT	RMB04260
RTFU00852	None	Unknown	-133	4.93	cGGTCCaGCCCCGGTCCT	RMB00348, RMT05630
RTFU01955	None	Unknown	-172	4.93	GGGGtCTGCCCCGGTCCC	
RTFU02009	None	Acyl-CoA-dehydrogenase	-60	4.75	cGGGaCGGCCCtGGTCCT	

Table 2. Members of the *ylaK*-related regulon with conserved candidate sites.

Gene name	Alias	Function	Site position	Site score	Site sequence
RMT04264	Rv1099c	GlpX protein	3	3.27	GGagCtGGTCCgGGTgac
RTFU00821			15	3.55	cGctCCGGTaCCGaTCCc
RMT00211	Rv2205c	Glycerate kinase	-55	3.62	GGggCCGGcaagcGaCtT
RTFU00282			-62	4.19	GtGaCCGGcCCCGcTCCc
RMT01107	proA	Gamma-glutamyl phosphate reductase (GPR) (EC 1.2.1.41)	6	3.35	cGTgCCaGcaCCGtcgCa
RTFU00343			-17	3.43	GGGtgCGGcCCCGcaCgT
RMT01470	suhB	Extragenic suppressor protein SuhB	-139	3.90	GGggCCGGTgCtGGTCaT
RTFU02892			-1	3.63	GtGaCCGtTCCCGaTCCg
RMT06072	fadE34	Acyl-CoA-dehydrogenase	-243	3.60	GGgAgCGcTaCtGGTgtT
RTFU02009			-60	4.75	cGGGaCGGcCCtGGTCCT
RMT03872	recR	Recombination protein RecR	9	3.93	GGgACCcGTCCaGGaCCT
RTFU00514			-66	3.33	aGGGtCctcaCCGGTtCc

Discussion

Positional analysis of *phoH* gene suggests several possibilities of the pathways it can be functionally linked to. It seems likely that *phoH* is functionally related to transformations of diacylglycerol released during biosynthesis of cell wall components. One way of this transformation can imply further phosphorylation by diacylglycerol kinase, whose gene is tightly coupled to *phoH*, and phosphatidate kinase. Phospholipid degradation/turnover also can be a possible functional link

to the *phoH* gene cluster. Glycerol and fatty acids can be released in this process and catabolised as a source of carbon and energy. According to our data, *ylaK* can be a member of a regulon related to fatty-acid beta oxidation or glycerol catabolism in Actinobacteria. Among enzymes likely to be co-regulated with *ylaK* in Actinobacteria, there are components of the fatty acid oxidation complex, acyl-CoA dehydrogenase and glycerate kinase. *glpX*, a member of the glycerol 3-phosphate regulon in *E. coli*, also seems to be co-regulated with *ylaK*. We are unable to define the exact functions of proteins encoded by the *phoH*-related cluster, though we made several general predictions. We hope that our data can be used as a starting point in experimental follow-up.

References

1. Felsenstein J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266, 418-427.
2. Kim S.K., Makino K., Amemura M., Shinagawa H., Nakata A. (1993) Molecular analysis of the phoH gene, belonging to the phosphate regulon in Escherichia coli. J. Bacteriol. 175, 1316-1324.
3. Kim S.A., Woo J.H., Hong S.D., Song B.H. (1997) Isolation of the Bacillus subtilis cdd downstream region and analysis of genetic structure around the cdd vicinity. Mol. Cells. 7, 648-654.
4. Mironov A.A., Koonin E.V., Roytberg M.A., Gelfand M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. Nucl. Acids Res. 27, 2981-2989.
5. Overbeek R., Larsen N., Pusch G.D., D'Souza M., Selkov E.Jr., Kyrpides N., Fonstein M., Maltsev N., Selkov E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucl. Acids Res. 28, 123-125.
6. Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl. Acids Res. 25, 4876-4882.

PARALLEL ALGORITHM FOR SEARCHING REGULATORY SIGNAL IN BACTERIAL GENOME

[1] Lyubetsky V.A., Rubanov L.I.

Institute for Information Transmission Problems, RAS, Moscow, Russia, e-mail: lyubetsk@iitp.ru
[1] Corresponding author.

Key words: *informational genetics, regulatory signal, parallel computing, MPI*

Resume

Motivation: The paper describes a method of transforming existing sequential algorithm of regulatory signal search into a parallel version suitable for contemporary parallel computers supporting MPI protocol. This parallel implementation of the algorithm does not strictly bounded to the number of available processors and features a linear dependence of calculation speed on the number of processors.

Results: The algorithm allows biologists to investigate bigger sets of genomic sequences than they can study earlier. That has been proved in real experiments with the parallel algorithm carried out on teraflops parallel supercomputer at JSC of RAS.

Availability: The software is available on request from the authors.

Introduction

The problem of searching a regulatory signal in a set of assumed regulatory domains (sequences of nucleotides) is well-known. For its solution various sequential algorithms were proposed. In particular, the algorithm described in (Danilova, Gorbunov et al., 2001) was effective for many natural samples. To solve the problem, powerful computing systems sometimes are required that mean parallel computers in practice. The role of such systems in computational genomics is expected to increase because they enable considering tasks and data dimensions that are impossible to analyze by existing sequential architectures in principle. So it is desirable to consider possible ways and methods of transition from consecutive algorithms to their parallel versions using as example some problems important for genomics. The parallelization is known to be nontrivial in many cases. We succeed in the transition for several algorithms, and consider here this subject as applied to the mentioned consecutive algorithm. Specifically, we have developed effective parallel algorithm for searching regulatory signals. Detail results of its application to natural and artificial samples (including comparison with those of consecutive algorithms) are being published in the online magazine *Informational Processes* at http://www.jip.ru.

Difficulties of parallel implementation

Parallelization of the existing algorithm is complicated with two factors which are typical for many algorithms in genomics: cyclic (i.e. consecutive) structure of the algorithm and the model of the shared memory it is implicitly based on. The latter complicates interchanging data in the multiprocessor system. Actually, the majority of modern parallel computers divide random access memory more or less equally among processors interfacing each with other via internal network. The communication is normally accomplished by short messages conforming to e.g. MPI standard. Apparently, this method of data sharing is more time-consuming than direct access to the shared memory. Now we shall discuss the first of the mentioned difficulties.

Our sequential algorithm has a structure outlined in Fig. 1. Here each repetition of the loop seeks for quasioptimal solution of the problem for some fixed way of numbering initial nucleotide sequences. For brevity this way of numbering we call *permutation*, and process of the problem solving will be called *assembling* of the given permutation. Once a permutation has been assembled, the algorithm generates next permutation, and so on. The optimal solution would be the best one found from all possible $n!$ permutations, that is of course unattainable.

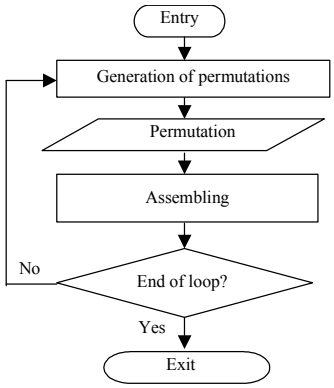


Fig. 1. Source algorithm.

Really we choose a termination condition of the loop so that the solution obtained from checked permutations would be close to absolute optimum in terms of some fixed functional of quality. For that we proposed two principles and two corresponding conditions of the algorithm termination. The first principle takes into account quality of the found signal and relative contribution of each sequence to joint quality: the more valuable site of the signal a sequence gives, the smaller serial number it has in the next permutation. Another principle takes into account completeness of covering the set of sequences by the principal edges of *G* tree which controls consecutive dichotomy of this set in that sequential algorithm.

The main burden of calculations (and, therefore, operating time of the algorithm) fall to the assembling phase. Since this block is located inside a loop (see Fig. 1), one can see the only opportunity to parallelize the algorithm: to perform parallel calculations inside the block. Though possible in principle, the assembling process is hardly scalable: vector-like data structures correspond to consecutive dichotomy of the set of source sequences. So dimension of a vector becomes rigidly connected to number n of sequences, and besides changeable at different levels of *G* tree, i.e. stages of assembling. In addition, time of computation for separate elements of such vector (i.e. sequence pairs) may vary over a wide range. But we need to wait until the slowest component complete prior to process the next level of the tree. Thus, an attempt to rigidly fasten separate stages or similar elements of assembling to processors would not be time effective, nor allow us to use all available processors and balance their workload. We come to necessity of essential changes in the very logic of consecutive algorithm.

Two-dimensional list of permutations and wavelike scheme of computation

When looking at algorithm (Danilova et al., 2001) as an optimization of given functional of quality, one can observe the following analogies. Each permutation is similar to a point in the space of optimum search, and assembling leads to the result similar to calculation of the functional value in this point. The consecutive algorithm is capable to generate points of two kinds corresponding to abovementioned principles. Namely, the second principle of coverage maximization provides new base points to search *global* optimum (we need them as we have no information on geometry of a response surface). And with the first principle we aspire to reach the best *local* solution, starting at the chosen base point.

The general idea of proposed parallelization method is that all available processors of the parallel computer system (except for one root processor) every moment are engaged to local optimization, each one starting at its own base point. All data necessary to apply the first principle are kept in local memory of the processor in charge. The dedicated root branch receives only results of local calculations (recognized signals along with their quality data). On completion of local optimization (this process we call *extension*), released processors receive new base permutations, and so forth. Termination criterion of this algorithm will be exhaustion of entire set of the base points generated with use of the second principle (maximum coverage), provided that all parallel branches finish their computations.

To implement this idea we shall generate not a straight queue of permutations like in consecutive algorithm, and the two-dimensional <*P, Q*> list (see Fig. 2). Its backbone is *P-list* of permutations P_1, P_2,…, P_k composed in the beginning of the algorithm; the length k of this list is determined by a desirable coverage degree of the set of base points. For example, if we demand that each pair of source genomic sequences appears at least once as principal edge at the upper level of *G* tree, i.e. these sequences fall into different halves at the very first dichotomy of entire set, then, obviously, $k = n\,(n\text{-}1)$:2.

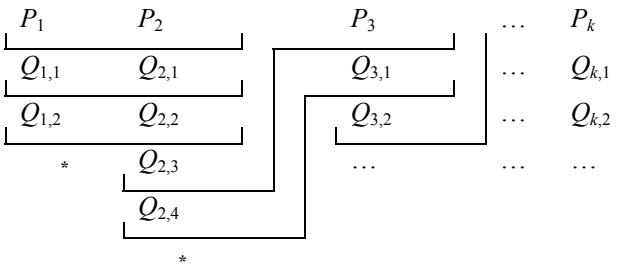


Fig. 2. <*P, Q*> list of permutations and assembling order.

Further, each element of the *P*-list initiates so called *Q-list* corresponding to the process of extension mentioned above. These *Q*-lists are being constructed dynamically during work of the algorithm. Specifically, after the analysis of quality of the signals already found in *i*-th *Q*-list (including its root P_i), the list either proceeds with a next permutation $Q_{i,j}$ or terminates (that is symbolically denoted by asterisk in Fig. 2). In the latter case the processor earlier serving *i*-th *Q*-list is released and takes the next unprocessed element of *P*-list, extending from it a new *Q*-list.

It is easy to see that in such two-dimensional set of permutations parallel processing is conducted like a computing wave propagating from the left top corner of the structure in a downward direction and to the right until whole <*P, Q*> list is processed. For the sake of simplicity Fig. 2 shows successive positions of the wave front for two processors working in parallel and with the assumption of constant duration of assembling any permutation, however one can easy imagine this algorithm in general case.

From the parallelization efficiency point of view the described scheme has many advantages from which we shall note three the most important: (1) it allows to put into operation at once the multitude of processors (at least as many as is the length *k* of *P*-list); (2) all secondary branches work on the same algorithm assembling their given permutations irrespectively of other branches; and (3) low-intensive data exchange between each of secondary branches and root branch of the algorithm: a permutation itself towards the secondary branch (*n* values), and found signal along with a quality of each its word (2*n* values) towards the root branch.

These advantages allowed us to create the parallel algorithm independent of any specific parallel computer and number of available processors; the only prerequisite is that the computer complex supports MPI standard (at least the first edition). We also avoid using shared memory due to move criteria checking and permutation generation to the single root branch. The only common data (a matrix of pair-wise proximities of all words from all sequences) is unchanged during work of the algorithm, so may be calculated and stored simultaneously in all parallel branches.

Implementation and Results

The described parallel algorithm of regulatory signal search has been implemented as 32 bit console application written in ANSI C. It may be compiled by e.g. gcc, pgcc, Borland C++ 5.02 compilers without any modification, and works in Windows 9x/NT4/ME/2k/ XP and Linux environments. Debugging and initial testing of the program were carried out on PCs with use of WMPI 1.54 software by Critical Software. Real experiments are being performed now on MBC1000M supercomputer in the Joint Supercomputer Centre of RAS et al. Results obtained for artificial and real examples show high efficiency of parallelization in the algorithm: busy time ratio of secondary processors equal 94-96%. Experiments also confirm theoretically predicted linear dependences of assembling time on number of source sequences (Fig. 4) and computation speed on number of available processors (Fig. 5). We could not recognize any tendency of deceleration this linear growth of performance.

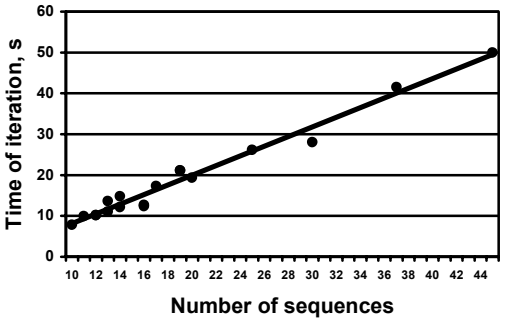


Fig. 4. Assembling time as a function of number of sequences.

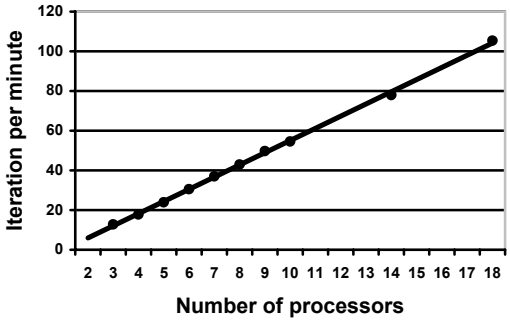


Fig. 5. Linear growth of parallel algorithm performance (*n*=14, *m*=200, *l*=20).

Reference

1. Danilova L.V., Gorbunov K.Yu., Gelfand M.S., Lyubetsky V.A. (2001) Algorithm for extraction of regulatory signals in DNA sequences (2). Mol. Biol. 35, 6, 987-995.

A GENETIC ALGORITHM FOR IDENTIFICATION OF REGULATORY SIGNALS

**** Stavrovskaya E.D., Mironov A.A.***

State Scientific Center GosNIIGenetika, 113545, Moscow, Russia
Integrated Genomics, POBox 348, 117333, Moscow, Russia
e-mail: esta191@fromru.com
*Corresponding author

Key words: *regulation signal, genetic algorithms*

Resume

There exist numerous algorithms for identification of regulatory signals in unaligned DNA fragments. Here we present a genetic algorithm for signal identification, describe its implementation and testing on simulated data. It is the first application of genetic algorithms in this area.

Introduction

The existing algorithms identification of regulatory sites can be divided into optimization and combinatorial ones. The former class includes greedy algorithms, expectation-maximization, DMS, MEME; and also stochastic algorithms: simulated annealing and the Gibbs sampler. The combinatorial algorithms are ConsInd and MatInd, WORDUP, CONSENSUS, WINNOWER pattern, graphs, and numerous other algorithms.

The genetic algorithm suggested here can be considered as the optimization one. We believe that it works faster then other stochastic algorithms of comparable recognition power.

Description of Algorithm

The following abstract concepts will be used (to avoid confusion with standard biological terms, they will be italicized): *genome*, *gene*, *allele*, *quality of genome*, *population*, *crossing*, *selection* and *mutation*. Consider set of DNA fragments. Each fragment corresponds to a *gene*, and each position, specifying a candidate site, is an *allele*. Thus a set of candidate sites, one in each fragment, generates a set of *alleles*, that is, a *genome*. Each *genome* is characterized by its *quality,* defined as the information content of the respective set of sites. At each step the algorithm processes *population*, that is a set of *genomes*, and performs the following operations:

Crossing: select at random a pair of *genomes* and generate new a one:

Genome1: $S_1, S_2, \ldots S_k S_{k+1} \ldots S_n$
Genome2: $T_1, T_2, \ldots T_k T_{k+1} \ldots T_n$

New genome: $S_1, S_2, \ldots S_k T_{k+1} \ldots T_n$

Position *k* of the cut is given by the random uniform distribution.

Selection: Delete the *genome* with the lowest quality.

Mutation: Select a random *gene* in a random *genome* and change the current *allele* to a random one. It is equivalent to selecting a random site in the corresponding fragment.

These steps are iterated for some fixed time.

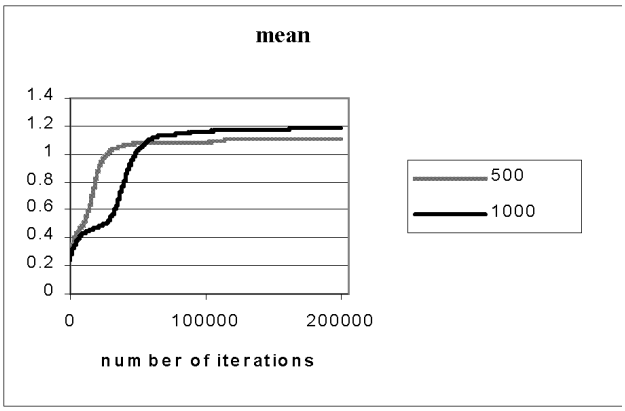
Results and Discussion

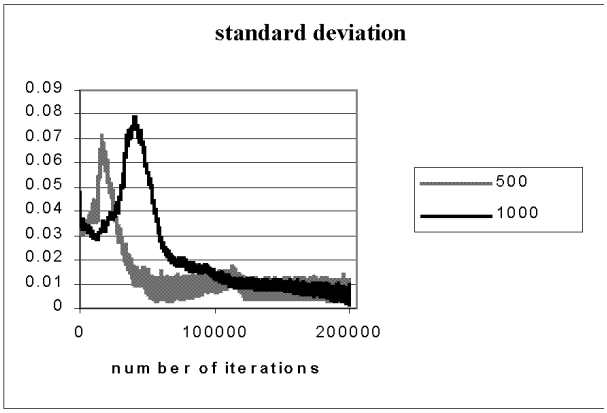
Each test file contained ten fragments of length 200. The signal was a fixed word of twenty nucleotides. The sites were modeled by introducing some mismatches into the signal word and then inserting the resulting word into the sequence fragments at random positions. The number of mismatches varied from one to seven. To model corrupted samples, some fragments did not contain the signal.

The results are presented in the Table.

Table. Probability of correct signal identification.

	Number of fragments with a signal						
Number of mismatches		10	9	8	7	6	5
	0	0.99	0.99	0.99	0.97	0.80	0.80
	1	0.99	0.96	0.90	1.0	0.90	0.70
	2	1.0	0.99	0.93	0.97	0.90	0.40
	3	0.84	0.74	0.81	0.81	0.90	0.66
	4	0.92	0.77	0.73	0.79	0.68	0.40
	5	0.71	0.53	0.44	0.44	0.43	0.54
	6	0.56	0.70	0.20	0.44	0.55	0.44
	7	0.17	0.27	0.23	0.36	0.22	0.0





The dependency of the mean quality and its standard deviation on the number of iterations for different population sizes (500 and 1000) are presented in the diagrams. It can be seen that in larger populations the mean is larger, the standard deviation is smaller, although the stationary values are reached later. The iteration process should be stopped when the mean becomes practically constant and standard deviation is close to zero.

If two signals are introduced simultaneously, only one of them is found (an arbitrary one).

In these tests the best results were obtained with the *population* of 1000 *genomes* at 200000 iterations.

References

1. Gelfand M.S. 1995. J. Comput. Biol. 2, 87-115.
2. Frech K., Quandt K., Werner T. 1997. Comput. Appl. Biosci. 13, 89-97.
3. Duret L., Bucher P. 1997. Curr. Opin. Struct. Biol. 7, 399-406.
4. Fickett J.W., Wasserman W.W. 2000. Curr. Opin. Biotechnol. 11, 19-24.

SITEPROB: YET ANOTHER ALGORITHM TO FIND REGULATORY SIGNALS IN NUCLEOTIDE SEQUENCES

**** Vinogradov D.V., Mironov A.A.***

Integrated Genomics, P.O. Box 348, 117333, Moscow, Russia
*Corresponding author, e-mail: darkfire@mccme.ru

Key words: *regulatory signal, probability theory, prefix tree, bacterial genomes, signal recognition*

Resume

Motivation: The recognition of regulatory signals in DNA sequences problem has great practical and theoretical value. Its main practical application is processing fully sequenced bacterial genomes, which completed would allow comparative analysis of regulatory interactions and mass application of DNA chips to the analysis of gene expression.

Results: New pattern-driven algorithm, based on statistical estimates of signal's importance, has been developed and implemented. Accomplished tests proved this method's high efficiency.

Availability: Algorithm was realized using Java™ language, making this implementation independent from specific platform. All sources are available from authors per e-mail request.

Introduction

Recognition of regulatory signals in DNA sequences is well known to be one of the oldest and most important, but still unsolved problems of computational biology. Its current popularity owes much to a huge number of completely sequenced bacterial genomes. This allows for the comparative analysis of regulatory interactions, as well as mass application of DNA chips to the analysis of gene expression.

There are many different algorithms, but still none can be considered as acceptable in all situations. These algorithms can be roughly divided into two groups. The profile-driven algorithms aim to maximize some function that measures the overall similarity of identified sites (e.g. MEME [1], DMS [2], Gibbs sampler [3], greedy algorithms [4]). The pattern-driven algorithms seek the word existing in every sequence from the sample with minimal deviation (e.g. CONSENSUS [5], WINNOWER [6], WORDUP [7], suffix trees [8]).

Here we describe an algorithm belonging to the latter group, but with some improvements to this approach. The distinctive feature of our algorithm is that it can process samples than contain strong sites in a small number of fragments, and relatively weak sites in the remaining fragments. This is achieved by maintaining a proper balance between the site scores and the number of fragments containing sites with such scores.

Methods and Algorithms

We use the following definitions: *a palindrome* is a nucleotide sequence which consist of two complementary words of equal length and a spacer between them. These words will be called *boxes*. *Alignment* is matching a palindrome to a long sequence.

Consider palindromes with box length *l* and the spacer length not exceeding *G*. The SiteProb algorithm scans all $4^l G$ such palindromes and calculates significance *P* of over-representation of each palindrome in the sample.

Consider uniform Bernoulli sequences with probability of each symbol being $P_a = 0.25$. The probability that a palindrome with a fixed spacer length will align at a fixed position with at most K_{err} mismatches is

$$P_{align} = \sum_{T-K_{err}}^{T} C_T^t p_a^t q_a^{T-t},$$

where $q_a = 1 - p_a$, C_T^t is the binomial coefficient, $T = 2l$ is total length of both boxes, $t = T - K_{err}$. Assuming independence of alignments (which is a reasonable approximation in our case), we obtain, that the probability of no alignments with less than *K* errors is $Q_{align}^{L-l} = (1 - P_{align})^{L-l}$, where *L* is the length of the nucleotide sequence fragment. Thus the probability of at least one good alignment for a given palindrome is $P_{word} = 1 - Q_{align}^{L-l}$. Finally, the probability of finding a good alignment in *m* out of *n* fragments in a sample is

$$C_n^m P_{word}^m (1 - P_{word})^{n-m},$$

and significance of observing at least *m* good alignments is

$$P = \sum_{i=m}^{n} C_n^i P_{word}^i (1 - P_{word})^{n-i} .$$

We use this value as a criterion for selection of palindromes that can be considered non-random; good alignments of these palindromes to the fragments from the sample are the predicted sites.

To decrease the running time of the algorithm, we consider only best alignments of each palindrome in each sequence, and apply the standard prefix tree technique.

Results and Discussion

The algorithm was tested on two samples from the *E. coli* genome. The sequence fragments were taken from regulatory regions and contained known sites. The "Purine" sample initially consisted of 19 PurR-binding sites, and the "Arginine" sample contained 11 ArgR-binding sites. Then we masked one by one the sites from the samples, but retained the corresponding sequence fragments, thus imitating contamination of the sample by spurious sequences. Each time the strongest site of the sample was deleted. The results are given in Tables 1 and 2.

Table 1. (Purine).

Number of known sites in sample (K)	19	18	17	16	15	14	13	12	11	10
Sites predicted by SiteProb (P)	14	13	12	11	10	14	13	8	7	9
Sensitivity (P/K)	73,68%	72,22%	70,59%	68,75%	66,67%	100%	100%	66,67%	63,64%	90,00%
Number of known sites in sample	9	8	7	6*	5*	4*	3			
Sites predicted by SiteProb	9	8	7	6	5	3	0			
Sensitivity (P/K)	100%	100%	100%	100%	100%	75,00%	0,00%			

Table 2. (Arginine).

Number of known sites in sample (K)	9	8	7	6	5	4	3
Sites predicted by SiteProb (P)	8	7	0	0	0	0	0
Sensitivity (P/K)	88,89%	87,50%	0,00%	0,00%	0,00%	0,00%	0,00%

Table 3. (Found signals consensuses).

Arginine	tgaata 2 tattca
Purine	Gcaaac 0 gtttgc

This results demonstrate algorithm's resistance to presence of sequence fragments containing no signal. That can be explained with one of the advantages of this algorithm – automatic choose of threshold. The ratio of fragments with signals to all fragments in the sample was 21% in Purine case, and 88% in Arginine case.

Finally, we want to note that for the moment program can search only for palindrome signals. This assuming was made, because more than 80% of known signals from bacterial genomes have structure of a palindrome. Among the rest most signals have tandem repeat structure, which is caused by cooperative way of regulatory protein binding with DNA. And existing program can be easily modified to find signals in this form.

Acknowledgements

We are grateful to Luda Danilova who shared with us the test samples and to Mikhail Gelfand, for his support in preparing this text.

References

1. Bailey T.L., Elkan C. (1994) Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. Proc. of the Second Int. Conf. on Intelligent Systems for Molecular Biology (ISMB'94), 28-36.
2. Hu Y.J., Sandmeyer S., McLaughlin C., Kibler D. (2000) Combinatorial motif analysis and hypothesis generation on a genomic scale. Bioinformatics. 16(3):222-232.
3. Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., Wootton J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 262(5131):208-214.

* When the number of sites in sample dropped below 7, algorithm returned the correct answer, but it was not the first in the list of results.

4. Stormo G.D., Hartzell G.W.III. (1989) Identifying protein-binding sites from unaligned DNA fragments. Proc. of the Natl Acad. of Sci. USA. 86:1183-1187.
5. Hertz G.Z., Stormo G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 15:563-577.
6. Pevzner P.A., Sze S.-H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. Proc. of 8th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB'2000), 269-278.
7. Pesole G., Prunella N., Liuni S., Attimonelli M., Saccone C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucl. Acids Res. 20(11):2871-2875.
8. Brazma A., Jonassen I., Vilo J., Ukkonen E. (1998) Predicting gene regulatory elements in silico on a genomic scale. Genome Res. 8(11):1202-1215.

YET ANOTHER DIGGING FOR DNA MOTIFS GIBBS SAMPLER

**** Favorov A.V., Gelfand M.S., Mironov A.A., Makeev V.J.***

State Scientific Centre "GosNIIGenetica", Moscow, 113545, Russia, e-mail: favorov@sensi.org
*Corresponding author

Key words: *DNA motifs detection, motif length determination, statistical model, weight matrix, Markov Chain Monte-Carlo, Gibbs sampler, Kullbak entropy distance*

Resume

Motivation: The problem of motif identification in a set of unaligned DNA sequences arises when a set of coregulated sequences is analysed to determine the sites, which are responsible for the common regulation. Heuristic algorithms, particularly, the Gibbs Monte-Carlo Markov chain (MCMC) sampling, have become the standard approach to this problem. The classic tool of this kind suggested by Lawrence (1993) does not provide an ultimate solution for a number of problems, e.g. the motif length determination. Also, it is reasonable to allow for a possibility of motif absence in a sequence, e.g. due to an artefact.

Results: We have developed and implemented an MCMC motif identification algorithm that is based on the same approach as the Lawrence one is. It is able to handle the situation of site absence in some sequences from the sample. It optimises the motif length "on fly". Rigorous analysis of the entropy distances between prior and posterior spatial and motif model distributions yields the identification of weak signals of unknown length. The program was tested on several samples of bacterial regulatory sites and proved its ability to detect weak signals without any prior assumptions about their length.

Availability: FreeBSD and Windows NT console executable files and a WWW interface for current version of the program are referred from http://favorov.hole.ru/gibbslfm.

Introduction

The methods of probabilistic data analysis, e.g. the Bayesian inference optimisation (Sivia, 1996) and the Monte-Carlo Markov chains (Besag et al., 1998), are adequate tools for the biopolymer data analysis in the era of biotechnology revolution that has resulted in huge amounts of raw genetic sequence data (e.g., see Liu, Lawrence, 1999).

Detection of common motifs (multiple local alignment) is a traditional area of the genetic data analysis. Due to the assumption that different, but functionally identical DNA regions probably have resembling nucleotide sequences, the motif recognition is a powerful tool for data interpretation in a wide range of DNA structure, function and evolution studies. Current algorithms devoted to the task are amazingly variable, due not only to the diversity of computational approaches, but mainly because of the intuitiveness of the motif (common pattern) concept. One of its possible formalisations is to treat a motif as a common probabilistic model of sequence generation. Such a model should characterise a set of sites in the data and must differ from the model of the remaining material (background). A very natural model for motifs of equal length is a nonuniform Bernoulli model that is defined by a position-base probability matrix (e.g., Hertz, Stormo, 1999; Bailey, Elkan, 1994, Lawrence et al., 1993).

Model and Algorithm

Basically, the approach of this study is a modification of the (Lawrence et al., 1993) methodology. We define a motif as a position-dependent Bernoulli model generating sequence fragments of fixed length. Each fragment that fits the model is referred to as site, and all non-site genetic material is the background, which is described by a simple, position-independent Bernoulli model. In other words, if we have the motif weight matrix q_i^r and the background symbol probabilities f^r, $r \in \{A,T,G,C\}$, $i = 1 \ldots L$ (r and i are the base and its position in the motif, respectively, L is the motif length), the likelihoods of sequence $r_0 r_1 \ldots r_{m-1}$ given that the sequence contains a motif site starting at p or given that it contains no site are:

$$P\left(r_0 r_1 \ldots r_p \ldots r_{p+L-1} \ldots r_{m-1} \mid \text{a site at } p\right) = \prod_{i=0}^{p-1} f^{r_i} \prod_{i=p}^{p+L-1} q_i^{r_i} \prod_{i=p+L}^{m-1} f^{r_i} \quad (1)$$

$$P\left(r_0 r_1 \ldots r_{l-i} \mid \text{no site}\right) = \prod_{i=0}^{m-1} f^{r_i} \quad (2)$$

When we look for a motif, we suppose that there is no more than one site per sequence. However, unlike (Lawrence et al., 1993), our algorithm allows for the possibility that there is no site in a sequence on every step.

The core of the algorithm (analogous to Lawrence et al., 1993) is a sequence-by-sequence cycle of site updates by the Gibbs algorithm (see Besag et al., 1998 for details about MCMC methods, the Gibbs sampling particularly). To update a current sequence site position, we count bases for all motif positions and for the background from all the remaining sequences and convert the obtained quantities into probabilities q_i^r and f^r by the standard formulae involving pseudocounts. The Bayes formula, which combines likelihoods counted by formulae (1) and (2), pre-given motif absence prior and flat prior for all site positions, gives us the posterior distribution for all site positions and site absence in our sequence. We sample anew the site position in the current sequence (or the site absence) from the distribution.

Although the Gibbs update algorithm is less prone to be locked in local maxima than deterministic algorithms like MEME (see Bailey, Elkan, 1994), there is one special case. If a set of sites found by the sampler is actually a shifted set of stronger motif sites, then the sampler will hardly understand it. To work around this problem (which stems from the nature of the data), we use (following Lawrence et al., 1993) the deterministic adjustment steps, when the site sets are shaken like a solid entity to obtain the best information gain (see below). In addition to (Lawrence et al., 1993), we use the same "shaking" to adjust the motif length in order to avoid a very slow procedure of starting the sampler with different *L*'s and then finding the best one among them.

The target function we are maximizing is the information gain per degree of freedom. It is combined from two components, the motif and spatial ones (Lawrence et al., 1993). The former is the Kullbak entropy distance between the motif weight matrix and the vector of the background symbols probabilities. The latter is the entropy distance between the spatial distribution of the site presence probability in the present motif model and the *a-priori* flat site probability distribution. The spatial component is necessary to compare the motif sets with different motif length; otherwise the motif information gain component is sufficient. We replaced the spatial component given in (Lawrence et al., 1993), which is just the entropy of model-described spatial distribution, by the Kullbak distance, and that increased the sampler sensitivity for weak motifs of unknown length.

Implementation and Results

The program is implemented in gnu c++ and compiled for FreeBSD UNIX and for Windows NT console. It has an option to consider both direct and complementary sequences on every step. Also, it can be switched to look for palindrome motifs.

Table. A typical result. 33 sites are found in 33 sequences of the CRP sample.

//motif length is 20			
Ec|aldB	16[1]	7.556[2]	agcgaggtaatcatcatttc
Ec|ansB	78	12.62	tgttacctgcctctaacttt
Ec|araC	73	13.9	tgtgacgccgtgcaaataat
Ec|cdd	74	17.02	tgcgatgcgtcgcgcatttt
Ec|crp	166	8.63	tgcaaaggacgtcacattac
Ec|cyaA	30	12.21	tgttaaattgatcacgtttt
Ec|cytR	56	12.57	tgcgaggcggatcgaaaaat
Ec|dadA	86	14.91	tgtgagccagctcaccataa
Ec|deoC	106	12.32	tgtgatgtgtatcgaagtgt
Ec|fur	45	11.84	tgtaagctgtgccacgtttt
Ec|galE	134	10.44	tgtcacacttttcgcatctt
Ec|glpE	108	12.49	agtgatatgtataacattat
Ec|glpF	61	11.95	tatgacgaggcacacacatt
Ec|glpT	174	17.06	tgtgcggcaattcacattta
Ec|ilvB	40	12.43	cgtgatcaacccctcaattt
Ec|lacZ	93	13.41	tgtgagttagctcactcatt
Ec|malE	42	12.67	tgtgcgcatctccacattac
Ec|malK	77	10.9	tgtggagatgcgcacataaa
Ec|malT	60	13.78	tgtgacacagtgcaaattca
Ec|melR	129	9.231	cgtgctcccactcgcagtca
Ec|mtlA	73	18.13	tgtgacactactcacattta
Ec|nagE	60	12.89	tgcgatacgaattaaatttt
Ec|nupG	84	15.76	tgttatccacatcacaattt
Ec|ompA	15	9.8	cctgacggagttcacacttg
Ec|ompR	20	9.354	cgtgatcatatcaacagaat
Ec|ppiA	99	11.95	tgtgatctgtttaaatgttt
Ec|ptsH	15	15.33	tgtggcctgcttcaaacttt
Ec|rhaB	76	12.48	tgtgaacatcatcacgttca
Ec|srlA_1	109	14.33	tgcgatcaaaataacacttt
Ec|tdcA	121	13.43	tgtgagtggtcgcacatatc
Ec|tnaL	108	18.41	tgtgattcgattcacattta
Ec|tsx	74	12.72	tgtgaaacgaaacatatttt
Ec|uxuA	16	13.39	tgtgatgtggttaaccaatt

[1] The site position.
[2] The site score of motif model fit.

It was tested on compilations of bacterial regulatory sequences and proved its ability to detect weak signals of unknown length in all three work modes, i.e., single DNA strand, double strand and palindromic signals. The table 1 is the program output for a sample compilation of 33 sequences, of 200 bp in length, containing CRP binding sites. We did not specify any motif length. The runtime was 15 seconds on 800MHz Thunderbird CPU.

Discussion

The Gibbs sampling is a traditional technique, which is widely used in the analysis of biopolymer motifs, hence any enhancement of its basic method (Lawrence et al., 1993) is certainly helpful for applications (e.g. Thijs et al., 2001).

The present algorithm is able to detect weak motifs without any prior information about their length due to improvement of the target function and more flexible management of the motif length compared to the initial version of the algorithm (Lawrence et al., 1993) and its current modifications (Thijs et al., 2001 and current version of sampler by Lawrence at http://bayesweb.wadsworth.org/gibbs/gibbs.html).

Unlike (Thijs et al., 2001), we do not process multiple sites on the sampling stage. Still, our sampler treats the site absence in a current sequence as a possibility on every Gibbs step, which is sufficient to detect spurious sequences. To find all remaining sites belonging to the motif, we can scan all data with the motif profile, collecting all sites that fit the profile better then the worst site forming the motif. We do not try to sample simultaneously a set of motif models (like the current version of sampler by Lawrence), for we suppose that "mask the present result and restart the search" tactics (like Thijs et al., 2001) is sufficient. On the other hand, these algorithm simplifications considerably reduce the runtime.

Currently, the program can look for general or for palindrome motifs. Still, the project architecture permits to involve other motif models, e.g. gapped, repeats, etc.

Acknowledgements

This research is partially supported by grants from the Howard Hughes Medical Institute (55000309), INTAS (99-1476) and the Russian Fund of Basic Research (00-15-99362, 02-04-49111). We are grateful to L.Danilova for the assistance with the data.

References

1. Bailey T.L., Elkan C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proc. of the 2nd Intern. Conf. on Intelligent Systems for Molecular Biology (ISMB), 28-36.
2. Besag J., Green P., Higdon D., Mengersen K. (1996) Bayesian computation and Stochastic Sytems. Statistical Sci. 10, 1, 3-66.
3. Hertz G.Z., Stormo G.D. (1999) Identifying DNA and Protein Patterns with Statistically Significant Alignments of Multiple Sequences. Bioiformatics. 1999, 15, 7/8, 563-577.
4. Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., Wootton J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 262, 208-214.
5. Liu J.S., Lawrence C.E. (1999) Bayesian inference on biopolymer models. Bioinformatics. 15, 38-52.
6. Sivia D.S. (1996) Data Analysis. A Bayesian tutorial. Clarendon Press, Oxford.
7. Thijs G., Marchall K., Lescot M., Rombauts S., De Moor B., Rouze P., Moreau Y. (2001) A Gibbs Sampling method to detect over-represented motifs in the upstream regions of co-expressed genes. Recomb. 2001.

HNF1-ALPHA BINDING SITES IN A COMPUTER SEARCH

[3] ***Lockwood C.R., Frayling T.M.***

Department of Diabetes and Vascular Medicine, University of Exeter, United Kingdom, e-mail: C.R.Lockwood@ex.ac.uk

Key words: *HNF1, transcription factor, binding site, knockout mouse, expression data*

Resume

Motivation: We wished to evaluate a new method of finding transcription factor binding sites in silico. This method used a novel combination of expression data from a TCF1 knockout mice (TCF1 codes for the transcription factor HNF1-alpha), and human and mouse genome sequences. The search was of 53 genes differentially expressed between wild type and TCF1 null mice. Each site was to have a p-value indicating its reliability.

Results: We identified 11 genes as candidates for being directly regulated by HNF1-alpha. All were downregulated in the knockout mouse, although half the genes searched were upregulated. 5 of the genes identified had experimental evidence for an HNF1-alpha binding site. Another 6 genes had unexpectedly high p-values, raising the possibility that HNF1-alpha sites had been suppressed from these genes. In conclusion, expression data from transgenic animals lacking a transcription factor can help identify binding sites.

Availability: Freely available from {{http://BindGene.ex.ac.uk}}.

Introduction

There are various methods for finding transcription factor binding sites in silico, but the false-match problem is serious.

One approach that has not seem to have been tried before is to assist the in-silico search using data from an animal missing the gene for a transcription factor. Shih et al. (2001) compared normal mice with TCF1 knock-outs (TCF1 is the gene for the transcription factor HNF1-alpha). They identified many genes whose expression levels differed (in the liver). It is likely that some of these genes will have an HNF1-alpha binding site, but others will not, as there are clearly ways in which the expression level of a gene could be indirectly affected by HNF1-alpha. We aimed to use an in-silico search to identify a subset of these genes that are likely to be directly regulated by HNF1-alpha.

HNF1-alpha belongs to a transcription factor network of medical interest. Many cases of Maturity Onset Diabetes of the Young (MODY) are caused by a mutation in a transcription factor, most commonly HNF1-alpha (see Frayling et al., 2001).

Because of the false-match problem, we thought each site identified should have a guide to its reliability, in the form of a p-value, as commonly found in statistical analysis.

Methods and Algorithms

Binding sites were scored using the position-weight matrix (PWM) method of Bucher (1990) and Tsunoda and Takagi (1999), with a smoothing parameter of 0.01. The PWM used, which represented the HNF1-alpha binding sequence, was essentially the "base frequency" matrix of Tronche et al. (1997), but not using their technique of replacing 0 by -99. An example of a best possible match is GGTTAATAATTACCA.

p-values were estimated using a development of a method described in Tronche et al. (1997). A "shuffled matrix" is generated by taking the PWM representing HNF1-alpha, and swapping the rows at random. Effectively, the shuffled matrix represents a fictional binding site, and any match to it can be regarded at a false match. A shuffled matrix will be identical to the real matrix in several properties (length, information content, AT:GC ratio); it is therefore hoped that the false-match rate will be similar.

Whereas Tronche et al. (1997) used only one such matrix, we used 1000 shuffled matrices. For a given gene, we scanned the upstream human sequence and mouse sequence with the *same* shuffled matrix, then averaged the best human score and the best mouse score. It was determined whether this exceeded the corresponding average obtained using the real matrix. The proportion of shuffled matrices that did so was regarded as the overall p-value.

For each gene, the search region was defined by the start of the sequence in a GenBank entry representing that gene. The region was 2000b upstream to 200b downstream.

Plausible HNF1 sites were checked, manually, by aligning the human and mouse sequences (using BLAST2 (Tatusova, Madden, 1999) {{http://www.ncbi.nlm.nih.gov/gorf/bl2.html}})

[3] Corresponding author.

Implementation and Results

A web program was written ({{http://BindGene.ex.ac.uk}}) that, given a GenBank accession number, would automatically retrieve the GenBank sequence, the upstream sequence (from the genome databases), and search it for HNF1 binding sites. 53 of the Shih et al. (2001) genes were searched.

5 of these genes had very low p-values ($p<=0.002$) and on this basis alone were thought very likely to be regulated by HNF1. The low p-values were due to the binding sites being very good matches that, it was calculated, were most unlikely to occur by chance.

Another 11 genes, although they had possible HNF1 sites, had p-values that were not very low (between 0.009 and 0.012). For each of these genes, it was calculated, there was a considerable risk of a false match. So these were tested further, by aligning the human and mouse upstream sequences. Possible HNF1 sites were retained only if they were aligned against a possible HNF1 site in the other species. By this means 5 genes were rejected, and 6 genes accepted as more candidates for being regulated by HNF1.

By these means the following "Candidate List" was obtained, of genes likely to be directly regulated by HNF1-alpha: LIPC, CRP, F13B, PRODH2, HSD17B2, SCL7A9, SLC16A7, PAH, SLC12A1, PLG, FABP1.

To assess the reliability of this list, two sources of evidence were used.

One source was experimental evidence for direct regulation, as reported in other papers. This was found for 5 of the genes listed (LIPC, CRP, PAH, PLG, FABP1).

The other source was the direction of the change of expression (when comparing wild-type mice with TCF1 knock-outs). All the 11 genes in the "Candidate List" were downregulated in the knock-out mouse. This contrasts with about half the genes searched being upregulated. As the direction of change was not used to select genes, this suggests the process was successful at only selecting genes directly regulated by HNF1. The result is consistent with HNF1 being an activator and not a repressor.

There were also an unexpected number of genes with very high p values (in this case, p-values for human or mouse sequence rather than an overall value). For six genes, one of these p-values was $>=0.99$ (3 based on mouse sequence and 3 based on human sequence). This is far more than one would expect by chance. Strikingly, all six genes were upregulated in the knock-out mouse.

So, in one species, these six genes do not contain HNF1-alpha sites in their upstream regions, not even the false sites expected to occur by chance - "HNF1-alpha deserts". But nothing was found to suggest that these "deserts" are conserved across species.

Discussion

Using expression data from a TCF1-knockout mouse, and human and mouse genome data, we have identified 11 genes likely to be directly regulated by HNF1. Two sources of evidence suggest that this list contains few false matches: first, experimental evidence for HNF1 sites for 5 of the genes; second, the fact that all these genes were downregulated in the knockout animal. This suggests the method has been successful.

The method could be applied to other transcription factors, if a knock-out animal was available, though it is not clear if the method would work with a transcription factor less specific than HNF1-alpha. It is not clear whether the method could be extended to find "composite sites" (which bind $>=2$ interacting transcription factors); but it is interesting that a mouse has already been produced with mutations in three transcription factor genes (coding HNF1-alpha, HNF3-beta, and PDX1) (Shih et al., 2002).

Each possible site was given a p-value. Compared with simply giving a site a score, the p-value is more useful, since it is more obvious if the match could easily have been obtained by chance. The p-value is a characteristic, not of the site alone, but of the site in relation to the region searched. (Increasing the size of the region searched increases the chance of a false match, and thus decreases the p-value).

Using a human-mouse alignment was useful, since it enabled us to select 6 genes whose p-values were too high to enable us to select them by p-value alone.

Six genes appeared to have "HNF1-alpha deserts" in their upstream regions, which did not contain even the weak HNF1-alpha sites expected to occur by chance. Perhaps these regions are exceptions to the idea of Tronche et al. (1997), that false sites appear throughout the genome and are not selected against. One explanation may be that these genes need to downregulate at the same time, and in the same cells, as HNF1-alpha is causing its target genes to upregulate. This may cause counter-selection against HNF1-alpha binding sites, but perhaps only in a limited group of promoters. However, it is not clear that these "deserts" are conserved between species, and it is possible they may be artefacts of the data analysis. If the "shuffled matrix" method overestimated the false-match rate, that would cause more than 1% of genes to have $p>0.99$.

Acknowledgements

We thank the EPSRC and Diabetes UK for funding. Genome data came from the Human and Mouse sequencing consortiums, accessed via UCSC (Santa Cruz) and Ensembl.

References

1. Bucher P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. 212: 563-578.
2. Frayling T.M., Evans J.C., Bulman M.P., Pearson E., Allen L., Owen K., Bingham C., Hannemann M., Shepherd M. et al. 2001. Beta-cell genes and diabetes: molecular and clinical characterization of mutations in transcription factors. Diabetes. 50: S94-100.
3. Shih D.Q., Bussen M., Sehayek E., Ananthanarayanan M., Shneider B.L., Suchy F.J., Shefer S., Bollileni J.S., Gonzalez F.J., Breslow J.L. et al. 2001. Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism. Nat. Genet. 27: 375-382.
4. Shih D.Q., Heimesaat M., Kuwajima S., Stein R., Wright C.V.E., Stoffel M. 2002. Profound defects in pancreatic beta-cell function in mice with combined heterozygous mutations in Pdx-1, Hnf-1alpha, and HNF-3beta. Proc. Natl Acad. Sci. USA. 99: 3818-3823.
5. Tatusova T.A., Madden T.L. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. 174: 247-250.
6. Tronche F., Ringeisen F., Blumenfeld M., Yaniv M., Pontoglio M. 1997. Analysis of the distribution of binding sites for a tissue specific transcription factor in the vertebrate genome. J. of Mol. Biol. 266: 231-245.
7. Tsunoda T., Takagi T. 1999. Estimating transcription factor bindability on DNA. Bioinformatics. 15: 622-630.

ANALYSYS TOOL FOR FINDING TRANSCRIPTION REGULATORY ELEMENTS, USING TRANSCRIPTION FACTOR DATA BASE (TFDB)

**** Mizushima H., Ichikawa H., Ohki M.***

Center for Medical Genomics and Cancer Genomics Division, National Cancer Center Research Institute, Tokyo, Japan
e-mail: hmizushi@ncc.go.jp
*Corresponding author

Key words: *gene regulation, transcription factor, binding site , gene comparison, harr plot, transcription factor database*

Resume

Motivation: As the human genome sequence has been almost sequenced, it is a key issue to analyze the transcription regulation system of the genome information. Regulation mechanisms of some genes has been analyzed by molecular biologists using deletion/mutation assay using reporter gene, however, there is little knowledge of the regulation system in genome wide range, and no consensus algorithm has been developed for unknown gene regulation. For this purpose, we have been maintaining Transcription Factor Database (TFDB) with binding site sequence information. Using this database, we tried to make a website for finding transcriptional regulatory elements.

Results: Using TFDB as a source for the binding site analysis, the initial program raised so many fault positive binding sites all around the gene. To decrease these candidates, we have compared the orthologus genes from human and mouse. We developed a web server which shows the binding site along with the homology and the CpG islands, which are suggested to be related to gene regulation.

Using the recent Microarray technology, the expression levels of many genes are easily measured. So, we also tried to search for the common sequence near the gene with similar expression patterns.

Availability: http://www.mizushima.info/hiroshi/bioinfo/tfds.html (temporaly).

Introduction

Draft sequence of human genome has been published analyzing the transcription regulation system of the genome information is becoming to be more important these days. Also, Microarray has been widely used to see large numbers of gene expression easily. Regulation mechanisms of some genes has been analyzed by molecular biologists using deletion/mutation assay using reporter gene, however, there is little knowledge of the regulation system in genome wide range, and no consensus algorithm has been developed for unknown gene regulation.

For this purpose, we have established and maintaining Transcription Factor Database (TFDB) with binding site sequence information (Mizushima et al., 1994, Okazaki et al., 1996). It was based on TFD database made by D.Ghosh at NCBI (Ghosh, 1992; Ghosh, 1993). We have made TFDB Maintenance system, which extracts data from Medline database, and molecular biologist can authorize the data through the web (Kaizawa et al., 1997; Kaizawa et al., 1998; Kaizawa et al., 1999). Using this database, we tried to make a website for finding transcriptional regulatory elements. This study is a trial to find the actual regulatory elements using TFDB data and conserved sequences between mouse and human.

Methods and Algorithms

TFDB data has been maintained by TFDB maintenance system (Kaizawa et al., 1997; Kaizawa et al., 1998; Kaizawa et al., 1999). Current database has about 3000 entries.

We have developed PC based program and e-mail server and opened it to the public (Mizushima et al., 1992). The match with the binding sequence is based on the percentage of the match and the numbers of the missmatch. These programs were useful however the false positives were too noisy. Using the algorithms in these programs as basic selection method, we made a checklist to select the desired factors on the list. Also we made the two-dimensioanl desplay for the maches so that we can see the conserved binding sites.

Web servers are constructed on Sun Netra Servers and PC Lynux(Red Hat 6.2) Servers as described anywhere. Programming of displaying applet was coded by JAVA.

Implementation and Results

First we made a system which shows the binding sequence within the submitted sequence. Fig. 1 shows the typical output of the system. Submitted sequence is searched for the binding sequence of the TFDB database with conditions specified by the query option. The table is shown (Fig. 1 rear window) and color or hiding for each factor can be selected. After

submitting the data, primary sequence with the distribution of the binding sites along with the CpG islands and CG contents are displayed using JAVA applet. (Fig. 1. front window). This system is capable of reading GENBANK formatted data with FEATURE tables, to show the features in the graphics.

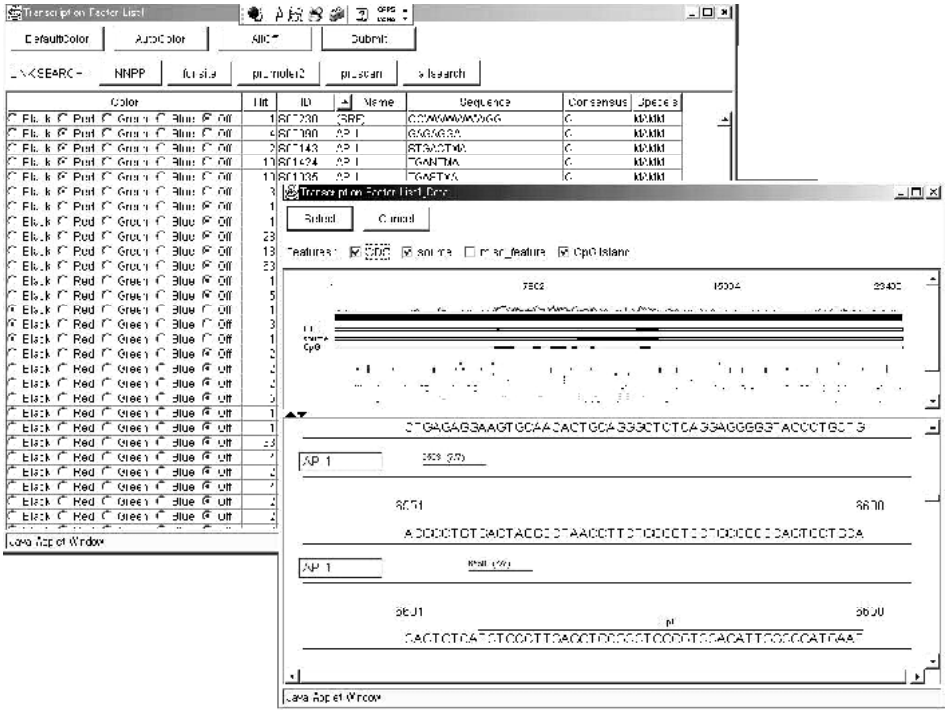


Fig. 1. Primary display of the transcription factor binding site in submitted sequence. Rear window shows the table of the binding site list, and front window shows the distribution of the binding sites selected in the rear window.

The usual consensus sequence of the transcription factor binding site is 6 to 10 bases, there are so many false positive candidates shown in the results. Actual regulation of the genes should have other information hidden in the genomic sequence. It is very difficult to select the functional binding sites just from the consensus binding sequence data.

Recently, genome sequence of other organisms became available. Using these data can help finding the important region of the genes, especially when the information is hidden somewhere. We used harr plot analysis to show the genome comparison.

As shown in Fig. 2, cut off value of the harr plot can be adjusted in the applet. Conserved binding sites of the selected transcription factor are shown by the colored plots.

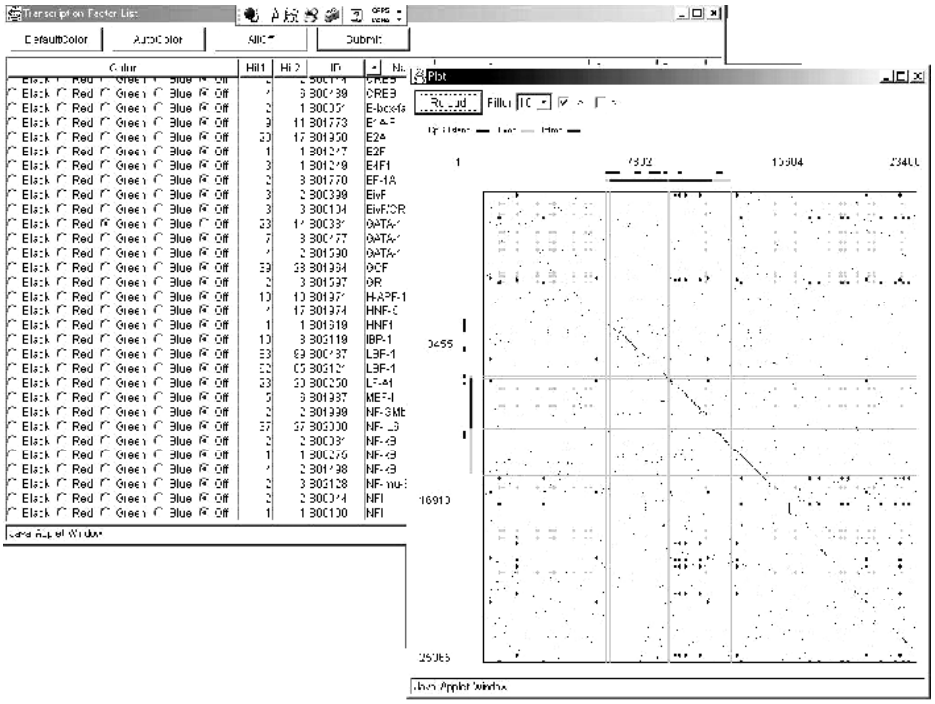


Fig. 2. Two-dimensional transcription factor binding site display. Rear window shows the list of the binding sites with numbers of binding sites in each sequence. Front window shows the conserved region with features (exon, intron) and CpG islands. Transcription factor binding sites are plotted at the conserved sites.

In Figures 1 and 2, TIS11B gene was used as an example. This gene seems to have 3 exons by the DNA sequence database and full length cDNA database. It is difficult to see which region is important in Figure 1. As shown in Figure 2, you can easily find the conserved region around the TIS11B gene. It is interesting that large region of intron is also conserved in this gene suggesting that there may be some important information in the intron region. There are also conserved region after the coding region, which suggests some hidden information after the gene. This kind of conserved region in introns and downstream of the gene was also observed in other genes.

Discussion

Venter reported that the human DNA binding proteins (transcription factors) also has protein-protein interaction motifs, suggesting that the transcription regulatory system is formed with multiple proteins with complicated regulatory mechanism (Venter et al., 2001).

Current system only can display 4 colors, which means 4 groups of transcription factors. As there are so many kind of transcription factors, and the combination of those factors seems to be important, we are planning to make it 10, so that more complex combination can be displayed.

Also, we are planning to change the system so that user can add some additional consensus sequence at the time of submitting the query. Also, we are planning to display the Open Reading Frame of the gene, Repeat sequences (such as Alu sequence, LINE sequence), SNPs, to make it as a multi-functional viewer.

As this system is availabe on the internet, anyone can use and send comments to the author.

Acknowledgements

The work was supported in part by the grant from Ministry of Education, Culture, Sports, Science and Technology. This study was also supported in part by the Program for Promotion of Fundamental Studies in Health Sciences of the Organization for Pharmaceutical Safety and Research of Japan. The authors are grateful to Mr. Takatsu and Mr. K.Kawahara of World Fusion Co. Ltd for coding the software. We also thank Dr. Lipman D., Dr. Landsman D., and Dr. Fujibuchi W. at National Center for Biotechnology Information (NCBI) for collaboration and helpful discussions.

References

1. Gohsh D. (1990) TFD: the transcription factors database. Nucl. Acid Res. 20, 2091-2093.
2. Gohsh D. (1993) Status of the transcription factors database (TFD). Nucl. Acid Res. 21, 3117-3118.
3. Mizushima H. (1992) Establishment of Trascriptional Factor Binding Site Searching Program. Proc. of 15th Japanese Molecular Biology Meeting.
4. Mizushima H., Hayashi K. (1994) Establishment of Transcription Factor Database and Human Mutaion Database. Genome Informatics Workshop, 1994, 5, 142-143.
5. Okazaki T., Kaizawa M., Mizushima H. (1996) Establishment and Management of Transcription Factor Database TFDB. Genome Informatics. 1996, 218-219.
6. Kaizawa M., Okazaki T., Mizushima H. (1997) Establishment of Transcription Factor Database TFDB Maintenance System. Genome Informatics. 1997, 292-293.
7. Kaizawa M., Watanabe S., Nobukuni T., Horikoshi M., Handa H., Kuchino Y., Sekiya T., Mizushima H. (1998) Maintenance of Transcription Factor Database TFDB by TFDB Maintenance System. Genome Informatics 1998. (Miyano S., Takagi T. Eds: ISSN:0919-9454), 316-318.
8. Kaizawa M., Watanabe S., Nobukuni T., Horikoshi M., Handa H., Kuchino Y., Sekiya T., Mizushima H. (1999) Maintenance of Transcription Factor Database TFDB. Genome Informatics 1999. (Asai K., Miyano S., Takagi T. Eds: ISSN:0919-9454), 269-271.
9. Venter C. et al. (2001) The sequence of the Human Genome. Science. 291, 1304-1351.

SITECON: A METHOD FOR RECOGNIZING TRANSCRIPTION FACTOR BINDING SITES BASING ON ANALYSIS OF THEIR CONSERVATIVE PHYSICOCHEMICAL AND CONFORMATIONAL PROPERTIES

[4] ***Oshchepkov D.Yu., Turnaev I.I., Vityaev E.E.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: diman@bionet.nsc.ru

Key words*: site recognition, conformational and physicochemical DNA properties, DNA–protein interactions, transcription factor*

Motivation: The local conformation of transcription factor binding sites resulting from their contexts is a factor determining the specificity of DNA–protein interactions. A set of conservative conformational and physicochemical parameters specific of sites for binding a particular transcriptional factor can be effectively used for their recognition.

Results: A method for recognizing potential transcription factor binding sites basing on a set of conservative context-dependent conformational and physicochemical properties determined for short regions of aligned functional DNA sequences is proposed in this work.

Introduction

Detection of potential sites for binding regulatory elements in genomic sequences is becoming a topical problem of the modern genomics. Similar to any affinity binding, the regulatory proteins bind specifically to DNA due to spatial structure complementarity of the interacting molecules, which, to a considerable degree, stems from conformational and physicochemical DNA properties (Starr et al., 1995; Meierhans et al., 1997). This imposes certain limitations on the sequence of a binding site, which are reflected in a partial conservation of its context. These assumptions formed the background of numerous methods for detecting and predicting regulatory regions, such as consensus method (Mulligan et al., 1984), weight matrices (Stormo et al., 1986), and the method of nucleotide frequency matrices, based on statistical physics approach to DNA–protein interactions (Berg, von Hippel, 1988).

Dickerson and Drew (1981) were the first to discover the dependence of the DNA conformation on its context using X-ray analysis of DNA dodecamers. An ever increasing data of structural analyses demonstrate both heterogeneity of conformational and physicochemical properties and their dependence on the nucleotide context (Frank et al., 1997; Suzuki et al., 1997). Consequently, as the local conformation of DNA molecules determined by the nucleotide sequence is a factor affecting the specificity of DNA–protein recognition, it is feasible to recognize the sites of interest not from the contextual characteristics, but basing on a set of properties of the DNA double helix.

Several mechanisms, which should be taken into account while analyzing the DNA sequences of functional sites and regulatory regions, can provide the preservation of specific properties. (1) Conservation of the nucleotide context, manifesting itself as a fixed pattern of nucleotides at certain positions in different site variants. The methods involving contextual analysis are suitable for studying this type of functional sites. (2) Conservation of certain conformational and physicochemical properties of particular regions within the site provided by the nucleotide substitution pattern preserving these properties. Such regions of functional sites may be detected using the methods analyzing the context-dependent conformational and physicochemical DNA properties.

The analysis involving determination of the mean value conservation of a property over a sequence region of 10–30 nucleotides long that either interacts with a protein or is located in the vicinity of the protein bound appeared very advantageous and allowed several methods for recognition and activity determination to be designed (Ponomarenko et al., 1997; Ponomarenko et al., 1999).

However, a more detailed analysis of a binding site sequence and study of specific DNA helix properties over shorter fragments of a functional sequence can aid the recognition accuracy. For example, analyzing the binding site for the transcription factor MetJ, Liu et al. (2001) demonstrated that the discriminatory power might be increased when DNA helix properties were taken into account in addition to conventional methods based on contextual analysis.

However, particular variants of genomic sequences interacting with specific regulatory proteins require preservation of only certain conformational and physicochemical properties. Study of the local properties of the DNA helix in the region of binding for their conservation provides more detailed information on specific conformational properties of this site, increasing the recognition accuracy of the analysis performed. In this case, analysis of a more complex pattern of conformational properties over the site sequence reflects more precisely the molecular mechanisms underlying the

*Corresponding author.

interactions of a functional sequence in question with the corresponding protein (Oshchepkov et al., this volume). Thus, such analysis allows us to determine the allowable conformations of the DNA regions, whose similarity suggests most efficient binding, that is, to determine finally those regions that are potential binding sites for the corresponding regulatory element.

In this work, a method for recognizing regulatory DNA regions is proposed. This method is based on comparison of context-dependent conformational and physicochemical properties of short fragments within functional DNA regions with the properties conserved in a set of transcription factor binding site sequences. The efficiency of this method is tested by an example of the binding site for the heterodimer E2F/DP (Zheng et al., 1999).

Methods and Algorithms

While analyzing sets of aligned DNA sequences of the sites, 38 conformational and physicochemical DNA properties (Ponomarenko et al., 1999) compiled in the database Property (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/) were used.

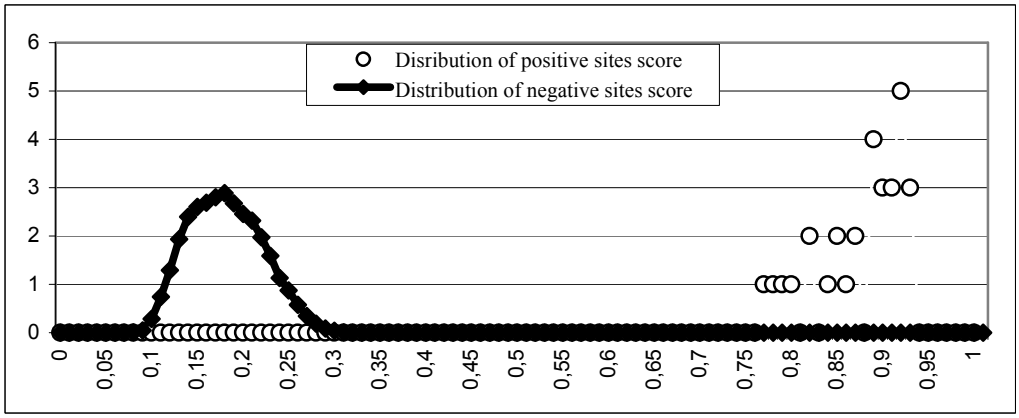
The essence of our approach is as follows. Conservative properties are determined for each position of a set of *N* aligned (phased) transcription factor binding site sequences with the length *L* (Oshchepkov et al., this volume); variance of a property at each position of the set is used as a measured of conservation.

We assume that if a particular property of a certain region within a sequence is important for the binding site function, the value of this property is conserved over all the sequences in a set, thereby providing a low variance compared with the variance for a set of random sequences. Thus, a low variance of a particular property suggests its conservation at the position in question (Oshchepkov et al., this volume). Significance of $\sigma_{F_{il}}$ is estimated using χ^2 test (Anderson, 1958). Then, we are assuming that the probability P_{il} of the *i*th property at position *L* of the sequence analyzed to take the value $\bar{F}_{il}$ required for the function at the value F_{iL} follows a Gaussian distribution:

$$P_{il} = \frac{1}{\sqrt{2\pi}\sigma_{F_{il}}} \exp((\overline{F_{il}} - F_{il})/\sigma_{F_{il}})^2 .$$

Let us select a sum P_{il} of all the significantly conservative properties normalized to the number of these properties as a measure of similarity between the sequences of the set and the sequence analyzed. This value corresponds to the probability of the properties of the DNA sequence analyzed to be close to the detected conservative properties of the sequences forming the initial set. Let us designate this value as the level of required conformational similarity.

Implementation and Results



To test the recognition method proposed, we selected a set of binding sites of the heterodimer E2F/DP with a length of 49 bp. The total number of sequences in the set amounted to 40. The overall distributions of the values of the required conformational similarity for the YES (positive) and NO (negative) sets are shown in the above figure. Note that the volume of the negative set amounted to 1.2×10^5; the negative sequences were generated by random shuffling of nucleotides in the initial site sequences; thus, the nucleotide compositions of both the positive and negative samples were identical. We selected the recognition threshold as 83% required conformational similarity; the discriminating power accuracy with reference to type I error was verified using a jack-knife technique with exclusion of 20% sequences from the learning set. The type I error (underrecognition) for the control amounted to 11%, as totally, 9 sequences absent in the learning set were not recognized in 10 series with exclusion of 20% sequences from the learning set. Type II error was controlled using recognition of the binding sites in a randomly generated sequence with a length of 73,300 bp (the recognition in both

directions was considered). Overall, 25 sequences were recognized as E2F binding sites, corresponding to the type II error equaling $3,4*10^{-2}$% (1/2900) for recognition in both strands.

Conclusion

A method for recognizing potential transcription factor binding sites basing on a set of conservative context-dependent conformational and physicochemical properties determined for short fragments of aligned functional DNA sequences is proposed and the effectiveness of the method is demonstrated.

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (grants № 00-04-49229 and 02-07-90355) and Siberian Branch of the Russian Academy of Sciences (integration project № 65). The authors are grateful to M.A.Pozdnyakov and O.V.Vishnevsky for helpful criticism and fruitful discussions.

References

1. Anderson T.W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons Inc. NY.
2. Berg O.G., von Hippel P.H. (1988). Selection of DNA binding sites by regulatory proteins II. The binding specificity of cyclic AMP receptor protein to recognition sites. J. Mol. Biol. 193:723-750.
3. Dickerson T.D., Drew H.R. (1981) Structure of B-DNA dodecamer. II. Influence of base sequence on helix structure. J. Mol. Biol. 149,761-786.
4. Frank D.E., Saecker R.M., Bond J.P., Capp M.W., Tsodikov O.V., Melcher S.E., Levandoski M.M., Record M.T. Jr. (1997). Thermodynamics of the interactions of Lac repressor with variants of the symmetric Lac operator: effects of converting a consensus site to a non-specific site. J. Mol. Biol. 267:1186-1206.
5. Liu R., Blackwell T.W., States D.J. (2001). Conformational model for binding site recognition by *E. coli* MetJ transcription factor. Bioinformatics. 17:(7):622-633.
6. Meierhans D., Sieber M., Allemann R.K. (1997). High affinity binding of MEF-2C correlates with DNA bending. Nucl. Acids Res. 25:4537-4544.
7. Mulligan M.E., Hawley D.K., Entriken R., McClure W.R. (1984). Escherichia coli promoter sequences predict *in vitro* RNA polymerase selectivity. Nucl. Acids Res. 12:789-800.
8. Oshchepkov D.Yu., Turnaev I.I., Vityaev E.E. (2002). Study of the context-dependent conformational and physicochemical properties of DNA functional sites. This volume.
9. Ponomarenko J.V., Ponomarenko M.P., Frolov A.S., Vorobyev D.G., Overton G.C., Kolchanov N.A. (1999). Conformational and physicochemical DNA features specific for transcription factor binding sites. Bioinformatics. 15(7/8):654-668.
10. Ponomarenko M.P., Ponomarenko Yu.V., Kel' A.E., Kolchanov N.A., Karas H., Wingender E., Sklenar H. (1997). Computer analysis of conformational features of the eukaryotic TATA-box DNA promoters. J. Mol. Biol. 31:733-740.
11. Starr D.B., Hoopes B.C., Hawley D.K. (1995). DNA bending is an important component of site-specific recognition by the TATA binding protein. J. Mol. Biol. 250:434-446.
12. Stormo G.D., Schneider T.D., Gild L. (1986.) Quantitative analysis of the relationship between nucleotide sequence and functional activity. Nucl. Acids Res. 14:6661-6679.
13. Suzuki M., Amano N., Kakinuma J., Tateno M. (1997). Use of 3D structure data for understanding sequence-dependent conformational aspects of DNA. J. Mol. Biol. 274:421-435.
14. Zheng N., Fraenkel E., Pabo C.O., Pavletich N.P. (1999.) Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes and Development. 13:666-674.

STUDY OF THE CONTEXT-DEPENDENT CONFORMATIONAL AND PHYSICOCHEMICAL PROPERTIES OF DNA FUNCTIONAL SITES

[5] ***Oshchepkov D.Yu., Turnaev I.I., Vityaev E.E.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: diman@bionet.nsc.ru

Key words*: conformational and physicochemical DNA properties, DNA–protein interactions, transcription factor*

Resume

Motivation: Research into molecular mechanisms underlying DNA–protein interactions using statistical analysis of nucleotide sequences of binding sites is most important for understanding the principles of gene expression regulation. The local conformation of transcription factor binding sites determined by their context is a factor responsible for specificity of the DNA–protein interactions. Analysis of the local conformations of a set of functional DNA sequences allows the conservative conformational and physicochemical properties reflecting molecular mechanisms of interactions to be determined.

Results: A method for determining the conservative context-dependent conformational and physicochemical properties in short regions of aligned functional DNA sequences is proposed in this work. The method was used to analyze binding sites of the heterodimeric complex E2F/DP. The discovered specific conformational properties for a set of these binding sites reflect the molecular mechanism of their interaction.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/.

Introduction

As a rule, each transcription factor is capable of interacting with regulatory DNA sequences differing in their context. Specific features of these interactions impose limitations on the sequences of binding sites, manifesting themselves in a partial conservation of the nucleotide sequence. These suggestions formed the background of numerous methods for detecting and predicting regulatory regions, such as consensus method (Mulligan et al., 1984), weight matrices (Stormo et al., 1986), and the method of nucleotide frequency matrices, based on statistical physics approach to DNA–protein interactions (Berg, von Hippel, 1988).

An increasing volume of experimental data suggests that the function of transcription factor binding sites is determined to a considerable degree by the conformational and physicochemical DNA properties (Starr et al., 1995; Meierhans et al., 1997). Dickerson and Drew (1981) were first to discover the dependence between the DNA conformation and its context using X-ray analysis of DNA dodecamers. An ever increasing data of structural analyses demonstrate both heterogeneity of conformational and physicochemical properties and their dependence on the nucleotide sequence (Frank et al., 1997; Suzuki et al., 1997). Thus, the local conformation of DNA molecules determined by the context is a factor affecting the specificity of DNA–protein recognition. This suggests that certain conformational and physicochemical properties of the variants of genomic sequences interacting with a certain regulatory protein should be preserved.

Several mechanisms, which should be taken into account while analyzing the DNA sequences of functional sites and regulatory regions, can provide the preservation of specific properties. (1) Conservation of the nucleotide context, manifesting itself as a fixed pattern of nucleotides at certain positions in different site variants. The methods involving contextual analysis are suitable for studying this type of functional sites. (2) Conservation of certain conformational and physicochemical properties of particular regions within the site provided by the substitution pattern preserving these properties. Such regions of functional sites may be detected using the methods considering the context-dependent conformational and physicochemical DNA properties. Thus, the conservation of nucleotide context may actually result from the limitations imposed on the values of particular properties at particular regions of the sequence. Hence, we hypothesize that analysis of local properties instead of the corresponding context may provide essentially more information on the structure of the site.

Two approaches are possible here. First, conservation of the mean value of a property over a sequence region of 10–30 nucleotides long that either interacts with a protein or is located in the vicinity of the protein bound is analyzed. This approach appeared very advantageous and allowed several methods for recognition and activity determination to be designed. Similar approach was realized in the system B-DNA-Video (Ponomarenko et al., 1999). However, this approach fails if individual conformational properties important for interaction are localized to short regions of the sequence studied.

* Corresponding author.

Second, the sequence of a site may be analyzed in more detail considering conservation of the properties over shorter regions of the functional sequence. In this case, analysis of a more intricate pattern of conformational properties over the site sequence may elucidate the molecular mechanisms involved in the interactions of the functional sequence analyzed with the corresponding protein. Generally speaking, such study aims to find the permissible conformations of a DNA region that provide most efficient binding, i.e. to determine finally the regions that are potential binding sites for the corresponding regulatory elements.

However, this approach is not free from limitations connected with both the accuracy of functional sequence phasing and the necessity of similar molecular interactions of the analyzed sequences with the protein. These limitations of the approach in question are related to the fact that conservation of properties is analyzed for the DNA regions that are localized to similar positions in the DNA–protein complexes with a certain transcription factor. Thus, when speaking about an analyzed position in alignment, we mean that the DNA regions (1) similar with reference to their positions in the DNA–protein complexes formed with a protein in question and (2) corresponding to particular position in alignment are analyzed.

In this work, we propose a method involving determination of conservative conformational and physicochemical properties in short regions of functional DNA sequences for studying DNA regulatory regions. Here, dispersion of conformational and physicochemical properties, compiled in the database Property (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/), at various positions within alignment of transcription factor binding sites was selected as a measure of conservation.

Materials

For the analysis, we selected binding sites for the heterodimer E2F/DP; the structure of E2F4/DP2–DNA complex was clarified by X-ray structure analysis (Zheng et al., 1999). E2F controls transcription of a group of genes whose expression is essential for the normal course of cell cycle, i.e. is actually a key regulator of the cell cycle. E2F forms heterodimers with proteins of a related DP family. Six E2F family proteins and two DP proteins are known for mammalians. E2F proteins differ structurally by the presence or absence of the sites for binding CDK (cycline-dependent kinase) and pRB (retinoblastoma protein) in various configurations. Six representatives of this family display the homology of 22 to 55%; DP1 and DP2, about 70%. The DNA-binding domain of E2F has a length of 70 aa, while this region of DP is longer, amounting to 90 aa; however, a fragment of 30 bp is identical in the DNA-binding domains of these proteins. The structure of this DNA-binding domain corresponds to winged-helix type (Jordan et al., 1994, Zheng et al., 1999); amino acid residues within structurally similar domains belonging to this type bind invariantly to DNA. This suggests that the winged-helix type domains of other E2F/DP combinations display similar specificities of DNA binding.

Methods and Algorithms

While analyzing sets of aligned DNA sequences of the sites, 38 conformational and physicochemical DNA properties compiled in the database Property (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/) were used.

The essence of our approach was as follows. A set of *N* aligned (phased) DNA sequences with the length *L* (without gaps) is considered. A value of a certain physicochemical or conformational property *f* is ascribed to each dinucleotide. Consequently, the matrix with a size *Nx(L-1)* is formed. An element of this matrix *f(l,m)* corresponds to the value of this particular property *f* of dinucleotide at the *m*th position in the sequence *l*.

The mean value of the property *i* at position *l* amounts to

$$\overline{F_{il}} = \frac{1}{N}\sum_{k=1}^{N} F_{ikl} \qquad (1)$$

Variance is used as a measure of conservation of the *i*-th property for each position *l*:

$$\sigma^2{}_{F_{il}} = \frac{1}{N-1}\sum_{k=1}^{N} (F_{ikl} - \overline{F}_{il})^2 \qquad (2)$$

It is assumed that if a particular property at particular location within the nucleotide sequence is important for the function of the binding site, the value of this property is conserved for all the sequences of the set, providing a low value of the variance compared with a sat of random sequences. Thus, a low variance of a particular property indicates its conservation at a particular position. The significance σ_{F_i} is estimated using χ^2 test (Anderson, 1958).

Implementation and Results

A set of 38 sites for binding E2F/DP heterodimer with a length of 49 bp was analyzed. The sequences with experimentally confirmed binding were extracted from TRRD (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/) and phased with respect to the consensus TTTCGCGCG without gaps.

Analysis of the set of E2F/DP binding sites for conservation of all the 38 properties allowed us to detect a number of specific DNA conformational properties, four of them being most interesting. The properties major/minor groove width/depth (Ponomarenko et al., 1997) and bendability towards the major groove (Gartenberg, Crothers, 1988) appeared conserved at all the positions within the site (8 bp at a 99% significance level according to chi-square test). Interestingly, conserved values of the major groove width (Fig. 1) for the region TTTCGCGCG and the minor groove width for the region TTTCGCGCG display the values exceeding considerably the mean values of these properties for random sample. This suggests that the properties in question are crucial for the DNA–protein interaction. However, the bendability profile is insufficiently coordinated, suggesting that the conservation discovered is likely to result from the interdependence between this property and the major groove width (correlation coefficients of these properties are 0.70 and 0.85, respectively).

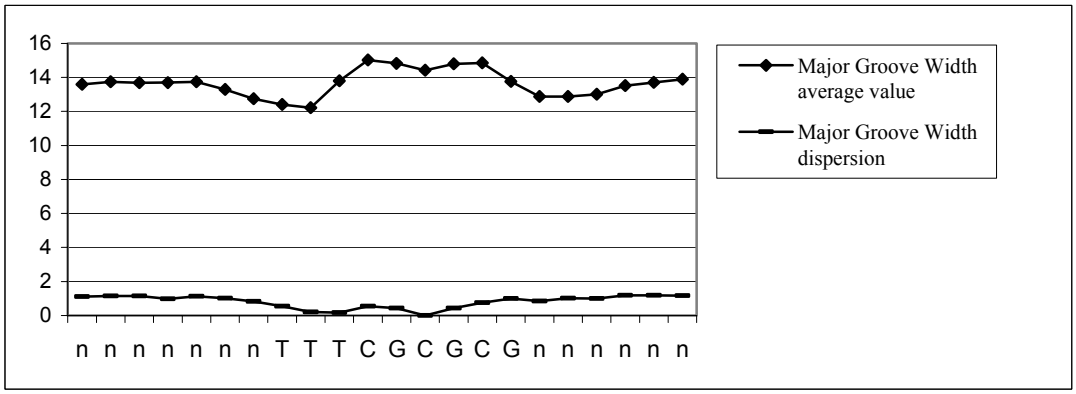


Fig. 1. The plots of major groove width at positions of the aligned E2F/DP biding sites, and its dispersion.

Discussion

The observations on the sizes of major and minor grooves comply with the X-ray structure analysis data. As was noted above, the transcription factor E2F belongs to the winged-helix type (Jordan et al., 1994). While binding, two recognition α helices, one from each of the heterodimer E2F/DP constituents, are localized to the major groove. In this process, the α helix of E2F DNA-binding domain lies in the region TTTCGCGCG, while the α helix of DP DNA-binding domain, in the region TTTCGCGCG. Thus, a widened major groove in the region TTTCGCGCG is likely to be a necessary condition for the heterodimer E2F/DP to recognize the binding site. Note that the dependence of this type is indeed the feature specific of E2F/DP binding sites, since the recognition α helices of both constituent proteins are spatially located one after the other (Fig. 2) and the bases contacting with these helices are also adjacent.

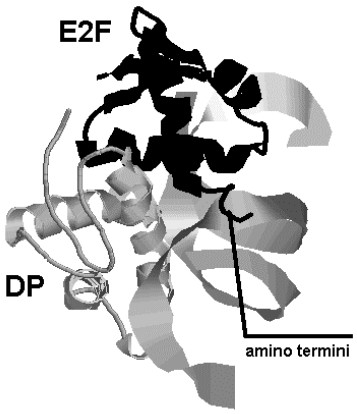


Fig. 2. E2F/DP complex with its target DNA sequence.

This complies with the data of Zheng et al. (1999) who demonstrated that these proteins bind to DNA already as heterodimer. It has been demonstrated that the T tract in TTTCGCGCG is critical for the binding (Jordan et al., 1994) and is necessary for insertion of the N-terminal of E2F recognition domain (Zheng et al., 1999), complying with the obtained data on a widened minor groove in the region TTTCGCGCG.

Conclusion

A method for detection and analysis of significant conformational DNA properties was developed. E2F heterodimer binding site was analyzed. The significant properties—sizes of major and minor grooves of two regions within the site—comply with the data on the mechanism of DNA–heterodimer binding.

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (grants № 00-04-49229 and 02-07-90355) and Siberian Branch of the Russian Academy of Sciences (integration project № 65). The authors are grateful to D.A.Afonnikov for helpful criticism and fruitful discussions.

References

1. Anderson T.W. (1958). An Introduction to Multivariate Statistical Analysis. NY: John Wiley & Sons Inc.
2. Afonnikov D.A., Oshchepkov D.Yu., Kolchanov N.A. (2001). Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with coadaptive substitutions. Bioinformatics. 17(11):1-12.
3. Afonnikov D.A., Oshchepkov D.Yu., Kolchanov N.A. (2000). Estimation of variances and covariances of protein physico-chemical characteristics in families of homologous sequences. Computational Technologies. 5(2):79-86.
4. Berg O.G., von Hippel P.H. (1988). Selection of DNA binding sites by regulatory proteins II. The binding specificity of cyclic AMP receptor protein to recognition sites. J. Mol. Biol. 193:723-750.
5. Frank D.E., Saecker R.M., Bond J.P., Capp M.W., Tsodikov O.V., Melcher S.E., Levandoski M.M., Record M.T.Jr. (1997). Thermodynamics of the interactions of Lac repressor with variants of the symmetric Lac operator: effects of converting a consensus site to a non-specific site. J. Mol. Biol. 267:1186-1206.
6. Gartenberg M.R., Crothers D.M. (1988). DNA sequence determinants of CAP-induced bending and protein binding affinity. Nature. 333:824-829.
7. Jordan K., Haas A., Logan T., Hall D. (1994). Detailed analysis of the basic domain of the E2F1 transcription factor indicate that it is unique among bHLH proteins. Oncogene. 9:1177-1185.
8. Meierhans D., Sieber M., Allemann R.K. (1997). High affinity binding of MEF-2C correlates with DNA bending. Nucl. Acids Res. 25:4537-4544.
9. Mulligan M.E., Hawley D.K., Entriken R., McClure W.R. (1984). *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. Nucl. Acids Res. 12:789-800.
10. Ponomarenko J.V., Ponomarenko M.P., Frolov A.S., Vorobyev D.G., Overton G.C., Kolchanov N.A. (1999). Conformational and physicochemical DNA features specific for transcription factor binding sites. Bioinformatics. 15(7/8):654-668.
11. Ponomarenko M.P., Ponomarenko Yu.V., Kel' A.E., Kolchanov N.A., Karas H.R.A., Wingender E., Sklenar H. (1997). Computer analysis of conformational features of the eukaryotic TATA-box DNA promoters. J. Mol. Biol. 31:733-740.
12. Starr D.B., Hoopes B.C., Hawley D.K. (1995). DNA bending is an important component of site-specific recognition by the TATA binding protein. J. Mol. Biol. 250:434-446.
13. Stormo G.D., Schneider T.D., Gild L. (1986). Quantitative analysis of the relationship between nucleotide sequence and functional activity. Nucl. Acids Res. 14:6661-6679.
14. Suzuki M., Amano N., Kakinuma J., Tateno M. (1997). Use of 3D structure data for understanding sequence – dependent conformational aspects of DNA. J. Mol. Biol. 274:421-435.
15. Zheng N., Fraenkel E., Pabo C.O., Pavletich N.P. (1999). Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes and Development. 13:666-674.

RECOGNITION OF E2F TRANSCRIPTION FACTOR BINDING SITES

[6] ***Turnaev I.I., Oshchepkov D.Yu., Podkolodnaya O.A.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: turn@bionet.nsc.ru

Key words: *cell cycle, site recognition, E2F site*

Resume

Motivation: Transcription factors of the E2F family play an important role in transcription regulation during the cell cycle. Thus, it is evident that prediction of potential binding sites for this factor in genes whose products are involved in this process is of great importance.

Results: The method SITECON, based on determination of conservative context-dependent conformational and physicochemical DNA properties in short DNA fragments, allowed us to detect potential sites for binding of E2F transcription factor in promoters of several genes involved in the cell cycle regulation. It was demonstrated that potential E2F binding sites were much more abundant in promoters cell cycle genes compared with promoters of other groups of genes.

Introduction

The E2F family of transcription factors plays an important role in regulation of gene expression at the G1/S transition of the eukaryotic cell cycle. Binding sites of the E2F transcription factor were found in promoters of genes whose products are essential for nucleotide synthesis, DNA replication, and cell cycle progression (2000). The transcription of genes regulated by E2F (heterodimer E2F/DP) increases in the G1 phase of cell cycle, reaches its maximum in the late G1 or early S phases, and decreases in the G2 and M phases. This pattern is mediated by the activity of the transcription factor. It binds to its sites in promoters of the corresponding genes and regulates their transcription according to cell cycle phases. At present, nine proteins of the E2F family are known for mammals; of them, six belong to the E2F subfamily (E2F1-6) and three, to the DP subfamily (DP1-3). The structures of E2F family members representatives differ in the presence or absence of cyclin-dependent kinase (CDK) and protein retinoblastome (pRB) binding sites in different configurations. The DNA-binding domains have the structure of winged helix (Jordan et al., 1994; Zheng et al., 1999). The proteins of these two families form the heterodimer E2F/DP, in which the DP subunit is essential for strengthening the binding of this transcription factor to DNA. Most E2F/DP heterodimers can activate transcription. An exception is E2F6/DP1,2,3, which suppresses transcription. Pocket proteins (pRB, p130, and p107) bind to the heterodimer E2F/DP, screen its activation domain, and thereby block its ability to activate transcription. Moreover, this is accompanied by attraction of histone deacetylase and the corresponding suppression of transcription. The transcription of the E2F-controlled genes in quiescent cells is blocked by both mechanisms. As the transcription factors of the E2F family play a key role in transcription regulation in the course of cell cycle, the prediction of potential binding sites of these factors in genes whose products are involved in this process is an important task.

Materials and Methods

The recognition program was trained on a sample containing 42 E2F sites of 49 bp in length, obtained from the TRRD database (Kolchanov et al., 2002). The sample contained only those sites whose binding to transcription factors of the E2F family was demonstrated experimentally. The recognition was performed by the SITECON method (Oshchepkov et al., 2002) using the published data on 38 conformational and physicochemical properties of dinucleotides stored in the PROPERTY database (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/). The method is based on determination of conservative context-dependent conformational and physicochemical properties of short DNA sequences. We chose the 83, 85, and 87% levels of required conformational similarity as recognition thresholds (Oshchepkov et al., 2002). The recognition quality for type I errors was checked by the jack-knife method with exclusion of 20% of sequences from the training sample. Control for type II errors was performed by recognition of binding sites in a randomly generated sequence with a length of 73,300 bp, and the recognition was carried out in both directions. Evaluation of type I and II errors for these levels of conformational similarity is shown in Table 1.

The sample of sequences for recognition of potential E2F binding sites included the promoter regions of 35 genes located from –800 to +300 relatively to the transcription starts of these genes. The sequences were chosen from the TRRD database. The criterion for the choice of these genes was that their product participated in cell cycle regulation but there was no experimental evidence for the presence of E2F binding sites in their promoter regions. Recognition was also performed in

*Corresponding author.

promoter samples of other groups of lipid metabolism genes, erythroid-specific genes, IFN-induced genes, and genes expressing in macrophages, which were also obtained from TRRD.

Table 1. Recognition errors for various levels of conformation similarity.

	83%	85%	87%
Type I error	0.112	0.156	0.187
Type II error	$3.4*10^{-4}$ (1/2932)	$1.6*10^{-4}$ (1/7330)	$5.4*10^{-5}$ (1/18325)

Results and Discussion

The training sample of E2F sites includes experimentally verified sites located in the promoter regions from –800 to +300 bp and aligned with the consensus TTTSGCGCSMDR (Kel et al., 2001). Of these genes, 13 have one E2F site each, 12 have two, and 1 gene (DNA polymerase alpha 180; DPOLA180) has three E2F sites. Most of the E2F sites (34 of 42) are located in the vicinity of the transcription start from –150 to +50 bp (Fig. 1). The central part of the E2F site contains a highly conservative GC sequence, to the right of it, in the 5' region, there is a TTT sequence, whereas the fragment to the left from the central part (3'-region) is less conservative. Eleven of the forty-two E2F sites of the sample are palindromic and have the consensus TTTSGCGCSAAA. Such sites are present in human, murine, and hamster *DHFR*; human and murine *E2F1;* human *PCNA* and *CMYC;* murine *HTF9a; Drosophila ORC1* and *DPOLA 180;* and adenovirus type 5 *E1A* genes.

Recognition of binding sites of the E2F transcription factor was performed in five samples of promoter sequences for groups of genes involved in various processes at the conformational similarity threshold of 83%.

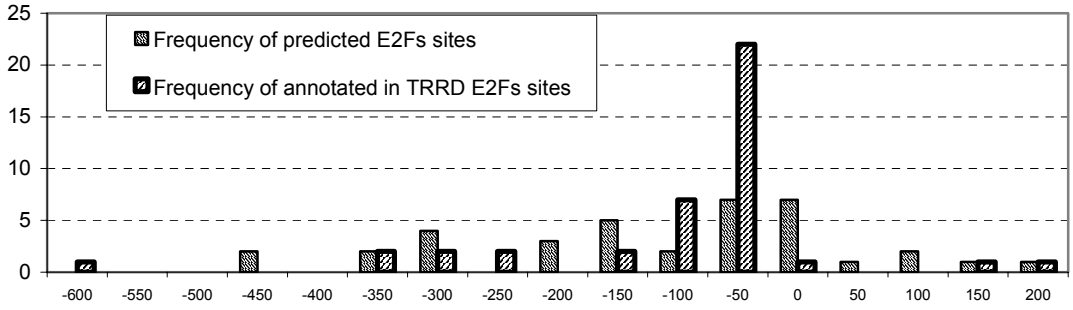


Fig. 1. Distribution of predicted and experimentally confirmed E2F sites in promoters of cell cycle-related genes. Axis of abscissas: positions relatively to the transcription start; axis of ordinates: number of sites per 50 bp.

Table 2. Results of recognition of E2F sites in the samples of promoters of various gene groups (from –600 to +100 bp relatively to the transcription start).

Promoter sample	Number of genes	Potential sites	Frequency (sites/bp)
Cell Cycle	35	35	1/700
Lipid. Met.	48	15	1/2200
Erythroid-specific	41	9	1/2505
IFN-induced	85	17	1/3596
EXP. Macroph.	47	7	1/4308

Promoters of cell cycle-related genes (from –600 to +100 bp) were found to contain 35 potential E2F sites in 21 of 35 promoters. According to these data, the density of E2F site localization is evaluated to be approximately one site per 700 bp. Extension of the region under consideration by 400 bp (–800 to +300 bp) revealed five additional E2F binding sites, which is insignificant.

The distribution of the density of potential E2F binding sites in the cell cycle promoters is shown in Fig. 1. Most of the recognized sites occur in the range from –300 to +50 bp relatively to the transcription start. Experimentally confirmed E2F sites, whose distribution is also shown in Fig. 1, have a similar localization pattern.

The frequency of potential E2F binding sites in promoters of other specifically expressed groups of genes is much less than in the cell cycle gene promoters. Promoters of lipid metabolism genes, erythroid-specific genes, IFN-induced genes, and genes expressing in macrophages contain less E2F sites by factors of 3.14, 3.58, 5.1, and 6.1, respectively. Apparently, potential E2F sites are significantly more abundant in cell cycle-related genes than in promoters of other gene groups under study.

Characteristic of the training sample of experimentally confirmed E2F sites is a pairwise localization of more than half of the sites (22 of 42), the distance between site centers being mostly within the range 10–20 bp. As in the training sample, a considerable proportion of the E2F sites recognized in promoters of cell cycle-related genes occur in clusters of 2–5 sites, whose centers are spaced 9–40 bp apart. Figure 2 shows the overall distribution of distances between close pairs of E2F sites, both experimentally confirmed and predicted. The mean distance in site pairs is 24.5 bp. Thus, the clusterwise occurrence of E2F sites appears to be a characteristic feature of the cell cycle-related genes. The COMPEL database contains two real composite elements (pairs of sites whose transcription factors are functioning cooperatively owing to specific DNA–protein and protein–protein interactions): E2F-E2F in the human *p107* (C00121) gene and the adenovirus type 2 *E2AE1* (C00193) gene (Kel-Margoulis et al., 2000). Thus, our study suggests that such pairwise composite elements occur much more widely in promoters of the cell cycle-related genes.

For additional verification of the recognition quality, we searched for potential E2F sites in promoter samples of genes of cell cycle and lipid metabolism at two higher threshold levels of the conformation similarity: 85 and 87%. The dependence of the frequency of occurrence of recognized sites on the recognition threshold value is shown in Fig. 3.

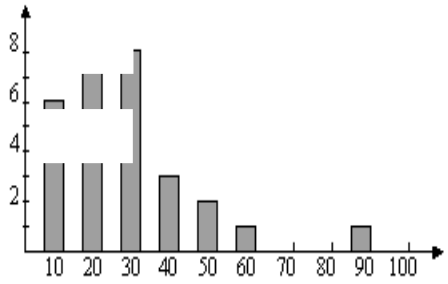


Fig. 2. The overall distribution of distances between closely located (centers ≤100 bp apart) E2F sites, experimentally confirmed and predicted. Axis of abscissas: positions relatively to the transcription start; axis of ordinates: the total numbers of experimental and predicted sites per 10 bp.

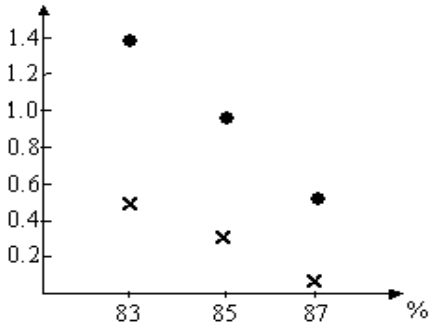


Fig. 3. The number of E2F sites predicted in promoters of genes of cell cycle and lipid metabolism. Axis of abscissas: the threshold of conformation similarity, %; axis of ordinates: the number of predicted sites per 1000 bp; CC: the number of predicted sites in promoters of cell cycle-related genes; and LM, in promoters of genes of lipid metabolism.

The comparison shows that E2F sites are significantly (no less than threefold) more abundant in promoters of the cell cycle-related genes than in any control sample for any recognition thresholds. An increase in the threshold of conformational similarity leads to an increase in the difference between the recognition frequency of potential E2F sites in the cell cycle promoters and the recognition frequency for the control set. This corresponds to a greater conformation similarity of the sites found in cell cycle promoters to the training sample.

The genes in which E2F sites have been predicted include those whose products are involved in cell cycle progression: rat cyclin *D1,* quail *E2F1,* and human *cdc25a* (G1/S transition and the S phase), cyclin *A1* (S phase and mitosis progression), and *cdc25c* (start of mitosis); genes involved in early mitogen-dependent activation of cyclin D transcription: murine *C-fos* and rat *N-Myc*; genes, whose products are essential for DNA duplication in the S phase: hamster *TK*, murine *PCNA*, and human *Dpola*; and histone genes: murine *h2ax* and human *H3*. The TRRD database contains experimental evidence for function of three of the predicted sites (murine PCNA (+10-22d{S1920}) and quail *E2F1* (-21-+1{S3143}, +7-17{S3144}) but not for binding of the E2F transcription factor to these sites. Some of the predicted sites occur in genes, whose homologues were formerly shown to contain functional E2F sites: human cyclin *A1* and DNA polymerase alpha, quail *E2F1*, rat *N-Myc*, and *cycD1*, and murine *PCNA*. Potential E2F binding sites have also been recognized in genes, for whose homologues the presence of functional E2F sites was not shown: (human and murine *IRF-1, C-fos, JUNB*, and *CDC7* and rat *IL6* and *ALDA*). This suggests the presence of new feedbacks for E2F/DP-dependent activation of the G1–S transition. Such activation of the genes *C-fos* or *P15INK4B* would, facilitate or limit, respectively, the progression of cell cycle.

Conclusion

The results of recognition have shown that the distribution of potential E2F binding sites in the investigated sequences of the cell cycle-related genes is similar to that of the known E2F binding sites. New potential E2F binding sites have been predicted. It has been shown that the concentration of potential E2F binding sites in promoters of the cell cycle-related genes is significantly higher than in promoters of other gene groups. Combination of experimental data with results of computer recognition of E2F binding sites has allowed a closer analysis of the promoter regions of cell cycle-related genes.

Acknowledgements

The authors are grateful to N.A.Kolchanov for fruitful discussion of the study, E.A.Ananko and E.V.Ignat'eva for supplying the promoter samples, and V.V.Gulevich for translating the manuscript into English. The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 00-07-90337); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Ivey-Hoyle M, Conroy R, Huber H.E., Goodhart P.J., Oliff A., Heimbrook D.C. (1993). Cloning and characterization of E2F-2, a novel protein with the biochemical properties of transcription factor E2F. Mol. Cell. Biol. 13:7802-7812.
2. Jordan K., Haas A., Logan T., Hall D. (1994). Detailed analysis of the basic domain of the E2F1 transcription factor indicates that it is unique among bHLH proteins. Oncogene. 9:1177-1185.
3. Kel A.E., Kel-Margoulis O.V., Farnham P.J., Bartley S.M., Wingender E., Zhang M.Q. (2001). Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. J. Mol. Biol. 309:99-120.
4. Kel-Margoulis O.V, Romashchenko A.G., Kolchanov N.A., Wingender E., Kel A.E. (2000). COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucl. Acids Res. 28:311-315.
5. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30:312-317.
6. Oshchepkov D.Yu., Turnaev I.I., Vityaev E.E. (2002). SITECON: a method for recognizing transcription factor binding sites basing on analysis of their conservative physicochemical and conformational properties. This volume.
7. Wells J., Boyd K.E., Fry C.J., Bartley S.M., Farnham P.J. (2000). Target gene specificity of E2F and pocket protein family members in living cells. Mol. Cell. Biol. 20:5797-5807.
8. Zheng N., Fraenkel E., Pabo C.O., Pavletich N.P. (1999). Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes and Development. 13:666-674.

RECOGNITION OF BINDING SITES FOR THE TRANSCRIPTION FACTORS SREBP, PPAR, HNF4, COUP-TF, AND SF-1 BY A GENETIC ALGORITHM BASED ON ITERATIVE DISCRIMINANT ANALYSIS

**** Levitsky V.G., Ignatieva E.V., Proscura A.L., Pozdnyakov M.A., Busygina T.V.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: levitsky@bionet.nsc.ru
*Corresponding author

Key words: *transcription factor binding sites recognition, lipid metabolism, endocrine system, genetic algorithm, discriminant analysis*

Summary

Motivation: Development of methods for search for transcription factor binding sites (TFBSs) is important for investigation of regulatory regions in eukaryotic genes. The accuracy of currently used methods is insufficient for correct recognition of binding sites of a given transcription factor (TF) or a group of factors of a single family, class, etc.

Results: We propose a new approach to search for TFBSs by the example of five site types: SRE, PPRE, HNF4, COUP-TF, and SF-1. The approach involves partitioning the regions including TFBSs into local regions and selection of the most suitable dinucleotides for each of the regions. Cross-validation tests of the functions of TFBS recognition allowed division of the sites into two groups which agree with the structure-functional subdivision of these TFs into two superclasses: (1) Basic Domains and (2) Zinc-coordinating DNA-binding domains.

Availability: The program for site recognition is included into the GeneExpress system; section "RegScan", http://wwwmgs.bionet.nsc.ru/mgs/programs/sitega/.

Introduction

The 5'-regulatory regions of genes are characterized by extremely complex structure and abundance of regulatory elements (Kolchanov et al., 2002). The regulatory regions of eukaryotic genes are following a block-hierarchical model of organization. Transcription-factor binding sites are the most important units of the structure. A single set of regulatory elements (blocks) present in a certain regulatory region can give rise to a great variety of transcription complexes and, as a result, implementation of various expression patterns. Core domains are recognized within a TFBS. There can be one or several of them, and they are separated with variable regions (spacers). The presence of cores is related to the fact that transcription factors have a block structure and may contain several domains or subunits, performing specific functions (Wingender, 1997). Taking into account the importance of the block structure of TFBSs, we used this principle as the basis for the algorithm for their recognition. To reveal the block structure of a site and its flanks, we followed an approach based on the determination and analysis of context-homogenous DNA regions. We analyzed samples of binding sites of five types: SRE, PPRE, HNF4, SF-1 and COUP-TF. Formerly, we showed that the transcription factors SREBP and PPAR, interacting with the sites SRE and PPRE, respectively, and the factor HNF4 are involved in the regulation of transcription of lipid metabolism genes (Ignatieva et al., 2000). The SF-1 factor of the nuclear receptor family is known to be a regulator of development and function of the hypothalamic-pituitary-gonadal complex and adrenals (Luo, 1999). This factor plays a crucial role in regulation of the transcription of genes controlling steroid production (Busygina et al., 2000). The transcription factor COUP-TF is a negative transcription regulator (Tsai, Tsai, 1997), and its binding site frequently occurs in the regulatory regions of genes of the lipid metabolism system, as well as in the regulatory regions of genes of the endocrine system (Ignatieva et al., 2000; Busygina et al., 2000).

Methods and Algorithms

The method we developed for recognition of TFBSs consists of two stages: (1) search for a partition of the region including the site and adjacent parts into local regions and (2) selection of major dinucleotides within each region. Both stages were implemented with the use of a genetic algorithm, which utilizes a linear discriminant function of dinucleotide frequencies characteristic of the local regions as a functional. Here we present analysis of samples of 120 bp long TFBS regions bearing the binding site of a certain TF.

Sample sizes are listed in Table. Random sequences obtained by mixing positive samples were used as negative samples.

The parameter of the method is p, the number of local regions into which the regions is partitioned. In our study, p=10. The method was described in more detail in (Levitskii, Katokhin, 2001; Levitsky et al., 2001; Levitsky, Katokhin, this issue).

For recognition of a TFBS in a nucleotide sequence, a 120 bp long sliding window are considered. Denote this fragment as X. For each position of the window, find the recognition function value according to the following equation:

$$\varphi(X) = \frac{1}{R^2} \times \sum_{n=1}^{N} \sum_{k=1}^{N} \{[f_n(X) - (1/2)] \times [f_n^{(2)} + f_n^{(1)}] \times S_{n,k}^{-1} \times [f_k^{(2)} - f_k^{(1)}]\} \quad (1)$$

Here, $f_n^{(1)} = f_{n(i,p)}^{(1)}$ is the frequency of the i^{th} dinucleotide for the p^{th} region averaged over the site sample; and $f_n^{(2)}$ is the corresponding frequency for the negative sample; and $f_n(X)$ is the dinucleotide frequency vector constructed with regard to the partition of the fragment X under study into local regions. The n(i, p) dependence is defined during the selection of the most significant dinucleotide frequencies for the partition regions. Denote the number of dinucleotides used for p_{th} region as $n_p \leq 16$. Then the total number of variables in (1) is determined as

$$N = \sum_{p=1}^{P} n_p \quad (2)$$

In equation (1) S^{-1} denotes the reverse matrix for the combined covariation matrix; $S = S^{(1)} + S^{(2)}$, $S^{(1)}$ и $S^{(2)}$ are covariation matrices for the positive and negative sequence samples for the dinucleotide frequency vectors $f_n^{(1)}$ и $f_n^{(2)}$; and R^2 is the Mahalanobis distance (Mahalanobis, 1936) between the samples of sites and random sequences:

$$R^2 = \sum_{k=1}^{N} \sum_{n=1}^{N} \{[f_n^{(2)} - f_n^{(1)}] * S_{n,k}^{-1} * [f_k^{(2)} - f_k^{(1)}]\} \quad (3)$$

Values of function φ(X) close to 1 correspond to higher probabilities of site recognition. For predicting TFBSs, the program uses the significance level index α. The TFBS recognition function φ(X) (obtained according to (1)) was converted as follows:

$$\varphi_\alpha(X) = \begin{cases} \dfrac{|1 - \varphi(X)|}{P_\alpha \times \sigma_\varphi}, & \text{if } |1-\varphi(X)| < P_\alpha \times \sigma_\varphi, \end{cases} \quad (4)$$

Here, P_α is the α quantile of the standard normal distribution (for example, $P_{0.95} = 1.96$), and σ_φ is the standard deviation of the values of the recognition function φ(X) over the site sequence sample. Nucleotide sequence regions with $\varphi_\alpha(X) > 0$ were considered to be potential TFBSs.

Implementation and Results

We compiled samples of TFBSs of five types from the data banks TRRD and EMBL (Table). The site samples are 120 bp long fragments of the regulatory regions of genes containing the sequence of a particular type in the center.

Table. TFBS samples used in the analysis.

Site name	Transcription factor interacting with the site	Number of sequences	Number of sequences for cross-validation test
SRE*	SREBP (Sterol Regulatory Element Binding Protein)	27	27
PPRE**	PPAR (Peroxisome Proliferator-Activated Receptor)	16	11
HNF4	HNF4 (Hepatic Nuclear Factor 4)	35	33
COUP-TF	COUP-TF (Chicken Ovalbumin Upstream Promoter Transcription Factor)	33	12
SF-1	SF-1 (Steroidogenic Factor 1)	39	30

SRE, Sterol Regulatory Element

** PPRE - Peroxisome Proliferator-Responsive Element

By the above-described method, we constructed recognition functions for sites of five types. The best partitions of site DNA sequences with adjacent flanks into local regions calculated for the five site types are shown in Fig. 1. The recognition functions were tested by cross-validation tests of DNA sequences from the samples collected by us. For these tests, we excluded sequences from different samples occurring in same genes and partly overlapping (Column 4 in Table) from the site samples used as learning samples for the recognition programs (Column 3 in Table).

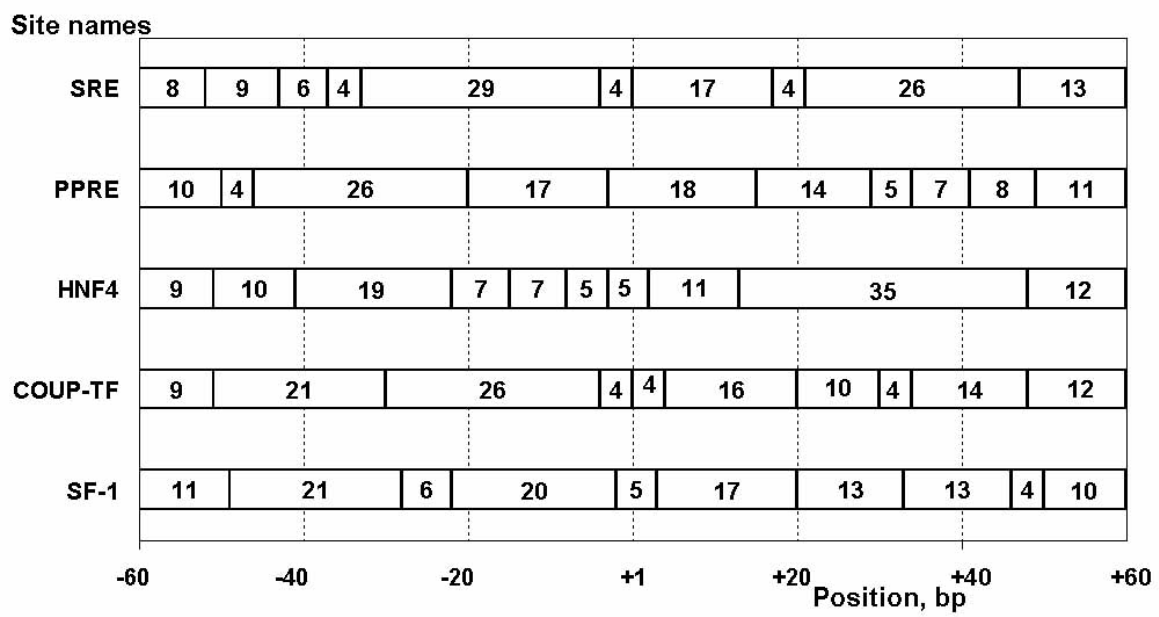


Fig. 1. Optimum partitions of the regions [–60; +60] of sites with adjacent regions into local regions calculated for five types of sites: SRE, PPRE, HNF4, COUP-TF, and SF-1.

For estimation of the accuracy of recognition, we calculated correlation coefficients (CC), which describe the accuracy of site recognition. An increase in CC points to an increase in site recognition accuracy. We concluded that the sites HNF4 and PPRE (CC are equal to 0.76 and 0.72, respectively) demonstrated a greater accuracy of recognition, and the accuracy of COUP-TF, SF-1, and SRE is lower (CC 0.63, 0.54, and 0.39, respectively).

The results of cross-validation tests of the recognition function performed for the five types of binding sites are shown in Fig. 2. A site was considered recognized if the value $\varphi_\alpha(X)$ (equation 4) was positive at the significance level α=0.95.

Consider the results of the cross-validation tests in more detail. It is seen from Fig. 2 that four sites (COUP-TF, PPRE, HNF4 and SF-1) are more efficiently recognized with the recognition functions of one another. On the contrary, the site SRE is less recognizable with the recognition functions of COUP-TF, PPRE, HNF4 and SF-1. According to Fig. 2, recognized sites constitute 7.4, 3.7, 7.4, and 11.1%, respectively.

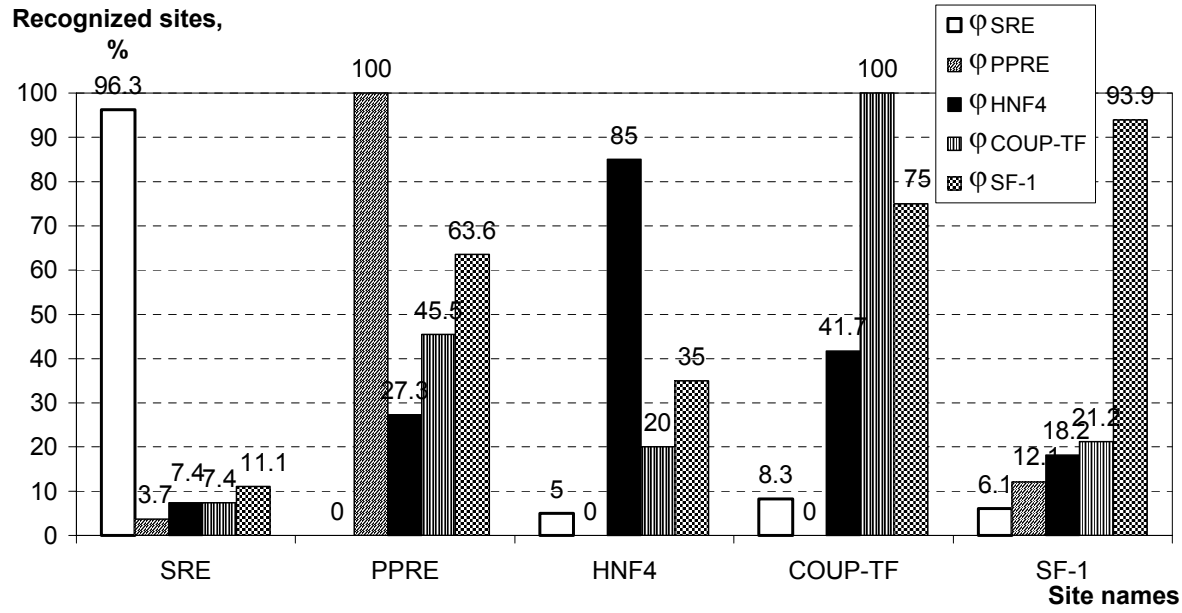


Fig. 2. Cross-validation tests of the recognition functions constructed for the samples of the binding sites COUP-TF, PPRE, SRE, HNF4, and SF-1. Y axis, the percentage of recognized sites. Equally filled bars denote recognition programs corresponding to learning samples of a particular site. Types of site samples for which the analysis was performed are indicated on the X-axis.

Discussion

Thus, we found that classification of TFBSs on the base of cross-validation tests of recognition functions into two groups—(1) COUP-TF, PPRE, HNF4, and SF-1 and (2) SRE—matches the structure-functional subdivision of the corresponding TFs into two classes: (1) Basic helix-loop-helix/leucine zipper factors (bHLH/ZIP), superclass Basic domains, and (2) Cys4 zinc finger of nuclear receptor type, superclass Zinc-coordinating DNA-binding domains. This appears to be related to the fact that the attribution of a factor to a particular class determines a context uniformity, which is revealed by cross-validation tests of recognition functions.

It should also be mentioned that the sites COUP-TF, PPRE, SRE and HNF4 include two cores, and the site SF-1 includes only one. This is reflected in the structure of partitions used for construction of site-recognition functions (Fig. 1). It is seen that the central part of the partition of the one-core site SF-1 contains a 5 bp long region bordered by longer regions, whereas the central parts of the two-core sites COUP-TF and HNF4 contain two partition regions. The partitions of the sites PPRE and SRE do not show any clear dependence of the number of cores on the partition structure in the central parts of the sites. Most probably, this illustrates the weak context signal of the site cores in comparison with the neighboring flanking regions.

Acknowledgements

The authors are grateful to Prof. N.A.Kolchanov for fruitful discussion and Dr. A.V.Osadchuk for supporting the work on the construction of the sample SF-1. The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 02-07-90355, and 00-04-49229); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institute of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Busygina T.V., Ignatieva E.V., Osadchuk A.V. Steroidogenesis-controlling gene transcription regulation: representation in TRRD database. Proc. Second Intern. Conf. on Bioinformatics of Genome Regulation and Structure. (BGRS'2000), Novosibirsk, 2000. 1, 41–44.
2. Ignatieva E.V., Likhoshvai V.A., Ratushny A.V., Kosarev P.S. Knowledge base on molecular-genetical foundations of lipid metabolism regulation: current state and perspective. Proc. Second Intern. Conf. on Bioinformatics of Genome Regulation and Structure. (BGRS'2000,), Novosibirsk, 2000. 3, 54–57.
3. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 2002, 30(1), 312–317.
4. Levitskii V.G., Katokhin A.V. Computer analysis and recognition of Drosophila melanogaster gene promoters. Mol. Bio. (Mosk.). 2001, 35(6), 970–978.
5. Levitskii V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis. Bioinformatics. 2001, 17, 998–1010.
6. Levitskii V.G., Katokhin A.V. Proc. Third International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002), this issue.
7. Luo X., Ikeda Y., Lala D., Rice D., Wong M., Parker K.L. Steroidogenic factor 1 (SF-1) is essential for endocrine development and function, J. Steroid. Biochem. Mol. Biol. 1999, 69, 13–18.
8. Tsai S.Y., Tsai M.J. Chick ovalbumin upstream promoter-transcription factors (COUP-TFs): coming of age. Endocr. Rev. 1997, 18, 229–240.
9. Wingender E. Classification of eukaryotic transcription factors. Mol. Biol. (Mosk.). 1997, 31(4), 584–600.

KERNEL METHOD FOR IDENTIFICATION OF LOCAL PATTERNS IN UNALIGNED SETS OF FUNCTIONAL SITES

[1] Tikunov Y., [1,2] Kel A.

[1] Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia, e-mail: tikunov@uiggm.nsc.ru
[2] BIOBASE GmbH, D-38304 Wolfenbüttel, Germany

Key words: *weight matrix, transcription factors, transcription factor binding sites, kernel method*

Resume

Motivation: Functional sites in nucleotide sequences, for instance sites for binding of transcription factors, often defer significantly in their primary sequence and can be hardly aligned. One or several hidden patterns that can be described by weight matrices need to be revealed from the sets of unaligned functional sites.

Results: We developed a new method for the analysis of unaligned sets of functional sites based on kernel estimation. The method is able to reveal "local patterns" – a set of weight matrices. Every weight matrix characterize a pattern that can be found in a significant subset of sites under analysis. The developed method has been applied for the analysis of a set of a mixture of binding sites for AP-1 and CREB transcription factors. The method was able to reveal two distinct patterns that correspond to the two considered TFs. Other available methods (CONSENSUS and Gibbs sampling) failed to differentiate these two patterns.

Availability: The program will be available for on-line use at: http://www.gene-regulation.com as well as at: http://compel.bionet.nsc.ru.

Introduction

Patterns shared by multiple nucleic acid sequences reflect their molecule structure, function, and evolution. So the recognition of such patterns is important problem in the analysis of nucleotide as well as protein sequences. In the context of this problem there may be formulated a problem which aim is to locate relatively short patterns shared by otherwise dissimilar sequences. Many different information measures were used for the solution of this problem (Hertz, Stormo, 1999; Lawrence et al., 1993). In the current presentation we show that the measure of probability density may be used more successfully. We developed a method based on the kernel estimation of probability density function. This method was used for the investigation of aligned sequences near start of transcription of eukaryotic genes (Tikunov, Kel, 2000). Here we present the adaptation of this method for the analysis of unaligned nucleotide sequences. The application of developed method for binding sites of transcription factors showed that it detects the conservative motifs more sensitively than program CONSENSUS-V6C.1 (Hertz, Stormo, 1999) or Gibbs sampling (Lawrence et al., 1993).

Kernel model for sequences of symbols

The basic intuition of our approach is the notion of probability density. Let us have a some sample $\mathbf{\Omega^*}$ of sequences ω ($\omega \in \mathbf{\Omega^*}$) with length m. The size of sample is n. Let us assign a weight $w(\omega)$ for every sequence ω from the sample $\mathbf{\Omega^*}$, like it is done in the strategy of Gibbs sampling. We call this weight function $w(\omega)$ as weight kernel. The sum S_n of all sequence weights from the sample may be considered as an amount of outcomes of the kernel w. (see 1a). We can also define a measure of volume V for kernel w. (see 1b). Thus the quotient of these two measures will be average density Φ_n of sequences in the kernel w:

$$S_n = \sum_{\omega \in \Omega^*} w(\omega) \qquad (1a)$$

$$V = \left(\sum_{\omega \in \mathbf{\Omega}} w^{1+h}(\omega) \right)^{1/(1+h)} \qquad (1b)$$

$$\Phi_n = \frac{S_n}{V} \qquad (1c)$$

The proposed functional of density Φ_n is remarkable because it runs up to maximum when $w^h(\omega)=c \cdot f(\omega)$, where $f(\omega)$ is the frequency of sequence ω; c is normalization factor. So we can reconstruct the probability space with the help of this functional by simply looking for the maximum. Under $h = 0$ the functional Φ_n is similar to Parsen-Rosenblatt kernel estimation of probability density which is often used in mathematical statistics for reconstruction of probability density. The proposed approach allows to reconstruct the probabilities and densities taking into consideration the more expected random

events with higher weights. Smoothing parameter h regulates the extent of weight deference for events with different expected probabilities. Using proposed functional Φ_n we can look for the best probability density function f(ω) suitable not for all space **Ω** but only for some its local compact part.

Let some consensus s_c is represented with a weight matrix $\| f_{jl} \|$. We propose that near the consensus the probability distribution of sequences s is described with this matrix in accordance with independent distribution of nucleotides in different position

$$f_s = \prod_{j=1}^{m} f_{j l_j^s} \tag{2}$$

f_s is the frequency of sequence s; f_{jl} is the frequency of letter l in position j (actually the elements of weight matrix $\| f_{jl} \|$); l^s_j is the letter of sequence s in position j. Weight matrix $\| fj_l \|$ defines consensus s_c built with the most probable letters of every position. One should define the weight kernel $w_s = (c \cdot f_s)^{1/h}$, as follows from above. Let us put the value of normalization factor equal to $1/ f_c$, where f_c is the frequency of consensus sequence. So the weight of consensus sequence is equal to 1 and the weights of others are equal or less than 1. Basing on the described above relations we can build the next equation system:

$$\begin{aligned} \gamma_{jl} &= \ln\left(\frac{f_{j l_j^c}}{f_{jl}} \right) \\ R_s &= \sum_{j=1}^{m} \gamma_{j l_j^s} \\ f_s &= e^{-R_s} \\ w_s &= e^{-R_s / h} \end{aligned} \tag{3}$$

here, l^c_j is the letter of consensus sequence in position j; R_s may be considered as a distance of sequence s to consensus, and γ_{jl} are the distance coefficients. The more the distance of the sequence to the consensus the less probability is assigned to this sequence in the model of given consensus. Smoothing parameter h regulates the dependence of sequence weights from the distance. Functional Φ_n reaches maximum when the next equation applies

$$f_{jl} = \left(c_f \cdot \sum_{\omega \in \Omega^*_{jl}}^{m} e^{-R_\omega / h} \right) / w_{jl} \tag{4}$$

$\mathbf{\Omega}^*_{jl}$ denotes the subset of set **Ω*** in which the position (j) is occupied with letter (l); c_f is normalization factor ensuring equality $\sum_l f_{jl} = 1$.

Algorithm

The local weight matrix is calculated on the basis of subsequences of length m picked up from the sample **Ω*** (one subsequence from each sequence). The present algorithm is initialized by choosing a random starting position within various sequences. There are three main steps in algorithm which proceed through many iterations.

STEP 1: The weight kernel matrix $\| f_{jl} \|$, distance coefficients $\| \gamma_{jl} \|$, distances R_s for every word from the sample are calculated in accordance with system (3) on the basis of the weight matrix $\| f_{jl} \|$.

STEP 2: One word s closest to consensus is picked up from every sequence of sample. So a matrix n×m of nucleotides is built.

STEP 3: The counts f_{jl} are calculated on the basis of picked distances R_s in accordance with equation (4).

The result of this algorithm is a weight matrix $\| f_{jl} \|$ that describes the distribution of picked sequences near consensus. Additional local shifting of the window along the sequences is performed to ensure of finding the maximum of Φ_n.

The result of one execution of this algorithm is a single solution (a weight matrix) which is compared with previously saved solutions. If it has no analogues among them we save the obtained solution. After making a reasonable number of executions we obtain all possible solutions. The set of patterns is outputted.

On the next step we perform classification of all subsequences of the set using the revealed set of patterns. Fore a selected consensus weight matrix all subsequences in the set may be classified as fitting to it or not. We build a mixture of probability functions for the consensus weight matrix and background weight matrix $P=k_cP_c+k_bP_b$. Every subsequence is classified in accordance with most probable estimated event to fit into the consensus weight matrix or to the background weight matrix. The background weight matrix and appropriate coefficients (k_c and k_b) are calculated by maximization of likelihood function $L = \sum_{s \in \Omega^*} P_s$.

Results

We take a mixture of binding sites for AP-1 and CREB transcription factors from database TRANSFAC. It is known that often these two different families of transcription factors binds to the same sites. The analyzed sample contains 155 sequences. Every sequence contains the TF binding site in the center (as it was described in the original paper) and add 10 nucleotide flans to both side of the site. We took into consideration both strands of nucleotide sequences: direct and complement.

We have applied the our program and have revealed two different patterns (Table 1). Analysis showed that the overwhelming majority of sequences contain a pattern of the length 7 that corresponds to the consensus: "**TGAGTCA**". The second pattern has the length 8 and corresponds to the consensus: "**TGACGTCA**" , which differs from the first one by insertion of letter "**C**" in the forth position (see Tab.1). It is important to mention that exactly these two patterns correspond to known consensi of AP-1 site (the first pattern) and CREB site (second pattern). Classification of sites shows that some binding sites contain both of these motifs located in different places not far one from other.

We have applied two other programs: CONSENSUS-V6C.1 and Gibbs sampling to the same set of sites. Both programs were not able to reveal two different patterns. Only one pattern was revealed that presents a mixture of the original two (Table 2, 3).

Table 1. Weight matrices revealed with kernel method (smoothing parameter h = 1.2).

Weight matrix 1 (113 sequences contain this motif)								
A	15	18	51	0	0	4	100	6
G	14	55	4	84	0	2	0	36
C	6	5	34	1	5	94	0	27
T	65	22	11	15	95		0	31
Consensus	T	G	A	G	T	C	A	
Weight matrix 2 (73 sequences contain this motif)								
A	12	1	75	7	23	0	2	100
G	8	77	10	5	62	0	0	0
C	5	3	5	88	15	3	98	0
T	75	19	10	0	0	97	0	0
Consensus	T	G	A	C	G	T	C	A

Table 2. The most optimal weight matrix (153 sampled words) resulted from run of program CONSENSUS-V6C.1 (Hertz, Stormo, 1999).

A	9	39	37	0	0	5	100	5
G	29	40	11	87	0	0	0	26
C	8	4	40	8	0	95	0	27
T	54	17	12	5	100	0	0	42
Consensus	t	g/a	c/a	g	t	c	a	T

Table 3. Weight matrix obtained with Gibbs sampling (Lawrence et al., 1993).

A		36	43				93	
G		42		82				
C			36			85		
T	52				92			39
Consensus	t	g/a	c/a	g	t	c	a	T

The high sensitivity of our method results from the fact that the estimation of sequence distribution probabilities is mainly built on the basis of sequences located near the consensus. Smoothing parameter h can regulate the compactness of analyzable space. In extreme case $n \to \infty$ and $h \to 0$ we can get exact local maximums. Gibbs sampling method resembles ours with h=1. The obtained results show that developed kernel method may be used with success for identification of respectively close imperfect motifs in the nucleotide sequences. The general approach presented at this paper can be applied for a wide range of problems.

Acknowledgements

Parts of this work was supported by Siberian Branch of Russian Academy of Sciences, by grant of Volkswagen-Stiftung (I/75941).

References

1. Hertz G.Z., Stormo G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 15, 563-77.
2. Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., Wootton J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 262, 208-14.
3. Tikunov Y., Kel A. Kernel method for estimation of functional site local consensi. Classification of transcription initiation sites in eukaryotic genes (2000). Proc. of the German Conf. on Bioinformatics (GCB00), October 5-7, 2000, Heidelberg, 83-88.

A GIBBS SAMPLING ALGORITHM TO DETECT CLUSTERED CIS-ELEMENTS

Frith M.C., Hansen U., * Weng Z.

Boston University, Boston, USA, e-mail: zhiping@bu.edu
*Corresponding author

Key words: *gibbs sampling, cis-element, enhancer, simulated annealing*

Resume

Motivation: The DNA sequence binding preferences of individual transcription factors are generally not informative enough for detection of their binding sites (cis-elements) in large DNA sequences. However, in higher eukaryotes cis-elements are observed to occur in clusters in the sequence, and multiple transcription factors typically bind them cooperatively to form an "enhanceosome". There has been increasing recent interest in computational detection of such clusters (Frith et al., 2001; Berman et al., 2002), but the main bottleneck in these approaches is the lack of knowledge of which cis-elements tend to cluster with one another.

Results: We have implemented a novel Gibbs sampling technique to detect unknown cis-elements that cluster with a given sample of known cis-elements of one type. The program was tested on a range of increasingly difficult problem scenarios. In cases where it fails, we dissect which aspect of the technique (objective function or search algorithm) is the cause of failure. When the algorithm is allowed to make several alternative predictions, it succeeds in most test cases.

Availability: The program is available from the authors on request.

Introduction

This work addresses the following problem: given a sample of experimentally verified cis-elements of one type ("known sites"), and their flanking sequences, can we identify other cis-element types ("helper cis-elements") that tend to cluster with the known one? A previous study has examined degenerate oligonucleotides that occur in the context of E2F binding sites (Kel et al., 2001). We feel that a matrix-based approach such as Gibbs sampling (Lawrence et al., 1993) possesses advantages. Matrices provide a more general model of cis-elements than do oligonucleotides, and while the oligonucleotides discovered in the earlier paper may be diagnostic of E2F sites, they seem too short to represent typical factor binding sites. The basic Gibbs sampling technique can be modified in many ways (various priors, site sampling vs. motif sampling, etc.) to optimize it for the specific problem in hand. Here, we describe a novel version of Gibbs sampling that is targeted to our problem.

Method

Any Gibbs sampling method requires definition of three basic ingredients: the search space of possible answers to the problem (i.e. possible locations of helper cis-elements), an objective function that assigns scores to these answers, and a search algorithm to traverse the search space. Our method finds one type of helper cis-element at a time, and assumes that each sequence contains zero or one helper. Although a more general algorithm would consider multiple helpers per sequence, it would pay the penalty of an increase in search space size. The objective function is a Bayesian posterior probability of helper cis-element locations given the sequence data, defined as follows:

$$\Pr(\text{locations} \mid \text{data}) \propto \Pr(\text{data} \mid \text{locations}) \times \text{Prior}(\text{locations}) \tag{0.1}$$

$$\Pr(\text{data} \mid \text{locations}) = \Pr(\text{helper sequences}) \times \Pr(\text{background DNA}) \tag{0.2}$$

Background DNA is modeled by a 5th order Markov chain trained on human chromosome 20, and Pr(helper sequences) is defined similarly to previous Gibbs samplers:

$$\Pr(\text{helper sequences}) = \int_{\{p_{ik}\}} \prod_{k,i} \left(p_{ik}^{\,c_{ik}}\right) \times \text{Prior}\{p_{ik}\} = \prod_k \left[\frac{6}{(H+3)!} \prod_i (c_{ik}!)\right] \tag{0.3}$$

Where c_{ik} is the count of nucleotide i in column k of the aligned helper sites, H is the number of sequences that contain a helper, and we have assumed a uniform prior for the nucleotide emission probabilities p_{ik}. We use a novel prior probability for helper locations, incorporating a geometric decay with increasing distance of the helpers from the known sites:

$$\text{Prior}(\text{locations}) \quad \propto \quad h^{H}(1-h)^{S-H} \times \prod_{s} \frac{1}{4}(1-g)\,g^{d_s-1} \tag{0.4}$$

Where S is the total number of sequences and d_s is the distance of the helper cis-element from the known cis-element in sequence s. The product is taken over sequences that contain a helper. h and g are tunable parameters. We refer to h as "helper probability", and we specify the geometric parameter g in terms of a half life: g = exp(1 / (half life – 1) * ln(0.5)). For the search algorithm we use simulated annealing rather than sampling.

Results

The method was tested using a set of 22 experimentally verified, mammalian estrogen response elements (EREs). We obtained flanking sequence of 500 bp either side of the ERE (except in a few cases where the full flanking sequence was not available). Since these regulatory regions have not been sufficiently well characterized to obtain a test set of helper cis-elements, we specified a fictitious "known cis-element" to occur at a random location in each sequence, and examined whether our algorithm was able to find the EREs. The difficulty of this test was varied in two ways. First, the fictitious "known cis-elements" were placed at geometrically distributed distances from the EREs, with half life varying from 10 bp in the easiest case, to 100 bp in the hardest. Second, a proportion of the sequences were replaced by other promoter sequences not thought to contain EREs, so that the fraction of sequences with an ERE ranged from 1 to 0.6. For each test case, we applied the Gibbs sampler and we counted the number of sequences for which it made a correct prediction: either locating the ERE correctly to within ±2 bp, or correctly predicting the absence of an ERE. We searched for helper sites of width 15 bp, but the results do not change substantially when this parameter is adjusted to 12 or 18. The results are shown in Figure 1. The horizontal axes show the half life and helper probability used in constructing the test set, whereas the values of these parameters given to the algorithm (Equation (0.4)) were fixed at 50 and 0.7 respectively. In these tests, the performance declines dramatically when more than 10% of the sequences do not contain an ERE. However, in similar tests where the EREs were embedded in synthetic DNA generated from a Markov model, most of the sequences received correct predictions under nearly all test conditions (Fig. 2). When the algorithm fails to find the known EREs, the predicted sites do not resemble low complexity sequence, and one interpretation, to be tested in future studies, is that these are genuine cis-elements, i.e. helpers of ERE.

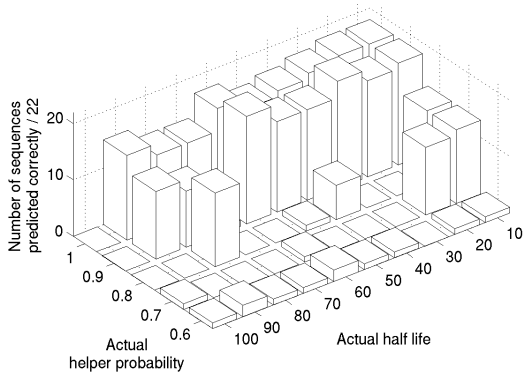


Fig. 1.

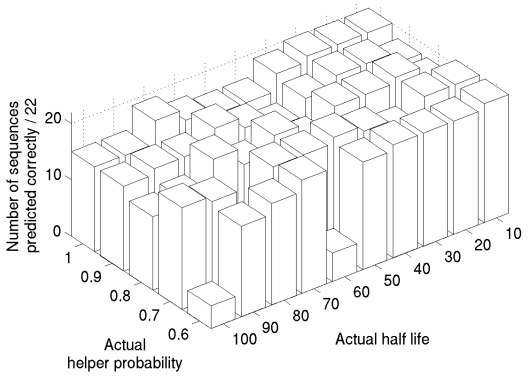


Fig. 2.

It is instructive to find out whether mispredictions are caused by the objective function or the search algorithm. To test this, we initialized the Gibbs sampler with the correct positions of the EREs, and conducted a few sampling iterations to reach a local maximum of the objective function. If this value were higher than any found from random starting positions, it would be clear that the search algorithm beginning from random starts has failed to find the optimum. However, we rarely observed this phenomenon (3 cases out of 50 for the natural sequence set), suggesting that the search algorithm performs well, but that the optimal cis-element configuration frequently does not match the known EREs.

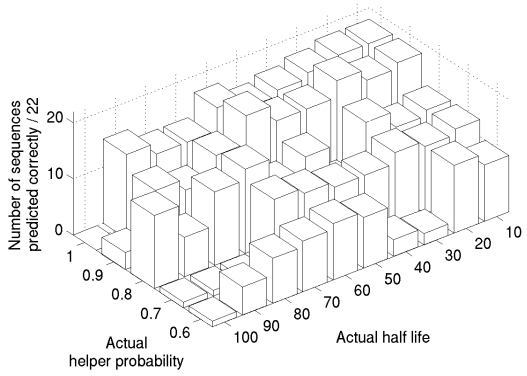


Fig. 3.

Since failure to find the EREs may be caused by detection of other cis-elements, we performed an iterative masking procedure, where a suboptimal cis-element configuration was searched for after masking the optimal one. Figure 3 shows, for each test condition, the best result among 6 such rounds of Gibbs sampling. Some correct predictions are now made for nearly all test cases.

Finally, we wished to assess whether the algorithm's predictions are sufficient to lead to successful experiments. We supposed that, given a prediction of helper locations, the experimenter would choose to mutate the one predicted helper that most strongly matched the consensus of all the others. How frequently is this cis-element correct (±2 bp)? If the experimenter is willing to test 6 alternative sites, the strongest helper could be chosen from each successive round of prediction. Among the test cases shown in Figure 3, a correct cis-element would be chosen in all cases where more than 3 sequences receive correct predictions, i.e. 43 cases out of 50.

Discussion

For the problem of detecting cis-elements that cluster with one another, we have both developed a Gibbs sampling approach and analyzed its performance for a range of relevant test cases. The method is capable of correct predictions in a wide range of scenarios. In contrast, we were unable to recover the EREs using a Gibbs sampler designed for other purposes (AlignACE (Roth et al., 1998)). We are currently working on two desirable additions to the algorithm: automatic detection of cis-element width, and a measure of significance of the prediction.

Acknowledgements

We thank R.O'Lone for assistance in obtaining the ERE data, and the research group of J.S.Liu for helpful comments. Martin Frith is a Howard Hughes Medical Institute Predoctoral Fellow.

References

1. Berman B.P., Nibu Y., Pfeiffer B.D., Tomancak P., Celniker S.E., Levine M., Rubin G.M., Eisen M.B. (2002) Exploiting transcription factor binding site clustering to identify cis- regulatory modules involved in pattern formation in the *Drosophila* genome. Proc. Natl Acad. Sci. USA. 99, 757-762.
2. Frith M.C., Hansen U., Weng Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics. 17, 878-889.
3. Kel A.E., Kel-Margoulis O.V., Farnham P.J., Bartley S.M., Wingender E., Zhang M.Q. (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. J. Mol. Biol. 309, 99-120.
4. Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., Wootton J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 262, 208-214.
5. Roth F.P., Hughes J.D., Estep P.W., Church G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol. 16, 939-945.

ARGO_VIEWER: A SYSTEM FOR RECOGNITION AND ANALYSIS OF GENE REGULATORY ELEMENTS IN EUKARYOTES

[1]* Vishnevsky O.V., [1] Ananko E.A., [1] Ignatieva E.V., [1] Podkolodnaya O.A., [1] Stepanenko I.L., [2] Vityaev E.E.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
e-mail: oleg@bionet.nsc.ru
*Corresponding author

Key words: *the structure and function of regulatory gene regions in eukaryotes, promoter recognition in eukaryotes, oligonucleotide vocabularies*

Resume

Motivation: Recognition of promoter gene regions in extended genome sequences is still an intriguing task. The regularities in structure and function of promoter gene regions that provide high specificity of gene expression call for further investigation too.

Results: We have analyzed five groups of promoters of tissue-specific genes. As a result, vocabularies of region-specific oligonucleotide motifs were developed. Based on the presence and distribution of particular oligonucleotide motifs, we suggest a procedure for recognition of tissue-specific gene promoters.

Introduction

Promoter regions refer to the most important regulatory elements that determine the level of gene expression. Assemblage of basal transcription complex, as well as tissue-and stage-specific peculiarities of eukaryotic gene transcription are dependent upon organization of the context- and structure of the core promoter and upon the presence in the 5'-regulatory region of transcription factor binding sites (TFBSs) (Ignatieva et al., 1997).

Most of modern approaches aimed at promoter recognition are based on searching for potential TFBS by means of weight matrices, consensuses, or other methods, which account for both representability of TFBSs (Prestrige, 1995) and their distribution along promoters (Kondrakhin et al., 1995; Fickett, Hatzigeorgiou, 1997; Werner, 1999). As shown, accounting for the TFBS's localization in promoter has increased an accuracy of promoter recognition. However, despite the huge variety of approaches suggested, promoter recognition on the basis of only TFBS analysis does not provide the accuracy necessary for recognition of these regulatory regions (Fickett, Hatzigeorgiou, 1997).

In this connection, some alternative approaches that need no information about TFBSs were suggested for promoter recognition (Solovyev, Salamov, 1997). These approaches are based on analysis of the oligonucleotide content of promoters and on accounting for a pattern of oligonucleotide distribution along promoter region. As was demonstrated, by taking into account regularities of oligonucleotide distribution along promoter, one may effectively increase an accuracy of promoter recognition (Zhang, 1998). Besides, the data are accumulated that give evidence about some similar properties in organization of promoters of genes with similar patterns of expression (e.g., promoters of genes expressing in a definite tissue). In particular, this fact is supported by similarity of the sets of TFBSs in promoters of functional groups of genes (Ignatieva et al., 1997; Podkolodnaya, Stepanenko, 1997; Anan'ko et al., 1997). Development of methods aimed at recognition of specific groups of promoters regulating transcription of genes with similar patterns of expression and, thus, with similar context organization is viewed as a perspective way to increase accuracy of recognition of these regulatory regions.

Methods and Algorithms

As an object of analysis, we have used 5 samples of tissue-specific gene promoter sequences extracted from the database on transcription regulatory regions, TRRD (http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4). The sequences were positioned within the region from –300 to +100 bp relatively transcription start. As a negative set of sequences, we have used 5 samples, of 1000 nucleotide sequences of length 400 bp, which were randomly generated with conservation of mononucleotide frequencies typical for respective types of promoter regions.

Complete sets of promoter sequences were divided into the training and control ones. The control sets were compiled from 20% of randomly selected tissue-specific promoter sequences. The training sets included the portion of the rest 80% of promoters. Evaluation of the false positive estimates was made on the control sets, whereas false negatives were calculated for the randomly generated sets of sequences of 100000 bp in length, with coincidence of mononucleotide frequencies.

The search for the region-specific significant oligonucleotide motifs in the sets of promoter sequences was made by the software package ARGO (http://wwwmgs.bionet.nsc.ru/mgs/programs/argo) (Vishnevsky, Vityaev, 2001). The method of searching for degenerate oligonucleotide motifs is based on viewing oligonucleotide vocabularies of each promoter with subsequent clusterization of similar perfect oligonucleotides, which are present in vocabularies of different promoters. In case the number of non-coincident letters between oligonucleotides of different promoters does not exceed the limiting value, these oligonucleotides are unified in a single group. Then, by iteration procedure, we form a consensus for such group of oligonucleotides. This consensus consists of the most significant letters of the 15-lettered IUEPAC code. Significance of each letter was evaluated by the binomial criterion:

$$W(k,K,x_i) = \sum_{i=k}^{K} C_K^i P^i(x_i)(1-P(x_i))^{K-i} ,$$

where $W(k,K,x_i)$ is a probability to observe the given letter x_i randomly in at least k out of K oligonucleotides of the group considered; $P(x_i)$ is a frequency of occurrence of the letter x_i from the 15-lettered alphabet in promoter sequences.

Oligonucleotide motif obtained by such a procedure is considered to be significant, if it satisfies to criterions given below:

$$\left.\begin{array}{l} a)\ F > f_0 \\ b)\ P(n,N) < p_0 \\ c)\ Q < q_0 \end{array}\right\} ,$$

where F is a share of promoters containing the motif considered;

f_0 is a threshold value of representability of a motif in a set of promoters;

P(n,N) is a probability to observe at random the motif in a given window in at least n out of N sequences;

p_0 is a threshold probability level;

Q is a share of sequences in a negative set, which contain the motif given;

q_0 is a threshold level of representability of a motif in a negative set.

Binomial probability P(n,N) is calculated as given below.

Let us consider an oligonucleotide motif $M=m_1,m_2,..,m_l$, of length l in extended 15-lettered IUEPAC code. The probability to observe the motif considered in some position of the sequence of the length L is estimated as:

$$P(M) = \prod_{i=1}^{l} P_i ,$$

where P_i is the frequency of the letter m_i calculated in accordance with the mononucleotide content of promoters.

Since P(M), as a rule, is sufficiently low, then the probability to observe this motif in a sequence considered at least once could be approximated by the Poisson distribution:

$$P = 1 - e^{-(L-l+1)*P(M)} .$$

The binomial probability $P(n,N)$ to observe the motif M in at least n out of N sequences equals to

$$P(n,N) = \sum_{i=n}^{N} C_N^i P^i (1-P)^{N-i} .$$

Based on the set R, of region-specific motifs, found in a way given above, by means of the software program ARGO_VIEWER (Vishnevsky, Vityaev, 2001) (http://wwwmgs.bionet.nsc.ru/mgs/programs/argo/argo_viewer.html), it is possible to recognize promoters of the type under study within extended genome sequences. An idea of the method is that at the first stage of analysis of each promoter sequence entering the training set, all specific motifs from the set R are found. For recognition of promoter within an arbitrary sequence, a comparison is made of the patterns of specific oligonucleotide motifs, with specific features of their representability and localization in promoters. In case there exists a promoter, which has a statistical similarity by these features to the sequence analyzed, then the solution is accepted that the sequence of interest is really a promoter.

Results and Discussion

The results of application of the methods described above to analysis of 5 samples of tissue-specific gene sequences are given in Table.

As could be seen from the Table, in four promoter groups out of five, the overprediction rate equals or less than 1 per 100000 bp, with the false negative equaling to 4-8%%. Promoters of different groups are characterized by different number of motifs detected. Notably, this characteristic depends not only upon the size of the training set. For example, in the group of Heat Shock-induced genes, which is compiled of 34 sequences, we have found only 45 motifs, whereas in the group of Erythroid-specific regulated genes, represented only by 26 sequences, 78 motifs were detected. Besides, the motifs found in promoters of the Heat Shock-induced genes are characterized by relatively low recognition ability in

comparison to the other groups of genes. Possibly, this result could be explained by the less similarity between these promoters, as well as by the fact that Heat Shock element binding site responsible for functional promoter specificity, may be compiled out of several (from 1 to 9) identical subunits, located on direct or complementary DNA chains, with varying distances between them. Such signals are difficult to describe by degenerous motifs of the ordered length, thus, giving rise to false positive and false negative estimates.

Table. Analysis and recognition of tissue-specific promoters in 5 groups of genes.

Promoters of tissue – specific genes	Number of sequences	Number of motifs obtained	Errors of recognition	
			False negative	False positive
Heat Shock-induced genes	34	45	0.09	$\sim 10^{-4}$
Interferon-inducible genes	41	131	0.07	$<10^{-5}$
Erythroid-specific regulated genes	26	78	0.08	$\sim 10^{-5}$
Genes of Lipid metabolism	50	281	0.04	$<10^{-5}$
Endocrine system genes	78	814	0.05	$<10^{-5}$

Acknowledgements

The authors are grateful to Professor Nikolay Kolchanov for the fruitful discussions.

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 07-90337, 00-02-07-90355, 00-04-49229, 00-04-49255), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), Siberian Branch of the Russian Academy of Sciences (Integration Project № 65), National Institutes of Health Grant № 2 R01-HG-01539-04A2, Department of Energy Subgrant № 535228 CFDA 81.049.

References

1. Ignatieva E.V., Merkulova T.I., Vishnevskii O.V., Kel A.E. Mol. Biol. (Mosc.). 1997, 31, 684-700.
2. Prestrige D.S. J. Mol. Biol. 1995, 249, 923-932.
3. Kondrakhin Y.V., Kel A.E., Kolchanov N.A., Romashchenko A.G., Milanesi L. Comp. Appl. Bioscien. 1995, 11, 477-488.
4. Fickett J.W., Hatzigeorgiou A.C. Genome Res. 1997, 7, 861-878.
5. Werner T. Mammal. Genome. 1999, 10, 168-175.
6. Solovyev V., Salamov A. Proc. Fifth Intern. Conf. on Intelligent Systems for Molecular Biology (ISMB-97), 1997, 294-302.
7. Zhang M.Q. Genome Res. 1998, 8, 319-326.
8. Podkolodnaya O.A., Stepanenko I.L. Mol. Biol. (Mosc.). 1997, 31, 671-683.
9. Anan'ko E.A., Bazhan E.A., Belova O.E., Kel A.E. Mol. Biol. (Mosc.). 1997, 31, 592-605.
10. Vishnevsky O.V., Vityaev E.E. Mol. Biol. (Mosc.). 2001, 35, 1-9.

PREDICTION OF POTENTIAL C/EBP/NF-κB COMPOSITE ELEMENTS USING THE MATRIX-BASED SEARCH METHODS

[1,3] Shelest E., [1] Kel A.E., [1] Gößling E., [1, 2]* Wingender E.

[1] GBF German Research Centre for Biotechnology, Mascheroder Weg 1, D-38124 Braunschweig, Germany
[2] BIOBASE GmbH, D-38304 Wolfenbüttel, Germany
[3] permanent address: Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia
e-mail: ewi@gbf.de
*Corresponding author

Key words: *promoters, regulatory sequence signals, transcription factors, transcription factor binding sites, composite elements, antibacterial response*

Resume

Motivation: Bacterial infections trigger a wide range of host cell responses. For the interaction of *Pseudomonas aeruginosa* and epithelial cells it is known that transcription factor NF-κB plays a central role, but its effects have to be specified by cooperation with additional factors. NF-κB containing composite element, e.g. with C/EBP, may be appropriate indicators for new antibacterial response genes.

Results: We refined matrix-based search methods for C/EBP, established a model for C/EBP/ NF-κB composite element, used it for scanning all known human 5'-flanking sequences and identified 135 new candidate genes.

Availability: The newly constructed C/EBP binding patterns will be available with the next release of the TRANSFAC database (http://www.gene-regulation.de).

Introduction

Binding of a bacteria to a eukaryotic cell triggers a complex network of interactions in and between both cells. *P. aeruginosa* is a pathogen that causes acute and chronic lung infections by interacting with the pulmonary epithelial cells (DiMango et al., 1998; Smith et al., 2001). We use this example for understanding the ways of triggering the eukaryotic cell(s) response, leading us to understanding the details of the inflammatory process in general. After adhesion of *P.aeruginosa* to the epithelial cells, the response of these cells is triggered by at least two distinct agents: bacterial lipopolysaccharides or bacterial pilins or flaggelins. Together with well-known signal transduction pathways through TLR4 (Zhang, Ghosh, 2001) to NF-κB and to gene activation, a different pathway through asialoGM1 receptor is also triggered (McNamara et al., 2001) leading to an increase of the intracellular calcium concentration which, in turn, leads again to activation of NF-κB. The latter pathway is not yet well understood.

The consequences of NF-κB activation seem to be multiple and diverse, but the response in every certain case is rather specific. That may mean that NF-κB is not the only factor responsible for the reply. Therefore, we searched for additional transcriptional factors that may cooperate with NF-κB and/or complement its effects. Identifying them would give a clue to realizing the mechanisms of their activation, in particular those connected with [Ca^{2+}] alteration. The information about the participating transcription factors would enable us to construct a model for searching other, still not identified, target genes which are potentially involved in defensive mechanisms.

Methods and Algorithms

Databases on transcription regulation): The databases used for this study are the TRANSFAC® (release 5.4, December 2001) and TRANSCompel™ (release 5.4, December 2001). TRANSFAC® is a database on transcription factors (TF), their expression patterns, genomic and artificial binding sites as well as positional weight matrices for many of the TF stored in the database (Wingender et al., 2001). TRANSCompel™ is a database on composite elements (Kel-Margoulis et al., 2000).

Subtractive approach of matrix generation: The whole set of sequences containing transcription factor binding sites has been put into GIBBS Motif Sampler for DNA, searching for motifs of 8-10 nucleotides length. The sequences with the motifs were aligned and put to «Matrix generation» subroutine of Match™ (Goessling et al., 2001; http://www.gene-regulation.de). Then the sequences used for the generation of the first matrix are subtracted from the whole set, and this restricted set is used again for the motif searching. The procedure was repeated until 90% of all true positive sites in the set were used for contructing a matrix.

Sequence analysis: A sequence set of putative promoter regions (-300/+100) was derived from the TRANSGENOME information resource of annotated human genome features. The reference sequence of TRANSGENOME is currently based on RefSeq as a data source for the nucleotide sequence and the NCBI annotation results, including gene models and SNP

information. For the analysis of these sequences we used the Match tool. Match™ is a weight matrix-based tool for searching putative transcription factor binding sites (TFS) in DNA sequences. In particular, Match™ uses the matrix library collected in TRANSFAC®. Match™ has been adapted for the analysis of long sequences and for the batch-wise analysis of large sequence sets. First we made profiles containing the matrices needed for the analysis and searched in the set of putative promoter regions with this profile using the adapted variant of the Match™ tool. On the next step we parsed the output for those pairs of C/EBP and NF-κB sites which match a defined model representing the relative orientation, distance and scoring of the constituents.

Implementation and Results

NF-κB often appears to be a part of composite elements (CE). Composite elements are combinations of two or more transcription factor (TF) binding sites which provide synergistic action of the TFs, qualitatively different from a purely additive effect. Screening the TRANSCompel database for CEs with NF-κB, we found that the most abundant CEs were those containing NF-κB together with C/EBP. Among 322 CEs documented in TRANSCompel (release 6.1) 64 contain a NF-κB moiety and 16 of them are of NF-κB/C/EBP type. C/EBP is an interesting TF in this context because of this and two further reasons: first, it is known to participate in immune response; second, it is one of target factors in the pathways triggered by increase of $[Ca^{2+}]$, which mediates the asialo-GM-dependent response to *P. aeruginosa* invasion.

To use the weight matrix approach we had first to re-evaluate the matrices for C/EBP binding sites. There are 350 entries for C/EBP binding site in TRANSFAC (release 6.1) and 8 matrices for them, but all of them exhibit rather weak consensi. We decided to make more precise matrices for C/EBP dividing the whole set into subgroups using the subtractive approach. We came up with a set of matrices, each of them searched for a subset of C/EBP sites. Comparison of the consensi derived from these matrices shows that they represent distinct sequence patterns.

Table 1. Comparison of the consensus sequences for C/EBP binding sites derived from positional weight matrices. CEBPB_01 is a previously defined matrix; CEBP_comp is the matrix derived from C/EBP/NF-κB CE, the others – those derived with the subtractive approach. Note that CEBP_comp and CEBP_sub1 represent nearly the same consensi.

Name of matrix	Consensus sequence
CEBPB_01 (M00109)	RNRTKDNGMAAKNN
CEBP_comp	SHNVNRTTGCACAA
CEBP_sub1	TTRCACAA
CEBP_sub2	CATTKCSYCAK
CEBP_sub3	NTNASCAAWCA
CEBP_sub4	DGCAGAGGTGAA

To make the search for binding sites more comprehensive we combined these matrices in such a way that for a defined rate of false negatives (FN) a minimal rate of false positives (FP) is achieved, the overlap between individual matrices being minimized. FP is represented here by f_r, the frequency of matches per nucleotide of random sequences (Table 2). As can be seen, the FP rate could be reduced by about two thirds compared to one of the previously used search patterns (CEBP_02).

Table 2. Frequencies of sites per nucleotide found in control set of random sequences, for the 50% threshold of true positives.

Combination	Matrices	f_r ($\times 10^{-3}$)
Single matrices	CEBP_02 (M00117)	1.6
	CEBP_sub1	1.6
	CEBP_sub2	2.4
	CEBP_sub3	4.2
Combination of 2 matrices	CEBP_sub1 CEBP_sub2	1.02
Combination of 3 matrices	CEBP_sub1 CEBP_sub2 CEBP_sub4	0.78
Combination of 4 matrices	CEBP_sub1 CEBP_sub2 CEBP_sub4 CEBP_sub3	0.51

Since we are particularly interested in identifying C/EBP sites within the CEs, we took only the sites contained in the CE together with NF-κB and constructed a specific matrix for this subset. Thorough analysis of the known composite elements of C/EBP / NF-κB type led us to develop a model describing the relative orientation, distance and scoring of the constituents. We used this model to search for potential CEs in the set of human genome promoter regions. The matches of the search indicate potential target genes.

We next analyzed our collection of human 5'-flanking sequences setting the parameters such to re-identify 80% of the true positives (20% FN). Under thesed conditions, we identified about 200 genes as harboring at least one potential C/EBP / NF-

κB composite element. Erasing all those that encode hypothetical products, we end up with a list of 135 genes which can be checked for plausibility. Interestingly, this list contains possibly relevant genes/gene products like calcium-binding proteins, interleukins, transcription factors, TNF receptor, apoptosis-related proteins.

We repeated the search with the combination of matrices described above instead of the one specific matrix constructed from the CE, also in combination with the NF-κB matrix. In this case, we found about twice as many potential targets, although the stringency was set to re-identify 50% of the known binding sites and thus was much lower than described above. Moreover, there was practically no overlap in the results between both search strategies, indicating again that both C/EBP search modes are targeting distinct sets of sites.

Discussion

Binding patterns for some transcription factors are extremely weak for different reasons, one of them may be that the training sets for deriving the patterns are too large and heterogeneous, representing different subgroups of binding factors and/or classes of binding sites. We classified the sequences of C/EBP binding sites in two ways: (i) by a sequence-analysis based subtractive approach and (ii) through a functional classification, the only functional group has been identified being the C/EBP sites of the CEs with NF-κB. It turned out that the first pattern derived from the subtractive approach was nearly the same as that constructed from the functional class of composite elements, thus confirming that the C/EBP sites within CEs represent a homogeneous class. Being specified for a certain class, the pattern for C/EBP within CEs can not be used for a general search for C/EBP binding sites. Therefore, for a comprehensive search other patterns should be provided. Up to now we did not identify functional correlations for the other subgroups of the binding sites derived by subtractive approach, neither in terms of the regulated genes nor of the C/EBP isoforms interacting with these sites (α, β, or others). So the subtractive approach appears to be more general and does not depend on detailed biological knowledge.

When applying our model for CE of C/EBP/NF-κB type to a whole promoter screening, we identified 135 potential new target genes. Among them, 25 seem to be tempting since they could play a role in the network we started our considerations with. Several interleukins are identified in addition to re-identification of IL-8 and IL-6 which have been reported before as having the CE of this type. Some calcium-binding proteins seem to be of particular interest because they could be a part of the Ca^{2+}-dependent pathway(s) of the network.

On the other hand, it is quite obvious that the list of the potential target genes should be further specified with adding some other transcription factors to the model. As a future perspective, we plan to develop such an enriched model, which would allow to predict the target genes as well as understand more about the details of the regulatory network of the epithelial cells' response to bacterial infection.

Acknowledgement

We want to thank O.V.Kel-Margoulis for fruitful discussions and Ingmar Reuter for technical support. Part of this work was financed by a grant from the German Federal Ministry of Education and Research (grant № 031U110A).

References

1. DiMango E., Ratner A.J., Bryan R., Tabibi S., Prince A. (1998) Activation of NF-κB by adherent Pseudomonas aeruginosa in normal and cystic fibrosis Respuratory epithelial cells. J. Clin. Invest. 101, 11, 2598-2606.
2. Goessling E., Kel-Margoulis O.V., Kel A.E., Wingender E. (2001) MATCH™ - a tool for searching transcription factor binding sites in DNA sequences. Application for the analysis of human chromosomes. Proc. of the German Conf. on Bioinformatics GCB 2001, Wingender E., Hofestädt R., Liebich I. (eds.), Braunschweig, 158-161.
3. Kel-Margoulis O.V., Romashchenko A.G., Kolchanov N.A., Wingender E., Kel A.E. (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucl. Acids Res. 28, 311-315.
4. McNamara N., Khong A., McKemy D., Caterina M., BoyerJ., Julius D., Basbaum C. (2001) ATP transduces signals from ASGM1, a glycolipid that functions as a bacterial receptor. Proc. Natl Acad. Sci. USA. 98, Issue 16, 9086-9091.
5. Smith R.S., Fedyk E.R., Springer T.A., Mukaida N., Iglewski B.H., Phipps R.P.(2001) IL-8 production in human lung fibroblasts and epithelial cells activated by the Pseudomonas aeruginosa autoinducer N-3-oxodododecanoyl homoserine lactone is transcriptionally regulated by NF-κB and activator protein-2. J. of Immunology. 167, 366-374.
6. Wingender E., Chen X., Fricke E., Geffers R., Hehl R., Liebich I., Krull M., Matys V., Michael H., Ohnhaeuser R., Prueß M., Schacherer F., Thiele S., Urbach S. (2001) The TRANSFAC system on gene expression regulation. Nucl. Acids Res. 29, 281-283.
7. Zhang G., Ghosh S. (2001) Toll-like receptor–mediated NF-B activation: a phylogenetically conserved paradigm in innate immunity. J. Clin. Invest. 107, 1, 13-19.

ANALYSIS OF THE REGULATORY REGIONS OF GENES INVOLVED IN THE IMMUNE SYSTEM OPERATION

**** Ananko E.A., Oshchepkov D.Yu., Levitsky V.G., Pozdnyakov M.A.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: eananko@bionet.nsc.ru
*Corresponding author

Key words: *interferon-regulated genes, macrophage expressed genes, transcription regulation, transcription factor binding sites, site recognition*

Resume

Motivation: Optimal operation of the immune system depends on cooperative interaction of various cells, such as T and B lymphocytes, macrophages, etc. It is well known that cytokines, such as interferons and interleukins secreted by immune cells, play an important role in these processes. To control the function of the immune system, it is necessary to study molecular genetic mechanisms regulating the protective activity of the body.

Results: Samples of interferon-regulated genes and genes expressed in macrophages were selected from the TRRD database. The information on regulation of these genes by various transcription factors was analyzed. On the basis of samples of transcription factor binding sites that play a key role in the regulation of the genes of NF-κB**, IRF, ISGF3, and STAT, two methods for recognizing these sites were developed.

Availability: The sample of interferon-regulated genes is available at http://wwmgs.bionet.nsc.ru/mgs/papers/ananko/iig-trrd/. The sample of genes expressed in macrophages is available at http://wwwmgs.bionet.nsc.ru/mgs/papers/ananko/macroph/.

Introduction

Macrophages playing an essential role in the immune system perform various functions, including the synthesis of cytokines, which ensure coordinate activities of immune cells. Interferons are key regulators not only of immune and anti-inflammatory responses, but also of the processes such as cell proliferation and differentiation as well as tumor growth and development. Interferons become involved into these processes by stimulating the transcription of IFN-inducible genes. At present, molecular mechanisms of transcription regulation of these genes are extensively studied. The TRRD database (Kolchanov et al., 2002) stores information on regulatory regions and specific features of expression of 119 interferon-inducible genes. It is known that genes are induced by interferons through the transcription factors IRF, ISGF3, STAT, and NF-κB. However, to detect other potential target genes of interferon induction, it is necessary to develop reliable methods for identifying the binding sites of these transcription factors.

The goal of this work was to analyze the regulatory regions of IFN-inducible genes and genes expressed in macrophages and develop reliable methods for identifying the binding sites of transcription factors IRFs, ISGF3, STATs, and NF-κB.

Methods and Algorithms

The following methods were used to recognize the transcription factor binding sites:

SITECON—a method for recognizing sites based on analysis of their conservative physicochemical and conformational properties (Oshchepkov et al., 2002) and

SiteGA—a method for site recognition that uses a genetic algorithm developed on the basis of iterative discriminant analysis (Levitsky, Katokhin, 2002).

As a recognition threshold, the SITECON method employs the level of necessary conformational similarity (Oshchepkov et al., 2002), which was 85% for NF-κB, 92% for STAT, 87% for ISGF3, and 93% for IRF. To optimize the recognition threshold for each site, type I error was minimized. When the SiteGA method was used, type I and type II errors were optimized taking into account the parameter "significance level of **α**", which defines an *a priori* probability of successful site recognition (Levitsky, Katokhin, 2002). All four types of sites were identified by the SiteGA method at $\alpha = 0.95$. Type II error was calculated using a random sequence with a length of 73300 bp.

** Abbreviations used in the text: NF-κB (Nuclear Factor-κB), IRF (Interferon Regulatory Factor), ISGF3 (Interferon-Stimulated Gene Factor 3), STAT (Signal Transducers and Activators of Transcription), and IFN (interferon).

Implementation and Results

The following two samples were selected from the information stored in the TRRD database: (1) IFN-regulated genes and (2) genes expressed in macrophages. The data on regulation of these genes by various transcription factors were analyzed. As is evident from Fig. 1, more than 15% of the genes from both samples are regulated by the transcription factors IRFs, ISGF3, NF-κB, and STATs. The transcription factor Sp1 also plays an important role in regulation of the genes expressed in macrophages (Fig. 1b).

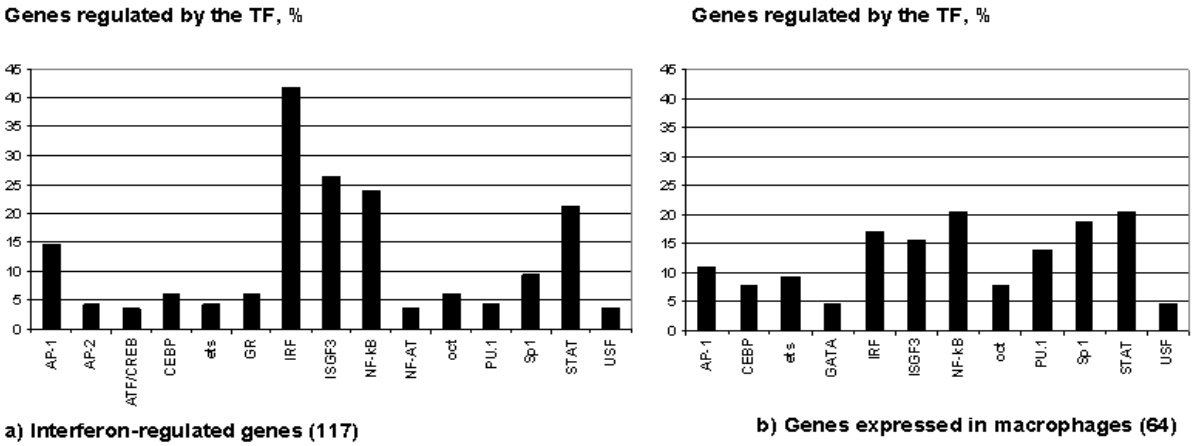


Fig. 1. Effect of various transcription factors on (a) IFN-regulated genes and (b) genes expressed in macrophages. A number of analyzed genes is given in parenthesis.

From the TRRDSITES database (Kolchanov et al., 1999), the binding sites of the transcription factors of the families IRF, NF-κB/Rel, and STAT and of the transcription factor ISGF3 were extracted. Training samples, necessary to develop methods for recognizing these sites, were constructed of the sites with sufficient experimental confirmation. On the basis of the training samples, we constructed two methods for recognizing the binding sites of these transcription factors: (1) SITECON is a method based on analysis of conservative physicochemical and conformational properties of a site (Oshchepkov et al., 2002) and (2) SiteGA is a recognition method that uses a genetic algorithm based on iterative discriminant analysis (Levitsky, Katokhin, 2002). A comparison of the methods is shown in Table 1.

Table 1. Comparison of the SITECON and SiteGA methods.

Factor	Type I error		Type II error		Sites predicted in a random sequence (1 site per bp)	
	SITECON	SiteGA	SITECON	SiteGA	SITECON	SiteGA
IRF	5.00%	6.20%	0.29%	0.02%	4581	1879
ISGF3	0	10.0%	0.04%	0.05%	555	9163
STAT	0	3.60%	1.24%	0.32%	85	289
NF-κB	2.50%	3.80%	0.51%	0.16%	163	482

The distribution of actual sites stored in the TRRD database was studied with respect to the transcription start. It was found that the majority of sites of all the four types are located in the promoter region from –1 to –200 with respect to the transcription start. All the four types occur within the distant 5' enhancers up to –5000, and the NF-κB sites are also met in the enhancers localized to introns up to +1500 with respect to the transcription start (Fig. 2).

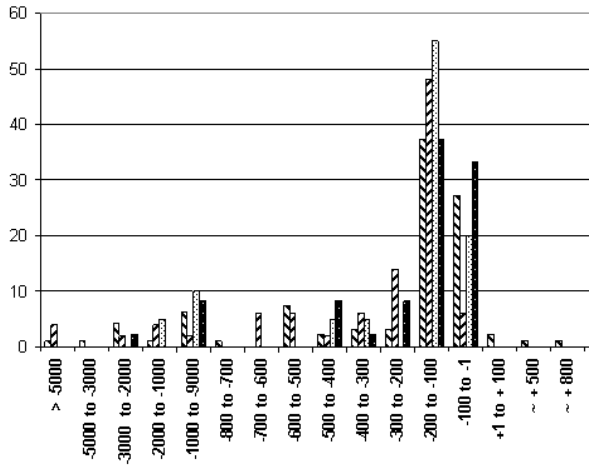


Fig. 2. Distribution of the transcription factor binding sites with respect to the transcription start. The ordinate axis shows a relative number of the sites in the sample (%).

NF-κB (96 sites);

STAT (50 sites);

IRF (48 sites);

ISGF3 (20 sites).

Discussion

The transcription factors STATs, IRFs, ISGF3, and NF-κB play an important role in regulation of genes expressed in cells of the immune system. ISGF3 is a key activator of early response gene transcription upon induction with type I interferons. The IRF-1 factor synthesized *de novo* in the induced cells is able to bind to the same sites as ISGF3, thus ensuring a later induction of the target genes. The factors of the STAT family are main activators of transcription upon induction by cytokines and growth factors. As a rule, the complexes of STAT dimers with DNA are unstable. It was found that a half-life of the Stat1/Stat1 complex with DNA equaled 3 min (Vinkemeier et al., 1996). However, if a factor binds to two adjacent sites and forms a tetramer, the stability of the complex *in vitro* is by more than an order of magnitude higher. In this case, the sites must be oriented in the same direction and have a distance between their centers of 18–23 bp (Vinkemeier et al., 1996). Using the developed methods for identifying STAT binding sites, we studied the flanking regions of 25 actual sites in the promoters of IFN-inducible genes (Table 2). For 23 sites, tandem repeats in the same orientation were predicted; moreover, 15 of these repeats were predicted by both methods (Fig. 3).

Table 2. Tandem repeats of sites in the promoters of IFN-inducible genes.

Factor	Sites studied	Tandem repeats predicted by			Tandem repeats confirmed experimentally	Site predicted by both methods in a random sequence	Distance between site centers
		SITECON	SiteGA	Both methods			
STAT	25	20	18	15	1	1/426 bp	13–63
ISGF3	33	15	4	3	8	1/14660 bp	13–67

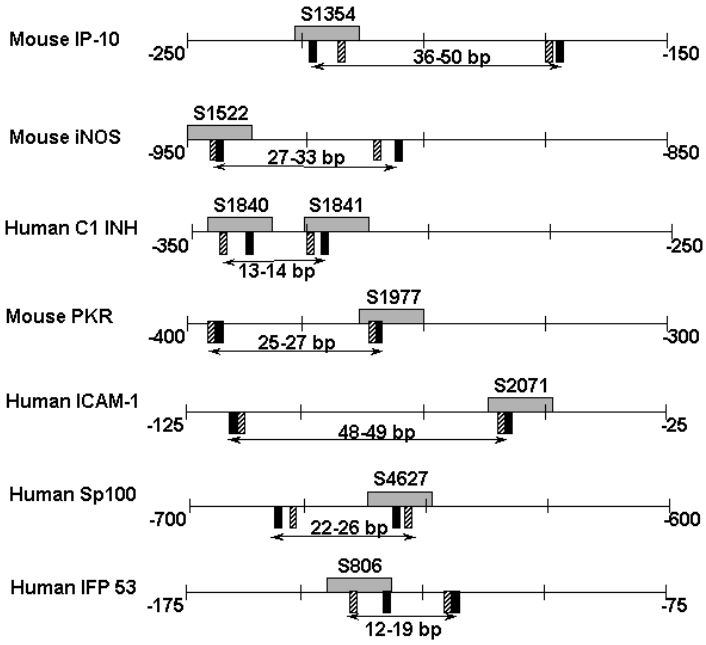


Fig. 3. Tandem repeats of the binding sites for STAT factors in the promoter regions of some IFN-inducible genes. Gray rectangles show the actual sites, stored in the TRRD database; shaded rectangles, the sites predicted by SITECON; and black rectangles, the sites predicted by SiteGA. Positions are given with respect to the transcription starts.

The available data suggest a higher affinity of the ISGF3 factor if two of its sites are located nearby and a multimer complex is formed of the two factors (Li et al., 1998). It is also known that tandemly repeated semicore sequences of IRF sites promote cooperative interaction of the factors (Fujii et al., 1999). Using recognition programs, we checked regions of 33 actual ISGF3 binding sites (Table 2). For these sites, the methods yielded mismatching results. However, note that the SiteGA method yielded a large number of unpredicted ISGF3 binding sites: 16 of 33 actual sites were not identified. The SITECON method predicted adjacent sites for 15 of the 33 actual ISGF3 binding sites in the region 20–80 bp, eight of which were confirmed experimentally. Among 37 actual IRF binding sites, the tandem repeat at a distance of 20–60 bp was predicted only for nine (data not shown).

A probability to obtain randomly 15 tandem repeats out of 25 studied sites for STAT and 8 out of 33 for ISGF3 with a distance between the repeats of 13–80 bp was determined by the χ^2 test. This probability was considerably lower than 10^{-10} for both sites. Thus, the result obtained is statistically significant.

Of great potential importance is a search for adjacent sites of various types in regulatory regions of genes of the immune system that can interact with one another, such as NF-κB and IRF-1, or the sites of tissue specific and ubiquitous transcription factors.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 00-07-90337, 00-04-49229, 00-04-49255, 01-07-90376, 01-07-90084, and 02-07-90359); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors thank I.V.Lokhova and L.V.Katokhina for their assistance with bibliography, D.A.Grigorovich for system administration, A.L.Proscura for a helpful assistance in extraction of samples from the TRRD database, and I.V.Filippova for translation into English.

References

1. Fujii Y., Shimizu T., Kusumoto M., Kyogoku Y., Taniguchi T., Hakoshima T. (1999). Crystal structure of an IRF-DNA complex reveals novel DNA recognition and cooperative binding to a tandem repeat of core sequences. EMBO J. 18:5028–5041.
2. Kolchanov N.A., Ananko E.A., Podkolodnaya O.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (1999). Transcription Regulatory Regions Database (TRRD): its status in 1999. Nucl. Acids Res. 27:303–306.
3. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30:312–317.
4. Levitsky V.G., Katokhin A.V. (2002). Recognition of eukaryotic promoters using genetic algorithm utilizing iterative discriminant analysis. Proc. BGRS-2002, this issue.
5. Li X., Leung S., Burns C., Stark J.R. (1998). Cooperative binding of Stat1-2 heterodimers and ISGF3 to tandem DNA elements. Biochimie. 80:703–710.
6. Oshchepkov D.Yu., Turnaev I.I., Vityaev E.E. (2002). SITECON: a method for recognizing transcription factor binding sites basing on analysis of their conservative physicochemical and conformational properties. Proc. BGRS-2002, this issue.
7. Vinkemeier U., Cohen S.L., Moarefi I., Chait B.T., Kuriyan J., Darnell J.E.Jr. (1996). DNA binding of *in vitro* activated Stat1 alpha, Stat1 beta and truncated Stat1: interaction between NH2-terminal domains stabilizes binding of two dimers to tandem DNA sites. EMBO J. 15:5616–5626.

RECOGNITION OF EUKARYOTIC PROMOTERS USING GENETIC ALGORITHM BASED ON ITERATIVE DISCRIMINANT ANALYSIS

Levitsky V.G. * Katokhin A.V., Lavryushev S.V.

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: levitsky@bionet.nsc.ru
*Corresponding author

Key words: *promoter recognition, genetic algorithm, discriminant analysis, nucleosome potential*

Resume

Motivation: The efficiency of methods for recognizing promoters of different types is proved to be dependent on account of their hidden context complexity. Combinations of various algorithms allow the recognition accuracy to be increased considerably; therefore, complex methods are most advantageous.

Results: A new approach to recognizing promoter regions of eukaryotic genes is proposed and illustrated by an example of *Drosophila melanogaster*. The essence of its novelty is in realizing the genetic algorithm to search for optimal partition of promoter region into local nonoverlapping fragments and selection of the most significant dinucleotide frequencies for the fragments obtained. The method developed was applied to recognizing TATA-containing (TATA+) and DPE-containing (DPE+) promoters of *Drosophila melanogaster* genes.

Availability: The program for promoter recognition is included into the GeneExpress system; section RegScan http://wwwmgs.bionet.nsc.ru/mgs/programs/proga/.

Introduction

The structure of core promoters displays a surprising diversity and is unique for each promoter, presumably reflecting the diversity of interactions between the proteins of transcription complex and promoter DNA (Goodrich et al., 1996; Kolchanov et al., 2002). Consequently, the research into various specific features of promoter context organization is acquiring an ever increasing importance (Hannenhalli, Levy, 2001). A number of methods for promoter recognition were proposed; these methods are based on detecting sets of specific context characteristics of promoter DNA taking into account their localization relative to the transcription start: models of Markov chains (Ohler et al., 1999), neural networks (Knudsen, 1999; Reese, 2001), and discriminant analysis (Davuluri et al., 2001). However the recent approaches combining several algorithms or utilizing both contextual and physicochemical properties of DNA for promoter recognition become the most efficient (Ohler et al., 2001).

In this work, we propose a new approach to recognizing eukaryotic promoters through detection of local contextual characteristics using the genetic algorithm (GA) based on iterative application of discriminant analysis. GA approved itself as an efficient tool for optimizing the functional dependent on numerous parameters (Willett, 1995). Here, the linear discriminant function defined by frequencies of dinucleotides within local promoter regions is used as a functional. This approach allowed us to recognize, along with other promoter types, the *Drosophila* TATA-less promoters lacking pronounced context signals.

Methods and Algorithms

The developed algorithm for promoter recognition utilizes a combination of several approaches and methods.

Optimization of parameters of promoter recognition functions. First, (i) a partition of promoter regions into local nonoverlapping fragments is searched for; then, (ii) the most significant frequencies of dinucleotides within the fragments obtained are selected. The GA utilizing iterative discriminant analysis of distribution of dinucleotide frequencies over the fragments of a current partition is used at both stages (Levitsky, Katokhin, 2001; Levitsky et al., 2001a).

The method searching for optimal partition starts from assigning in a random manner a certain set of initial partitions (Fig. 1a). Let us specify $N = 16 * P$ values of dinucleotide frequencies for P local regions (here, $P = 12$). Samples of (1) promoters and (2) random sequences obtained by shuffling individual promoter sequences were used to construct the recognition function. Let us determine the Mahalanobis distance R^2 (Mahalanobis, 1936) for samples (1) and (2):

$$R^2 = \sum_{k=1}^{N}\sum_{n=1}^{N}\{[f_n^{(2)} - f_n^{(1)}] * S_{n,k}^{-1} * [f_k^{(2)} - f_k^{(1)}]\}. \quad (1)$$

Here, $f_n^{(1)} = f_{i,p}^{(1)}$ is the mean frequency of the *i*th dinucleotide in the *p*th partition fragment for the sample of promoter sequences; $f_n^{(2)}$, the corresponding frequency for the sample of random sequences ($n = (p-1) \times 16 + i$, $p = 1,...12$, $i = 1,...16$, $n = 1,..,N$); and matrix S^{-1} is calculated according to the dinucleotide frequencies $f_n^{(1)}$ and $f_n^{(2)}$.

GA is constructed using elementary operations of two types: "mutations" (changes in the positions of borders between regions of the same partition, with constant number of these regions and constant minimal size of any region; Fig. 1b–e) and "recombinations" (exchange of fragments between two partitions; Fig. 2). The partition most suitable for recognition is determined as a result of successive "mutations" and "recombinations" of the partitions analyzed.

The method for selecting the most significant contextual characteristics. A subset of $m_p \leq 16$ dinucleotides is determined for each partition regions:

$$M = \sum_{p=1}^{P} m_p \tag{2}$$

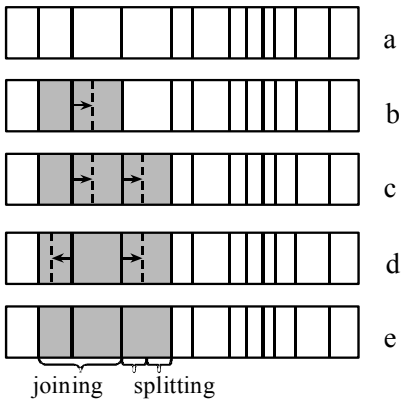


Fig. 1. Examples of modifications used by the optimal partition searching: (a) arbitrary distribution; (b) shift of the border between adjacent regions; (c) shift of a region relatively to the neighbor regions; (d) symmetrical shift of the region's borders relatively to its center; and (e) joining and splitting.

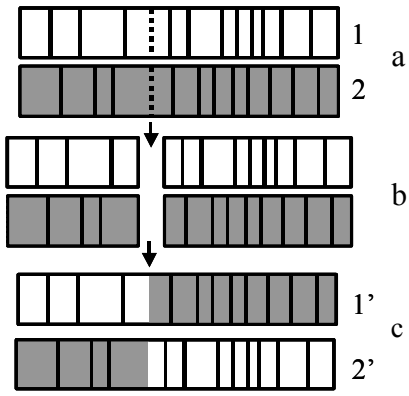


Fig. 2. Examples of crossovers used by the optimal partition searching: (a) introduction of a "break" (dotted lines) into two initial partitions 1 and 2 (indicated by different colors); (b) fragments of partitions after the break; and (c) exchange of fragments, removal of the break, and forming of the final partitions 1' and 2'.

Construction of the promoter recognition function. The value of recognition function is calculated for an arbitrary nucleotide sequence at each position of the window with a length of 400 bp (fragment *X*):

$$\varphi(X) = \frac{1}{R^2} \times \sum_{n=1}^{M} \sum_{k=1}^{M} \{[f_n(X) - (1/2) \times [f_n^{(2)} + f_n^{(1)}] \times S_{n,k}^{-1} \times [f_k^{(2)} - f_k^{(1)}]\}, \tag{3}$$

where $f_n(X)$ is the dinucleotide frequencies with account of the partition of the fragment *X*. The distance R^2 is calculated using equation (1) by summing over *M* selected dinucleotides. A higher probability of promoter recognition corresponds to the values of the function $\varphi(X)$ close to +1. The recognition function $\varphi(X)$ (3) was transformed as follows to recognize promoters with a specified significance level α:

$$\varphi_\alpha(X) = \begin{cases} \dfrac{|1 - \varphi(X)|}{P_\alpha \times \sigma_\varphi}, \text{ if } |1-\varphi(X)| < P_\alpha \times \sigma_\varphi, \\ 0, \text{ otherwise.} \end{cases} \tag{4}$$

Here, P_α is an α-quantile of the standard normal distribution (for example, $P_{0.95} = 1.96$) and σ_φ, a standard deviation of the recognition function $\varphi(X)$ values over the sample of promoter sequences.

Assessment of the recognition accuracy. The correlation coefficient (*CC*) characterizes the integral recognition accuracy with account of both the rate of false positives and the rate of false negatives:

$$CC = \frac{TP*TN - FN*FP}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}}, \quad (5)$$

where *TP* and *FP* are the numbers of true and false promoter prediction; *TN* and *FN,* the numbers of true and false "non-promoters" prediction.

Recognition of promoters using calculated recognition functions. While solving this problem for an arbitrary nucleotide sequence, the profile of nucleosome potential (NP; Levitsky, Katokhin, 2000) was calculated in addition to the promoter recognition function (4). Use of the NP profile allows the predictions of function (4) for the sequence positions failing to display NP values specific of promoter regions to be discarded. We have earlier demonstrated that the region [–50; +1] relative to the transcription start exhibits decreased mean values of NP (Levitsky, Katokhin, 2000; Levitsky et al., 2001a).

We used the mean NP_{PR} values over the region [–50; +1] relative to the transcription start for the promoter samples studied in this work with their standard deviations σ_{PR}, the mean NP value for the sample of *D. melanogaster* introns NP_{INT} with its standard deviation σ_{IN} for calculating the threshold values NP_{ST} while recognizing promoters:

$$NP_{ST} = \frac{NP_{PR} \times \sigma_{IN} + NP_{IN} \times \sigma_{PR}}{\sigma_{IN} + \sigma_{PR}}. \quad (6)$$

The values NP_{INT} were taken into account, as we discovered that the majority of false predictions fell into intron regions. However, we have earlier demonstrated that introns display a high NP (Levitsky et al., 2001b). Consequently, involvement of NP_{INT} allows false predictions of the promoter recognition function (4) to be excluded. Thus, according to the combined approach developed, a position *X* in the sequence analyzed is recognized as promoter if the two following conditions are met:

$$\varphi_{\alpha}(X) > 0 \quad (7a);$$

$$NP(X) < NP_{ST} \quad (7b)$$

Here, $\varphi_{\alpha}(X)$ is the promoter recognition function (4); $NP(X)$, the mean NP value over the region [–50; +1] relative to the putative transcription start (position *X*); and NP_{ST} is calculated according to equation (6).

Results and Discussion

To construct the promoter recognition function, we took 236 [–300; +100] fragments of *D. melanogaster* promoter sequences, phased relative to the transcription start, from the database EnDPD (Katokhin, Levitsky, 2000). Two promoter samples—TATA-containing (TATA+) and DPE-containing (DPE+)—were formed. The TATA+ promoter sample comprised 68 sequences displaying $Score_{TATA} \geq -5$, i.e., the value of weight matrix (Bucher, 1990) in the region [–40; –5] relative to the transcription start. Specific of the DPE-containing promoters is occurrence of a weaker (compared with the TATA box) context signal DPE (Downstream Promoter Element; Kutach, Kadonaga, 2000). Earlier, these promoters belonged to the heterogeneous group of TATA-less promoters. The DPE+ sample comprised 31 sequences from the Drosophila Core Promoter Database (http://www-biology.ucsd.edu/labs/Kadonaga/DCPD.html; Kutach, Kadonaga, 2000), tested by BLAST-analysis with respect to 5' ends of the corresponding ESTs.

Application of the GA utilizing iterative discriminant analysis at the first stage (search for optimal partition) resulted in the following accuracy estimates compared with the random sequences: $CC_{TATA+} = 0.92$ and $CC_{DPE+} = 0.70$. Upon the second GA stage (selection of the most significant contextual characteristics), the estimates amounted to $CC_{TATA+} = 0.92$ and $CC_{DPE+} = 0.82$. Note that the two-stage GA allowed the recognition accuracy of DPE+-type promoters to be essentially increased. Thus, we have demonstrated that selection of significant contextual characteristics is most efficient for increasing the recognition accuracy of the promoters with weakly pronounced contextual signals.

Further increase in the promoter recognition accuracy is achieved through a combined consideration of both the values of promoter recognition function (4) and nucleosome potential to discard the predictions of the function in question for the positions failing to display the NP values typical of promoters.

The mean values of NP_{PR} for the promoter samples used in this work amount to $NP_{TATA+} = 0.04$ and $NP_{DPE+} = -0.625$; the standard deviations (σ_{PR}) equal 0.73 and 0.85, respectively. The mean NP value for the sample of intron fragments amounts to $NP_{INT} = 0.534$ with a standard deviation of $\sigma_{IN} = 1.05$. Thus, according to equation (6), we obtain the following threshold values: $NP_{ST\,TATA+} = 0.33$ and $NP_{ST\,DPE+} = 0.02$.

The combined approach for eukaryotic promoter recognition is implement as an web-available program included into the GeneExpress system, section RegScan http://wwwmgs.bionet.nsc.ru/mgs/programs/proga/. It allows to determine position of putative promoter in sequences of length up to 32000 bp. The choice between different promoter types and usage of NP filter are customized.

Let us illustrate the efficiency of the combined approach developed. Figure 3 shows the profiles of functions $\varphi_{\alpha}(X)$ (4) at $\alpha = 0.95$ for promoters of the genes *zen* (TATA+; Fig. 3a) and *Cyt-C2* (DPE+; Fig. 3c) and profiles of their nucleosome

potential (Figs. 3b and 3d, respectively). The annotated sequences of the genes *zen* (FBgn0004053) and *Cyt-C2* (FBgn0000409) were retrieved from FlyBase (http://flybase.bio.indiana.edu/genes/).

Pronounced minimums of the nucleosome potentials, presumably corresponding to the regions with less dense nucleosome packaging (Levitsky et al., 2001a), are evident in the regions of gene transcription starts. It is also apparent that the peaks of the promoter recognition function fall into these regions. Note that the peaks of recognition function in the case of *Cyt-C2* gene promoters located near positions 1200, 1660, and 1700 (indicated with crosses; Fig. 3c) were discarded, as the corresponding NP values (Fig. 3c) fail to meet the condition (7b).

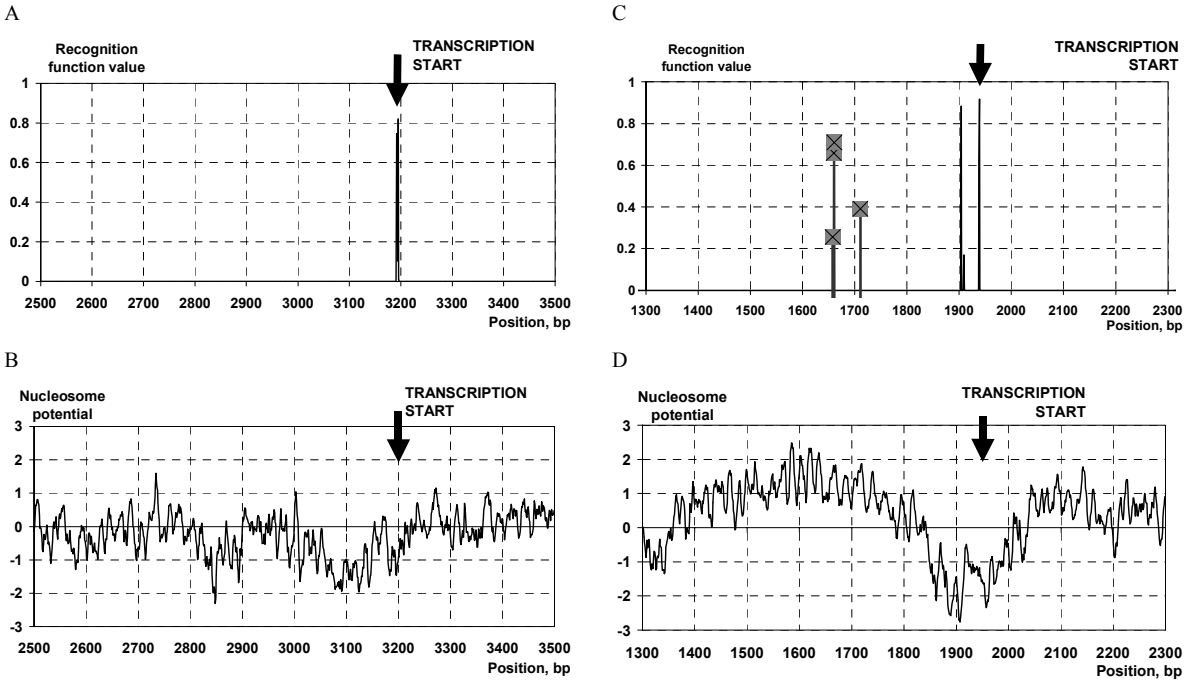


Fig. 3. Profiles of (a) the TATA+ promoter recognition function calculated for the gene *zen;* (b) its nucleosome potential; (c) DPE+ promoter recognition function calculated for the gene *Cyt-C2;* and (d) its nucleosome potential: arrows indicate transcription starts; crosses, the recognition function peaks discarded according to the condition (7b).

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 02-07-90355, and 00-04-49229); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Bucher P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. 212, 563-578.
2. Davuluri R.V, Grosse I., Zhang M.Q. (2001). Computational identification of promoters and first exons in the human genome. Nat. Genet. 29(4), 412-417.
3. Goodrich J.A., Cutler G., Tjian R. (1996). Contacts in context: promoter specificity and macromolecular interactions in transcription. Cell. 84(6), 825-830.
4. Hannenhalli S., Levy S. (2001). Promoter prediction in the human genome. Bioinformatics. 17, Suppl. 1, S90-S96.
5. Katokhin A.V., Levitsky V.G. (2000). Drosophila Promoter Database EnDPD: project and the first steps of its realization. In: Proc. IId Int. Conf. on Bioinf. Genome Regulation and Structure, Novosibirsk (eds. Kolchanov N.A. et al.), III, 105-108.
6. Knudsen S. (1999) Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics. 15, 356-361.
7. Kolchanov N.A, Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I,L, Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30(1), 312-317.
8. Kutach A.K., Kadonaga J.T. (2000). The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol. Cell. Biol. 20(13), 4754-4764.
9. Levitskii V.G., Katokhin A.V. (2001). Computer analysis and recognition of *Drosophila melanogaster* gene promoters. Mol. Biol. (Mosk.). 35(6), 970-978.

10. Levitsky V.G., Katokhin A.V. (2000). Characteristic modular promoter structure and its application to development of recognition program software. In: Proc. IId Intern. Conf. on Bioinf. Genome Regulation and Structure, Novosibirsk (eds. Kolchanov N.A. et al.), I, 86-89.
11. Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. (2001a). Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis. Bioinformatics. 17, 998-1010.
12. Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. (2001b). Nucleosome formation potential of exons, introns, and Alu repeats. Bioinformatics. 17, 1062-1064.
13. Mahalanobis P.C. (1936). On the generalised distance in statistics. Proc. Natl Inst. Sci. India. 12, 49-55.
14. Ohler U., Harbeck S., Niemann H., Noth E., Reese M.G. (1999) Interpolated Markov chains for eukaryotic promoter recognition. Bioinformatics. 15, 362-369.
15. Ohler U., Niemann H., Liao G., Rubin G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics. 17(Suppl.1), S199-206.
16. Reese M.G. (2001). Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. Comput. Chem. 26(1), 51-56.
17. Willett P. (1995). Genetic algorithms in molecular recognition and design. Trends Biotechnol. 13(12), 516-521.

PHYLOGENETIC FOOTPRINT. A NEW METHOD FOR PROMOTER ALIGNMENT

[1] Cheremushkin E., [1,2] Kel A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: cher@bionet.nsc.ru
[2] BIOBASE GmbH, 38304, Wolfenbuettel, Germany, e-mail: ake@biobase.de

Key words: *phylogenetic footprint, promoter alignment, TF sites, weight matrices*

Resume

Motivation: Phylogenetic Footprint is a new approach for revealing potential transcription factor binding sites in promoter sequences. The idea is based on an assumption that functional sites in promoters should evolve much slower then other regions that do not bear any conservative function. Therefore, potential transcription factor (TF) binding sites that are found in the evolutionally conservative regions of promoters have more chances to be considered as "real" sites. The most difficult step of the Phylogenetic Footprint is alignment of promoter sequences between different organisms (f.e. human and mouse). The conventional alignment methods often can not align promoters due to the high level of sequence variability.

Results: We have developed a new alignment method that takes into account similarity in distribution of potential binding sites. This method has been used effectively for promoter alignment and for revealing new potential binding sites for various transcription factors. We have developed a database of predicted potential TF binding sites in human genome by analyzing human/mouse conserved non-coding sequences (CNS).

Availability: http://compel.bionet.nsc.ru/FunSite/footprint/

Methods and Algorithms

We developed a new method of alignment of regulatory sequences that include information about TF binding sites. To search the sites we apply position weight matrices (PWM) from TRANSFAC database (www.biobase.de) (Wingender et al., 2001). Every nucleotide in a sequence can potentially be belong to one or several TF binding sites. We estimate the probability $w_p(\bar{S},k)$ of k-th nucleotide of sequence $\bar{S}$ to be belong to a binding site of a factor T_p ($p \in [1,P]$) by using the following formulism:

$$w_p(\bar{S},k) = \sum_{j=k-L+1}^{k} F\left(\mathrm{X}_p(\bar{S},j)\right), \quad \vec{w}(\bar{S},k) = \langle w_1(\bar{S},k),\ldots,w_P(\bar{S},k)\rangle$$

where $\mathrm{X}_p(\bar{S},j)$ - score of p-th matrix at j-th position of sequence, L - length of $\bar{S}$.

$$F(x) = \frac{\exp(\lambda \cdot x)}{\exp(\lambda)}$$

The corresponding scores for different weight matrices can be seen in the Figure 1. We use different smoothing functions that weight differently the core positions of the sites (Fig. 1 b and c).

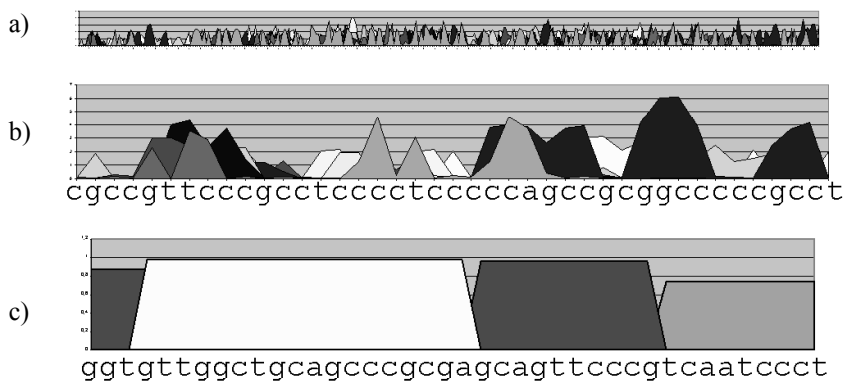


Fig. 1. Distribution of nucleotide weights in a sequence. For each nucleotide in sequence we compute a vector of weights that reflects the probability of the nucleotide to be belong to a TF binding site. Different colors correspond to different TFs. (b,c) – usage of two different smoothing functions.

It is known that the library of weight matrices contains matrices that are similar to each other. These are different matrices for the same transcription factor or for the transcription factors that are very similar in their DNA binding signature. We consider a similarity matrix M that takes into account similarities between weight matrices.

$\vec{\varphi}(\bar{S},k)=\vec{w}(\bar{S},k)\cdot \mathrm{M}$, where M - $P\times Q$ similarity matrix. We will use $\vec{\varphi}(a)$ instead of $\vec{\varphi}(\bar{S},k)$, where $a\in\Sigma\times\Phi$ - sequence element, $\gamma(a)\in\Sigma$ - nucleotide for this element.

Alignment Algorithm

We have developed an alignment algorithm for pair-wise and multiple alignment of nucleotide sequences (Cheremushkin, Kel, 2002). The algorithm is similar to the generally accepted Needleman-Wunsch dynamic programming algorithm. A major modification is made in the way of calculating the nucleotide substitution weights and gap penalty. The PWM scores were considered at every sequence positions in order to compute the corresponding substitution weights and gap penalty (see Fig. 2).

Fig. 2. We consider alignment as a favorable one, if sites are aligned to each other.

Gap penalty, while inserting gap in $\bar{S}^1$ between $k-1$ and k under position l in $\bar{S}^2$:

$$GAP(\bar{S}^1,\bar{S}^2,k,l)=\frac{G(\bar{S}^1,k)+R(\bar{S}^2,l)}{2},$$

Substitution weight:

$$SUB(\bar{S}^1,\bar{S}^2,k,l)=\mathrm{Z}(s^1_k,s^2_l),$$

where

$$G(\bar{S}^1,k)=\mathrm{Y}(s^1_{k-1},s^1_k),$$

$$R(\bar{S}^2,l)=\frac{\mathrm{Y}(s^2_{l-1},s^2_l)+\mathrm{Y}(s^2_l,s^2_{l+1})}{2}$$

$$\mathrm{Y}(a,b)=\frac{C_{gap}}{N}+W_{gap}\cdot s_{gap}(a,b),$$

$$\mathrm{Z}(a,b)=\frac{\Delta}{N}\cdot C_{sub}-W_{sub}\cdot\frac{\sum_{i=1}^{3}\lambda_i\cdot s_i(a,b)}{\sum_{i=1}^{3}\lambda_i},\ \text{for } a,b\in\Sigma\times\Phi,\ \text{where,}$$

$$\Delta=\begin{cases}1,\gamma(a)\neq\gamma(b)\\0,\gamma(a)=\gamma(b)\end{cases},\quad s_{gap}(a,b)=\begin{cases}(\vec{\varphi}(a)+\vec{\varphi}(b))^2,\gamma(a)=\gamma(b)\\\vec{\varphi}(a)^2+\vec{\varphi}(b)^2,\gamma(a)\neq\gamma(b)\end{cases},$$

$$s_1(a,b)=s_{gap}(a,b),\quad s_2(a,b)=\begin{cases}0,m>C_{\min}\\(C_{\min}-m)/C_{\min},m\le C_{\min}\end{cases},\ \text{where } m=\min_i|\varphi_i(a)-\varphi_i(b)|,$$

$$s_3(a,b)=\max_i(\varphi_i(a)\cdot\varphi_i(b)),$$

$\gamma(a) \in \Sigma$ - nucleotide, $\vec{\varphi}(a) \in \Phi$ - matrices weight vector, C_{corr}, C_{gap}, W_{corr}, W_{gap}, λ_i - constants.

N - number of sequences.

In the Figure 3 we present an example of alignment of two sequences that is done by the algorithm. The score values of the aligned sequences are shown above and under the sequences correspondingly. One can see that the score picks are aligned to each other.

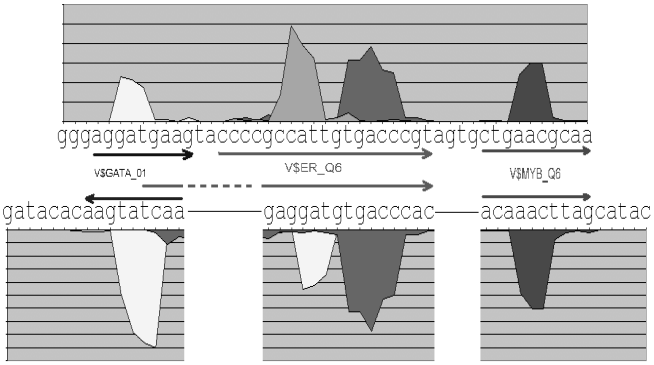


Fig. 3. Example of alignment of a sequences. Graphical repres entation with the nucleotide weights in the alignment.

Implementation and Results

The algorithm was implemented as a Java standalone program. It takes two sequences an input and align them. First it runs an Match algorithm that finds potential TF binding sites in the sequences. Specific collection of weight matrices with predefined cut-off values for every matrix can be specified by the user: taxon-dependent collection, tissue or function specific, minimizing false positive or false negative error. User can build his own profile with the help of TRANSPLORER program (http://www.biobase.de/pages/products/transplorer.html).

Testing of the alignment using a model of orthologous promoter sequences.

In order to validate the developed alignment algorithm we have constructed a computer model of evolution of promoter sequences. An ancestor sequence of a length L is randomly created. In this sequence we implant N_{sites} binding sites with $N_{sites}+1$ spacers between them and on flanks. From this sequence we generate two descendant sequences by introducing a R_{spacer} random mutations (insertions, deletions and substitutions) in the spacer regions and R_{site} substitutions in the sites. We require that after each iteration all sites should remain "functional". For that, we check the PWM score for each of them and discard cases when the score drops below a certain cut-off (CO_{site}). Then, these two sequences are aligned and positions of the alignment blocks are compared with the sites that were originally implanted. In the case of misalignment of one of the sites we report a failure.

We have compared the developed alignment algorithms with the ClustalW by counting the percentage of failures. Our algorithm shows much better performance in finding correct alignment. With the homology of sequences equals to homology between human and mouse, the failure rate of our algorithms was about 0.1% whereas ClustalW gives approximately 2.5% of failures.

Phylogenetic footprint of human/mouse conserved non-coding sequences (CNS)

Evolutionary conserved non-coding regulatory sequences (CNS) could serve as good landmarks on genome to find functionally important promoters, enhancers or silencers (Duret, Bucher, 1997). Phylogenetic footprinting of CNS will help us to reveal TF binding sites and assign a regulatory function to the regulatory regions and to the adjacent genes. We use results of the Berkeley Genome Pipeline (http://pipeline.lbl.gov/) of the global comparison of human and mouse genomes. We have download the complete list of CNS and made the phylogenetic footprint of all of them. Two types of alignment were used. First, we used the original alignment that was done by the VISTA program (http://www-gsd.lbl.gov/vista/), and second, we made our own alignment using the developed algorithm. Phylogenetic footprint was done by the previously developed tool (http://compel.bionet.nsc.ru/FunSite/footprint/) that takes two or several aligned sequences, finds conservative binding sites and display them. Binding sites with the score acceding a predefined cut-off, for transcription factors that belong to the same family and that have overlapping location on the alignment are considered as the positive match at the phylogenetic footprint. The list of 17117 CNS of the total length of alignment 2418267 bp was analysed. We applied a set of 240 weight matrices from TRANSFAC rel. 5.3 with the cut-offs optimized to minimize the sum of false positive and false negative errors. Using VISTA alignments we found 54075 conservative TF binding sites. Using alignment by our own algorithm we found 58106 conservative TF binding sites. So, our algorithm that includes information about potential TF sites at the very early stage of analysis allows us to

reveal 4031 more binding sites then using standard alignment algorithm. In the figure 4 one can see the comparison of the number of revealed sites using VISTA alignment versus our alignment.

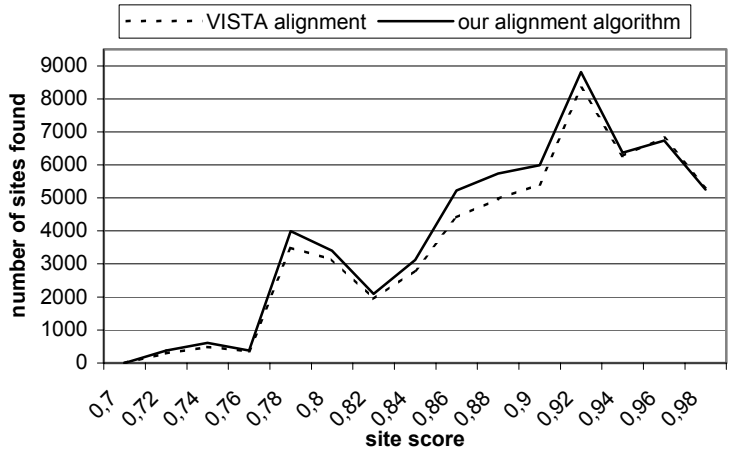


Fig. 4. **Comparison of the number of revealed sites using VISTA alignment versus our alignment algorithm. More sites with the score values from 0.78 to 0.92 can be revealed by our alignment algorithm.**

It is interesting to observe that our alignment algorithm helps to reveal more sites with the score values from 0.78 to 0.92, which are the most functionally relevant sites. Low scoring sites (lower then 0.72) and pick scoring (higher then 0.92) are revealed in the same amount as using the VISTA alignment.

We have developed a database of predicted potential TF binding sites in human genome by analyzing the human/mouse CNS. Using this database user can retrieve all conservative sites for a selected chromosome or for a region at the chromosome and can visualize gene information for the nearest upstream and downstream genes, that can be targets for regulation through found TF binding sites. Using the developed database molecular biologists can plan their experiments for validation of found target genes and can make regulatory functional annotation of human and mouse genome.

Acknowledgements

The authors are indebted to Edgar Wingender for fruitful discussion of the results. Parts of this work was supported by Siberian Branch of Russian Academy of Sciences and by the grant of Volkswagen-Stiftung (I/75941).

References

1. Cheremushkin E., Kel A. (2002) PromoterFootprint: A new method for alignment of regulatory genomic sequences. Phylogenetic footprinting of TF binding sites. In Liliana Florea, Brian Walenz, Sridhar Hannenhalli (eds) Currents in Computational Molecular Biology 2002. RECOMB 2002, Washington D.C. 40-41.
2. Duret L., Bucher P. (1997). Searching for regulatory elements in human noncoding sequences. Curr. Opin. Struct. Biol. **7**, 399-406.
3. 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., Urbach S. (2001) The TRANSFAC system on gene expression regulation. Nucl. Acids Res. 29, 281-283.

SPECIFIC STRUCTURAL FEATURES OF THE PROMOTERS IN THE EUKARYOTIC tRNA GENES OF DIFFERENT TYPES

Kondrakhin Yu.V., * Yudin N.S., Rogozin I.B., Naykova T.M., Voevoda M.I., Romaschenko A.G.

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: yudin@bionet.nsc.ru
*To whom correspondence should be addressed

Key words: *structural variants of the tRNA gene promoters, RNA polymerase III, evolutionary aspects of the transformation of promoters*

Resume

Motivation. It has been previously shown that groups of adjacent positions, modules, are the elementary units of the A and B promoter boxes in the tRNA genes. Each module is represented by a limited amount of structural variants, whose number is smaller (N<20) than that of the corresponding types of the tRNA genes in eukaryotes. Type-specific correlative relations were found between the structural variants of the particular modules that are not present in the tRNA genes of prokaryotes.

Results. The distribution of polymorphic DNA modules in the A and B promoter boxes of different tRNA gene types was analysed for two kingdoms of eukaryotes. It was shown that combinations of the structural variants of the DNA modules are characteristics of particular tRNA gene types and that these combinations are evolutionarily conserved, as a rule, within two multicellular kingdoms. In unicellulars, some of the combinations were revealed in the tRNA genes of other types. This was indirect evidence that the structural variants of the DNA modules may recombine during the evolutionary transition from unicellular to multicellular organisms.

Availability: http://www.bionet.nsc.ru/bgrs2002/

Introduction

In the genomes of eukaryotes, the tRNA genes are usually repeated many times and are scattered throughout a large number of chromosomes as single copies or clusters of different sizes. Identification of functioning tRNA genes among the very numerous homologous nucleotide sequences (pseudogenes) in the genome requires use of an accurate method for the recognition of functioning promoters. Furthermore, the large number of genes served by RNA polymerase II contain dispersed repeats (SINEs, LINEs) with DNA sequences homologous to the A and B boxes of the tRNA genes type 2 promoters. Some of them function to provide the synthesis of RNA; possibly performing epigenetic missions in the genome. Identification of RNA polymerase III type 2 promoters in these occasionally widely diverged from each other repeats would facilitate resolution of issues regarding the detection of functional copies of repeats within the various RNA polymerase II genes.

Here, we present the results of a thorough study on the structure of the intragenic promoters of different tRNA gene types from 3 eukaryotic kingdoms (Single Cell, Plant, Animal) and also on the prokaryotic tRNA genes whose promoters are outside the gene. We proceeded from the assumption that the differences may be rather due to the present promoter than to the tRNAs they encode.

Methods and Algorithms

The tRNA gene sequences from the database available via Internet at http://www.uni-bayreuth.de/departments/biochemie/trna/ were utilized as training samples. The tRNA gene type was determined by conversion of anticodon to the corresponding codon. The presence of tRNA in cytoplasm was taken as an indication of the expression of functional activity by these promoters. Samples of the A and B box DNAs from all tRNA gene types representing species of each kingdom were set up: 179 from 15 unicellular, 74 from 17 plant and 291 from 21 metazoan species. For prokaryotes, the samples were composed of 167 DNA samples from 25 Archaebacteria species and 409 DNA samples from 63 Eubacteria species. Comparative analysis of the abundance of the different variants of the DNA modules was performed using the SPSS 8.0 software.

Results and Discussion

Each A and B box was subdivided into 3 modules taking into account the correspondence of the levels of correlative relations between different positions of the DNA boxes and the distribution pattern of the invariant and relatively less conserved nucleotides. In the case of the A box, highly conserved adenine at position 7 was disregarded. From the results (Naykova et al., 2002) it followed that during the evolution from prokaryotes to eukaryotes the number of DNA variants corresponding to module 3 of the A box reduced considerably from 40 homologous DNA fragments in prokaryotes to 13 in eukaryotes. It should be noted that analysis revealed that the majority (35 variants) is represented in prokaryotes by homologous DNA fragments 5 bp in size (long variants). In eukaryotes, the number of the long DNA variants in module 3

was reduced to 8. The 5 tetrameric DNA variants of this module, widespread in prokaryotes, too, were found to be predominant in the different types of the eukaryotic tRNA genes. Combinations of the structural DNA variants of the different A and B box modules in the promoters of concrete type of the tRNA genes from unicellular species were in their majority different from those of the multicellular species whose combinations were, as a rule, highly conserved for particular tRNA gene types in all the analysed species of the two kingdoms (Animal, Plant) (Tables 1 and 2, truncate variants).

Table 1. The structural variants of the box A and B modules in the tRNA genes from multicellular organisms.

tRNA gene type	Box A			Box B		
	module 1, positions 1, 2, 3	module 2, positions 4, 5, 6	module 3, positions 8, 9, 10, 11, 12	module 1, positions 1, 2, 3, 4, 5	module 2, positions 6, 7, 8	module 3, positions 9, 10, 11
$tRNA^{Lys}$(AAA)	-TAG-	-CTC-	-GTCGG-	-GGTTC-	-AAG-	-TCC-
$tRNA^{Met}$(ATG)	-TGG- -TAG-	-CGC-	-GCGG- -GTGG- -GTAGG-	-GGATC- -AGTTC-	-GAA- -GAT-	-ACC- -CCT-
$tRNA^{Asp}$(GAC)	-TAG-	-TAT-	-GTGG-	-GGTTC-	-GAT- (M) -AAT-	-TCC-
$tRNA^{Gln}$(CAA)	-TGG-	-TGT-	-ATGG-	-AGTTC-	-AAA-	-TCT-
$tRNA^{Gln}$(CAG)	-TGG-	-TGT-	-ATGG-	-AGTTC-	-AAA-	-TCT-
$tRNA^{Ser}$(TCT)	-TGG-	-CCG-	-GTGG-	-GGTTC-	-GAA-	-TCC-
$tRNA^{Ser}$(TCA)	-TGG- -TGT- (C)	-CCG-	-GTGG-	-GGTTC-	-GAA-	-TCC- -CCC-
$tRNA^{Ser}$(TCG)	-TGG-	-CCG-	-GTGG-	-GGTTC-	-GAA-	-TCC-
$tRNA^{Ser}$(AGC)	-TGG-	-CCG-	-GTGG-	-GGTTC-	-GAA-	-TCC-
$tRNA^{His}$(CAC)	-TCG-	-TAT-	-GTGG-	-GGTTC-	-GAA-	-TCC-

Note: C – Caenorhabditis elegans, M – mammalian.

Table 2. The most typical structural variants of the box A and B modules in the tRNA genes from unicellular organisms.

tRNA gene type	Box A			Box B		
	module 1, positions 1, 2, 3	module 2, positions 4, 5, 6	module 3, positions 8, 9, 10, 11, 12	module 1, positions 1, 2, 3, 4, 5	module 2, positions 6, 7, 8	module 3, positions 9, 10, 11
$tRNA^{Lys}$(AAA)	-TAG-	-CTC-	-GTCGG- -GTTGG- -GTGG-	-GGTTC-	-GAG- -GAT-	-CCC-
$tRNA^{Met}$(ATG)	-TAG-(7) -TAA-(1) -TGG-(7)	-AGC-(1) -CGA-(1) -CGC-(6) -CTC-(4) -GAG-(3)	-ATGG-(1) -GGGG-(1) -GTAGG-(3) -GTGG-(9)	-AGTTC-(3) -GGATC-(10) -GGATG-(1) -GGTTC-(1)	-AAT-(1) -GAA-(14)	-ACC-(11) -CCT-(3) -TCC-(1)
$tRNA^{Asp}$(GAC)	-TAG-(6)	-TAT-(3) -TTT-(3)	-ATGG-(3) -GGGG-(1) -GTGG-(2)	-GGTTC-(6)	-AAT-(5) -GAA-(1)	-TCC-(6)
$tRNA^{Gln}$(CAA)	-TAG-(6) -TGG-(1)	-CTC-(1) -TGT-(60	-ATTGG-(1) -GCGG-(1) -GTGG-(5)	-AGTTC-(1) -GGTTC-(6)	-AAA-(1) -GAA-(5) -GAG-(1)	-CCC-(1) -TCC-(5) -TCT-(1)
$tRNA^{Gln}$(CAG)	-TAG-(1)	-TGT-(1)	-GTGG-	-AGTTC-(1)	-GAA-(1)	-TCC-(1)
$tRNA^{Ser}$(TCT)	-TAG-(1) -TGG-(4) -TGT-(1)	-CAA-(1) -CCG-(5)	-GTGG-(6)	-GGTTC-(6)	-GAA-(2) -AAA-(1) -GAG-(3)	-TCC-(5) -CCC-(1)
$tRNA^{Ser}$(TCA)	-TGG-(5) -TGT-(2) -TAG-(1)	-CCG-(5) -CGA-(1) -CAG-(2)	-GTGG-(8)	-GGTTC-(8)	-GAA-(3) -AAA-(5)	-TCC-(6) -CCC-(2)
$tRNA^{Ser}$(TCG)	-TGG-(2) -TGT-(1)	-CCG-(3)	-GTGG-(3)	-GGTTC-(3)	-AAA-(2) -AAT-(1)	-TCC-(2) -CCC-(1)
$tRNA^{Ser}$(AGC)	-TGG-(1) -TAG-(1)	-CCG-(1) -CAA-(1)	-GTGG-(2)	-GGTTC-(2)	-GAA-(2)	-TCC-(1) -CCC-(1)
$tRNA^{His}$(CAC)	-TAG-(4) -TGG-(1)	-TAT-(4) -TCC-(1)	-GTGG-(5)	-GGTTC-(5)	-GAT-(5)	-TCC-(2) -TCT-(3)

It was found that the long variants of the A box (the A1 subclass) occur only in a small number of the tRNA gene types in Metazoa to which among the genes we analysed the $tRNA^{Lys}$(AAA), $tRNA^{Lys}$(AAG), $tRNA^{Phe}$(TTC), $tRNA^{Tyr}$(TAC), $tRNA^{Asn}$(AAC), $tRNA^{Thr}$(ACT), $tRNA^{Ala}$(GCT) and $tRNA^{Ile}$(ATT) belong. In that part of the tRNA genes of this subclass, which instead of –TAG- variant of the module 1 have –TGG-, the DNA variant of module 2 is altered, too.

In the $tRNA^{Met}$ gene, short and long DNA variants of the A box were detected. In the case of the short variant of the A box, the DNA variants of all the modules of these boxes, except the module 2 variant of the A box, were modified. The promoters of the $tRNA^{Lys}$(AAA) and $tRNA^{Lys}$(AAG) genes have identical A boxes, but differ by combinations of the variants of modules the 2 and 3of the B box.

The short variants of the A box prevail in the majority of types of the tRNA genes. Those with -TAG- variant of module 1 fall into 2 classes; some combine with the -CTC- variant of module 2 (for example $tRNA^{Cys}$(TGC), $tRNA^{Ala}$(GCA) and $tRNA^{Thr}$(ACA)); others combine with the -TGT- variant ($tRNAs^{Val}$(GTT, GTA, GTG) and -TAT- variant ($tRNA^{Asp}$(GAC)). Most short DNA sequences of the A box have the -TGG- variant of module 1, which combines with the variants of module 1 CGC-, -CCG-, -TGT-, -TCT-, -TAT- and -TTC- rarely occurring in the DNA sequence of the A1 subclass and only in combination with the -TGG- variants of the module 1. Exceptions are the $tRNA^{Met}$ genes in which the -CGC- variant is present in both isoforms of the A box.In the list of the tRNA genes (Table 1) the $tRNA^{His}$(CAC) gene differs from the others by the minor structural variant of module 1 -TCG-. This variant of module 1 is conserved in the $tRNA^{His}$(CAC) genes isolated from the Drosophila, mouse, sheep and human genomes. In Drosophila, the-TCG- variant combines with the -TCT- variant of module 2 and with the -TAT- variant in other organisms.

Certain polymorphic short DNA variants of module 3, in contrast to long occur, as a rule, in several tRNA gene types. For example, the -GGGG- variant was found in the $tRNA^{Thr}$(ACA), $tRNA^{Pro}$(CCT) and $tRNA^{Cys}$(TGC) genes. The -ATGG- variant is represented in the $tRNAs^{Arg}$(CGT, CGA, AGA), $tRNA^{Trp}$(TGG), $tRNAs^{Gln}$(CAA, CAG) genes. Such variant, like -GCGG-,was found in the $tRNAs^{Val}$(GTG, GTA), $tRNAs^{Leu}$ and $tRNA^{Met}$ genes. The variant –GTGG- occur in the most number of particular types of the tRNA genes. Summarizing the data in Table 1, it may be stated that the particular type of the tRNA genes, in most cases, differs by the structure of at least a single module.

For unicellular organisms, compared to multicellular, a characteristic feature is wide diversity of the DNA variants of the different modules of the A and B boxes. The diversity of module 2 DNA in unicellulars is a prominent feature. Certain module combinations in unicellular organisms occur in gene types other than those in multicellular organisms. This is indirect evidence that recombination of the structural variants of the modules was a possibility during evolution.

We attempted to understand why certain combinations of the structural DNA variants of different modules in various types of the tRNA gene were under selection. It proved that, in most cases, elements are formed at the junction of the structural variants of two modules can change the parameters of the DNA helix. This is most obvious for the nucleotide sequence of the A box with respect to a number of the identified combinations of structural variants of the modules (Tables 3 and 4). For example, the -TGG- variant of module 1 combined with the -CGC-, -CCT-, -CCG- variants of 2 form the palindromes GCGC- and -GGCC-, respectively. The same -TGG- variant in combination with the module 2-TGT- -TAT-, -TCT- variants form a repeat with mirror symmetry, -TGGT-. The widespread variants -TAG- of module 1 and -CTC- of module 2 form the palindrome -AGCT-. Combination of the -TAG- and -TGT- variants, in the case of the $tRNA^{Val}$ genes, form the tetraoligonucleotide -TAGT- repeated at the junction of the same -TGT- with A at position 7 and the variant GTGG of module 3. The list of such examples for the DNA sequences of the A box is actually longer (see Table 3). The same features were observed for the structure of the B box. All the variants of module 1 of this box form palindromes with the variants of module 2 having guanine at position 6. In the case of the structural variants -GAG- and -GAT- (module 2), the palindromes -TCGA- is formed and the hexanucleotide -TTCGAA- is formed in the presence of the -GAA-variant. The imperfect palindrome -TTCAA-is formed in the case of the-AAA- and -AAT- variant.

Table 3 presents the results for the distribution of palindromes in pooled samples of the eukaryotic tRNA genes in the case of short nucleotide sequences of the A box and homologous prokaryotic DNA fragments. From the comparative results it is obvious that the occurrence frequencies of the palindromes and the repeats with mirror symmetry increased during the evolutionary transition from prokaryotes to unicellular eukaryotes, reaching maximum values in multicellular organisms (Table 4). Features of this kind are essential for the core of a large number of known cis-elements with which the specific trans-acting protein factors interact. It cannot be excluded that saturation with such elements of the coding part of the tRNA genes is due to the formation in this region of an intragenic promoter with whose DNA the transcription factor interacts.

Table 3. The observed (p1) and expected (p2) probabilities of palindromes at the different positions of the A2 subclass of the box A.

Kingdom	Proba-bilities	Box positions							
		1-4	2-5	3-6	4-7	5-8	6-9	7-10	8-11
Eukaryotes	p_1	-	0.258	0.156	0.084	0.174	-	-	-
	p_2	-	0.170	0.049	0.126	0.177	-	-	-
	p_1/p_2	-	1.52	3.20	0.67	0.98	-	-	-
Eukaryotes Without Single Cell	p_1	-	0.248	0.214	0.092	0.168	-	-	-
	p_2	-	0.185	0.055	0.130	0.176	-	-	-
	p_1/p_2	-	1.34	3.89	0.71	0.96	-	-	-
Prokaryotes	p_1	-	0.078	0.01	0.04	0.127	-	-	-
	p_2	-	0.122	0.025	0.098	0.141	-	-	-
	p_1/p_2	-	0.64	0.044	0.418	0.90	-	-	-

Table 4. The observed (p1) and expected (p2) probabilities of the repeats with mirror symmetry at the different positions of the A2 subclass of the box A.

Kingdom	Proba-bilities	Box positions							
		1-4	2-5	3-6	4-7	5-8	6-9	7-10	8-11
Eukaryotes	p_1	0.309	0.033	0.197	-	-	0.141	-	0.046
	p_2	0.356	0.012	0.073	-	-	0.099	-	0.035
	p_1/p_2	0.87	2.71	2.69	-	-	1.41	-	1.32
Eukaryotes without Single Cell	p_1	0.344	0.042	0.206	-	-	0.143	-	0.046
	p_2	0.390	0.014	0.072	-	-	0.091	-	0.035
	p_1/p_2	0.88	2.91	2.86	-	-	1.58	-	1.31
Prokaryotes	p_1	0.123	0.052	0.071	-	0.090	0.071	-	0.041
	p_2	0.150	0.031	0.0079	-	0.075	0.048	-	0.032
	p_1/p_2	0.82	1.71	0.90	-	1.20	1.46	-	1.28

References

1. Naykova T.M., Kondrakhin Yu.V, Rogozin I.B., Voevoda M.I., Yudin N.S., Romaschenko A.G. (2002) This issue.

STRUCTURAL REORGANIZATION RESULTING IN THE APPEARANCE OF INTRAGENIC PROMOTER SPECIFIC TO DIFFERENT tRNA GENE TYPES IN EUKARYOTES

**** Naykova T.M., Kondrakhin Yu.V., Rogozin I.B., Voevoda M.I., Yudin N.S., Romaschenko A.G.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: naykova@bionet.nsc.ru
*Corresponding author

Key words: *intragenic promoters of the different gene types, structural organization of the A and B boxes, DNA polymorphism of the A box*

Resume

Motivation: During the evolutionary transition from prokaryotes to eukaryotes, the emergence of the nucleus was associated with dramatic changes in the cell. The eukaryotic genes started to be transcribed by three different polymerases instead of a single RNA polymerase, which was common to all the prokaryotic genes. In eukaryotes, the tRNA genes became transcribed by RNA polymerase III and their promoter composed of the A and B boxes appeared within the coding sequences of the tRNA genes. Hence, the DNA in the location site of the A and B boxes in the eukaryotic tRNA genes started to accommodate both the promoter elements and the information about the structural features of particular tRNA molecules. To estimate the permissive limits of structural variability of the nucleotides in the DNA region of the A and B boxes compatible with promoter function and differential gene expression, we analyzed the A and B box DNAs in all the tRNA gene types for 3 eukaryotic kingdoms and compared them with the corresponding regions of the prokaryotic tRNA genes.

Result: It was found that in eukaryotes the A and B box DNAs are subdivided into modules on the basis of highly significant intragroup correlative relations of the closely adjacent nucleotides. The number of structural variants of each module is restricted, much smaller than 20, and the number of the combinations of each structural variant of the modules is fixed. As a result, two promoter subclasses of the tRNA genes differing by the structure and size of the A box were identified.

Introduction

Transfer RNAs are classified by codon/anticodon interaction of aminoacetylated tRNA molecules with mRNA. A definite tRNA type or a set (in the case of the presence of isoacceptor molecules) corresponds to each of the 20 different amino acids. The composition of tRNA molecules may vary widely both qualitatively and quantitatively, depending on the type of the cells and their differentiation stage (Ikemura, 1985). During the prokaryote-eukaryote transition, with the emergence of the nucleus, the tRNA genes became transcribed by one of the three newly arisen RNA polymerases, RNA polymerase III, whereas in prokaryotes these genes remained transcribed by a polymerase common to all the genes (Geiduschek, Tocchini-Valentini, 1988). The promoter occurs in the coding part of the all types tRNA genes in eukaryotes, while it is located upstream from the transcription start site in the prokaryotic tRNA genes. The similarity of the secondary and tertiary structures between the prokaryotic and nuclears tRNA suggests strong selection was targeted at maintenance of the standard molecular architecture in all the types tRNA (Soll, RajBhandary, 1995). However, RNA polymerase III initiates transcription when it form a complex with the basal protein factors recognizing the specific nucleotide blocks in DNA (Willis, 1993). For this reason, characteristic structural changes would be expected in the region where intragenic promoters appeared in the all types tRNA genes in eukaryotes; in contrast to the homologous tRNA genes in prokaryotes whose structure remained unaltered.

The intragenic promoters of the eukaryotic tRNA genes are comprised of spatially separate elements termed the A and B boxes (Geiduschek, Tocchini-Valentini, 1988). The variable 11-12 bp A box DNA occupies the region of the tRNA gene that corresponds to the sequence of the D arm of the three-dimensional tRNA structure (the nucleotide sequence downstream from the acceptor stem stretched inclusive across 3/4 of the D loop) of the tRNA molecules. The B box is in the DNA region that corresponds to the T loop with dinucleotides of the T stem flanking it at both ends in the molecule tRNA (Rich, RajBhandary, 1976).

Materials and Methods

The structure of the A and B box DNAs of the tRNA genes from three eukaryotic kingdoms (Protozoa, Plant, Metazoa) and the corresponding regions of the tRNA genes from prokaryotic species (Archae, Eubacteria) were compared. Estimates of changes in DNA structure were based on comparisons of total samples from the A and B box sequences composed of all the types of the tRNA genes from different eukaryotic species with the homologous DNA fragments of different prokaryotic tRNA genes. The estimates were also based on comparisons of sets of the A and B box DNAs belonging only to a particular

type of tRNA genes from eukaryotic species with the corresponding DNA fragments of the same type of the tRNA gene from prokaryotic species. The nucleotide sequences of the tRNA genes from the database (Sprinzl et al., 1996) available at http://www.unibayreuth.de/departments/biochemie// were subjected to analysis. The frequencies of the structural variants for different regions of the A and B box DNAs and those of their combinations were estimated using the SPSS 8.0 software.

Results and Discussion

The DNA structure of the B box site proved to be relatively conserved (except for 2 or 3 positions) in the tRNA genes of all the studied prokaryotic and eukaryotic species. In contrast, the DNA structure in the A box region, especially in its 3' part, underwent drastic changes during the evolutionary transition from prokaryotes to eukaryotes.

Two DNA subclasses of the A box were identified among the sequences DNA of the A boxes in the total sample of the eukaryotic tRNA genes: long 12 bp and short 11 bp variants differing by the length of the 3' part of the A box DNA. In a previous analysis, we established specific distribution patterns for the long and short variants of the A box DNA among certain gene types in different eukaryotic tRNA genes. This was contrary to prokaryotes showing characteristic heterogeneity among homologous fragments DNA within the particular types of the tRNA genes. This may be explained by assuming that during the prokaryote-eukaryote transition certain DNA sequence variants corresponding to the 3' part of the A box were favored by selection. Selection for concrete structural variants of this DNA region was presumably determined by the amino acid specificity of the tRNA molecules coded by the tRNA genes.

Table 1. Distribution of the long and short variants of the A box DNA among some of the types of the eukaryotic tRNA genes.

№	Types of the tRNA genes	Long variants			Short variants			Number of sequences
		Metazoa	Plant	Protozoa	Metazoa	Plant	Protozoa	
1	$tRNA^{Lys}$(CUU)	13	1	6				20
2	$tRNA^{Phe}$(GAA)	16	6	7				29
3	$tRNA^{Tyr}$(GUA)	10	9	5			1	25
4	$tRNA^{Ile}$(AAU)	4	1	5			1	11
5	$tRNA^{Lys}$(UUU)	6		4			1	11
6	$tRNA^{Ala}$(AGC)	6		5	1	1		13
7	$tRNA^{Ile}$(UAU)	1		1			1	3
8	$tRNA^{Thr}$ (AGU)	2		2	2		1	7
9	$tRNA^{Asn}$(GUU)	5		5	1	2	2	15
10	$tRNA^{Val}$(AAC)		1	5	8	1		15
11	$tRNA^{Met}$(CAU)	3	4	3	15	4	9	38
12	$tRNA^{Ala}$(UGC)		1	1	5		1	8
13	$tRNA^{Glu}$(CUU)		1	1	7		4	13
14	$tRNA^{Glu}$(CUC)		2		6		2	10
15	$tRNA^{Trp}$(CCA)		2		5	1	4	12
16	$tRNA^{Arg}$(UCU)			1	2		4	7
17	$tRNA^{Arg}$(ACG)			1	5	1	8	15
18	$tRNA^{Leu}$(UAA)			1	1	1	4	7
19	$tRNA^{Val}$(UAC)			1	2		5	8
20	$tRNA^{Gln}$(UUG)			1	6		6	13
21	$tRNA^{Ser}$(CGA)	1			2	1	3	7
22	$tRNA^{Ser}$(GCU)				3	2	2	7
23	$tRNA^{Ser}$(UGA)				5	2	7	14
24	$tRNA^{Val}$(CAC)				5		2	7
25	$tRNA^{Pro}$(UGG)				3	2	3	8
26	$tRNA^{Pro}$(AGG)				5	2		7
27	$tRNA^{Asp}$(GUC)				8	1	5	14
28	$tRNA^{Gln}$(CUG)				7		1	8
29	$tRNA^{Gly}$(GCC)				6	4	2	12
30	$tRNA^{His}$(GUG)				6	1	4	11
31	$tRNA^{Leu}$(CAA)				3	1	4	8
32	$tRNA^{Leu}$(CAG)				7		3	10
33	$tRNA^{Leu}$(AAG)				5	1	2	8

Comparative analysis of the 3' part of the A box DNA of the eukaryotic tRNA genes with the corresponding DNA fragments of the prokaryotic tRNA genes revealed that the number of structural variants in this DNA region reduced considerably from 40 in prokaryotes to 13 in eukaryotes and also that the total number of long DNA variants prevailing in prokaryotes redistributed in favor of short variants DNA in eukaryotes (Table 2).

Table 2. Abundance of the structural variants in the 3' part of the block of nucleotides in the A box DNA (positions 8, 9, 10, 11 and 12) in the promoters of the eukaryotic tRNA genes and in the homologous DNA sequences of the prokaryotic tRNA genes.

Structural variants							
	Long (12 bp)					Short (11 bp)	
	GNNHG n=17	HBNWG n=6	RNHGG n=10	YMHGG n=5	RHHGG n=8	RBGG n=5	NNG*N n=17
Archae	59.5%	1.4%	1.4%	1.8%	11%	23.5%	1.4%
Eubacteria	9.3%	0.2%	7.3%	0.2%	49.5%	29.8%	3.7%
Protozoa	0.5%	0.5%	0.5%	0.5%	32%	64%	2%
Plant	-	-	2.7%	1.3%	40.6%	54.1%	1.3%
Metazoa	0.4%	0.8%	1.1%	0.4%	26%	69.4%	1.9%

Note:
8 9 10 11 12
-RHHGG- (-GTTGG-; -GTCGG-; -GCTGG-; -ATCGG-; -ATTGG-; -AATGG-; -GATGG-; -GTAGG-) - the long structural variants of the oligonucleotides occurring in the eukaryotic tRNA genes in the 3' part of the A box DNA;
-RNHGG- - nucleotide sequences without the -RHHGG- variants;
8 9 10 11
-RBGG- (-ACGG-; -ATGG-; -GCGG-; -GGGG-; -GTGG-) - sequences shorter by one nucleotide than the long -RHHGG- sequence variants of the eukaryotic tRNA genes;
-NNG*N- - - short oligonucleotides without the -RBGG- variants;

* Of the 17 rare identified structural short variants of oligonucleotides satisfying the requirements for -NNGN- was the sole -GTTA- revealed in Archae having G instead of T at the next to last position; n - the number of structural variants. Designations - R: A,G; Y: T, C; M: A, C; W: A, T; H: A, T, C; B: T, G, C.

From analysis of the data in tables 1 and 2 it may be concluded that during the transition from prokaryotes to eukaryotes:

1. The structural variability of the 3' part of the A box DNA in all the tRNA gene types reduced sharply in eukaryotes. Both in unicellular and multicellular organisms, the reduction was mainly restricted to 8 long (-RHHGG-) variants DNA and 5 short ones (-RBGG-) instead of the 40 homologous fragment variants in prokaryotes.

2. In many tRNA gene types long variants were substituted by short ones.

3. The distribution of the long and short variants of the A box DNA among a large number of the tRNA genes was evolutionary conserved among species of the different eukaryotic kingdoms.

Fixation either long or short DNA variants in the 3' part of the particular type of the A box DNA implied the existence of correlative structural variations also in the 5' part of this promoter element.

Based on analysis of weight matrices, it was found that the 5' part of the A box DNA may be divided into 2 parts according to DNA heterogeneity. In all eukaryotes, the first triplet is represented mainly by two variants, -TAG- (preferential for the long variants of the A box DNA) and -TGG- (preferential for the short variants of the A box DNA), and by several minor variants. The second triplet of the 5' part of the A box DNA is very heterogeneous: 10 variants in multicellular and 15 variants in unicellular organisms. The differences in polymorphism level between the first and second triplet are presumably due to their different contribution to the formation of the tertiary structure of the molecule tRNA. It proved that the variants of the first and second triplets are dependent on each other and on the length of the 3' part of the A box DNA (Table 3).

Table 3. Combination frequencies of the structural variants of the first and second triplets in the 5' part of the DNA sequence of the long and short A boxes in the tRNA genes from three eukaryotic kingdoms.

Structural variants			Variants of the second triplet of the A box DNA														
			CTC	CGC	CCG	TCT	TGT	TAT	CTT	CCT	TTC	CGG	TTT	CGT	CCC	TCC	CCA
Variants of the first triplet of the A box	A box (12 bp)	TAG(100)	88	1		3	1		2				2		1		2
		TGG(44)	1	16		4	8		2				1	3	4		5
	A box (11 bp)	TAG(68)	22			1	20	17	2				5		1		
		TGG(198)	6	38	62	33	22	4		7	7		2	3	6	8	
		TAA(4)	1						1				2				
		TCG(6)				3		3									
		TGC(7)										7					
		TGT(6)		6													

Note: the structural variants of the second triplet occurring only in the tRNA genes of unicellular eukaryotes are in italics.

Summarizing, the existence of two subclasses of the tRNA gene promoters differing by the size and structure of the A box DNA may be postulated. The variants of the first subclass of the promoters are characterized by greater length and higher structural heterogeneity in the 3' part of the A box DNA, which is in contrast to the relative structural homogeneity of its 5' part. The variants of the second subclass of the tRNA genes promoters are distinguished by lower structural DNA diversity of 3' part of the A box DNA and higher structural heterogeneity of its 5' part (Table 3).

Thus, the A box DNA consists of three subfragments (modules): the first, the second triplets and the 3' part of the A box DNA differing by structural heterogeneity. The certain structural variants of each module combine in preference with each other. The obtained data are consistent with current knowledge about the formation of transcription complexes on the promoters of RNA polymerase III; moreover, they demonstrate the potential capacities for differential and group expression of the tRNA genes of different types.

References

1. Geiduschek E.P., Tocchini-Valentini G.P. (1988) Transcription by RNA polymerase III. Ann. Rev. Biochem. 57, 873-914.
2. Ikemura T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13-34.
3. Rich A., RajBhandary U.L. (1976) Transfer RNA: molecular structure, sequence and properties. Ann. Rev. Biochem. 45, 805-860.
4. Soll D., RajBhandary U.L. (1995) tRNA: structure, biosynthesis and function. ASM Press, Washington, D.C.
5. Sprinzl M., Steegborn C., Hubel F., Steinberg S. (1996) Compilation of tRNA sequences and sequences of tRNA genes. Nucl. Acids Res. 24, 68-72.
6. Willis I.M. (1993) RNA polymerase III. Genes, factors and transcriptional specificity. Eur. J. Biochem. 212, 1-11.

DO DROSOPHILA RETROTRANSPOSON LTRs CONTAIN FUNCTIONAL SITES CAPABLE OF PROVIDING HEAT SHOCK-INDUCIBLE TRANSCRIPTION?

**** Furman D.P., Katokhin A.V., Oshchepkov D.Yu., Stepanenko I.L.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: furman@bionet.nsc.ru
*Corresponding author

Key words: *D. melanogaster, retrotransposon, LTR, heat shock element*

Abstract

Motivation: Considerable increases in the transposition rate of *Drosophila* retrotransposons under the effect of various stress factors, in particular, heat shock, have been described. The proposed interpretation of this phenomenon (heat shock–activated transcription) as well as the question on whether the observed changes in retrotransposon distribution in the genome actually result from genuine *de novo* transpositions, that is, whether they are really induced by the effects experienced, yet remain controversial (Arnault et al., 1997). The heat shock–induced transcription requires the presence of specialized regulatory sites—the so-called heat shock elements (HSEs)—within the corresponding promoters. If retrotransposons are capable of being stress-activated according to the heat shock response mechanism, the LTRs whereto the regulatory elements necessary for initiating the transcription of retrotransposons are localized should contain HSEs as well. The goal of this work was to search retrotransposon LTRs for HSEs and evaluate their functionality.

Results: The results of computer analysis of retrotransposon LTRs within 11 families using specialized techniques for functional site recognition in combination with the published data on promoter structures of heat shock–activated genes prevents from speaking about functionally robust HSEs in retrotransposon LTRs.

Introduction

Mobile elements are an important component of eukaryotic genomes, representing, for instance, up to 10–15% of fruit fly genome (Vieira et al., 1999). The number of mobile elements and their localization in the genome are usually rather stable: the rates of their spontaneous transpositions are comparable with those of spontaneous mutations, amounting to ~10^{-5} per site per generation (Nuzhdin, Mackay, 1995; Dominguez, Albornoz, 1996). However, these rates might increase drastically under particular conditions. The ability of mobile elements—retrotransposons, in particular—to move within the genome makes them a mighty source of mutation-based variation.

Several publications report a considerable increase in the transposition rate of retrotransposons upon exposures to stress (the review by Arnault, Dufournel, 1994 and references therein). It is known that the cell responds to any stress exposure by switching on the universal heat shock mechanism involving a protein, the heat shock factor (HSF), whose active trimer binds to the corresponding functional elements—heat shock elements (HSEs), localized to the promoter regions of heat shock–activated genes, and initiates their transcription (Lis, Wu, 1993; Morimoto, 1998). The same mechanism is proposed for explaining the stress induction of transpositions (Vasil'eva et al., 1997); this implies the presence of the corresponding HSEs in retrotransposons, more precisely, in their promoters localized to LTRs.

The goal of this work was to analyze the structure of LTRs and evaluate the functionality of potential HSEs once they were found.

Materials and Methods

The potential HSEs in LTR sequences were recognized using the program SITECON (Oshchepkov et al. (a), this issue). The parameters involved in recognition were optimized to maximally reduce type I error (false negatives) to zero.

LTR sequences of eleven *D. melanogaster* retrotransposon families, retrieved from the FlyBase database (http://flybase.harvard.edu:7081/transposons/lk/melanogaster-transposon.html)—*297, 17.6, yoyo, HMSBeagle, mdg1, mdg3, Dm412 (mdg2), copia, blood, roo (B104),* and *tirant*—were analyzed (test sample).

As a training sample, 62 HSE sequences from the promoters of heat shock–induced genes whose functionality had been confirmed experimentally in the literature (http://wwwtest.bionet.nsc.ru/mgs/ papers/stepanenko/hs-trrd/) were used; as a negative control sample, a set of randomly generated sequences.

Results

All the known functionally active HSEs from fruit fly genes of heat shock proteins (*hsp*) exhibit the following distinctive features (Fig. A):

1) Contain at least three motifs for binding to three heat shock factor subunits (nGAAnnTTCnnGAAn or nTTCnnGAAnnTTCn); moreover, at least two of them stringently conform to the consensus (Lis, Wu, 1993);

2) As a rule, contain GAGA factor (GAF) recognition sites or are encompassed by them; GAGAG/CTCTC is the canonical GAE (GAF recognition element) motif; however, certain publications report that GAG/CTC is also capable of binding GAF (Wilkins, Lis, 1997); and

3) Are clustered in TATA+ promoters 50 to 300 bp upstream of the transcription start (Fernandes et al., 1995).

Analysis of LTR structures within eleven retrotransposon families by SITECON detected one potential HSE in each of seven families, namely, *gypsy, mdg3, 17.6, Dm412(mdg2), blood, tirant,* and *yoyo* (Fig. B, Table).

Functionalities of the putative HSEs detected were then evaluated qualitatively according to the criteria listed above.

Four HSEs—from *gypsy, mdg3, tirant,* and *yoyo* LTRs—meet the first condition: the combinations of motifs they contain are similar to functional HSEs from fruit fly *hsp* genes (Fig.).

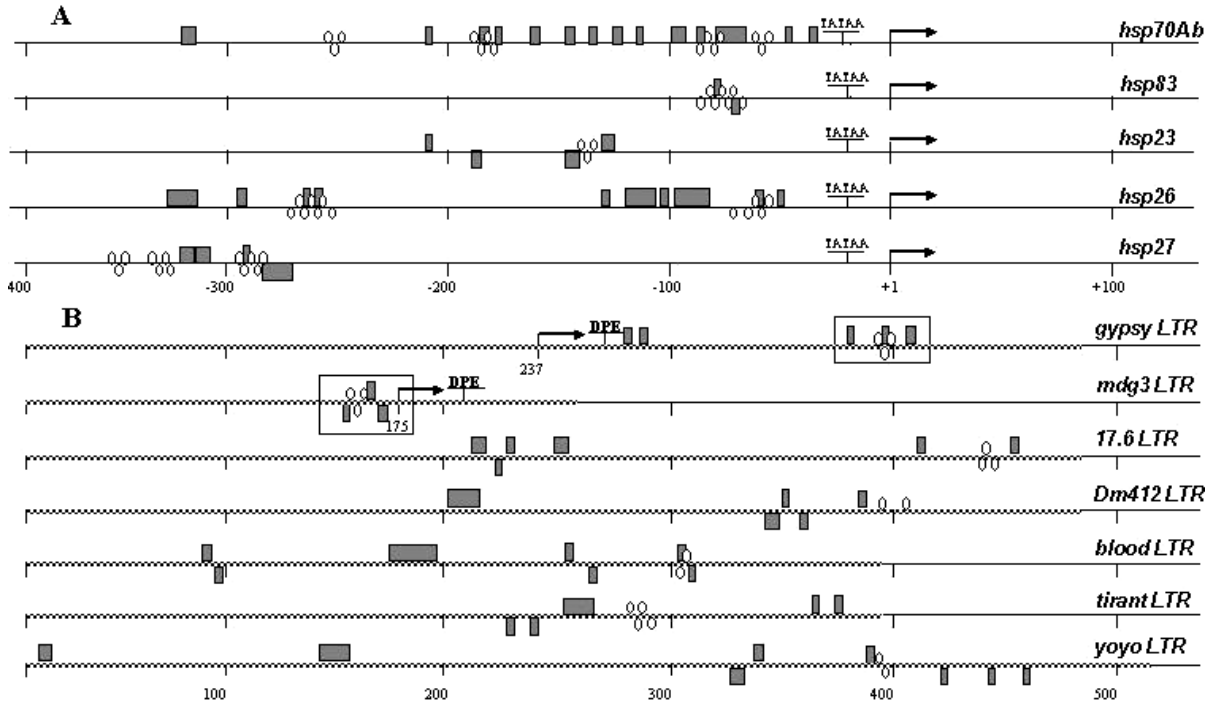


Fig. (A) Map of HSEs and GAEs in promoters of fruit fly *hsp* genes and (B) map of potential HSEs and GAEs in LTRs of fruit fly retrotransposons: positions are indicated in bp; thin lines, fragments of gene sequences; heavy lines, LTR sequences; ovals, HSEs (GAA motif, above the line; TTC, below); shaded rectangles, GAEs (CTC motif, above the line; GAG, below); the lengths of rectangle reflect approximately the lengths of GAE motif clusters; right-angle arrows, transcription starts; positions of TATA boxes and DPEs are indicated; and big empty rectangles cover the two best recognized HSEs.

Table. Potential HSEs detected in retrotransposon LTR sequences.

Retrotransposon	Putative HSEs detected*
gypsy	agac**GAA**cc**TC**agc**GAA**agaa
mdg3	tcga**GAA**ac**TT**atc**GA**ctaat
17.6	cata**TTC**g**GAA**cgg**TC**cattt
Dm412	aaaa**GAA**gaaatt**GAA**taaat
blood	ctgc**Ta**C**tc****GAA**gagataaga
tirant	ttttGAAcTTCaaGAAagTCa
yoyo	cgtaaa**C**tc**GAA**cc**TTC**ttaa

*The trinucleotide motifs for binding HSF monomers are marked with capitals; bi- and tripartite structures formed by binding motifs are underlined.

Two of the four HSEs, those from *gypsy* and *mdg3,* meet the second criterion; however, applying the third condition to evaluation of their functionality, we must admit that they also could hardly be considered as functionally robust. Note that the HSE from *gypsy* LTR is located approximately 160 bp downstream of the transcription start, that was localized experimentally with a high accuracy (Arkhipova, Il'in, 1991). Moreover, the *gypsy* promoter does not belong to the TATA-containing external promoters; on the contrary, it is an internal DPE-containing promoter (Arkhipova, Il'in, 1991; Burke, Kadonaga, 1997; Fig. B).

The HSE discovered in the *mdg3* LTR is located approximately 30 bp upstream of the transcription start, determined reliably (Arkhipova, Il'in, 1991). However, note that none of the functional HSEs of the fruit fly *hsp* promoters studied are located so close to the transcription start. In addition, the *mdg3* promoter also does not belong to the TATA-containing external promoters (Arkhipova, Il'in, 1991).

These two last HSEs are most likely candidates for functionally robust elements of this type, provided experimental evidence of fruit fly heat shock–induced promoters other than external TATA-containing type (yet unknown). Anyway, the functionalities of the predicted candidate sites require direct experimental confirmation.

Discussion

Thus, none of the HSEs detected meet the totality of criteria applied. This suggests that the LTRs of eleven retrotransposon families analyzed are not likely to contain functionally robust HSEs, In turn, this disputes the possibility of retrotransposon transposition activation in the fruit fly genome through initiation of their transcription according to the heat shock response mechanism with involvement of HSFs. If the changes in retrotransposon distribution pattern occurring upon temperature and/or other stress exposures actually are the result of their mobilization, alternative interpretations of this phenomenon should be searched for. For example, one variant is a weakening of the transcription autoregulation mediated by the silencers located in LTR 3' regions of certain retrotransposons (Faure, 1999 and references therein). Another possibility is an impaired control of subsequent retrotransposition stages at the level of post-transcription (Aravin et al., 2001).

Finally, it has been demonstrated that nucleosomal organization and local conformation of genomic DNA at the insertion sites play certain role in integration of various retrotransposons (Katokhin et al., this issue; Oshchepkov et al. (b), this issue). Thus, it is not unreasonable that changes in these parameters under a stress exposure are also capable of influencing essentially the transposition of retrotransposons.

Both incompleteness and inconsistency of these data allow only the hypothetical behavior patterns of retrotransposons belonging to various families under stress exposures to be discussed. Evidently, the final answer to the question on what mechanisms actually determine the retrotransposon behavior under such conditions requires further experimental studies.

References

1. Aravin A.A., Naumova N.M., Tulin A.V. et al. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. Curr. Biol. 11:1017-1027.
2. Arkhipova I.R., Il'in Yu.V. (1991). Organization of promoter regions in *Drosophila* retrotransposons. Mol. Biol. (Mosk.). 25(1):69-76.
3. Arnault C., Loevenbruck C., Biemont C. (1997). Transposable element mobilization is not induced by heat shocks in *Drosophila melanogaster*. Naturwissenschaften. 84(9):410-414.
4. Arnault C., Dufournel I. (1994). Genome and stresses: reactions against aggressions, behavior of transposable elements. Genetica. 93(1-3):149-160.
5. Burke T.W., Kadonaga J.T. (1997). The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of Drosophila. Genes Dev. 11:3020-3031.
6. Dominguez A., Albornoz J. (1996). Rates of movement of transposable elements in *Drosophila melanogaster*. Mol. Gen. Genet. 251(2):130-138.
7. Fernandes M., Xiao H., Lis J.T. (1995). Binding of heat shock factor to and transcriptional activation of heat shock genes in *Drosophila*. Nucl. Acids Res. 23(23):4799-804.
8. Faure E. (1999). A sequence of the U5 region of *Drosophila* 1731 retrotransposon long terminal repeat (LTR) trans-represses the LTR-directed transcription. Biokhim. (Mosk.). 64(6):678-92.
9. Katokhin A.V., Furman D.P., Levitsky V.G., Katokhina L.V. Nucleosomal organization of *Drosophila* retrotransposon insertion sites. This issue.
10. Lis J., Wu C. (1993). Protein traffic on the heat shock promoter: parking, stalling, and trucking along. Cell. 74(1):1-4.
11. Morimoto R.I. (1998). Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators. Genes Dev. 12(24):3788-3796.
12. Nuzhdin S.V., Mackay T.F. (1995). The genomic rate of transposable element movement in *Drosophila melanogaster*. Mol. Biol. Evol. 12(1):180-181.
13. Oshchepkov D.Yu., Turnaev, I.I., Vityaev E.E. (2002 a). SITECON: a method for recognizing transcription factor binding sites basing on analysis of their conservative physicochemical and conformational properties. This issue.
14. Oshchepkov D.Yu., Furman D.P., Katokhin A.V., Katokhina L.V. (2002 b). Detection of conservative conformational properties of insertion sites for Drosophila retrotransposons. This issue.
15. Vasil'eva L.A., Ratner V.A., Bubenshchikova E.V. (1997). Stress induction of retrotransposon transposition in *Drosophila:* reality of the phenomenon, characteristic features, and possible role in rapid evolution. Genetika. 33(8):1083-1093.
16. Vieira C., Lepetit D., Dumont S., Biemont C. (1999). Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol. Biol. Evol. 16(9):1251–1255.
17. Wilkins R.C., Lis J.T. (1997). Dynamics of potentiation and activation: GAGA factor and its role in heat shock gene regulation. Nucl. Acids Res. 25:3963-3963.

ANALYSIS OF TOPOLOGICAL REPRESENTATIONS OF TRANSCRIPTIONAL REGULATORY REGIONS

[1] Sand O., [2] Vu T.D., [1] Gilbert D., [1] Viksna J.

[1] Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, Glasgow, UK
[2] Department of Computing, School of Informatics, City University. London, UK
e-mail: osand@brc.dcs.gla.ac.uk

Key words: *transcriptional regulatory region, composite element, pattern discovery, topological representation, graph*

Resume

Motivation: Transcriptional regulatory regions of genes contain the information necessary to modulate their selective expression in various conditions. Understanding how the structure of these regions influences the expression of genes is a fundamental problem in biology.

Results: A topological representation of transcriptional regulatory regions (TREGS) has been developed. Information on regulatory regions was collected from publicly available databases and stored in a relational database. The data has been cleaned to remove synonyms, alternative spellings and typographical errors. Different approaches are now tried to generate a TREGS pattern database.

Introduction

Transcriptional regulatory regions (TRR) of genes play an essential role in genomes, since they mediate the selective synthesis of proteins in response to the availability of metabolite in the external medium, the developmental stage, the presence of a stress, etc. The ability of a sequence to regulate the level of transcription of a neighboring gene is due to the action of very short segments that are specifically recognized by transcription factors. In higher organisms, regulatory sites tend to aggregate in so called composite elements (CE), i.e. DNA regions where several transcription factors bind simultaneously and interact either synergistically or competitively, contributing to a highly specific pattern of gene transcriptional regulation (Kel et al., 1995).

Methods and Algorithms

We have developed a topological representation of transcriptional regulatory regions (TREGS) providing a high-level abstraction we are using for pattern matching and pattern discovery. This representation is in the form of a context-sensitive grammar, describing binding sites and their associated factors as well as interactions between bound factors. We have generalized this representation to permit the definitions of patterns over TREGS, in a manner similar to that which we have developed for protein topology (Gilbert et al., 1999). These patterns are defined in terms of regulatory element aggregations rather than nucleotide sequences.

We have collected information on regulatory regions from publicly available databases: COMPEL (http://compel.bionet.nsc.ru/), TRRD (http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/trrdintro.html) and TRANSFAC (http://transfac.gbf.de.transfac/index.html) and stored this in a relational database. The data has been cleaned to remove synonyms, alternative spellings and typographical errors, using a program that we have developed based on our FURY programming system (Gilbert and Schroeder, 2000). In addition, we have developed a method to automatically generate a graphical representation ("cartoon") of a TREGS.

We are now developing a TREGS pattern database. To start, we have compiled a set of patterns by hand, and stored them in that database, together with an automatically generated cartoon for each pattern, annotated with the TREGS it describes. We are now designing methods to automatically discover TREGS patterns from TRR data. Two approaches are considered. The first one is sequence based (Brazma et al., 1998). It is focusing on a subset of our pattern language, comprising regular expressions over TRR binding element sequences and uses a dynamic programming algorithm to find the longest common subsequence (LCS) of each pair of sequences (Pevzner, 2000). The other is graph based. It uses the Bron-Kerbosch algorithm (Bron, Kerbosch, 1973) to find the maximal clique of the edge product graph of each pair of instance graphs.

Implementation and Results

The relational database was implemented in the MySQL DBMS. The LCS algorithm (Pevzner, 2000) and maximal clique algorithm (Bron, Kerbosch, 1973) were implemented in Java.

We have grouped the 607 genes in our database by performing an all against all pair-wise pattern discovery over the TREGS and hierarchically clustering using the OC program (Barton, 1993). The clustering is done using similarity

measures and means linkage analysis. We then generated unique non-null TREGS patterns for the clusters and computed their compression. We have evaluated the goodness of those patterns with reference to the entire TREGS database.

Our initial results have shown that

(i) Many of our groups of genes are associated with 'good' (characteristic) patterns.

(ii) The overlapping nature of transcription factors binding sites has very little effect on the goodness of the patterns discovered.

Discussion

The two algorithmic approaches produce different clusters and the LCS based pattern discovery method produces more clusters than the clique based method. It is mainly because the LCS method allows inserts between the sites and the clique one does not. Another limitation of the graph-based approach is the very limited number of known composite elements (graph edges). The genes for which there are known composite elements are therefore not representative of the whole dataset. A general limitation of this research is that the number of transcription factors is high relative to the number of known binding sites for them. This reduces the chance to find patterns. Hopefully, the identification of new binding sites will be accelerated in the near future, thanks to *in silico* help.

We will work next on evaluating the relationship between the discovered TREGS patterns and the expression profiles of the genes sharing the patterns.

Acknowledgements

This research was supported by a European Community Marie Curie Fellowship. The European Commission is not responsible for any views or results expressed.

References

1. Barton C.J. (1993) OC- A cluster analysis program. [http://www.compbio.dundee.ac.uk/Software/OC/oc.html]
2. Brazma A., Jonassen I., Eidhammer I., Gilbert D. (1998) Approaches to the automatic discovery of patterns in biosequences. J. of Computational Biol. 5(2), 277-303.
3. Bron C., Kerbosch J. (1973) Algorithm 457 – finding all cliques of an undirected graph. Commun. ACM. 16, 575-577.
4. Gilbert D.R., Westhead D.R., Nagano N., Thornton J.M. (1999) Motif-based searching in TOPS protein topology databases. Bioinformatics. 15(4), 317-326.
5. Gilbert D., Schroeder M. (2000) FURY: Fuzzy unification and resolution based on edit distance. BIBE 2000: IEEE International Symposium on Bio-Informatics and Biomedical Engineering, November 8-10 2000.
6. Kel O.V., Romaschenko A.G., Kel A.E., Wingender E., Kolchanov N.A. (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. Nucl. Acid Res. 23, 4097-4103.
7. Pevzner P.A. (2000) Computational molecular biology: an algorithmic approach. In Istrail S., Pevzner P., Waterman M. (Eds). MIT Press, Cambridge, Massachusetts. 96-97.

VISUALIZATION OF DNA SEQUENCES BY COLOR CUBE TRANSFORMATION

[1] Cheremushkin E.S., [1,2] Kel A.E., [3] Lobiv I.V., [3] Murzin F.A., [3] Polovinko O.N.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: cher@bionet.nsc.ru
[2] BIOBASE GmbH, Halchtersche Strasse 33, 38304Wolfenbuettel, Germany, e-mail: ake@biobase.de
[3] Institute of Informatics Systems, SB RAS, Novosibirsk, Russia, e-mail: murzin@iis.nsk.su
Corresponding author: cher@bionet.nsc.ru

Key words: *visualization of DNA, motif search*

Resume

Motivation: The fast grows of DNA sequence information rises a challenge to the bioinformatic community to develop new ways of efficient analysis of DNA sequence structure. There is vast variety of methods and algorithms that were developed in the field of image analysis and signal processing that can be efficiently used for analysis of DNA sequences. New convenient methods for representation and visualizing the information containing in DNA sequences are necessary.

Results: We have developed several methods of visualization of the primary structure of DNA sequences. These methods were applied for analysis of promoter sequences. We believe that these methods could serve as a basis for instrumental tools suitable for better understanding of DNA sequence data.

Availability: http://www.skypiece.ru

Introduction

Fast growing of DNA sequence information opens a new field of research devoted to the analysis of regularities in the structure of DNA sequences (Doolittle, 1997; Maley, Marshall, 1998). Such regular features of DNA sequence correspond to different functional important structures in genome. Coding regions of genes bring specific regular features that correspond to the triplet code and to the amino acid compositions of the encoded proteins. This results in emerging of short repeats in DNA sequences (Kolchanov et al., 1988) as well as long correlations (Herzel et al., 1999). Observation of regular DNA features in non-coding regions of genomes can help in understanding of the mechanisms of functioning and evolution of genomes.

The stable development of a life of organisms suggests, that there are more dependencies in a genome, than it is known at present (Jeffrey, 1990).

A convenient representation of information containing in a genome can help us to understand these dependences. In particular for this purpose, the visualization can be efficiently used (Burma et al., 1992; Goldman, 1993; Solovyev, 1993; Oliver et al., 1993; Deschavanne et al., 1999).

We have developed several methods of visualization of the primary structure of DNA sequences using a color cube transformation. Using this methods we have analyzed promoters of liver enriched genes. The problem of comparison promoter structures become actual in the course of functional annotation of genomes (Cheremushkin, Kel, 2002). The created color images of the alignment of these promoters help us to identify sub-regions of elevated similarity that may correspond to specific functionally important signals in these promoters.

Methods

Visualization of DNA sequences by color cube transformation.

Let $\overline{S}$ be a sequence in an alphabet consisting of four letters A, C, G, T. The k - th element of the sequence will be denoted by s_k , and M be a length of the sequence.

Suppose a positive number N is given. Denote by $BL_N[i]$ a subsequence of $\overline{S}$ having a length equal to N and beginning of i - th position, i.e. $BL_N[i] = s_i \ldots s_{N+i-1}$.

Let $n_A[i,N]$, $n_C[i,N]$, $n_G[i,N]$, $n_T[i,N]$ be the numbers of letters A, C, G, T in the considered subsequence $BL_N[i]$ respectively. If i, N are fixed we will write for brevity n_A, n_C, n_G, n_T.

It is easy to see that $n_T = N - (n_A + n_C + n_G)$. It means that it is sufficient to study only three components. From here frequencies $p_A = n_A / N$, $p_C = n_C / N$, $p_G = n_G / N$ can be calculated.

There exists the natural function from an interval [0,1] of real numbers onto a set of integers $\{i : 0 \le i \le 255\}$ defined by the formula $f(x) = \text{int}(255 \times x)$. Let us introduce $\overline{p}_A = f(p_A)$, $\overline{p}_C = f(p_C)$, $\overline{p}_G = f(p_G)$. Than a triple $\langle \overline{p}_A, \overline{p}_C, \overline{p}_G \rangle$ may be considered as the components of colors $\langle R, G, B \rangle$ respectively.

The color image may be defined by three matrices $S = (S_R, S_G, S_B)$, $S_R = S_R(i,j)$, $S_G = S_G(i,j)$ $S_B = S_B(i,j)$, $0 \le$ i $\le$ n–1, 0$\le$ j $\le$m–1.Usually the values of $S_R(i,j)$, $S_G(i,j)$, $S_B(i,j)$ change from 0 to 255. Where n and m correspond to the width and the height of the image. A set of triples $\{(r,g,b) : 0 \le r,g,b < 255\}$ is called the color cube. We make a transformation of the frequencies of nucleotides into colors of the color cube.

The first algorithm of visualization

Let us fix the width and height of the image: *n* and *m.* Suppose that two positions i_1, i_2 on a sequence $\overline{S}$ are given, $i_1 - i_2 \le n \cdot m$ and $i_1 \le k \le i_2$. Let us consider a window $BL_N[k]$ of the size *N* that is moving along the given sequence $\overline{S}$. Then, we compute the corresponding triple $\langle \overline{p}_A, \overline{p}_C, \overline{p}_G \rangle$ for every position k.

Therefore we write

$\langle \overline{p}_A, \overline{p}_C, \overline{p}_G \rangle = \langle \overline{p}_A(k), \overline{p}_C(k), \overline{p}_G(k) \rangle = \langle R(k), G(k), B(k) \rangle$.

Now we can construct the following image

$$S_R(i,j) = \begin{cases} R(i_1 + m \cdot i + j - 1) & \text{if } i_1 + m \cdot i + j - 1 \le n \cdot m; \\ 0 & \text{otherwise}; \end{cases}$$

$$S_G(i,j) = \begin{cases} G(i_1 + m \cdot i + j - 1) & \text{if } i_1 + m \cdot i + j - 1 \le n \cdot m; \\ 0 & \text{otherwise}; \end{cases} \quad (1)$$

$$S_B(i,j) = \begin{cases} B(i_1 + m \cdot i + j - 1) & \text{if } i_1 + m \cdot i + j - 1 \le n \cdot m; \\ 0 & \text{otherwise}. \end{cases}$$

It means that we fill the pixels in the image in a process of obtaining the components $\langle R, G, B \rangle$. At the beginning we are filling the upper row, i.e. $i = 0$. Than we are filling the first row and so on. If the number of necessary components of color is not sufficient, then we fill $\langle 0,0,0 \rangle$, i.e. the black color.

By taking one sequence after another, we can construct the second image, the third image etc. As a result we obtain a sequence of images. They are called frames. The sequence of frames allows us to construct a film, which can be represented in a form of AVI-file.

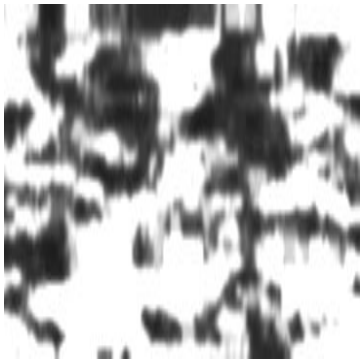


Fig. 1. The successive filling of pixels in the image in the process of obtaining the components $\langle R, G, B \rangle$ represented on gray scale picture. The color depth is calculated via calculating the nucleotide frequencies in the 10 bp window. The more A+T rich regions of the sequence are characterized by the more dark color of the image.

The second algorithm of visualization

Suppose we have any function $g : \{A, C, G, T\} \to \{i : 0 \le i \le 255\}$. Than our sequence $\overline{S}$ generates a sequence of integers constructed by the following rule

$g[\overline{S}] = g(s_1)g(s_2)g(s_3)\ldots$

Thus every three numbers staying near each other can be considered as the color components, i.e. we obtain the following sequence of triples

$\langle g(s_1)g(s_2)g(s_3) \rangle, \langle g(s_4)g(s_5)g(s_6) \rangle, \langle g(s_7)g(s_8)g(s_9) \rangle \ldots = = \langle R_1, G_1, B_1 \rangle, \langle R_2, G_2, B_2 \rangle, \langle R_3, G_3, B_3 \rangle, \ldots$

Analogously moving along the given sequence we construct a sequence of images.

Also we can consider more complicated functions

$g_2 : \{A,C,G,T\}^2 \to \{i : 0 \le i \le 255\}$ or $g_3 : \{A,C,G,T\}^3 \to \{i : 0 \le i \le 255\}$.

In these cases pairs or triples of letters in the initial sequence $\overline{S}$ are investigated. It is well known that pairs or triples are more informative [8].

The third algorithm of visualization

Consider the sequence of triples obtained in the first algorithm of visualization $\langle \overline{p}_A(k), \overline{p}_C(k), \overline{p}_G(k) \rangle$, $k \ge 1$. They can be considered as coordinates in three-dimensional space.

Suppose the function $h : [0,1]^3 \to \{i : 0 \le i \le 255\}^3$ is given. This function may be represented in the form $h(x,y,z) = \langle h_R(x,y,z),\ h_G(x,y,z),\ h_B(x,y,z) \rangle$. Thus we obtain some image in three-dimensional space which can be useful to see better a structure of the sequence $\overline{S}$. Examples of such visualization are given in the Fig. 2. It can be used for visual comparison of different sequences.

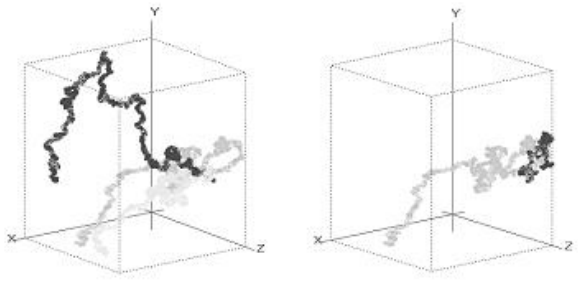


Fig. 2. Visualization of promoters with the help of the function $h : [0,1]^3 \to \{i : 0 \le i \le 255\}^3$. a) Visualization of three sequences for 5' regions of c-myc genes for human, mouse and rat; b) comparison of human c-myc gene 5' region with the coding region of mouse a'-actin skeletal gene.

Results and Discussion

We analyze a set of liver specific promoters that contains 66 sequences of the same length L=100 (from –100 to –1). The goal was to reveal potential common signals in this sequences applying local alignment and visualization algorithms. To do local alignment we developed a straightforward approach by using a sliding window and by comparison oligonucleotides in the window. The description of the formalism is given below.

For a given window size *w* and for a window location variation *d* we calculate a similarity between two promoter sequences $\overline{S}'$ and $\overline{S}''$ using the following formula:

$$sim(\overline{S}', \overline{S}'') = \frac{1}{L-w+1} \cdot \left[\sum_{i=1}^{L-w+1} D_i(\overline{S}', \overline{S}'') \right] \qquad (2)$$

which is the sum of local similarity *D* through all positions of the sequence. Here, local similarity $D_i(\overline{S}', \overline{S}'') = \frac{1}{K} \cdot \left[\sum_{j=-d}^{d} \frac{1}{|j|+1} H(\overline{S}', \overline{S}'', i, i+j) \right]$ is calculated as a sum of similarities between oligonucleotide in the position *i* of the first sequence and every oligonucleotide in the second sequence located in the positions from *i - d* to *i + d*. Similarity between two oligonucleotides is

$$H(\overline{S}', \overline{S}'', i, i+j) = \frac{1}{w} \cdot \left[\sum_{s=0}^{w-1} \begin{cases} 1, & if\ \overline{S}'_{i+s} = \overline{S}''_{i+s+j} \\ 0, & otherwise \end{cases} \right]; \left(K = \sum_{j=-L'}^{L'} \frac{1}{|j|+1} \right).$$

We have applied the described formalism to the set liver specific promoters ($\overline{S}^1 \ldots \overline{S}^K$) with *w=15* and *d=15*. Then we search for a path ($p_1..p_{K-1}$) with a property that $\sum_{i=1}^{K-1} sim(\overline{S}^{p_i}, \overline{S}^{p_{i+1}}, L) \to \max$, where sim is the similarity between promoters p_i and p_{i+1}.

To do this we apply some heuristic analogous to the so-called greedy algorithms. The list of promoter sequences was then ordered according to the found similarity path ($p_1..p_{K-1}$). This list was visualized then using the first visualization algorithm described above with the window length for nucleotide frequency calculation *N=10* and the width of the image *m=100*. So, it means that in every line of the image we see a visualization of one promoter sequence (see Fig. 3). Since the promoters were ordered according their pair wise similarity the unicolor places in the image can correspond to the common signals in this promoters that are located close to each other in similar promoters. To check that, for all pairs of sequences $\overline{S}^{p_i}, \overline{S}^{p_{i+1}}$

we search for T not intersected fragments ($B_1 \ldots B_T$) of length P<L with a maximal similarity $sim^*(i,j) = sim(\overline{S}^{p_i}, \overline{S}^{p_{i+1}}, j, P)$, where i is the sequence number in the path ($p_1..p_{K-1}$) and j is the first position of the fragment. We mapped some of found fragments to the image (see the outlined regions in the Fig. 3). One can see the good correspondence of the found fragments to the unicolor regions of the image.

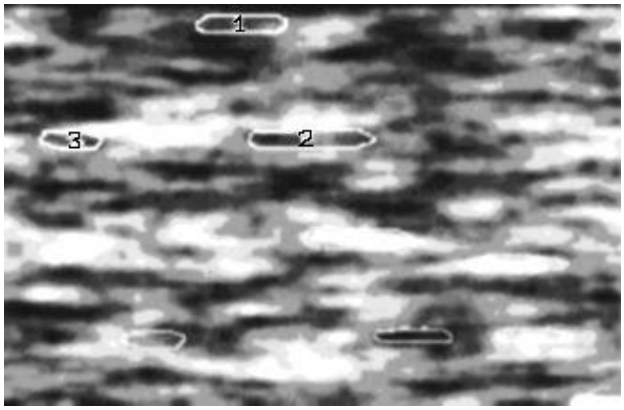


Fig. 3. Visualization of the alignment of the set of liver specific promoters. The unicolor regions correspond to the most similar regions of the promoters.

The developed technique can be used for analysis of the sets of aligned regulatory sequences to present visually the information about similarity of local regions, their nucleotide compositions and overall sequence structure.

In future the interactive techniques of using such images for navigating through the sets of aligned sequences and for finding potential signals in local regions will be developed. Several image analysis and image recognition approaches can be applied for comparison of sets of sequences.

Acknowledgements

Parts of this work was supported by Siberian Branch of Russian Academy of Sciences and by the grant of Volkswagen-Stiftung (I/75941).

References

1. Doolittle R.F. (1997) Microbial genomes opened up. Nature. 392, 339–342.
2. Maley L.E., Marshall C.R. (1998) The coming of age of molecular systematics. Science. 279, 505–506.
3. Kolchanov N.A., Kel' A.E., Solov'ev V.V. (1988) Convergent origin of repeats in the genes coding for globular proteins. The modelling of the convergent origin of direct repeats. [Article in Russian] Zh. Obshch. Biol. 49, 723-728.
4. Herzel H., Weiss O., Trifonov E.N. (1999) 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics. 15, 187-193.
5. Jeffrey H.J. (1990) Chaos game representation of gene structure. Nucl. Acids Res. 18, 2163–2170.
6. Burma P.K., Raj A., Deb J.K., Brahmachari S.K. (1992) Genome analysis: a new approach for visualization of sequence organization in genomes. J. Biosci. 17, 395–411.
7. Goldman N. (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. Nucl. Acids Res. 21, 2487–2491.
8. Solovyev V.V. (1993) Fractal graphical representation and analysis of DNA and protein sequences. Biosystems. 30, 137-160.
9. Oliver J.L., Bernaola-Galva P., Guerrero G., Roma R. (1993) Entropicprofiles of DNA sequences through chaos-game-derived images. J. Theor. Biol. 160, 457–470.
10. Deschavanne P.J., Giron A., Vilain J., Fagot G., Fertil B. (1999) Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences. Mol. Biol. Evol. 10, 1391-9.
11. Cheremushkin E., Kel A. (2002) PromoterFootprint: A new method for alignment of regulatory genomic sequences. Phylogenetic footprinting of TF binding sites. In Liliana Florea, Brian Walenz, Sridhar Hannenhalli (eds) Currents in Computational Molecular Biology 2002. RECOMB 2002, Washington D.C. 40-41.

BINDING SITES FOR TRANSCRIPTION FACTORS: WHAT CAN TELL AN ISOLATED SEGMENT OF REGULATORY DNA?

[3] Lifanov A.P., [1] Nazina A.G., [2*] Makeev V.J., [1] Papatsenko D.A., [4] Regnier M., [1] Desplan C.

[1] Department of Biology, New York University, New York, USA
[2] State Scientific Centre NIIGenetika, Moscow, USA
[3] Institute of Chemical Physics, RAS, Moscow, Russia
[4] INRIA, Roquencour, France
[*]The corresponding author. e-mail: makeev@imb.ac.ru

Key words: *Drosophila melanogaster, early developmental enhancers, binding sites, transcription factors*

Motivation: Basal promoters and enhancers, the regulatory regions of higher eukaryotes, often contain important regulatory elements in many copies. This opens a possibility of extraction of binding sites for transcription factors (BSTF) by a word count analysis of an isolated regulatory segment without any cross-specie comparison. This opportunity is particularly promising for analysis of tissue specific transcription, for which construction of a good collection of co-regulated genes is a very difficult experimental problem.

Results: We have developed a conceptually simple algorithm ScanSeq, which extracts frequent motifs from a given region. Selecting as a model a well-characterized system of early developmental enhancers of *Drosophila melanogaster* we have compared experimental BSTF maps with those constructed from highly redundant words. It is found that clustering of words similar to the strong binding sites within an enhancer provides sufficient information to restore the majority of functional binding sites. Our approach also yields good candidates for further experimental verification.

Availability: see http:\\homepages.nyu.edu\~dap5\

Introduction

Most of the currently accepted strategies for identification of binding sites for transcription factor (BSTF) rely on extraction of binding sites from a set of unaligned, but functionally related regulatory sequences e.g. (Lawrence et al., 1993; Bailey, Elkan, 1995). This strategy is excellent for prokaryotic regulatory elements with their long and conservative BSTF. In higher eukaryotes this strategy fails because their BSTF are short and divergent. However, there are evidences that in many cases the oligonucleotides that can be recognized by transcription factors are found in many copies in the eukaryotic regulatory modules (Lewis et al., 1995; Wagner, 1999; Berman et al., 2002; Markstein et al., 2002). Hence, one can hope to extract BSTF from an individual sequence exploring motif redundancy. The main problem here is the lack of training data, since the experimentally verified maps of BSTF in eukaryotic regulatory sequences are often incomplete and BSTF positioning is often unreliable.

Data: We built experimental annotations for BSTF positioning of a set of early developmental enhancers from *Drosophila melanogaster*. A library of footprint for 7 BSTF types (Bicoid, Caudal, Ftz, Hunchback, Knirps, Kruppel, and Tramtrak) was collected. For each BSTF types the footprints were aligned with CLUSTAL, and the conservative region of local multiple alignment was identified with column information content greater than $-\log_2(3/4)$. Positional Weighted Matrices were built from conservative segment of alignment (see details in Papatsenko et al., 2002). Two kinds of reference maps were built with PWM: (i) the segments covered by footprints were search with PWM to built "the refined map"; (ii) the whole sequence of each regulatory model were search with PWM to built "the consistent map". We used the difference between "the refined map" and "the consistent map" to evaluate the possible error in the experimental data. In the end the training sampling of ten developmental enhancers with known positions of regulatory sites and the reference sampling of two enhancers known to be regulated with the same set of transcription factors but with the experimental BSTF distribution unknown was created see (Papatsenko et al., 2002).

The algorithm for prediction of BSTF includes three basic stages. In the first stage, for each *m*-letter word in the sequence (the "seed" word) all the similar words with no more than *k* mismatches are counted. The resulting word family forms the initial motif assigned to the "seeded" word. All such motifs containing no less than 10 words are aligned and positional weighted matrices are built from the alignment. In the second stage the motif search is performed with this PWM. The expectation and the variance of the number of occurrences of PWM motif in the random sequence are also calculated in this stage and the Z score is calculated and assigned to the "seed" word. We used a precise formulae for variation of the number of occurrences of a motif, that take into consideration possible word overlapping. The "seed" words with the highest Z-scores are assembled in a map of potential BSTF distribution. The algorithm is implemented as a ScanSeq program, which can be found together with a number of programs at http:\\homepages.nyu.edu\~dap5.

Comparison with prediction was performed with three discrepancy measures: (i) Pearson association coefficient, *CC*, (Lakin, 1980), overlap quality, OQ, (Gelfand et al., 1996), and the logarithmic gain in the overlap quality, *PQ*, the log-ration of OQ and the expected OQ for random prediction with the given coverage with positives.

Results

We have found that in many cases the idea that the binding sites for transcription factors are represented in many ("weak") copies in the regulatory regions can be a rather good basis for prediction of transcription factor binding sites. The quality of prediction, however, dramatically depends on the parameters of the search. Almost in every sequence taken separately the seeded words with best Z-scores overlapped strongly with functional binding sites. However, our hope to find an efficient threshold on the Z-score, which would allow one to identify the correct binding sites in all sequences has failed. In this case, a more stable parameter is an expected coverage of the sequence with the BSTF. Although there are exception, in the surprising majority of enchancers BSTF covers about 25% of the overall length (Table 1).

We have performed several rounds of calculation with different values of the following parameters: the length of the seeded word, *m*, the number of mismatches allowed, *k*, the minimal number of shadows in the motif, and the expected enhancer with BSTF, *c*. It was found that in every case it was sufficient to allow no less than five occurrences of each BSTF type. In contrast the values of *m, k,* and *c* for which the best agreement was observed with the experimental and the predicted maps were different from sequence to sequence. In this case we increased recognition by piling together results obtained for different *m* in some range between m_{min} and m_{max} (Table 1).

Table 1. The comparison between the consistent and the predicted maps. c_{map} is the coverage of the consistent map with sites, *L* is the length of the enhancer, other notation is explained in the text. More results can be found at http:\\homepages.nyu.edu\~dap5

Sequence			Statistics			Best parameters				
NAME	L	*c*-MAP	CC	OQ	PQ	m_{min}	m_{max}	k_{max}	Z	c
EVE2	728	0.15	0.62	0.51	0.80	9	9	2	9.7	0.15
HAIRY6	547	0.65	0.55	0.59	0.05	7	9	2	6.3	0.73
HAIRY7	932	0.16	0.53	0.41	0.77	8	9	1	11	0.11
EVE37	508	0.29	0.52	0.46	0.43	8	9	1	4.9	0.29
TLL	480	0.15	0.46	0.37	0.65	11	12	2	3.7	0.16
IAB2	1745	0.07	0.46	0.33	0.89	9	11	4	22.3	0.10
KR730	718	0.32	0.43	0.40	0.31	8	9	1	3.8	0.33
SAL	516	0.22	0.42	0.32	0.24	12	14	4	8.4	0.54
FTZPROX	396	0.23	0.41	0.34	0.24	9	10	4	7.6	0.55
ENINT	900	0.20	0.34	0.29	0.39	7	7	1	4	0.23
AVERAGE	784	0.24	0.47	0.39	0.32				8.8	0.33

To evaluate potential of prediction of our method at unknown sequences we also evaluated the prediction quality over the whole set of the sequences with the same parameter set. It was found that the predicted map with piled results for 7–9 b.p. of the seed length and with 0 and 1 mismatches proved to yield the best agreement. The expected coverage was set as the average 0.24. The results, summarized in Table 3 indicate that the "default" parameters still worked well for most examples from the training set. Here we also compared the predicted results with the consistent map built for two enhancers, **runt5** and **eve4+6**¸ for which there is no experimental map of BSTF distribution.

Table 2. Prediction with the optimal set of parameters values. The "experimental set" has no experimentally verified BSTF.

Training set	Statistics			
	CC	OQ	PQ	Z
EVE2	0.50	0.38	0.59	4.7
HAIRY7	0.39	0.31	0.48	8.6
HAIRY6	0.38	0.32	0.20	9.5
TLL	0.35	0.28	0.46	4.5
ENINT	0.33	0.28	0.37	4.5
EVE37	0.32	0.29	0.30	7.7
KR730	0.30	0.29	0.27	6.5
IAB2	0.24	0.16	0.47	8.5
FTZPROX	0.23	0.23	0.24	3.2
SAL	0.18	0.19	0.19	4.1
EXPERIMENTAL SET				
EVE46	0.61	0.52	0.64	9.3
RUNT5	0.15	0.15	0.20	9

Discussion

One can see that at least for the system of early developmental enhancers our method works reasonably well. There were examples reported on BSTF clustering in other systems (Kondrakhin et al., 1995; Wagner, 1999) belonging to other species (*H. sapience* and *S. cerevisae* respectively). Thus there is a hope that this method would work for some other systems too. This is supported with our experiments at Drosophila rodopsine promoters (http:\\homepages.nyu.edu\~dap5).

The main difficulty in recognition arises from the dramatic variation in the native coverage of enhancers with BSTF. Indeed, one can compare iab2 with hairy6 with c equal to 0.07 and 0.65 respectively. This would hinder recognition in large collections of regulatory sequences or in a complete genome at least until some methods for evaluating the coverage would appear.

The other important question is the functional role of the multiple sites. Although this role is unclear, one can speculate that multiple sites are needed for cooperative recognition with many molecules of the factor, or for increasing the local concentration of protein in the enhancer environment. However, the experimental data supporting both models are poor.

Acknowledgements

Authors are pleased to M.Gelfand for many discussions. This work was supported by grants from National Science Foundation (IBN 0002958) and National Institutes of Health/National Eye Institute (EY13010) (D.P., A.N., and C.D.), Hughes Medical Institution grant (55000309), INTAS (99-1476) and the Russian Fund of Basic Research (02-04-49111) (V.M., and A.L) and French-Russian Lyapunov Institute grant 005 (M.R.)

References

1. Bailey T.L., Elkan C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning. 21, 51-80.
2. Berman B.P., Nibu Y. et al. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl Acad. Sci. 99, 757-62.
3. Kondrakhin Y.V., Kel A.E., Kolchanov N.A., Romashchenko A.G., Milanesi L. (1995). Eukaryotic promoter recognition by binding sites for transcription factors. CABIOS. 11, 477-488.
4. Gelfand M.S., Mironov A.A., Pevzner P.A. (1996). Gene recognition via spliced sequence alignment. Proc. Natl Acad. Sci. USA. 93, 9061-6.
5. Lakin G.F. (1980), Biometriya (The biometry). Moscow, Vyschaya shkola.
6. Lawrence C.E., Altschul S.F. et al. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 262, 208-214.
7. Lewis E.B., Knafels J.D., Mathog D.R., Celniker, S. E.(1995). Sequence analysis of the cis-regulatory regions of the bithorax complex of Drosophila. Proc. Natl Acad. Sci. USA. 92, 8403-7.
8. Markstein M., Markstein P., Markstein V., Levine M.S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. Proc. Natl Acad. Sci. 99, 763-8.
9. Papatsenko D.A., Makeev V.J., Lifanov A.P., Regnier M., Nazina A., Desplan C. (2002) Extraction of Functional Binding Sites from Unique Regulatory Regions: The Drosophila Early Developmental Enhancers. Genome Res. 12, 470–481.
10. Wagner A. (1999). Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics. 15, 776-84.

ALL HUMAN GENOME ANALYSIS FOR REGULATORY SIGNALS

[1, 2]* Wingender E., [1] Kel A.E., [1] Gößling E., [1] Hornischer K., [1] Lewicki-Potapov B., [1] Tchekmenev D., [1] Kel-Margoulis O.V.

[1] BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany
[2] GBF German Research Centre for Biotechnology, Mascheroder Weg 1, D-38124 Braunschweig, Germany
e-mail: ewi@gbf.de
*Corresponding author

Key words: *human genome, upstream regions, promoters, regulatory sequence signals, transcription factors, transcription factor binding sites*

Resume

Motivation: The availability of the complete human genome enables us to initiate a comprehensive and systematic investigation of its regulatory features and, thus, of modeling the dynamics of teh biological system it defines. For that purpose, we developed a set of database and software tools and applied them on a whole genome analysis for regulatory characteristics.

Results: A number of transcription factor binding patterns are significantly overrepresented in regulatory regions in general, whereas certain others are greatly underrepresented and therefore provide a much higher specificity of gene regulation.

Availability: Public versions of the tools mentioned in this article are available under http://www.gene-regulation.de . Professional versions of the databases and the software tools described are distributed by BIOBASE GmbH (http://www.biobase.de).

Introduction

All research on functional genomics heavily depends on thorough computational analyses of the genome under consideration. An important part of the biological function of a whole genome is represented by the way its genes are regulated. Over many years, information that is relevant for gene regulation in eukaryotes has been collected and stored in databases such as the Eukaryotic Promoter Database (EPD; Praz et al., 2002), TRANSFAC® (Wingender et al., 2001), TRRD (Kolchanov et al., 2002) or COMPEL (Kel et al., 1995; Kel-Margoulis et al., 2002). On the basis of their contents, programs were developed which enable to identify individual regulatory sequence elements (transcription factor binding sites, TFS; e. g. Quandt et al., 1995; Goessling et al., 2001), complex compositions of regulatory regions (Frech et al., 1998; Kel et al., 1999), or whole promoters with highly variable degree of accuracy among the different tools (for review, see Fickett, Hatzigeorgiou, 1997). Now, with the (nearly) complete sequence of the human genome at hand, we can start to systematically apply what we have learned so far about individual regulatory sequence signals and their context and, thus, to evaluate the regulatory potential of a complex genome as a whole.

Methods and Algorithms

The TRANSGENOME resource: TRANSGENOME is an information resource which provides an overall annotation of the human genome with emphasis on its regulatory characteristics. The reference sequence of the genomic database is currently based on RefSeq as a data source for the nucleotide sequence and the NCBI annotation results, including gene models and SNP information. Nomenclature information is added from HGNC (Human Genome Nomenclature Committee), automatically updated as soon as changes are available. Propietary information, e.g. from the TRANSFAC database, is matched with RefSeq as well.

Databases on transcription regulation (TRANSFAC® and TRANSCompel™): The databases used for this study are TRANSFAC® and TRANSCompel™ (both as release 5.4, December 2001) TRANSFAC® is a database on transcription factors (TF), their expression patterns, genomic and artificial binding sites as well as positional weight matrices for many of the TF stored in the database (Wingender et al., 2001). TRANSCompel™ is the professional version of the COMPEL database on composite elements (Kel et al., 1995, Kel-Margoulis et al., 2002).

Programs for the analysis of regulatory regions (Match™): Match™ is a weight matrix-based tool for searching putative transcription factor binding sites (TFS) in DNA sequences. It uses the matrix library of TRANSFAC® . The user may construct and save specific profiles which are selected subsets of matrices including default or user-defined cut-off values. The algorithm has been described elsewhere (Goessling et al., 2001).

Implementation and Results

5'-flanking regions of the human genome exhibit an elevated GC-content: It is known that the human genome has a GC-content that is below 50%, but still different between the individual chromosomes, i.e. between 0.38 for chromosomes 4 and

13 and 0.48 for chromosomes 19 and 22. We have composes different sequence sets with the help of TRANSGENOME and found that in particular the immediate 5'-flanking regions (–300/+50, with the transcription start site at +1) exhibit the highest GC content when compared with short or long introns, exons, or 3'-flanking regions. For instance, in the AT-rich chromosome 13, the overall GC-content of immediate upstream sequences is 0.49 and that of the GC-rich chromosome 22 is around 0.58.

Despite of their generally enhanced GC content, 5'-flanking regions exhibit a rather asymmetric distribution of this feature: There is a considerable number of slightly more AT-rich upstream sequences than the average, but a smaller number of sequences that are extremely GC-rich. The overall range of GC-contents lies between 0.24 and 0.82 in the AT-rich chromosome 13, and 0.22 and 0.88 in the GC-rich chromosome 22, respectively.

Presence of potential transcription factor binding sites in human chromosomes: We investigated the occurrence of potential binding sites for a set of selected transcription factors (TF) with their positional weight matrices (PWD) as they are given in the TRANSFAC database. The analysis was done with the Match™ routine described in the Methods section. The threshold was set to FN50, i.e. we applied relatively stringent conditions which allowed for the detection of (only) 50% of the known genomic sites, but significantly reduced the number of false positive matches.

Expectedly, the number of hits for individual TF DNA-binding profiles largely varies amongst the different patterns as well as between the chromosomes. First, the frequency with which each pattern matches reflects its stringency which is determined by both its width and conservation. The "weakly" defined TATA-box pattern matches with the highest frequency, whereas the matching frequency of the well-defined E2F-binding site is nearly three orders of magnitude lower. Also the presence of a conserved CpG dinucleotide in the pattern influences the probability of its appearance. Moreover, the general GC content of both the pattern as well as the analyzed sequence determine the match frequency. When we corrected the frequency of TF binding patterns for the GC content of the analyzed chromosomes, most of them appear as expected from the general nucleotide composition. However, some are significantly overrepresented, such as TATA and GC boxes, but some of them are extremely rarely found such as CRE-BP1 and E2F sites.

Presence of potential transcription factor sites in 5'-flanking regions of human genes: We have investigated in greater detail the occurrence of TF binding patterns in immediate upstream regions compared with random sequences of the same GC content, and made a comparable analysis for exons and random sequences of identical nucleotide composition. In exons, most TF patterns occur as to be expected from the GC content, with a few exceptions. In 5'-flanking sequences, however, TATA and GC boxes are significantly overrepresented, as well as the transcription initiator element-binding factor YY1 (to a lesser extent). Again, E2F sites prove to be extremely rare, in exons even more than in 5'-flanking sequences, whereas CRE-BP1 sites occur in 5'-flanks with the expected frequency, but are underrepresented in exons.

Discussion

The tool TRANSGENOME we have introduced here provided us with reliable sequence sets of defined functional regions of the human genome. While the assigment of intron and exon sequences is relatively robust, that of 5'-flanking regions may be subject to future revisions. Suzuki et al. (2002) showed that about one third of the 5'-gene ends indicated in RefSeq have to be extended, but half of them by less than 100 nucleotides thus affecting the results discussed here not too much. At least we consider our set of 5'-flanking to be highly enriched in putative promoter sequences.

Our results about the elevated GC-content in the immediate 5'-flanking regions are consistent with previous reports on the *Drosophila* genome (Ohler et al., 2001), and the asymmetric distribution of upstream regions according to their GC content coincides with that reported for the whole genome (International Human Genome Sequencing Consortium, 2001).

When correcting the occurrence of TF binding patterns for the overall GC-content of chromosomes or of individual functional regions, it becomes obvious that for some of these patterns the general chromosomal sequence background is particularly favorable for some of these pattern such as TATA and GC boxes, allowing for an even more pronounced enrichment of these elements in 5'-flanking regions. In contrast, other elements are clearly counterselected in the whole genome as well as in 5'-flanking regions, though being present there in higher concentration than in other regions, probably to avoid pseudo-sites which could exert or gain during evolution an erroneous function.

Further analyses will give us more insights into the regulatory sequence features of the whole human genome as well as into the mechanisms of co-evolution of genomic regulatory sequence signals and the DNA-binding domains recognizing them.

References

1. Fickett J.W., Hatzigeorgiou A.G. (1997) Eukaryotic promoter recognition. Genome Res. **7**, 861-878.
2. Frech K., Quandt K., Werner T. (1998). Muscle actin genes: a first step towards computational classification of tissue specific promoters. In Silico Biol. 1, 0005.
3. Goessling E., Kel-Margoulis O.V., Kel A.E., Wingender E. (2001) MATCH™ - a tool for searching transcription factor binding sites in DNA sequences. Application for the analysis of human chromosomes. Proc. of the German Conference on Bioinformatics GCB 2001. Wingender, E., Hofestädt, R., and Liebich, I. (eds.), Braunschweig, 158-161.
4. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature. 409, 860-921.

5. Kel O.V., Romaschenko A.G., Kel A.E., Wingender E., Kolchanov N.A. (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. Nucl. Acids Res. 23, 4097-4103.
6. Kel A., Kel-Margoulis O., Babenko V., Wingender E. (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. J. Mol. Biol. 288, 353-376.
7. Kel-Margoulis O.V., Romashchenko A.G., Kolchanov N.A., Wingender E., Kel A.E. (2000) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. Nucl. Acids Res. 30, 332-334.
8. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30, 312-317.
9. Ohler U., Niemann H., Liao G.-c., Rubin G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics. 17, S199-S206.
10. Praz V., Perier R., Bonnard C., Bucher P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucl. Acids Res. 30, 322-324.
11. Quandt K., Frech K., Karas H., Wingender E., Werner T. (1995) MatInd and MatInspector - New fast and sensitive tools for detection of consensus matches in nucleotide sequence data. Nucl. Acids Res. 23, 4878-4884.
12. Suzuki Y., Yamashita R., Nakai K., Sugano S. (2002). DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. Nucl. Acids Res. 30, 328-331.
13. Wingender E., Chen X., Fricke E., Geffers R., Hehl R., Liebich I., Krull M., Matys V., Michael H., Ohnhaeuser R., Prueß M., Schacherer F., Thiele S., Urbach S. (2001) The TRANSFAC system on gene expression regulation. Nucl. Acids Res. 29, 281-283.

GENERALIZED SPECTRAL PORTRAIT OF *ESCHERICHIA COLI*- TYPE REPLICATION ORIGIN'S PRIMARY STRUCTURE

**** Kravatskaya G.I., Esipova N.G.***

Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia
* Corresponding author: gk@imb.ac.ru

Key words: *DNA replication,* oriC, *primary structure*

Resume

Motivation: Compositional asymmetry between two DNA strands has been used to locate origins of replication in certain bacterial and large viral genomes (Lobry, 1996; Grigoriev, 1998). But there are many genomes, in which strand compositional asymmetry is not observed (Mrazek, Karlin, 1998). Therefore search for new approaches to replication origins prediction is important. Approaches contributing to our understanding of the process of replication are especially interesting. In this paper we develop new way of replication origins description - spectral portrait of their primary structure.

Results: We have compared Fourier spectra of the nucleotide sequences of several different bacterial DNA replication origins and derived generalized spectral portrait of primary structure of replication origins that function in *Escherichia coli*.

Introduction

Chromosomal replication in eubacteria is initiated at the replication origin *oriC*. Understanding the nature of DNA sites where replication begins is one of the central moments in understanding how cells regulate the replication of their genomes. Periodic patterns in the sequence organization of the replication origin of *E.coli* K-12 chromosome we have studied in (Esipova et al., 2000, Kravatskaya, Esipova, 2000). It was shown that nucleotide distribution of *oriC* contributes to DNA destabilization and DNA unwinding near the replication origin. In this paper we spread our approach to the analysis of other bacterial chromosomal replication origins.

Methods and Algorithms

We considered nucleotide sequences from several bacterial origins of replication. Fourier images of their primary structure were obtained with the aid of the Matrix Fourier analysis. Fourier transform of the nucleotide sequence with the length M is defined as $x_\alpha(T_n) = \frac{\sum_{m=1}^{M} x_{m,\alpha} \exp(-i\frac{2\pi m}{T_n})}{\sqrt{M}}$, where $n=0, 1, ...,M-1$, $\alpha \in$ A = {a, c, g, t}, $m=1,..., M$. $x_{m,\alpha} = 1$, if the nucleotide of type α is in the m-th position of the sequence, $x_{m,\alpha}=0$ - otherwise. The spectral power of a harmonic corresponding to the period T_n $F(T_n) = \sum_{\alpha,\beta \in A} L_{\alpha\beta} x_\alpha(T_n) x^*_\beta(T_n)$, where $L_{\alpha\beta}$ is a correlation matrix. For more details see (Makeev et al., 1996) and references herein. All values of spectral power were normalized to that ones corresponding to the nucleotide sequence with binomial distribution. For multiple sequence alignment we used T-Coffee program (Notredame et al., 2000). For the interpretation and analysis of Fourier spectra we used ad hoc Perl program.

Implementation and Results

We have obtained Fourier spectra of several bacterial origins (from *Shigella boydii, Shigella dysenteriae, Shigella flexneri, Shigella sonnei, Salmonella typhimurium, Enterobacter aerogenes, Klebsiella pneumoniae, Erwinia carotovora, Vibrio harveyi, Escherichia coli*). Peaks corresponding to the period T=2, 11, 27, 70-81, 85-100 nucleotides are common for Fourier-spectra of all origins under consideration. Peak corresponding to the period T=13 nucleotides is present in all but one spectrum (*Vibrio harveyi*). On the basis of analysis of Fourier-spectra of six bacterial chromosomal replication origins (*Salmonella typhimurium, Enterobacter aerogenes, Klebsiella pneumoniae, Erwinia carotovora, Vibrio harveyi, Escherichia coli*) we derived generalized portrait of *E. coli*-type replication origin by means of averaging-out. Pre-eminent values of spectral power F correspond to T=2, 11, 17, 27, 86-105 nucleotides (Fig. 1). Less pronounced peaks correspond to T=9, 13, 14, 18, 19, 28, 33-35, 45-47, 74-85, 106-110 nucleotides.

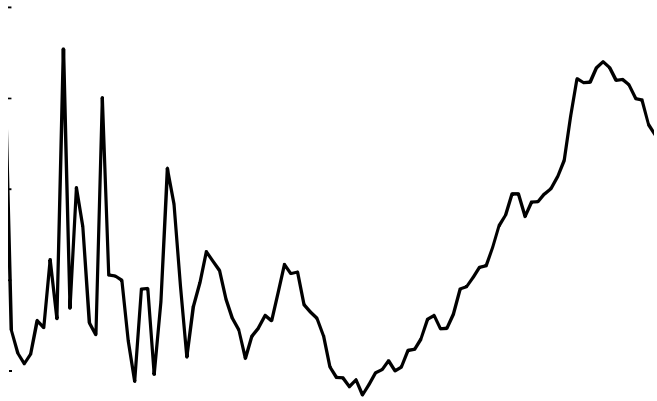


Fig. 1. Generalized spectral portrait of *E. coli*-like *oriC* in terms of a, c, g, t nucleotides occurrences. T – the length of the period, F – spectral power.

Discussion

We also analyzed chromosomal replication origins from *Shigella boydii, Shigella dysenteriae, Shigella flexneri, Shigella sonnei*. These *oriC* are almost identical to *E.coli oriC* and as a consequence have very similar to *E.coli oriC* Fourier spectra. We did not use these sequences in the generalized portrait of *E. coli*-type replication origin construction. We considered six origins of replication that function in *Escherichia coli* (Zyskind et al., 1983). In spite of differences in their primary structures, these origins of replication have Fourier-spectra with several common features. **As** we can see from multiple alignment of these bacterial chromosomal replication origins (Fig. 2), conservative sites are clustered and correspond to DnaA (R1, M, R2, R3, R4), IHF and FIS proteins binding sites. There are also several conservative GATC sites. The differences in the primary structures of these six fragments are mainly substitutions. There are only few deletions in the A+T-rich region of origins (5'-end). It seems that the role of non-conserved clusters of nucleotides in *oriC* is to provide a specific system of distances between conserved regions or, perhaps, a special system of periodic patterns in their organization. On the basis of our preliminary results on *Sacchromyces cerevisiae* chromosomes, we suppose that our approach will be useful for the analysis of eukaryotic genomes as well.

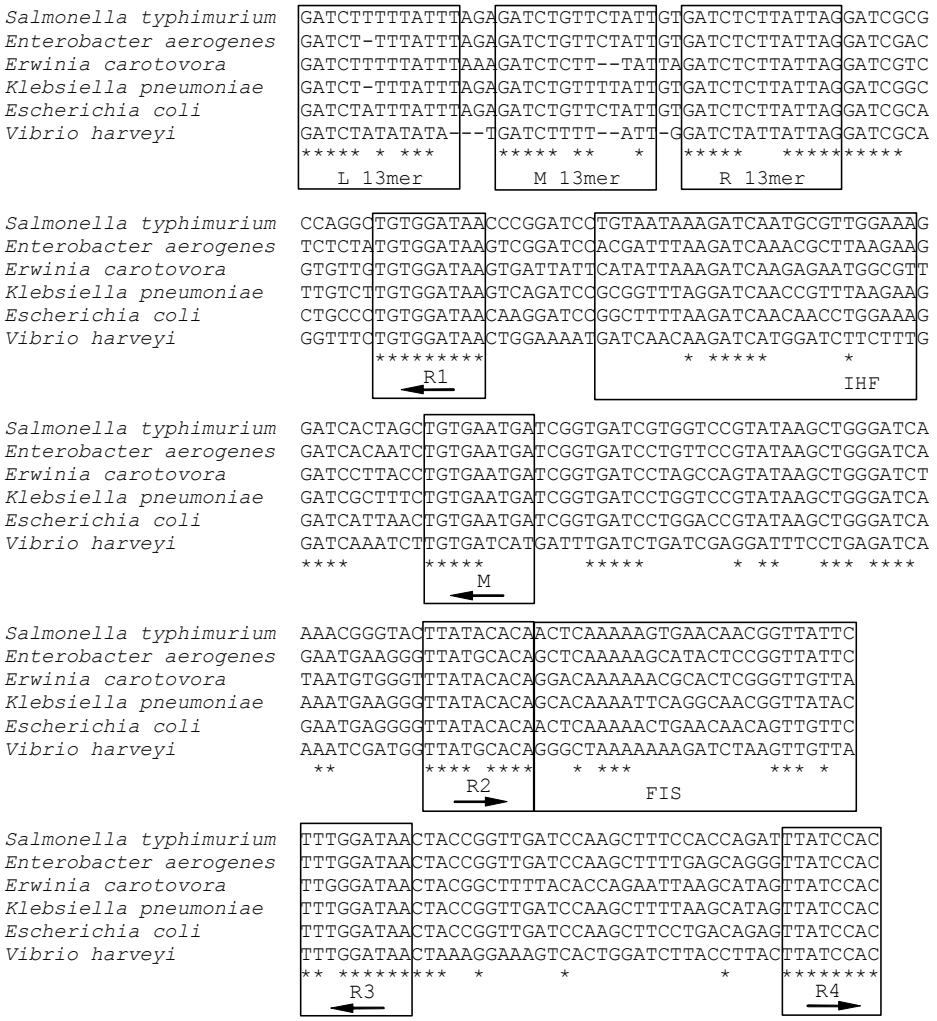


Fig. 2. Multiple alignment of minimal *E. coli oriC* with other bacterial origin sequences that function in *E. coli*. Binding sites of proteins and A/T-rich 13-mers are enclosed in boxes (positions as for *E.coli* in Woelker, Messer, 1993).

Acknowledgements

This work was supported by grant 02-04-06581 and grant 00-04-48351 from Russian Foundation of Basic Research (RFBR).

References

1. Grigoriev A. (1998). Analyzing genomes with cumulative skew diagrams. Nucl. Acids Res. 26, 2286-2290.
2. Esipova N.G., Kutuzova G.I., Makeev V.Yu., Frank G.K., Balandina A.V., Kamashev D.E., Karpov V.L. (2000). Patterns of Nucleotide Distribution in the *Escherichia coli* Replication Origin *oriC*. Biophysics. 45, 421-426.
3. Kravatskaya G.I., Esipova N.G. (2000). Periodic Patterns in Sequence Organization of Replication origin of *Escherichia coli* K-12 Chromosome. Proc. of the BGRS 2000. 2, 74-76.
4. Lobry J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660-665.
5. Makeev V.Ju., Frank G.K. , Tumanyan V.G. (1996) Statistics of periodic patterns in the sequences of human introns. Biophysics. 41, 1, 263-268.
6. Mrazek J., Karlin S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl Acad. Sci. USA. 95, 3720-3725.
7. Notredame C., Higgins D., Heringa J. (2000) T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. J. Mol. Biol. 302, 205-217.
8. Woelker B., Messer W. (1993) The structure of the initiation complex at the replication origin, *oriC*, of *Escherichia coli*. Nucl. Acids Res. 21, 22, 5025-5033.
9. Zyskind Ju., Cleary J., Brusilow W., Harding N., Smith D. (1983). Chromosomal replication origin from the marine bacterium *Vibrio harveyi* functions in *Escherichia coli* : *oriC* consensus sequence. Proc. Natl Acad. Sci. USA. 80, 1164-1168.

STRUCTURE-FUNCTION RELATIONSHIP IN PROTEIN-DNA RECOGNITION

Sarai A.

RIKEN Tsukuba Institute, Tsukuba 305-0074 Japan, e-mail: sarai@rtc.riken.go.jp

Key words: *protein-DNA recognition; structure-function relation; transcription factors; target prediction*

Resume

Motivation: Although the structural data on the protein-DNA complex have been rapidly increasing, the mechanism of DNA sequence recognition by proteins has been poorly understood, and thus the accurate prediction of their targets at the genome level is not yet possible. We try to extract functional information from structural data of protein-DNA complexes, and utilize the information for quantifying the specificity of protein-DNA recognition, for examining structure-function relationship, and for predicting target sites of transcription factors.

Methods: We made a statistical analysis of structural database of protein-DNA complex, and derived empirical potential functions for the specific interactions between bases and amino acids. Then, we used a sequence-structure threading to quantify the specificity in protein-DNA recognition. We calculated Z-score as a measure of specificity for a given protein-DNA complex against random sequences by the sequence-structure threading. We also evaluated the fitness of DNA sequence against DNA structure to examine the role of indirect readout mechanism.

Results: The statistical potentials showed different characteristics for different combinations of bases and amino acids, indicating that these potentials can be used to quantify specificity in the base-amino acid interactions. We applied the method to examine the relationship between structure and specificity in protein-DNA recognition; cooperativity, symmetry/asymmetry in binding, and congnate/noncognate binding. These results show how the structures affect the specificity. We could show relative contribution of direct and indirect readout mechanisms in the recognition. We also applied the method to predict the target sequences of transcription factors at the genome level.

Introduction

Regulation of gene expression in higher organisms is achieved by a specific recognition of target DNA sequences by DNA-binding proteins. Due to the progress of X-ray crystallography and NMR spectroscopy techniques, structural data on the protein-DNA complexes have been rapidly increasing. However, the mechanism of DNA sequence recognition by proteins has been poorly understood, and thus the accurate prediction of their targets at the genome level is not yet possible. This situation implies that the structural information has not been fully utilized. Understanding the molecular mechanism and its application to genome-wide prediction are essential for the analysis of gene regulation network. Here, we try to extract functional information from structural data of protein-DNA complexes. We made a statistical analysis of structural database of protein-DNA complex, and derived empirical potential functions for the specific interactions between bases and amino acids. Then, we used a sequence-structure threading to examine the relationship between structure and specificity in protein-DNA recognition, and to predict target sites of transcription factors in real genome sequences. We also evaluated the fitness of DNA sequence against DNA structure to examine the role of indirect readout mechanism. We show how the structural features are related to the specificity, and discuss relative roles of direct and indirect readout mechanisms in the recognition. We also discuss the strategy to predict the target sequences of transcription factors at the genome level.

Methods

We extracted interacting pairs of bases and amino acids from a refined set of non-redundant protein-DNA complexes (Kono, Sarai, 1999). The distant-dependent statistical potentials for the specific base-amino acid interactions were derived from the spatial distributions of C_α atoms of amino acids around a base. The potential function for each pairs of base and amino acid in a particular protein-DNA complex was summed to derive a total potential energy. By threading a set of random DNA sequences onto the template structure, we calculated the Z-score of the specific sequences against the random sequences, which represent the specificity of the complex. We have also derived statistical potential functions for conformational energy of DNA from the protein-DNA complex structural data to evaluate the fitness of sequences to a particular conformation of DNA. Then, these potentials can be used to quantify the specificity of indirect readout mechanism of protein-DNA recognition, by using the same threading procedure. The threading procedure was also applied to the real genome sequences in order to find potential target sites.

Results and Discussion

We first tested the method by calculating the Z-scores for NF-κB, glucocorticoid receptor DNA binding domain, EcoRV endonuclease and BamHI endonuclease, for which both the cognate and non-cognate complex structures are available in

PDB. The method could distinguish the two structures as differences in the Z-scores as well as statistical poteitials. Thus, the subtle differences in specificity of these structures could be detected by the analysis of energy potentials. Proteins often bind to DNA as homodimers, which leads to subtle structural differences between the two subunits. Thus, we examined in detail the structural effects of asymmetric binding on specificity. Marked differences in the specificity of DNA binding were observed for the two subunits of λ repressor, the glucocorticoid receptor, and for transcription factors containing a Zn_2Cys_6 binuclear cluster domain, which are known to bind asymmetrically to DNA. We also applied this method to examine the relationship between structure and specificity in cooperative protein-DNA binding. The effect of cooperative binding was examined by comparing the monomer and heterodimer complexes of MATa1/α2, MCM1/MATα2 and NFAT/AP-1 transcription factors. We found that the heterodimer binding enhance the specificity in a non-additive manner. This result indicates that the conformational changes introduced by the heterodimer binding play an important role in enhancing the specificity. We have derived statistical potential functions for conformational energy of DNA to quantify the specificity of indirect readout mechanism of protein-DNA recognition. Combining the potentials with the above base-amino acid potentials, we can evaluate the contribution of direct and indirect readout mechanisms. We have analyzed various protein-DNA complexes systematically, and found that both the direct and indirect mechanisms make significant contribution to the specificity. The relative contributions depend on the types of DNA-binding proteins.

The threading procedure was used to find target sites of transcription factors in real genome sequences. As an example of such applications, we could identify the experimentally-verified binding sites of the transcription factor MATa1/α2 in the promoter of *HO* gene successfully. We have attempted to identify target sites and genes of MCM1/MATα2 in the whole yeast genome. So far, we could identify potential binding sites in the promoters of target genes. We are comparing the predicted results with experimental data to assess the reliability of the method.

The increase in the structural data, which will be accelerated by structural genomics project, will make the present structure-based method promising for revealing the structure-function relationship in protein-DNA recognition and for predicting targets of transcription factors. This method can also be applied to proteins of unknown structure having substantial sequence similarity to known proteins, on the basis of which structures can be modeled and binding sites can be predicted. Such analysis would contribute significantly to the functional genomics in the era of post-genome science.

Acknowledgements

This work is the result of collaboration with Drs. H.Kono, S.Selvaraj, M.Gromiha and J.G.Siebers.

References

1. Kono H., Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. Proteins. 1999. 35, 114-131.

THE IMPORTANCE OF GENOME REGULATION AND STRUCTURE TO THE PHARMACEUTICAL INDUSTRY

Hodgman T.C.

GlaxoSmithKline, Stevenage, UK, e-mail: ch18380@gsk.com

Key words: *Drug discovery, drug development, genetic variation, gene expression, information map*

Resume

Motivation: Pharmaceutical research and development generates a broad variety of data ranging from chemical assays through biomolecular structure/activity determination to disease/therapeutic effects on the whole human population. To increase efficiency and reduce costs, it is essential to integrate this data and develop predictive computational tools.

Results: A structured "information space" has been developed which maps pharmaceutical data in a way that shows they, and the tools for their analysis, inter-relate in the development of pharmaceutical products. The map can be used in various ways and will be exemplified by showing how the Bioinformatics of Genome Regulation and Structure benefits the pharmaceutical industry

Availability: The author welcomes discussion and can be contacted at the above address.

Introduction

The drug discovery process has now been optimised to a well defined pathway that belies its complexity. Diseases, for which there is a clinical need for new medicines, are studied to discover appropriate "Targets" for new drugs. These Targets are used in screening for compounds that have the desired effect. These "Leads" are moved into development where, in the pre-clinical phase, they are subject to a battery of tests to determine their suitability as medicines (on safety, pharmacological and cost grounds), followed by tests in clinical trials. There is a general perception that only 10% of leads survive to become commercial products. Pharmaceutical companies now have priorities in reducing the time required to discover leads and improving the success rate of leads during development.

The data generated during these different stages are very varied and have largely been confined to the parts of companies that generated them. If predictive methods are to be developed to help companies make the best choices of which potential targets and leads should be used, then data integration is an essential first step. The data refer to physical entities (and their interactions) that range in size from small molecules through individual patients to the whole human population (10^{-9} to 10^{6} metres), and very short to very long timescales (10^{-9} to 10^{9} seconds). This work describes further the information map and information-flow model of pharmaceutical research and development (Hodgman, 2001) by showing how genome regulation and structure has an important role to play.

Methods and Algorithms

The common ways of depicting the drug discovery process are usually based on the stages of product development (linear pathways) or line management (tree structures). However, by viewing what is happening from the perspective of information generation and transfer, it becomes clear that the process is circular, because information from the clinic determines which new diseases should enter a company's portfolio. This results in the two-dimensional respresentation shown in Figure 1. This figure can be considered to be a Venn diagram where the position on the vertical axis has meaning. Every position in the figure refers to some data, and the higher a point is up the figure, the greater the physical size of the object(s) to which the data refer. This grading results in a natural partitioning of information into medical, biological and chemical domains; but there are other consequences. The vertical axis also denotes increasing complexity, longer times for experiments/trials, and hence greater costs in producing the information.

Lines drawn on this map denote the flow of information from one stage to another. The line in figure 1 depicts the needed to develop a new small-molecule medicine. There are other, progressively shorter paths for biopharmaceuticals (i.e. therapeutic proteins and DNA vaccines), diagnostics and surrogate markers (i.e. indicators that correlate with the stages of disease progression and therapeutic effect, which are especially useful for examining chronic diseases, such as Alzheimer's). The stages leading to identification of leads involve successive reduction in the complexity and variability of the objects under investigation, whereas the remainder of the cycle involves increasing variability into ever more complex systems. This explains why so many leads fail during development, because there are factors in patients and populations that could not be predicted from knowledge of the drug and its target alone.

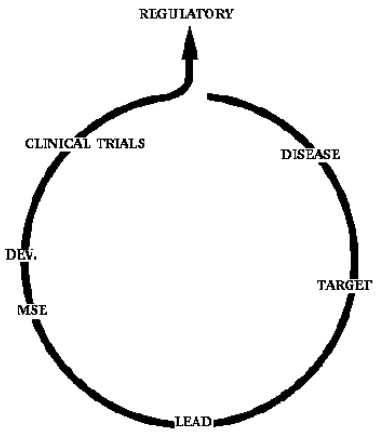


Fig. 1. The pharmaceutical information space. Each point refers to items of information and lines show where information is transferred (and generated) in the research and development of small molecule medicines. The names refer to the stages in the process, with MSE and DEV referring respectively to Medical Safety Evaluation and drug development (in all its forms).

The final point to note is that there is a third dimension to this information space, which corresponds to portfolios, or sets or populations at this point. The height in this third dimension corresponds to the number of the object concerned: tens of diseases and clinical trials, thousands of Targets, millions of compounds. This height also equates with the number of records in a database.

Results and Discussion

Genome regulation and structure contribute to the drug discovery process in the same areas molecular biology and genetics (see Fig. 2). Molecular biology has provided invaluable assistance in the characterisation of Targets, and this is where Bioinformatics has thus far most contributed to drug discovery. However, with the publication of the draft human genome, the focus of Bioinformatics is making a major shift. Genetics links the genotype of an organism to the phenotypes it displays under various circumstances. Thus, it intrinsically spans the divide between the medical and biological domains. The relationship between activities in the pharmaceutical industry, general biological research, and the bioinformatics of genome regulation and structure have been mapped in Figure 3.

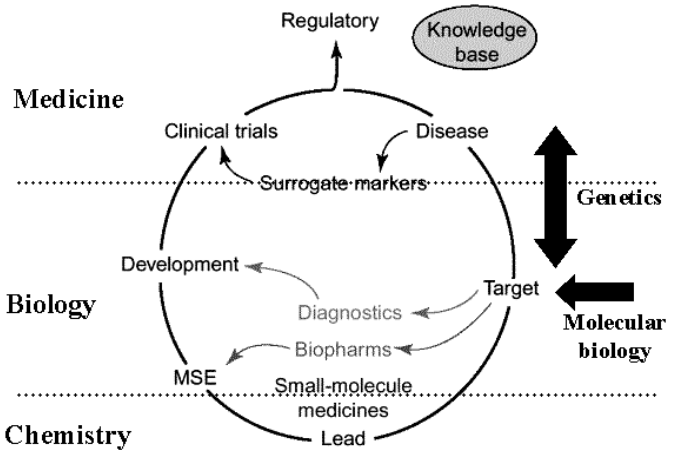


Fig. 2. Information partitioning and the regions involving molecular biology and genetics.

The path for small-molecule medicine development has been supplemented by the extra paths pertaining to the development of biopharmaceuticals, diagnostics and surrogate markers. The only pharmaceutical company product that does not have any path associated with it is a knowledge base. The dotted lines show the approximate boundaries between medical, biological and chemical information. The large arrows show the zone (on the vertical axis) where genetics and molecular biology contribute to the information. Note that they apply to both left and right hand sides of the cycle.

Techniques to interpret RNA/protein/metabolite profiles and the development of network models both impact upon the most number of activities. Much effort worldwide is going into the former, but there is still a great deal to do in terms of developing appropriate network models. Regulatory region prediction and quantitative sequence/activity relationships relate to a fewer number of general research activities, but are still associated with most if not all of the pharmaceutical industry activities. The other bioinformatics contributions (polymorphism detection, protein property prediction and gene identification) have been employed some years already. However, their focus is changing.

Polymorphism identification was employed almost entirely to focus on gene pre-disposing towards disease, whereas now they are also being used to look for genes conferring adverse drug reactions or differential response to therapeutic regimes (pharmaco-genomics). Many of the human genes have now been identified in the genome sequence, even if their products have not yet been characterised. The focus of gene identification is shifting towards genes of pathogens and organisms that are used as models for disease or medical-safety screening. Protein property prediction is likewise moving its attention from

plain function characterisation, to identification of interacting ligands or, in the case of biopharmaceuticals, properties associated with its stability during bulk manufacture.

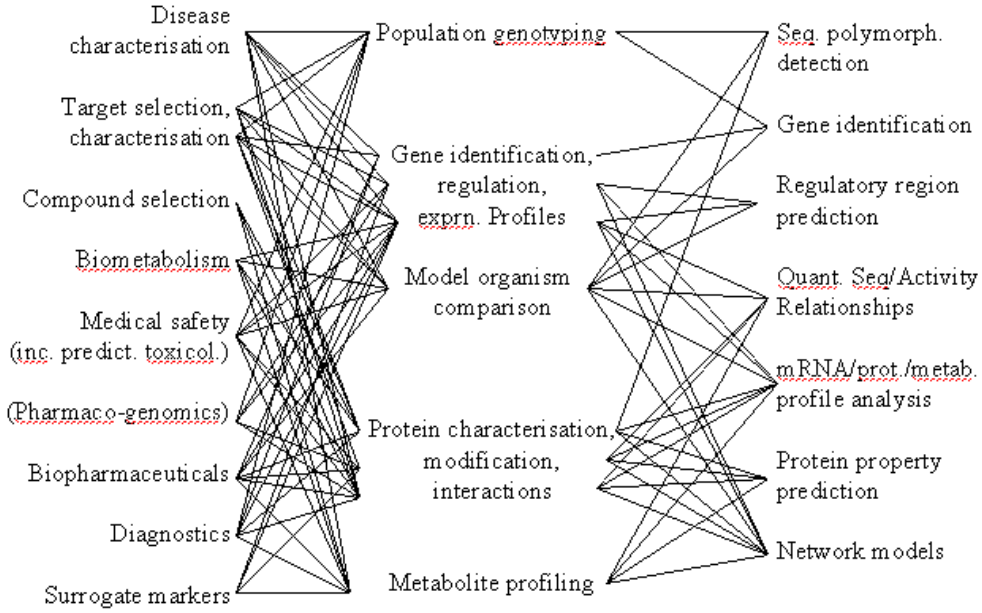
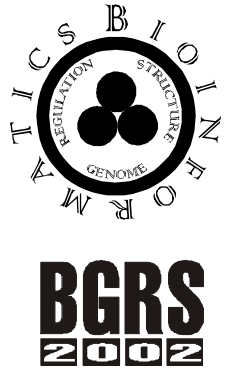


Fig. 3. From left to right, the 3 columns refer activities in pharmaceutical development, general research and bioinformatics. Lines show the connections between these respective groups of activities, and help the reader to see where their bioinformatics expertise might be applied in a pharmaceutical industry context.

It is clear that bioinformatics contributes in many important ways to pharmaceutical research and development, but that this relationship is complex. Apart from the general role of helping companies to organise, store and analyse their data, it can specifically help prioritise Targets, suggest compounds for drug screening, and develop the resources to predict which compounds are likely to fail during development.

References

1. Hodgman C. (2001) An information-flow model of the pharmaceutical industry.
2. Drug Discov. Today, 6**,** 1256-1258.



GENOME STRUCTURE AND FUNCTION

EXACT MAPPING OF PROKARYOTIC GENE STARTS

[1] Baytaluk M.V., [2] Gelfand M.S., [2] Mironov A.A.*

[1] Institute of Molecular Biology, RAS, Moscow, Russia
[2] Integrated Genomics – Moscow, P.O.Box 348, 117333, Moscow, Russia.
* corresponding author: e-mail: gelfand@integratedgenomics.ru, phone +7-(095)-135-20-41, fax: +7-(095)-132-60-80

Key words: *gene, genomics, gene recognition, reading frame, start of translation, computer analysis, prokaryotes*

Resume

Motivation: It is known that while the programs used to find genes in prokaryotic genomes reliably map protein-coding regions, they often fail in the exact determination of gene starts. This problem is further aggravated by sequencing errors, most notably, insertions and deletions leading to frameshifts.

Here describe a new algorithm for correction of gene starts and identification of frameshifts in prokaryotic genomes. The algorithm is based on the comparison of nucleotide and protein sequences of homologous genes from related organisms, using the assumption that the rate of evolutionary changes in protein-coding regions is lower than that in non-coding regions. A modification of the Smith-Waterman algorithm is used to align protein sequences obtained by the formal translation of genomic nucleotide sequences. The possibility of frameshifts is taken into account.

Results: The algorithm was tested on several groups of related organisms: gamma-proteobacteria, the *Bacillus/Clostridium* group, and three *Pyrococcuss* genomes.

The testing demonstrated that, dependent on a genome, 1 through 10% of genes have incorrect starts or contain frameshifts.

The algorithm is implemented as the program package Orthologator-GeneCorrector.

Introduction

Systematic analysis of the performance of available software for gene recognition, using HMM techniques (GeneMark, GLIMMER, GeneHacker Plus) and comparative analysis (ORPHEUS), highlighted that while the current programs perform well at identifying genes (as opposed to random open reading frames), gene starts are predicted with lower accuracy.

An additional problem complicating gene recognition arises from frameshifts that interrupt reading frames.

Materials and Methods

The prediction is done in three steps:

1.1 Building the tables of ortologs.

1.2 Applying the modified Smith-Waterman alignment algoritm to the pairs of orthologous genes.

1.3 Filtering of the results and identification of suspicious gene starts and possible frameshifts.

Data. The algorithm was tested on three groups of genomes: 1) Escherichia coli, Vibrio cholerae, Haemophilius influenzae, Buchnera sp., Xylella fastidiosa; 2) Bacillus subtilis, Bacillus halodurans, Clostridium acetobutylicum; 3) Pyrococcus horikoshii, Pyrococcus abyssi , Pyrococcus furiosus.

Building the ortholog tables

The ortholog tables were constructed using the program ORTHOLOGATOR, which is a part of the created software package (Baitaluk et al., 2000).

The modified Smith-Waterman algorithm for alignment of gene starts. In a pair of extended sequences, all potential starts (ATG, TTG, GTG) and stops (TGA, TAA, TAG) around gene starts ([$s1$-$n1$, $s1$+$n1$] for the first gene, [$s2$-$n2$, $s2$+$n2$] for the second gene) and ends ([$e1$-$m1$, $e1$+$m1$] for the first gene, [$e2$-$m2$, $e2$+$m2$] for the second gene) are marked. For two nucleotide sequences [$s1$-$n1$, $e1$+$m1$] and [$s2$-$n2$, $e2$+$m2$] the Smith-Waterman dynamic programming algorithm (Smith et al., 1981) is modified to align the protein sequences generated by the formal translation of the nucleotide sequences in all three reading frames, with account to possible frameshifts.

The recursions for the alignment are:

$$S_{i,j} = \max \begin{cases} S_{i-3,j-3} + d(x_i, y_j) \\ Q_{i,j} \\ R_{i,j} \\ S_{i-2,j} + f \\ S_{i,j-2} + f \\ S_{i,j-1} + f \end{cases}$$

where:

$$Q_{i,j} = \max \begin{cases} S_{i,j-3} + a \\ Q_{i,j-3} + b \end{cases}$$

$$R_{i,j} = \max \begin{cases} S_{i-3,j} + a \\ R_{i-3,j} + b \end{cases}$$

Here $S_{i,j}$ is the score of the alignment at point (x_i,y_j); $d(x_i,y_j)$ is the weight of matching two amino acids encoded by codons (x_{i-2},x_{i-1},x_i) and (y_{j-2},y_{j-1},y_j) codons; a is the deletion initiation penalty; b is the deletion extension penalty; f is the frameshift penalty.

Filtering. At the filtering step possible explanations for the observed differences between the aligned and annotated gene termini are considered.

Results and Discussions

Verification using a third genome was made. If a correction made by the comparison of the BaseGenome (BG) with an additional genome AGj was confirmed in the comparison with one more additional genome AGk, the new gene coordinates were accepted as final.

Table shows types of errors found in the annotations of the base genomes. The majority of corrections in the GenBank annotation, approximately 80% (dependent of the genome), were confirmed by the SWISS-PROT databank, thus demonstrating high accuracy of the obtained results.

Table. Types of errors found in the annotations of base genomes.

Base genomes	Number of genes	Number of corrections	Old/obsolete annotation	Conflicting annotation (confirmed by SwissProt)	Hypothetical genes
E. coli	4288	25	2	21	2
V. cholerae	2736	33	4	26	3
B. subtilis	4097	30	4	23	3
P. horikoshii	2058	32	8	18	6

Application of the comparative approaches lead to a number of interesting observations. Three types of genomic sequencing and annotation errors were identified:

1) Genes that had been sequenced and annotated long ago and not revised ever since. More accurate analysis showed in some cases that gene starts had been mapped incorrectly.

2) Hypothetical genes for which there is no experimental information. In such cases the comparison corrects the results of the statistical annotation.

3) Genes with conflicting annotation in different databases.

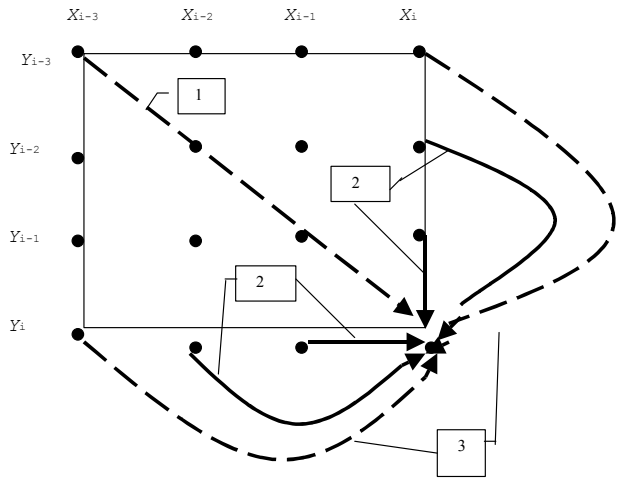


Fig. Fragment of the alignment graph with all types of transitions: (Xi and Yi - nucleotides).

1 - from a pair of amino acids encoded by codons (x_i-6, x_i-5, x_i-4) and (y_i-6, y_i-5, y_i-4) into a pair of amino acids encoded by codons (x_i-3, x_i-2, x_i-1) and (y_i-3, y_i-2, y_i-1);

2 - frameshift;

3 - deletion of amino acids encoded by codons (x_i-3, x_i-2, x_i-1) or (y_i-3, y_i-2, y_i-1).

Acknowledgements

We are grateful to Pavel S. Novichkov and Michael Fonstein for useful discussions.

This work was partially supported by grants from INTAS (99-1476), HHMI (55000309), and the Ludwig Institute for Cancer Reserch (CRDF RB0-1268).

The complete list of corrections is available at the: www.imb.ac.ru/ig_papers.

References

1. Baitaluk M.V., Novichkov P.S., Mironov A.A., Gelfand M.D. (2000) Software for orthoogy analysis in complete bacterial genomes BGRS'2000'. Proc. 2nd Intern. Conf. on Bioinformatics of Genome Regulation and Structure. 2, 26-27.

GENE PREDICTION IN GENOMIC DNA OF ASPERGILLUS

**** Neverov A.D., Gelfand M.S., Mironov A.A.***

State Scientific Center "GosNIIGenetika", 113545, Moscow, Russia, e-mail: neva_2000@mail.ru
Corresponding author

***Key words**: gene prediction, Aspergillus, Hidden Markov Model, HMM*

Resume

Motivation: Currently there are no programs for gene prediction in non-yeast fungi. We developed a tool for gene recognition in *Aspergillus* DNA sequences. It is based on the HMM approach.

Results: The gene prediction program was tested on a set of single-gene fragments and on a set of semi-artificial multi-gene fragments**.** The accuracy statistics of our program are close to the GENSCAN statistics on human single gene fragments, and predictably weaker on multi-gene fragments.

Availability: *available on request from the authors.*

Introduction

The Aspergillus genus of fungi contains a number if important human and plant pathogens. Among human and animal diseases caused by *Aspergillus* species are aspergillosises and mycoses; also some of the *Aspergillus* species cause the food decay. The genomes of *Aspergillus nidulans*, *A. niger*, *A. orysae* and *A. parasiticus* species are currently being sequenced and thus the problem of *Aspergillus* genome annotation arises. Statistics-based programs may be useful for gene searching in DNA regions without strong homology to known proteins such genes are expected to occur in exotic genomes like *Aspergillus*.

We have developed a computer program for ab initio gene prediction in eukariotic DNA. The program predicts the exon-intron structure in sequences fragments containing partial genes, complete genes or multiple genes on both DNA strands. The program is based on a general probabilistic model of the Hidden Markov (HMM) type. Introns, exons, intergenic regions, single-exon gene are modeled as hidden states of the Markov model from (Burge, Karlin, 1997). Three exon types considered in our model are initial exon (from a translation start to a donor splice site), internal exon (from an acceptor splice site to a donor splice site), and terminal exon (from a donor splice site to a stop codon). The donor and acceptor splicing sites, 5'- and 3'-untranslated regions are modeled separately.

Statistical parameters were estimated on a training set of 193 *Emericella nidulans* (*Aspergillus nidulans*) genes. These parameters were shown to be the same in all *Aspergillus* species.

Models

The model of a DNA is sufficiently like in (Burge, Karlin, 1997).

State models (exon, intron and intergenic region models)

The exon state was modeled as the probability of generating the observed protein-coding sequence multiplied by the length probability given by a specific distribution for each type of exons derived from the training set. The protein-coding sequence was modeled by the Markov three-periodic third-order model.

The intron state was modeled as the probability of generating the observed non-coding sequence multiplied by the length probability given by the length distribution. That distribution was close to the geometric with mean 62,5 bp. Intergenic regions were modeled in the same way with mean ~ 2000 base pairs for the length distribution. Non-coding sequences were modeled by the Markov third-order model.

Site models

Matrices of dependence between positions in donor and acceptor splice sites were constructed as in (Burge, Karlin, 1997). The donor sites have strong dependencies between both adjacent and not-adjacent positions. Thus we used the Maximum Dependence Decomposition model for the donor splice sites (Burge, Karlin, 1997). The acceptor splice signal of *Aspergillus* is very weak. The acceptor sites were modeled by nucleotide positional probabilities. The 5'and 3'–untranslated regions (UTR) we described by the first order Markov model. The 5' UTR was assumed to occupy 50 base pairs upstream of the translation start, the 3'-UTR was modeled as 100 base pairs downstream of the stop codon.

The choice of the statistical model for given HMM state depends on the size of a data set. The complicated models may improve the accuracy but need large data sets for parameter estimation. The other problem is the biases in the training set what lead to parameter overestimation. Such a model will be overtrained on the training set and unsuitable in practice. The

lack of data especially for insufficiently studied genomes puts the strict constrains on using the HMM-based gene prediction models. We plan to introduce the fifth order Markov model for non-coding regions and three–periodic fifth order model for exon states as the training set volume increases. See (Guigo, 1999) for a review of the statistical approaches to the gene recognition.

Results

Gene prediction programs are usually tested on single gene DNA fragments, the only exception a recent study by Guigo et al. (2000). It confirms the observation that the gene prediction accuracy is weaker on multi-gene fragments, because (1) long intergenic regions contain ORFs that may be over predicted as short genes and (2) it is difficult to properly predict gene boundaries. To account for these phenomena, we have performed test on simulated multi-gene "genomic" DNA fragments in addition to standard single-gene fragments.

The program was tested on 285 *Aspergillus* genes not included in the training set. Genomic fragments were modeled by random concatenation of single gene DNA sequences from the testing set. At that, complementary fragments were used with probability 1/2. Intergenic regions were constructed from non-coding DNA fragments with average length ~2000 base pairs (bp). The set of thus created 100 fragments is called the genomic set. The gene prediction accuracy for these sets is summarized in the table. The standard accuracy statistics on the nucleotide level were used: CC (Correlation Coefficient), AC (Approximate Correlation), Sn (sensitivity), Sp (specificity) and QQ (overlap). To calculate the accuracy statistics, each nucleotide from a test sequence is categorized as predicted positive (PP) if it is in a predicted coding region or predicted negative (PN) otherwise. If nucleotide is coding according to the annotation, it is denoted as actual positive (AP); non-coding nucleotides are actual negative (AN). The number of true positive (TP) nucleotides is calculated from TP=PP&AP, true negative (TN) from TN=PN&AN, false positive (FP) from FP=PP&AN, and false negative (FN) from FN=PN&AP. The accuracy statistics (CC, AC, Sn, Sp, QQ) are defined by

$$CC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}} \qquad AC = 1/2[\frac{TP}{AP}+\frac{TP}{PP}+\frac{TN}{AN}+\frac{TN}{PN}]-1$$

$$QQ = TP/(PP+AP-TP)$$

$$Sp = TP/PP \quad Sn = AP/PP$$

The exon accuracy statistics were Sn and Sp, where only exactly predicted exons considered, as well as the proportion of missed exons (MissE), the proportion of true exons not overlapped by any predicted exon, and Wrong Exons (WrongE), the proportion of predicted exons not overlapped by any real exon.

set:	CC_nuc	AC_nuc	Sen_nuc	Sp_nuc	QQ	Sn_exon	Sp_exon	MissE	WrongE
Training	0,87	0,87	0,94	0,91	0,88	0,66	0,63	0,11	0,14
Testing	0,88	0,88	0,95	0,95	0,91	0,71	0,68	0,07	0,1
Genomic	0,88	0,88	0,95	0,91	0,87	0,7	0,63	0,08	0,16

The accuracy statistics are close to the GENSCAN statistics on human single gene fragments (CC_nuc=0,93) for testing set. The accuracy statistics for multi-gene fragments are slightly weaker. There is an increase of WrongE in genomic set (0,16) to testing set (0,10); Sp_exon decreases from (0,68) to (0,63). The nucleotide accuracy isn't worse. The run time for the program is quadratic with sequence length (100 kb is processed nearly 8 minutes on Athlon 550 processor).

Discussion

A nice feature of the HMM approach is the fact that it is simple to implement new states and probabilistic models, also the HMM model let as to incorporate the state durations (lengths) using the empirical length distributions derived from training data set. The second fact is very important and has been accounted in the sophisticated alignment model (Pachter et al., 2001). The most powerful statistical-based gene-finding program GENSCAN (Burge, Karlin, 1997) incorporated the majority of advantages of HMM. The accuracy of the all statistical-based gene prediction programs is much weaker than homology-based programs on large genomic sequences with long intergenic regions (Guigo et al., 2000). However the statistical-based programs are indispensable when studying genes not homologous to known proteins, e.g. species- or lineage-specific genes.

References

1. Holden D.W., Tang C.M., Smith J.M., (1994). Molecular genetics of Aspergillus pathogenicity. Antonie Van Leeuwenhoek. 65(3): 251-255.

2. Burge C., Karlin S. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 1997. Apr 25 268:1 78-94.
3. Guigó R. (2000). DNA composition, codon usage and exon prediction. M.Bishop, editor: Genetic Databases. Academic Press. 53-80.
4. Guigo R., Agarwal P. et al. (2000) An assessment of gene prediction accuracy in large DNA sequences. Genome Res. 200 Oct; 10(10): 1631-42.
5. Pachter L., Alexandersson M., Cawley S. (2001) Application of Generalized Pair Hidden Markov Models to Alignment and Gene Finding Problems. RECOMB.

IDENTIFICATION OF CODING REGIONS IN GENOMES OF LOWER EUKARYOTES BY COMPOSITIONAL SEGMENTATION OF COMPLETE GENOMES

[1] Paskhin A.I., [2] Ramensky V.E., [3] Gelfand M.S., [3] Makeev V.J.

[1] Moscow Institute of Physics and Technology, Dolgoprudny, 141700, Moscow Region, Russia
[2] Institute of Molecular Biology, RAS, 119991, Moscow, Russia
[3] GosNIIGenetika, 113545, Moscow, Russia

***Key words**: complete genome, nucleotide composition, segmentation, gene recognition*

Resume

Motivation: It is known that the protein-coding regions exhibit a lower degree of compositional variation than the intergenic regions or introns. In this study we tried to consider an "inverse problem", that is to segment the sequence of a complete genome into regions of homogeneous composition and then to compare those regions with various functional units of a genome.

Results: We applied our program of compositional segmentation into compositionally homogeneous domains to several sequences of complete chromosomes of lower eukaryotes. We have identified DNA segments of the three main types: a) coding segments, b) long repeats, c) AT-rich intergenic segments. After filtering out the AT-rich regions and the repeats, we have obtained the quality of the prediction of coding sequences at the nucleotide level comparable to that of specialized gene finding programs. Several additional candidate coding regions have been identified.

Introduction

Genomic DNA exhibits a well-pronounced segmented structure. Certain segments code for proteins; others have regulatory functions; long AT-rich sequences can be found in the intergenic space. The problem of correctly identifying these main functional segments in the newly sequenced genomes is of significant interest for bioinformatics.

It is well known that compositional variations in a genome often correlated with functional units (Guigy, Fickett, 1995; Herzel, Grosse, 1997). Recently, a number of algorithms to assess compositional heterogeneity in a symbolic sequence have appeared (Oliver et al., 1999; Ramensky, Makeev, 2001). With such a tool at hand one can try to solve an "inverse problem" i.e. to find all compositionally homogeneous regions in a genomic sequence and compare them with functionally annotated units.

The structure of genomes of higher eukaryotes is very complex but lower eukaryotes have much shorter chromosomes, and the average number of exons in a gene is usually small. A good object for a study of this kind is the genome of *P. falciparum*, which is currently being sequenced. This genome has a very high percentage of AT, and one can hope that the protein composition apply such constrains on the coding sequences that the latter can be immediately identified by the compositional segmentation method. Another protozoan genome found in the GeneBank is the I chromosome of *L. major*, which, in contrast, is rather GC-rich. In this case the difference between coding and non-coding regions is small and one can only hope that the intergenic segments display a greater degree of compositional variation.

Methods

We used our segmentation software BASIO (Ramensky, Makeev, 2001) to identify coding DNA segments in several sequences of complete chromosomes of *P. falciparum* and *L. major* taken from GeneBank (www.ncbi.nlm.nih.gov). BASIO is based on the algorithm for sequence segmentation into domains with homogeneous composition that combines the direct Bayesian estimator with dynamic programming approach.

The main parameter that determines the behavior and efficiency of the program is *BIP* – the penalty for introducing a new inter-segment boundary. The value of this parameter was determined in our experiments on decomposition with artificial block-random sequences with known location and composition of compositional domains. The best performance was obtained for BIP=3.0–5.0 (Makeev et al., 2001).

The segmentation was performed as follows:

1. The complete genome was cut into 11,000 bp long fragments, with 1,000 bp overlaps between subsequent fragments.

2. The trivial DNA compositional segments (short segments with high compositional bias, e.g. containing only one or two symbols) were identified within those fragments by segmentation with a small value of *BIP* (0.3 – 0.5), thus dramatically reducing the amount of allowed boundaries.

3. Fragments were merged back in a single file, containing the set of the allowed boundaries. In the overlapping regions, the boundaries were allowed if at least one of the overlapping sequence contained a boundary.

4. The merged sequence was segmented again with a higher *BIP* (3–5) with only the boundaries obtained in the previous step allowed.

The predicted DNA segments were sorted by their length and GC-content, and the segments with low GC-content were discarded. The obtained list of fragments was compared with the list of coding DNA segments (CDS), as described in the GeneBank annotation. The correlation between the annotated segments and the long homogeneous segments with moderate AT composition was estimated using Pearson association coefficient (Lakin, 1980):

$$CC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP the number of correctly identified nucleotides within the CDS (True Positives), and TN, FP, and FN are true negatives, false positives, and false negatives respectively.

Results

It was found that the best agreement for *P. falciparum* (CC above 0.65) was achieved with the parameters *bip_part*=0.5, *bip_whole*=3.5, and the GC content cutoff at 17%. In this case, the number FNC of not identified CDS nucleotides was less than 1%. The analysis of falsely identified (FPC) CDS revealed that many segments contained open reading frames (ORFs) longer than 300 bp (see Table 1). Applying of BLAST search to those predicted ORFs resulted in identification of several new genes not annotated by the GeneBank.

Table 1.

Genome	Genome length	Number of CDS	CC	FNC	FPC	ORF
P. falciparum 2	947103	349	0.667	3	77	14
P. falciparum 3	1060106	405	0.669	7	39	10
L. major 1	268984	79	0.717	0	35	21

Table 2 displays a number of BLAST comparisons of the ORFs found with non-redundant protein database. In the vast majority of examples proteins were found in other versions of annotation of *P. falciparum* (shown with italic; there are several versions of annotation of *P.falciparum* chromosome III in the GeneBank). The most interesting case is AL031747, which has a closest "relative" in the first chromosome of *P. falciparum.* The other interesting case is NC_001905 of *L. major*, which exhibit a strong homology to the other part of the same chromosome. This can be a recent duplication or a sequence assembling error.

Table 2.

Genome	Segment		E value	BLAST similarity
	begin	end		
P. falciparum 3	46448	46900	8 E-31	AL031747
P. falciparum 3	47161	47631	4 E-43	NC_000910
P. falciparum 3	228467	228979	2 E-53	AL034558
P. falciparum 3	251951	252295	7 E-20	AL034558
P. falciparum 3	282254	282787	3 E-55	AL034558
P. falciparum 3	457854	458315	3 E-55	AL008970
P. falciparum 3	734207	734563	6 E-49	NC_000521
P. falciparum 3	899743	900057	8 E-40	AL034559
L. major 1	136524	137108	4 E-22	NC_001905

Discussion

It is clear that the results of so conceptually simple a method as compositional segmentation displays a very high correlation with coding segments in the chromosomes of lower eukaryotes. This fact can be explained by a very simple structure of the genes, which usually contain one or two exons, one of which is often very long. Our procedure identifies these long exons. Two comments should be made here. (i) No information on signals (stop codons, initiation codons, splice sites) was included, hence the usual deviation of about 10% of segment boundaries from the CDS annotation, which significantly reduced CC. (ii) Some of FPC of *P. falciparum* can still contain genes consisting of several shorter exons.

All in all, the statistics of the compositional variations in the coding and in the intergenic regions is dramatically different. Long homogeneous sequences fall into one of three classes: coding, divergent repeats, and AT-rich tracts, and all other genome regions exhibit a much higher degree of compositional variation. Such observations were already reported (Guigo,

Fickett, 1995; Herzel, Grosse, 1997) by comparison of compilations of coding vs non-coding sequences. However, we believe, that the demonstration of the fact that the "inverse problem" can be solved is of interest.

It should be noted that gene prediction in a sequence containing many genes (e.g. a complete chromosome sequence) still has room for improvement. The best software tools for statistical gene identification (Glimmer (Salzbert et al., 1998), GeneMark.HMM (Borodovsky, Lukashin, 1998), GeneScan (Burge, Karlin, 1997) and others) display in this case a substantial drop in the gene identification quality as compared to the experiments on sets of sequences containing a single gene (Guigy et al., 2000). In this experiments of artificially generated "genomic fragments" with random intergenic regions GeneScan demonstrated average CC~0.76 on the set of ~170,00 b.p. long sequences (comp. to CC~0.97 for a single gene set). These sequences did not contain pseudogenes, repeats and huge introns, thus the reported quality might still be overestimated. True that the experiments described related to the human genes, with their much more complex exon-intron structure and our simple approach cannot be used to identify genes in higher eukaryotes. However, our algorithm is not specialized particularly for gene finding, it is global at the genomic scale, reasonably fast, and so it can be used for initial studies of newly sequenced genomes with expected simple gene structure.

The application of our segmentation method to the GeneBank sequences resulted in two interesting observations. First, many coding fragment could be subdivided into a few sub-fragments with a homogeneous DNA contents but different GC-concentration. Perhaps, such sub-structure reveals the domain organization of proteins. Second, the coding segments proved to be always more GC-rich than the non-coding ones. Even in *L. major* with its 66% GC genome a good agreement with the annotation occurs when the segments with lower GC are filtered out.

Acknowledgements

This research was partially supported by grants from the Howard Hughes Medical Institute (55000309), INTAS (99-1476) and the Russian Fund of Basic Research (00-15-99362, 02-04-49111).

References

1. Ramensky V.E., Makeev V.J., Roytberg M.A., Tumanyan V.G. (2001) Segmentation of long genomic sequences into domains with homogeneous composition with BASIO software. Bioinformatics. 17(11): 1065-6.
2. Makeev V.J., Ramensky V.E., Gelfand M.S., Roytberg M.A., Tumanyan V.G. (2001) Bayesian approach to DNA segmentation into regions with different average nucleotide composition. Lecture Notes in Computer Science. 2066 Springer–Verlag. 57–74.
3. Burge C., Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.
4. Guigo R., Fickett J.W. (1995) Distinctive sequence features in protein coding, genic non-coding and intergenic human DNA. J. Mol. Biol. 253, 51-60.
5. Guigo R., Agarwal P., Abril J.F., Burset M., Fickett J.W. (2000) An assessment of gene prediction accuracy in large DNA sequences. Genome Res. 10:1631-1642.
6. Herzel H., Grosse I. (1997) Correaltions in DNA sequences: The role of protein coding segments. Phys. Rev. E. 55, 800-810.
7. Lakin G.F. (1980). Biometrija (*The biometry*). Moscow. Vysshaja shkola.
8. Lukashin A.V., Borodovsky M. (1998) GeneMark.hmm: new solution for gene finding. Nucl. Acids Res. 26:1107-1115.
9. Oliver J.L., Román-Roldán R., Pérez J., Bernaola-Galván P. (1999) SEGMENT: identifying compositional domains in DNA sequences. Bioinformatics. 15, 974-979.
10. Pizzi E., Frontali C. (2000) Low-complexity regions in *Plasmodium falciparum* proteins. Genome Res. 11: 218-229.
11. Salzberg S.L., Delcher A.L., Kasif S., White O. (1998) Microbial gene identification using interpolated Markov models. Nucl. Acids Res. 26:544-548.

SYNONYMOUS CODON USAGE PECULIARITIES IN *ESCHERICHIA COLI* PROTEIN-CODING GENES AND NUCLEOTIDE FREQUENCY DISTRIBUTION IN HOMOLOGOUS GENES OF ATP-SYNTHASE

[1] Ermagambetov A.M., [*2] Ivashchenko T.A., [1] Ivashchenko A.T., [1] Gabdulina Zh., [1] Goncharova A.V., [1] Karpenjuk T.A.

[1] al-Farabi Kazakh National University, Almaty, 480078, Kazakhstan, e-mail: bbpp.kafedra@mailkazsu.uni.sci.kz
[2] Ajtkhozhin Institute of Molecular Biology and Biochemistry, Almaty, 480012, Kazakhstan, e-mail: timour@itte.kz
[*]Corresponding author

Key words: *synonymous codons, protein coding genes, expression, translation, genome, Escherichia coli, ATP synthase, mitochondria, chloroplasts, eubacteria*

Resume

Motivation: Synonymous codon usage in different genes within a single genome as well as in homologous genes of different genomes remains a topical problem that is far from resolution. The present work is devoted to the study of the interdependence of synonymous codon usage changes in the genes of *Escherichia coli* and nucleotide frequency distribution at the third codon position in the genes of ATP synthase subunits.

Results: Based on the protein coding gene sampling of *E. coli,* the correlation in synonymous codon usage for some amino acids has been detected. The changes in the ratios of synonymous codons AAC/AAU (Asn), UGC/UGU (Cys), and AGC/AGU (Ser) in subgroups of genes are characterized by positive correlation coefficient (r = 0.985–0.996). The changes in the ratios of synonymous codons AAC/AAU (Asn) and the ratios of synonymous codons of amino acid group GUC/GUU (Val), GCC/GCU (Ala), UCC/UCU (Ser), CCC/CCU (Pro) negatively correlate (r = –0.885 to –0.985). For the amino acid group AGG/AGA (Arg), CAG/CAA (Gln), CUG/CUA (Leu), CCG/CCA (Pro) with purine nucleotides at the third codon position, a high correlation in the changes of codon ratios NNG/NNA are also detected (r = 0.917–0.978). It was shown that nucleotide usage at the third codon position in the genes of ATP synthase A, B, D, E, F, G, H, and I subunits in chloroplasts, mitochondria, and bacteria conform to the "four-nucleotide-rule" (FNR), which is described by the linear regression equation: $(fA/fT - fC/fG) = \boldsymbol{a}(fA + fG - 0.5) + \boldsymbol{b}$, where ***a*** and ***b,*** are the regression parameters. The synonymous codon usage in the genes studied is discussed in connection to the FNR pattern.

Availability: The C++ source code is available free on the request to the author: timour@itte.kz.

Introduction

The problem of synonymous codon usage is interesting in several aspects. It is accepted that some codons are required for an accelerated protein synthesis while others reduce the translation rate. This problem is linked to the tRNA pool that, in turn, correlates with the amino acid usage in proteins encoded in a given genome. There are several peculiarities of synonymous codon usage not only in different genomes but also in different genes within the same genome (Antezana, Kreitman, 1999; Holm, 1986; Morton, 1998; Seetharaman, Srinivasan, 1995; Shimada, Sugiura, 1999). In connection to the established limitations on the nucleotide frequencies in single-stranded DNA (Ivashchenko et al., 1999; 2000a; 2000b), including the third codon position, it seems interesting to ascertain how this limitation affects the selective synonymous codon usage. The aim of this study is to reveal the peculiarities in synonymous codon usage in protein coding genes of *Escherichia coli* and nucleotide frequency distribution at the third codon position in the genes of ATP synthase A, B, D, E, F, G, H, and I subunits in chloroplasts, mitochondria, and bacteria.

Methods

For each gene (mRNA), we calculated the codon frequencies of synonymous codons NNG, NNA, NNU, or NNC (N, any nucleotide). Then, for the whole gene, we calculated the ratios of the codon pares with purine (NNG/NNA) and pyrimidine (NNC/NNU) nucleotides at the third codon position. For the synonymous codon pares of each amino acid, the ratios of NNG/NNA or NNC/NNU were determined and expressed as the percentage to NNG/NNA and NNC/ NNU values for the whole gene.

Implementation and Results

In randomly selected 300 genes of *E. coli* with lengths of more than 300 nucleotides, codon frequencies were calculated and all the genes within a group ranged by the ratio NNC/NNU of the codon sums corresponding to the amino acids.

Using this test, all the genes were divided into 5 groups with respect to the intervals of NNC/NNU values: 0–100%, 100–150%, 150–200%, 200–250%, and over 250%. Then, we calculated mean values of synonymous codon ratios of all the amino acids for each group. These data for the five groups are shown in table. The results indicate that for some synonymous codon pares, there is an increase in NNG/NNA and NNC/NNU values from group to group, and for the other codon pares, there is a significant decrease or insignificant changes take place. Thus, for the AAC/AAU (Asn), AGC/AGU (Ser), and UGC/UGU (Cys) codon pares, the ratio increased fore–fivefold, while for the GUC/GUU (Val), GCC/GCU (Ala), UCC/UCU (Ser), and CCC/CCU(Pro) pares, the decrease was twofold. The significant changes in synonymous codon ratios with purine nucleotides at the third codon position were also detected for Leu: CUG/CUA. All these changes have rather high correlation. Thus, the correlation coefficients between the AAC/AAU (Asn) and AGC/AGU (Ser), UGC/UGU (Cys), CUG/CUA (Leu) are r = 0.996, 0.986, and 0.966, respectively. It is apparent that these changes are not random and are determined by some limitations on the nucleotide usage at the third codon position. Differential synonymous codon usage in protein-coding genes of *E. coli* suggested to be linked to gene expression (Holm, 1986), and there is a correlation between the codon usage and availability of the corresponding tRNAs. However, the data from the table indicate that there is no significant correlation for a major number of synonymous codon pares. For instance, the changes of synonymous codon ratios GAC/GAU, GGC/GGU, AUC/AUU, UAC/UAU, CAC/CAU, UUC/UUU, CUC/CUU, ACC/ACU, AAG/AAA, GGG/GGA, GUG/GUA, and GCG/GCA were either detected or opposite to the tRNA gene number. We have been shown that the nucleotide content in protein coding genes from chloroplasts, mitochondria, and bacteria conform to the four-nucleotide-rule pattern (Ivashchenko et al., 1999; 2000a; 2000b) described by the linear regression: $(fA/fT - fC/fG) = \mathbf{a}(fA + fG - 0.5) + \mathbf{b}$, where **a** and **b** are regression coefficients; *fA, fT, fC,* and *fG,* nucleotide frequencies in genes.

The ratio of the tRNA genes with synonymous anticodons in *E. coli* genome and the synonymous codon ratios for different amino acids expressed as percentage to the NNC/NNU and NNG/NNA values for the whole gene.

tRNA	Acid	Codons	1	2	3	4	5
4/0	Asn	AAC/AAU	58	98	153	196	324
3/0	Asp	GAC/GAU	58	57	52	48	73
4/0	Gly	GGC/GGU	140	138	151	115	88
1/0	Ser	AGC/AGU	111	175	213	279	452
3/0	Ile	AUC/AUU	78	73	68	86	82
3/0	Tyr	UAC/UAU	77	71	67	74	100
0/4	Arg	CGC/CGU	137	127	87	96	74
1/0	Cys	UGC/UGU	77	109	162	215	282
1/0	His	CAC/CAU	68	57	70	81	76
2/0	Fhe	UUC/UUU	60	56	68	82	95
1/0	Leu	CUC/CUU	99	103	99	118	95
2/0	Val	GUC/GUU	119	97	94	61	51
2/1	Thr	ACC/ACU	303	246	271	257	237
2/0	Ala	GCC/GCU	226	186	173	153	100
2/0	Ser	UCC/UCU	147	110	115	89	79
1/0	Pro	CCC/CCU	110	80	68	63	46
0/6	Lys	AAG/AAA	29	20	23	20	17
0/4	Glu	GAG/GAA	44	31	30	29	24
1/1	Gly	GGG/GGA	84	90	98	108	81
1/1	Arg	AGG/AGA	29	41	40	52	59
1/0	Arg	CGG/CGA	90	104	94	138	97
2/2	Gln	CAG/CAA	97	110	127	132	152
0/5	Val	GUG/GUA	133	147	187	146	139
1/1	Leu	UUG/UUA	68	54	56	61	66
4/1	Leu	CUG/CUA	634	614	910	1130	2331
1/0	Thr	ACG/ACA	124	123	109	189	195
0/3	Ala	GCG/GCA	90	104	94	138	97
1/1	Ser	UCG/UCA	59	81	94	85	80
1/1	Pro	CCG/CCA	123	164	177	232	242

Since this rule is true for any representative sampling of nucleotides, it must be applicable to the nucleotides at the third codon position, i.e. for synonymous codons. This proposal was confirmed on the ATP synthase homologous genes A, B, D, E, F, G, H, and I subunits in chloroplasts, mitochondria, and bacteria sampling. It was shown that the nucleotide frequencies at the third codon position form the gene pattern in such a way that the nucleotide content for the whole gene to a higher extent conform to the four-nucleotide-rule with the ***a*** coefficient close to 8, which is characteristic of protein-coding genes with arbitrary codon usage. The results of the present work demonstrate that nucleotide substitutions at the third codon position are not random but limited by the four-nucleotide-rule. From this rule it follows that for the *fA/fT* and *fC/fG* constancy for the whole genome, it is necessary to balance the increase/decrease in *fA/fT* and *fC/fG* values for different amino acids. It can be seen from the table that only a part of all the codon pares markedly changes for the maintenance of $(fA/fT - fC/fG) = \mathbf{a}(fA + fG - 0.5) + \mathbf{b}$. From a number of publications (Sharp et al., 1986; Sharp, Li, 1987; Shield et al., 1988; Shimada, Sugiura, 1999; Zhang et al., 1991), it follows that the synonymous codon ratio significantly changes depending on the expression (translation) rate in codon pares of only several amino acids. The most significant changes are detected for the following pares: AAC/AAU (Asn), AGC/AGU (Ser), UGC/UGU (Cys), GUC/GUU (Val), CCU/CCC (Pro), GAG/GAA (Glu), and CUG/CUA (Leu). The list of the objects (genomes of organelles, prokaryotes, and eukaryotes) is an evidence of the universality of the rule indicated above.

References

1. Antezana M.A., Kreitman M. (1999). The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. 49, 36-43.
2. Holm L. (1986). Codon usage and gene expression. Nucl. Acids Res. 14, 3075-3086.
3. Ivashchenko T.A., Ivashchenko A.T., Aitkhozhina N.A. (1999). Regularity of nucleotides usage in DNA. II. Peculiarities of chloroplast genomes of *Pinus thunbergii, Nicotiana tabaccum, Orysa sativum* and *Zea mays*. Biotechnology. Theory and practice. 11-12, 103-109.
4. Ivashchenko T.A., Goncharova A.V., Ivashchenko A.T. (2000a). Regularity of nucleotides usage in DNA. III. Four-nucleotide-rule in *Marchatia polymorpha* and *Arabidopsis thaliana* mitochondrial DNA. Biotechnology. Theory and practice. 13, 137-141.
5. Ivashchenko T.A., Kurmasheva R.T., Ivashchenko A.T. (2000b). Regularity of nucleotides usage in DNA. IV. Four-nucleotide-rule in DNA of bacteria. Biotechnology. Theory and practice. 13, 142-145.
6. Morton B.R. (1998). Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J. Mol. Evol. 46, 449-459.
7. Seetharaman J., Srinivasan R. (1995). Analysis of codon usage: positional preference in various organisms. Indian J. Biochem. Biophys. 32, 156-160.
8. Sharp P.M., Li W.-H. (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucl. Acids Res. 15, 1281-1295.
9. Sharp P.M., Tuohy T.M.F., Mosurski K.R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucl. Acids Res. 14, 5125-5139.
10. Shield D.C., Sharp P.M., Higgins D.G., Wright F. (1988). "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5(6), 704-716.
11. Shimada H., Sugiura M. (1999). Fine structural of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucl. Acids Res. 19, 983-995.
12. Zhang S., Zubay G., Goldman E. (1991). Low-usage codons in *Escherichia coli*, yeast, fruit fly and primates. Gene. 105, 61-72.

MUTATIONAL HOTSPOTS IN THE P53 GENE REVEALED BY CLASSIFICATION ANALYSIS

* ***Glazko G.V, Rogozin I.B.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk 630090, Russia, e-mail: gvg2@psu.edu
*Corresponding author:

Key words*: hotspot, p53 gene, mutation, classification, tumor, cancer, Li-Fraumeni syndrome*

Motivation: The p53 protein is a multifunctional transcription factor: it regulates cell cycle progression and interacts with several key proteins involved in DNA replication, transcription, repair and apoptosis. Nearly 60% of all human cancers of different origins are accompanied by mutational events in the p53 gene.

Results: We compared the mutational spectra in the p53 gene from different tumors, observed in germline cancer-prone families, in tumors of different histogenesis and their derived cell lines and second examine the p53 hotspots map in the evolutionary perspective. While spectra from different solid tumors share common hotspots with the germline spectrum (CpG sites), they also contain unique sets of new hotspots that are not observed in the germline spectrum (new CpG and non-CpG sites). Our analysis suggested that the distribution of hotspots in the p53 gene could be influenced by cell growth conditions, as well as the specificity of the mutagenesis and the bulk hotspots analysis do not take into account these differences.

Introduction

The p53 protein is a multifunctional transcription factor, nearly 60% of all human cancers of different origins are accompanied by mutational events in the p53 gene (reviewed by Agarwal et al., 1998). Levels and/or activity of p53 increase in response to DNA damaging agents, decreased oxygen, oncogenic stimuli, cell adhesion, altered ribonucleotide pools and redox stress. It is generally assumed that the tumor suppressor properties of the p53 protein are associated with committing cells with multiple DNA damages to apoptosis (Linke et al., 1997). Deletions, insertions and base substitutions in the p53 gene can violate its function and cells with any multiple DNA damages are not eliminated but instead accumulate mutations, which contribute to the tumor development. Experiments with null (p53-/-) mice show that the absence of p53 increases the probability of tumor susceptibility and this probability in turn depends on different genetic background and many other parameters. However, our understanding of how p53 mutations and even loss contribute to tumor initiation and progression remains incomplete.

It is well-known that many factors concur for contribution to the specific mutational spectrum: polymerase replication errors, DNA damages due to endogenous processes (e.g., deamination of methylated cytosine at CpG sites), DNA damages due to exogenous mutagens, and inefficient DNA repair. Further fixation of mutations in p53 may depend on the several conditions: 1) cell origin (tissue-specificity); 2) the stage (grade) of tumor progression (for spectra obtained from tumor cells); 3) the cell growth conditions (f. e. *in vitro* and *in vivo*, type of tumor, etc.). Besides, mutational spectra under consideration relate only to those "admitted" mutational events, which are compatible with leaving cell.

The influence of 1-3 factors mentioned above on the mutational spectra of the p53 gene may be discarded (Walker et al., 1999). By contrast, to reveal the finer resolution of p53 hotspots map this influence should be taken into account as full as possible from current data. Obviously, the 10000 mutations in the IARC database (release 3, 1999) is a large number only when considered as a whole; the number of mutations observed within a single tumor class is usually quit small (Hollstein et al., 1999). Several earlier studies showed that mutational spectra of the p53 gene in different types of tumors are different. For example, in skin and colon cancer the most frequently observed mutations are GC->AT transitions. While majority of the transitions found in skin cancer are at dipyrimidine sites, which are the primary sites for UV photo lesions, the transitions observed in colon cancer are localized in CpG sites (Beroud, Soussi 1998; Hainaut et al., 1998). Transversions, such as GC->TA, which are predominant in lung cancer, are not found in skin cancer (Beroud, Soussi, 1998). Differences between mutational spectra in lung, bladder and all other tumors of different origin were found by Dogliotti et al. (1998). It was shown that p53 mutation profile in brain tumors closely resembles the set of germline mutations identified in Li-Fraumeni cancer syndrome families. This data and the assumption that germ cells, from which founder mutation arise, are presumably better protected from the external environment than somatic epithelial cells from which most sporadic human cancer arise, may provide together the first hint on the approximation of spontaneous mutations in the p53 gene (Hollstein et al., 1999). It was shown also, that mutational spectra of the p53 gene are related is some organs and in some other do not (Lutz et al., 1998). However, the classification of different p53 mutation frequencies and comparison of hotspots distribution taking into account the possible influence of 1-3 factors (in different tumors (*in vivo*) and their derived cell lines (*in vitro*), as well as in tumors on different stages of tumor progression) have not yet been done. But to reveal the influence of environmental carcinogenic risk factors, for example, it would be useful to know first the tissue-specific and spontaneous mutation patterns.

Methods and Algorithms

The analysis of mutational spectra in the p53 gene was carried out using the CLUSTERM and CLUSTERP computer programs (Glazko et al., 1998; Rogozin et al., 2001; URL:www.itba.mi.cnr.it/webmutation). The original CLUSTERM program separates a mutational spectrum into statistically homogeneous classes. It was shown that the distribution of mutational events can be approximated by a mixture of binomial distributions with different parameters (probabilities and weights). A single distribution corresponds to a single class of sites (positions with observed mutations) with equal probabilities for mutations to be observed. This model permits operate with relatively small sample size (approximately 30-200 mutations), otherwise Poisson approximation of binomial distribution can be used (CLUSTERP program). The class of hotspots is defined as the class with the greatest binomial probability. However, this definition is not sufficient for mutational spectra in the p53 gene and will be discussed below.

Implementation and Results

In this work we analyzed 10 mutational spectra in p53 gene. Numbers of mutations were different for different tumors (from 43 to 398 mutations). For correct classification by CLUSTERM it is important to define the number of zero-sites, that is the sites where mutations would be observed (in the case of p53 could be related to tumor progression, for example) but were not revealed so far due to restricted sample sizes (Rogozin et al., 2001). This problem is extremely complicated for the p53 gene since biological meaning of zero-sites in this case is not clear. We tested various approaches for estimation of zero-sites number. The simplest one was an heuristic approach when all 10 spectra were merged together and only sites in which more than 1 mutation was observed were accounted. Number of sites with 2 or more mutations (S2 = 193) is significantly smaller in comparison with total number of sites where at least one mutation was revealed (S1 = 301), however we used S2 since sites with one mutation may represent sequencing errors and rare polymorphic sites. Number of zero-sites was calculated for each spectrum as S2 - M, where M - number of sites within this spectrum. With these numbers of zero-sites significant deviation from the expected distribution has been found for Blood and Bl_cell spectra, signal that in these cases the numbers of zero-sites were probably overestimated (113 and 166 zero-sites, respectively).

Discussion

We used Poisson approximation of binomial distribution and Poisson-derived numbers of zero-sites (Rogozin et al., 2001) For one Poisson distribution the expected number of zero-sites $Z(0) = Z(1)^2/Z(2)\times 2$, where Z(i) is a number of sites where i mutations were observed. This equation is a result of simple transformation of the relationship $m = (i)\times Z(i)/Z(i-1)$, where m is the mean of the Poisson distribution (Topal et al., 1986). In this approach we assume that zero-sites are drawn from the classes with 1 and 2 mutations, which in turn almost always are contained in a class with minimum mutation frequency. Occasionally some low-frequency sites can be misclassified but when there are hotspots such cases should be rare. Using this approach for all spectra we did not found any significant deviations from expected distributions, that is Poisson-derived numbers of zero-sites fitted the mixture of Poisson distributions quite well. Comparing two approaches (Poisson-derived and heuristic number of zero-sites) one can see that in some cases the last gave strong overestimate of the expected numbers (Table). Poisson-derived approach was implemented in a CLUSTERP program, the version of CLUSTERM program developed for separation of a mixture of Poisson distributions using Poisson-derived estimates of zero-sites.

Table. The numbers of substitutions for tumors of different origin in databases for p53 mutations (Hainaut et al., 1998; Beroud, Soussi, 1998). C – carcinoma; A – adenocarcinoma.

Spectrum identifier		S[1)]	N[2)]	Number of zero-sites[3)] estimated using different approaches	
				Heuristic	Poisson-derived
Germline		105	182	137	126
SOLID:					
Bladder	A	81	147	134	88
	C	212	265	84	156
Colon	A	133	146	131	84
	C	238	224	111	142
Lung	A	229	233	77	117
	C	398	231	50	88
SUSPENSIONAL:					
Lung_cell		78	136	149	87
Blood		186	123	113	36
Bl_cell		43	59	166	32

[1)]Total number of substitution; [2)] Total number of different sites per tumor violated by substitutions; [3)] Sites without substitutions.

The majority of mutational spectra in p53 gene has been divided by the CLUSTERP program into two classes. Usually the first class includes mutations with relatively low probability of occurrence: "cold spots", the number of mutations 1-7 and zero-sites; the second class includes mutations with higher probability of occurrence: hotspots, the number of mutations 8-21. In the germline sample the number of mutation in the second class is 5-9.

Conclusion

In summary, the following results are obtained from our investigation of the p53 gene hotspots: (1) sets of overlapping hotspots are observed that are present in both the germline spectrum and spectra from tumors of different origin (CpG sites); (2) some of the germinal hotspots are absent in different tumors; (3) new hotspots absent in the germline spectrum are observed for solid tumors (CpG and non-CpG sites); (4) hotspots are nearly absent in the mutational spectra from lymphomas and cell lines *in vitro* (5) on the protein level in all tumors hotspots were observed in the regions which constitute the p53 protein-consensus DNA interface in the p53-target genes but in suspensional tumors their number was reduced to two DNA-contact residues (R273, R248) and (6) the evolutionary substitution rate was inversly correlated with mutational rate in tumors. Presumably tumor-supressor function of p53 does not depend on the presence of mutations in sites, which were revealed as hotspots for patients with Li-Fraumeni syndrome. The obtained data suggests that one of important factors, influencing the distribution of hotspots in the p53 gene, is peculiarities of cell growth condition in solid tumors of different origins. The cell-specific constraints (cell origin and growing condition) might provide the loss- and gain-of-function mutants with benefits for tumor development. That is, the tumor-specificity of p53 mutational spectra presumably depends on the condition of tumor cell proliferation, as well as on specificity of mutagenesis.

Acknowledgements

We are thankful to T.T.Glazko and B.A.Rogozin for stimulating discussions and helpful comments. This work was supported by the Russian Fund of Fundamental Research (grant № 02-04-48342).

References

1. Agarwal M.L., Taylor W.R., Chernov M.V., Chernova O.B., Stark G.R. (1998) The p53 network. J. Biol. Chem. 273, 1-4.
2. Beroud C., Soussi T. (1998) p53 gene mutation: software and database. Nucl. Acids Res. 26, 200-204.
3. Dogliotti E., Hainaut P., Hernandez T., D'Errico M., DeMarini D.M. (1998) Mutation spectra resulting from carcinogenic exposure: from model systems to cancer-related genes. Recent Results Cancer Res.154, 97-124.
4. Glazko G.V., Milanesi L., Rogozin I.B. (1998) Subclass approach for mutational spectrum analysis: application of the SEM algoritm. J. Theor. Biol. 192, 475-487.
5. Hollstein M., Hergenhahn M., Yang Q., Bartsch H., Wang Z.Q., Hainaut P. (1999) New approaches to understanding p53 gene tumor mutation spectra. Mut. Res. 431, 199-209.
6. Linke S.P., Clarkin K.C., Whal G.M. (1997) p53 mediates permanent arrest over multiple cell cycles in response to gamma-irradiation. Cancer Res. 57, 1171-1179.
7. Lutz W.K., Fekete T., Vamvakas S. (1998) Position- and base pair-specific comparison of p53 mutation spectra in human tumors: elucidation of relationships between organs for cancer etiology. Environ. Health Perspect. 106, 207-211.
8. Rogozin I.B., Kondrashov F.A., Glazko G.V. (2001) Use of mutation spectra analysis software. Hum. Mutat. 17, 83-102.
9. Walker D.R., Bond J.P., Tarone R.E., Harris C.C., Makalowski W., Boguski M.S., Greenblatt M.S. (1999) Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 structural and functional features. Oncogene. 18, 211-217.

MECHANISMS OF MUTAGENESIS AND THE ROLE OF LOCAL DNA SEQUENCE COMPLEXITY

*[1,2] ***Chuzhanova N.A.,*** [3] ***Cooper D.N.***

[1] Department of Computer Science, Cardiff University, PO Box 916, Cardiff CF24 3XF, UK
[2] Institute of Mathematics, RAS, 630090, Novosibirsk, Russia
[3] Institute of Medical Genetics, University of Wales College of Medicine, Cardiff CF14 4XN, UK
e-mail: nadia.chuzhanova@cs.cardiff.ac.uk

Key words: *sequence complexity, micro-deletion, micro-insertion, indel, mutagenesis*

Resume

Motivation: A detailed study of deletion and insertion mutagenesis could improve our understanding of the molecular mechanisms underlying micro-insertions, micro-deletions and indels and could be an invaluable aid to the optimisation of mutation search strategies in molecular diagnostic medicine.

Results: In the present study we examine the local DNA sequence complexity and its role in previously postulated mechanisms for deletion and insertion mutagenesis such as slipped mispairing and strand switching, secondary loop excision and quasi-palindrome correction, Moebius loop resolution and excision. A novel mutational mechanism mediated by the insertions of inverted repeats is proposed. It was found that the mechanism of mutation and the type of repeat involved into micro-insertions and micro-deletions depend upon the change in a certain complexity measure. Data from the Human Gene Mutation Database (Krawczak et al., 2000a) is used to compare and contrast 3767 micro-deletions, 213 different indels and a small amount of insertions in order to propose mechanistic processes that could account for their genesis.

Introduction

The most common types of mutation causing human genetic disease are single base-pair substitutions and micro-deletions (Antonarakis et al., 2001). The remainder comprises an assortment of larger deletions, insertions, inversions, expansions and complex rearrangements. One relatively uncommon type of mutation is *indel,* a combined micro-deletion/ micro-insertions that result in the apparent replacement of one or more base-pairs by others, not necessarily the same number. Although it is likely that the component micro-insertion and micro-deletion events occur contemporaneously (i.e. as part of the same complex mutational event), this need not necessarily be so. The study of mutational lesions causing human genetic disease has revealed that, irrespective of their type, the nature, frequency and location of mutations are invariably non-random (Cooper, Krawczak, 1993; Antonarakis et al., 2001), being strongly influenced by the complexity of the local DNA sequence environment (Krawczak et al., 2000a). Complexity analysis was used to examine the change in complexity and the subsequent involvement of different types of repeat in deletion/insertion event.

Methods and Algorithms

Complexity analysis, as devised by Gusev et al. (1999), was used to examine the potential contribution of the local DNA sequence complexity to the several postulated mechanisms of deletion/insertion mutagenesis and also to the two-step process of indel formation.

Complexity analysis is based on a definition of sequence complexity by taking into account different types of repetitive element including direct and inverted repeats and inversions thereof occurring in a given sequence. One can conceive of the sequence as being decomposed into "words", where each word is the longest among all possible words for which a direct or inverted repeat, or an inversion thereof, occurs somewhere upstream of the current position. An overlap between two repeat copies is also permitted. The length of the first fragment is always 1. It is apparent that this decomposition, H(S), contains the minimal number of words. The number of words in this minimal decomposition, H(S), is called the complexity of S. We shall denote respectively by C1(S), C3(S), C4(S), C5(S) and C2(S) the complexities or number of words in decompositions H1(S) computed with respect to direct repeats, H3(S) with inverted repeats, H4(S) with symmetric elements, H5(S) with inversions of the inverted repeats and H2(S) computed using a combination of all the above mentioned types of repeat.

Let us consider as an illustrative example a DNA fragment S from the *RET* gene with reported indel delCCinsGG (the precise location of the indel and the deleted nucleotides are indicated by lower case letters). Decompositions of S into words and the corresponding complexities are:

* Corresponding author.

H1: A-C-G-A-G-C-T-G-TG-c-cG-C-ACG-GTG-A-T; C1(S)=16;
H3: A-C-G-A-GCT-GT-Gc-cG-CAC-GG-TG-A-T; C3(S)=13;
H4: A-C-GAGC-T-GT-Gc-cG-CACG-GTG-A-T; C4(S)=11;
H5: A-C-G-A-GCT-G-TGc-cG-CACGG-TG-A-T; C5(S)=12;
H2: A-C-GAGC-TG-TGc-cG-CACGG-TG-A-T; C2(S)=10.

Some of the repeated words are underlined. Different types of repeats are marked by arrows in order to indicate their orientation. Parameters C1–C5 represent suitable measures of complexity (regularity) of a sequence S, since any abundance in S of direct and inverted repeats and inversions thereof serves to reduce the corresponding complexity. As can be seen from the above example, the minimum complexity among C1, C3–C5 is achieved with measure C4. The second measure in ascending order is measure C5. Fragment S is thus comparatively rich in symmetric repeats and inversions of the inverted repeats that together make a major contribution to measure C2. To be able to compare the complexities of two or more sequences that differ in length, one can use the *complexity per base*, c=C/N, where N is the length of a given sequence in base-pairs.

Intuitively, complexity should decrease as a consequence of deletion and increase as a result of an insertion but this is not invariably so. Whether the complexity is increased or decreased can be strongly influenced by the appearance or disappearance of prominent repeats through either the insertion or deletion of bases. We show later that the mechanism of mutation and the type of repeat involved into micro-insertions and micro-deletions depend upon the change in a certain complexity measure.

Implementation and Results

Direct repeats and slipped mispairing models. One mechanism frequently implicated in the generation of micro-deletions and micro-insertions is slipped mispairing that involves the misalignment of short direct repeats. During DNA replication, the template strand can slip forward, producing a single-stranded loop that can subsequently be excised and repaired thereby fixing a micro-deletion. This event serves to decrease the C1 complexity of a sequence. Thus, for example, a short micro-deletion in the *APC* gene gives rise to a slight decrease in complexity C1 from 13 to 12. The decompositions are as follows (two direct repeats causing micro-deletion are underlined):

H1 before deletion: A-AAA-G-AA-T-AGa-tag-T-C-T-TC-CTT-TA, C1=13, c1=0.54;
H1 after deletion: A-AAA-G-AA-T-AG-T-C-T-TC-CTT-TA, C1=12, c1=0.60.

Conversely, one of the nascent strands may slip backwards thereby templating a micro-insertion. In this case, the complexity of a sequence bearing an insertion must either remain the same or can increase slightly owing to the presence of imperfect direct repeats. For example, an insertion of gcg into the *AAP* gene does not change the complexity of the fragment. The decompositions are as follows:

H1 before insertion: G-G-A-GG-C-GGCGGCGGC-C-A-CCA, C1=9, c1=0.45;

H1 after insertion: G-G-A-GG-C-GGCGgcgGCGGC-C-A-CCA, C1=9, c1=0.39.

A modified slipped mispairing model was proposed by Krawczak and Cooper (1991) to account for deletions not readily explicable by the standard model. If the DNA sequence flanking the deleted bases also occurs as a contiguous sequence in the immediate vicinity, the intervening non-homologous bases may loop out thereby potentiating the formation of a second direct repeat copy. Transient misalignment of the two repeats may then allow the deletion of the intervening bases before strand alignment is restored. The juxtaposition of two repeat copies as a consequence of the deletion serves to decrease the complexity of the sequence. For example, in the *CFTR* gene, a short deletion [ATTCTGTTCTcaGTTTTCCTGG] leads to the creation of a second copy of TTCTGTT with a concomitant decrease in complexity from 12 to 9 (complexity per base also decreases from 0.55 to 0.45).

Inverted repeats and secondary structure. Inverted repeats have also been implicated in the generation of micro-deletions and micro-insertions (Cooper, Krawczak 1993). An inverted repeat or palindrome comprises a series of bases that are complementary to another contiguous sequence upstream on the same DNA strand. By definition, therefore, an inverted repeat allows hairpin loop formation; excision repair of such a loop may yield a micro-deletion. Inverted repeats may also promote slipped mispairing of the nascent strand and subsequent duplication of downstream sequence. In the case of a micro-deletion in the *KIT* gene this led to a decrease in complexity:

H3 before deletion: T-C-T-GA-A-C-TCA-a-aGT-C-CTGAGTT, C3=12, c3=0.55;
H3 after deletion: T-C-T-GA-A-C-TCAG-TC-CTGAGTT, C3=9, c3=0.45.

In the case of a micro-insertion in the *F9* gene this led to an increase in complexity:

H3 before insertion:T-ATA-C-C-A-A-GGTAT-CC-CGG-TAT-TG-CAA,C3=12, c3=0.5;

H3 after insertion:T-ATA-C-C-A-A-GGTAT-CC-C-a-a-ggta-cc-a-a-GGTAT-TG-T-CAA,
C3= 20, c3=0.47.

A number of mechanisms that can account for both micro-deletions and micro-insertions in the vicinity of inverted repeats involve quasi-palindromic sequences (imperfect inverted repeats). Quasi-palindromes are thought to promote "strand-switching", or aberrant templating in the formation of the nascent strand.

Symmetric elements and Moebius loops. Symmetric elements, sequences that possess an axis of internal symmetry or which are symmetrical to another contiguous sequence on the same DNA strand upstream, have also been implicated in the generation of deletions and insertions, being thought to facilitate the formation of secondary structure intermediates. Krawczak and Cooper (1991) proposed that such an intermediate could be a Moebius loop-like structure formed after strand separation, twisting and re-annealing to the opposite strand in reverse orientation. Mismatched bases could loop out in such a structure thereby facilitating their own excision. An example of this process is provided by the 1bp micro-deletion in the *APC* gene responsible for the appearance of an 8bp self-symmetric fragment that decreases C4 complexity:

H4 before deletion: C-T-TC-A-TC-AC-A-g-A-AA-CA-G-TC-A-T, C4=15, c4=0.71;

H4 after deletion: C-T-TC-A-TC-AC-AAAACA-G-TC-A-T, C4=11, c4=0.55.

Alternatively, such a Moebius loop may partially resolve if one of the DNA strands disconnects and breaks. The repair of this region by a DNA polymerase would effectively result in the duplication of a sequence from the end of the symmetric element that initially broke off.

Inversions of inverted repeats and a novel mutational mechanism. Inversions of inverted repeats were found to be over-represented in the genomes of various organisms (Tchurikov, 1992). It was suggested that these elements may facilitate the formation of secondary structures, henceforth to be termed *knots*. Such pseudo-palindromic intermediates are structurally similar to hairpin-loops and may be implicated in the generation of micro-deletions and micro-insertions. Thus excision repair of such a loop may yield a micro-deletion whilst slipped mispairing of the nascent strand could lead to duplication of downstream sequence. For example, in a micro-deletion in the *BRCA1* gene, a mismatched base that looped out of the knot could have facilitated its own excision:

H5 before deletion: AAAATATTTGgGAAAACCTAT, C5=12, c5=0.57;

H5 after deletion: AAAATATTTGGAAAACCTAT, C5=10, c5=0.50.

The existence of this knot structure may help to resolve the ambiguity in deletion position that occurs as a consequence of the repetitive nature of the deletion-prone site. It is therefore likely that it was the third G that was deleted.

Repeats involved in insertion/deletion events and the possible path of mutation. Each indel may be regarded as having been the result of a two-step insertion/deletion process; the first step transforming the wild-type sequence to an intermediate, the second step transforming the intermediate to the final mutated sequence. There are essentially three possibilities. As it was shown above, the insertions increase the complexity while the deletions decrease it. This means that the complexity of a fragment remains more or less the same before and after the insertion/deletion event, i.e. if during the first step, a decrease in complexity is observed, then the second step must reverse this process leading to an increase in complexity and *vice versa*.

Discussion

The analysis demonstrated that changes in local DNA sequence complexity can be accounted for the involvement of a certain postulated mutational mechanism in micro-insertion and micro-deletion events and in indel formation; it can predict both the number and identity of the bases deleted and/or inserted. Proposed approach could also be applied to the analysis of gross rearrangements, which have so far been refractory to analysis. This is the short summary of a new approach; the detailed results will be reported elsewhere.

References

1. Antonarakis S.E., Krawczak M., Cooper D.N. (2001) The nature and mechanisms of human gene mutation. In Scriver C.R., Beaudet A.L., Sly W.S., Valle D. (eds). The Metabolic & Molecular Bases of Inherited disease, 8th ed. McGraw-Hill, New York, 259-291.
2. Cooper D.N., Krawczak M. (1993) Human Gene Mutation. BIOS Scientific, Oxford.
3. Gusev V.D., Nemytikova L.A., Chuzhanova N.A. (1999) On the complexity measures of genetic sequences. Bioinformatics. 15, 994-999.
4. Krawczak M., Cooper D.N. (1991) Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. Hum Genet. 86, 425-441.
5. Krawczak M., Chuzhanova N.A., Stenson P., Johansen B., Ball E., Cooper D.N. (2000) Changes in primary DNA sequence complexity influence the phenotypic consequences of mutations in human gene regulatory regions. Hum. Genet. 107, 362-365.
6. Krawczak M., Ball E., Fenton I., Stenson P.D., Abeysinghe S., Thomas N., Cooper D.N. (2000a) Human Gene Mutation Database - a biomedical information and research resource. Hum. Mutat. 15, 45-51.
7. Tchurikov N.A., Schyolkina A.K., Borisova O.F., Chernov B.K. (1992) Southern molecular hybridization experiments with parallel complementary DNA probes. FEBS Letts. 297, 233-236.

SUBWORDS GRAPHS, GENERATED BY GENETIC SEQUENCES

Evdokimov A.A., * Levin A.A.

Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia, e-mail: evdok@math.nsc.ru, levin@math.nsc.ru
*Corresponding author

Key words: *computer analysis, symbolic sequences, genetic sequences, combinatorial complexity, de Bruijn graph*

Resume

We suggest an approach to the analysis of combinatorial properties of symbolic sequences on the basis of their visualization on the de Bruijn graphs. A functioning program realized on the language JAVA is presented. It allows to study mathematical and genetic sequences of various types and to observe the dynamics of alterations of their structural „portraits".

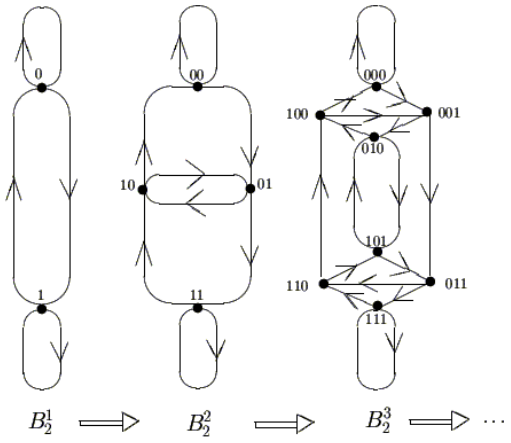
Availability: http://www.math.nsc.ru/LBRT/k3/Graph/Bruijn.html

Introduction

For many problems of combinatorics on symbolic sequences it is useful to study the correlation between the sequences properties and properties of various structures defined by subwords sets (fragments) of the sequence. Identifying such structures and studying process of their modification with growth of fragment length and increase of the sequence length can give valuable information concerning combinatorial and statistical properties of sequences. Comparing the structures with known structures and themselves helps to detect approaches to the description of separate sequences and their classes. In particular, the words overlaping graphs are examples of such structures, introduced by de Bruijn in 1946 and now named after him (Bruijn, 1946).

Methods and Algorithms

The node set in the de Bruijn graph B_m^n of dimension *n* consist from various words of length *n* in the alphabet of *m* symbols. Two nodes $\alpha = (\alpha_1,\ldots,\alpha_n)$ and $\beta = (\beta_1,\ldots,\beta_n)$ are connected by an arc oriented from α to β if and only if $\alpha_2 = \beta_1$, $\alpha_3 = \beta_2, \ldots, \alpha_n = \beta_{n-1}$ i.e., if the words α and β are overlapped on n-1 characters. The graph B_m^n has *m* loops in nodes corresponding to the constant-words consisting of a single symbol of the alphabet. It is connected and regular and such that indegree and outdegree of each node is equal to *m*. To draw the graphs sequence on the plane B_m^n for *n=1,2,3...* it is possible to construct them by induction on *n*. This procedure is based on the fact that B_m^{n+1} is the line graph of the graph B_m^n.



Any finite or infinite sequence $X=x_1,x_2,x_3,...$ of symbols in the *m*-alphabet corresponds to a path in the graph B_m^n that starts at the node $(x_1,...,x_n)$ and sequentially passes through the nodes $(x_i,...,x_{i+n-1})$ for $i=2,3,\ldots$ The subgraph of the graph B_m^n covered by this path is called the graph of *n*-subwords in the sequence *X* or the factor graph of dimension *n* and is denoted by $G^n(X)$. The set of nodes $V^n(X)$ of the graph $G^n(X)$ is a set of all *n*-subwords in *X*, and the set of arcs $E^n(X)$ is the set

all subwords of length *n+1* in *X*. A drawing of the supplement graph $B_m^n \setminus G^n(X)$ allows to observe the structure that is formed by those *n*-subwords that are absent in the sequence *X*.

The function $f(n,X) = |V^n(X)|$ is called the combinatorial complexity of a sequence *X*. It characterizes variety of a subword set. The function $f(n,X)$ is a nondecreasing function, which is strictly increasing for any infinite nonperiodic sequence and is satisfies $n+1 \le f(n) \le m^n$. It attains both extreme values on the sequences whose constructions are known (Evdokimov, 1983; Lothaire, 1983).

The experimental" study of interrelation of properties of a sequence and structure of the subwords graph depends on the ways drawing the graphs on the display. It leads to the problems of embedding graphs preserving certain structural properties on the plane - metrical, algebraic or combinatorial (Evdokimov, 2000); in particular, problems concerning embeddings of the graphs B_m^n and $G^n(X)$ on a plane preserving the distance of near nodes but the distances between distant nodes exceed some threshold. It is essential to take into account a symmetry of the graphs B_m^n, cyclic structures, various modes of the representation, effectiveness of an embedding with growth of dimension *n*.

For the graphs of DNA-sequences having large combinatorial complexity one has to use to „large block" representations of the graphs. We have found some embeddings of the graphs B_m^n on a plane for 2, 3 and 4-symbol alphabets which help to observe various properties of sequences.

Another direction of research is analysis of correlation between ways of generation of sequences from various classes and manifestation of their complexity and structural properties on a sequence of „portraits" $G^n(X)$, $n = 1,2,3,\ldots$. The developed code for visualization allows to carry out experiments with various classes of sequences:

- mathematical, defined by constructive procedures (recurrent; the DOL-sequences generated by iterations of substituting subwords for symbols etc. The DOL-sequence belongs to the class of L-systems defined by Lindenmayer to study models of the development of organisms in biology (Lindenmayer Aristid, 1975);
- pseudorandom, generated, for example, circuits of shift registers or random-number generators;
- precise but have errors in some positions;
- the genetic sequences (in experiments we use a database containing exons, introns, promoters, sites of binding matrix DNA and other sequences with known functional properties).

Results

The analysis of symbolic sequences from various classes has revealed a number of interesting and theoretically predicted properties corresponding to specifical features of structure of subwords set.

The sequence of images is an important characteristic of the structure of a symbolic sequence and allows to analyze dynamics of a modification of the subwords graphs $G^n(X)$, $n = 1,2,3,\ldots$. It should take into account the combinatorial complexity, the frequency of subwords , the cyclical structure of the sequence and contains the information on a local structure of the sequence and on properties of the sequence „as a whole".

Applications of the researches on visualization of symbolic sequences consist in the extensions of methods and toolkit for the analysis of structuredness of the texts both natural by origin, for example genetic (Woterman, 1999), or artificially generated.

For visualization of the subwords graphs and research of various sequences we develop the code VIZ. The code are implemented in JAVA and therefore can function on any computer with a virtual processor JAVA. The demo version of the code can be found on the server of Institute of mathematics: http://www.math.nsc.ru/LBRT/k3/Graph/Bruijn.html.

The program VIZ creates the graph B_m^n of overlapping of words (de Bruijn graph) for the following given parameters: a size *m* of the alphabet of the sequence and a length *n* of words (the dimension of the graph). The sequence *X* under consideration and its graph $G^n(X)$ is then represented on the graph. In the code it is possible to vary the image of structural portraits $G^n(X)$ on the screen. Changing parameters of the process can be caried out directly during observation.

The user can change a disposition of nodes of the graph on the screen, move the graph, and also change a size of the image of the graph. After pressing a control button the program reads out the next symbol (the whole word at the beginning of a sequence) and forms a new word from the previous one by adding that symbol to the end of previous word.

The node whose name coincides with the obtained word, and the arc connecting the previous node to this node is marked and is included in the passed chain, and their counters increase by 1. The passed chain is marked on the image and is called the „snake". The program allows to change the dimension of the de Bruijn graph where the studied sequence is placed on.

To increase the dimension of the graph we compute the line graph of the graph or of the passed part of the graph only. To decrease the dimension we return to the graph the line graph was constructed from.

If the combinatorial complexity of a sequence (the number of various subwords) is small, then it is possible to increase the dimension of the graph via multiple construction of the line graph for a passed part.

References

1. Bruijn de N.G. (1946) A combinatorial problem. Proc. Kon. Ned. Akad. v. Wet., 49(7), 758-764.
2. Evdokimov A.A. (1983) Complete sets of words and their numerical performances. In: methods of the discrete analysis in research of extreme structures. Novosibirsk, Institute of mathematics SD AS USSR. 39, 7-19.
3. Evdokimov A.A. (2000) Coding of the structured information and enclosures of discrete spaces. In: Discrete analysis and Operations research. A series 1. 7(4), 48-58.
4. Lindenmayer Aristid. (1975) Developmental systems and languages in their biological context. In: Herman G.T. and Rozenberg G. Developmental system and languages, North-Holland Publ.Co. 1975, 1-40.
5. Lothaire M. (1983) Combinatorics on words. Encyclopedia of mathematics and its applications. Addison - Wesley Publ. Company.
6. Woterman M.S. (1999) (ed.) Mathematical methods for analysis of DNA sequences.

INFORMATION CONCEPTION OF PERIODICITY OF SYMBOLIC TEXTS

Korotkov E.V., Korotkova M.A., Kudryashov N.A.

Center of Bioengineering RAS, Moscow, Russia
Moscow Physical Engineering Institute, Moscow, Russia
e-mail: katrin22@mtu-net.ru; kudr@dampe.mephi.ru
*Corresponding author

Key words: *latent periodicity, information decomposition, gene structure*

Resume

Motivation: The main purpose of the present work is to introduce the concept of information decomposition (ID) of a symbolical sequence that allows the finding of all available cases of periodicity in a sequence, as well as to develop an algorithm which will allow the calculation of ID for a symbolical sequence based on any possible alphabet.

Results:. We show the stability of the ID method in the case of a large number of random letter changes in an analyzed symbolic sequence. We demonstrate the efficiency of the method, analyzing both poems, and DNA and protein sequences. In poems of A.Pushkin and W.Shakespeare we found a latent periodicity of different lengths that can be reflections of periodicity of poem sounds. In DNA and protein sequences we show the existence of many DNA and amino acid sequences with different types and lengths of latent periodicity. All revealed sequences with latent periodicity have been accumulated in a newly created data bank, LPD (Latent Periodicity Database), which is systematically replenished according to the process of increase of Genbank and Swiss-Prot. The database contains more than $1.5x10^6$ sequences with various types of triplet periodicity, more than $2x10^6$ DNA sequences with period length from 2 up to 200 bases (without triplet periodicity) and more than 12,000 cases of latent periodicity of protein sequences. The possible origin of latent periodicity for different symbolical sequences is discussed.

Availability: http://bioinf.narod.ru/

Introduction

The development of mathematical methods for the study of symbolical sequence periodicity is given special significance nowadays. Earlier comprehensive mathematical methods were developed for the study of periodicity of continuous and discrete numerical and symbolical sequences, using Fourier transformation. However, such application of Fourier transformation demands presentation of a symbolical sequence as a numerical sequence in which the properties of any symbolical text should be displayed unequivocally. In our opinion the given method only works well for the study of periodicity of symbolical sequences with a relatively short length (which is smaller than the size of the symbolical sequence alphabet). For periods with a length greater than the size of the symbolical sequence alphabet, there is the possibility of "decomposition" of the statistical importance of the longer periods in favor of the shorter ones. We shall explain this with the following example. Let a symbolical sequence be given with the period YRTDFT repeated 50 times. For this sequence we have 5 numerical sequences consisting of the numbers 0 and 1 (according to the alphabet used). In this case, for the letters Y, R, D, and F the Fourier-harmonics show the length of the symbolical sequence period equal to 6 symbols, but for the letter T the period equal to 3 letters is found. This reduces the statistical importance of the 6-letter period by the value of the statistical importance of the 3-letter period. This effect will increase with the growth of the relation of period length to the size of the alphabet used. Thus, it turns out that the statistical importance of the longer period is a kind of "spread" onto the statistical importance of the shorter periods, i.e. there is an effect of attenuation of harmonics with longer periods in favor of harmonics with shorter periods. This effect will be even stronger for cases where there are several replacements in periodic sequences - in such sequences periods could not be simply identical. In this study, we are developing the Information decomposition of symbolical sequences (ID) and apply ID method for analysis of poetic texts, DNA sequences and amino acid sequences. The ID concept allowed us to discover latent periodicity in many poetic texts, genes and various proteins. We also discuss the possible origin of latent periodicity in symbolical sequences of various origin.

Model

Under ID, we understand a spectrum representing the statistical importance of mutual information for periods of various lengths in the analyzed symbolical sequence. Mutual information between the investigated symbolical sequence and artificial symbolical periodic sequences can be used for obtaining an ID spectrum. Let us take a symbolical sequence a(i), i = 1,2, … n. We generate a set of artificial symbolical sequences with periods from 2 up to n/2 of the same length as well as an initial symbolical sequence. We use numbers as symbols of artificial symbolical sequences. The artificial symbolical sequence with period length equal to 2 symbols can be presented as follows: 1,2,1,2,1,2,1,2,1,2,1,2,1,2,…; the sequence with period length equal to 3 symbols can be presented as: 1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,…; the sequence with period length

equal to k symbols can be presented as: 1,2,…k, 1,2,…,k, 1,2…,k…. Further, we can determine the mutual information between the analyzed sequence a(i) and each of the artificial periodic sequences. Values of the mutual information define the ID spectrum for the analyzed symbolical sequence. We fill a matrix M' with dimensions (nxk) for the value of the mutual information, where **n** shows the period length of the artificial periodic sequence used, and **k** is the size of the alphabet of the analyzed symbolical sequence. The value of the mutual information is calculated as follows:

$$I = \sum_{1}^{n}\sum_{1}^{k} m(i,j)\ln m(i,j) - \sum_{1}^{n} x(i)\ln x(i) - \sum_{1}^{k} y(j)\ln y(j) + L\ln L \qquad (1)$$

where matrix M' shows the numbers of coincidences of ij (i=1,2..,n; j=1,2…n) type between compared sequences (L is the length of the analyzed symbolical sequence, x(i), i=1,2, …, n are the frequencies of symbols 1,2, …, n in the artificial periodic symbolical sequence; y(j), j=1,2, …, k are the frequencies of symbols in the analyzed symbolical sequence).

One of the properties of mutual information is its orthogonality. This means that I(a, b)=0, in conditions where a and b are symbolical periodic sequences with period lengths representing prime numbers. Another important property of an information spectrum is the nesting of mutual information for various periods one into another [1, 2]. This means that mutual information I(n) for composite period $n = n_1 \times n_2 \times n_3 \ldots \times n_t$ (n_1, n_2, n_3,…,n_t – prime numbers) is equal to:

$$I(n)=I(n_1)+I(n_2)+I(n_3)+\ldots+I(n_t)+I'(n) \qquad (2)$$

where $I(n_1)$, $I(n_2)$, $I(n_3)$,…,$I(n_t)$ are the values of mutual information between artificial periodic sequences with lengths n_1,n_2,n_3,…,n_t and the analyzed symbolical sequence; I'(n) - is "pure" mutual information between the artificial periodic sequence with period length equal to **n** and the analyzed symbolical sequence, which could not arise from periodicity with period lengths equal to n_1, n_2, n_3, …, n_t. The formula (2) is easily deduced from the correlation of mutual information for three sequences [6], taking into account that the mutual information between artificial periodic sequences is equal to zero, if the period lengths of them represents various prime numbers.

Formula (2) shows that short ID periods don't provide the effect of "damping" of longer period statistical significance. On the contrary, ID allows accumulation of mutual information of "prime periods" (period lengths are given by various prime numbers) in "composite periods". This ID property is attractive for the detection of long latent periodicity (where the period length value exceeds the value of the alphabet size of the analyzed symbolical sequence). For ID construction it is necessary to take into account that the value 2I(n,k) is distributed as χ^2 with the value of degree of freedom equal to (n-1)(k-1). The average value of a mutual information of two random symbolical sequences with alphabet size **n** and **k**, is equal to (n-1)(k-1) correspondingly. This means that a I(n,k) dependence from **n** at the constant **k** has the linear component equal to (k-1)(n-1). Therefore, it is more convenient to show on the diagram the dependence J(n,k)=I(n,k)-(k-1)(n-1), which should resemble a diagram of the Fourier transformation. Such a relation is shown in our previous publications [1, 2]. This relation is similar to Fourier harmonics, but only for relatively short periods. For longer periods, the mutual information determined by the formula (1) begins to deviate from the χ^2 distribution if each element of the matrix M becomes less than 10. This deviation from the χ^2 distribution results in an increase of the average value of J(n,k) with an increasing **n** value. Two approaches can be used for taking into account such a deviation of I(n,k) from the χ^2 distribution in conditions of small sample statistics. The first approach allows direct calculation of the probability of the fact that the relation of symbols in the artificial periodic sequence and in the analyzed symbolical sequence is caused by random factors only, instead of calculation of mutual information. The second approach is based on a Monte-Carlo method for the estimation of the statistical importance of J(n, k) by means of the value Z(n,k) calculation [20, 21, 22]:

$$Z(n,k) = \{J(n,k) - \overline{J(n,k)}\} / \sqrt{D(J(n,k))} \qquad (3)$$

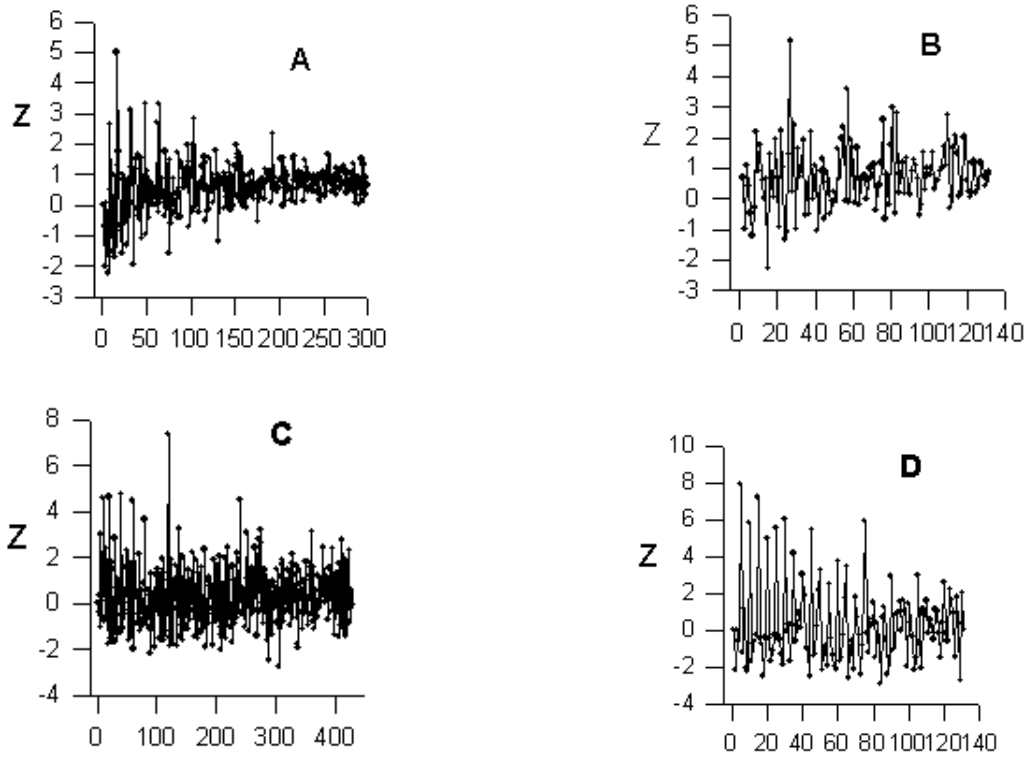
where $\overline{J(n,k)}$ and D(J(n,k)) show the average value and deviation of the J(n,k) value, for a set of random matrixes with the same sums x(i) and y(j) as in the initial matrix M(n,k). The results of the study of periodicity in various symbolical sequences are presented below. In our study, we used a Monte-Carlo method that permits the execution of the calculations relatively quickly [3,4,5]. Information decomposition of a symbolical sequence we shall represent as a spectrum Z(n,k).

The spectrum Z(n,k) is similar to a spectrum of Fourier transformation for numerical sequences, but has the following advantages: 1. The calculation of the spectrum does not require any transformation of a symbolical sequence to numerical sequences; 2. ID allows the revealing of both the obvious periodicity and the latent periodicity of a symbolical sequence in which there is no statistically important similarity between any two periods; 3. The statistical importance of long periods is not spread onto the statistical importance of shorter periods; 4. On the basis of the matrix M it is possible to determine the type of periodicity. The proposed approach also allows us to refrain from using a fixed size window when searching for periodicity in a symbolical sequence of large size. For this purpose, we search in the analyzed window for the sub-sequence having the maximum Z value for every tested period length. This approach allows us to realize a segmentation of the analyzed symbolical sequence, depending on the presence of periodicity in various sites of the sequence. Using this approach, periodicity with various period lengths and of various, types can be revealed [1, 2, 3, 4, 5].

Results and Discussion

We shows some examples our results of application of ID method. It seems quite natural to expect that poetic texts will have some degree of periodicity, since such periodicity is quite evident at a perusal of poetic texts. At first we analyzed several poems of the classic Russian poet, A.S.Pushkin.

The ID of the A.S. Pushkin poem "I remember a wonderful moment…" is shown in Fig. A. The latent periodicity was also found in other poems of A.S. Pushkin and in poems of many other authors. In Fig. B the latent periodicity of fragments in poems W.Shakespeare "A Midsummer-Night's Dream" is shown. On the whole, the results obtained from the present study demonstrated that ID is capable of revealing the structure of poetic texts, most probably caused by the sound periodicity of poems.



Examples of DNA sequences periodicity shown in the Fig. C and D. Fig. C shows latent periodicity of Deinococcus radiodurans gene for c-di-GMP phosphodiesterase (2867-5239 base pairs) from sequence AE002006. DNA sequence from 3108 to 3963 bases has the latent period equal to 120 bases and Z(120,4) equal to 9,1. Fig. D shows latent periodicity of the gene coding region of the high-sulphur wool matrix protein B2A from sheep (73-561 base pairs) from SHPWMPBB. DNA sequence from 373 to 634 bases has the latent period equal to 5 bases and Z(5,4) equal to 7,5.

The developed method of information decomposition (ID) of symbolical sequences proved to be capable of revealing the latent structure of various symbolical sequences. The method intended for detection of periodicity appeared to be rather tolerant to a relatively large number of symbol replacements. The results obtained using this method demonstrated that a large number of known genetic texts contain sequences with latent periodicity of various lengths and various types, which could not have been revealed earlier. The origin of latent periodicity in genetic texts might be connected both with the evolution of the genome and protein molecules, and with the functional meaning of various sequences. Periodicity equal to 21 bases is usually connected with α-helix formation protein molecules. The longer periodicity could be determined by domain organization formation in proteins; it could also be involved in the process of nucleosome binding with DNA. However, we observed a great variety of period lengths and types in DNA and amino acid sequences. It testifies that some other protein spatial structures have characteristic periods of their own [5]. It is possible to assume that the ID method is able to "see" certain structural characteristics of gene sequences, reflecting spatial organization of the corresponding proteins. In this regard, ID is obviously an important method, allowing the connection of the origin of certain protein structures with the presence of certain latent periodicity in corresponding DNA and amino acid sequences**.**

All revealed sequences with latent periodicity have been accumulated in a new data bank, named LPD (Latent Periodicity Database). At present the volume of the database developed exceeds 20 Gb. The database contains more than 10^6 sequences with various types of triplet periodicity, more than $2x10^6$ DNA sequences with period length from 2 up to 200 bases (without triplet periodicity) and more than 12,000 cases of latent periodicity of protein sequences.

The developed ID method, as well as the methods based on Fourier transformation, are not able to detect latent periodicity with deletions or insertions of symbols. We think that this often leads to the impossibility of revealing all cases of latent periodicity in symbolical texts. However, deletions and insertions are mutation events, frequently met in genetic texts. It permits us to assume that the number of sequences with latent periodicity in genetic texts is actually considerably greater than can be revealed with the ID method at present. We are currently improving the ID method, using some of the methods of profile analysis and dynamic programming. In the near future the LPD database will be completed with a large number of genetic texts with latent periodicity, obtained in the presence of deletions and insertions of symbols

Acknowledgements

The work was supported by the grant № 1379 of the International Science and Technology Center (ISTC).

References

1. Korotkov E.V., Korotkova M.A. (1995) DNA regions with latent periodicity in some human clones. DNA Sequence. 5**,** 353-358.
2. Korotkov E.V., Korotkova M.A., Tulko J.S. (1997) Latent sequence periodicity of some oncogenes and DNA-binding protein genes. CABIOS. 13**,** 37-44.
3. Korotkov E.V., Korotkova M.A., Rudenko V.M., Skryabin K.G. (1999) Regions with the latent periodicity in the amino acid sequences of the many proteins. Mol. Biol. (Russian). 33, 1-8.
4. Chaley M.B., Korotkov E.V., Skryabin K.G. (1999) Method reavealing latent periodicity of the nucleotide sequences modified for a case of small samples. DNA Res. 6, 153-163.
5. Korotkova M.A., Korotkov E.V., Rudenko V.M. Latent periodicity of protein sequences**.** J. of Mol. Modelling. 5**,** 103-115.
6. Yaglom A.M., Yaglom I.M. (1960) Probability and Information. M.: Nauka press.

A CHARACTERISTIC TYPE OF LATENT PERIODICITY OF 21 BPS FOUND IN BACTERIAL GENES OF THE TRANSMEMBRANE CHEMORECEPTORS (MCP II)

[1*] Chaley M.B, [1] Korotkov E.V., [2] Kudryashov N.A.

[1] Center "Bioengineering" RAS, 117312, Moscow, Russia
[2] Moscow Physical Engineering Institute, 115409, Moscow, Russia
e-mail: mariam@biengi.ac.ru

Key words: *chemoreceptor, transmembrane receptor, bacterial genome, periodicity*

Resume

Motivation: A search of periodicity in the known genomes can provide a revealing of various structural and functional determinants, which further may be useful for more stringent functional assignment and identification of ORFs, and for determination of regulatory and DNA conformational sites. Here we present an effort to identify the genes of bacterial transmembrane chemoreceptors (MCP II) with a characteristic periodicity of 21 base pairs revealed in the Tar chemoreceptor of *E. coli.*

Results

Among all found bacterial genes, having the characteristic periodicity of 21 bps, the genes of transmembrane chemoreceptors have made the much numerical group. However, this periodicity cannot be considered as an exclusive determinant of such chemoreceptors. As one can see it should at first study what impact of probable α-helix structure is present in such a periodicity and what is typical only for the tansmembrane regions of the chemoreceptors.

Introduction

An earlier proposed method for search of the latent periodicity in DNA and amino acid sequences (Korotkov, Korotkova, 1995; Korotkov et al., 1999) did not allow to reveal periodicity if indels were occurred in the sequences. By this reason at least 50% of the latent periodicity is still being unrevealed. Nevertheless this problem may be solved if to search for the latent periodicity of a known kind, determined with weight position-specific matrix (a profile), in applying a dynamic programming method to alignment of the periodically repeated profile with an analyzed nucleotide or amino acid sequence. A particular interest is to search the latent periodicity, which serves as a functional or structural discriminator of gene or protein. We have found earlier that in some bacterial chemoreceptor genes the regions corresponding to the transmembrane domains had periodicity of 21 base pairs (Chaley et al., 1999). A chemoreceptor gene's periodicity of 21 bps may be quite correspond to α-helix structures of the transmembrane domains, because it is a conformation of α-helixes for the aspartat receptor that has been determined by X-ray method (Milburn et al., 1991); and besides this the latent period of 21 bps provides a potential basis for its translation into a coiled-coiled α-helix protein structure (Trifonov, 1998). In this particular investigation we have intended to find out whether the revealed 21 bps periodicity in the bacterial chemoreceptor genes is due to probable α-helix protein structure or it may serve as the characteristic discriminator for the chemoreceptors.

Methods

A sequence of an earlier revealed 21 bps periodicity's region in the Tar chemoreceptor gene of *E. coli* has been served for the periodicity profile creation. The Tar chemoreceptor has been chosen not casually, because its whole 2D structure, consisting of α-helixes, is well known (Mowbray, Sandgren, 1998). The profile has been periodically repeated up to a size of scanning window equal to 252 bases. The size was limited by the possibilities of a dynamic profile alignment program, which has been already described (Korotkov et al., 2000). Having such a window size setting, the program occupies at least 40 Mb of PC operation memory, so the window elongation would need more memory extension accompanied with a nonlinear decrease in performance. The dynamic profile alignment approach used in the program is an application of dynamic programming method like Smith and Waterman algorithm (1981) to alignment of common symbolic string with a profile. The original periodicity profile was calculated as a position-specific matrix denoting the weight of each symbol (A, T, C, G) in a fixed period position as

*Corresponding author.

$$w(i,j) = f(i,j)\ln\{f(i,j)/p(i)\} \quad (1)$$

Here i is a kind of symbol, j denotes a position in period, f(i,j) shows the frequency of the symbol i in position j, p(i) is a frequency of the symbol i over all periodic sequence of the Tar chemoreceptor gene. Such a profile was repeated 12 times to modulate latent periodicity region of 21 bases at the length of scanning window. So, this new periodic profile has been used in the search for statistical significant alignments in the bacterial genomes from the GenBank-127. Total weight W of an alignment was calculated as a sum of appropriate symbols' weights w(i,j) over the aligned positions. Statistical significance was appreciated for each alignment by the Monte Carlo method. A generation of a random sequence with the same frequencies and triplet correlation of the bases as in a bacterial sequence from the GenBank allowed us to calculate mean weight W_m and variance D(W). The random sequence exceeded the total size of the bacterial DNA sequences in the GenBank in about 20 times. We have used the formula (2) to estimate significance of tested alignment.

$$Z = \frac{W - W_m}{\sqrt{D(W)}}. \quad (2)$$

No alignments were found in the random sequences with Z>7.0. So, we assumed the alignments to be significant if their Z-values were greater than 7.0.

The dynamic profile alignment program (Korotkov et al., 2000) allows one to search for a local alignment of maximal weight within the size of scanning window. Further, we have filtered the significant results excluding the overlapping alignments, if their cross area exceeded 30%, leaving the alignment with the greater Z-value.

Results

An analysis of the statistically significant alignments for about 600 bacterial genes with known functions, in which the latent periodicity of 21 bps of the same kind was found, has shown that transmembrane chemoreceptors constitute the much numerical group among the genes (18.3% of total findings having the average length of alignment equal to 170 bps). It is interesting that in some genes of regulatory proteins, transferases, dehydrogenases, and etc. the same kind of latent periodicity has been found with length about 100 bps. The revealed characteristic periodicity of 21 bps in the 16S ribosomal RNAs at length about 40 bps appears unexpected enough (13% of total findings). Table 1 provides a review of the alignments' figures in various functional protein groups. The groups, counting less than 1% of totally found similarities to the characteristic periodicity; have been incorporated into a single group.

Table 1. A distribution of the figures of statistical significant alignments is shown through the general protein groups.

General protein group	Group's contribution to the total number of findings with characteristic periodicity of 21 bps
Proteases	1%
Nucleases	1%
Phototaxis transducers	1%
Synthases	1.3%
Hydrolases	1.4%
Recombination and DNA repair proteins	1.5%
DNA gyrases	2.2%
Synthetases	2.4%
Kinases	2.6%
Dehydrogenases	2.7%
Reductases	3%
GTP/ATP-binding proteins	3%
DNA polymerases	3.6%
Regulatory proteins	4.3%
Transferases	4.3%
16S ribosomal RNAs	13%
Chemotaxis transmembrane receptors	18.3%
Other groups (including <<1% of the total findings)	33.4%

Discussion

As one can conclude from the content of Table 1, the kind of 21 bps periodicity used in this investigation preferably reveals the genes of transmembrane chemoreceptors, though a numerical pick of similarities to the investigated periodicity of 21 bps in the 16S ribosomal RNAs needs an additional study. In the case of the rest protein groups, one may suppose that the revealed periodicity is probably due to their α-helix regions. In general, more detailed analysis of all revealed similarities to the characteristic periodicity of 21 bps should be done before it will become possible to appreciate the characteristic value of the studied periodicity.

References

1. Chaley M.B., Korotkov E.V., Skryabin K.G. (1999) Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples. DNA Res. 6, 153-163.
2. Korotkov E.V., Korotkova M.A. (1995) Latent periodicity of DNA sequences from some human gene regions. DNA Sequence. 5, 353-358.
3. Korotkov E.V., Korotkova M.A., Rudenko V.M. (2000) MIR-family of repeats common for vertebrate genomes. Mol. Biol. (Mosk). 34(4), 553-559.
4. Korotkova M.A., Korotkov E.V., Rudenko V.M. (1999) Latent periodicity of protein sequences. J. Mol. Model. 5 103-115.
5. Milburn M.V., Prive G.G. et al. (1991) Three-dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand. Science. 254(5036), 1342-1347.
6. Mowbray S.L., Sandgren M.O.J. (1998) Chemotaxis receptors: a progress report on structure and function. Struct Biol. 124, 257-275.
7. Smith T.F., Waterman M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol. 147, 195-197.
8. Trifonov E.N. (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. Physica A. 249, 511-516.

PERICENTROMERIC ALPHA SATELLITES: NON-RANDOM DISTRIBUTION OF STRUCTURAL REARRANGEMENTS AND INSERTIONS OF DISPERSED ELEMENTS ALONG THE MONOMER

**** Oparina N.J., Lacroix M.-H., Mashkova T.D.***

Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia, e-mail: nixie@eimb.ru
*Corresponding author

Key words: *alpha satellite DNA, centromere, heterochromatin, recombination, dispersed elements*

Resume

Motivation: Pericentromeric regions of human chromosomes is the favorite place for many processes, leading to chromosomal rearrangements. The predominant class of centromeric satellites - alpha satellite (alphoid) DNA - is a good model for studying the traces of former recombinational, retropositional and other types of events.

Results: We have screened nucleotide sequences, containing alphoid/non-alphoid junctions from current databases. All sequences were analyzed, and various rearrangements disrupting the regular tandem structure were found. What types of such the events we can distinguish? Structural rearrangements (deletions, duplications, inversions, expansions of short oligonucleotide motifs) and insertions of different dispersed elements (mostly L1). We have revealed the non-random distribution of events sites along alphoid monomer, as well as their preferential occurrence at kinkable DNA sites.

Availability: Supplementary material on this study is available through the URL http://www.imb.ac.ru/nixie/.

Introduction

Centromeres of all primates contain large arrays of specific tandem repeats. The predominant class of those repeats, so-called alpha satellites or alphoid DNAs, is characterized by tandemly repeated 171-bp monomers. Slightly different monomers copies are usually organized into higher-order repeats units. Different types of alphoid DNAs, showing dimeric, pentameric or single-monomeric organization, are described. "Classic" alpha satellites, located to the inner centromere regions, demonstrate dimeric organization, very low degree of inter-monomer divergence, and the presence of the CENP-B (centromeric protein B) binding site (reviewed by Jorgensen, 1997; Lee et al., 1997). Centromeric regions, studied on human chromosomes 7 and 21 consist of two non-overlapping arrays - loci α-I and α-II (Ikeno et al., 1994). Alpha satellites of α-II locus are known to lack higher order repeat organization and to consist from single -monomeric alphoid family, devoid of CENP-B boxes. It is also known, that α-II locus is the place of concentrations of different sequence alterations in alphoid tandems (Mashkova et al., 1998). One type of such alterations, the dispersed elements (*Alu* and L1) insertions, was reported many times (Jackson et al., 1992; Laurent et al., 1999).

The sequence specificity of L1-encoded endonuclease EN is still the hot field of study. Most of the insertions sites identified have a purine or a short run of purines immediately 3' and a run of pyrimidines immediately 5' to the insertion point. Short interspersed repeats (SINEs), with the most frequent human representative - *Alu* element, do not encode their own reverse transcriptase of endonuclease. It has been suggested that SINEs can use the active LINE integration machinery (Jurka, 1997). Further analysis of mammalian retroposons integration sites showed their tendency to integrate at kinkable DNA sites, rich in TA or CA/TG dinucleotides (Jurka et al., 1998).

Here we study integration sites of different found retroposons in alpha satellite arrays, as well as structural rearrangments in alphoid tandems.

Methods

For large-scale screening of nucleotide databases (GENBANK and EMBL, including HTG, GSS, EST and STS subdivisions) we used both BLASTN and FASTA programs. We searched for sequences homologous to M consensus alphoid monomer, A-type (lacking CENP-B box) and B-type (containing CENP-B box) consensus alphoid monomers. Also we generated listfiles of sequences, homologous to L1, *Alu* consensus sequences, deca-satellite, satellites II and III and other human repetitive elements (RepBase Update, available at http://www.girinst.org). All resulted listfiles were compared to alphoid-homologous sequences listfiles. Sequences, showed homology both to alpha satellites and to any other repeats, were used for further analysis. All numerous alphoid sequences adjacent to non-alphoid DNA were subdivided into monomers. Base position 1 was assigned according to *Bam*HI cleavage sites in the X-chromosome specific alphoid repeat (Waye, Willard, 1985). All alpha satellites were aligned, and for sequences, contained more than 10 monomers, their own consensi (with the 51% of nucleotide frequency threshold) were designed.

Results and Discussion

The picked-up bordered alphoid sequences were divided into monomers and aligned. The detailed analysis of more than 1100 alphoid monomers (~200 kbp) revealed 47 cases of internal deletions in monomers, 12 inversions, 8 duplications of short alphoid sequences, 17 cases of short oligonucleotides expansions and numerous joinings with interspersed repeats of other satellites. All rearrangement and insertion events were analyzed according to their flanking sequences and their positions distribution along the consensus alphoid monomer. Detailed sequence comparison showed that most of these points coincided with TG, CA or TA dinucleotides (see the example on Fig. 1, showing flanks of some deletions):

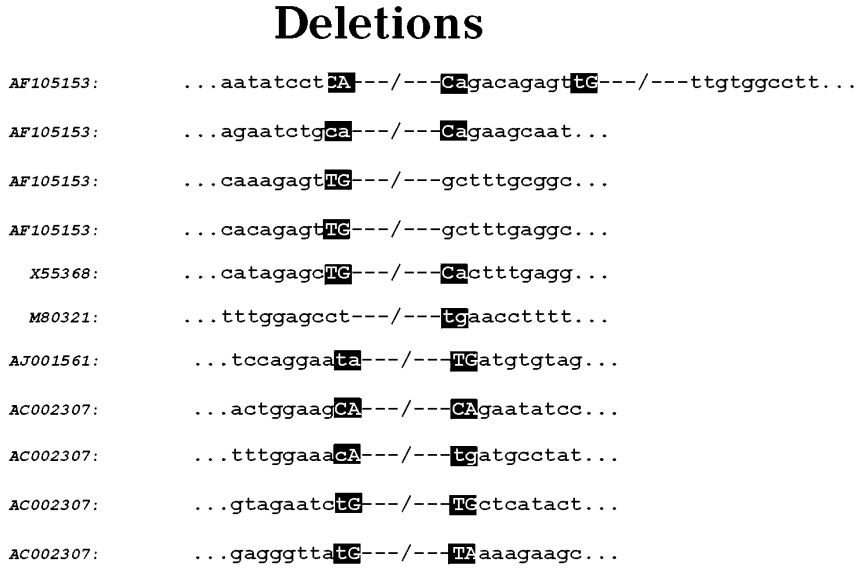


Fig. 1. Examples of deletions-type rearrangements in alphoid DNA. Nucleotides immediately adjacent to deletion, and lost during these events, but presented in intact alphoid consensus monomer are shown in uppercase. Border kinkable dinucleotides are shaded.

In order to verify the non-randomness of this coincidence we constructed the 17-kbp (~100 monomeric) test sequence composed from fragments of all bordered sequences in proportion of their alphoid parts length. The observed frequencies of all 16 dinucleotides in the test sequence were counted and compared to their expected frequencies. Observed TA frequency was 1.95 times lower that expected, TG and CA frequencies were 1.42 and 1.26 times higher than expected, respectively. The above-mentioned dinucleotides were not predominant in studied alphoids.

By the example of TA dinucleotide the possible ways of alterations events points (shown as ↓) coincidence with the particular dinucleotide could be represented as NNN↓T↓A↓NNN, which corresponded with 4 bp sliding window. There were three possible locations of the TA dinucleotide in the 4 bp sliding window: T↓A↓N*N, N↓T↓A↓N, N*N↓T↓A. Probable locations of the alterations events points are shown as ↓ or * (9 variants). Only 7 of all alterations events points (shown as ↓) could be considered as coinciding with TA dinucleotide. Consequently, the probability of random coincidence of alterations points with all three kinkable dinucleotides (TA, CA or TG) could be calculated as $P_{TG} + P_{CA} + P_{TA}$ = F(TG) x 3 x 7/9 + F(CA) x 3 x 7/9 + F(TA) x 3 x 7/9 = 0.48197 (~48%) (where dinucleotide frequencies corresponded to the test sequence). In contrast to expected 48% probability, we observed the 92% frequency of TG, CA or TA location at the border of all analyzed deletions breakpoints, 90% of inversion breakpoints and 100% of expansions. Most of insertions points (96% L1 and 93% *Alu* insertions) also coincided with kinkable dinucleotides. In comparison to our date, 1-3 bp deletions in alphoid consensus monomers J2 (AC AJ130754) and W1 (AC AJ130755) occurred at TG and TA dinucleotides. The sequence rearrangements induced by homologous recombination often involve the Holliday structure creation and cleavage by specific endonucleases. Some of them are known to cleave the DNA at 5'-AC<u>T</u>↓<u>A</u> (for *S.cerevisiae* CCE1 endonuclease (Schofield et al., 1998) or 5'-WTT↓<u>CA</u> > 5'-WT<u>T</u>↓<u>G</u>A ~ 5'-WT<u>T</u>↓<u>A</u>A (for *E. coli* RuvC endonuclease (Fogg et al., 1999). Kinkable dinucleotides are not evenly distributed along alphoid monomer and form clusters (Fig 2b). But the non-randomness of structural alteration point distribution along alphoid monomer (Fig. 2a) could not be explained by dinucleotides distributions only and is the field of interest for further investigations.

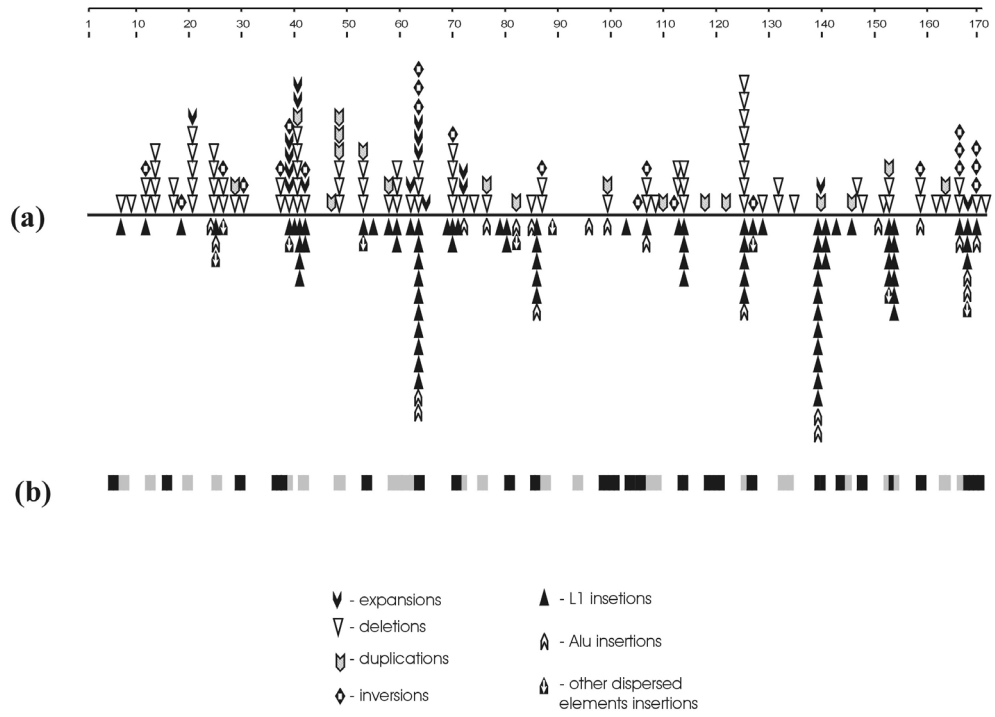


Fig. 2. Location of structural rearrangements (above the line) and insertional events (below the line) along the consensus alphoid monomer **(a)**. Distribution of the kinkable dinucleotides in alphoid consensi is also shown **(b)**. Conserved dinucleotides are black-shaded and variable - gray-shaded.

References

1. Fogg J.M., Schofield M.J., White M.F., Lilley D.M.J. (1999) Sequence and functional-group specificity for cleavage of DNA junctions by RuvC of *Escherichia coli*. Biochemistry. 38, 11349-11358.
2. Ikeno M., Masumoto H., Okazaki T. (1994) Distribution of CENP-B boxes reflected in CREST centromere antigenic sites on long-range α-satellite DNA arrays of human chromosome 21. Human Mol. Genet. 3, 1245-1257.
3. Jackson A.S., Mole S.E., Ponder B.A.J. (1992) Characterization of a boundary between satellite III and alphoid sequences on human chromosome 10. Nucl. Acids Res. 20, 4781-4787.
4. Jorgensen A.L. (1997) Alphoid repetitive DNA in human chromosomes. Danish Med. Bull. 44, 522-534.
5. Jurka J. (1997) Sequence patterns indicates an enzymatic involvement in integration of mammalian retroposons. Proc. Natl Acad. Sci. USA. 94, 1872-1877.
6. Jurka J., Klonowski P., Trifonov E. (1998) Mammalian retroposons integrate at kinkable DNA sites. J. Biomol. Struct. Dynam. 15, 717-721.
7. Laurent A.-M., Puecheberty, Roizes G. (1999) Hypothesis; for the worst and for the best, L1Hs retroposons actively participate in the evolution of the human centromeric alphoid sequences. Chromosome Res. 7, 305-317.
8. Lee C., Wevrick R., Fisher R.B., Ferguson-Smith M.A., Lin C.C. (1997) Human centromeric DNAs. Human Genet. 100, 291-304.
9. Mashkova T.D., Oparina N.Yu., Alexandrov I.A., Zinovieva O.L., Marusina A.I., Yurov Y.B., Lacroix M.-H., Kisselev L.L. (1998) Unequal crossing-over is involved in human alpha satellite DNA rearrangements on a border of satellite domain. FEBS Letters. 441, 451-457.
10. Schofield M.J., Lilley D.M.J., White M.F. (1998) Dissection of the sequence specificity of the Holliday junction endonuclease CCE1. Biochemistry. 37, 7733-7740.
11. Waye J.S., Willard H.F. (1985) Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobase-pair repeat from the human X chromosome. Nucl. Acids Res. 13, 2731-2743.

CLUSTERS OF LONG TERMINAL REPEATS OF HUMAN ENDOGENOUS RETROVIRUSES (K-FAMILY)

**** Artamonova I.I., Gorodentseva T.N., Sverdlov E.D.***

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, 117871, Russia, e-mail: irena@humgen.siobc.ras.ru
*Corresponding author

Key words: *human endogenous retroviruses-K (HERV-K); HML-2; long terminal repeat (LTR)*

Resume

Motivation: More than 8% of the human genome consists of the endogenous retroviruses sequences and their fragments. So the whole genome analysis of HERV distribution could reveal their functional role in the genome and peculiarities of the mechanism of retrotransposition.

Results: Here we describe the distribution of the human endogenous retroviruses HERV-K (HML-2) long terminal repeats in the human genome and announce the existence of LTR clusters within separate chromosomes.

Introduction

About 40% of human DNA sequences are derived from different types of transposable elements and more than 8% of them are endogenous retroviruses sequences and their fragments (IHGSC, 2001), in a majority they are results of past retroviral infections of the germline cells of primates and following spreading in the genome by means of retrotransposition. Currently HERVs comprise about 200 distinct groups (Jurka, 2000).

The structure of endogenous retroviruses is the key to the interest in their evolution and functional role in the genome because they include many regulator elements such as promoters, enhancers, poly-A sites, hormone responsive elements etc.

One of the most widespread and functionally kept families of endogenous retroviruses, HERV-K (HML-2) family, is represented in the genome not only by retroviruse-like sequences but also by much more numerous solitary long terminal repeats. The main goal of this investigation is evolutionary analysis of HERV-K LTRs distribution.

Methods and Algorithms

Database compiling. The preliminary selection of LTR sequences was formed based on GenBank Release 122.0 using our own software, BLASTn (available at http://www.ncbi.nlm.nih.gov/BLAST) and RepeatMasker2 (available at http://ftp.genome.washington.edu). Further supplements and specifications are based on Human Genome Browser (available at http://genome.ucsc.edu) using their special tools.

LTR classification. We suppose that retroviruses spreading over the genome was not uniform but it was the result of some separate events. So the aim of LTR classification is to cluster HERV-K LTRs into groups of independently mutated copies.

The algorithm of dividing was the following. We began by taking all the known LTR sequences from GenBank. Multiple alignment was built by ClustalW (version 1.7, available at http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html), then we edited this alignment by our own software for editing of multiple alignments. A matrix of numbers of pair-wise differences (excluding mutations of CpG-dinucleotides) was built by the same software and written to an output file in a format that can be read by the distance matrix phylogeny programs FITCH and KITCH from Phyllip software package (version 3.5c, Felsenstein, 1993). Then phylogenetic trees were constructed by Fitch program using different orders of LTR copies in the input data and visualized using TreeView1.6 (http://taxonomy.zoology.gla.ac.uk/rod/rod.html). This tree was clustered into 2 groups of sequences, then the procedure was repeated for sequences from these groups. In such a way LTRs were clustered into 13 groups containing sequences that are supposed to have accumulated mutations randomly and independently. Trees of these groups look like "bushes". Then consensuses based on multiple alignments of these clusters were built that allowed us to attribute every new sequence to certain family according to the closest consensus.

Results and Discussion

The sequences under consideration are widely spread in the genome. We estimate the total amount of HERV-K LTR sequences as 1500 copies per haploid genome.

Distribution of LTR sequences in individual chromosome was studied. It was proved to be irregular with respect to chromosome lengths. The density of LTR location varies significantly for different chromosomes. So all human chromosomes may be divided into 2 classes with relatively high or low LTR content.

Next point was the investigation of the relative location of LTR sequences within every chromosome. About one half of all known sequences form clusters with the level of local LTR density more than 5 times higher than that for the whole chromosome. Almost every chromosome contains 1-5 LTR clusters. However, there is no direct correlation between the proximity of LTRs in the cluster and their structural divergence (for chromosomes 19, 21 and 22, see Lavrentieva et al., 1998; Kurdyukov et al., 2001; Artamonova et al., 2000, respectively). This is the ground for a hypothesis that successive infections were independent but human chromosomes contain regions preferable for retrotransposition.

So it should be interesting to bring out correlation between chromosomal distribution of LTR sequences and local genome features such as average GC-content, densities of genes, other repeats and single nucleotide polymorphisms. We haven't found any correlation between LTR clusters positions and local GC-content. But chromosome segments containing LTR clusters are distinguished by having high density of SNPs and genes (for chromosome 21, see Kurdyukov et al., 2001). For example, chromosome 19 has both the highest gene density and the largest number of LTR sequences. The following figure combines the SNP density histogram and the ideogram with LTR locations for human chromosome 5.

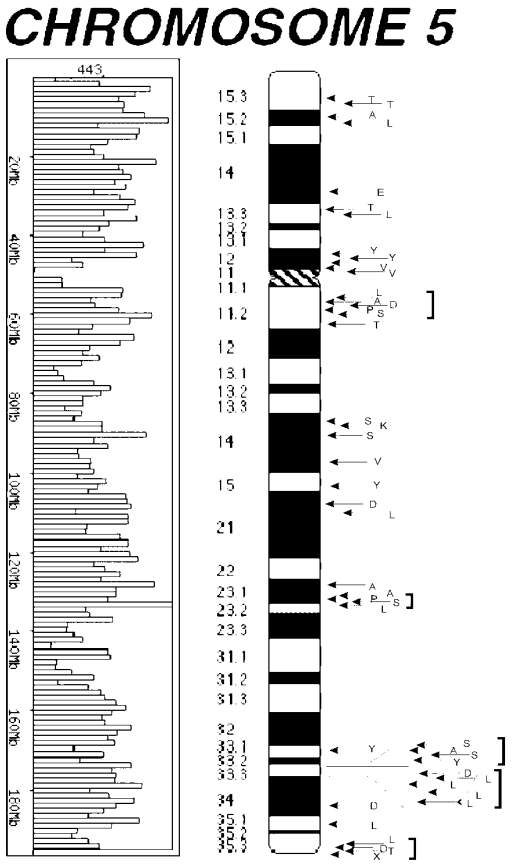


Fig. A SNP density histogram and an ideogram with LTR locations for human chromosome 5. Each arrow corresponds to LTR sequence location. Latin letter to the right marks belonging to one of 12 structurally homogeneous families. Square brackets indicate LTR clusters.

Acknowledgements

This work was partially supported by the Russian Foundation for Basic Research (project № 00-04-48166). We thank V.Ruda for assistance in translation and Dr. M.S.Gelfand for helpful discussions.

References

1. Artamonova I.I., Gorodentseva T.N., Lebedev Y.B., Sverdlov E.D. (2000) Nonrandom distribution of the endogenous retroviral regulatory elements HERV-K LTR on human chromosome 22. Dokl. Biochem. 372, 87-89.
2. Felsenstein J. (1993) PHYLIP version 3.5c. Distributed by the author. Based program in Felsenstein, J. (1989) PHYLIP – Phylogeny Inference Package. Cladistics. 5, 164-166.
3. International Human Genome Sequencing Consortium (2001) Initial Sequencing and analysis of the human genome. Nature. 409, 860-921.
4. Jurka J. (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 16, 418-20.
5. Kurdyukov S., Lebedev Y., Artamonova I., Gorodentseva T., Batrak A., Mamedov I., Azhikina T., Legchilina S., Efimenko I., Gardiner K., Sverdlov E. (2001) Full-sized HERV-K (HML-2) human endogenous retroviral LTR sequences on human chromosome 21: map locations and evolutionary history. Gene. 273, 51-61.
6. Lavrentieva I., Khil P., Vinogradova T., Akhmedov A., Lapuk A., Shakhova O., Lebedev Y., Monastyrskaya G., Sverdlov E. D. (1998) Subfamilies and nearest-neighbour dendrogram for the LTRs of human endogenous retroviruses HERV-K mapped on human chromosome 19: physical neighbourhood does not correlate with identity level. Hum. Genet. 102, 107-116.

DISTRIBUTION OF SHORT INVERTED REPEATS FLANKING DNA FRAGMENTS IN CEREAL CHLOROPLAST GENOMES AND THEIR APPLICATION FOR PCR-FINGERPRINTING

[1*] Ignatov A.N., [2] Mischenko A.S., [2] Yambartsev A., [3] Shimkevich A.V., [3] Goloenko I.M., [1] Dorokhov D.B., [1] Skryabin K.G., [3] Davydenko O.V.

[1] Centre «Bioengineering» RAS, 117312, Moscow, Russia
[2] Moscow State University, Russia
[3] Institute of Genetics and Cytology, NAS of Belarus, 220072, Minsk, Belarus, e-mail: ignatov@biengi.ac.ru
*Corresponding author

Key words: *chloroplast genome, cereals, barley, inverted repeats, genetic polymorphism, PCR*

Resume

Cultivated in temporary area grasses have evolved from common ancestor about 1.5 million years ago and maintain many similarities especially in cytoplasm genome. Chloroplast genome of those grasses consist of some 135,000 bp and approximately 100 functional genes. Genetic similarity of chloroplast genomes was evaluated by DNA hybridization with conservative probes and by sequencing of "hot spots" in intergeneric regions. Those methods are less useful for intraspecies studies, and PCR-fingerprinting was offered for this particular case. Information content of PCR markers may be increased by application of primers complementary to (a) DNA repeats abundant in genome or (b) DNA fragments participating in modification events. Distribution of small (7-10bp) inverted DNA repeats located in 100-4000bp from each other and thus capable to serve as PCR primers was studied in complete cereal genomes. The repeats flanking from 10 to 30 PCR-able fragments of the chloroplast DNA were chosen as AP-PCR primers and real PCR fingerprints obtained for wheat and 15 alloplasmic barley lines were compared to expected ones. Comparing to PCR analysis with random primers (RAPD), the selected primers produced bands of significantly (3-4 folds) higher PIC (Polymorphism Information Content), and GD (Genetic Diversity). Thus, PCR-fingerprinting with primers homologous for selected from sequence inverted repeats proved to be useful for genome variability studies.

Introduction

Cultivated in temporary area grasses: wheat, barley, rye, and oats (family *Poaceae*) have evolved from common ancestor about 1.5 million years ago and maintain many similarities especially in cytoplasm genome (Ogihara, et al., 1991). Chloroplast genome of those grasses consist of some 135,000 bp and approximately 100 functional genes. Several "hot spots" of variability resulted in length and composition changes of the genome were found in chloroplast DNA of grasses. Genetic similarity of cereal chloroplast genomes was evaluated by restriction analysis (Clegg et al., 1984; Ichikawa et al., 1989), DNA hybridization with conservative probes (Terachi et al., 1986; Hartmann et al., 1989) and by sequencing of "hot spots" in intergeneric regions (Shimada, Sugiura, 1991; Cosner et al., 1997; Hiratsuka et al., 1989; Belcour et al., 1997; Hill, Singh, 1997) and by PCR-RFLP analysis of conservative genes, introns, and microsatellites (Ishii, McCouch, 2000; Weising, Gardner, 1999; Bryan et al., 1999; Morton, Clegg, 1993; Belcour et al., 1997; Ogihara et al., 1991).Those methods are less useful for intraspecies studies, and PCR-fingerprinting was offered for this particular case. Information content of PCR markers may be increased by application of primers complementary to (a) DNA repeats abundant in genome or (b) DNA fragments participating in modification events. Minor genetic changes common for interspecies level can be grouped as (a) single nucleotide changes, and (b) insertions or deletions. Short dispersed direct and inverted repeats of DNA sequences can be "hot spots" for recombination (Cosner et al., 1997; Hiratsuka et al., 1989), and moreover, may be moved by transposase-proteins synthesized by transposons of II type, present in cereal genome.

Methods and Algorithms

Available in the GeneBank complete and partial sequences of grasses chloroplasts were analyzed by selection in ring sequence different types of short (7-10 bp) Inverted Repeats Flanked DNA Fragments (IRFFs) of 10-2000 bp in length. Only repeats suitable for PCR amplification were selected. Two less diverged chloroplast genomes: wheat and rice were compared by theoretical spectra given by short IRFFs and most informative inverted repeats were selected for PCR analysis as primers. Random sequence of 135,000bp with CG content equal to grass chloroplast genome was generated as control. The IRFFs frequency were compared by chi-square and Kolmogorov-Smirnov analysis. Genetic Distance between different species and accessions within one species was calculated as described by Nei and Li (1979): $(GS)_{ij} = 2\ N_{ij} / (N_i + N_j)$,

$GD_{ij}=1-(GS)_{lj}$ where (GS) – genetic similarity, N_{ij} - a number of common IRFFs for one inverted repeat for accessions i and j, N_i and N_j – total number of fragments for accessions i and j. The Polymorphism Information Content (PIC) was calculated for each primer according to Nei (1973): (PIC) =1 - $\sum x_k^2$ where x_k is a frequency of κ-IRFFs.

Results

In average, some 30,000 DNA fragments from 100 to 4000bp flanked by exact 7bp-long inverted repeats and only about 1800 IRFFs flanked by exact 10bp-long repeats were found in chloroplast genomes of cereals. Large part of the IRFFs (от 29 до 82%) was present in a single copy only, giving one amplifiable fragment at PCR analysis. Their distribution was almost random over the entire genome except two duplicated regions containing rRNA gene clusters. Most abundant IRFFs (10-50 repeats per genome) possess high content of single- or two-nucleotide stretches, mostly $(A)_n$ or $(T)_n$, and their frequency was significantly higher comparing to random nucleotide sequence. It has explanation in DNA-slippage, frequent in the chloroplast genome. Such IRFFs were located mostly in intergeneric regions and were most polymorphic between the chloroplast genomes of different species, but could not be used as primers for arbitrary-primered PCR analysis. The IRFF with intermediate (4-10 repeats) frequency were concentrated in intergeneric regions and in the regions coding ribosomal RNA and protein genes. Several inverted repeats from 7 to 10bp-long flanking DNA fragments from 100 to 2000 of most variable size with CG content 40-60% and without single-nucleotide stretches >3bp those were found to be highly polymorphic between cereal chloroplast genomes were used as AP-PCR primers. Primers shorter than 10bp were elongated by degenerated CG nucleotides from 5'-end till length of 10. AP-PCR was made on chloroplast DNA of wheat and 15 alloplasmic lines of barley, representative for distinct cytoplasm types within *Hordeum* species. Amplified in AP-PCR loci were significantly (3-4 folds) more informative according to PIC (Polymorphism Information Content) and GD (Genetic Diversity) indexes comparing to Random Amplified DNA (RAPD) analysis with 10-bp primers of commercial design. Most of loci were present in DNA of both genera (*Hordeum* and *Triticum*) but nuclear DNA.

Discussion

As it was mentioned before, the dispersed short repeats of DNA sequences in chloroplast genome can be "hot spots" for recombination, can be transposed on MITE-fashion and modified by DNA-slippage. Application of those repeats, as template for PCR-primers will have a great potential for taxonomic studies and identification of mutation events within genomes of the same species. Complete-sequence fascinated computer analysis have enabled to select inverted repeats of greatest polymorphism between related species of cereal plants those were proved to be "hot-spots" of changes on intra-species level as well. Thus, AP-PCR analysis with primers homologous to short dispersed inverted repeats in chloroplast genome may be either an effective method for evaluation of genetic variability within the same species of plant or can provide valuable information about localization of genetic modifications in chloroplast genome.

Acknowledgements

We thank Dr. G.E.Pozmogova for excellent primers synthesis. This work was financed from RFBR grant № 00-04-81098-Bel2000_a.

References

1. Bryan G.J. et al. (1999) Polymorphic simple sequence repeat markers in chloroplast genomes of *Solanaceous* plants. TAG. 99, 859-867.
2. Morton B.R., Clegg M.T. (1993) A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near rbcL in the grass family (*Poaceae*). Curr Genet. 1993. 34, 357-65.
3. Hiratsuka J. et al. (1989). The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. 1989. 217, 185-94.
4. Ogihara Y. et al. (1991) Molecular analysis of the hot spot region related to length mutations in wheat chloroplast DNAs. I. Nucleotide divergence of genes and intergenic spacer regions located in the hot spot region. Genetics. 129, 873-84
5. Cosner M.E. et al. (1997) The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Current Genet. 31(5), 419-429.
6. Nei M. (1973) Analysis of gene diversity in subdivided populations. PNAS. 70, 3321-3323.
7. Nei M., Li W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. PNAS. 76, 5269-5273.
8. Shimada H., Sugiura M. (1991) Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucl. Acids Res. 19, 983-95.
9. Terachi T., Tsunewaki K. (1986) The molecular basis of genetic diversity among cytoplasms of *Triticum* and Aegilops. 5. Mitochondrial genome diversity among Aegilops species having identical chloroplast genomes. TAG. 73, 175-181.
10. Weising K., Gardner R.C. (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome. 42, 9-19.

IN SILICO ANALYSIS OF HUMAN GENOMIC SEQUENCES, ADJACENT TO HPV16 INTEGRATION SITES

**** Klimov E.A., Rakhmanaliev E.R.***

N.I.Vavilov Institute of General Genetics, RAS, Moscow, Russia, e-mail: klimov_eugeney@mail.ru
*Corresponding author

Key words: *HPV16, RH-mapping, computer analysis, exon-intron structure*

Resume

Motivation: Earlier we localized HPV16 integration sites in human chromosomes using RH-mapping. Homologies of nucleotide sequences of human genome adjacent to virus integration sites with human genes and/or ESTs were demonstrated. In this work, we analyzed *in silico* human nucleotide sequences, detected as papilloma virus integration sites.

Results: Four hypothetical genes (*LOC151128*, *LOC161154*, *KIAA1808* and *KIAA0887*) and gene similar to *WASF2*, which include HPV16 integration sites, were characterized *in silico*. Exon-intron structure and the presence of necessary regulator elements in promoter region of these genes were detected. For protein products of genes (KIAA1808, KIAA0887 and protein of gene similar to *WASF2*) homologies were found and hypothetical functions were determined.

Introduction

We performed analysis of nucleotide sequences of human genome, adjacent to integration sites of human papilloma virus type 16 (HPV16). HPV16 is an etiology factor in development of cervical cancer. The virus genome contains genes E6 and E7, products of which are able to interact with products of genes-suppressors (p53 and Rb105, accordingly), evoking inactivation of these proteins. The dysfunction of p53 and Rb105 is crucial factor for transformation of cells, since these genes play the key role in regulation of cell circle. The expression of viral genes can occur both in episomal and in integration virus forms. In the case of integrative form, viral mRNA with poly-A tail from 3'-end is linked with cellular sequence, poly(A)-site sequences are present downstream to 3'-end of integrated viral DNA (Klaes et al., 1999; Kiselev et al., 2001). In the laboratory of molecular biology of viruses of Blokhin Institute of Carcenogenesis, Cancer Research Center (Moscow) cellular expressed sequences adjacent to virus (the INT markers) were isolated, cloned and sequenced, using the observation.

Methods

The homologies were searched by the BLASTN-program provided by NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). Exon-intron structure of novel genes was created by BLASTN (NCBI) and GENSCAN (http://genes.mit.edu/GENSCAN.html) programs. Promoter regions were identified using PromoterInspector (http://genomatix.gsf.de/cgi-bin/promoterinspector/promoterinspector.pl) and Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html) programs. For search of binding sites of transcription factors, we used MatInspector program (http://transfac.gbf.de/cgi-bin/matSearch/matsearch.pl). Amino acid sequences encoded by novel genes were translated *in silico* using DNA=>Protein service (http://cn.expasy.org/tools/dna.html). We also used programs at GeneBee server (http://genebee.msu.ru/) for the search of amino acid homologies, construction of the full local similarity maps for hypothetical proteins and phylogenetic trees. Information concerning proteins, encoded by novel genes was obtained from OMIM database (http://www.ncbi.nlm.nih.gov/OMIM/).

Results

Earlier we have localized five INT markers in genome of cervical carcinoma cells using RH-mapping method. The results of RH-mapping have demonstrated that viral DNA integrates in high-density gene regions of different chromosomes. Using the BLASTN program we obtained homologies for INT markers with human genomic sequences and ESTs.

Searching homologies with human genomic sequences confirmed the results of RH-mapping for five markers and allowed us to localize seven earlier non-mapped markers. From 12 INT markers, two have multiple chromosomal localization. The result of searching homologies of human ESTs for marker INT467 allowed us to detect, that given sequence is similar to mRNA of multicopy gene 40S ribosomal protein S27 (*MPS1*), which is localized on several chromosomes (Table 1). Marker INT290 is localized by RH-mapping on Xp11.3, however, it has the high level of homology with two genomic sequences, localized on chromosome 1 (region p36) and X (region p11.3).

To increase the resolution of searching homologies of INT markers and human ESTs, we used the genomic sequence, flanking INT markers from both 3'- and 5'-ends (2 kb). As result of this search, two markers were obtained (INT475 and INT423), which had homology with sequences of the known genes (*NIMP* and *GLS*, accordingly). Four markers (INT290, INT505, INT466 and INT467) have homology with the sequences that are similar to known genes. For the markers

INT259, INT477 and INT421, the high level homology was shown with transcripts of hypothetical genes. Only for three markers (INT254, INT431 and INT407), no homologies were obtained with human genes or ESTs (Table 1).

Based on data from MapViewer program, for two markers (INT254 and INT407) without any homologies with genes or ESTs, there was determined that virus was integrated in introns of the hypothetical genes (*LOC161154* and *KIAA0887*). Marker INT431 had homology with genomic sequence disposed between hypothetical genes *LOC122134* and *LOC122133* (Table 1).

Table 1. Homology of INT markers and adjacent cellular nucleotide sequences with genes and site of integration (exon or intron number), in which HPV16 integrated.

INT (length, bp)	Homology level with human genome sequences		Chromosome localization	Homology with human genes / ESTs	Site of integration
	Length, bp	%			
254 (349)	352	95	14q23.2	LOC161154	Intron 2
259 (213)	210	98	4p16.1	Gene for KIAA1808 protein	Intron 12
431 (175)	175	100	13q21.23	Between LOC122134 and LOC122133	-
407 (150)	148	98	5q35.3	Gene for KIAA0887 protein	Intron 1
290 (272)	268	98	Xp11.3	*LOC158537,* similar to human WASF2 protein	Exon 3
505 (249)	248	93	10q23.32	Similar to myoferlin, *MYOF*	Terminal exon
466 (385)	125 149	99 92	3q22.3	*LOC152028*, similar to human interferon alpha/beta receptor 1	Exon 5
477 (384)	372	94	7q21.11	LOC168295	Intron 5
467 (296)	-	-	1,2,3,4,5,6,7,11,12,15,18,19	Similar to human 40S ribosomal protein S27 (*MPS1*), mRNA	Exon
421 (322)	326	98	2q22.1	LOC151128	Intron 3
475 (320)	306	99	6q21	NOGO-interacting mitochondrial protein (*NIMP*)	Intron 6
423 (337)	338	98	2q32.3	Glutaminase (*GLS*)	Terminal exon

We have made the attempt to describe the sequences of four hypothetical genes (*LOC161154*, *LOC151128*, *KIAA1808* and *KIAA0887*) and the sequence similar to *WASF2* gene. For these genes, exon-intron structures were described. For all genes, start- (ATG) and stop-codons, poly-A sites and canonical splicing sites (AG…GT) were found. Basic binding sites of transcription factors in promoter regions of all genes were found. Exon-intron structure and necessary regulator elements for gene expression were presented in Table 2.

We determined the presupposed function for two hypothetical proteins (KIAA1808 and KIAA0887) by searching of amino acid homologies and construction of the full local similarity maps (Table 2).

Table 2. Basic elements of structure of the studied hypothetical genes and homologs for encoded proteins.

Genes	Basic transcription sites	Stop codon	PolyA-sites	Number of exons	Protein homology
LOC161154	GATA; TATA; OCT1	TGA	AATAAA	6 exons	-
KIAA1808	GATA; OCT1	TAA	AATAAA	17 exons	ABLIM 57%
Similar to *WASF2*	OCT1; TATA; GATA	TAA	AATAAT	3 exons	WASF2 88%
KIAA0887	GATA; SP1; OCT1	TGA	AATAAA	11 exons	Faf-P1 (*D.m.*) 37%
LOC151128	GATA; TATA; OCT1	TAG	AATAAA	4 exons	-

The ABLIM protein (LIM actin-binding protein 1) contains LIM domain, playing a key role in regulation of development. Protein Faf-P1 (*D. melanogaster*) contains UBX domain. The WASF2 protein is a member of the GTP-ase family proteins

carrying signal to actin cytoskeleton. The data obtained allow us to refer the genes, encoding these proteins, to the group of house keeping genes, since they are required for vital activity of all cell types of organism.

Discussion

The results of RH-mapping and searching homologies within databases with human genomic sequences allow us to confirm previous observation, that integration of virus is non-specific and can occur to the different regions of cellular genome. The data are in agreement with published results of other authors (Thorland et al., 2001; Wentzensen et al., 2002). However, we suggest, that viral DNA integrates in actively transcribed genes, i.e., in decompactizated genomic regions (data in press). This suggestion was made on the basis of fact that homologies human ESTs presented in databases in several tens of clones. The studied hypothetical genes (*KIAA1808* and *KIAA0887*) are expressed in the most cell types and their proposed function also allowed us to consider these genes as house keeping genes.

Acknowledgements

This work was supported by "Human Genome Project" (project № 89'99) and RFBR (project № 00-15-97777). The authors are grateful to prof. Sulimova G.E. for helpful discussions.

References

1. Kiselev F.L., Kiseleva N.P., Kobzeva V.K., Gritsko T.M., Semenova L.A., Pavlova L.S., Klaes R., von Knebel Doeberitz M. (2001) Status of the human DNA papillomavirus in cervical tumors. Mol. Biol. 35. 470-476. (In Russ.).
2. Klaes R., Woerner S.M., Ridder R., Wentzetzen N., Duerst M., Schneider A., Lotz B., Melscheimer P., von Knebel Doeberitz M. (1999) Detection of high-risk cervical intraepithelial neoplasia and cervical cancer by amplification of transcripts derived from integrated papillomavirus oncogenes. Cancer Res. 59, 6132-6136.
3. Thorland E.C., Myers S.L., Persing D.H., Sarkar G., McGovern R.M., Gostout B.S., Smith D.I. (2000) Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites. Cancer Res. 60, 5916-5921.
4. Wentzensen N., Ridder R., Klaes R., Vinokurova S., Schaefer U., Doeberitz M.K. (2002) Characterization of viral-cellular fusion transcripts in a large series of HPV16 and 18 positive anogenital lesions. Oncogene. 21, 419-426.

DETECTION OF CONSERVATIVE CONFORMATIONAL PROPERTIES OF INSERTION SITES FOR DROSOPHILA RETROTRANSPOSONS

**** Oshchepkov D.Yu., Furman D.P., Katokhin A.V., Katokhina L.V.***

Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia, e-mail: diman@bionet.nsc.ru
*Corresponding author

Key words: *D. melanogaster, retrotransposon, conformational and physicochemical DNA properties*

Resume

Motivation: The molecular mechanisms underlying the localization (choosing) of the potential site for retrotransposon insertion into the host genome require further studies. Conformational characteristics of the target site DNA might be a factor essential for their selection. The goal of this work was to detect significant conformational DNA properties in potential target sites for a number of *Drosophila* retrotransposons.

Results: Analysis of samples of potential retrotransposon insertion sites for five retrotransposon families (*297, 17.6, yoyo, tirant,* and *roo*) by the SITECON technique has demonstrated that a set of DNA conformational properties within a region of ±10 bp from the insertion site center is conserved. The data obtained suggest molecular mechanisms involved in choosing of the target sites while retrotransposon integration into the host genome.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/.

Introduction

Typical of the retrotransposons, widely represented in eukaryotic genomes and, in particular, *Drosophila* genome, is their capability of transposing spontaneously or upon certain induction, which makes them an essential source of mutation-based variation. Thus, a detailed research into fine mechanisms of retrotransposon transposition and the factors affecting it is evidently necessary.

It is known that the integration of retrotransposons into new chromosomal sites proceeds via formation of an integration complex, comprising a copy of retrotransposon as a DNA intermediate and integrase, encoded by retrotransposon ORF *pol.* The integrase bound to LTR sequences of the DNA intermediate, first, generates a double-strand step break in the chromosomal target site and then, links covalently the ends of intermediate to chromosomal DNA. In this process, the target site is duplicated (Labrador, Corces, 1997). The molecular mechanisms underlying the localization (choosing) of the potential site for retrotransposon insertion into the host genome require further studies. For example, it is known that certain host proteins may be involved in the retrotransposon integration; however, their roles are yet vague (Labrador, Corces, 2001).

As a rule, the regions encompassing target sites of retrotransposons belonging to various families display no contextual specificities (Dzhumagaliev et al., 1986; Labrador, Corces, 1997). However, the contextual specificity may be encoded in a more complex manner at the level of secondary DNA structure in a form of conformational signals representing markers of specific DNA–protein interactions (Trifonov, 1997). Indeed, such signals were discovered for insertion sites of *D. melanogaster P* element (Liao et al., 2000). A growing volume of experimental data suggests that the function of active DNA sites is determined to a considerable degree by their conformational and physicochemical properties (Meierhans et al., 1997). Data of structural analysis demonstrate that conformational and physicochemical properties of DNA depend on the nucleotide sequence (Dickerson, Drew, 1981; Suzuki et al., 1997).

Thus, the local conformation of DNA molecules determined by the context is a factor controlling the specificity of active DNA sites. This suggests that the local DNA regions contacting with a particular protein within active sites should display similar conformational and physicochemical properties; moreover, this property will be stable independently of a certain context variation within such regions.

An original method SITECON (Oshchepkov et al., this issue), involving detection of conservative conformational and physicochemical DNA properties in samples of aligned sequences, was used to analyze retrotransposon insertion sites.

Analysis of samples of potential retrotransposon insertion sites for five retrotransposon families (*297, 17.6, yoyo, tirant,* and *roo*) has demonstrated that a set of DNA conformational properties within a region of ±10 bp from the insertion site center is conserved. The data obtained allow possible molecular mechanisms involved in choosing of the target sites while retrotransposon integration into the host genome to be discussed.

Materials and Methods

The reconstructed euchromatic sequence of *D. melanogaster* genome (Adams et al., 2000) and sequences of retrotransposons belonging to 11 families (*297, 17.6, yoyo, HMSBeagle, mdg1, mdg3, Dm412 (mdg2), copia, blood, roo (B104),* and *tirant)* retrieved from FlyBase (http://flybase.harvard.edu:7081/transposons/lk/melanogaster-transposon.html) were used for the analysis.

The retrotransposons were localized in the sequenced genome by the program BLASTn (http://www.ncbi.nlm.nih.gov/BLAST/ and http://www.fruitfly.org/blast/index.html) using the parameters proposed.

The sets of insertion sites of retrotransposons from 11 families were composed of sequences with a length of 60 bp (±30 bp of the target site center), each containing one copy of the target site with the flanking fragments of genomic DNA encompassing the site. The total number of sequences analyzed amounted to 328 (Katokhin et al., this issue).

The samples of insertion site sequences aligned with respect to the target site were analyzed by the SITECON technique (Oshchepkov et al., this issue). The method utilizes the published data on 38 conformational and physicochemical properties of dinucleotides compiled in the database PROPERTY (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/) and allows the conservation of these properties in short DNA fragments to be detected. Dispersion of the values of these properties at individual positions within the aligned retrotransposon insertion site sequences was used as a measure of their conservation.

Results and Discussion

Overall, 11 samples of retrotransposon insertion sites were analyzed by SITECON. Five families (samples)—*297, 17.6, yoyo, tirant,* and *roo*—displayed sets (patterns) of local conservative conformational and physicochemical properties within the region of ±10 bp from the target site center. The corresponding plots are shown in Fig. 1.

Let us consider the pattern of conservative conformational and physicochemical properties of retrotransposon target sites by the example of the family *297* (Fig. 1a). In this case, all the 38 properties studied display conservation within the region of [–1;+2] (a column of white cells with a width of three positions), corresponding to the context conservation at positions [–2;+2] or, in other words, to three conservative dinucleotides of the target site. However, it is evident that certain properties display conservation beyond the target site as well, for example, the properties Roll and Twist. The last two properties exhibit a significant (95%) conservation even at a distance of 12 bp from the target site center. Note also the property Slide, which displays conservation at all the positions closer than 4 bp to the target site center, and the property Tilt with a region of conservation over 4 bp at a distance of 13 bp upstream of the target site center.

It is evident that in the case of retrotransposon *17.6* target sites, similar to the retrotransposon *297,* the sequence of the target site itself is conservative at the level of mononucleotides and all the positions within the target site display conservation of all the 38 conformational properties.

Similarly, certain properties are conserved beyond the target sites: the property Roll (No. 28) at positions ±8 bp from the *297* site center and at positions ±3, 4, and 6 from the *17.6* site center (Fig. 1a, b).

It is essential that individual conservative conformational properties are detected at certain positions within the neighborhood of target site even if the site sequence itself is degenerate (target sites of retrotransposons *yoyo, tirant,* and *roo*), in particular, the property Twist (No. 18) at positions ±3–4 for *yoyo* (Fig. 1c); Tilt (No. 12) at positions ±3 for *tirant* (Fig. 1d); and Roll (No. 10) at positions ±5–6 for *roo* (Fig. 2). Thus, it may be assumed that occurrence of positions displaying conservative conformational properties in the neighborhood of the target site is a factor determining selection of the site for retrotransposon insertion into the host genome.

This fact suggests that the positions detected in the regions encompassing the target sites may serve as conformational signals for either the integrase itself or host DNA-binding proteins whose activity is utilized by integrase. Precedents are known. For example, the specificity of *gypsy* insertion within 1.3 kbp of the *ovo* gene 5'-region is mediated by an interaction between the protein OvoA and components of *gypsy* integration complex (Labrador, Corces, 2001).

The described diversity of the patterns of local conservative DNA properties in the regions adjacent to insertion sites of retrotransposons from various families (Fig. 1) may reflect both differences in the structures of integrases synthesized by different retrotransposons and involvement of a set of host DNA-binding proteins in the retrotransposon integration.

The fact that no stable conformational signals were detected in the neighborhood of insertion sites of retrotransposons from the rest six families may indicate, for example, a nonspecific interaction of their integration complexes with DNA or a key role of heterochromatin packaging, in particular, its nucleosomal organization (Katokhin et al., this issue), in the retrotransposon integration.

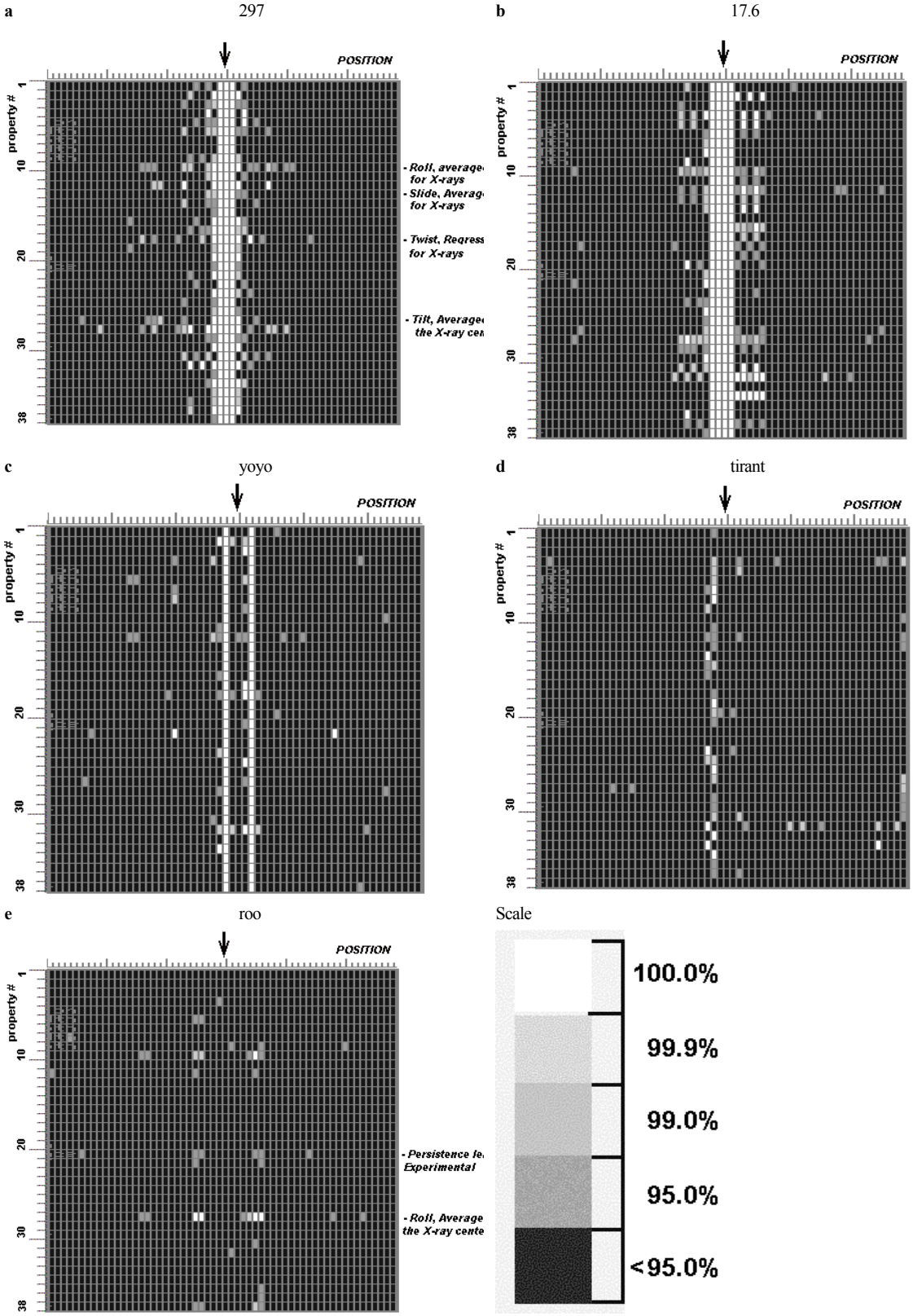


Fig. 1. Patterns of conservative conformational and physicochemical DNA properties of phased samples of insertion sites for retrotransposons belonging to (**a**) 297, (**b**) 17.6, (**c**) yoyo, (**d**) tyrant, and (**e**) roo families: the ordinate, numbers of properties in the database PROPERTY; the abscissa, positions (bp); the center of insertion site at position 30 (arrow). Color code: black cells indicate the positions whereat the corresponding property displays no conservation; gray tone intensity, significance of the detected conservation estimated with χ^2 test; and white cells, a complete conservation.

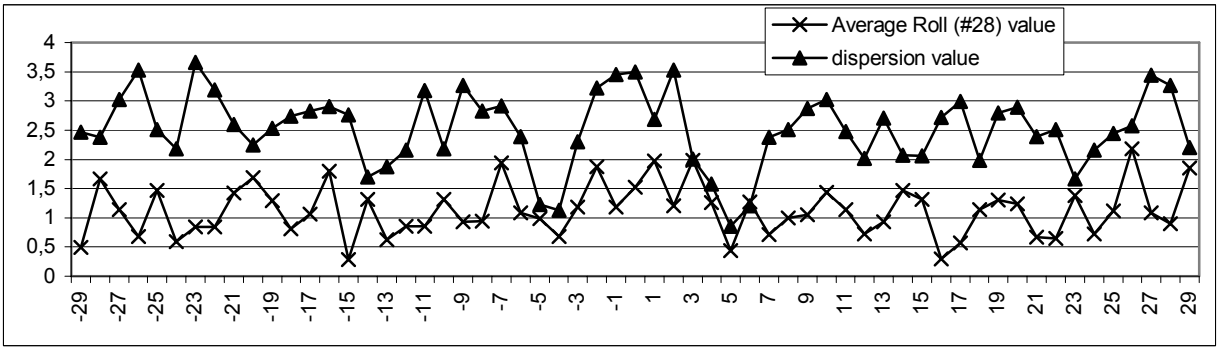


Fig. 2. An example of conservation of the property Roll (No. 28 according to PROPERTY database) in the neighborhood of a degenerate target site of a roo family retrotransposon: the plots of average Roll value and its dispersion value

References

1. Adams M.D., Celniker S.E., Holt R.A. et al. (2000). The genome sequence of *Drosophila melanogaster*. Science. 287:2185-2195.
2. Dzhumagaliev E.B., Mazo A.V., Bayev A.A. Jr., Gorelova T.V., Arkhipova I.R., Schuppe N.G., Ilyin Yu.V. (1986). The structure of long terminal repeats of transcriptionally active and inactive copies of Drosophila mobile dispersed genetic elements MDG3. Genetika (Russ.). 22:368-377.
3. Dickerson T.D., Drew H.R. (1981). Structure of B-DNA dodecamer. II. Influence of base sequence on helix structure. J. Mol. Biol. 149:761-786.
4. Katokhin A.V., Furman D.P., Levitsky V.G., Katokhina L.V. (2002). Nucleosomal organization of Drosophila retrotransposon insertion sites. This issue.
5. Labrador M., Corces V.G. (1997). Transposable elements–host interactions: regulation of insertion and excision. Ann. Rev. Genet. 31:381-404.
6. Labrador M., Corces V.G. (2001). Protein determinants of insertional specificity for the Drosophila gypsy retrovirus. Genetics. 158(3):1101-1110.
7. Liao G.C, Rehm E.J., Rubin G.M. (2000) Insertion site preferences of the P transposable element in *Drosophila melanogaster*. Proc. Natl Acad. Sci. USA. 97(7):3347-3351.
8. Meierhans D., Sieber M., Allemann R.K. (1997). High affinity binding of MEF-2C correlates with DNA bending. Nucl. Acids Res. 25:4537-4544.
9. Oshchepkov D.Yu., Turnaev I.I., Vityaev E.E. (2002). Study of the context-dependent conformational and physicochemical properties of DNA functional sites. This issue.
10. Suzuki M., Amano N., Kakinuma J., Tateno M. (1997). Use of 3D structure data for understanding sequence-dependent conformational aspects of DNA. J. Mol. Biol. 274:421-435.
11. Trifonov E.N. (1997). Genetic level of DNA sequences is determined by superposition of many codes. Mol. Biol. (Mosk.). 31(4):759-767.

NUCLEOSOMAL ORGANIZATION OF *DROSOPHILA* RETROTRANSPOSON INSERTION SITES

**** Katokhin A.V., Furman D.P., Levitsky V.G., Katokhina L.V.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: katokhin@bionet.nsc.ru
*Corresponding author

Key words: *D. melanogaster, retrotransposon, insertion site, nucleosomal potential, nucleosome context*

Resume

Motivation: It is known that nucleosomal organization of chromatin plays an important role in regulation of gene expression. Changes in nucleosome positioning and the consequent changes in expression of genes in the region of insertions are a possible aftermath upon insertion of retrotransposons (REs) into the host genome. This aspect of the interactions between the genome and retrotransposons yet remains vague. The goal of this work is to clarify, first, whether the nucleosome potential profile is specific of the genomic regions housing potential insertion sites for various retrotransposons and, second, whether the nucleosome potential profiles of LTRs of retrotransposons belonging to different families and the corresponding genomic regions encompassing the insertion sites are different.

Results: Nucleosome potential profiles of insertion sites (genomic regions where insertions were detected) for retrotransposons from eleven families were for the first time studied using computer tools and compared with the corresponding profiles of potential insertion sites ("native" sequences obtained by *in silico* reconstruction). It was demonstrated that (1) potential insertion sites of several REs displayed specific nucleosome profiles correlating with nucleotide compositions of both the target sites themselves and the inverted terminal repeats (ITRs), forming the 5' and 3' ends of each retrotransposon LTR, and (2) the profiles of nucleosome potentials of retrotransposon LTRs from different RE families differed from one another as well as from the profiles of the genomic regions encompassing these inserts.

Introduction

Mobile elements (transposons) constitute a considerable part of the eukaryotic genomes. The mobile elements in general and retrotransposons (REs) in particular represent a mighty source of genotypic variation due to their influence on the gene structure–function organization in the regions of insertions.

As a rule, the main attention of the corresponding research is focused on the effects of RE insertions on the gene structure–function organization stemming from the changes in primary DNA sequence. However, it is known that expression of genes depends as well on chromatin packaging (Bonifer, 1999). Thus, it is natural to assume that REs are capable of changing function of the genes localized to the regions of insertions not only via the interactions at the DNA level, but also in an indirect manner through modulating the chromatin structure, in particular, its nucleosomal organization. In turn, this suggests that the nucleosomal organization of a DNA fragment wherein a certain retrotransposon has inserted might also be of importance to the integration itself.

Availability of Drosophila complete genomic sequence in combination with adequate computer methods for detecting and analyzing contextual and conformational properties of DNA sequences allowed these questions to be studied theoretically. The goal of this work was (1) to find out whether the nucleosome potential (NP) profiles of genomic regions containing potential sites for insertion of retrotransposons from various families are specific and (2) to detect the differences between the NP profiles of various retrotransposon LTRs and those of genomic regions encompassing the inserts.

Materials and Methods

The reconstructed euchromatic sequence of *D. melanogaster* genome (Adams et al., 2000) and the RE sequences with the copy numbers exceeding seven retrieved from FlyBase (http://flybase.harvard.edu:7081/transposons/lk/melanogaster-transposon.html) were used for the analysis.

The REs were localized in the genomic sequence by the program BLASTn (http://www.ncbi.nlm.nih.gov/BLAST/ and http://www.fruitfly.org/blast/index.html) using the default parameters.

The NP profiles were studied using sets of sequences with a length of 1200 bp (±600 bp relative to the site of RE insertion) divided into two following groups: (1) "5'flanks + LTRs", comprising genomic sequences adjacent to left LTRs together with the LTR sequences, and (2) "5'flanks + 3'flanks", comprising genomic fragments adjacent to left LTRs attached to the sequences adjacent to right LTRs. The total number of analyzed sequences for REs belonging to 11 families (Table) amounted to 328. The method RECON (Levitsky et al., 2001) was used to calculate NP.

Table. Characteristics of *D. melanogaster* retrotransposons analyzed.

RE family	Target site	ITR	LTR (bp)	Copy number
297	ATAT	AG/CT	415	20
17.6	ATAT	AG/TT	512	7
yoyo	**T**(**At**)(**Ta**)(**At**)	AG/CT	519	8
HMSBeagle	**A**(**Ga**)(**Tc**)**A**	AG/CT	266	7
mdg1	(**C**ag)(**AT**)(**Ag**)(**T**cg)	TG/CA	442	8
mdg3	(**Ga**)(**Tc**)(**AT**)(**Tg**)	TG/AA	267	7
Dm412 (mdg2)	(**A**gc)(**Tg**)(**At**)(**Gc**)	TG/CA	481	16
copia	(**Ag**)(**Ta**)(**Ta**)(**TA**)**C**	TG/CA	276	14
blood	(**Ga**c)**T**(**Ag**)(**G**ct)	TG/CA	400	13
roo (B104)	(**A**gc)(**T**cg)(**T**ac)(**A**gt)(**C**tg)	TG/CA	429	36
tirant	CGCG	AG/CT	417	10

Results and Discussion

Comparative analysis of the data obtained demonstrated a diversity of NP profiles of the sequences from both sets. Most typical profiles are shown in Fig.

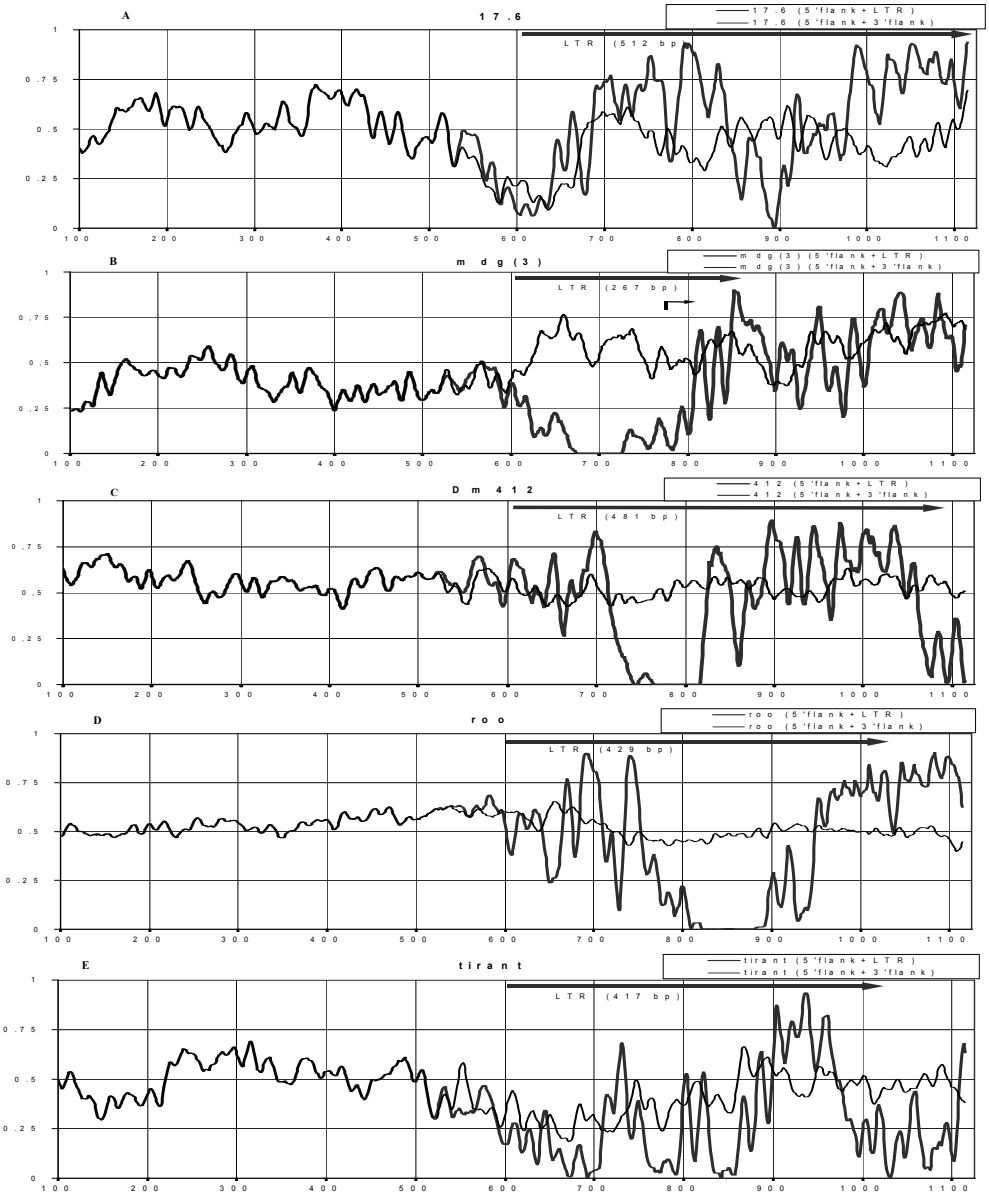


Fig. NP profiles of insertion sites of (**A**) *17.6*, (**B**) *mdg3*, (**C**) *Dm412*, (**D**) *roo*, and (**E**) *tirant* retrotransposons: the ordinate, value of NP; the abscissa, position (bp); the insertions correspond to position 600; gray line, NP profiles of group (1) "5'flanks + LTRs"; black line, NP profiles of group (1). "5'flanks + 3'flanks"; arrows above NP profiles indicate LTR lengths; and right-angle arrow marks the transcription start for *mdg3* (**B**).

1) Properties of "5'flanks + 3'flanks" sequences:

The sequences of this type free of *17.6* and *297* inserts (*17.6*-less and *297*-less) displayed a decrease in the value of NP in the region of ±100 bp relative to the site center (Fig. A). The analogous *tirant* sequences (Fig. E) manifested this property to a lesser degree. The sequences free of the rest REs did not display this pattern.

The elements *17.6* and *297* have identical ITRs and target sites (Table), suggesting that this underlies the observed similarity in their nucleosomal potential profiles.

The *tirant* ITR displays the same composition but a different target site (CGCG instead of ATAT). However, both sites are purine–pyrimidine motifs; hence, they are similar in certain conformational features (Fitzgerald, Anderson, 1999). Consequently, the specific features of NP profiles stemming from the presence of purine–pyrimidine tracts are likely to be important for integration of the REs with ITRs of AG/TC composition into the genome.

2) Properties of "5'flanks + LTRs" sequences:

The "5'flanks + LTRs" sequences with *17. 6* (Fig. A), *297,* and *copia* (data not shown) REs exhibit elevated NP over their LTR fragments compared with the corresponding sequences of "5'flanks + 3'flanks" type, that is, with the encompassing genomic context. This inference is experimentally confirmed for *copia:* it was demonstrated that a high nucleosome density of a *copia* sequence is comparable with the density typical of heterochromatin (Sun et al., 2001).

On the contrary, LTRs of all the rest REs - *mdg3, Dm412, roo*, *tirant* (Fig. B–E), *yoyo, mdg1, blood,* and *HMSBeagle* (data not shown)—display decreased values of the NP within the inner LTR regions. The NP profile observed reflects the presence of two nucleosomes at both LTR ends and an intermediate region of "open" chromatin. As a rule, such nucleosome-free regions correspond to gene regulatory regions, in particular, promoters (Bonifer, 1999; Levitsky et al., 2001). Promoter of *mdg3* is actually located precisely in this region (Mazo et al., 1986). The LTRs of the retroviruses MMTV and HIV-1 appeared to exhibit similar architecture (Bonifer, 1999; Marzio, Giacca, 1999).

References

1. Adams M.D., Celniker S.E., Holt R.A. et al. (2000). The genome sequence of *Drosophila melanogaster*. Science. 287:2185-2195.
2. Bonifer C. (1999). Long-distance chromatin mechanisms controlling tissue-specific gene locus activation. Gene. 238(2):277-289.
3. Fitzgerald D.J., Anderson J.N. (1999). DNA distortion as a factor in nucleosome positioning. J. Mol. Biol. 293(3):477-91.
4. Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. (2001). Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis. Bioinformatics. 17(11):998-1010.
5. Marzio G., Giacca M. (1999) Chromatin control of HIV-1 gene expression. Genetica. 106(1-2):125-130.
6. Mazo A.M., Arkhipova I.R., Cherkasova V.A., Gorelova T.V., Shuppe N.G. (1986). Fine structure of long terminal repeats and stages of reverse transcription of mobile dispersed genes in *Drosophila.* Genetika. 22(3):378-389.
7. Arkhipova I.R., Mazo A.M., Cherkasova V.A., Gorelova T.V., Schuppe N.G., Ilyin Y.V. (1986). The steps of reverse transcription of *Drosophila* mobile dispersed genetic elements and U3-R-U5 structure of their LTRs. Cell. 44(4):555-563.
8. Sun F.L., Cuaycong M.H., Elgin S.C. (2001) Long-range nucleosome ordering is associated with gene silencing in *Drosophila melanogaster* pericentric heterochromatin. Mol. Cell Biol. 21(8):2867-2879.

ANALYSIS OF THE NUCLEOSOME POTENTIAL OF DNA SEQUENCES GENERATED BY SELEX-EXPERIMENTS AND POSSESSING BY THE LOW AND HIGH AFFINITY TO HISTONE OCTAMER

**** Levitsky V.G., Podkolodny N.A.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: levitsky@bionet.nsc.ru
*Corresponding author

Key words: *nucleosome positioning, increased and decreased affinity to histone octamer*

Resume

Motivation: Nucleosome packaging of genome DNA makes an integral part necessary for DNA functioning *in vivo*. Previously, we have designed the software program for calculating nucleosome potential for a nucleotide sequence, which is a quantitative characteristics reflecting an ability of DNA to pack into a nucleosome (Levitsky *et al.,* 2001). For testing this program, it is important to study its ability to recognize DNA regions with high or low affinity to histone octamer within the samples of DNA sequences obtained experimentally.

Results: We have performed an analysis of two obtained by SELEX-experiments sets of DNA sequences with different potential to form a nucleosome. As shown, the program for calculating nucleosome potential gives proper classification of nucleotide sequences possessing by high or low affinity to histone octamer. This means that nucleosome potential value characterizes the affinity of the core histones to the DNA region under study.

Availability: The program for recognition of nucleosome sites is a part of the GeneExpress system, the section "DNA Nucleosomal Organization", http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/.

Introduction

The discriminative feature of eukaryotic genome is the complexly organized chromatin that provides compactness of genome DNA. The basal level of this packaging is a nucleosome organization of DNA. In this connection, development and estimation of the efficacy of computer-assisted methods aimed at searching for reliable signals in DNA, which support nucleosome positioning, are of essential importance. Recently, development of the methods of computer-assisted analysis and recognition of nucleosomal sites became of topical interest due to cardinal success in a mass sequencing of eukaryotic genomes, which brought a necessity to search for peculiarities of their structure-functional organization by computer methods.

Methods and Algorithms

For constructing a function characterizing an ability of DNA to form a nucleosome (hereinafter referred to as nucleosome potential), we have used a method based on discriminant analysis and on calculation of dinucleotide frequencies in the local regions of nucleosomal sites (Levitsky et al., 2001). This method leans upon searching for the blockwise structure of a nucleosomal site in the process of its partitioning into local regions with specific dinucleotide context. For constructing the method, we have searched for optimal partitioning of a nucleosomal site into local parts, which are characterized by more homogeneous dinucleotide content in comparison to that of the site in a whole.

Let us consider two sets of sequences: (1) the sites forming nucleosomes, (2) random sequences. As partitioning $\Omega(b_1,b_2,\ldots,b_{p-1})$ of a site [a,b], we understand the set of non-overlapping local regions $[a_p,b_p]$ (p=,1,…,P), which satisfy to the following conditions: (1) $a_1 = a$; (2) $a_{p+1} = b_p$, for p=1,…,P-1; (3) $b_P = b$. The search for the optimal partitioning has a goal to minimize the recognition errors. As the estimation of the quality of partitioning, we use the Mahalanobis distance, R^2 (Mahalanobis, 1936), between distributions in the sets (1) and (2):

$$R^2 = \sum_{k=1}^{N}\sum_{n=1}^{N}\{[f_n^{(2)} - f_n^{(1)}] * S_{n,k}^{-1} * [f_k^{(2)} - f_k^{(1)}]\} \qquad (1)$$

Here $f_n^{(1)} = f_{i,p}^{(1)}$ is an average frequency of the i-th dinucleotide for the p-th region of partitioning for the set of nucleosomal sites, whereas $f_n^{(2)}$ is a corresponding frequency in the set of random sequences (n=(p-1)×16+i, p = 1,…12, i = 1,…16, n=1,..,N), S^{-1} denotes the inverse matrix for the united covariation matrix, $S = S^{(1)} + S^{(2)}$, $S^{(1)}$ and $S^{(2)}$ are

covariation matrices, for the positive and negative sets of sequences, of vectors of dinucleotide frequencies, $f_n^{(1)}$ and $f_n^{(2)}$. The value R^2 depends upon N=16×P variables, being the dinucleotide frequencies in the regions of partitioning (number of dinucleotides is 16). The growth of R^2 corresponds to the mutual remoteness of the centers of distributions in the sets (1) and (2).

For analysis of an arbitrary DNA sequence, in each position of the sliding window, (the fragment X, 160 bp), we calculate the function φ(X):

$$\varphi(X) = \frac{1}{R^2} \times \sum_{n=1}^{N} \sum_{k=1}^{N} \{[f_n(X) - (1/2) \times [f_n^{(2)} + f_n^{(1)}] \times S_{n,k}^{-1} \times [f_k^{(2)} - f_k^{(1)}]\} \quad (2)$$

Here $f_n(X)$ is a vector of dinucleotide frequencies, which was constructed by accounting partitioning of the fragment X under study into local regions. Nucleosome potential φ(X) is constructed so that in the learning sample of sites forming the nucleosomes, its average value equals to +1, whereas in the negative learning sample (random sequences with dinucleotide frequencies equalling to 0.25), this value is -1.

For analysis, we have used two experimentally obtained sets of nucleotide sequences that differ by ability to form a nucleosome (Table).

Table. Sets of sequences used for analysis.

	Name of a set	Reference	Sample volume
1	Stable sites	Widlund *et al.*, 1997	86
2	Anti-nucleosomes	Cao *et al.*, 1998	40

In the first set, the sequences from the mouse genome are accumulated that are characterized by the maximal affinity to histone octamer ("stable nucleosomes"). The second set is represented by synthetic sequences with the least affinity to histone octamer ("anti-nucleosomes", or DNA fragments, for which nucleosome formation is at most difficult). For the program that we have developed for constructing nucleosome potential, the length of the sequence to be analyzed should be at least 160 bp. In the set of "stable nucleosomes", the sequence lengths vary from 109 to 151 bp, for "anti-nucleosomes", from 76 to 126 bp. For analysis of these sets, the flanking regions of sequences were added to the unified length of 160 bp by random sequences compiled from corresponding real sequences with preservation of their dinucleotide content. As we have shown previously by modeling of various types of random sequences (Levitsky et al., 2001), such addition does not deviate or change essentially the distribution of nucleosome potential.

Implementation and Results

For studying the sets of sequences differing by affinity to histone octamer (Table), we have analyzed the competence region, with respect to the input data, for our program calculating nucleosome potential.

Nucleotide content of the learning sample of nucleosomal sites equals to 59.3±11.3% (A+T). We suppose that in accordance with the Student's criterion, in the 99% interval [29.9; 88.6] (A+T)%, our recognition could be considered as reliable. Nucleotide content of the sample of "stable nucleosomes" is 54.9±7.6% (A+T), of that of "anti-nucleosomes", 50.3±3.7% (A+T). Both sets satisfy the criterion mentioned above for the nucleotide content.

For restriction of variability of dinucleotide content, we have used the dinucleotide measure introduced by Karlin, $\delta(X,X_{nuc})$ (Karlin, Ladunga, 1994), where X is a sequence analyzed, X_{nuc} is a sequence obtained by the unity of the learning samples of nucleosomal sites. Dinucleotide measure, $\delta(X,X_{nuc})$, is calculated according to the dinucleotide and nucleotide contents of sequences.

Following the Karlin's method (Karlin, Ladunga, 1994), we have averaged the dinucleotide content by complementary dinucleotides. For example, instead of dinucleotide frequencies, f_{AA} and f_{TT} , we use the frequency $f^*_{AA} = f^*_{TT} = 1/2 * (f_{AA} + f_{TT})$. By this averaging, it is possible to account the complementary chain. Then for each dinucleotide, XY, we determine the value, g_{XY} (Karlin, Ladunga, 1994), as follows:

$$g_{XY} = \frac{f^*_{XY}}{f^*_X * f^*_Y} \quad (3)$$

Here f^*_X and f^*_Y are the frequencies of nucleotides X and Y, calculated by accounting complementarities. For example, instead of nucleotide frequencies f_A and f_T , we use the frequency $f^*_A = f^*_T = 1/2 * (f_A + f_T)$. The value g_{XY} determines non-homogeneity of representation of the dinucleotide XY in comparison to that expected due to the nucleotide content. The measure $\delta(X_1,X_2)$ for the sequences X_1 and X_2 is determined by the following formula (Karlin, Ladunga, 1994):

$$\delta(X_1, X_2) = \frac{1}{16} * \sum_{i=1}^{16} | g_i(X_1) - g_i(X_2) | \quad (4)$$

For the sample of nucleosomal sites, $\delta = 0.208 \pm 0.003$, for the negative learning sample (random sequences), $\delta = 0.209 \pm 0.004$. For both these samples, $\delta < 0.5$ for all the sequences. We have used the sequences of "stable nucleosomes" and "anti-nucleosomes", which satisfy to condition $\delta < 0.5$ (60 out of 86 and 26 out of 40, respectively). For rejected sequences, the dinucleotide content is typical and it is strongly differing from that of the learning sample (35% of "anti-nucleosomes" are represented by repeats $(TGGA)_n$, with $\delta > 0.7$ for each of them).

Thus, for application of our program, the sequence should satisfy two conditions: (i) nucleotide content, (A+T)%, ranges from 29.9 to 88.6; (ii) for dinucleotide content, the Karlin's measure should be $\delta(X,X_{nuc}) < 0.5$.

Distributions of nucleosome potential for the sets of "stable nucleosomes" and "anti-nucleosomes" are illustrated in Figure. The mean value in the set of "stable nucleosomes" equals to 0.44 ± 0.02, in the sample of "anti-nucleosomes", 0.88 ± 0.04, the difference between the mean values is significant according to the Student's criterion, $p<10^{-5}$.

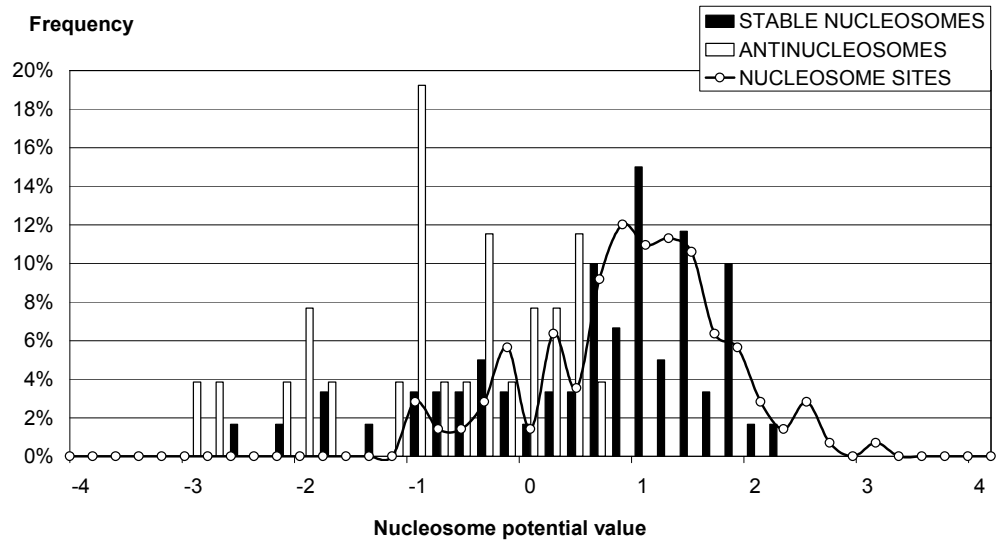


Fig. Distribution of nucleosome formation potential, $\varphi(X)$, in the samples of "stable nucleosomes" and "anti-nucleosomes" in comparison to distribution in the learning sample of nucleosome sites. The mean values in the samples "stable nucleosomes" and "anti-nucleosomes" differ significantly by the Student's criterion, $p<10^{-5}$.

Discussion

It should be noted that the interval with increased values of distribution frequencies in "stable nucleosomes" (Fig.) falls in the vicinity of +1, which is the mean value in the learning sample of nucleosomal sites. Notably, the distribution for "anti-nucleosomes" (Fig. 1) is essentially shifted to the left in comparison to distribution for the "stable nucleosomes".

The result obtained gives evidence on the appropriateness of the method developed for calculation of nucleosome potential and on its possible usage for the quantitative estimation of DNA's ability to form nucleosome in positions given. The result obtained could be also interpreted as another argument in favor of hypothesis claiming that an essential element of nucleosome positioning code in genome DNA is the regular location of dinucleotides along nucleosomal site.

Factually, the materials presented mean that the value of nucleosome potential calculated by (2) characterizes the affinity of the core histones to DNA region under study.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (№ 01-07-90376, 02-07-90355, 00-04-49229), Russian Ministry of Industry, Sciences and Technologies (№ 43.073.1.1.1501), Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65), National Institutes of Health USA (№ 2 R01-HG-01539-04A2), The Department of Energy USA (№ 535228 CFDA 81.049).

References

1. Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. (2001) Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis. Bioinformatics. 17, 998-1010.
2. Cao H., Widlund H.R., Simonsson T., Kubista M. (1998) TGGA repeats impair nucleosome formation. J. Mol. Biol. 281(2), 253-260.
3. Karlin S., Ladunga I. (1994) Comparisons of eukaryotic genomic sequences. Proc. Natl Acad. Sci. USA. 91, 12832-12836.
4. Widlund H.R., Cao H., Simonsson S., Magnusson E., Simonsson T., Nielsen P.E., Kahn J.D., Crothers D.M., Kubista M. (1997) Identification and characterization of genomic nucleosome-positioning sequences. J. Mol. Biol. 267, 807-817.

STUDY OF THE CONTEXT-DEPENDENT CONFORMATIONAL AND PHYSICOCHEMICAL PROPERTIES OF DNA TOPOISOMERASE I CLEAVAGE SITES

[1] Oshchepkov D.Yu., [2] Bugreev D.V., [1] Vityaev E.E., [2] Nevinsky G.A.

[1] Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia, e-mail: diman@bionet.nsc.ru
[2] Novosibirsk Institute of Bioorganic Chemistry, SB RAS, 630090, Novosibirsk, Russia

Key words: *conformational and physicochemical DNA properties, DNA–protein interactions, parameters of DNA recognized by human DNA topoisomerase I*

Resume

Motivation: DNA topoisomerase I (topo) is a sequence-specific enzyme; however, it is able to cleave DNA sequences differing from the optimal sequence with lower efficiency. The sequence and, consequently, conformational properties of the cleavage sites may play a crucial role in their ability to serve as substrates for topo.

Results: Analysis of DNA conformational properties of sites cleaved by topo revealed a number of specific double helix features that can be important for their efficient recognition and cleavage by the enzyme.

Introduction

DNA topoisomerases are enzymes responsible for maintaining and controlling topology of the DNA helix. In eukaryotes, topoisomerase I (topo) is essential for transcription and replication. The enzyme removes supercoiled turns, thereby providing the needed topological state of DNA during various functional processes (Wang, 1996). Human topo can relax both positively and negatively supercoiled DNA by introducing a temporary single-strand break with subsequent religation in a sequence-dependent manner. The rate of cleavage of sequences corresponding to the motif

```
                A          |     A      AAA
5'-AAAAAA   ACTT▼AG   AAA       -3'
                G          G     TTT
```

exceeds the average DNA cleavage rate by approximately three orders of magnitude. However, topo is also able to bind and cleave other DNA sequences that differ significantly in their structure (Perez-Stable et al., 1988). In this process, different topo cleavage sequences are selected according to the principles of structural conformity when realization of certain conformational rearrangements provides both stable binding and adjustment of reactive orbitals of the enzyme and the substrate with an accuracy of 10–15^{o} and allows highly efficient catalysis (Bugreev et al., 1999). Thus, even topo sites strongly divergent in their primary structures may display common structural characteristics, conformationally significant for binding and cleavage.

A growing body of experimental data suggests that the function of specific sites in DNA is determined to a high degree by their conformational and physicochemical properties (Starr et al., 1995; Meierhans et al., 1997). Dependence of DNA conformation on the context was first discovered by X-ray analysis of DNA dodecamers (Dickerson, Drew, 1981). Increasing amount of structural data demonstrated heterogeneity of conformational and physicochemical properties of DNA, and their dependence on nucleotide sequence (Frank et al., 1997; Suzuki et al., 1997). Thus, the context-dependent local conformation of DNA molecules is one of the factors controlling specificities of DNA sites. Thus, the properties of DNA that are most important for binding and function of a protein might be refelcted in similar conformational and physicochemical parameters characterizing its specific sites. In other words, conservation of certain conformational and physicochemical characteristics should be observed for a variety of genomic sequences interacting with a certain enzyme. Two mechanisms may underlie the conservation of properties and should be considered when analyzing DNA sequences of functional and regulatory sites:

1) Nucleotide context conservation, when site variants contain a fixed set of nucleotides in defined positions. Functional sites of this type can be analyzed by context analysis techniques (Mulligan et al., 1984; Stormo et al., 1986);

2) Conservation of certain conformational and physicochemical properties within a definied region of a site, resulting from the restriction of the full set of possible nucleotide substitutions to those that fail to change these properties. Analysis of such DNA sites based on their conformational and physicochemical characteristics (Oshchepkov et al., this issue) can yield more detailed information on structural similarity of or difference between DNA regions corresponding to the sites hydrolyzed by topo.

For analyzing structural features, we have chosen the DNA sequences that are bound and hydrolyzed by topo with high efficiency. The results were compared with the data on X-ray structural analysis of the enzyme (Stewart et al., 1998) and thermodynamic and kinetic analysis of topo action ([1]Bugreev et al., 2002; [2]Bugreev et al., 2002; Bugreev et al., in press).

In the present work, a method based on determination of the conservative context-dependent conformational and physicochemical properties over short regions of functional DNA sequences was used. Dispersion (variance) of conformational and physicochemical properties from the PROPERTY database (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/) was chosen as a measure of conservation at individual positions of a phased sample of functional sites.

Materials and Methods

A sample of 45 human topo binding sites 32 bp long was used for the analysis. This sample was extracted from the SAMPLES database (http://wwwmgs.bionet.nsc.ru/mgs/dbases/nsamples/), phased according to the DNA cleavage point, selected for actual experimental confirmation of binding and cleavage, and filtered for presence of homologous sequences.

When analyzing the samples of aligned nucleotide sequences, 38 conformational and physicochemical DNA properties from the PROPERTY database (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/) were used.

The essence of our approach is in the following. For a sample of N aligned (phased) DNA sequences of the length L, a value of a certain physicochemical or conformational property f is ascribed to every dinucleotide. A matrix of size $Nx(L-1)$ is constructed as a result. The element $f(k,l)$ of the matrix corresponds to the value of the property f of dinucleotide at the position l of the sequence k.

The average value of the property i at position l is

$$\overline{F_{il}} = \frac{1}{N}\sum_{k=1}^{N} F_{ikl}\,, \tag{1}$$

Variance is used as a measure of conservation of the ith property for every position within l:

$$\sigma^2{}_{F_{il}} = \frac{1}{N-1}\sum_{k=1}^{N}(F_{ikl} - \overline{F}_{il})^2\,, \tag{2}$$

It is assumed that if a particular property within a certain sequence region is important for the function of this active site, then the value of this property is conserved for all the sequences of the sample, generating lower variance compared with the sample of random sequences. Thus, a low variance of a property indicates its conservation in a particular position. Significance of σ_{F_i} is estimated by χ^2 test.

Results and Discussion

A number of common specific conformational features of the sequences of the constructed sample was detected by testing all 38 properties for their conservation. Significantly close values were found for several DNA helix parameters, such as roll (averaged for the X-ray centers; Gorin et al., 1995) of dinucleotides at positions 0…–1 (i.e. directly at the cleavage site); slide (averaged for the X-ray centers; Gorin et al., 1995); twist (regressed for X-rays; Shpigelman et al., 1993); and rise (averaged; Ponomarenko et al., 1997), the last three also at cleavage sites. In addition, significantly close values of the parameter characterizing the occurrence of sterically disadvantageous contacts between N3 and NH_2 of guanines and N3 of adenines were detected (the property "Minor groove distance"; Gorin et al., 1995). This property corresponds to the presence of Py–Pu sequences at the cleavage site. Moreover, a decrease in the double helix melting temperature is observed in the cleavage site region of all the sequences interacting with topo (Hogan, Austin, 1987).

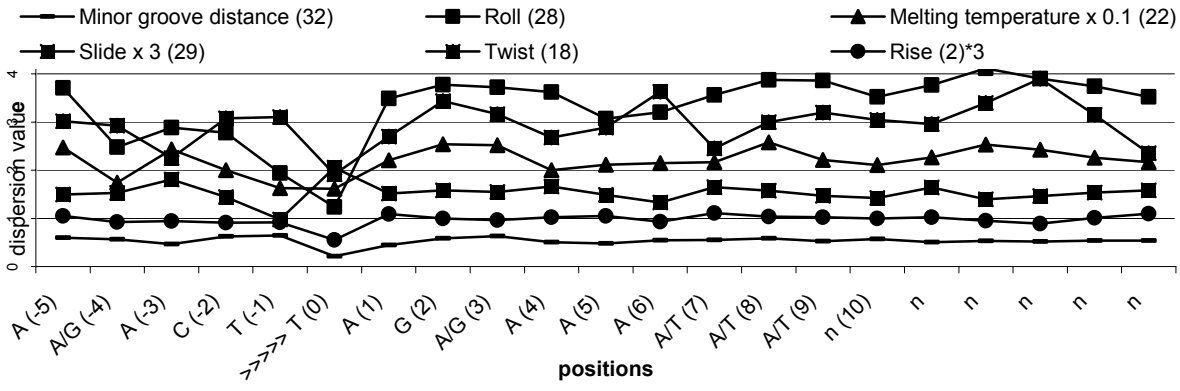


Fig. 1. Distribution of dispersion σ_{F_i} of the properties that displayed significantly lower values of a given parameter at certain positions of the sample. Numbers of properties (in parentheses) correspond to their numbers in the PROPERTY database.

Topo simultaneously contacts 10 nucleotide pairs of the duplex. However, the analysis of the sequence sample has demonstrated that the conformation properties show specificity mainly in the DNA region located directly at the cleavage site. The observed specificity of structural characteristics at the cleavage site is nonrandom and suggests that a set of particular parameters of this region may be important for both DNA recognition and subsequent conformation change and catalytic cleavage.

As evident from X-ray data, topo indeed forms a specific contact at the cleavage site (Stewart et al., 1998), which, according to thermodynamic analysis results, together with the contact of the enzyme with the nearest phosphate group provides a significant factor of enzyme affinity for the specific sequence (Bugreev et al., 2002) Therefore, rise and twist of the helix may provide for optimal interaction of the DNA region with topo, since they determine the spatial arrangement of enzyme amino acid residues relative to specific groups of the duplex bound in the topo DNA-recognizing center. The roles of other parameters may be reduced to facilitating the necessary conformational rearrangements of the enzyme and DNA during their mutual adaptation, exceptionally important for the cleavage rate.

Studies of topo recognition and catalytic mechanisms have demonstrated that the DNA-binding center of the enzyme is bent, most sharply at the cleavage site region (Fig. 2d; Bugreev et al., 2002; Bugreev et al., 2002). We suggest that this is required for DNA to be bent after binding the enzyme, correctly orienting the reacting groups at the active center and increasing DNA binding efficacy due to additional contacts with the enzyme's linker domain (Bugreev et al., in press).

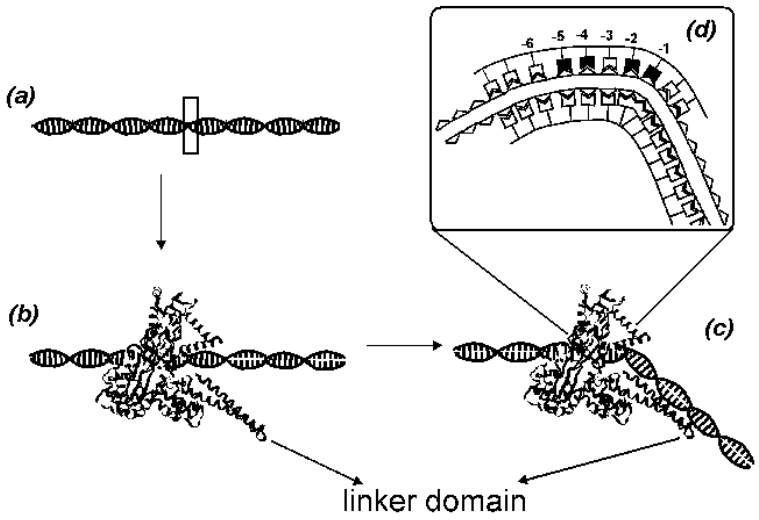


Fig. 2. DNA conformation changes after binding topo I: (a) a specific DNA conformational feature; (b, c) DNA bending at the cleavage site; (d), scheme of the enzyme DNA-recognizing center. In this figure, triangles stand for internucleoside DNA phosphate groups; and rectangles contacting with them, for amino acid residues at the DNA-recognizing center of the enzyme. The DNA-binding center of the enzyme is bent with the most pronounced bend at the cleavage site.

DNA bends are characterized by several structural features. Axes of the helices adjacent to the bent region are shifted relative to each other, stacking disruption occurs, and the neighbor base pairs at the bend are inclined relative to each other (Zaenger, 1984). Evidently, such conformational changes will proceed easier in the DNA regions displaying certain similar structural properties. From this point of view, an increase in slide, a local decrease in melting temperature and minor groove distance are exactly the parameters meeting these requirements. In addition, high conservation of the minor groove distance facilitates bending towards the major groove, providing correct interaction of the corresponding DNA region with the topo linker domain.

Thus, despite the optimal specific sequence for topo (Wang, 1996) contains a specific set of base pairs and, therefore, has a certain unique conformation, the sequence and full conformation does not appear to be critical for DNA binding and cleavage. Our calculations suggest that the specific structural features of the DNA helix at the cleavage site, presumably allowing it the DNA conformation to be readily changed in the direction specified by the enzyme, represent the major limitation for the DNA activity as a substrate. The dependence of DNA specificity on the structure of its rather short region indicates that the enzyme selects the DNA sequence to be cleaved mainly at the stages of conformational rearrangements and catalysis, since the difference between the efficiencies of binding to specific and nonspecific DNA sequences does not exceed one–two orders of magnitude (Bugreev et al., 1999). However, the optimal conformation of the other specific sequence regions provides more stable binding as well as the 100–1000-fold increase in the reaction rate.

Resume

Conformational features of human topoisomerase I cleavage sites were analyzed. The detected specificity of the structure characteristics at the cleavage site is shown to be nonrandom. This suggests that a set of certain parameters of this region is important for both DNA recognition and the subsequent conformational rearrangement and catalytic cleavage.

References

1. Bugreev D.V., Nevinsky G.A. (1999) Possibilities of the method of step-by-step complication of the ligand structure in the investigation of the protein-nucleic acid interaction: The mechanisms of functioning of some repair, replication, topoisomerization, and restriction enzymes. Biokhimiya (Mosk.). 64, 291-305.
1. [1]Bugreev D.V., Buneva V.N., Sinitsyna O.I., Nevinsky G.A. The mechanism of recognition of supercoiled DNA by eukaryotic type I topoisomerases. I. Interaction of enzymes with nonspecific oligonucleotides. 2002). Bioorg. Khim., in press.
3. [2]Bugreev D.V., Buneva V.N., Sinitsyna O.I., Nevinsky G.A. (2002) The mechanism of recognition of supercoiled DNA by eukaryotic type I topoisomerases. II. Comparison of interaction of enzymes with specific and nonspecific oligonucleotides. Bioorg. Khim., in press.
4. Bugreev D.V., Buneva V.N., Nevinsky G.A. The mechanism of specific cleavage of supercoiled DNA by human topoisomerases I: ligand structure influence on catalytic stage of reaction. Mol. Biol. (Mosk.), in press.
5. Gorin A.A., Zhurkin V.B., Olson W.K. (1995) B-DNA twisting correlates with base-pair morphology. J. Mol. Biol. 247, 34-48.
6. Hogan M.E., Austin R.H. (1987) Importance of DNA stiffness in protein-DNA binding specificity. Nature. 329, 263-266.
7. Frank D.E., Saecker R.M., Bond J.P., Capp M.W., Tsodikov O.V., Melcher S.E., Levandoski M.M., Record M.T.Jr. (1997) Thermodynamics of the interactions of Lac repressor with variants of the symmetric Lac operator: effects of converting a consensus site to a non-specific site. J. Mol. Biol. 267, 1186-1206.
8. Meierhans D., Sieber M., Allemann R.K. (1997) High affinity binding of MEF-2C correlates with DNA bending. Nucl. Acids Res. 25, 4537-4544.
9. Mulligan M.E., Hawley D.K., Entriken R., McClure W.R. (1984) Escherichia Coli promoter sequences predict in vitro RNA polymerase selectivity. Nucl. Acids Res. 12, 789-800.
10. Oshchepkov D.Yu., Turnaev I.I., Vityaev E.E. Study of the context-dependent conformational and physicochemical properties of DNA functional sites (this issue).
11. Perez-Stable C., Shen C.C., Shen C-K.J. (1988) Enrichment and depletion of HeLa topoisomerase I recognition sites among specific types of DNA elements. Nucl. Acids Res. 16, 7975-7993.
12. Ponomarenko M.P., Ponomarenko Iu.V., Kel' A.E., Kolchanov N.A., Karas H., Wingender E., Sklenar H. (1997) Computer analysis of conformational features of the eukaryotic TATA-box DNA promotors. Mol. Biol. (Mosk.). 31, 733-740.
13. Saenger W. Principles of Nucleic Acid Structure. Springer-Verlag New York Inc., 1984.
14. Shpigelman E.S., Trifonov E.N., Bolshoy A. (1993) CURVATURE: software for the analysis of curved DNA. Comput. Appl. Biosci. 9, 435-440.
15. Starr D.B., Hoopes B.C., Hawley D.K. (1995) DNA bending is an important component of site-specific recognition by the TATA binding protein. J. Mol. Biol. 250, 434 –446.
16. Stewart L., Redinbo M.R., Qiu X., Hol W.G., Hampoux J.J. (1998) A model for the mechanism of human topoisomerase I. Science. 279, 1534-1541.
17. Stormo G.D., Schneider T.D., Gild L. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. Nucl. Acids Res. 14, 6661-6679.
18. Suzuki M., Amano N., Kakinuma J., Tateno M. (1997) Use of 3D structure data for understanding sequence – dependent conformational aspects of DNA. J. Mol. Biol. 274, 421-435.
19. Tsui S., Anderson M.E., Tegtmeyer P. (1989) Topoisomerase I sites cluster asymmetrically at the ends of the simian virus 40 core origin of replication. J. Virol. 63, 5175-5183.
20. Wang J.C. (1996) DNA topoisomerases. Ann. Rev. Biochem. 65, 635-692.

FOUR-NUCLEOTIDE-RULE. VIRAL GENOMES

[1] Kramskova Zh.D., [2] Ivashchenko T.A., [1] Ivashchenko A.T.*

[1] al-Farabi's Kazakh National University, Almaty, 480078, Kazakhstan, e-mail: bbpp.kafedra@mailkazsu.uni.sci.kz
[2] M.A.Ajtkhozhin's Institute of Molecular Biology and Biochemistry, Almaty, 480012, Kazakhstan, e-mail: timour@itte.kz
[*]Corresponding author

Key words: *viral genomes, nucleic acids, nucleotide composition*

Resume

Motivation: Nucleotide frequencies in RNA and DNA viral genomes vary in wide ranges. It seems important to reveal at what extent these variations follow the nucleotide usage regularity discovered on the DNA of prokaryotic and eukaryotic genomes.

Results: The variability of nucleotide frequencies in DNA and RNA of viral genomes conforms to the following regularity: the nucleotide ratios difference is proportional to the purine nucleotide content in single-stranded nucleic acid: fA/fT(or fU) – fC/fG = **a** (fA + fG – 0.5) + **b**, where **a** and **b** – linear regression parameters. Established regularity (four-nucleotide-rule) is true for all 56 RNA- and 36 DNA-containing viral genomes studied. The linear regression parameters variability demonstrated with the model DNAs.

Available: The C++ source code is available free by the request to the author: timour@itte.kz.

Introduction

The significant heterogeneity of nucleotide composition in DNA stretches reflects the different functions they play in a cell (Karling et al., 1994). Despite of the heterogeneity, the nucleotide composition in DNA obeys the regularity – the four-nucleotide-rule (FNR). It consists in: the complementary nucleotide ratios difference is proportional to the purine nucleotide content in single-stranded nucleic acid: fA/fT(or fU) –fC/fG = **a** (fR–0.5) + **b**, where **a** and **b** – linear regression parameters, fR=fA+fG. FNR describes the nucleotide composition variability in different sites of chloroplast, mitochondrial and bacterial DNAs rather well (Ivashchenko et al., 1999, 2000). It seems very important to test this statement on the RNA- and DNA-containing genomes of viruses.

Methods

Nucleotide sequences have been downloaded from the GenBank database using Entrez search system (www.ncbi.nlm.nih.gov/Entrez). To analyze the nucleotide frequencies (fA, fT, fU, fG, fC) these genome sequences have been divided into pieces with the length of 300 n. or 3000 n. in such a way to get the reliability of the correlation coefficient not less than 0.001. We extracted protein coding gene sequences (CDSs) from several viral genomes. From the intron-containing genes we have formed a sampling of introns. The nucleotide sequences between the adjacent CDSs called by us as intergenic regions longer than 100 n. have formed a separate sampling.

Implementation and Results

For the complete genomes the linear regression parameters are shown in Table. The data presented show significant heterogeneity of regression parameters and high correlation coefficient. The majority of RNA- and DNA-containing viruses studied have **b** parameter close to zero, while **a** parameter varies from 4,89 to 13,2. In order to reveal the reasons of regression parameters variability we considered model DNAs with different nucleotide frequency ratios. For model DNAs with the equal ratio of fT and fG and changing ratio of fA and fC, corresponding FNR dependencies are shown in Fig. a. In the Figure **y**–axis is (fA/fU–fC/fG) and **x**-axis is (fR–0.5). It is seen from these dependences that in a DNA with such nucleotide frequency ratios **a** parameter depends on fT and fG while **b** parameter is zero: 1 - fA/fT–fC/fG=5.71(fR–0.5) with fT=fG=0.350; 2 - fA/fT–fC/fG=6.67(fR–0.5) with fT=fG=0.300; 3 - fA/fT–fC/fG=8,00(fR–0.5) with fT=fG=0.250; 4 - fA/fT–fC/fG=10.0(fR–0.5) with fT=fG=0.200; 5 - fA/fT–fC/fG=13.3(fR–0.5) with fT=fG=0.150. Such FNR dependences (with **b**=0) are observed for model DNAs with different fixed ratios of fA and fC and varying ratios of fT and fG. These model DNAs are complementary to DNAs for which the FNR dependences are shown in Fig. a, and the corresponding regression equations are: 1- fA/fT–fC/fG=47.6(fR–0.5) with fA=fC=0.350; 2 - fA/fT–fC/fG=18.7(fR–0.5) with fA=fC=0.300; 3 - fA/fT–fC/fG=9.07(fR–0.5) with fA=fC=0.250; 4 - fA/fT–fC/fG =4.86(fR–0.5) with fA=fC=0.200; 5- fA/fT–fC/fG=2.60(fR–0.5) with fA=fC=0.150. High value of **a** parameter in original DNA strand turns into lower one in the complementary strand. For instance, to models, shown in Fig. a the DNA of the following genomes corresponds well: 1 - Diaporthe ambigua RNA virus 1 (fT=fG), 2 - Hepatitis C virus (fG=fC), 3 –Sagiyama virus (fA=fG). Viral DNAs with **b** parameter different from zero can be modeled by the nucleotide sequences with fixed but not equal ratio of fA and fC with varying ratio of fT and fG. It been known that if fA higher than fC, the **b** is above zero, while if fA lower than fC, the **b**

parameter is negative (Fig. c: 1 – fA=0.400 and fC=0.200; 2 – fG=0.200 and fT=0.400; 3 – fA=0.200 and fC=0.400; 4 – fT=0.200 and fG=0.400). The corresponding equations are as following: 1 - fA/fT–fC/fG =16.3(fR–0.5)-0.584; 2 - fA/fT–fC/fG =7.5(fR–0.5)+0.250; 3 - fA/fT–fC/fG =18.7(fR–0.5)+0.7343; 4 - fA/fT–fC/fG =7.5(fR–0.5) –0.250, for instance, for some DNA-viruses (1 - Human immunodeficiency virus, 2 - TTV-like mini virus, 3 - Chyote mosaic tymovirus, 4 - Human T-cell lymphotropic virus) shown in Fig. d. Thus, the difference in **a** and **b** parameters for the RNA and DNA of different viruses is determined by non-equality of nucleotide frequency mean values of at least 2 nucleotides.

Table. Linear regression parameters (**a, b)** and correlation coefficient (**r**) between the nucleotide frequency ratios difference and purine nucleotide content in RNA- and DNA-containing viruses.

Object	GenBank code	Length, n.	Regression parameters		r
			a	b	
RNA-viruses					
Murine hepatitis virus	NC_001846	31357	6.11	0.075	0.96
Ebola virus	AF272001	18959	10.75	-0.042	0.91
Hendra virus	NC_001906	18234	9.23	-0.045	0.99
Bovine parainfluenzea virus 3	NC_002161	15456	13.59	-0.265	0.85
Mups virus	AF280799	15384	10.28	-0.058	0.97
Sendai virus	AB039658	15384	9.57	-0.031	0.96
West Nile virus	AF196835	11029	8.93	-0.024	0.97
Japanese yam mosaic virus	NC_000947	9760	10.20	-0.118	0.98
Rous sarcoma virus	J02342	9625	8.35	0.005	0.97
Hepatitis C virus	AB049088	9616	6.26	-0.005	0.95
Soybean mosaic virus N	D00507	9588	8.64	-0.030	0.99
Potato virus A	NC_001649	9585	9.80	-0.073	0.97
Sacbrood virus	NC_002066	8832	6.63	0.065	0.98
Human immunodeficiency virus	AY008718	8806	10.13	-0.126	0.97
Garlic virus B	NC_001800	8106	12.86	-0.080	0.95
Avian leukosis virus	NC_001408	7286	7.78	-0.002	0.99
Sheep astrovirus	NC_002469	6640	7.24	0.017	0.99
Chyote mosaic tymovirus	NC_002588	6364	15.53	0.560	0.94
Poinsettia mosaic virus	NC_002161	6099	10.80	0.100	0.97
Tomato bushy stant virus	TBU80935	4776	7.18	0.017	0.99
Melon necrotic spot virus	NC_001504	4266	6.83	0.040	0.99
Sesbania mosaic virus	NC_002568	4149	6.90	-0.008	0.99
Diaporthe ambigua RNA virus1	NC_001278	4113	4.64	-0.045	0.98
Cockafoot mottle virus	NC_002618	4083	7.94	0.011	0.98
Olive latent virus 1	NC_001721	3699	8.24	-0.014	0.99
DNA-viruses					
Neisseria meningitidis	NMA2Z2491	349061	7.47	-0.005	0.97
Fowlopox virus	NC_002188	288539	7.91	0.003	0.99
Bacteriophage T4	AF158101	168897	8.45	0.034	0.99
Human herpesvirus 2	NC_001798	154746	7.81	0.007	0.93
Human herpesvirus 1	NC_001806	152261	7.87	-0.003	0.96
Gallid herpesvirus 2	NC_002229	138675	8.22	-0.004	0.98
Bovine herpesvirus	NC_001847	135301	7.63	-0.022	0.95
Bacteriophage lambda	J02459	48502	8.52	0.006	0.98
Frog adenovirus 1	NC_002501	26163	7.06	0.050	0.96
Bovine ephemeral fever virus	NC_002526	14900	8.72	-0.037	0.99
Human T-cell lympotropic virus	NC_001877	8960	13.12	0.070	0.92
Bovian leukemia virus	AF257515	8588	9.91	0.042	0.98
Papio cynocephalus provirus	NC_000863	8572	11.73	-0.030	0.98
Soybean chlorotyc mottle virus	NC_001739	8178	10.66	0.188	0.95
Polyomavirus	NC_001515	5297	7.60	0.036	0.99
Simian virus 40	NC_001669	5243	10.36	-0.067	0.93
JC virus	AF295731	4854	8.68	-0.043	0.97
TT virus	AB041958	3798	8.93	0.050	0.98
Ground squirrel hepatitis virus	K02715	3311	6.79	-0.008	0.95
Hepatitis B virus	AF305327	3182	8.18	0.003	0.96
Beet curly top virus	U56975	2930	6.69	0.062	0.99
TTV-like mini virus	NC_002195	2915	13.97	-0.389	0.98
Sugarcane streak virus	NC_000897	2735	7.87	0.018	0.98

Footnote. The genomes larger than 100000 n. have been divided into pieces of 3000 n. long, and those shorter - into 300 n. long.

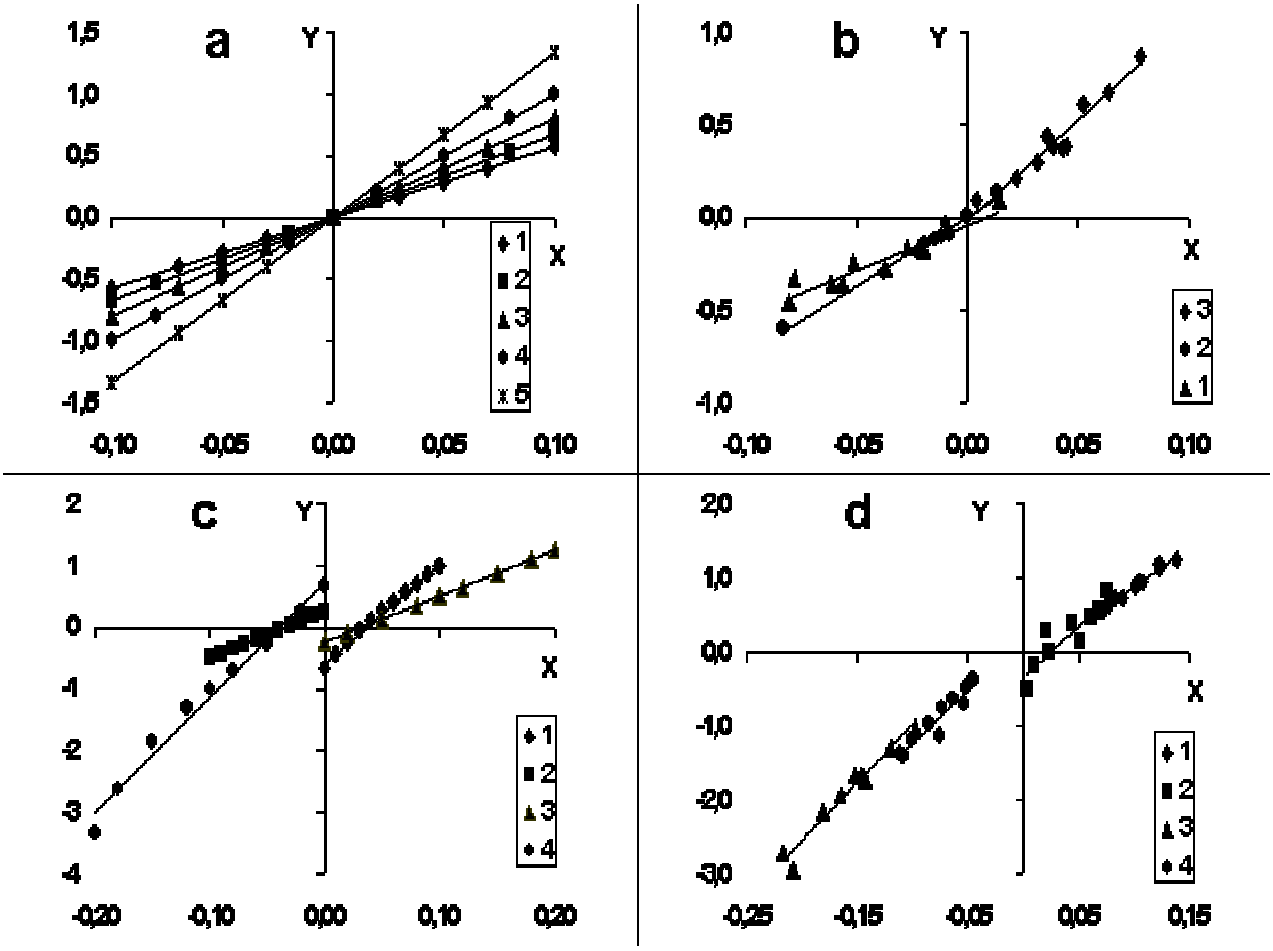


Fig. Interdependencies of nucleotide frequencies for viral genomes and model DNAs with different ratios of fA, fT, fG, fC. The designation is in the text.

The nucleotide frequencies of CDSs, introns, intergenic regions of T4 phage and other viral genomes follow the FNR. The results of this study report that four-nucleotide-rule can be applied to the genomic nucleotide sequences of all organism types, starting from viruses and ending with eukaryots. The rule works also rather well for short nucleotide sequences, for instance tRNA genes (70-90 n.) and short intergenic spacers (Ivashchenko et al., 2000).

Refrences

1. Ivashchenko T.A., Ivashchenko A.T., Aitkhozhina N.A. (1999) Regularity of nucleotides usage in DNA. II. Peculiarities of chloroplast genomes of *Pinus thunbergii, Nicotiana tabacum, Orysa sativum* and *Zea mays*. Biotechnology. Theory and practice. 11-12, 103-109.
2. Ivashchenko T.A., Kurmasheva R.T., Ivashchenko A. T. (2000) Regularity of nucleotides usage in DNA. IV. Four-nucleotide-rule in DNA of bacteria. Biotechnology. Theory and practice. 13, 142-145.
3. Karling S., Ladunga I., Blaisdell B.E. (1994) Heterogeneity of genomes: measures and values. Proc. Natl Acad. Sci. USA. 91, 12837-12841.

IDENTIFICATION OF FOUR GENES ON HUMAN CHROMOSOME 3 HOMOLOGOUS TO THE KNOWN GENES ON OTHER CHROMOSOMES BY *IN SILICO* ANALYSIS

**** Rakhmanaliev E.R., Klimov E.A.***

Vavilov Institute of General Genetics, RAS, Moscow, Russia, e-mail: elian-rr@newmail.ru
*Corresponding author

Key words: *NotI-clones, NotI-STSs, gene-markers, gene homologs*

Resume

Motivation: In this work, we analyzed the structure of genomic sequences of human chromosome 3, marked by *Not*I-STSs. The *Not*I-STSs had homology with genes localized on other chromosomes that allowed us to suppose the presence of homologous sequences for these genes on chromosome 3.

Results: Four nucleotide sequences of human chromosome 3, marked by *Not*I-STSs (NB1-100, NL3-004, NLM-246 and NRL-402) which have high homology with genes, earlier localized in other genomic regions, were characterized *in silico*. We have shown that *RINZF* gene earlier localized on 8q13-q21.1 has the full-length copy on human chromosome 3. For three *Not*I-STSs (NL3-004, NLM-246 and NRL-402), which were markers of genes *LOC132160*, *ATP11B* and *ITGA9* on chromosome 3, genes *KIAA1157* (12q14.1), *HSA9947* (1p36) and *SCYA5* (17q11.2-q12) were determined as having homology to the *Not*I-STS sequences, respectively. Similarity of regulatory regions for three pairs of genes, (*LOC132160* / *KIAA1157*, *ATP11B* / *HSA9947* and *ITGA9* / *SCYA5*), marked by the *Not*I-STSs, was shown.

Introduction

The endonuclease *Not*I restriction sites (5'-GCGGCCGC-3') are located in CpG-island, which are associated with 5'-UTR of genes. Therefore, STSs (sequenced tagged site), created on base *Not*I-sites might be considered as the universal markers of genes. Library of *Not*I-clones of human chromosomes 3 was created earlier (Zabarovsky et al., 1990; 1996). In our laboratory, 113 *Not*I-STSs for 84 *Not*I-clones were created (Sulimova et al., 1999). We have determined the physical localization of 30 *Not*I-STSs by radiation hybrid mapping method and constructed *Not*I-map of human chromosome 3, including 60 *Not*I-STSs (data in press). The search of homologies for the localized *Not*I-clone sequences with corresponding nucleotide sequences, presented in public databases (GenBank, EMBL and TIGR) by the program BLAST has revealed the high level of associations (91,7%) of *Not*I-STSs with human genes or ESTs. The localization of the majority *Not*I-STSs and corresponding homologous genes and ESTs completely coincided. However, four genes significant homology with *Not*I-STSs, were earlier localized on other human chromosomes. These results allowed us to suppose the presence of homologs (or pseudogenes) of these genes on human chromosome 3. To confirm these suggestions, we performed *in silico* analysis of genomic sequences of human chromosome 3, adjacent to sites of *Not*I-STSs localization.

Methods

The homologies were searched by the BLAST-program provided by NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). Exon-intron structure of novel genes was created by BLASTN (NCBI) and GENSCAN (http://genes.mit.edu/GENSCAN.html) programs. Promoter regions were identified using PromoterInspector (http://genomatix.gsf.de/cgi-bin/promoterinspector/promoterinspector.pl) and Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html) programs. We also used programs at GeneBee server (http://genebee.msu.ru/) for the search of amino acid homologies and construction of the full local similarity maps for hypothetical proteins. Information concerning proteins, encoded by novel genes, was obtained from OMIM database (http://www.ncbi.nlm.nih.gov/OMIM/).

Results and Discussion

Screening of human genomic sequences for homologous sequences to earlier RH-mapped *Not*I-STSs was revealed. The data allowed us to identify four nucleotide sequences on human chromosome 3, homologous to the genes previously localized in other genome regions (Table). It allowed us to suggest presence of four earlier non-described gene-homologs (or pseudogenes) on chromosome 3.

The *Not*I-STS NB1-100 has 99% homology with mRNA *RINZF* gene, encoding protein with yet unknown function, containing “zinc fingers” domain. The *RINZF* gene was earlier localized on chromosome 8. However, marker NB1-100 also has 99% homology with fragment of clone AC009812, localized on chromosome 3. The comparative analysis of *RINZF* gene and clone AC009812 has revealed that clone AC009812 includes nucleotide sequence, identical to the

sequence of *RINZF* gene. The exon-intron structure of a copy of the *RINZF* gene constructed by BLAST and GENSCAN programs was identical (in the number, length and nucleotide sequence of exons and introns) the exon-intron *RINZF* gene structure, predicted by computer analysis performed with the BLAST program and presented in database MapViewer. The gene identical to *RINZF* gene has all necessary regulatory elements present in any gene: TATA-box, promoter region, poly(A)-sites and splicing sites. Therefore we can suggest that full-length copies of *RINZF* gene are present on both human chromosome 3 and 8. This is not an artifact, since localization of *Not*I-STS NB1-100 and localization of homologous clone AC005812 on chromosome 3 completely coincided.

Table. Localization of *Not*I-STSs and genes homologous to the *Not*I-STSs.

*Not*I-STS	*Not*I-STS localization		Gene on chromosome 3 marked by *Not*I-STS	Detected gene-homolog	Localization of gene-homolog
	position on GM99'-GB4 in cR_{3000}	cytogenetic localization			
NB1-100	133.5	3p21.33	-*	Human zinc finger protein RINZF (*RINZF*)	8q13-q21.1
NL3-004	189.0	3p21.1	Human hypothetical gene *LOC132160*	Human gene for KIAA1157 protein (*KIAA1157*)	12q14.1
NLM-246	687.0	3q27.2	ATPase, Class VI, type 11B (*ATP11B*)	Human putative ATPase gene (*HSA9947*)	1p36
NRL-402	127.1	3p21.3	Human integrin, alpha 9 gene (*ITGA9*)	Human small inducible cytokine A5 (RANTES) gene (*SCYA5*)	17q11.2-q12

Footnote: *We detected full-length copy of *RINZF* gene from chromosome 8, which yet untitled.

The hypothetical gene *LOC132160* (4,5 kb in length) is present on chromosome 3, near the site of localization of *Not*I-STS NL3-004. Marker NL3-004 has homology with a fragment of the hypothetical gene *KIAA1157* (290 kb) (chromosome 12). These genes (*LOC132160* and *KIAA1157*) have different exon-intron structure and cDNA sequences. In spite of these differences, genes *LOC132160* and *KIAA1157* encoded similar proteins (identity 55%). Using the protein-protein BLAST, we revealed the nearest homolog for these proteins – protein phosphatase 2C (identity 85% for *LOC132160* and 54% for *KIAA1157*). Probably, two considered genes encoded proteins, which referred to the class of proteins PP2C, from serin/threonin phosphatases family (Marley et al., 1998). The products of these genes can be involved in the same metabolic way. Therefore similarity of promoter regions of these genes, might be due to probable similar regulation at the stages of initiations of transcriptions.

Analogous results were received for *Not*I-STS NLM-246, localized in 5'UTR of ATPase gene, Class VI, Type 11B (*ATP11B*) (chromosome 3). The gene has also homology with the hypothetical gene *HSA9947* (chromosome 1). Between the genes, no reliable homology was detected on nucleotide level. However, both genes encoded proteins (ATPases), from one protein family.

The *Not*I-STS NRL-402 is a marker of the gene integrin (*ITGA9*), localized on chromosome 3, and has homology with the gene chemokine (*SCYA5*) localized on chromosome 17. Between these genes, no homology was observed on nucleotide or protein levels. Homology of NRL-402 with gene *SCYA5* is explained by the presence of conservative regions in 5'UTR of the genes.

Therefore, we characterized *in silico* four nucleotide sequences of human chromosome 3, marked by *Not*I-STSs (NB1-100, NL3-004, NLM-246 and NRL-402) which have high homology with genes, earlier localized in other genomic regions. We have shown, that earlier localized on 8q13-q21.1 gene *RINZF* has the full-length copy on human chromosome 3. For three *Not*I-STSs (NL3-004, NLM-246 and NRL-402), which were marked of genes *LOC132160*, *ATP11B* and *ITGA9* on chromosome 3, genes *KIAA1157* (12q14.1), *HSA9947* (1p36) and *SCYA5* (17q11.2-q12) were determined as having homology to the *Not*I-STS sequences, respectively. Similarity of regulatory regions for three pairs of genes, (*LOC132160* / *KIAA1157*, *ATP11B* / *HSA9947* and *ITGA9* / *SCYA5*), marked by the *Not*I-STSs of chromosome 3, was shown.

Acknowledgements

This work was supported by "Human Genome Project" (project № 89'99) and RFBR (project № 00-15-97777). The authors are grateful to prof. Sulimova G.E. for helpful discussions.

References

1. Marley A.E., Kline A., Crabtree G., Sullivan J.E., Beri R.K. (1998) The cloning expression and tissue distribution of human PP2C-beta. FEBS Lett. 431, 121-124.
2. Sulimova G.E., Udina I.G., Kunizheva S.S., Kompaniitsev A.A. (1999) Creation of *Not*I-STS markers for human chromosome 3. Mol. Biol. (Mosk). 33, 791-796.
3. Zabarovsky E.R., Boldog F., Thompson T., Scanlon D., Winberg G., Marcsek Z., Erlansson R., Stanbridge E., Klein G., Sumegi J. (1990) Constraction of a human chromosome 3 specific *Not*I linking library using a novel cloning procedure. Nucl. Acids Res. 18, 6319-6324.

4. Zabarovsky E.R., Kashuba V.I., Gizatullin R.Z., Winberg G., Zabarovska V.I., Erlandsson R., Domninsky D.A., Bannikov V.M., Pokrovskaya E., Kholodnyuk I., Petrov N., Zakharyev V.M., Kisselev L.L., Klein G. (1996) *Not*I jumping and linking clones as a tool for genome mapping and analysis of chromosome rearrangements in different tumors. Cancer Detect Prev. 20, 1-10.

A MACROMOLECULAR MODELING AS A TOOL TO EXPAND BIOINFORMATICS DATABASES

Vorobjev Y.N.

Novosibirsk Institute of Bioorganic Chemistry, SB RAN, e-mail: ynvorob@niboch.nsc.ru

Key words*: protein stability, free energy calculation, molecular dynamics, DNA conformational dynamics*

Resume

Motivation: The development and use of efficient computational tools can enhance a completeness of bioinformatics databases and our understanding of molecular structure and relation between structure and function. A reliable macromolecular modeling can served as a tools to expand (experimental) bioinformatics data bases beyond the capacity of experimental methods and provide a new knowledge.

Results: an approach that combines explicit solvent (ES) and implicit solvent (IS) models has been developed for free energy calculation of protein conformations.

Availability: access to software can be negotiated. Mail to ynvorob@niboch.nsc.ru

Introduction

A recent progress in the macromolecular modeling methods gives more evidence that the advanced modeling methods can be used to obtain a simulated bioinformatics data to fill up 'gapes' in the data bases. When properly designed and used, these tools can aid experimental scientist in more accurate three dimensional (3D) molecular structure determination and refinement and reveal structure details and internal conformational dynamics at different thermodynamic conditions well beyond the capacity of experimental methods. There are several fields for a productive application of the macromolecular modeling methods:

- 3D *structure reconstruction* of the missing structural information for a some elements of the macromolecule, i.e. coordinates of atoms which are undefined in the pdb data base, e.g. flexible loops on a protein surface.
- 3D *structure determination* using conformational restraints based on experimental measurements.
- Internal *conformational dynamics* due to thermal fluctuations.
- 3D *structure prediction* - "protein folding" and "nucleic acid folding" problems and stability of different conformations.

Macromolecular modeling methods can provide a tool to 'interpolate' and 'extrapolate' missing bioinformatics knowledge based on the available knowledge. Protein and nucleic acid conformation, protein-ligand, protein-protein and protein-nucleic acid binding are guided by the total free energy of the system, and respective forces, in solvent. Therefore the importance of being able to estimate accurate free energies is obvious from the applications. However, the development of an accurate and fast method to calculate the free energy of a macromolecular complex in solution on the basis of strictly statistical-mechanical principles is a task of great complexity.

Methods

We have developed an approach that combines explicit solvent (ES) and implicit solvent (IS) models. The key to success is to obtain accurate representative conformations by a simulation with explicit solvent model, and then to estimate the free energy of the protein-solvent interactions with an implicit solvent model. The ES/IS method suggests a consistent approach taking advantage of a simulations of a set of microscopic structures which are compatible with the molecular structure of an aqueous solution and reasonably reliable physical implicit model of solvent for calculation of a solvation free energy (Vorobjev et al., 1998, Vorobjev, Hermans, 2001). The implicit solvation model includes an empirical parameters, born radii to define a molecular surface and dielectric surface interface. A consistent calibration of the radii parameters of the implicit model of the ES/IS method, with free energy calculations by a free energy perturbation method for a reference data base molecules has been done to obtain the force-field consistent ES/IS method and evaluate the born radii set for the GROMOS force field. The free energy F_A of a solute molecule in macroscopic conformation A

$$F_A = \langle U_m(\mathbf{x}) \rangle_A - TS_{conf,A} + \langle W(\mathbf{x}) \rangle_A \qquad (1)$$

where $\langle\rangle_A$ denotes an average over micro-configurations of the conformation A, U_m represents the intra-protein conformational energy and S_A is the entropy of the conformation A. In the ES/IS method, a representative set of microscopic configurations $\mathbf{x}_{A,i}$ of a solute in a solvent is generated by MD simulation with explicit solvent along a

relatively short trajectory (50-100 ps). The solvation free energy W(**x**) can be written as a sum of terms for cavity formation, solute-water van der Waals interactions and electrostatic polarization of solvent by the polar components of the solute. As a result, eq. (1) becomes

$$F_A = \langle U_{m,sh} \rangle_A + \langle U_{m,coul} \rangle_A - TS_{conf,A} + \langle G_{cav} \rangle_A + \langle G_{s,vdw} \rangle_A + \langle G_{pol} \rangle_A \quad (2)$$

where the intra molecular potential energy U_m has been represented as a sum of short-range (i.e. angle deformation and van der Waals) terms, $U_{m,sh}$ and electrostatic Coulombic interactions, $U_{m,coul}$. Of the six terms in eq. (2), three, namely $\langle U_{m,sh} \rangle$, $\langle U_{m,coul} \rangle$ and $\langle G_{s,vdw} \rangle$ are accumulated as averages during the molecular dynamics simulation: as will be discussed, the free energy, $G_{s,vdw}$ of van der Waals interactions between solute and solvent can be accurately approximated by the potential energy of these interactions

$$G_{s,vdw} = U_{s,vdw} \quad (3)$$

The entropic term, $TS_{conf,A}$ is estimated in the harmonic approximation from the covariance matrix of the positional fluctuations during the dynamics trajectory. This term pertains only to the conformational fluctuations of the solute molecule. The remaining two terms, the free energy for formation of the cavity G_{cav} and the free energy of solvent polarization, G_{pol}, which are difficult to simulate microscopically, are found with models in which the solvent is treated implicitly with an appropriate physical model, as a continuum or a dielectric continuum. As will be discussed, the free energy for formation of the cavity G_{cav} is well approximated with a term given by the product of the molecular surface and a microscopic surface tension and the free energy of solvent polarization, G_{pol} is found by modeling the solvent as a continuum dielectric, with Poisson's equation. A validity of the continuum dielectric approximation is analyzed via a simulation of a polarization free energy of charging by the slow charging method, and then an optimal set of atomic radii is derived for the ES/IS model by fitting the IS free energies of solvations with the free energies of microscopical simulations.

An advanced modeling method technique can be used as a computational tools to produce a new bioinformatical knowledge which can not be obtained via experimental methods. An example of this is a context dependent internal conformational dynamics of the DNA fragments. A protein binding sites of a genomic DNA should have an appropriate conformation and conformational mobility to effectively interact with cell proteins. Therefore the context-dependent conformational properties of the DNA sequence can serve as a natural set of the descriptors to characterize a probability of site to interact with proteins (Ponomarenko et al., 1999).

The available context-dependent static conformational parameters of DNA extracted for the experimental 3D structures of DNA fragment in crystal state is not complete and accurate due to a distortion effect of crystal packing and detectable conformational differences in a crystal and a solution for a some sequences (Dornberger et al., 1999). The context dependent conformational properties of the DNA sequences can be obtained in a solvent environment via the molecular dynamics simulations. The covariance fluctuation matrix **C** (Vorobjev et al., 1998).

$$\mathbf{C}_{ij} = \langle (\mathbf{x}_i - \langle \mathbf{x}_i \rangle)(\mathbf{x}_j - \langle \mathbf{x}_j \rangle) \rangle_A \quad (3)$$

where **x**(t) are the atomic coordinates along the MD trajectory. If the conformation is stable, then the molecular motion along the MD trajectory can be approximated as quasi-harmonic to give a statistical description of the real anharmonic intramolecular motions over a large number of microstates of the molecule. The eigenvectors defines normal modes and eigenvalues λ_i of the (mass-weighted) covariance matrix **C** define frequencies and positional fluctuations of the normal coordinates. Considering a set of short 10-12 base pares DNA fragments with all possible double and triple sequences at the center of duplex and random sequences at the ends of duplexes, a context dependent conformational parameters of pair, triplets etc., can be investigated. The spiral parameters Ω,τ,ρ (twist, tilt, roll) for each base-pair step of DNA duplex can be calculated from atomic coordinates **x**(t) and the average values and amplitudes of thermal fluctuation can be calculated. To obtain a statistically reliable values of conformational parameters, the MD simulations should be made until a convergence of the fluctuation matrix **C**. A required convergence can be obtained for a simulation of ~ 1 ns length. A reliability of MD simulations depends on computational method and force field.

Results

The ES/IS method's accuracy has been demonstrated by comparing hundreds of non-native conformations ("decoys") of a several proteins from Park&Levitt decoy set and models presented at CASP3 with their native structures, and finding that the latter had lower free energy for all proteins (Vorobjev, Hermans, 2001).

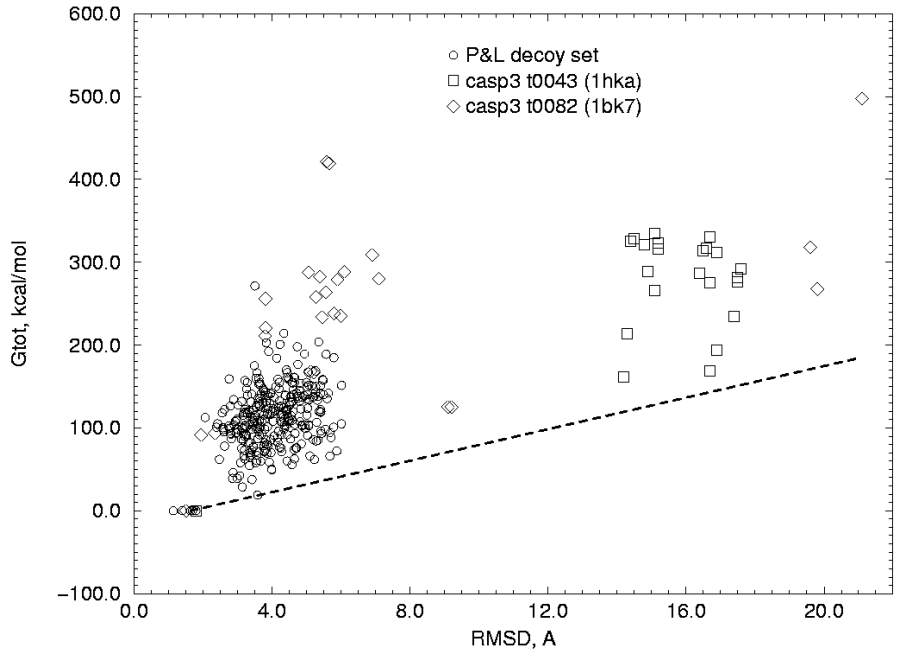


Fig. The total excess free energy of decoys versus RMSD from the respective native structure. Dashed line is the minimum discrimination line.

The required molecular dynamics simulations of proteins and DNA duplexes are done with the Gromos96 and Amber6.0 MD simulation packages. For a charged macromolecules a proper account of a long-range electrostatic interactions as the Ewald sum is very important for a stability and accuracy of the MD simulations. We investigate the Gromos96 and Amber6.0 MD simulation package on stability of long simulations. The test MD simulations for the DNA duplex show that the Amber6.0 package provide a stable MD nanosecond trajectory while the gromos96 force field gives less satisfactory results.

References

1. Dornberger U., Leijon M., Fritzsche H. (1999) Solution structure and base pare opening rates of GGCC containing oliginucleotides. J. Biomol. Struct. and Dynamics. 16: 1251.
2. Ponomarenko M.P., Pomonomarenko J.V., Frolov A.S., Podkolodny N.L., Savinkova L.K., Kolchanov N.A., Overton G.C. (1999) Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins. Bioinformatics. 15: 687.
3. Vorobjev Y.N., Almagro J.C., Hermans Jan. (1998) Discrimination between native and intentionally misfolded conformations of proteins: ES/IS a new method for calculating conformational free energy. Proteins: Struct. Funct. Gen. 32:399-413.
4. Vorobjev Y.N., Hermans J. (2001) Free energies of protein decoys provide insight into determinants of protein stability. Protein Sci. 10: 2498-2506.

LARGE PROPELLER DEFORMATIONS OF NUCLEOTIDE STEPS IN SHORT DNA DOUBLE HELIXES: QUANTUM-CHEMICAL MNDO/PM3 STUDY

Kabanov A.V., Komarov V.M., Yakushevich L.V., * Teplukhin A.V.

Institute of Cell Biophysics, RAS, 142290, Pushchino, Moscow Region, Russia
*Institute of Mathematical Problems of Biology, RAS, 142290 Pushchino, Moscow Region, Russia

Key words: *DNA structure, base pairs, internal polymorphism, MNDO/PM3 calculations, oligonucleotide duplexes, propeller twisting, pairs buckling, end effects, bounded water*

Stability of "propeller-like" and "step-like" forms of nitrous base pairs in the structure of elongating oligonucleotide duplexes is examined by using semiempirical MNDO/PM3 technique. We give a substantiation of the important role of the primordial non-coplanarity of Watson-Crick base pairs in the initiation of sequence dependence of DNA helix curve. The influence of end effects of double strands and of incorporated water on the process of nucleotides packing is discussed.

Intoduction

Estimation of the physical factors limiting the process of structure-function organization of DNA molecule, is a major problem of modern physico-chemical biology [1]. Theoretical studies of electronic structure of nucleic acids constituents are extremely actual. They permit to elucidate the specificity of nucleotide chains hydrogen bonding, their thermodynamics and mechanisms of nucleotide sequence dependence of the double helix secondary structure.

Many quantum-chemical searches for the relation between DNA structure and energetic parameters of coupling nitrous bases are known (see, for example, refs. [2-6] and references therein), and the hypothesis about stacking perturbation nature of deformations of planar base pairing in double helix is widely accepted. This conception, however, cannot ambiguously describe a sizeable distortion of hydrogen bonding of the pairs observed. For instance, in the structure of single crystals of nucleosides and nucleotides, "propeller" twisting of the bases reaches an angle of 49 degrees [7], and in the native as well as in the synthetic forms of DNA molecules, the "propeller" and "buckling" pair deformation remains also rather large [1, 8-10] resulting in a twisting angle as large as 33-39 degrees [10].

In the given theoretical study we discuss the appearance of "step-like" and "propeller-like" structures of the pairs with large buckling in short oligonucleotide duplexes as the result of packing of different polymorphic forms of complementary base H-pairing.

Early, we have theoretically shown [11-14] that non-coplanarity of complementary and hoogsteen pairs as well as the non-uniqueness of their H-coupling at the same hydrogen bonds (i.e. their hidden polymorphism) are probably internal, fundamental properties of single nitrous base pairs. They are initiated by bistable, sp^3-hybridized character of valence bonds of N-atom of the amino groups involved in bases H-binding.

Main goal of our computer simulating now is to show the crucial role of accumulation of the initial non-planarity of Watson-Crick nucleotide steps in the formation of nucleotide sequence dependence of DNA helix curve.

The influence of end effects of the chains and of bounded water molecules on the base packing are also considered.

Method

To examine the energetic and structural peculiarities of nucleotide packing in different types of dimer-, trimer- and tetramer-nucleotide duplexes we utilized well-known semiempirical quantum-chemical MNDO technique with PM3 approach [15,16] of MOPAC7.01 software. This, low-cost semiempirical method allows ones to estimate [17-21,28,29] (sometimes, unfortunately, to overestimate) the order of internal molecular energy barriers and other integral characteristics of electronic structure of different molecular systems.

The choice of the method is explained by the following reasons.

Application of high-level ab initio quantum chemistry methods for structural biological investigations is unfortunately restricted by rather small molecules due to very high computational resources demand. Usually the size of the calculated molecular complexes is no more than several tens of atoms [2-6, 21-27]. In our case the size of the duplexes under consideration is about 260 atoms. Therefore, the application of very popular, for instance, MP2/6-31G(d,p), technique to comprehensive analysis of the structural changes of elongating double helixes remains practically impossible even if one uses modern high-end computer systems.

In our theoretical simulations 3'-5' antiparallel double chains such as d(ApA).d(TpT), d(ApT).d(TpA), d(GpG).d(CpC), d(CpG).d(CpG), d(ApApA).d(TpTpT), d(TpApT).d(ApTpA), d(GpGpG).d(CpCpC), d(GpCpG).d(CpGpC), d(ApApApA).d(TpTpTpT), d(ApTpApT).d(TpApTpA), d(GpGpGpG).d(CpCpCpC) and d(GpCpGpC).d(CpGpCpG) were

designed as neutral species. The charge neutrality of sugar-phosphate backbone was modelled by placing protons (H^+) on appropriate anionic oxygen (=O^-) of the phosphate groups. The nucleotides were built in the *anti*-orientation about glycosyl $C_{1'}$ sugar-N base linkage.

"Supermolecule" approximation was used in all fully geometry-optimization calculations. All structures were considered in stationary states and in the true minima of potential energy surfaces. Reaching the optimal geometry was checked by the extrema of calculated heat of formation and by the lack of imaginary frequencies in the spectrum of normal modes of a complex.

The enthalpy of hydrogen bonding was calculated by comparing the heats of formation of the duplex with 1SCF heats of formation for separate chains "frozen" in their duplex-optimized geometry [17].

To simulate the influence of bound water on the base packing in sugar-phosphate double strands, we studied, as an example, the short duplex d(ApA).d(TpT) with uncompensated charge state of

$\Delta Ze = -2$. Two variants of aqueous "shell" containing greater and smaller number of water molecule were tested.

Results and Conclusions

The main findings of our calculations are illustrated in figures 1,2.

The results obtained show, that:

1) "propeller-like" specificity of base H-pairing with rater large buckling of the base planes, which is a characteristic of single Watson-Crick AT and GC pairs [11-14], remains the major factor of structure organization of short double helixes;

2) the end effects are fair for all isolated structures of the duplexes [14];

3) tetranucleotide double chain is the most ordered sequence with noticeable paralleling of the basis in the middle of a chain;

4) the appearance in the duplex structure of the bound water reduces the end effects of the chains due to water "cross-links" of the bases and improves their stack packing.

So, the internal peculiarities of AT and GC pairs non-planarity initiate the noticeable dependence of a mini-spiral secondary form on the nucleotide sequence.

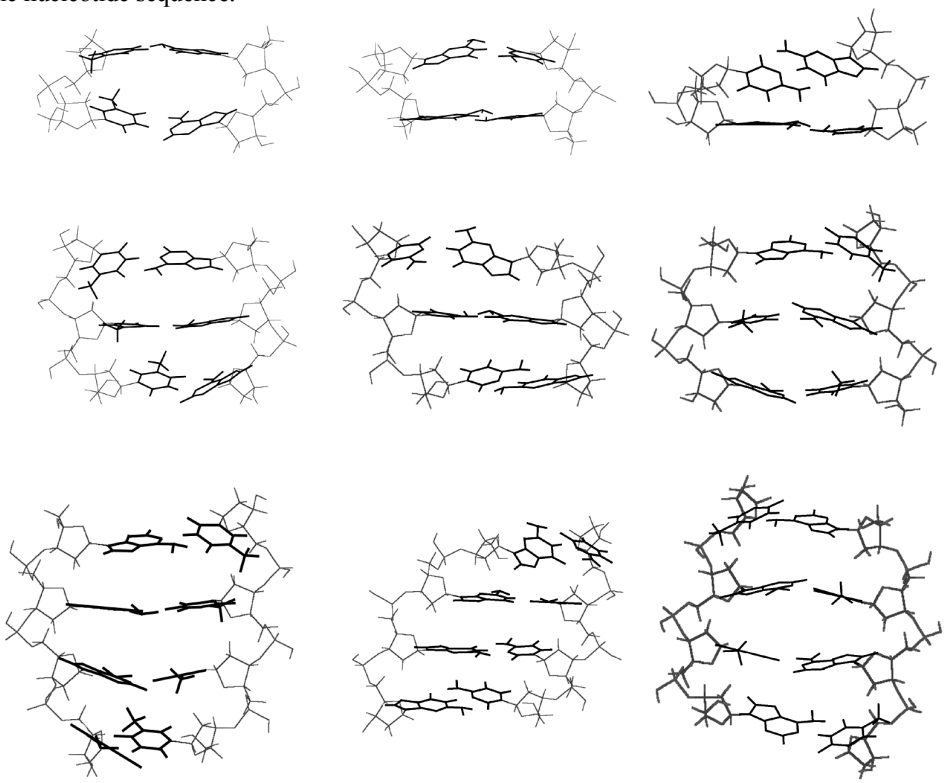


Fig. 1. MNDO/PM3 optimized structure of dinucleotide - d(TpT).d(ApA), d(GpG).d(CpC), d(CpG).d(CpG), trinucleotide - d(TpTpT).d(ApApA), d(CpCpC).d(GpGpG), d(ApTpA).d(TpApT) and tetranucleotide - d(ApApApA).d(TpTpTpT), d(GpGpGpG).d(CpCpCpC), d(TpApTpA).d(ApTpApT) duplexes.

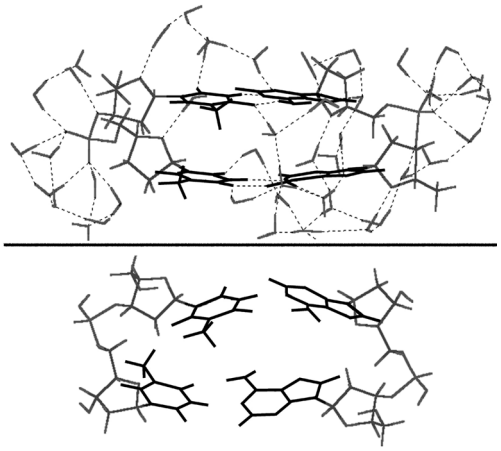


Fig. 2. MNDO/PM3 optimized structure of duplex d(TpT).d(ApA) in aqueous "shell" with $43H_2O$ molecules and without water.

One can also expect, that the obtained large deformation of the base pairs of 3'-5' antiparallel double chains should induce an uncompensated large component of electrical dipole moment along spiral axis [14]. Thus, the results obtained can shed light on the problem of nucleic acids functioning and DNA-protein recognition process.

This work was supported by Russian Fond for Basic Research, Grant № 99-04-48162.

References

1. Saenger W. Principles of Nucleic Acid Structure, New York, Springer-Verlag, 1984.
2. Hobza P., Sponer J., Chem. Rev. 99, 3247 (1999).
3. Sponer J., Leszczynski J., Hobza P., J. Biomol. Struct. Dyn. 14, 117 (1996).
4. Florian J., Sponer J., Warshel A., J. Phys. Chem. B103, 884 (1999).
5. Aida M. J. Comput. Chem. 9, 362 (1988).
6. Santamaria R., Vazquez A. J. Comput. Chem. 15, 981 (1994).
7. Wilson C.C. Nucl. Acid Res. 15, 8577 (1987).
8. El Hassan M.A., Calladine C.R. J. Mol. Biol. 259, 95 (1996).
9. Heineman U., Alings C., Hahn M. Biophys. Chem. 50, 157 (1994).
10. Jursa J., Kypr J. Gen. Physiol. Biophys. 12, 401 (1993).
11. Komarov V.M., Mevkh G.N. J. Phyz. Khim. 69, 1419 (1995) (in Russian).
12. V.M.Komarov, Biophysics, 43, 917 (1998)
13. V.M.Komarov, J.Biol.Phys., 24, 167 (1999)
14. A.V.Kabanov, V.M.Komarov, Int.J.Quant.Chem. 88 (5) (2002)(in press)
15. J.J.P.Stewart, J.Comput.Chem., 10, 209 (1989)
16. J.J.P.Stewart, J.Comput.Chem., 10, 221 (1989)
17. T.N.Lively, M.W.Jurema, G.Shields, Int.J.Quant.Chem.Quant.Biol.Symp. 21, 95 (1994)
18. H.D.Dos Santos, W.B.De Almeida, J.Molec.Struct.(Theochem), 335, 129 (1995)
19. I.Juranic, H.S.Rzepa, Yan Yi Min, J.Chem.Soc.Perkin Trans.2, 877 (1990)
20. V.Hroda, J.Florian, P.Hobza, J.Phys.Chem., 97, 1542 (1993)
21. P.Hobza, F.Hubalek, M.Kabelac, Mejzlik, J.Sponer, J.Vondrasek, Chem.Phys.Let., 257, 31 (1996)
22. M.Elstner, Th.Frauenheim, E.Kaxiras, G.Sheifert, S.Suhai, Phys.Stat.Sol., 217b, 357 (2000)
23. J.Florian, J.Lezsczynski, Int.J.Quant.Chem.Quant.Biol.Symp. 22, 207 (1995)
24. L.Gorb, J.Leszczynsky, in *"Computational Molecular Biology"* of Theoretical Chemistry
26. Book Series, v8, 167-209 (1999), Elsevier
25. J.Sponer, M.Sabat, L.Gorb, J.Leszczynski, B.Lippert, P.Hobza, J.Phys.Chem. B104, 7535 (2000)
26. Y.Podolyan, Y.V.Rubin, J.Leszczynski, J.Phys.Chem., A104, 9964 (2000)
27. C.F.Guerra, F.M.Bickelhaupt, J.G.Snijders, E.J.Baerends, J.Am.Chem.Soc., 122, 4117 (2000)
28. G.Barone, M.C.Ramusino, R.Barbieri, G.La Manna, J.Molec.Struct.(Theochem), 335, 129 (1995)
29. L.Gorb, A.Korkin, J.Leszczynski, V.Varnek, F.Mark, K.Schaffer, J.Mol.Struct.(Theochem), 425, 137 (1998)

EXAMPLE OF A RECONSTRUCTION OF EVOLUTION OF THE GENETIC CODE (GC)

Lenski S.V.

Ekaterinburg, Russia, e-mail: ucag@mail.ur.ru

Key words: *genetic code, evolution, codon*

Resume

Motivation: The finding - out of stages of origin of GC is necessary for understanding fundamental evolutionary processes.

Results: The table of GC has necessary information for reconstruction a scheme of evolution of GC without logical contradicts. It is possible to observe the evolution of GC from mono to triplet. The triplet was originated from second place, next was first place and third place of triplet.

Availability: http://ucag.web.ur.ru

Introduction

In this research was used a consensus equivalent of known tables of GC. Names of amino acids are absent. This was determined the investigation of evolution only the codon without a investigation a evolution of whole way of the translation (Chapeville, Haenni, 1974). Base properties of GC are next: GC was packed; GC has two constants. First constant is line U-C-A-G. Second constant is the order of filling the table GC: first filling on second place of triplet; second filling on first place; third on third. F.H.C.Crick used these constants for forming the first table of GC (Crick, 1966, 1968) after decoding (Fig. 1). On Fig. 1 were used next designations:

- triangles for pyrimidines;
- circumferences for purines;
- white background for U, A;
- black background for C, G.

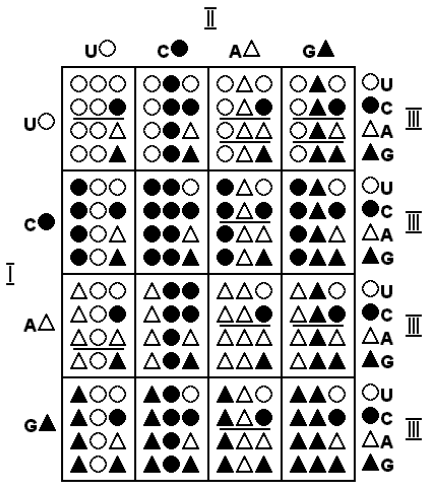


Fig. 1. Table of the Genetic Code – Standard.

According to historical reasons H.F.C.Crick examined the table "Standard". Unpacked variants of the table of GC are losing some of necessary properties. First of all, the changing an order in line U-C-A-G change space relations into GC. Only reverse of order into G-A-C-U is equivalent U-C-A-G conformably to space relations.

Methods

On the base of discrete mathematics, an apparatus for evolutionary research was worked out. This apparatus may be applied to some combinatorial space. This discrete space must be formed by a combinatorial operator, similarly n^k (placed n in k cells with repeats). This combinatorial space should satisfy some conditions:

- the space must be filled completely and be degenerated;

- in the space, there must be a particular full order concerning some optimal rules.

The apparatus works in three stages. At the first stage, we cancel the surplus information. At the second stage, we perform a combinatorial composition of all possible options. At the third stage, we perform a search of a real version out of all the possible options. The combinatorial space of GC satisfies all necessary conditions.

Results

Fig. 2 show the table of GC without different between pyrimidines and purines. In such a case there were appeared tautologies - the writing of codons with the equal sequences of triangles and circumferences. With the cancellation of all tautologies up to the direction of arrows on the diagram, - the brief notation of the table of GC has only eight codons.

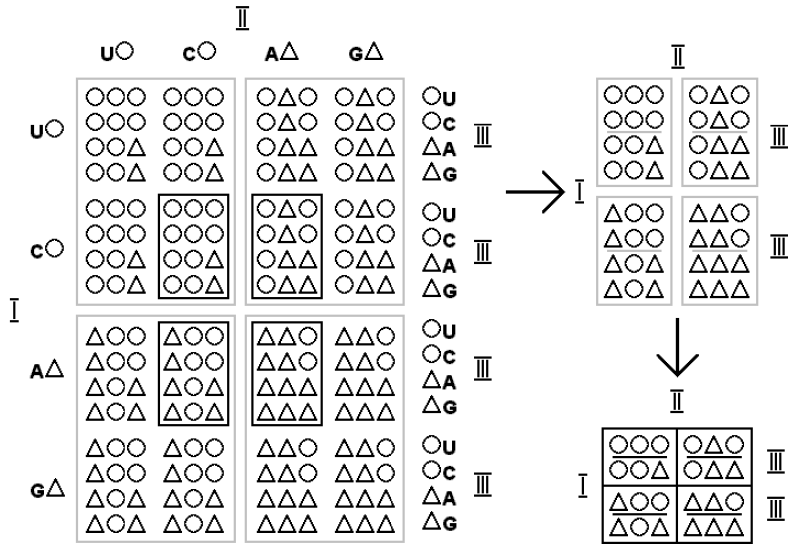


Fig. 2. The cancellation with U=C and A=G.

This result is necessary to identify among all theoretical possible notations evolution GC from mono to triplet. Possible notations mono→duple→triplet only 3! = 6. For identification it is necessary to make a procedure of creating all six possible notations, and then to compare with the brief notation GC. This procedure was showed in Fig. 3. The step mono (painted position) is possible only with the independent filling of one position by some base. The movement to next step -> duple (position with shadowing background) is possible as addition the right position (upper line of tables), and as addition the left position (lower line of tables). The movement to the step triplet (positions with white background) is possible by addition the position from the right, the left, or to the center by the relation on the step = duple. Notation that is gotten as result of combinatorial model an evolution from mono to triplet coinciding with the notation GC from Fig. 2 (is underlined). The notation equivalent the brief notation GC was created by consecutive addition from mono to duple → from the left, (lower line of tables), and from duple to triplet - by addition next position from the right.

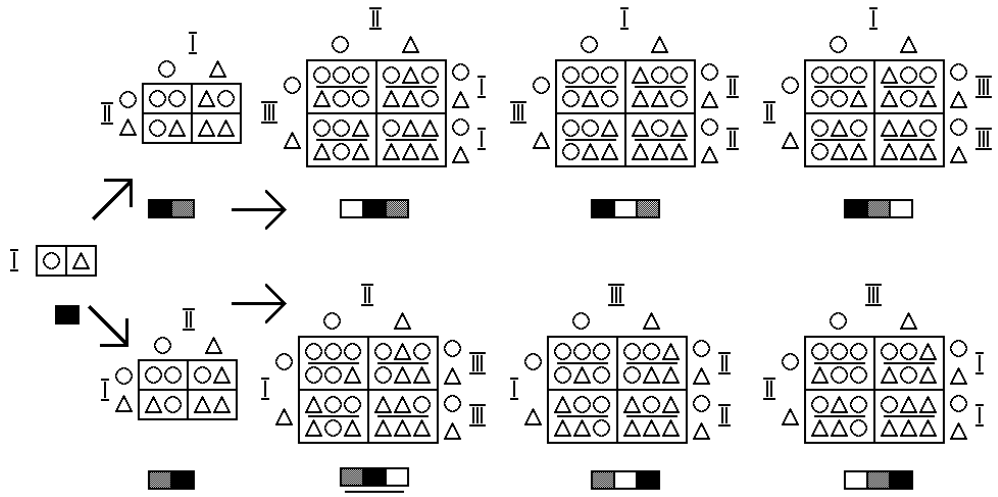


Fig. 3. The combinatorial modeling of triplet evolution.

Discussion

Some of stages of evolutionary process may be reconstructed by theoretical approach (Trifonov, Bettecken, 1997). The method of combinatorial modeling may be used for theoretical researches. On big data the method of combinatorial modeling may be used with computer. The authenticity results big data is very high. In examined example the authenticity is very low. It is only 1 case against 6 possible. Consequently an additional support this result from the theoretical or/and from another approaches is necessary.

References

1. The Genetic Code. Cold Spring Harbor Symposia in Quant. Biol. Cold Spring Harbor. N.Y. 31 (1966).
2. Crick F.H.C. (1968) The Origin of the Genetic Code. J. Mol. Biol. 38, 367-379.
3. Chapeville F., Haenni A. (1974) Biosynthese des proteines.Traduction genetique. Hermann Collection Paris Methodes.
4. Trifonov E.N., Bettecken T. (1997) Sequence fossils, triplet expansion and reconstruction of earliest codons. Gene. 205, 1-6.

MULTIPLATFORM INTEGRATED PROGRAM PACKAGE JGENOMEEXPLORER FOR GENOMIC ANALYSIS

Dolgopolov A.Y., Dachtchian K.A., Novichkov P.S., Mironov A.A.

Integrated Genomics, 117333, P.O. Box 348, Moscow, Russia
Corresponding author: e-mail: phe13@mail.ru

Key words: *regulation, comparative analysis, regulatory site, computer analysis*

Resume

Motivation: Recognition of regulatory sites is an important part of genome annotation.

Results: We have created the program package JgenomeExplorer that integrates a number of necessary tools for analysis of regulation and has a convenient graphic and web interface. Software is developed on the Java platform and can be ported to different types of computers and operating systems.

Availability: The web part of package: http://212.48.144.189/genexp/signal_search.jsp.

Introduction

Recognition of regulatory sites is an important part of genome annotation. The analysis of regulation not only is interesting in itself, but also allows one make more specific annotation of gene function. It is often assumed that regulons (sets of co-regulated genes) are conserved in related genomes. Thus, if the gene function is characterized in one genome, and this gene has an upstream regulatory site, it is possible to select a homolog retaining the site in a related genome. Thus, it becomes possible not only to resolve the orthology relationships and thus to assign the exact cellular function to this homolog, but also to predict the regulation in the new genome. Similar considerations can be used to enhance computational predictions in the absence of experimental data. Availability of a large number of bacterial genomes makes the comparative analysis of regulation a powerful tool of genome annotation. This approach has been used to describe various regulatory systems (see the abstract by Ravcheev et al., Permina et al., Panina et al., Vitreschak et al., Gerasimova et al., Kotelnikova et al. In this volume).

Methods and Algoritms

This analysis requires a number of specific tools. Firstly, one identifies the regulatory signal common to genes forming a regulon. Secondly, the derived recognition rule is applied to find candidate sites in the analyzed genomes. Then one needs to identify orthologs in the sets of candidate regulon members. Finally, it convenient to have a simple way to revise the existing annotation.

The existing program GenomeExplorer for Windows (Mironov et al., 2000) addressed some of the above problems. However it has a number of deficiencies. First, it can work only on the Windows platform. Second, it cannot distribute computations. Finally, this programm lacks a number of necessary tools (in particular: multiple alignment, using combinations of profiles for identification of candidate sites in genomes, taking into account signals of more diverse types).

Implementation and Result

Therefore, we have created the program package JgenomeExplorer that integrates all the necessary tools and has a convenient graphic interface. This simplifies and accelerates the researcher's work. Application of the JAVA technology allows one to run the program on different platforms under different operating systems. The three-level structure allows for easy distribution of tasks over several computers. Being based on standard interfaces, the package is easily extendable, and integration of new functionalities is simple and convenient. The data can be stored in any database supporting the ANSI SQL standard.

The current version has the following functionalities:

- search over annotation by gene name, product, keywords, sequence;
- generation of orthogs for a given gene using pre-processed ortholog tables;
- similarity search for a nucleotide or amino acid sequence over genome, proteome, or translated genome using BLAST with subsequent alignment using the Smith-Waterman algorithm;
- multiple alignment using CLUSTAL;

- finding signals in unaligned sequences and construction of recognition profiles (search for signals of the following types is possible: word, palindrome, inverted repeat, tandem repeat, common repeat, double palindrome);
- identification of candidate sites in genomes using single profiles or combinations of profiles (with conditions on multiple site occurrences, distances between sites, etc.);
- "computational subtractive hybridization" leading to identification of genes present in a given set of genomes and missing in other genomes.

The screenshots demonstrate the graphic interface.

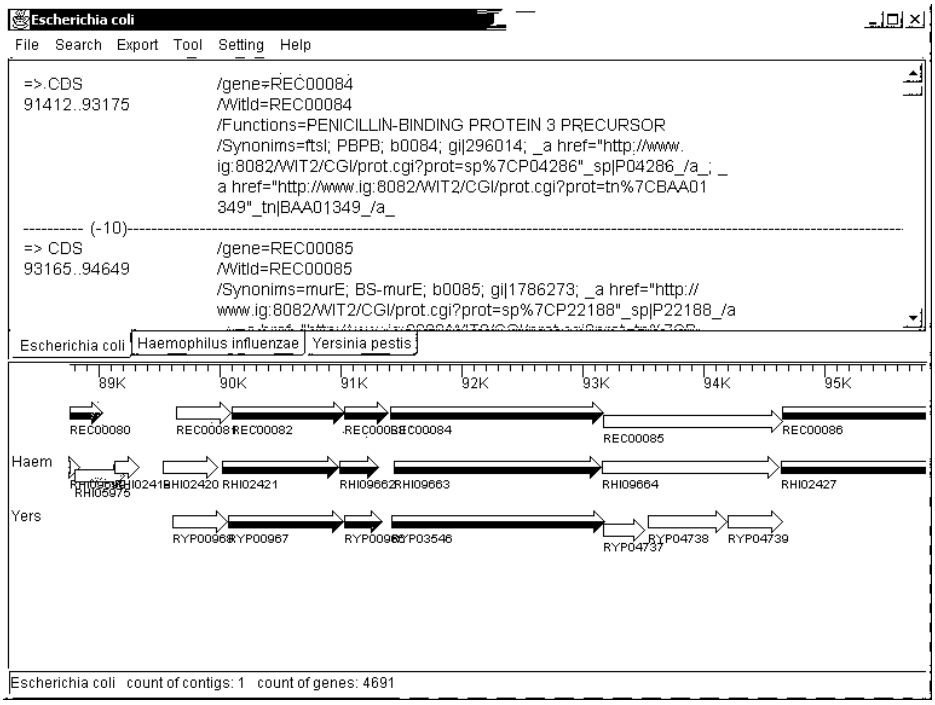


Fig. 1. General layout: genomic map.

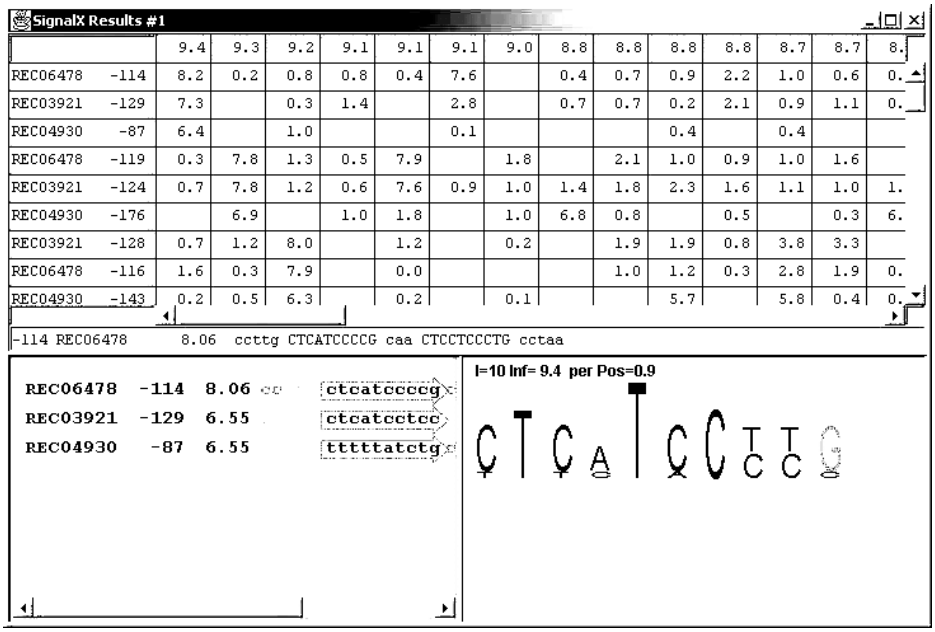


Fig. 2. Output of the signal identification procedure.

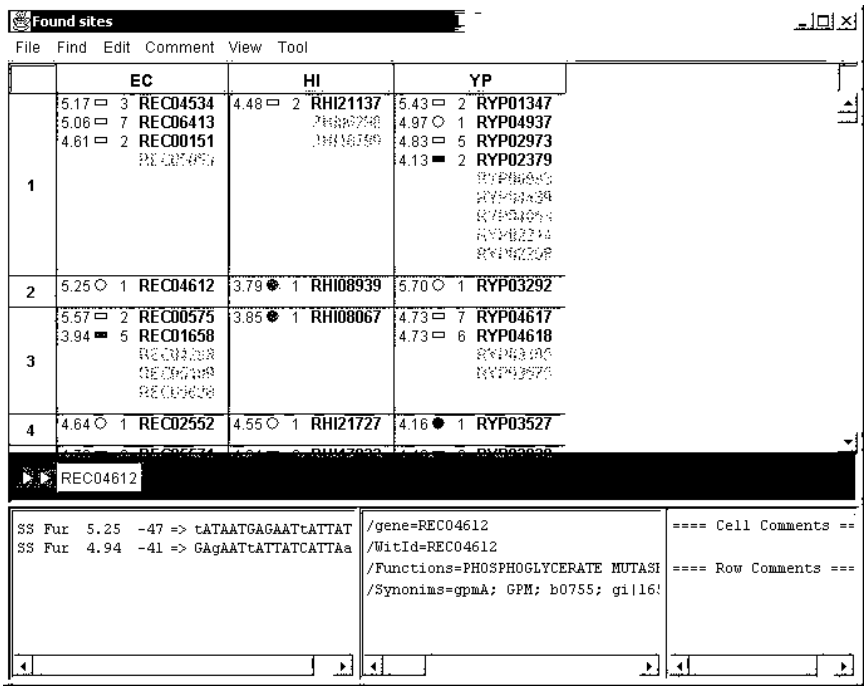


Fig. 3. Results of site search in the three genomes.

Acknowledgements

The authors are grateful to M.S.Gelfand, V.E.Brodyansky and A.E.Kazakov for helpful discussions and testing of the package.

References

1. Mironov A.A., Vinokurova N.P., Gelfand M.S. (2000) The software of the analysis bacterial genomes. Mol. Biol. (Mosk). 34, 253-362.

MANUAL ANNOTATION OF THE HUMAN AND MOUSE GENE INDEX: WWW.ALLGENES.ORG

[1] Brunk B., [1] Crabtree J., [1] Diskin S., [1*] Mazzarelli J., [1] Zigouras N., [2] Alkalaeva E., [2] Bogdanova V., [2] Trifonoff V., [2] Vorobjeva N., [2*] Katokhin A., [2] Kolchanov N., [1] Stoeckert C.

[1] Computational Biology and Informatics Laboratory, Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104
[2] Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia,
e-mail: mazz@SNOWBALL.pcbi.upenn.edu, katokhin@bionet.nsc.ru
*Corresponding authors

Key words: *human and mouse gene index, ESTs clustering, GO function prediction, integrated database*

Resume

Motivation: With the recent announcement of the completion of a "draft" of the human genome, and of a new plan to complete a (public) low-coverage draft sequence of the mouse genome, there is an increasing need for databases and tools to support the analysis and interpretation of partially and fully sequenced mammalian genomes.

Results: allgenes.org is a web site whose primary goal is to provide access to an integrated database of every known and predicted human and mouse gene, using only publicly-available data. allgenes.org is its focus on integrating the various types of data (e.g., EST sequences, genomic sequence, expression data, functional annotation) and on doing so in a structured manner. To this end allgenes.org relies on a relational database that uses controlled vocabularies and ontologies to ensure that biologically meaningful queries can be posed in a uniform fashion..

Availability: http://www.allgenes.org/

Introduction

With the recent completion of a rough "draft" of the human genome (International Human Genome Sequencing Consortium 2001; Venter et al., 2001), with the finished sequence for *D. melanogaster* (Adams et al., 2000), *C. elegans* (The *C. elegans* Sequencing Consortium, 1998), and *S. cerevisiae* (Mewes et al., 1997), and with numerous other sequencing projects in progress, vast amounts of genomic data have become available for further refinement and analysis.

However, the heterogeneity of the data sources, together with their frequently unconventional implementation, makes accessing genomic data across multiple data sources extremely difficult. Researchers are faced with the problem of integrated access to heterogeneous data sources (Davidson et al., 2001).

Over the past ten years, a variety of techniques have been developed to overcome the problem within the genomic community. Several *link-driven federations* have been created, in wich users start by extracting entries of interest in one data source and then hop to other related data source via Web links that have been explicitly created by the developers of the system. In the systems implementing a *view integration* appproach, the schemas of collection of underlying data sources are merged to form a global schema in some common model (such as the relational, complex value, or object-oriented model) (Wong, 2000, Davidson et al., 2001, Stoeckert et al., 2001).

allgenes.org is a web interface providing access to the assembled EST and mRNA sequences, or DoTS RNA transcripts, contained within GUS (Genomics Unified Schema), a relational database.

Implementation and Results

allgenes.org is a web site whose primary goal is to provide access to an integrated database of every known and predicted human and mouse gene, using only publicly-available data. There are a number of similar efforts, some public and some private (i.e., for-profit), that also provide gene indexes for human, mouse, and other organisms. What we hope distinguishes allgenes.org is its focus on integrating the various types of data (e.g., EST sequences, genomic sequence, expression data, functional annotation) and on doing so in a structured manner. To this end allgenes.org relies on a relational database that uses controlled vocabularies and ontologies to ensure that biologically meaningful queries can be posed in a uniform fashion.

The heart of **allgenes.org** is a comprehensive index of predicted human and mouse genes. At present these gene predictions are drawn from transcripts predicted by clustering and assembling EST and mRNA sequences. These EST and mRNA clusters are those from the latest release of the Database Of Transcribed Sequences (DoTS), developed here in the Computational Biology and Informatics Laboratory at the University of Pennsylvania. The DoTS transcripts integrate annotation from cDNA libraries (tissue source) and RH mapping data also stored in GUS. Automated annotation has been applied to the DoTS transcripts to determine their predicted gene ownership, protein sequences and GO Functions.

Manual annotation efforts have focused on validating the automated annotation and adding additional gene information. Manual annotation of the gene index utilizes an annotation tool, the GUS annotator interface, which directly updates the GUS database.

Functional features of the interface which allow defined annotation tasks to be performed by the annotator include: determination of transcript gene membership using BLAST similarities and transcript alignments to genomic sequence, assignment of approved (HUGO or MGI) gene symbol, gene synonyms and confirmation/addition of protein GO Function assignments (Fig.). Evidence for the automated annotation is stored in GUS and provided to the annotator to assist in the validation of the assignments. Evidence is also manually added by the annotator for each assignment and is stored in GUS. The mouse and human DoTS transcripts have been aligned on the UCSC Golden path contigs allowing for the identification of new genes, alternative transcript forms and annotation of the genome.

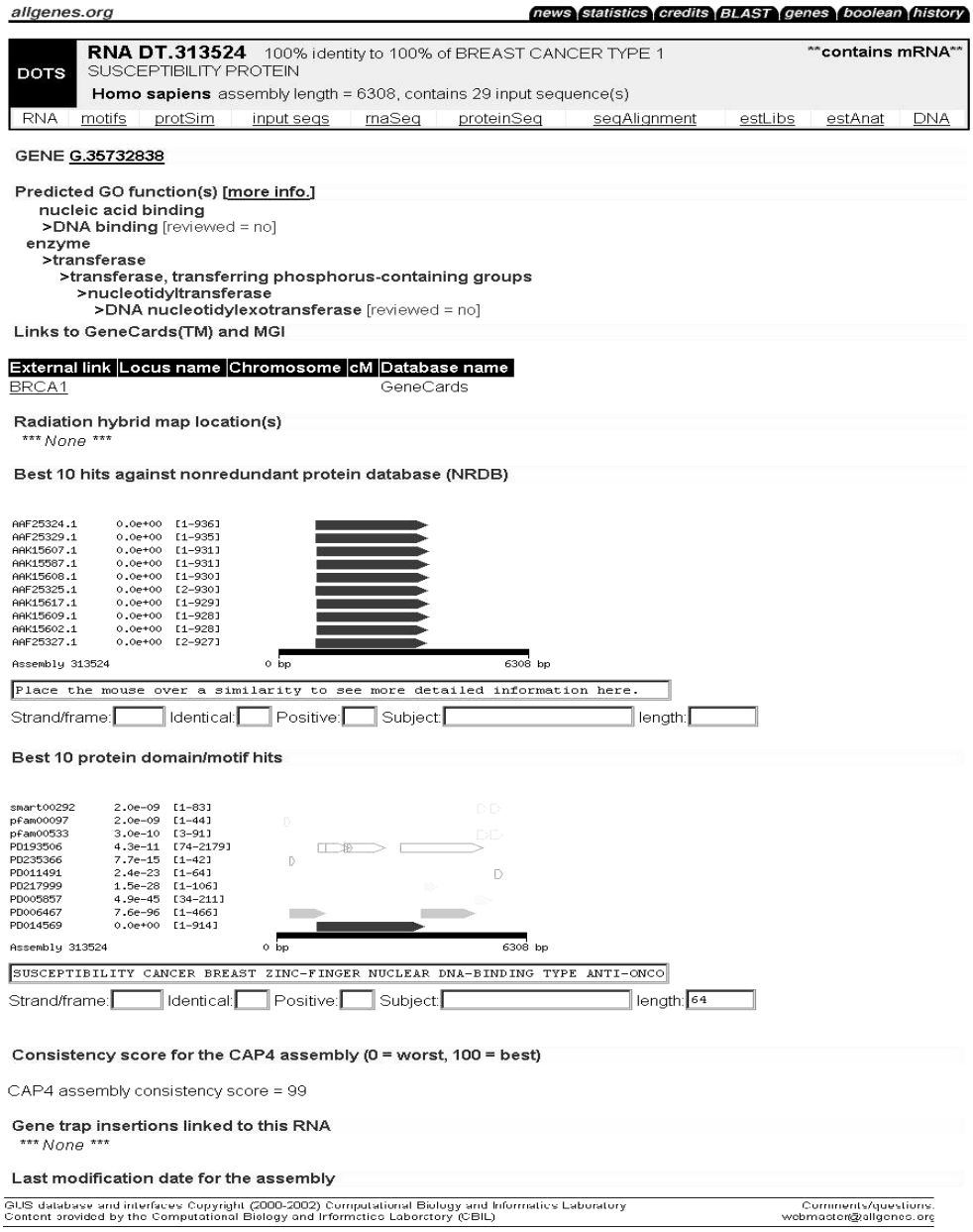


Fig. 1. An example of allgenes entry for BRCA1 gene cluster.

Manual annotation efforts have focused on mouse genes important for pancreatic development and function and human genes contained within a region deleted on chromosome 22, causing DiGeorge syndrome, a developmental disorder. Efforts are currently underway to develop an improved annotation tool allowing the creation of gene models, the definition of alternative RNA transcripts and the linkage of tissue expression with these potential protein isoforms.

The gene index contains over 3 million human and nearly 2 million mouse ESTs and mRNAs as of September, 2001 that have clustered into 150,006 human and 74,024 mouse "genes" (a new build of the index is underway). Approximately half the human and mouse genes have similarity to a known protein sequence and of these, we have been able to predict a Gene Ontology (GO) molecular function for 31% of the human and 45% of the mouse genes (Schug et al., 2002). Manual annotation is used to better structure the data (e.g., assign libraries to an anatomy ontology), confirm automated annotation (e.g., check GO assignments), and add new information (e.g., assign gene symbols and synonyms). Nearly 2000 human and mouse assemblies have been manually reviewed as of October, 2001 and this number is expected to greatly increase.

The sequences, their contained accessions, predicted protein translations and predicted GO functions can be downloaded at the **allgenes.org** site.

Acknowledgements

Work was supported by grant of National Institutes of Health USA (№ 2 R01-HG-01539-04A2).

References

1. Adams M.D., Celniker S.E., Holt R.A. et al. (2000) The genome sequence of *Drosophila melanogaster*. Science. 287:2185-2195.
2. Davidson S.B., Crabtree J., Brunk B.P. et al. (2001) K2/Kleisli and GUS: experiments in integrated access to genomic data sources. IBM Systems J. 40(2):512-531.
3. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. Nature. 409:860–921.
4. Mewes H.W., Albermann K., Bahr M. et al. (1997) Overview of the yeast genome. Nature. 387(6632 Suppl):7-65.
5. Schug J., Diskins S., Mazzarelli J., Brunk B., Stoeckert C. (2002) Predicting Gene Onthology Functions from ProDom and CDD Protein Domains. Genome Res. 12:648-655.
6. Stoeckert C., Pizarro A., Manduchi E. et al. (2001) A relational schema for both array-based and SAGE gene e[pression experiments. Bioinformatics. 17(4):300-308.
7. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science. 282, 2012-2018.
8. Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A. et al. (2001) The sequence of the human genome. Science. 291(5507): 1304–1351.
9. Wong L. (2000) Kleisli, a functional query system. J. Functional Programming. 10(1):19-56.