# MODELLING & COMPUTER METHODS IN MOLECULAR BIOLOGY & GENETICS

Reports of International Conference
Novosibirsk, August 28–September, 1990

# MODELLING AND COMPUTER METHODS IN MOLECULAR BIOLOGY AND GENETICS

Edited by
V.A. Ratner and N.A. Kolchanov

NOVA SCIENCE PUBLISHERS, INC.

Nova Science Publishers, Inc.
6080 Jericho Turnpike, Suite 207
Commack, New York 11725

Book Production Manager: Janet Glanzman White
Graphics: Elenor Kallberg, Christopher Concannon, and Michael A. Masotti

*Library of Congress Cataloging-in-Publication Data available upon request*

ISBN 1-56072-077-8



*Printed in the United States of America*

# MODELLING AND COMPUTER METHODS IN MOLECULAR BIOLOGY AND GENETICS

# PREFACE

An International Conference "Modelling and Computer Methods in Molecular Biology & Genetics" was organized by the Theoretical Department of the Institute of Gytology & Genetics, SB AS USSR, together with the All-Union scientific and technological program "GENINFORM". It was held on August 28 - September 1, 1990 in Akademgorodok , Novosibirsk Scientific Center of Siberian Branch of the USSR Academy of Sciences, at Maly Zal (the Small Hall) of the House of Scientists [I].

Being the I-st International Conference on these topics, it was actually the 4-th All-Union Conference, because the soviet participants meet once in two years since 1984 in Akademgorodok under the patronage of Siberian Branch of the USSR Academy of Sciences and the "GENINFORM" program. Here we had found a compact range of cooperated institutes and persons, discussed many common works, methods, results and technical projects. The subject of all these conferences was the usage of all available software, data-banks, methods for solution of some problems of the Theory of Molecular Genetic Regulatory Systems (MGRS), analysis of primary and spacial structure of macromolecules, and development of the Theory of Molecular Evolution (TME) (see [2-5]).

To continue this tradition the present Conference had the meetings and poster sessions on the following subjects:

1. Computer analysis of polynucleotide sequences;
2. Data Banks and program packages;
3. Computer analysis and Modelling of Protein Structure;
4. Molecular Genetic Sistem Modelling;
5. Theory of Molecular Evolution.

Moreover, there was the evening Round Table for Discussions devoted to the two main problems:

a). Genetical Language as a linguistic system;
b). Unified Theory of Molecular Evolution.

We decided not to present here the text of discussion, but our final paper [6] which contains the list of the problems for discussion and some comments.

In general, we are satisfied with the results of the Conference. We hope, the opinions of our guests are the same. It seems to me that they were pleased to find in so a distant place as Novosibirsk such an active community of mathematical and computer geneticists. And the weather was also fine.

In my final words I want to welcome all the authors of this Issue, all the participants of our Conference and interested readers. My best regards to our Publisher, Mr. Frank Columbus (NOVA SCIENCE PUBLISHERS), who gave his "yes" just after the first words about this Conference were said. The emblem of the Conference was designed by Dr. Andrey Rzhetsky, the participant of the Conference. I must also thank all our colleagues who helped us to organize this meeting.

## REFERENCES


[1] Abstr. of Intern. Conf. "Modelling and Computer Methods in Molecular Biology & Genetics", Eds. V.A.Ratner, N.A.Kolchanov, ICG-Press, Novosibirsk, (1990), v. I,2.

[2] Proc. of I-st All-Union Workshop "Theoretical Investigations and Data Banks in Molecular Biology & Genetics", Ed. V.A. Ratner, ICG-Press, Novosibirsk, (1986) (Russ.).

[3] Abstr. of 3-rd All-Union Workshop "Theoretical Investigations and Data Banks in Molecular Biology & Genetics", Ed. V.A.Ratner, ICG-Press, Novosibirsk, (1988) (Russ.).

[4] Computer Analysis of Structure, Function and Evolution of Genetic Macromolecules, Ed. N.A. Kolchanov, ICG-Press, Novosibirsk, (1989) (Russ.).

[5] Computer Analysis of Genetic Texts. Ed. M.D. Frank-Kamenetzky, Nauka, Moscow (1990) (Russ.).

[6] V.A. Ratner, this volume,



Prof. Vadim A. Ratner
Head of Theoretical Department,
Vice-President of Organizing
Commettee of the Conference


November, 1990
Novosibirsk

# Contents

Modelling and Computer
Methods in Molecular
Biology and Genetics


ISBN 1-56072-077-8

# CURRENT PROBLEMS OF COMPUTER ANALYSIS OF GENETIC TEXTS

N.A.Kolchanov

Institute of Cytology and Genetics of the USSR Academy of Sciences, Novosibirsk, 630090, USSR

The development of highly efficient methods of DNA sequencing stimulated working out methods of genetic text computer analysis. A number of such program packages are designed up to date [1,2]. However, they cannot provide the rate of analysis compared with that of DNA sequencing. In order to improve the situation, more efficient technique is to be developed. The most complicated problems of genetic text analysis are associated with the "Human genome" project. In this case even the primary data analysis requires highly productive computer systems, able to work in an automatic mode.

It should be emphasized that some fundamental problems are also associated with the genetic text analysis. They are:

1) The theory of structural & functional organization of genomes, which involves:
   a) Theory of genome regulatory regions and functional sites.
   b) Theory of mobile genetic elements.
   c) Theory of gene structural & functional organization.
2) Theory of mutational and recombinational processes in genomes.
3) Theory of genetic text evolution.
4) Theory of protein structure & function prediction on the base of their amino acid sequences.

The most efficient way for these problems to be solved is the use of a new informational technique, which is oriented to the creation of artificial intelligence systems and knowledge bases in the corresponding fields of molecular biology and genetics.

This paper is devoted to new approaches in the genetic text analysis at the example of the results obtained in the Laboratory of Theoretical Molecular Genetics of the Institute of Cytology and Genetics of the USSR Academy of Sciences.

| TYPE | ELEMENT | |
|---|---|---|
| OBLIGATORY ELEMENTS | TATA-BOX | |
| | -35-BOX | |
| MODULATORS | COMPLEMENTARY PALINDROMS | [3] |
| | REGIONS WITH HIGH AND LOW ENERGY OF DNA MELTING | [3] |
| | TANDEM NONPERFECT REPEATS | [4] |

Fig. 1. Examples of obligatory elements and modulators in prokaryotic promoters

Let us consider the problem of functional site computer analysis of polynucleotide sequences. Developing of artificial intelligence systems begins from the creation of conceptual models of phenomena in study. As for functional sites, the below concept is represented

We suppose that in case of a site to function normally, the two types of elements are essential:

1) Obligatory elements, which are vitally important for the site normal activity. As a rule (but not always), various types of consensuses or some conservative blocks of consensuses in functional sites are considered as obligatory elements. In most cases functional sites can be characterized by the more or less exact location of some obligatory elements.

2) Modulators - second type of elements, which are presented in functional sites with variable location and number.

It is supposed, that the location and number of modulators is the essential factor of site functioning. We have found various types of such modulators in functional sites (Fig. 1). Solovyev et al. [3] have shown that prokaryotic promoters contain such type of modulators as complementary palindromes.The greater the number of these palindromes, the more is the strength of the promoter. Modulators are regions of high and low energy of DNA melting and their location within promoters is different [3]. Kel' et al. [4] have shown that

promoter strength depends on saturation with tandem nonperfect repeats. So, obligatory elements and modulators are actual and their origin is different.

It should be noted, that the mechanisms of site functioning are unknown in most cases. Apparently it is difficult to suppose a certain type of structural elements to define functioning of a definite type of sites. Consequently,it is necessary to generate a wide range of hypotheses upon the nature of these elements while studying functional sites. This approach corresponds to one of the main principles of artificial intelligence - the principle of impartiality [5], according to which the hypothesis should never be neglected without serious reasons.This very approach was realized in a computer system designed by Kel' et al. [6,7].

This system has a number of standard blocks (see fig.4 in paper [7] ). The first block is an intellectual interface, which was designed for input and output of information and also for the system operating.

The second block is a database, containing information about functional sites ( type of functional site, its polynucleotide sequence, location, etc.).

The third block is the system of hypothesis generation about the role of various structural elements in site functioning. Here this block generates a hypothesis of the role of the two types of elements: 1) oligonucleotides ; 2) various types of repeats and symmetrical sequences.

The fourth block is used in statistical examination of hypotheses. The hypotheses, whose justice is stated as being proved are considered as

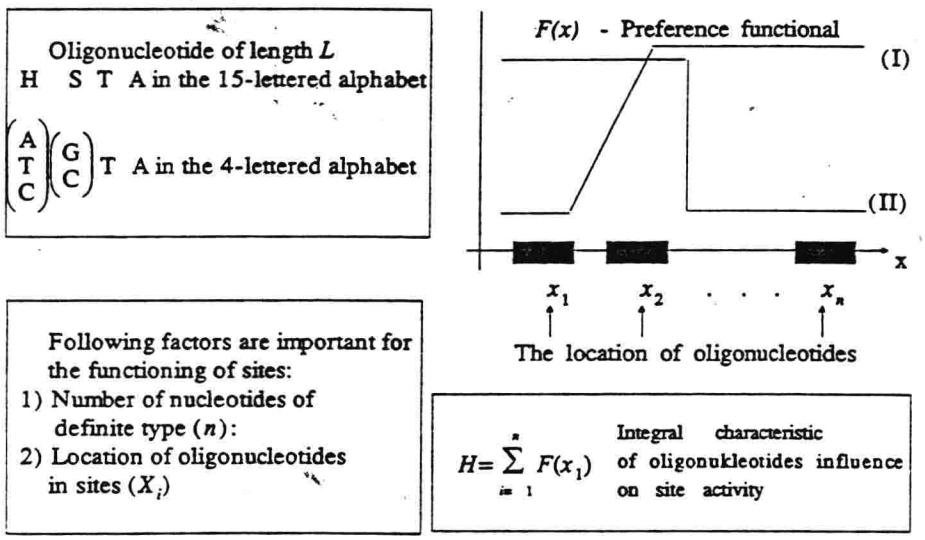



Fig. 2. Formal description of oligonucleotide influence on site activity

knowledges. Therefore, the above blocks are included in the system of knowledge production.

It is the knowledge base that distinguishes the artificial intelligence system from ordinary packages of programs.The knowledge obtained in the course of the system work is stored in the knowledge base.

In this system the two types of knowledge are stored in the knowledge base: 1) knowledge about structural elements which provide site functioning; 2) knowledge about the methods of those sites recognition. The methods are produced by the block of recognition. This block is also the part of the knowledge production system, as well as the intellectual interface controlling the process of their construction.

Let us consider the principles of system work at the example of oligonucleotides as structural elements (Fig. 2). Oligonucleotide is a fragment of DNA of length $L$ and it has certain nucleotide sequence. It can be presented in the 15lettered code or in 4-lettered one, as it is shown in the upper left part of Fig. 2.

Let us suppose that functioning of every type of sites depends on a specific set of oligonucleotides. In this case, for the site functioning the following parameters are important:

1) Number of oligonucleotides of a certain type in the site ($N$).
2) Their localization in the site ($X_i$).

It is supposed that the influence of the certain type of oligonucleotides on the site activity depends upon their position in the site.

It is the preference functional that is used in our system for the qualitative description of this influence. It is schematically represented in the right upper part of Fig. 2. For example, functional I reflects the increasing of oligonucleotide influence on the site activity from the 5'- to the 3'- direction. Functional II reflects the decreasing of the oligonucleotide influence on the site activity from its 5'- to the 3'-direction.

It is supposed that the contribution to the total site activity carried out by the oligonucleotide at the $X_i$ -position is proportional to the quantity of the preference function $F(X_i)$. So, value $H$ is the integral characteristic of a certain type of oligonucleotides to influence the site activity (Fig. 2).

Thus, the hypothesis about oligonucleotide influence on the site functioning can be described by the following parameters:

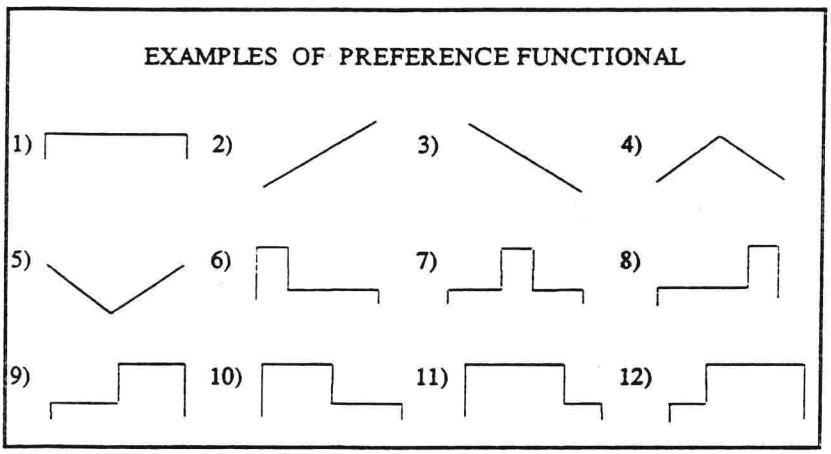
1)Oligonucleotide length $L$;
2)Oligonucleotide type;
3)Function of preference.

The given set of parameters describes the model of oligonucleotide influence on the site functioning.

25 various variants of the preference function are examined in the system (Fig. 3). One could note that when the oligonucleotide length varies within the 1

PARAMETERS OF THE MODEL

1) Length of oligonucleotides $L = 4$
2) Variants of oligonucleotides $V_0 = 15^L$
3) Variants of the preference functional $V_f = 25$




Total number of models: $N = V_0 \times V_f = 15^4 \times 25 = 1.434.375$

Fig. 3. Variants of the preference function.

to $L$ range and when 25 preference functionals are under consideration, the total number of model variants accounts to billions.

Let us fix a variant of the oligonucleotide, the variant of a preference function and consider a sample of sites (Fig. 4) to calculate $H$ for each site. Now let us do the same for a sample of "non-sites", i.e. for sequences not having this activity . We calculate $H$ for each "non-site".

These procedures are accomplished with the block of hypothesis generation. Then starts the block of hypothesis examination which compare the distributions of $H$ and $H^*$ as it is shown in Fig. 4.The more the difference between the distributions, the more significance the given type of oligonucleotides has for the site activity. Such oligonucleotides are more useful for constructing the method of this site recognition.

In our system the difference between $H$ and $H^*$ distributions is integrally estimated by value $U$ which is called "Utility". The utility is calculated on the base of a special expert system designed in the Laboratory [8]. It contains knowledges on the theory of pattern recognition, theory of classification, methods of cluster analysis and the knowledge on mathematical statistics.