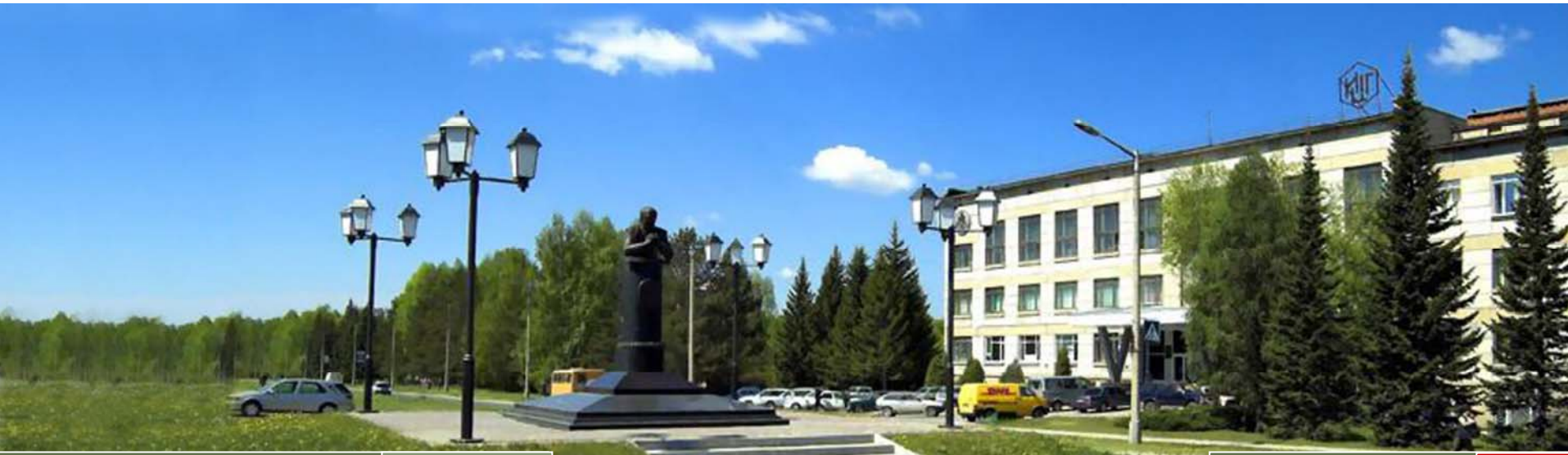
COMPUTATIONAL SYSTEMS BIOLOGY

in the Institute of Cytology and Genetics, Siberian Bransh of the Russian Academy of Sciences (ICG SB RAS)

Eds.
Nikolay A. Kolchanov,
Dagmara P. Furman and
Viatcheslav A. Mordvinov



2006

Design by Andrey V. Kharkevich

Annotation

The Laboratory of Theoretical Genetics, SB of RAS (Novosibirsk, Russia) proudly presents the most important results of research it has been conducting into computational systems biology over years.

The presentation contains the introduction and seven chapters devoted to the different trends in research developed by the Laboratory.

The Overview (Slides 4 - 12) gives an overview of the Institute of Cytology and Genetics, SB of RAS, of which the Laboratory is a part; provides a listing of the scientific institutes of the Siberian Branch of the RAS that are active in doing bioinformatics research; contains data on within-institute, Russian and international contacts of the Laboratory. The conferences organized by the Laboratory are extensively covered too.

There is an Introduction (Slides 13 - 15), a general Contents (Slide 16) and, eventually, the results of work arranged into chapters according to where they belong. For easier navigation, each chapter has the index.

Finally, there is a list of publications by Laboratory staff, a listing of grants awarded in support of research and a listing of databases and software that have been developed in the Laboratory and can be accessible via the Internet.

Contact addresses:
Institute of Cytology and Genetics, SB of RAS
http://www.bionet.nsc.ru/indexEngl.html

Nikolay A. Kolchanov, Doctor of Biological Sciences, Professor,
Academician of the RAS,
Director of the IC&G, Head of the Department of Systems Biology,
Head of the Chair of Informational Biology of the Novosibirsk State University
http://wwwmgs.bionet.nsc.ru/mgs/gnw/
kol@bionet.nsc.ru

Dagmara P. Furman, Doctor of Biological Sciences, Vice-Head of the Chair of Informational Biology of the Novosibirsk State University
furman@bionet.nsc.ru

Viatcheslav A. Mordvinov, Ph.D., Head of the Sector of Functional Genomics of IC&G, SB RAS
mordvin@bionet.nsc.ru

Bioinformatics in the Siberian Branch of the Russian Academy of Sciences

A range of institutes and the Novosibirsk State University are involved

Research Institutes

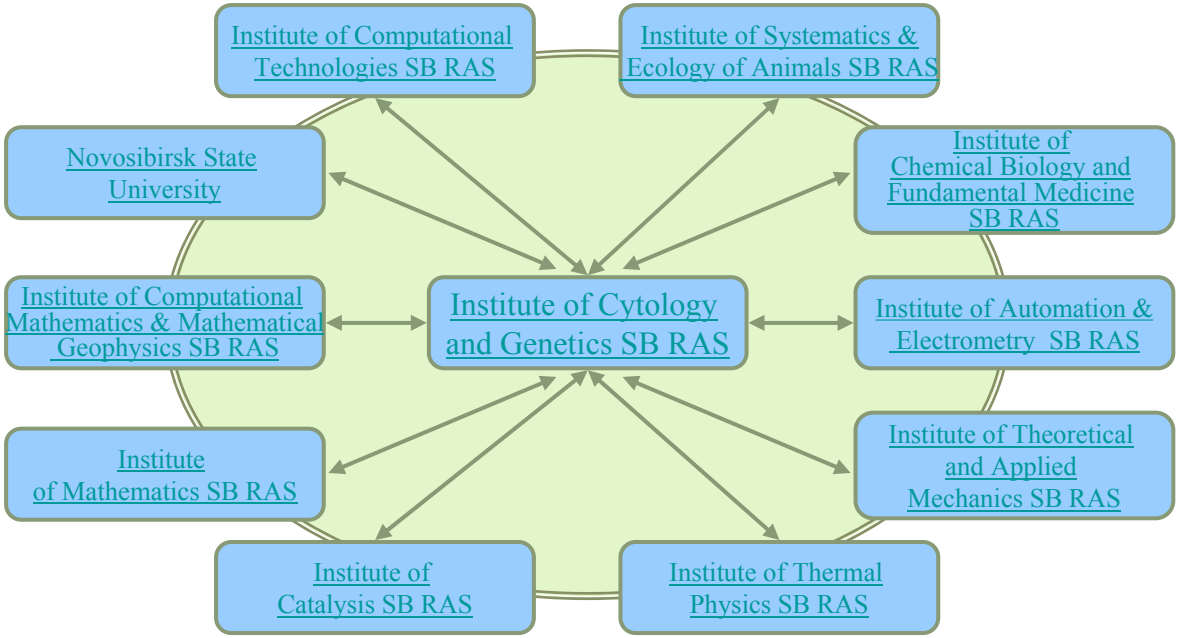
- Institute of Cytology and Genetics (Novosibirsk);
- S.L. Sobolev's Institute of Mathematics (Novosibirsk);
- Institute of Theoretical and Applied Mechanics (Novosibirsk);
- Institute of Thermal Physics by name of S.S. Kutateladze (Novosibirsk);
- Institute of Computational Mathematics and Mathematical Geophysics (Novosibirsk);
- Institute of Computational Technologies (Novosibirsk);
- Institute of Systematics & Ecology of Animals (Novosibirsk);
- Institute of Chemical Biology and Fundamental Medicine (Novosibirsk);
- Institute of Catalysis (Novosibirsk);
- Institute of Automation and Electrometry (Novosibirsk);
- Urga Scientific-research Institute of Informational Technologies (Khanty-Mansyisk);
- Institute of Biophysics (Krasnoyarsk).

Novosibirsk State University

- Department of Natural Sciences,
- Department of Information Technologies,
- Department of Mechanics and Mathematics,
- Department of Physics,
- Chair of Information Biology of the Department of Natural Sciences

Bioinformatics in the Siberian Branch of the Russian Academy of Sciences

A range of institutes and the Novosibirsk State University are involved



Bioinformatics network staff of the SB RAS is about 200 persons, including permanent research positions, students, magisters, postgraduate students.

Bioinformatics in the Novosibirsk state university

http://www.nsu.ru/english/

Rector:
Nikolay S. Dikansky, Professor, Corresponding Member of RAS
Pro-Rector for Information Technologies and Computer Equipment
Anatoly M. Fedotov, Professor, Corresponding Member of RAS
e-mail:aak@srd.nsu.ru,

DEPARTMENT OF NATURAL SCIENCES
Dean: Dr.Sci., Professor Vladimir A. Reznikov
e-mail: decan@fen.nsu.ru
URL:http://www.fen.nsu.ru

CHAIR OF INFORMATIONAL BIOLOGY
Head: Professor, Academician of RAS, Nikolay A. Kolchanov
URL:http://www.bionet.nsc.ru/chair/cib/
e-mail: kol@bionet.nsc.ru
Deputy Head: Dr.Sci., Dagmara P. Furman
e-mail: furman@bionet.nsc.ru

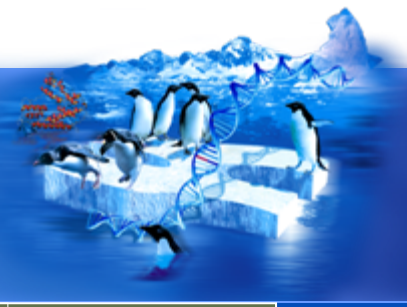
DEPARTMENT OF INFORMATION TECHNOLOGIES
Dean: Professor, Dr.Sci., Michail M. Lavrentiev
e-mail: dekanat@ccfit.nsu.ru
URL :http://www.fit.nsu.ru

DEPARTMENT OF MECHANICS AND MATHEMATICS

Dean: Professor, Dr. Sci., Corresponding member of RAS Sergey S. Goncharov
e-mail: mmf@nsu.ru
URL:http://mmfd.nsu.ru

DEPARTMENT OF PHYSICS
Dean: Professor, Dr.Sci., Andrey V. Arzhannikov
e-mail: dean@phys.nsu.ru
URL :http://www.phys.nsu.ru

BIOINFORMATICS network of the Novosibirsk State University provides training for more than 60 students, undergraduates, post-graduated students



Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences is the leading Siberian institute for bioinformatics

General Information
Director: Nikolay A. Kolchanov, Professor, Academician of the Russian Academy of Sciences

Deputy Directors:
- Suren M. Zakian, Professor, Doctor of Biological Sciences
- Sergey G. Veprev, PhD

Scientific Secretary: Doctor of Biological Sciences Alexander V. Osadchuk

Staff: about 1,000
The Institute sees integration of molecular, cell, ontogenetic, populational and computer-assisted studies for a better understanding of genetic mechanisms of variability and evolution as its main mission.
The main trends in research are:
- In silico methods in systems biology.
- Structural and functional organization of genetic material at a genomic, chromosomal and genetic level.
- Genome reconstruction; transgenesis in animals and plants.
- Molecular genetic and genetic evolutionary bases for the operation of physiological systems responsible for the most vital functions. Chromosomal and gene diagnostics of inherited and multifactorial diseases.
- Genetic-evolutionary and ecological bases of population biology and biodiversity.

Computational systems biology in IC&G
Department of Systems Biology

http://wwwmgs.bionet.nsc.ru/mgs/gnw/

Head: Dr.Sci., Professor, Academician of RAS, Nikolay A. Kolchanov

Main research areas:

Gene Networks: reconstruction, computer analysis and modeling

- Computational genomics
- Computational transcriptomics
- Computational proteomics
- Molecular pathologies
- Computational evolutionary biology
- Plant and animal development: computer analysis and modeling
- Transgenesis optimization
- Knowledge discovery and data mining in bioinformatics
- High-performance calculations in bioinformatics

Computational systems biology in IC&G
Department of Systems Biology

Grants awarded to the laboratory of theoretical genetics since 2000

Russian Foundation for Basic Research – 40 grants awarded

International Foundations - 50 grants awarded

Grants awarded and projects completed over the past 5 years

http://wwwmgs.bionet.nsc.ru/mgs/gnw/
Head of the Laboratory–Dr. Sci., professor, Corresponding Member of RAS, Nikolay A. Kolchanov

Conferences and workshops arranged by the IC&G and the Laboratory of theoretical genetics

1984, 1986, and 1988:
1st, 2nd and 3rd Symposia "Theoretical Research and Data Banks in Molecular Biology and Genetics", Novosibirsk, Russia

International Conference "Modeling and computer methods in molecular biology and genetics". August 24-31, 1990, Novosibirsk, Russia

The First International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'1998), August 24-31, Novosibirsk, Russia
http://www.bionet.nsc.ru/bgrs/

Workshop "ACTUAL PROBLEMS OF INFORMATIONAL BIOLOGY", August, 6, 1999, Novosibirsk, Russia,
http://wwwmgs.bionet.nsc.ru/mgs/info/workshop99.html

"BIODIVERSITY AND DYNAMICS OF ECOSYSTEMS IN NORTH EURASIA" (BDENE'2000)
August 21—26, 2000, Novosibirsk, Russia
http://www.bionet.nsc.ru/meeting/bdne2000/

The Second International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2000), August 7 - 11, Novosibirsk, Russia
http://www.bionet.nsc.ru/bgrs2000/index_local.html

First Workshop on Information Technologies Application to Problems of Biodiversity and Dynamics of Ecosystems in North Eurasia (WITA-2001), July 9 -14, 2001, Novosibirsk, Russia
http://www.bionet.nsc.ru/meeting/bdne2001/index_eng.html

The Third International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002), July 14 - 20, 2002, Novosibirsk, Russia
http://www.bionet.nsc.ru/meeting/bgrs2002/index_local.html

INTAS Workshop «Entangling Mathematics, Life Sciences and Information Technology: Biomatics, a Science of multi-scale complexity», July 19 - 20, 2002, Novosibirsk, Russia

The Forth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2004), July 25-30, Novosibirsk, Russia
http://www.bionet.nsc.ru/meeting/bgrs2004/

«INTAS/ FP6 'EU-NIS Partnering in Bio-Informatics' Event», July 29 - 30, 2004, Novosibirsk, Russia

The Fifth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2006), July 16-22, 2006, Novosibirsk, Russia
http://www.bionet.nsc.ru/meeting/bgrs2006/

Bioinformatics: collaboration of the Department of Systems Biology with Russian and foreign partners

1. S.L. Sobolev's Institute of Mathematics of SB RAS;
2. Institute of Theoretical and Applied Mechanics of SB RAS;
3. Institute of Thermal Physics by name of S.S. Kutateladze of SB RAS;
4. Institute of Computational Mathematics and Mathematical Geophysics of SB RAS;
5. Institute of Computational Technologies of SB RAS;
6. Institute of Systematics & Ecology of Animals of SB RAS;
7. Urga Scientific-research Institute of Informational Technologies (Khanty-Mansyisk) of SB RAS;
8. Institute of Biophysics (Krasnoyarsk) of SB RAS;
9. Institute of Chemical Biology and Fundamental Medicine of SB RAS;
10. ISMAC – Institute of Macromolecule Research, Genoa, Italy;
11. Institute of Genomics and Bioinformatics, University of California, Irvine, USA;
12. Institute of Modern Biomedical Technologies, Milan, Italy;
13. Center of Bioinformatics of the Pennsylvania University, USA;
14. Pharmaceutical Company GlaxoSmithKline, UK;
15. University of Bilefield, Germany;
16. Nottingham University, UK;
17. Technological Institute Kiotsu, Japan;
18. Institute of Experimental Pathology, Oncology, and Radiobiology of NAS of Ukraine, Kiev.

Bioinformatics network of the Institute of Cytology and Genetics

- Laboratory of Theoretical Genetics (Nikolay A. Kolchanov, Dr.Sci., Professor, Corresponding Member of RAS);
- Laboratory of Animal Molecular Genetics (Aida G. Romaschenko, Ph.D.);
- Laboratory of Gene Expression Control (Tatiana I. Merkulova, Dr.Sci.);
- Laboratory of Evolutionary Cell Biology (Elina M. Baricheva, Ph.D.);
- Laboratory of Cell Cycle Genetics (Leonid V. Omelyanchuk, Dr.Sci.);
- Laboratory of Morphology and Function of Cell Structures (Nikolay B. Rubtsov, Dr. Sci.);
- Laboratory of Human and Animal Genetics (Alexander S. Grafodatsky, Dr.Sci., Professor);
- Plant Gene Engineering Laboratory (Alekxey V. Kochetov, Ph.D.);
- Laboratory of Plant Heterosis (Vladimir K. Shumny, Dr.Sci., Professor, Full Member of RAS);
- Laboratory of Genetic Recombination and Segregation (Pavel M. Borodin, Dr.Sci., Professor);
- Sector of Molecular Evolution (Yury G. Matushkin, Ph.D.);
- Sector of Molecular Genetic Mechanisms of Protein-Nuclein Intercations (Ludmila K. Savinkova, Ph.D.);
- Sector of Functional Genomics (Viatcheslav A. Mordvinov, Ph.D.);
- Sector of Medical Genetics (Natalya G. Kolosova, Dr.Sci.);
- Sector of Wheat Genetics (Nikolay P. Goncharov, Dr.Sci.);
- Sector of Plant Breeding and Genetics (Vassily S. Koval , Ph.D.).

Introduction

The last 15-20 years in development of biology were marked with accumulation of unprecedentedly huge arrays of experimental data. The information was amassed with exclusively high rates due to the advent of highly efficient experimental technologies that provided for high throughput genomic sequencing; of functional genomics technologies allowing investigation of expression dynamics of large groups of genes using expression DNA chips; of proteomics methods giving the possibility to analyze protein compositions of cells, tissues, and organs and to determine their primary and spatial structures, assess the dynamics of the cell proteome (changes in protein concentrations in the cells), and reconstruct the networks of protein–protein interactions; and of metabolomics, in particular, high resolution mass spectrometry study of cell metabolites, determination of their concentrations, and distribution of metabolic fluxes in the cells with a concurrent investigation of the dynamics of thousands metabolites in an individual cell.

Analysis, comprehension, and use of the tremendous volumes of experimental data reflecting the intricate processes underlying the functioning of molecular genetic systems are unfeasible in principle without the systems approach and involvement of the state-of-the-art information and computer technologies and efficient mathematical methods for data analysis and simulation of biological systems and processes.

The need in solving these problems initiated the birth of a new science—postgenomic bioinformatics or systems biology *in silico*.

Welcome to BGRS\SB-2010

Dear colleagues,

The Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences will be hosting the International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2010) in Novosibirsk, Russia, from 20-27 June 2010. This Conference is the seventh in the series since the first BGRS event held in 1998.

As one of the key disciplines in modern biology, bioinformatics is a rapidly developing science. Consequently, each of the past BGRS events was focused on the most important topics of that time. To keep this tradition going, BGRS\SB-2010 will be centered on bioinformatics and systems biology.

Systems biology largely focuses on the study of the organization and operation of the biological systems at various levels: molecular genetic entities, cells, tissues, organs and organisms on the basis of information encoded in their genomes.

Systems biology strongly depends on high-performance experimental technologies:

- sequencing of genomic DNA, analysis of its between-population and evolutionary variation;
- study of the expression of genes and gene complexes using biochips-based modalities;
- structural and functional analysis of proteins and metabolites using mass spectrometric methods;
- study of the structural and functional organization of biological objects (macromolecules, chromosomes, cells, tissues, organs, organisms) using modern microscopic methods;
- construction of artificial molecular genetic systems using genetic engineering techniques.

In systems biology, bioinformatics methods play by far the most important role. With them, the researcher can:

- accumulate and integrate experimental information in databases;
- bring this information to computer analysis;
- perform mathematical modeling of the structural and functional organization of living systems;
- predict new properties of living systems;
- design new rounds of experimental research.

Welcome to BGRS\SB-2010

Systems biology follows in the steps of physics where no experiment or its interpretation is possible until profound theoretical and computer-aided analyses of the systems and processes being studied are made.

Consequently, BGRS\SB-2010 will have special focus on research efforts that are based on integration of experimental and computer-based/theoretical approaches.

The following are the particular studies, in which bioinformatics and systems biology meet and which are of special interest to the Conference:

- genomics;
- chromosomics;
- transcriptomics;
- proteomics;
- metabolomics;
- reconstruction and modeling of gene networks;
- cell biology;
- physiological genetics;
- developmental biology;
- evolutionary biology;
- synthetic biology;
- medical biology and pharmacology;
- biotechnology.

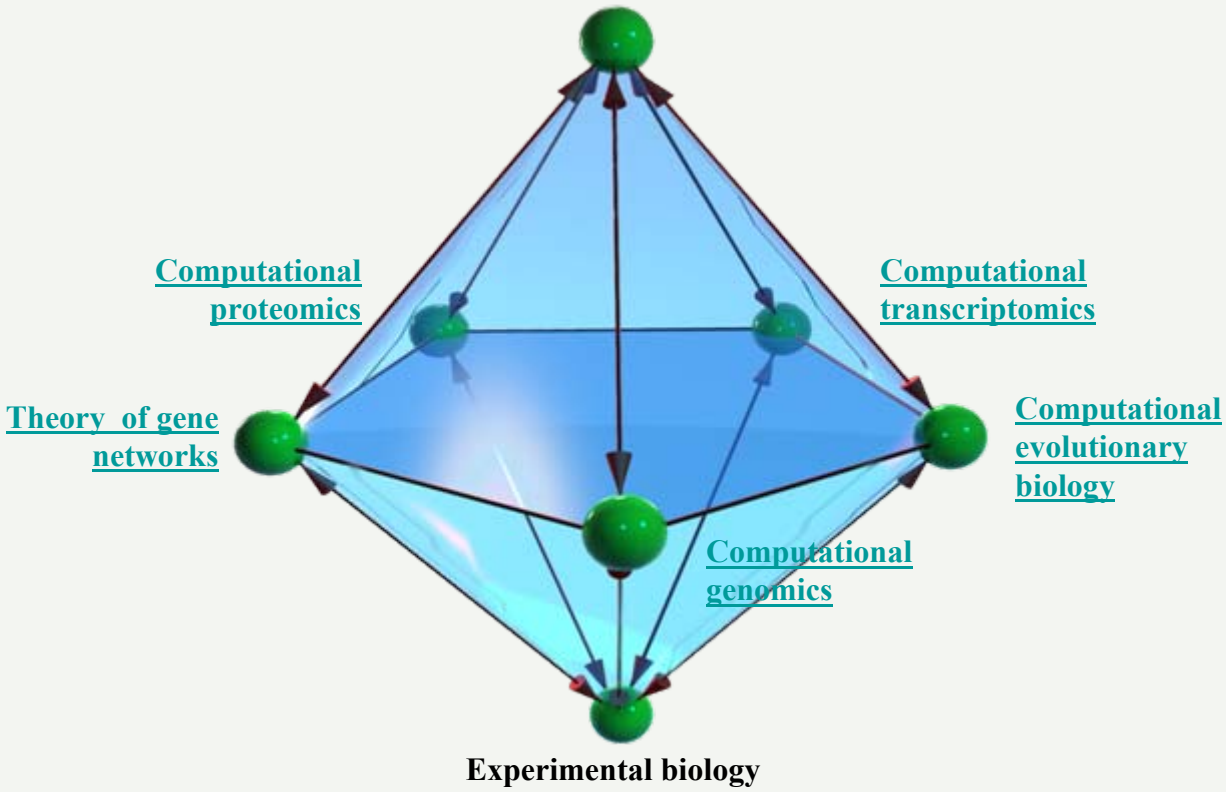
The results of the most recent research in these fields will be presented. The Conference program will include plenary papers, session papers and round tables. As previously, we are hoping to hear from those who wish to step down as Session Chairs and about their suggestions for the sessions they wish to chair. The Session Chairs will be offered special privileges at the Conference.

You are very welcome to participate in the 7th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology - BGRS\SB-2010.

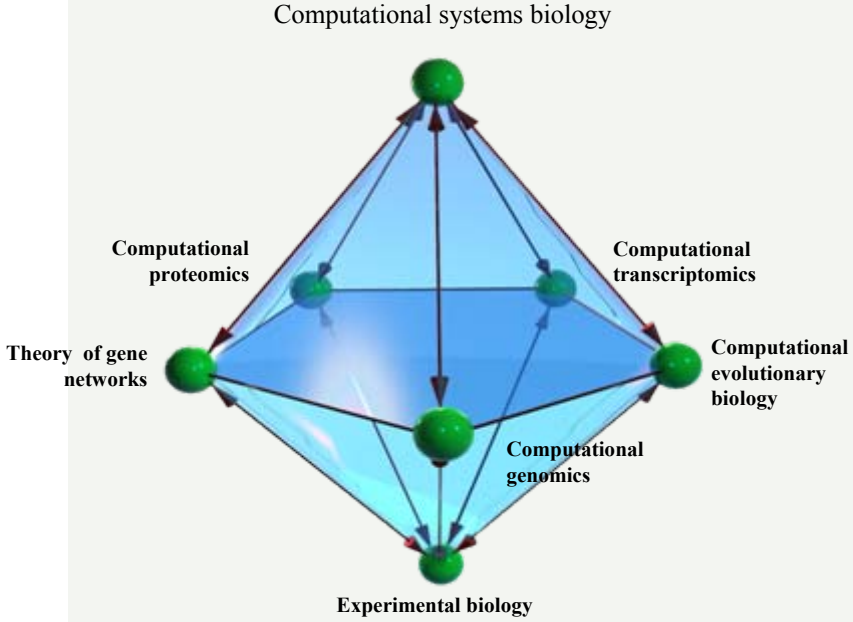
The Conference's official site is http://www.bionet.nsc.ru/meeting/bgrs2010/index.html
The email address is bgrs_sb2010@bionet.nsc.ru

Contents

Computational systems biology: from analysis to synthesis



Computational systems biology: from analysis to synthesis



GOAL:

Reconstruction, on the basis of the information contained in the genomes, of knowledge about the systems and processes responsible for CELL and ORGANISM replication: functioning and interaction with the environment.

METHODOLOGY:

Integration of methods used in bioinformatics and modern experimental biology

BIOINFORMATICS:

computer-based integration of experimental data obtained using analytical molecular biology methods in; mathematical modeling of molecular genetic systems and processes

EXPERIMENTAL BIOLOGY:

Structural and functional genomics,
transcriptomics,
proteomics,
metabolomics,
cell biology.

Chapter 1

GENE NETWORKS: reconstruction, computer analysis and modeling

1.1. Gene Networks: principles of organization and mechanisms of operation and integration

1.1.1. GeneNet technology
1.1.2. GeneNet database: fragments of reconstructed gene networks

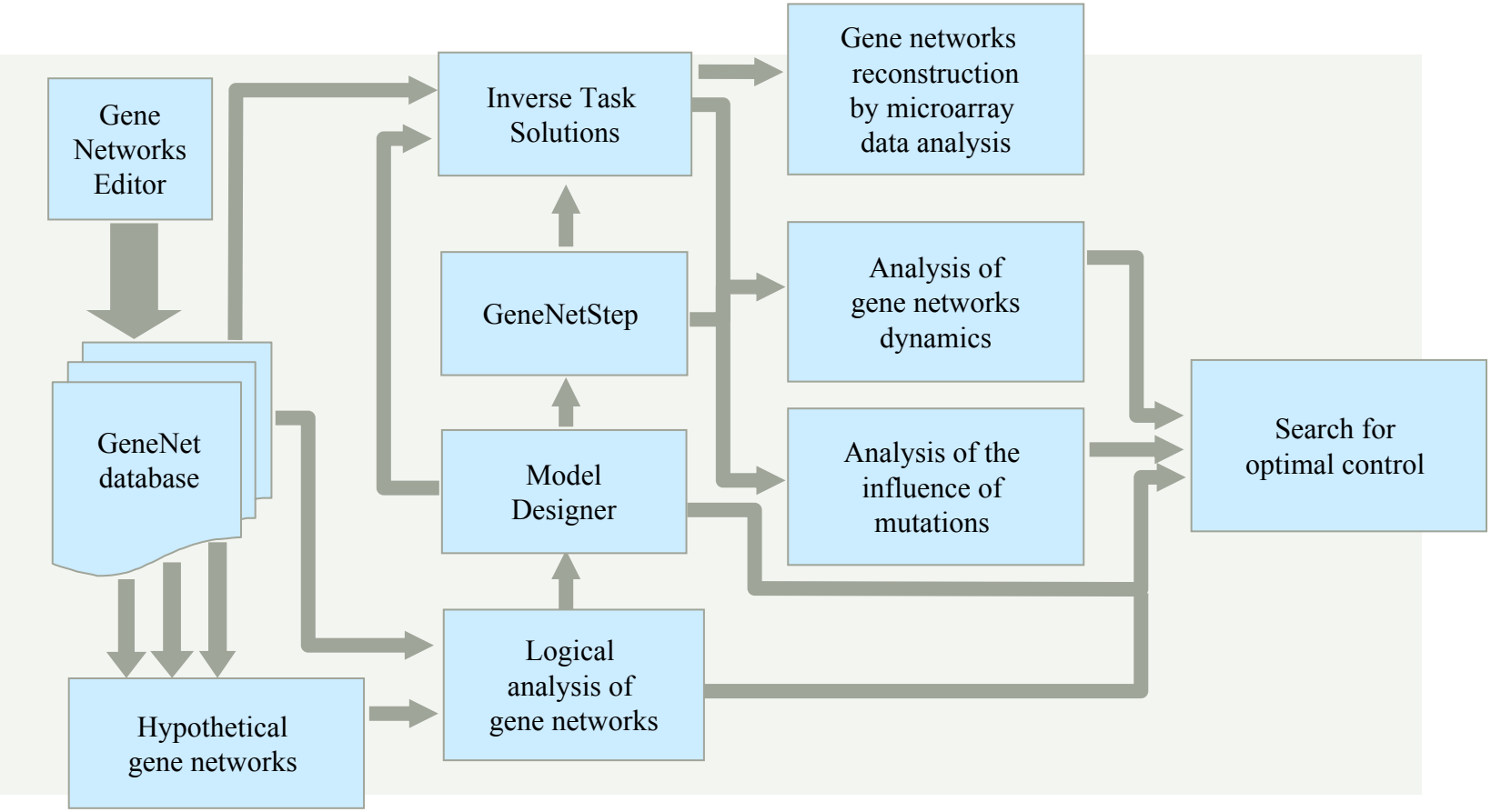
The gene network: the central object in systems biology

1. A gene network is a groups of genes that function in coordination with each other and control the development of a particular phenotypic character of an organism: molecular, biochemical, physiological, morphological, behavioral etc.

2. Any gene network includes:
genes; RNA and proteins they encode; metabolites; signal transmission pathways; metabolic pathways; positive and negative feedback regulatory circuitries.

Over one million works on the organization and function of gene networks have been published to date. To make the information contained in these publications manageable, special databases are required. Fact is, only 5-7 % of the publications have been processed to the effect. There is an urgent need for computer technologies with which the gene networks could be reconstructed by summing up experimental data available from scientific publications. GeneNetDiscovery, a technology that has been developed in the Laboratory of Theoretical Genetics, is one that does that.

GeneNetDiscovery: tools for gene networks reconstruction, computer analysis and modeling



GeneNetDiscovery:

- Accumulation of experimental data on structure-functional organization of gene networks controlling molecular genetic, biochemical, physiological, morphological, and other characteristics of organisms.
- Reconstruction of gene networks and metabolic pathways on the basis of annotation of experimental data represented in scientific publications and electronic databases.
- Analysis of gene networks structural and functional organization.
- Calculation of gene network micro-structural parameters:
  - Search for critical gene network elements, computer-assisted analysis and modeling of gene networks;
  - Search for a strongly connected sub-networks in the gene network graph;
  - Search for regulatory circuits of gene networks and the points of their intersections.
- Computer simulation of gene networks dynamics.
- Inverse task solutions.
- Search for optimal control.
- Analysis of mutations effect on gene networks functioning.
- Gene networks reconstruction by microarray data analysis.
- Hypothetical gene networks theory.

GeneNet Database

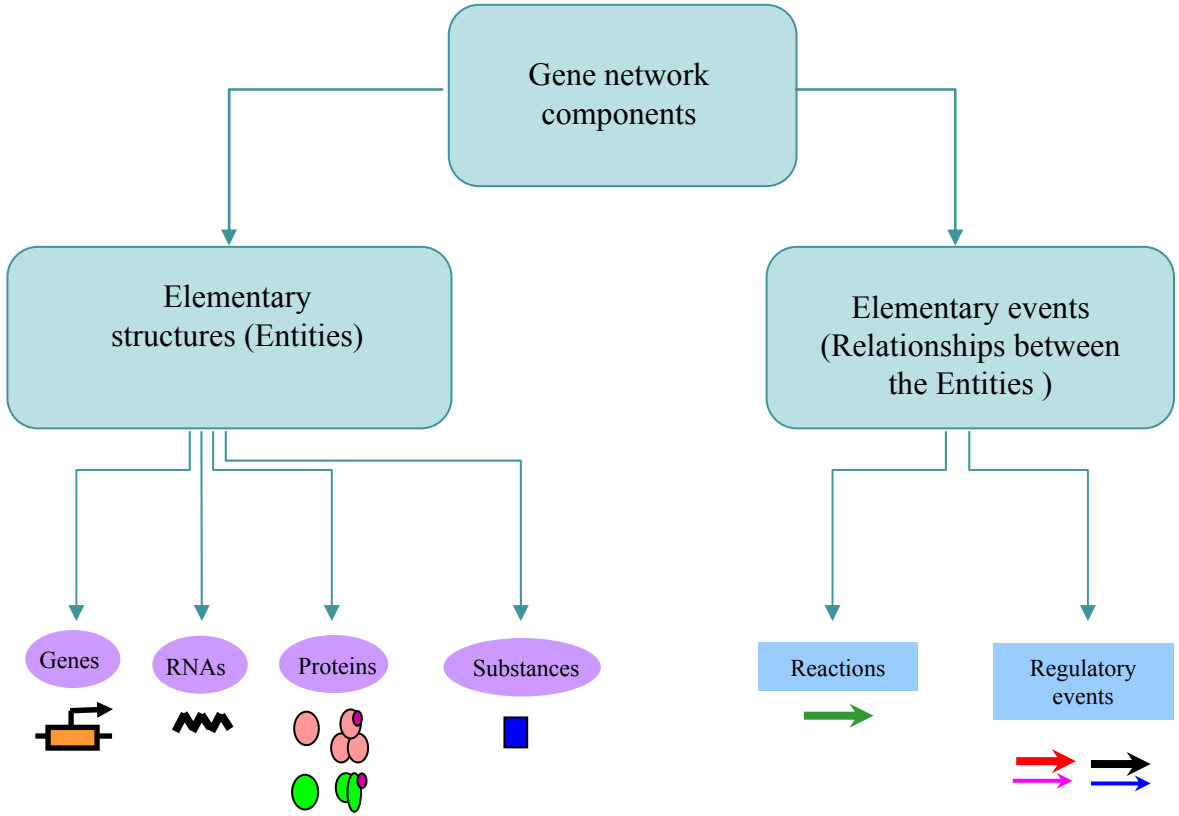
Missions

Reconstruction of gene networks and metabolic pathways on the basis of annotation of experimental data represented in scientific publications and electronic databases.

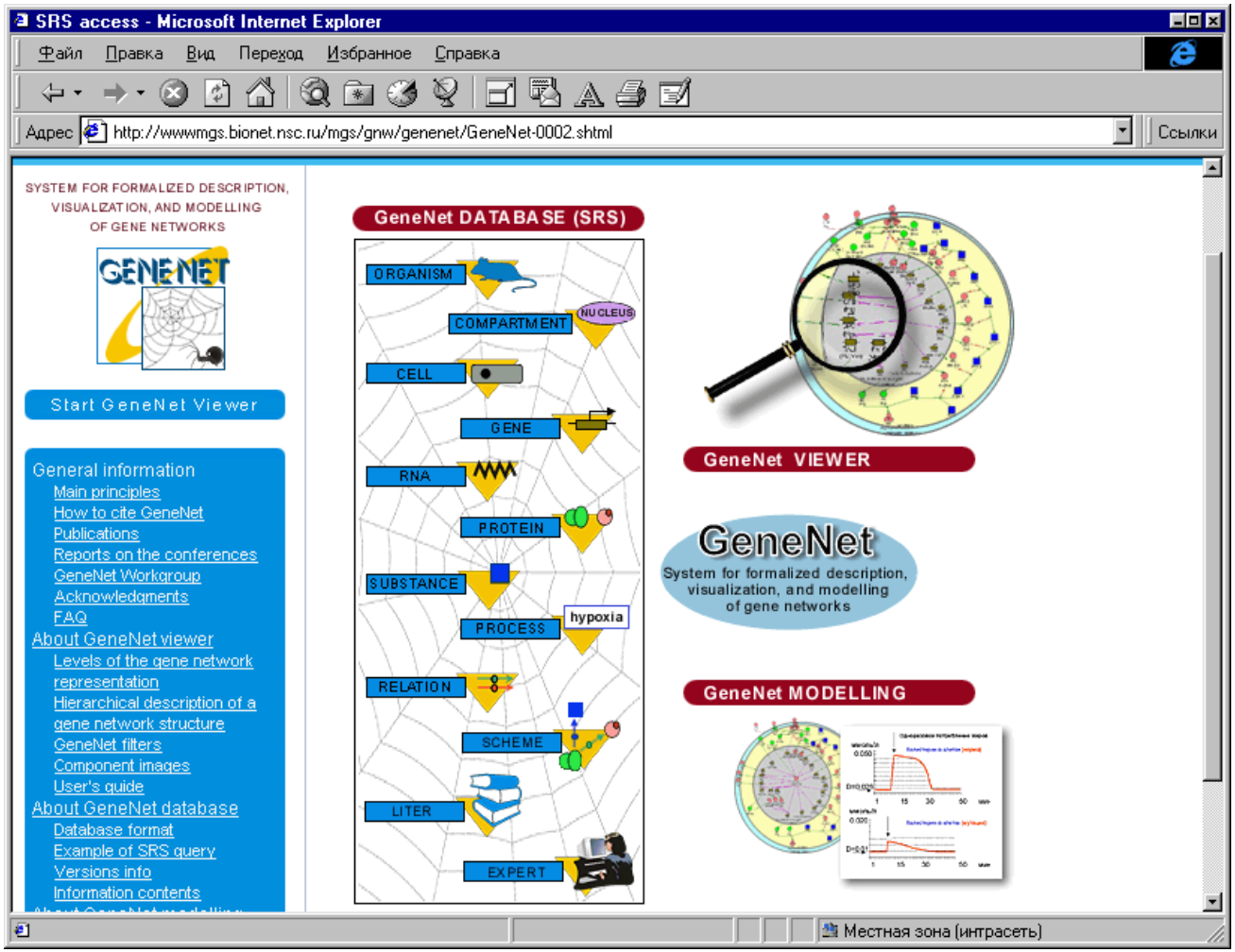
Accumulation of experimental data on structure-functional organization of gene networks controlling molecular genetic, biochemical, physiological, morphological,
and other characteristics of organisms.

1.1.1. GeneNet technology

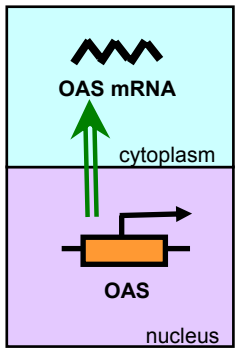
GeneNet technology: object-oriented approach; class hierarchy in the GeneNet database



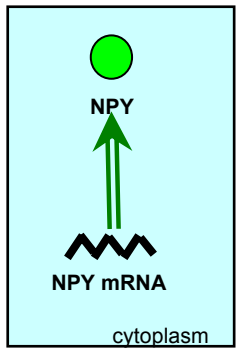
GeneNet technology



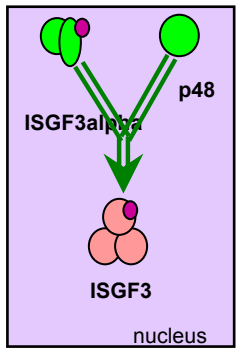
GeneNet technology: examples of elementary structures and events significant for gene network operation



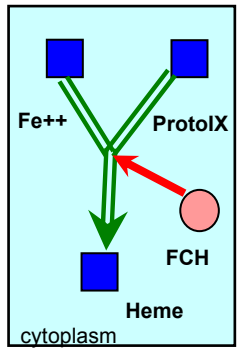
Transcription



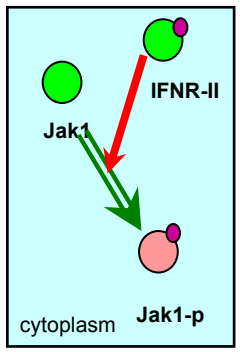
Translation



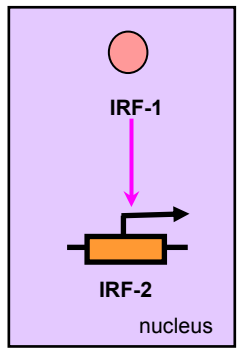
Multimerization



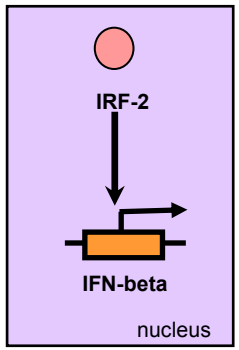
Enzymatic synthesis



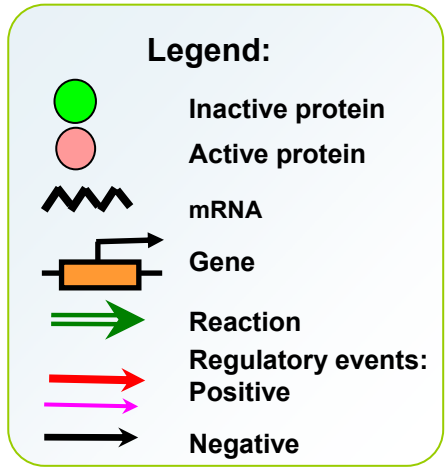
Phosphorylation



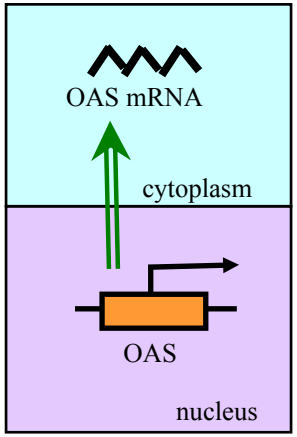
Positive effect on the gene expression



Repression of the gene



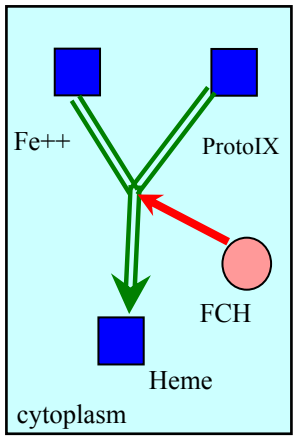
GeneNet technology: encoding of the gene network operations



Transcription

The protein encoded by the human 2'-5' oligoadenylate synthetase gene is expressed in the cell cytoplasm. The process is indirect (intermediate processes, such as transcription, RNA processing, and splicing are missing).

```
ID  <gene>Hs:OAS^nucleus -> <protein>Hs:OAS^cytoplasm
DT  17.5.1999; Ananko E.; created.
EF  indirect
RF  Wathelet M. et al., 1986
//
```

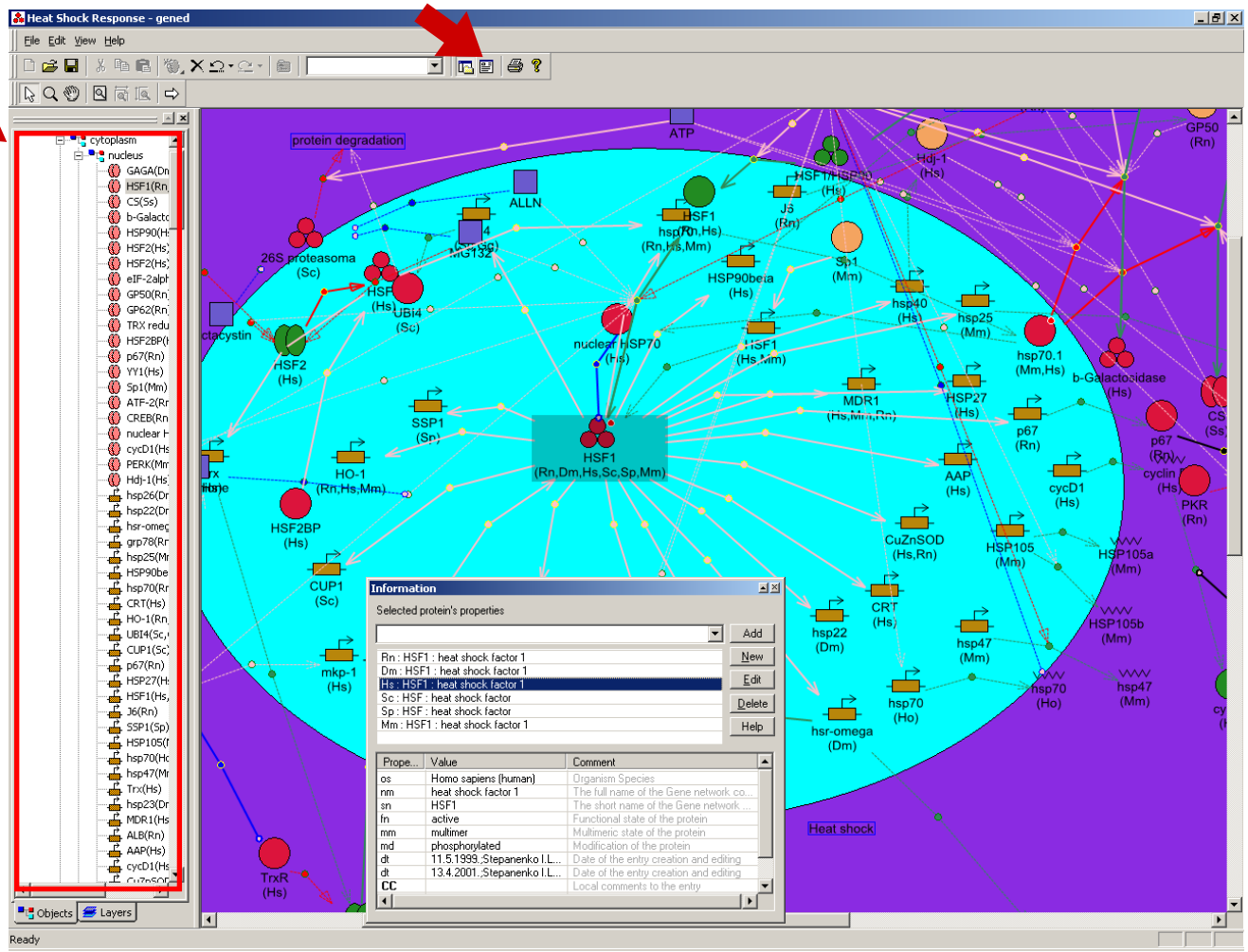


Enzymatic catalysis

In human mitochondrion, the ferrochelatase (FCH) catalyzes the Heme synthesis out of precursors (Fe++, Proto IX). Interaction between the substrate and enzyme is direct.

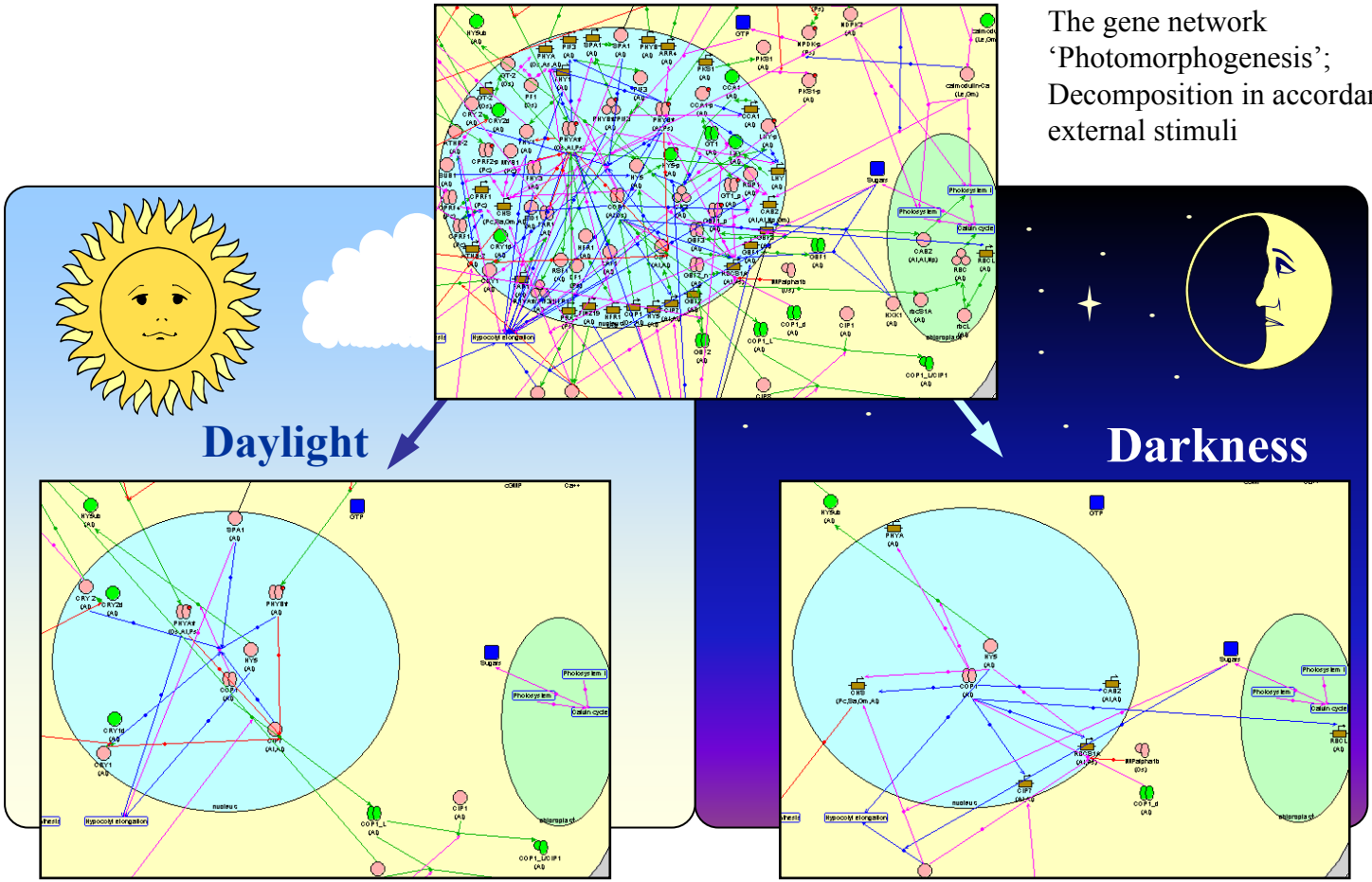
```
ID  <protein>Hs:FCH^mitochondrion ->>
<substance>Fe++^mitochondrion,
<substance>ProtoIX^mitochondrion ->
<substance>Heme^mitochondrion
DT  23.6.1999; Podkolodnaya O.A.; created.
AT  switch on
EF  direct
RF  Ponka P., 1997
//
```

Edit component properties



GeneNet technology: Interface of the GenEd editor for gene network formalized description, reconstruction and visualization

GeneNet technology: decomposition of the gene network



The gene network ‘Photomorphogenesis’; Decomposition in accordance with external stimuli

Informational content of the GeneNet database

Section	*Number of gene networks*
Cell cycle	4
Homeostasis	7
Endocrine regulation	4
Morphogenesis	14
Response of an organism to the external stimuli	13
Total	42

The latest publication on the GeneNet database

Nucleic Acids Research, 2005, Vol. 33, Database issue ***D425–D427***
doi:10.1093/nar/gki077

GeneNet in 2005

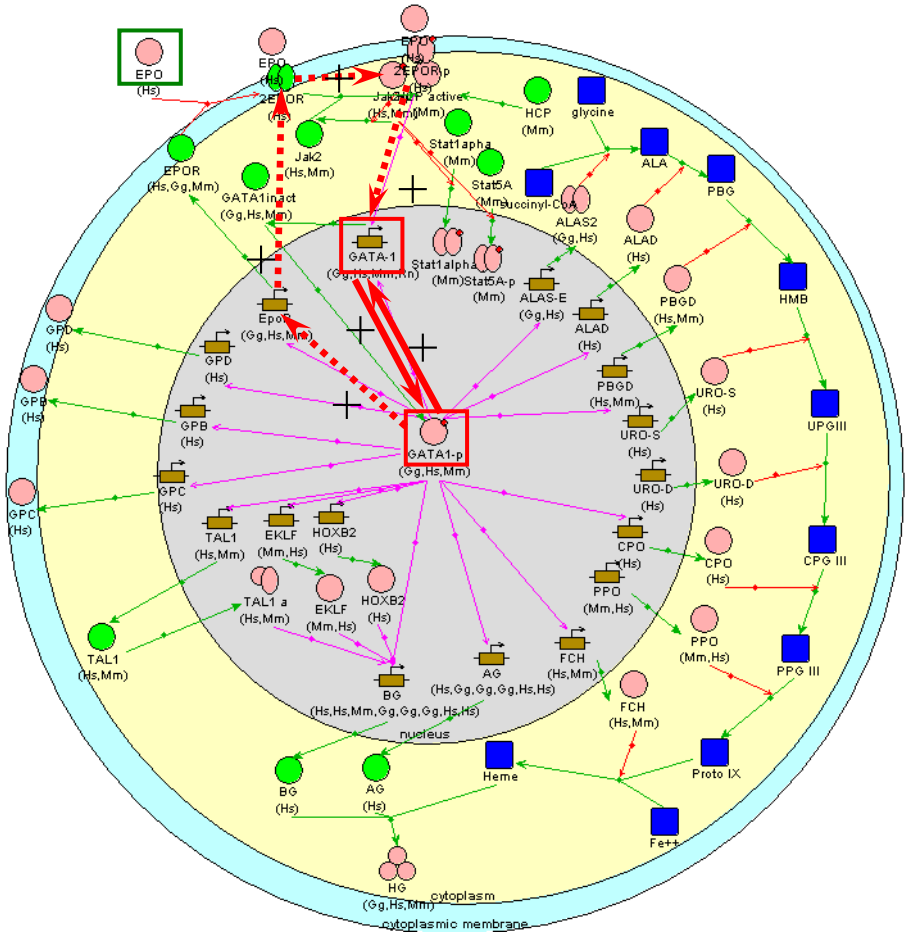
E. A. Ananko*, N. L. Podkolodny, I. L. Stepanenko, O. A. Podkolodnaya, D. A. Rasskazov, D. S. Miginsky, V. A. Likhoshvai, A. V. Ratushny, N. N. Podkolodnaya and N. A. Kolchanov

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Lavrentiev Avenue 10, Novosibirsk 630090, Russia

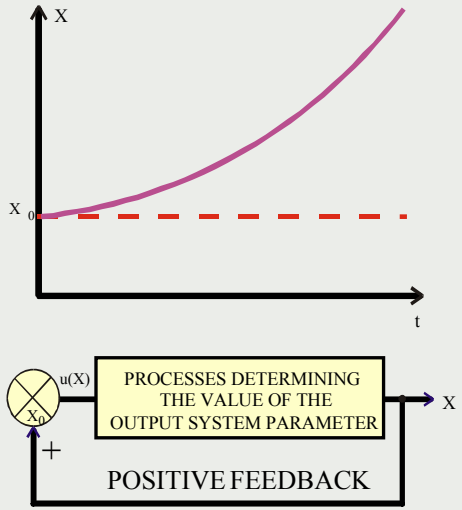
Received September 15, 2004; Revised and Accepted October 8, 2004

1.1.2. GeneNet database: fragments of reconstructed gene networks

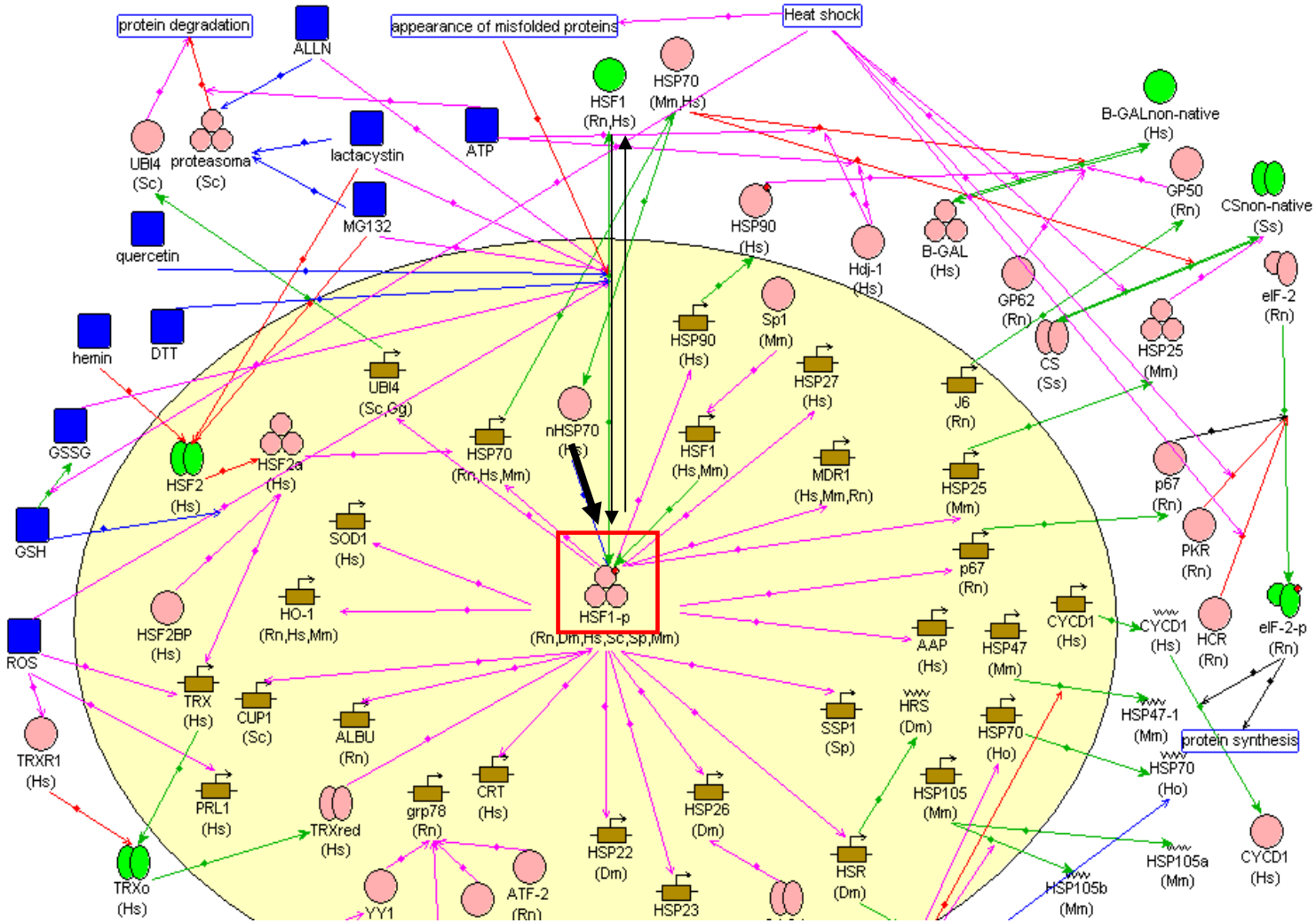
GeneNet database: a fragment of the gene network controlling erythrocyte maturation induced by Erythropoietin



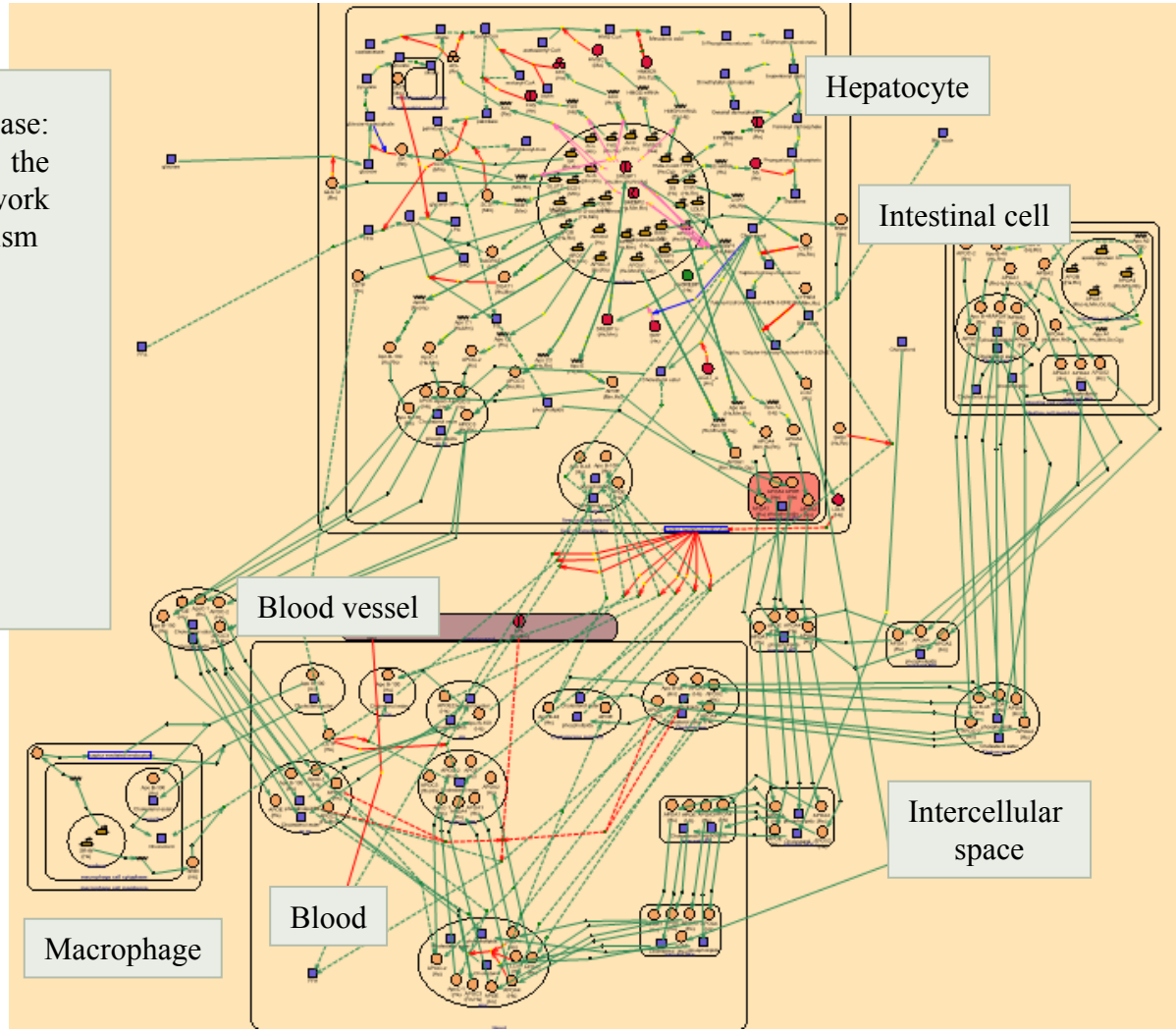
Principle scheme of regulatory contour with the positive feedback



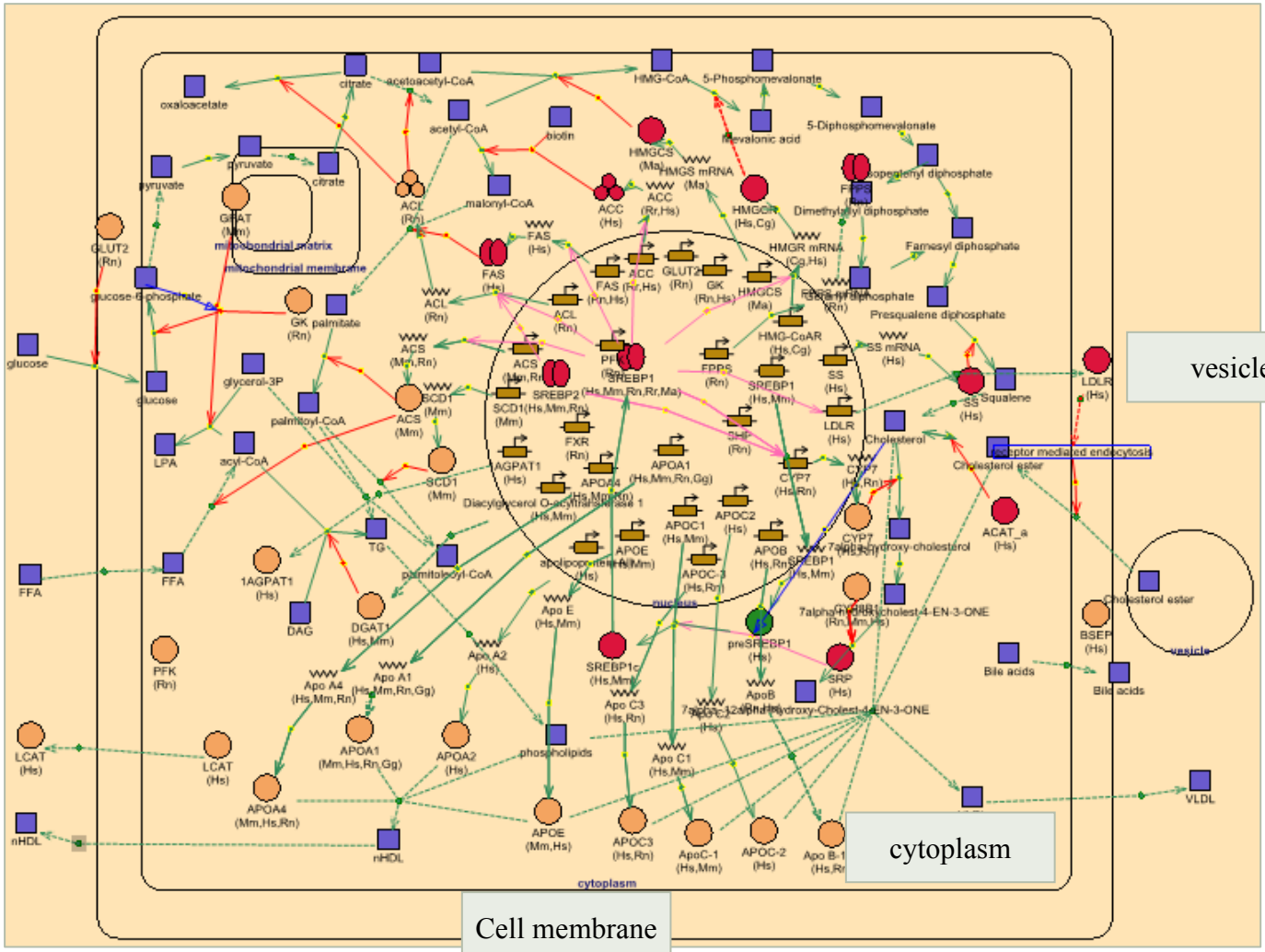
GeneNet database: a fragment of the heat shock response gene network



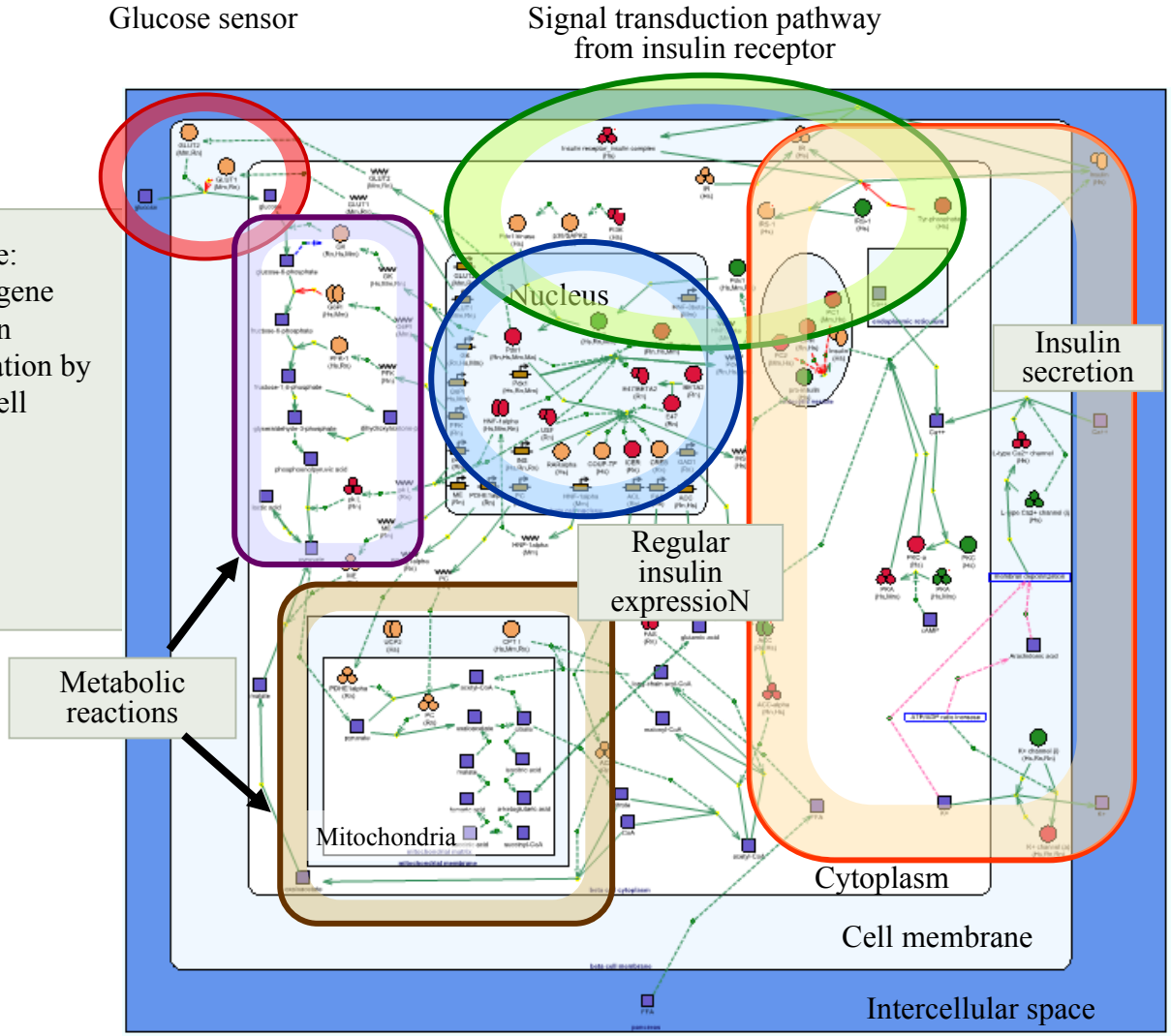
GeneNet database: fragments of the global gene network of lipid metabolism



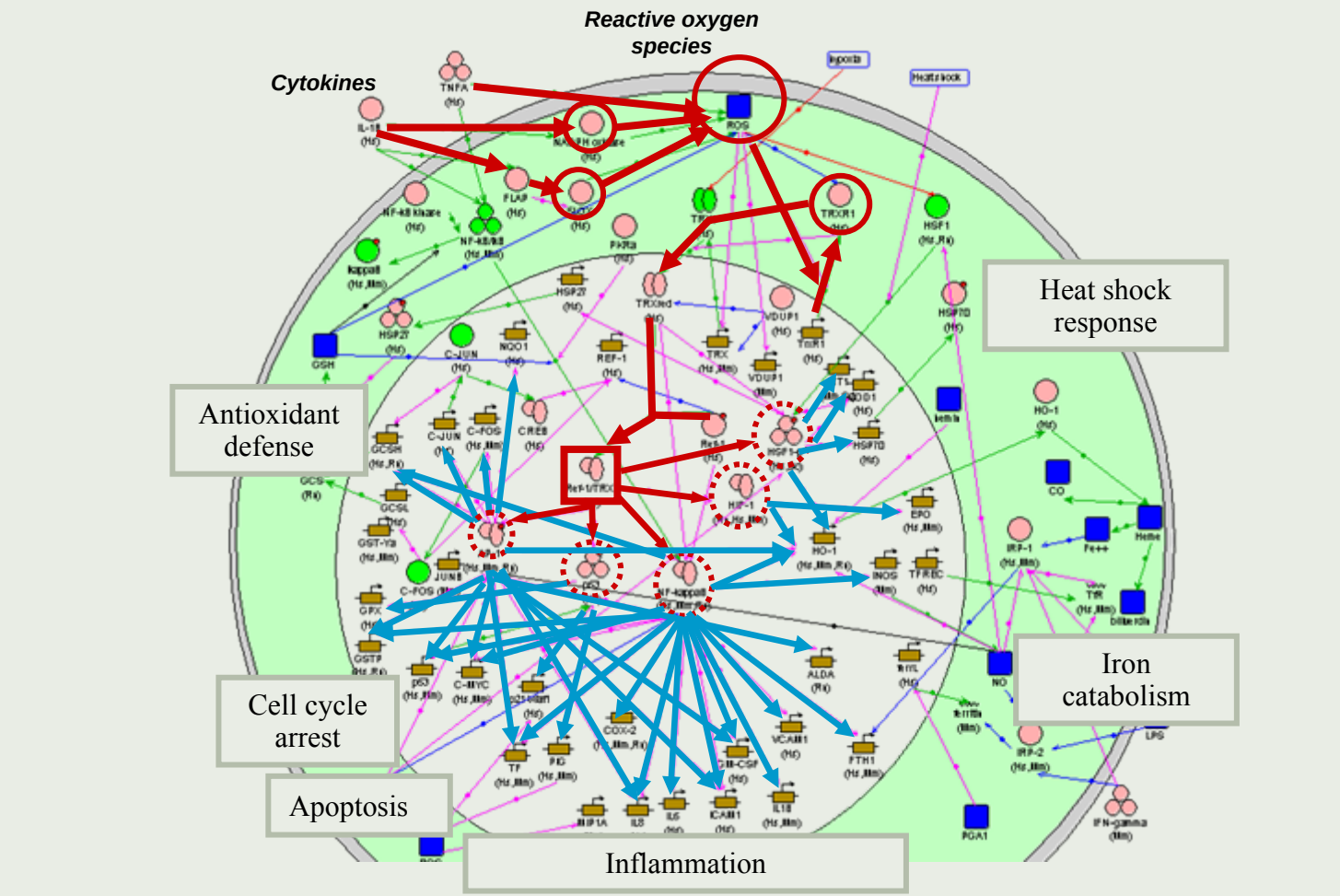
GeneNet database: a fragment of the network of lipid metabolism in the liver cell



GeneNet database: fragments of the gene network of insulin production regulation by pancreatic beta-cell



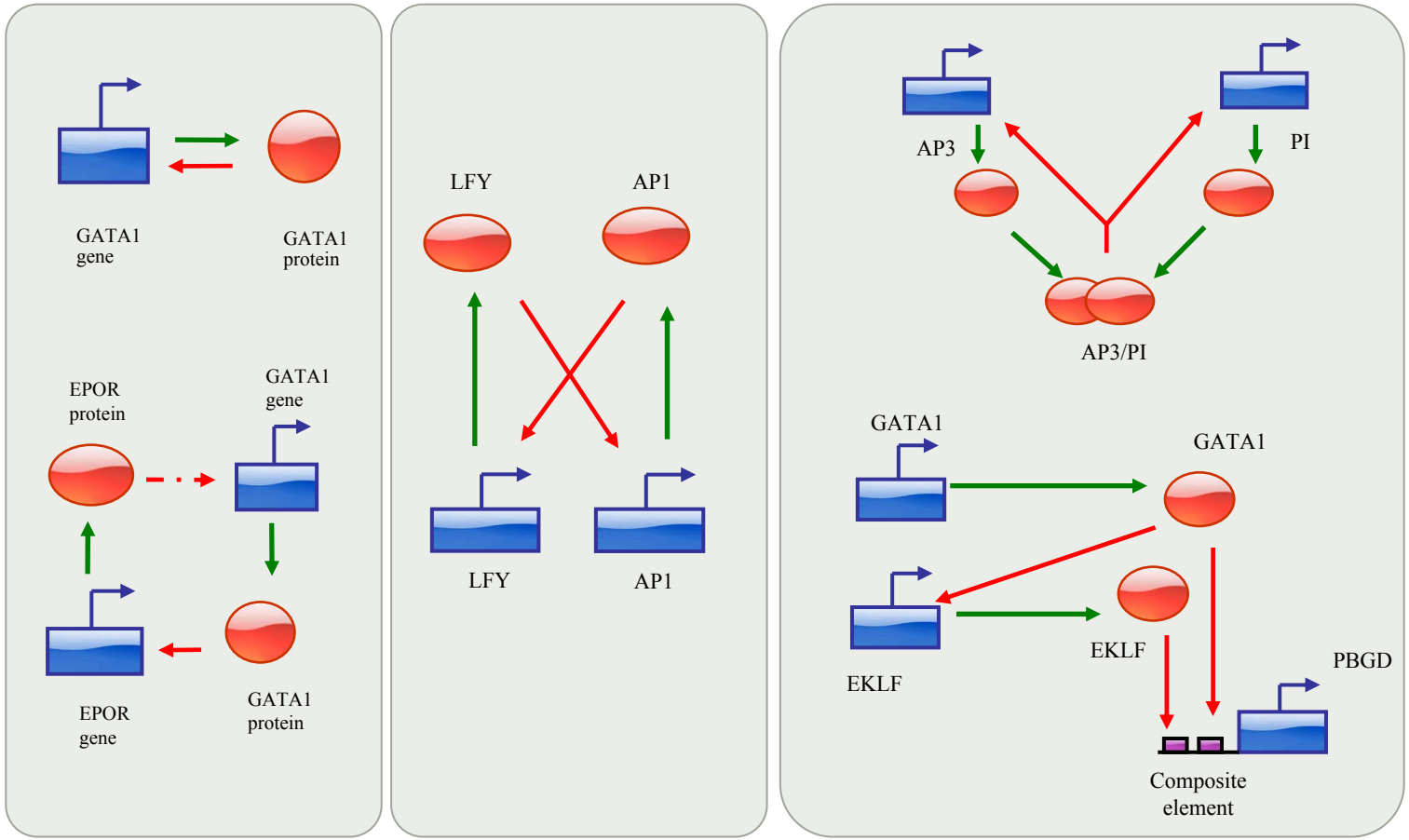
Gene networks that integrate: gene network of redox-regulation and cassette activation of gene networks interfering with it



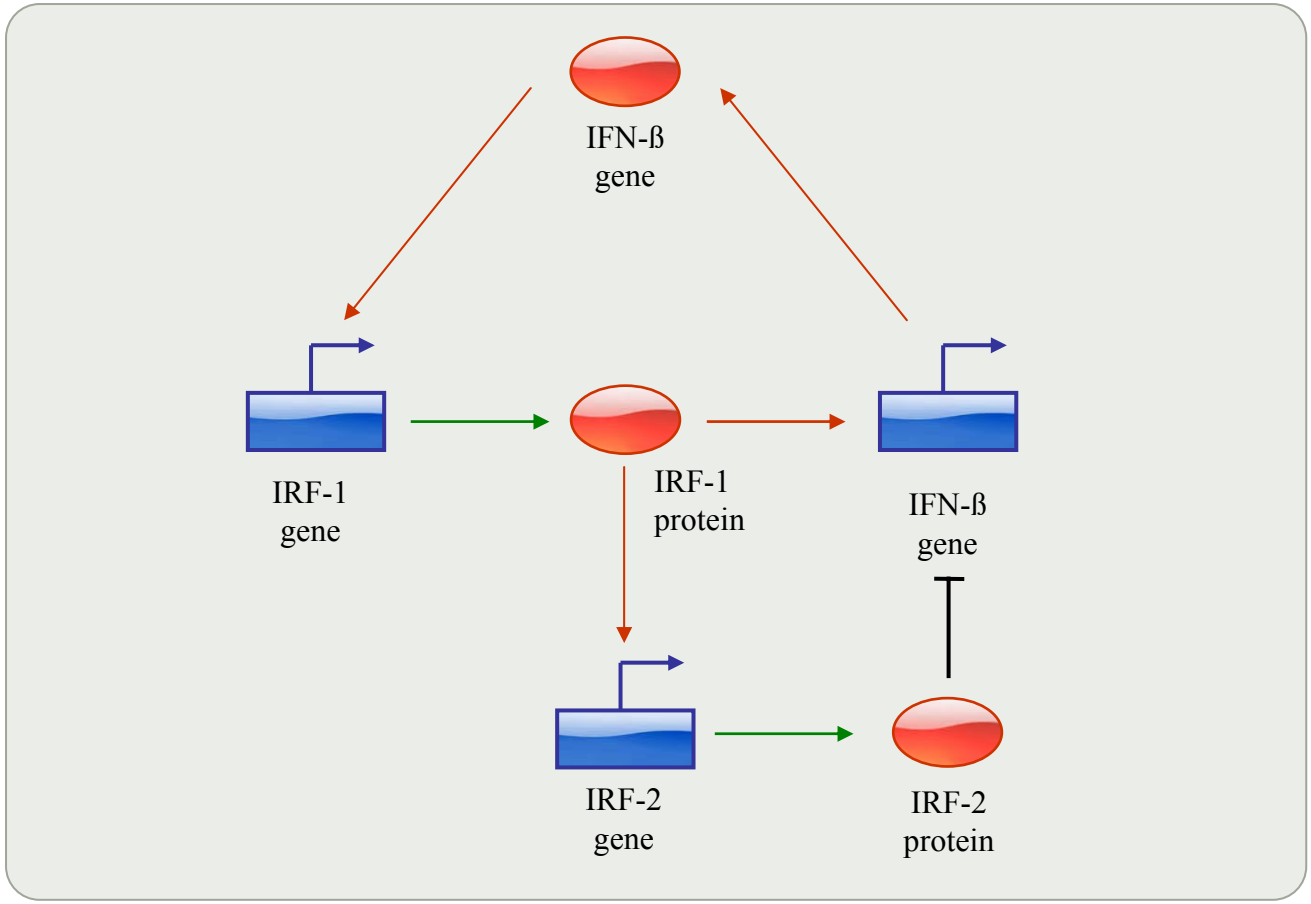
MACROSYSTEMIC MUTATIONS:

Mutationally modified function of an integrating gene network may at once result in a modification to the functions of many associated gene networks and, consequently, a coordinated change of the phenotype characters they control.

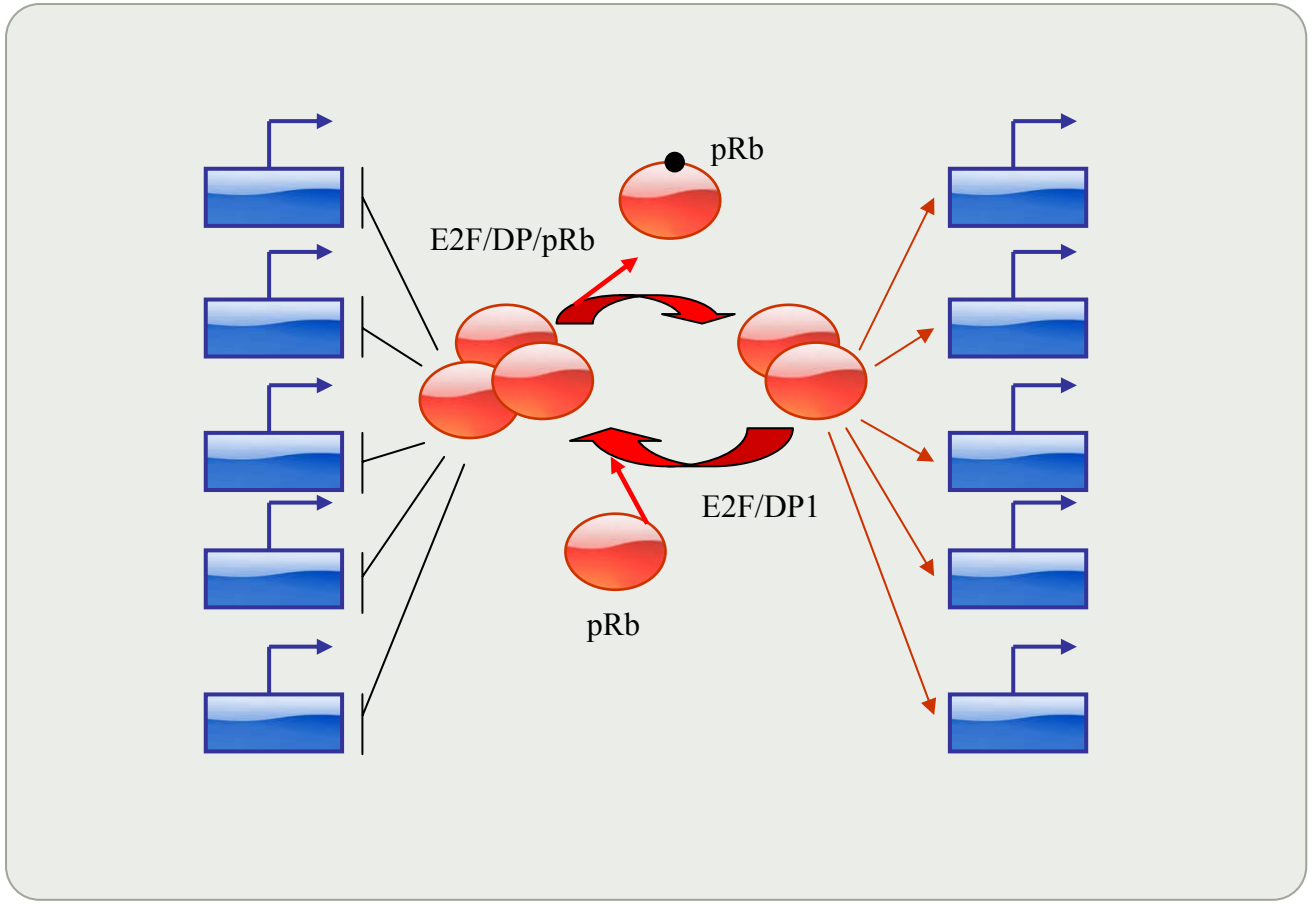
Gene network functional elements: enhancement of the signal of the central regulator according to the positive feedback mechanism



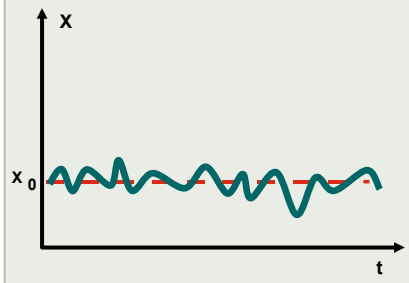
Gene network functional elements: interaction of contours with positive and negative feedbacks



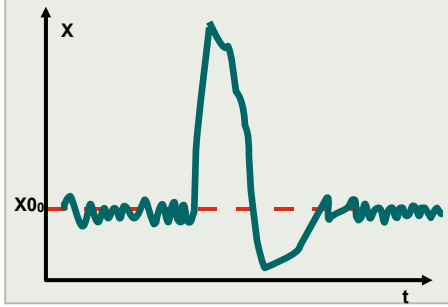
Gene network functional elements: interaction of contours with positive and negative feedbacks



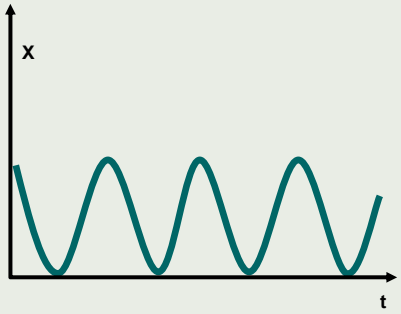
Types of dynamics of processes controlled by gene networks



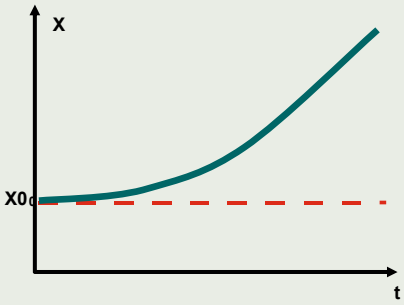
Gene networks of homeostasis (constancy of a parameter under control)



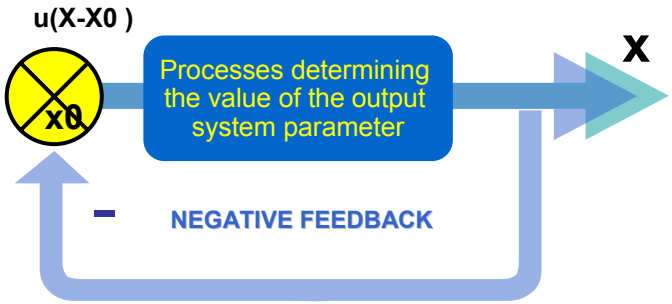
Gene networks of stress response (a pronounced deviation of a parameter followed by its restoration to the initial state)

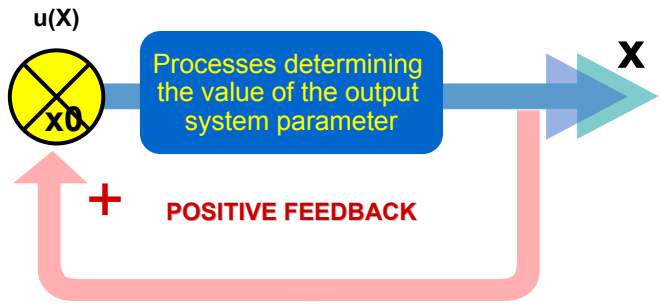


Gene networks of cyclic processes (oscillation of a parameter)



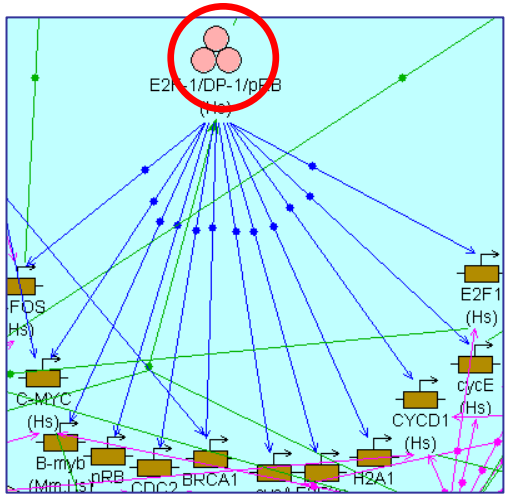
Gene networks controlling cell differentiation and morphogenesis of tissues and organs (a monotonic drift of a parameter from the current state)





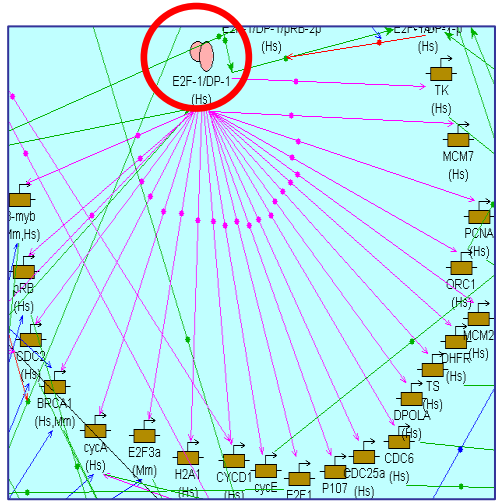
Gene network functional elements:

cassette activation and repression of a large groups of genes

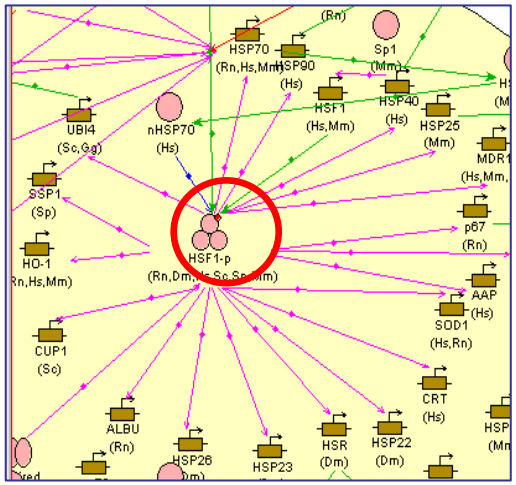


Gene network of cell cycle regulation: repression of genes by factors

E2F1/DP1/pRB

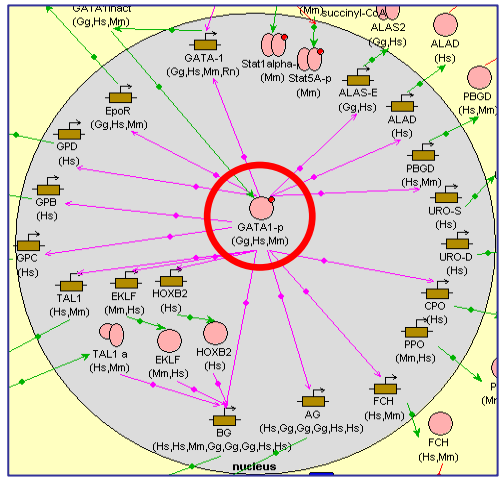


Gene network of cell cycle regulation: activation of genes by factors E2F1/DP1



HSF1

Gene network of heat shock response regulation

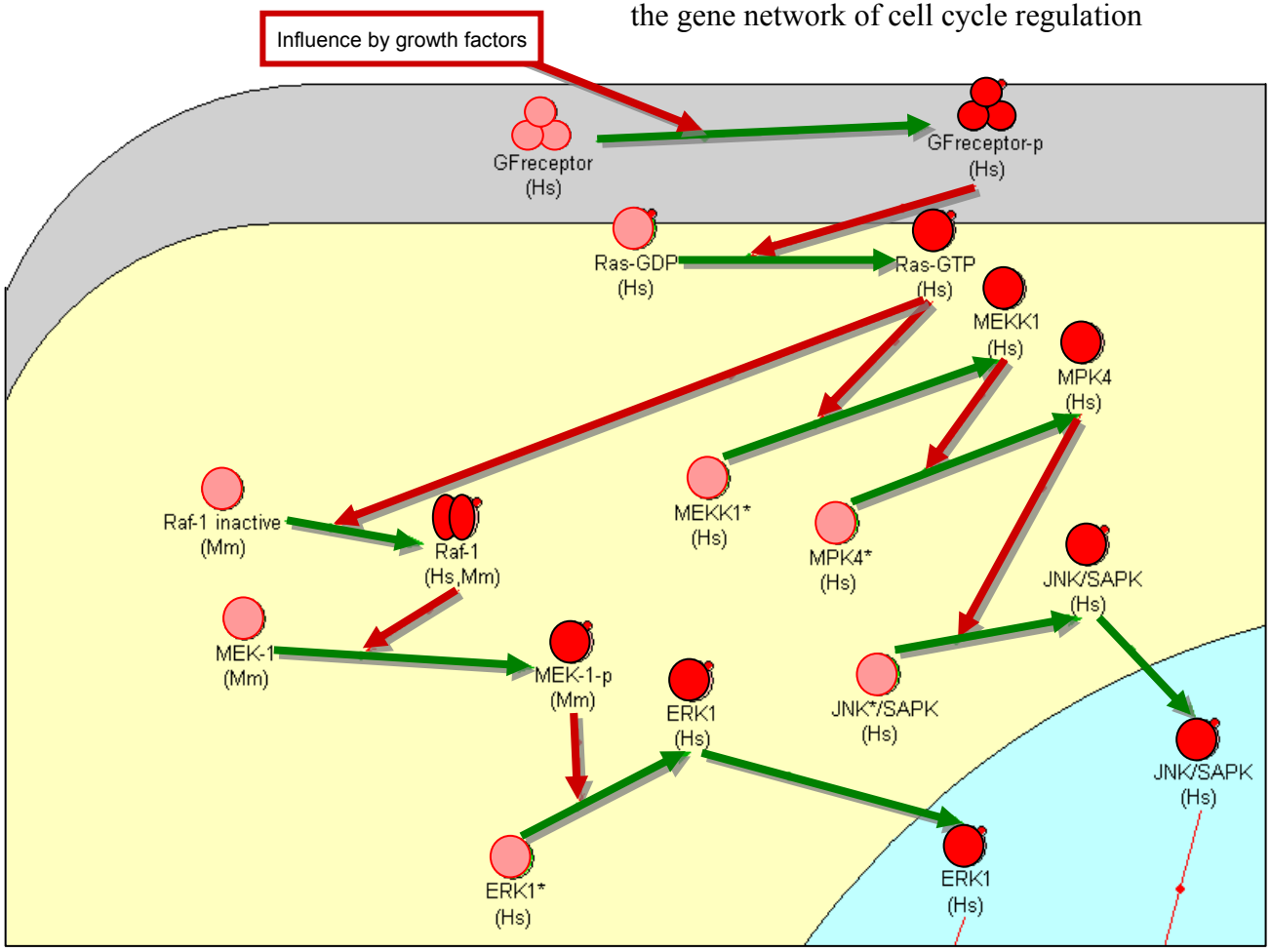


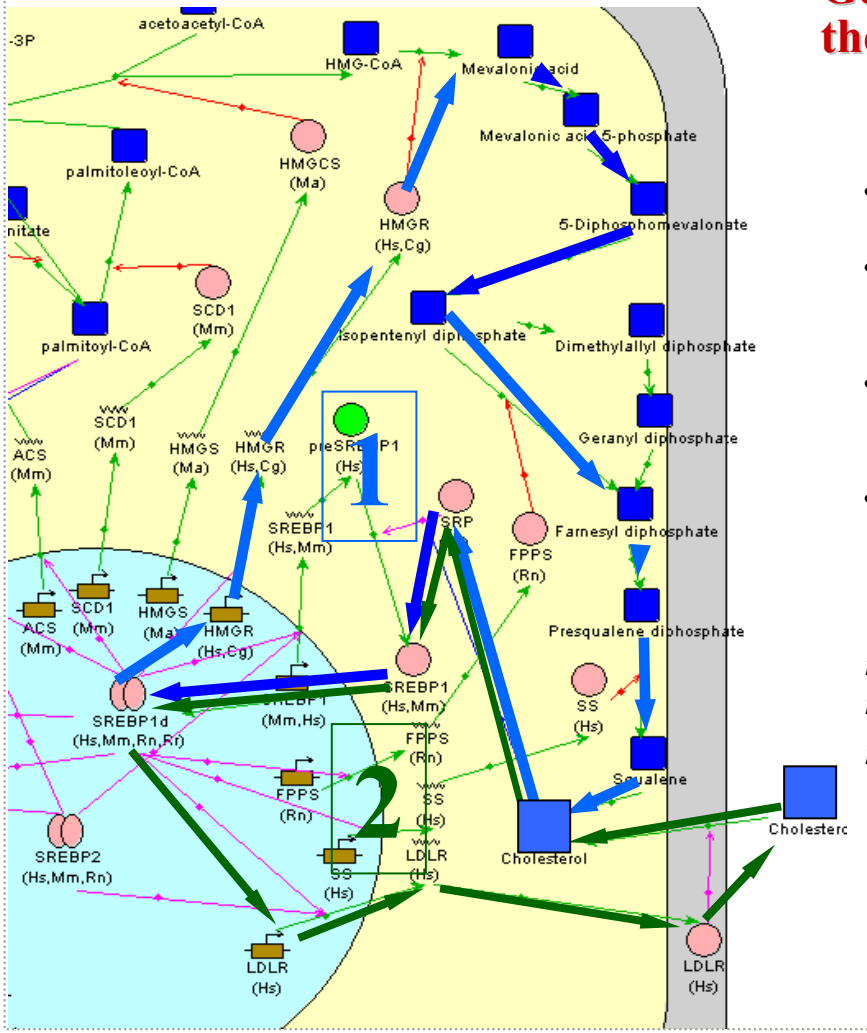
GATA1

Gene network of erythrocyte differentiatio n regulation

Gene network functional elements:

MAP-kinase pathway transforming the signal into the gene network of cell cycle regulation





Gene networks analysis using graph-theoretical approach:

- Calculation of gene network microstructural parameters
- Search for critical gene network elements (graph cutpoints)
- Search for a strongly connected subnetworks in the gene network graph
- Search for regulatory circuits of gene networks and the points of their intersections

Two closed regulatory circuits - negative feedbacks regulating SREBP2 gene expression level have been revealed in the cholesterol biosynthesis gene network

1.2. Computer analysis and modeling (continued)

1.2.4. Modeling of gene networks by using generalized chemical kinetics approach (GCKA)

Specific features of gene networks as object of modeling

These systems may comprise hundreds and thousands of components organized and operating in a complex, hierarchical manner

- **Biochemical level** *(biochemical processes and reactions)*
- **Genetic level** *(arrangement of genes and regulatory elements and their orientations, regulation of gene expression, template oriented processes, polyallelism, etc.)*
- **Subcellular and cellular levels** *(compartmentalization, processes of intercompartmental exchange of energy and substances, active and passive transport etc.)*
- **Organ, tissue, and overall body levels** *(supercompartments, interaction with an external environment etc.)*
- **Population level** *(interaction between individuals and with an environment, age structure of a population, evolutionary processes etc.)*

In the majority of cases, such objects display a complex nonlinear behavior due to negative and positive feedbacks

Methods used for simulating gene network dynamics

- Discrete methods
- Continuous methods
- Hybrid methods

- Chemical kinetic approach
- Stochastic approach
- Logical simulation

- Chemical kinetic approach
- Modeling of genes expression regulation in terms of generalized Hill functions
- Petri nets
- Boolean nets
- Threshold models
- Stochastic simulation
- Etc.

At the Institute of Cytology and Genetics, a generalized chemical kinetic simulation (GCKS) approach is being developed oriented at a formalized, first and foremost, portrayed, description of operation of arbitrary biological systems

Basic principles of gene network simulation

Formalization in the context of GCKS utilizes a block principle

- The elementary subsystems are described in terms of elementary processes;
- The major principle is local independence of the processes;
- Elementary processes are described using a set of formal blocks; and
- These blocks are unambiguously characterized with (1) ordered list of formal dynamic

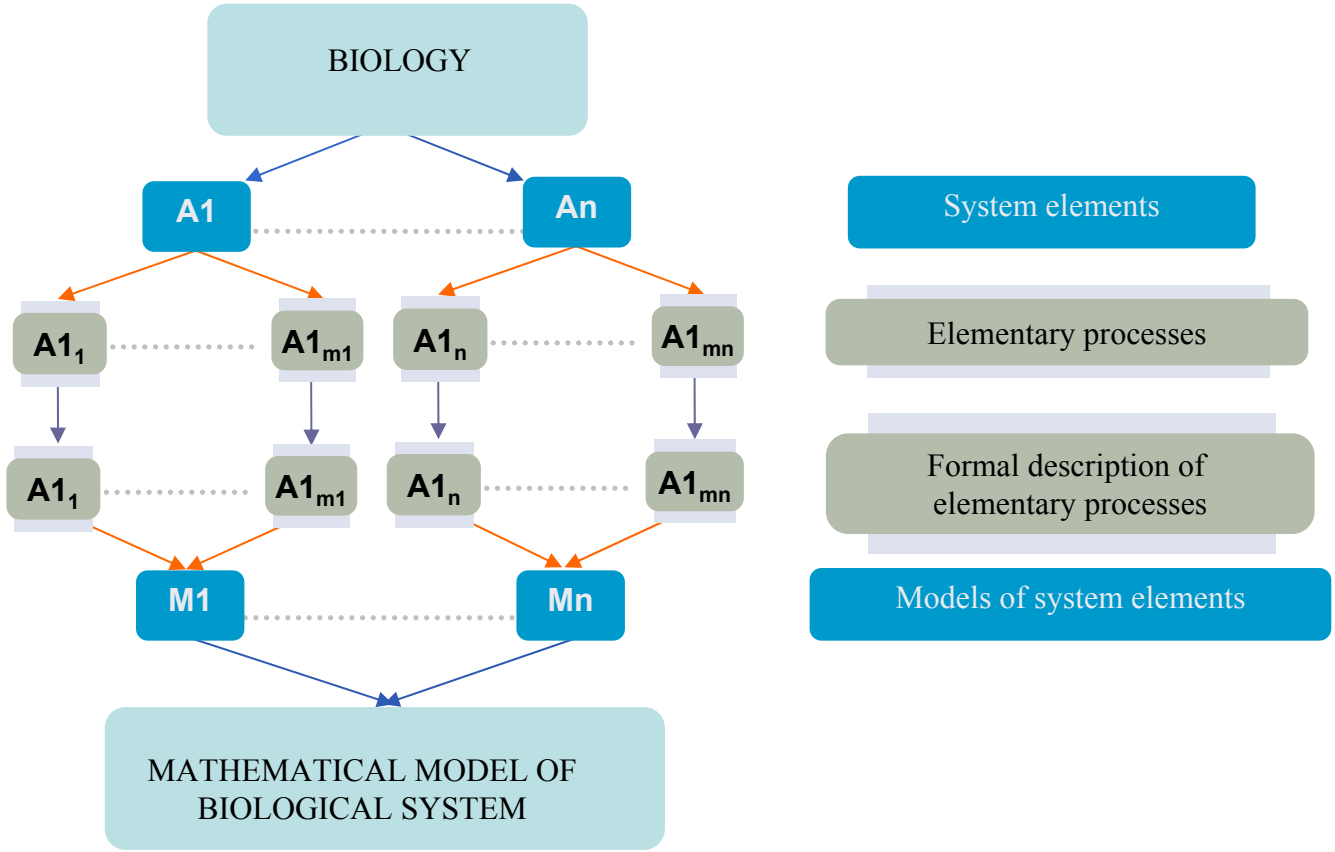
variables (X), (2) ordered list of formal parameters (P), and (3) the rule for transforming information (F)

Generalized Chemical Kinetics Approach functionality

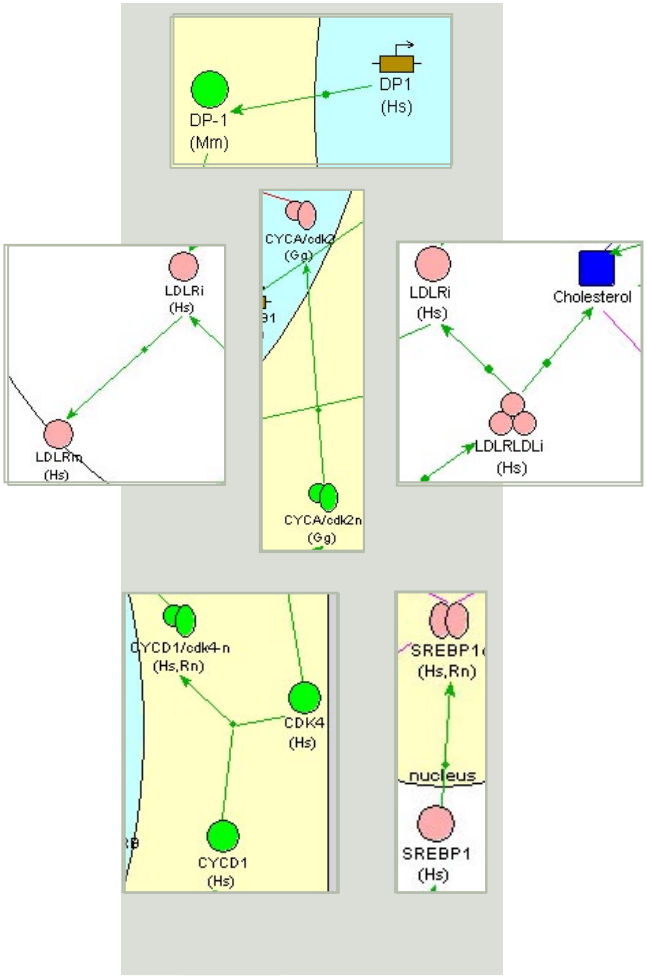
In the context of GCKS, a set of tools are either already available or being developed that allow modeling of the following specific structure-function features of gene networks:

- *Cis*- and *trans*-effects;
- Mutual arrangement and orientations of genes and their regulatory sites;
- A template nature of basic processes (replication, transcription, and translation);
- Polyploidy;
- Genetic rearrangements, recombination, and crossover;
- Mutations;
- Multiple compartmentalization;
- Multivariance;
- Etc.

Modeling methodology in the context of GCKA



Examples of formal descriptions of elementary processes (1)



Constitutive synthesis :

$$\xrightarrow{k} x$$

$$\overline{X} = (x),\ \overline{P} = k, F: \frac{dx}{dt} = k$$

Monomolecular irreversible reaction:

$$x \xrightarrow{k} y_1 + y_2 + \ldots + y_n$$

$$\overline{X} = (y_1, y_2, \ldots, y_n),\ \overline{P} = k,$$

$$F: \frac{dx}{dt} = -\frac{dy_1}{dt} = -\frac{dy_2}{dt} = \ldots = -\frac{dy_n}{dt} = -k \cdot x,\ n \geq 0$$

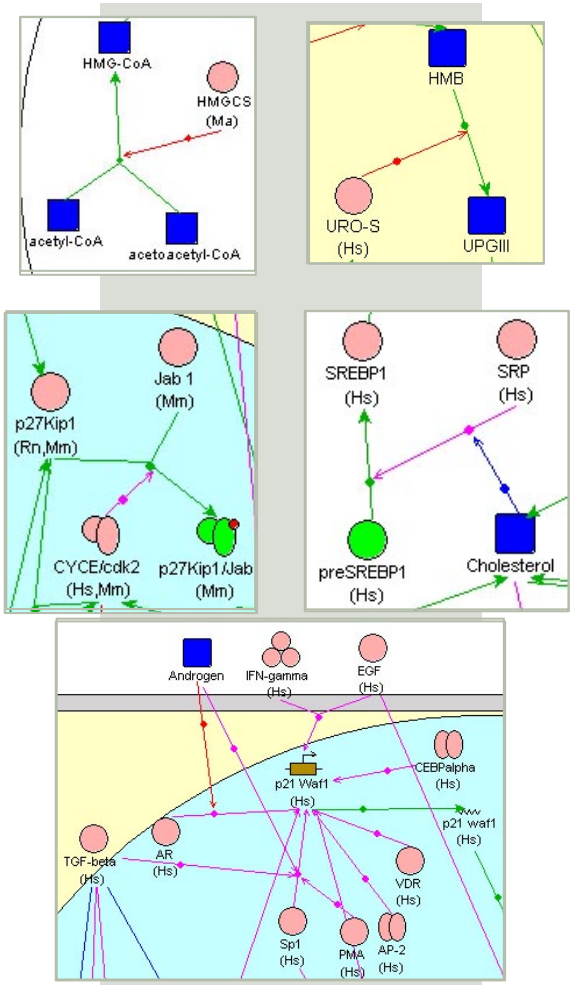
Bimolecular reversible reaction:

$$x_1 + x_2 \underset{k_2}{\overset{k_1}{\Longleftrightarrow}} x_3$$

$$\overline{X} = (x_1, x_2, x_3),\quad \overline{P} = (k_1, k_2),$$

$$F: \frac{dx_1}{dt} = \frac{dx_2}{dt} = -\frac{dx_3}{dt} = k_2 \cdot x_3 - k_1 \cdot x_1 \cdot x_2$$

Examples of formal descriptions of elementary processes (2)



Generalized enzymatic reactions:

$$\overline{X} = (x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n),$$

$$\overline{P} = (m, k_i, k_d, a_1, \dots, a_m, b_1, \dots, b_n),$$

$$F: \frac{dx_j}{dt} = -a_j \cdot k_i \cdot Z,\ j = 1, \dots, m,\ m \geq 2,$$

$$\frac{dy_l}{dt} = b_l \cdot k_i \cdot Z,\ l = 1, \dots, n,\ n \geq 0$$

$$Z = \frac{k_d \cdot x_1 \cdot \dots \cdot x_m}{(k_d + x_1) \cdot \dots \cdot (k_d + x_m) - x_1 \cdot \dots \cdot x_m}$$

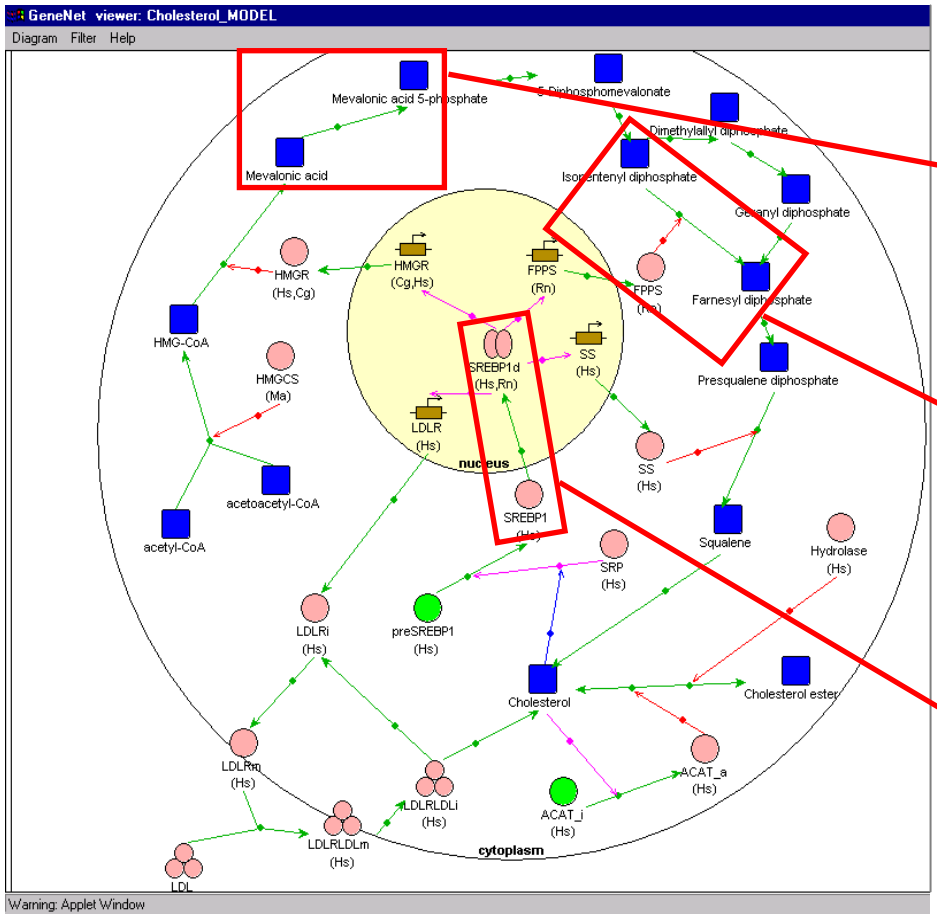
Regulatory processes:

$$-\frac{dX}{dt} = \frac{dY}{dt} = k \cdot \frac{k_1 + \sum_{i=1}^{n_1} \left(\frac{A_i}{\alpha_{1i}} \right)^{h_{1i}}}{k_2 + \sum_{i=1}^{m_1} \left(\frac{I_i}{\beta_i} \right)^{\gamma_i} + \sum_{i=1}^{n_1} \left(\frac{A_i}{\alpha_{2i}} \right)^{h_{2i}}} \cdot$$

$$\cdot \prod_{j=1}^{n_2} \frac{\alpha_{0j} + \left(\frac{A_j}{\alpha_{1j}} \right)^{h_{1j}}}{1 + \left(\frac{A_j}{\alpha_{2j}} \right)^{h_{2j}}} \cdot \prod_{j=1}^{m_2} \frac{1}{1 + \left(\frac{I_j}{\beta_j} \right)^{\gamma_j}} \cdot X,$$

Examples of formal description of elementary processes (3):

gene network regulating cholesterol biosynthesis in the cell



Monomolecular reaction:

$$A \xrightarrow{k} B, \quad a = [A], b = [B]$$

$$-da/db = db/dt = ka$$

Enzymatic reaction:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P,$$

$$e=[E], s = [S], e_a=[ES], p = [P]$$

$$ds/dt = -dp/dt = -k_3 e_o s/(K_m + s),$$

$$e_o = e + e_a, \quad K_m = (k_{-1} + k_2)/k_1$$

Bimolecular reaction:

$$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C, \; a = [A], b = [B], c=[C]$$

$$da/dt = db/dt = -dc/dt = -k_1 ab + k_2 c$$

http://wwwmgs.bionet.nsc.ru/mgs/gnw/gn_model/

A fragment of the system of differential equations

mathematical model of cholesterol biosynthesis in the cell

……

d(Cholin)/dt =
+ {*KISBin**Squalene} – {*KISBestr**Cholin*ACAT/(*KMSBestr* + Cholin + ACAT)} + {*KISBhydr**Cholestr*Hydrolas/(*KMSBhydr* + Cholester + Hydrolas)} + { 1785**KISBfree** LDLRLDLi} + {-*Kin1prot**Cholin*SRPprot + *Kin2prot**SRP-Chol} + {*Ksrpcdeg**SRP-Chol} – {*Kutichol* * Cholin}

d(LDLR____)/dt =
+ { -*Kclaw1**LDLR____*LDL____ + *Kclaw2**LDLRLDLo}+ {*Kfre____t** LDLRin} – {*KLDLRdeg* * LDLR____}

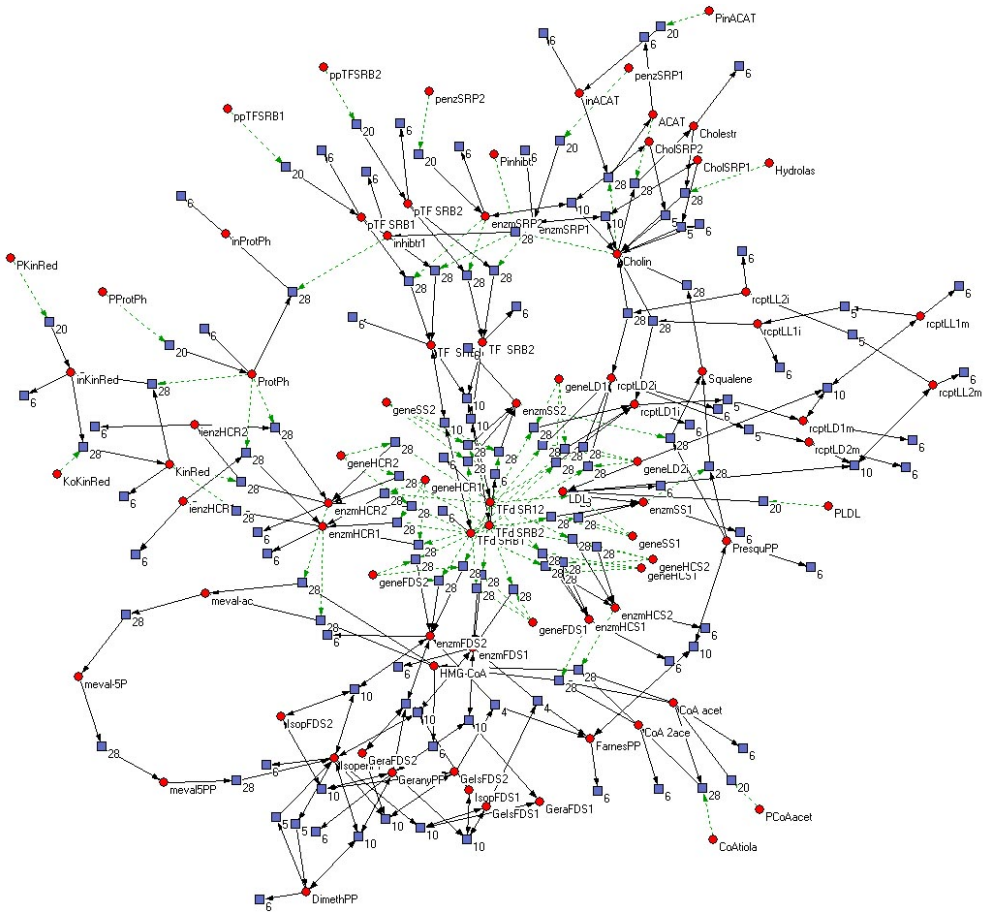
d(LDL ____)/dt =
+ { -*Kclaw1**LDLR____*LDL ____ + *Kclaw2**LDLRLDLo} + { *KsynLDL* * *PLDL* ____ } – {*KutiLDL** LDL ____}

d(LDLRin)/dt =
+ {*KISBldlr***LDLRgene**SREBPdim/(*KMSBldlr* +*LDLRgene* + SREBPdim)}
– {*Kfre____t**LDLRin} + {*Kfree** LDLRLDLi} – {*KdegLRin**LDLRin}

……

Constants of monomolecular reactions, sec^{-1}	*KISBfree = 1*10^{-02}, Ksrpcdeg = 2.5*10^{-05}, Kfre____ = 1*10^{-02}, KLDLRdeg = 1.93*10^{-6}, KsynLDL = 3.32*10^{-02}, KutiLDL = 3.21*10^{-06}, KdegLRin = 1.93*10^{-06}*	
Constants of bimolecular reactions	$l \times mmol^{-1} \times sec^{-1}$	sec^{-1}
	*Kclaw1 = 3.3*10^{-01}*	*Kclaw2 =1*10^{-03}*
Constants of enzymatic reactions	Constants of enzyme turnover, sec^{-1}	Michaelis constants, mmol/l
	*KISBestr = 1*10^{+02},* *KISBhydr = 1.9*10^{+03},* *KISBldlr = 1*10^{-1}*	*KMSBestr = 6.67*10^{-03},* *KMSBhydr = 7.6*10^{-01},* *KMSBldlr = 3.3*10^{-05}*

Bipartite graph of the computer model of diploid gene network regulating intracellular cholesterol homeostasis



- 57 dynamic variables
- 139 processes

Involvement in a process with a nonzero stoichiometry

Involvement in a process with a zero stoichiometry

1 Constitutive synthesis

2 Utilization or degradation

3 Monomolecular reaction

4 Bimolecular reaction

5 Generalized Michaelis-Menten reaction

1.2. Computer analysis and modeling (continued)

1.2.5. Inverse task solution: verification of gene network model parameters

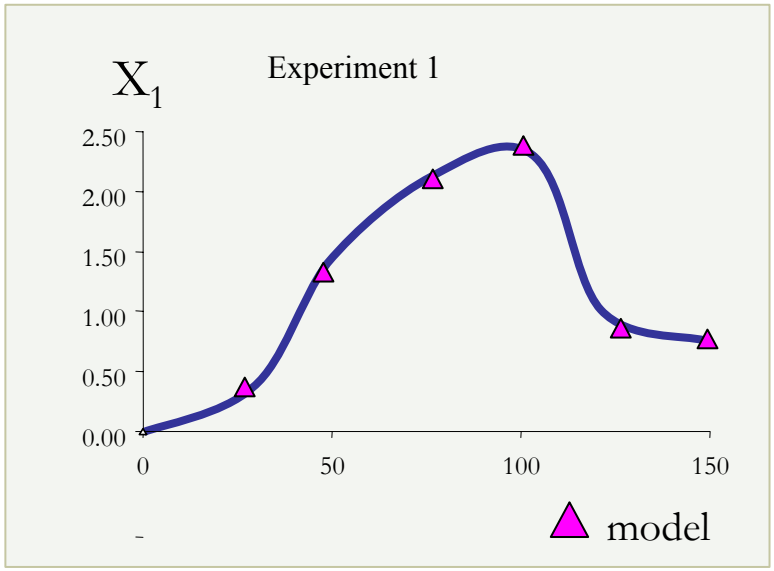
Inverse task solution: key problems

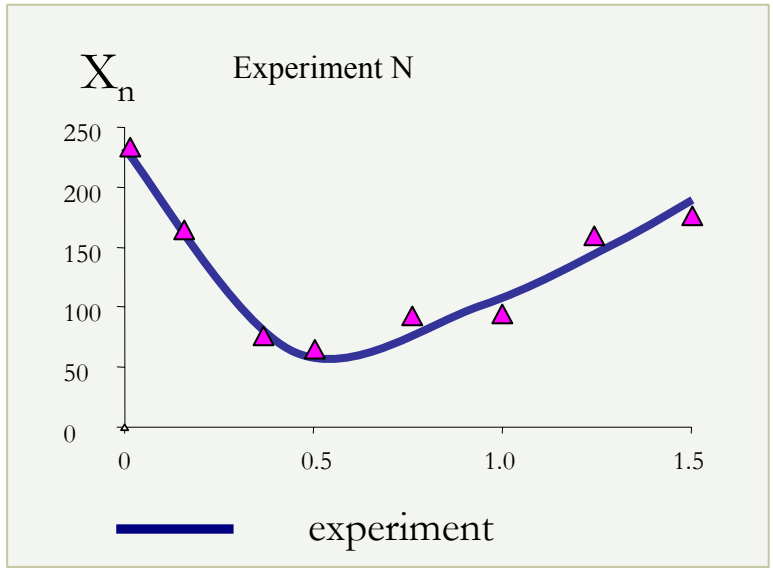
- **Gaps in the knowledge on the structure–function organization of certain gene network fragments (***mediators, mechanisms of processes, etc.***);**
- **Insufficiency of the quantitative data suitable for adapting a particular model;**
- **Descriptive qualitative information is typically compiled in databases** (*information on the structure–function organization of gene network; mechanisms of various processes, organization of gene regulatory regions, etc.*)**, as well as static quantitative information** (*constants of enzymatic reactions, molecular weights of proteins , length of nucleotide sequences, etc.);*
- **Distribution of the kinetic data in multitudes of scientific publications (***semistructured data*)**;**
- **Heterogeneity of the data** (*sets of experimental dynamical data are obtained in various type experiments, under various experimental effects, at different time points, etc.*)

Verification of gene network model parameters

- **Constants** of many **enzymatic reactions** are available in databases, such as
  WIT - *http://www-unix.mcs.anl.gov/compbio/*
  BRENDA - *http://brenda.bc.uni-koeln.de/*
  etc.
- **Constants** of **macroprocesses** *(replication, transcription, translation, etc.) Vtranslation~0.1 sec-1*
- **Constants** characterizing **gene network interaction** with its cellular and organismal **environments**:
  - equilibrium concentrations равновесные концентрации;
  - lifespans or half-lives of the system's components;
  - integral characteristics интегральные характеристики;
  - etc.

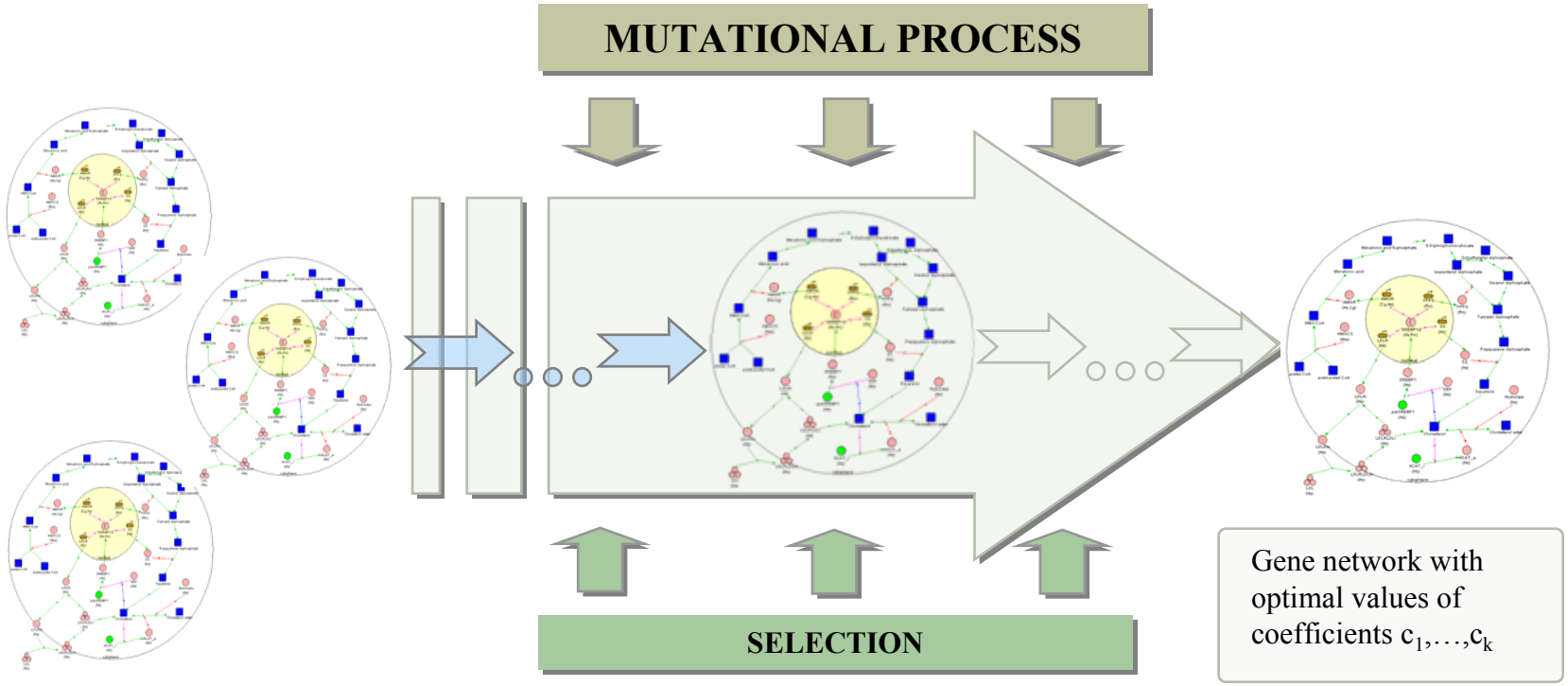
Verification of gene network model parameters





We are searching for such values of c_1, …,c_k constants, which provide maximal correspondence between calculated and observed dynamic behavior of the gene network for a large number of experiments

Solutions of the Inverse Task for a Gene Network: genetic algorithm



Gene network with optimal values of coefficients $c_1,\ldots,c_k$

Initial population of gene networks: not all the constants $c_1,\ldots, c_k$ are known

Optimized functional: W=1/F, where
W – is an adaptability of an organism in particular environmental conditions;
F – is the measure of deviation of calculated characteristics from the ordered ones

$$F(k_1,\ldots k_m)=\sqrt{\sum_{i,j(i)}(X_{ij}^{calc.}-X_{ij}^{exp.})^2} \text{ or } F(k_1,\ldots k_m)=\sum_{i,j(i)}\left[\frac{X_{exp.}^{ij}}{X_{calc.(k_1\ldots k_m)}^{i}}+\frac{X_{calc.(k_1\ldots k_m)}^{ij}}{X_{exp.}^{ij}}-2\right]$$

The inverse task for a gene network: experimental data on gene networks dynamics in the GeneNet database

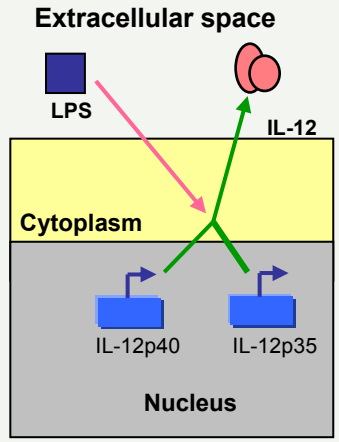
General information

Type	Protein
Name_brief	IL-12
Name_full	Interleukin-12
Organism	Mouse – mus musculus
Cells	Peritoneal macrophages
StageCellDifferentiation	Terminally differentiated
OrganismStatus	Norm
ExpressionDetectionDevice	Relative protein level
Reference	Nomura F. et al., 2000
Comments	C57BL/6J mice

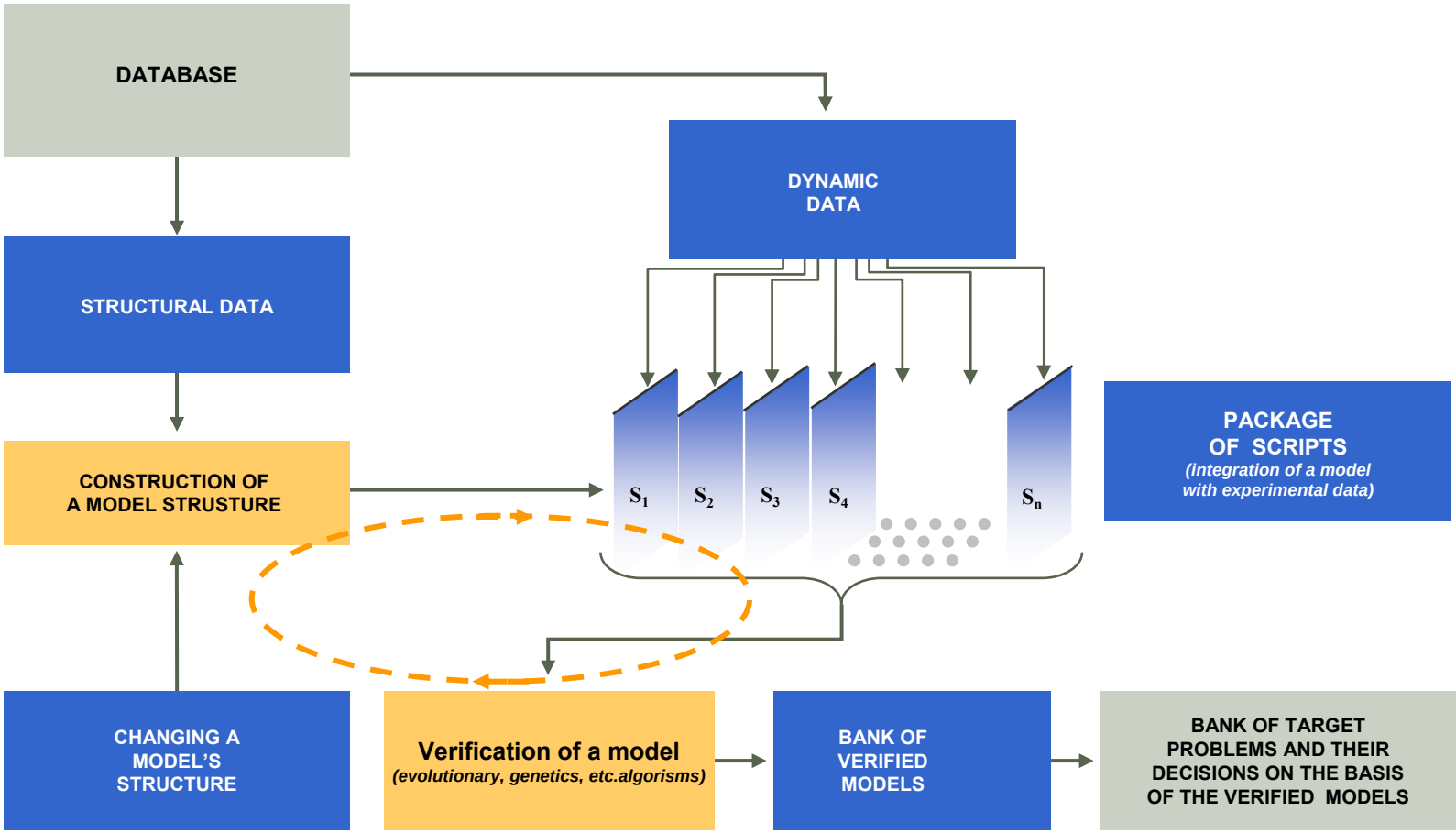
Peritoneal macrophages were preincubated with LPS for the indicated periods, then washed with HBSS twice, and then stimulated with LPS.

Experimental conditions

ID	External Factor	Specification	Time			Concentration	
			(initial point)	(exposure time)	Units	Value	Units
1	LPS	Escherichia coli O55:B5	0	6	hours	10	ng/ml
2	LPS	Escherichia coli O55:B5	0	1	hours	100	ng/ml
2	LPS	Escherichia coli O55:B5	1	7	hours	10	ng/ml
3	…	…	…	…	…	…	…

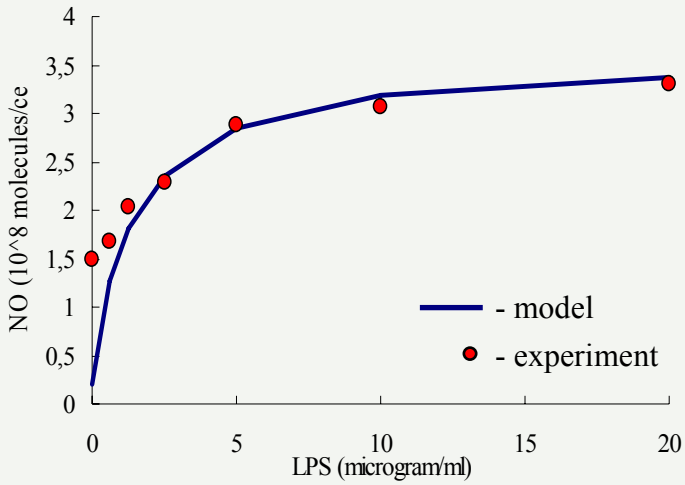


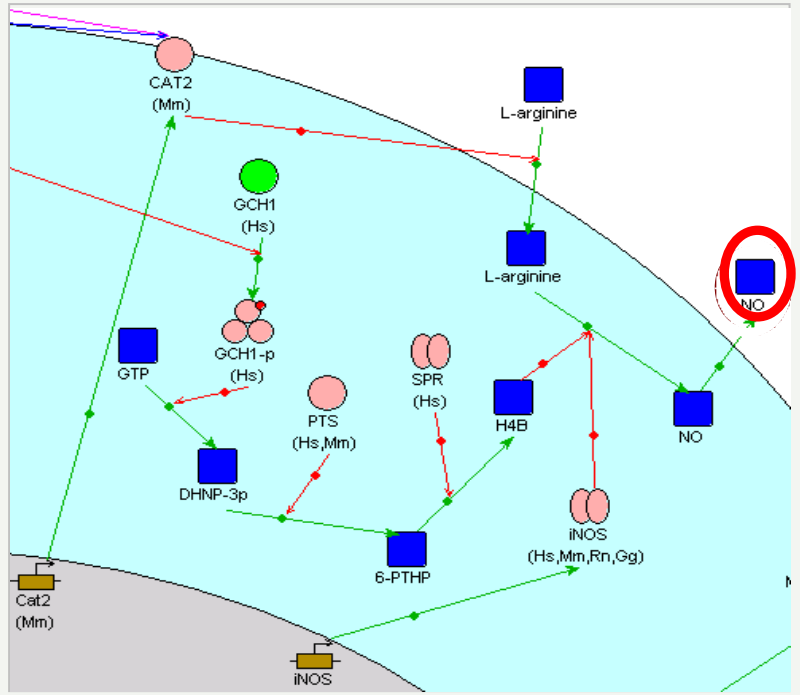
The principle scheme of the script construction for inverse task solution



Emergence of new data incompliant with the model!

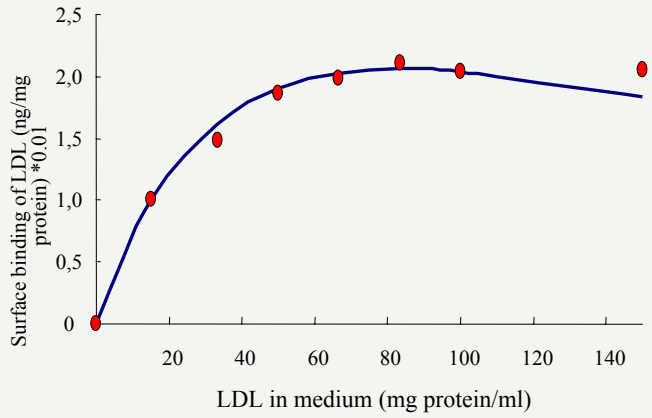
Macrophage activation gene network: inverse task solution





Dependence of the amount of NO evolved on LPS concentration

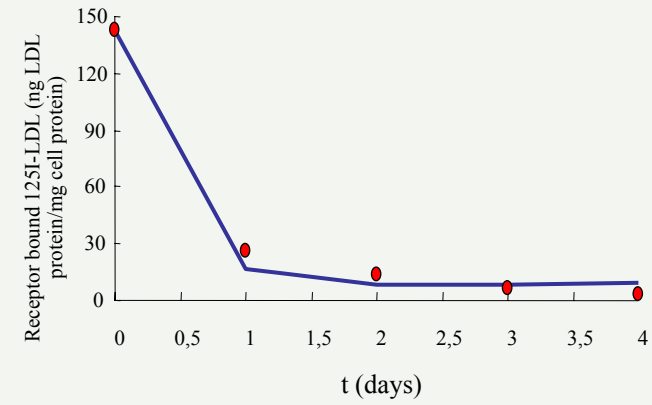
Gene network regulating intracellular cholesterol homeostasis: inverse task solution



Surface binding of LDL

Experiment conditions: Monolayers of cells were grown in the lipoprotein-deficient serum (LPDS) for 2 days and then were incubated in the medium containing the indicated concentration of ^{125}I-LDL at 37^0 C. After 5 hr, the amount of ^{125}I-LDL bound to the cell was measured.

- *experimental data*
  (Brown M.S. and Goldstein J.L., PNAS 1979, 76(7):3330-7)
- *calculated by the model*

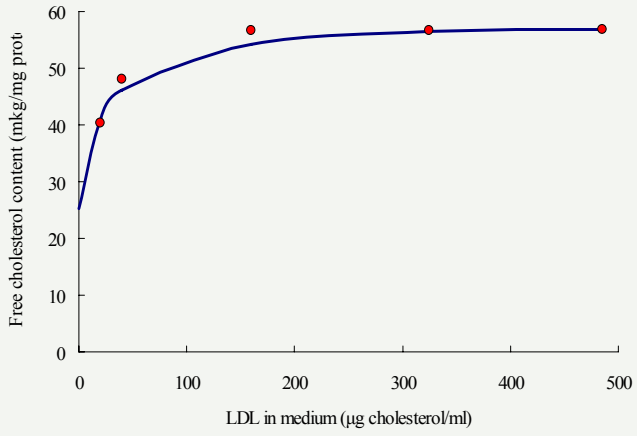


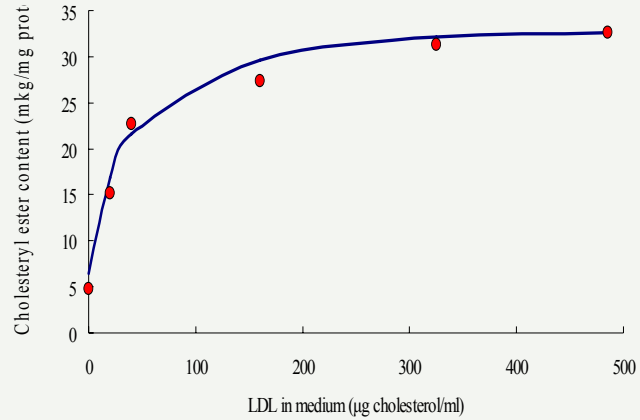
Suppression of LDL receptor synthesis

Experiment conditions: Monolayers of cells were grown in LPDS for 3 days. On day 4 of cell growth (zero time). Each dish of nonconfluent cells received 2 ml of growth medium containing 10 μg protein/ml of unlabeled LDL. At the indicated time, the medium was replaced with 2 ml of medium containing 25 μg protein/ml of ^{125}I-LDL. After incubation at 37^0 C for 2 hr, the specific heparin releasable ^{125}I radioactivity was determined.

- *experimental data*
  (Goldstein J.L.and Brown M.S., Annu.Rev.Biochem., 1977, 46:897-930)
- *calculated by the model*

Gene network regulating intracellular cholesterol homeostasis: inverse task solution





Effect of increasing concentrations of LDL on the content of free and esterified cholesterol in normal fibroblasts
Experiment conditions: The cells were incubated in fresh growth medium containing 10% fetal calf serum. After 3 days, the medium was replaced with 2 ml of fresh growth medium containing 5% human LPDS. After 24 hr, the medium was replace with 2 ml of fresh growth medium containing 5% human LPDS and indicated concentration of LDL. After 24 hr, each cell monolayer was washed and harvested and the sterols content were measured.

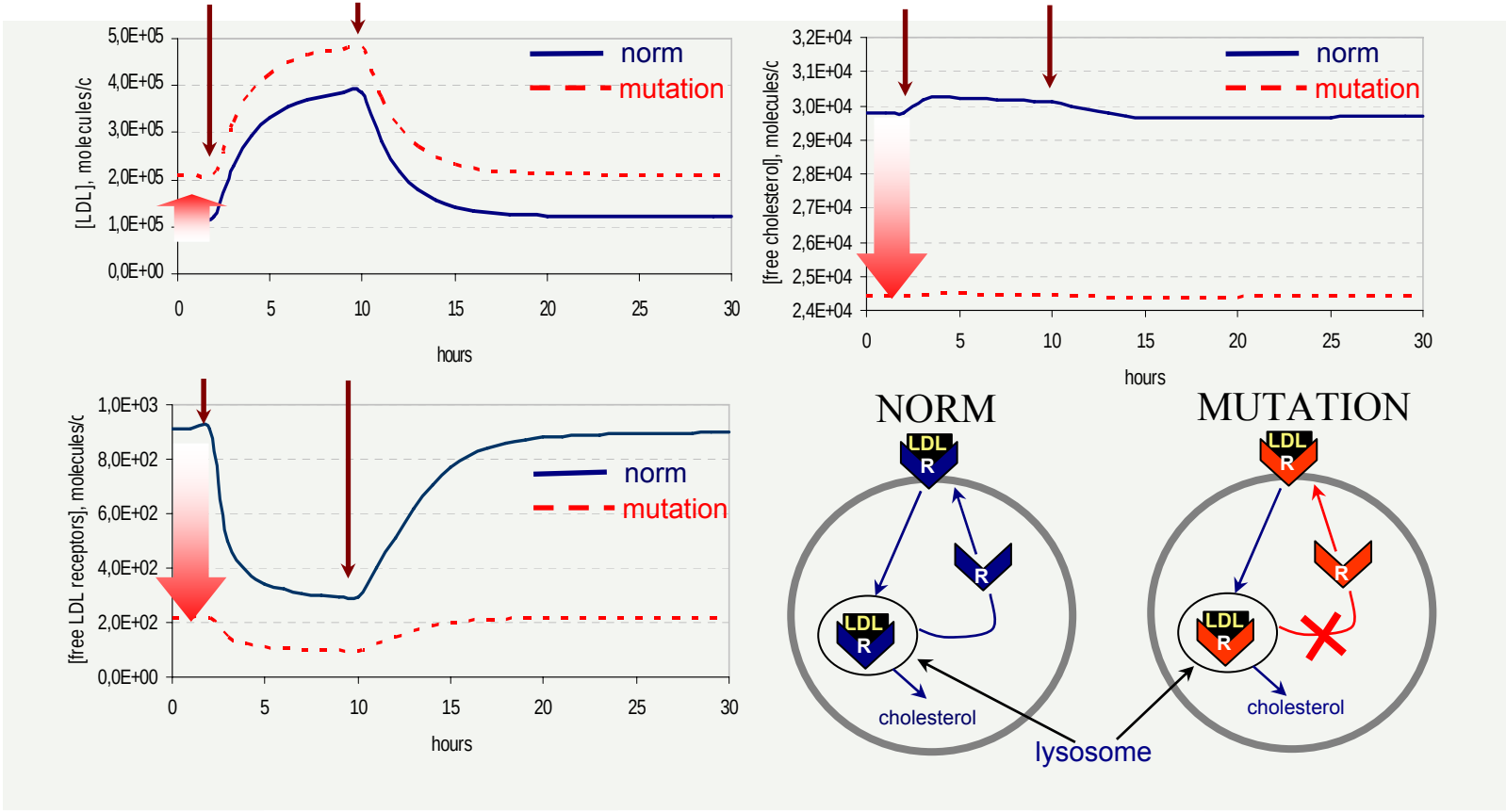
● *experimental data*
(Goldstein J.L. et al., J. Biol. Chem., 1977, 250(21):8487-95)
— *calculated by the model*

1.2. Computer analysis and modeling (continued)

1.2.6. Analysis of mutation effects on gene networks functioning

Analysis of mutation effects on the gene network regulating intracellular cholesterol homeostasis

Simulating the response of the gene network controlling cholesterol biosynthesis in the cell to a twofold increase in LDL inflow into blood plasma during 8 hours in the presence of a mutation decreasing the ability of receptors to release LDL in endosomes, resulting in a tenfold-increased receptor degradation



Analysis of the effects of mutations on the gene network regulating intracellular cholesterol homeostasis

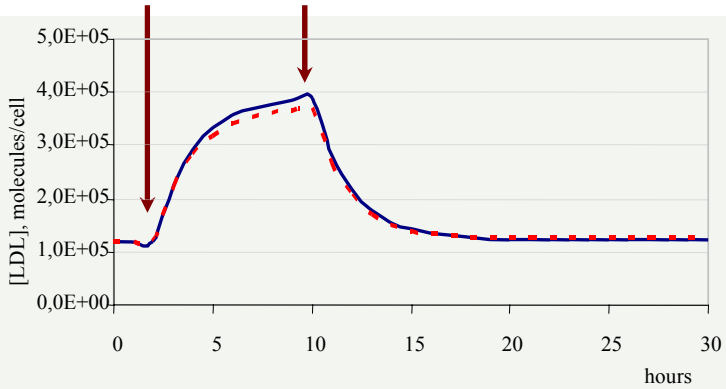
The model allows stationary characteristics and dynamics of the gene network, both in norm and in the presence of mutations, upon various effects to be studied.

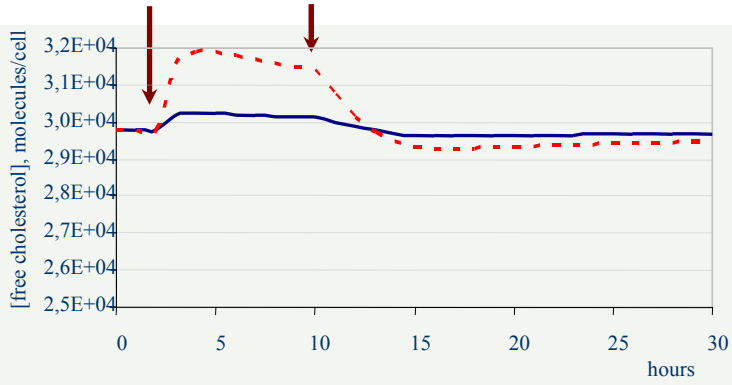
Bold lines in Figure demonstrate the calculated response of the studied gene network in norm to a twofold increase in the inflow of LDL into blood plasma continuing over 8 h. These conditions cause a monotonic increase in blood LDL, reaching an approximately fourfold level by 10 h of the experiment. In this process, the concentration of free receptors on the cell surface decreases, whereas the concentration of cholesterol in the cell changes insufficiently, which is explained by the negative feedback decreasing the rate of cholesterol biosynthesis in the cell upon its increased inflow from outside the cell. All the variables of this system take stationary values approximately 6 h after the internal effect is stopped.

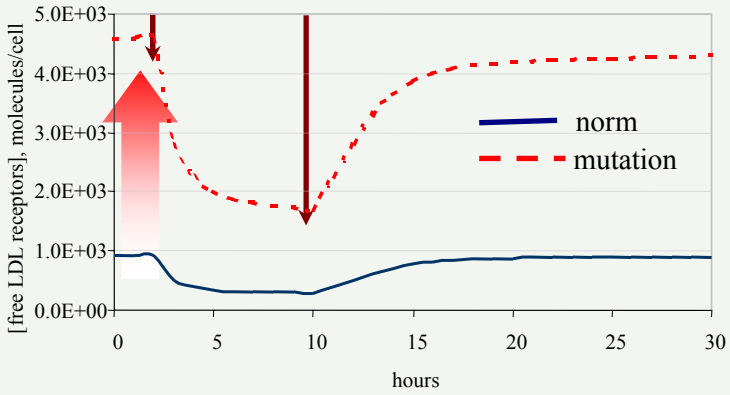
Cleavage of LDL receptor from its ligand in the acid medium of endosomes and its return to the cell surface complete the receptor conversion cycle in the cell. The mutation variant brings about formation of a truncated LDL receptor protein. This truncated receptor loses the ability to release LDL in endosomes, resulting in receptor degradation. The degradation rates of the LDL receptors impaired by mutations of this class may grow 5–10-fold, decreasing considerably the number of receptors on the cell surface (Fourie et al., 1992). The model allowed us to analyze the response to a tenfold increase in the receptor degradation in the cell relative to the normal rate. Figure demonstrates that the stationary LDL concentration in blood increases approximately twofold, accompanied by a drop in the cell sensitivity to changes in the external LDL concentration. The stationary number of free receptors on the cell surface reduces approximately 4.5-fold; the concentration of free cholesterol, by approximately 25%.

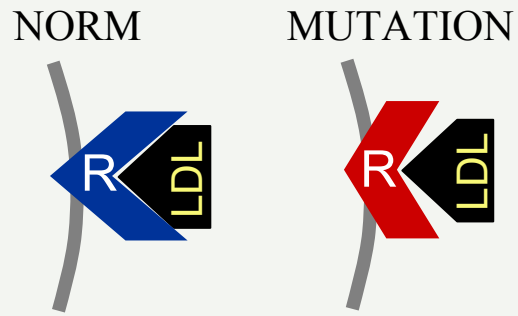
Analysis of the effects of mutations on the gene network regulating intracellular cholesterol homeostasis

Simulating the response of the gene network controlling cholesterol biosynthesis in the cell to a twofold increase in LDL inflow into blood plasma during 8 hours in the presence of mutation decreasing fivefold the receptor ability to bind LDL



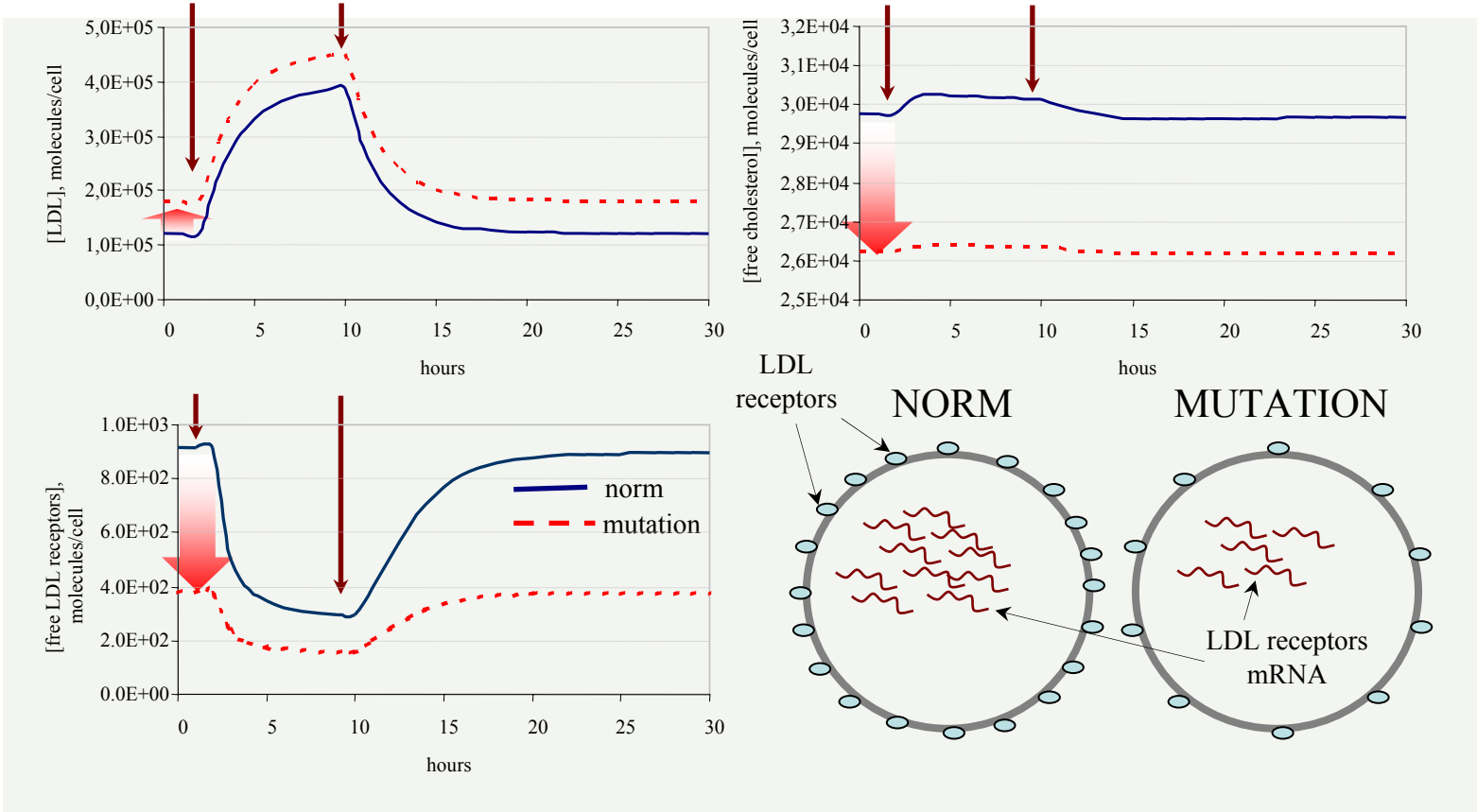






Analysis of the effects of mutations on the gene network regulating cholesterol biosynthesis in the cell

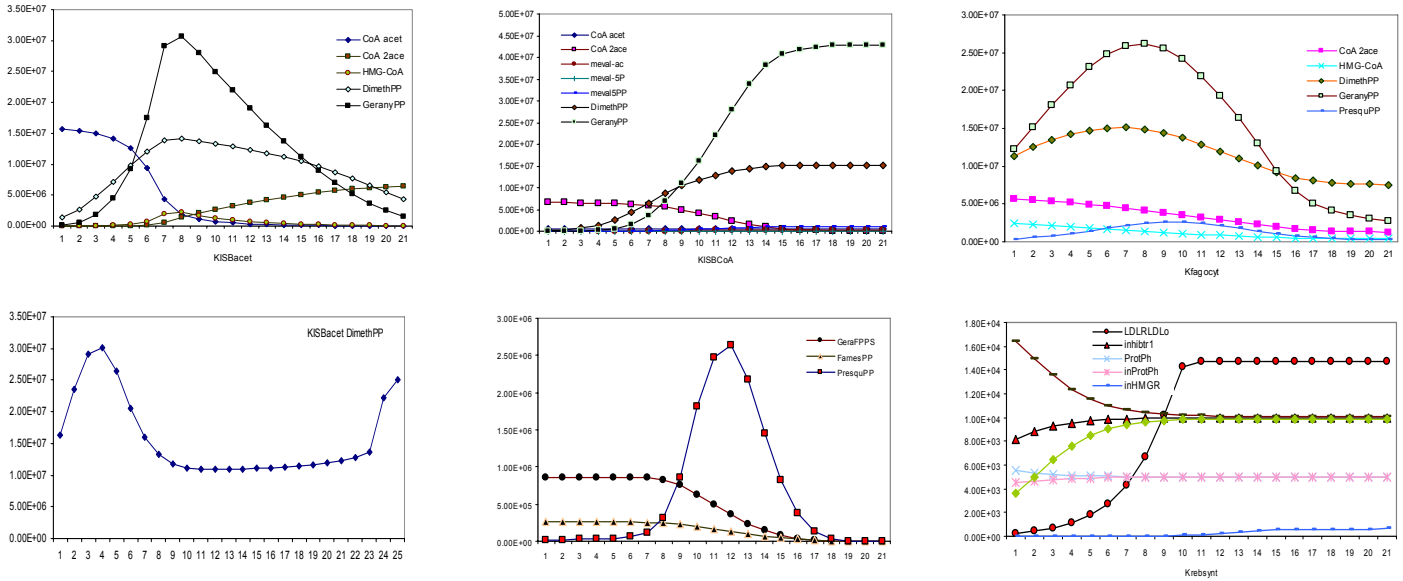
Simulating the response of the gene network controlling cholesterol biosynthesis in the cell to a twofold increase in LDL inflow into blood plasma during 8 hours in the presence of mutation decreasing the LDL gene expression rate twofold



A mutational portrait of the gene network regulating intracellular cholesterol homeostasis

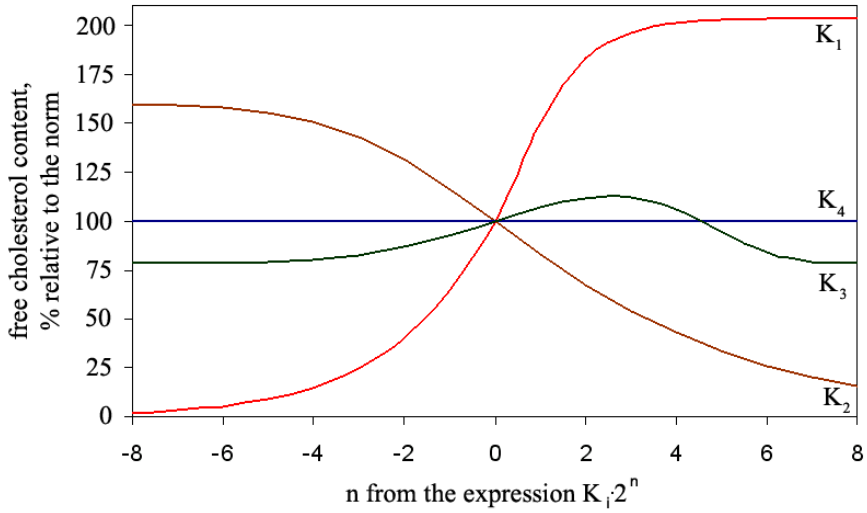
A more detailed study of the effects of mutations on the system to detect the key processes of the gene network and analyze the system's behavior in various pathological situations require investigating of mutations with varying degrees of manifestations in all the gene network components.

"Mutational portrait" of a gene network is a set of its stationary states and dynamic characteristics obtained through mutational variation of each elementary process composing the gene network within the specified rate limits.



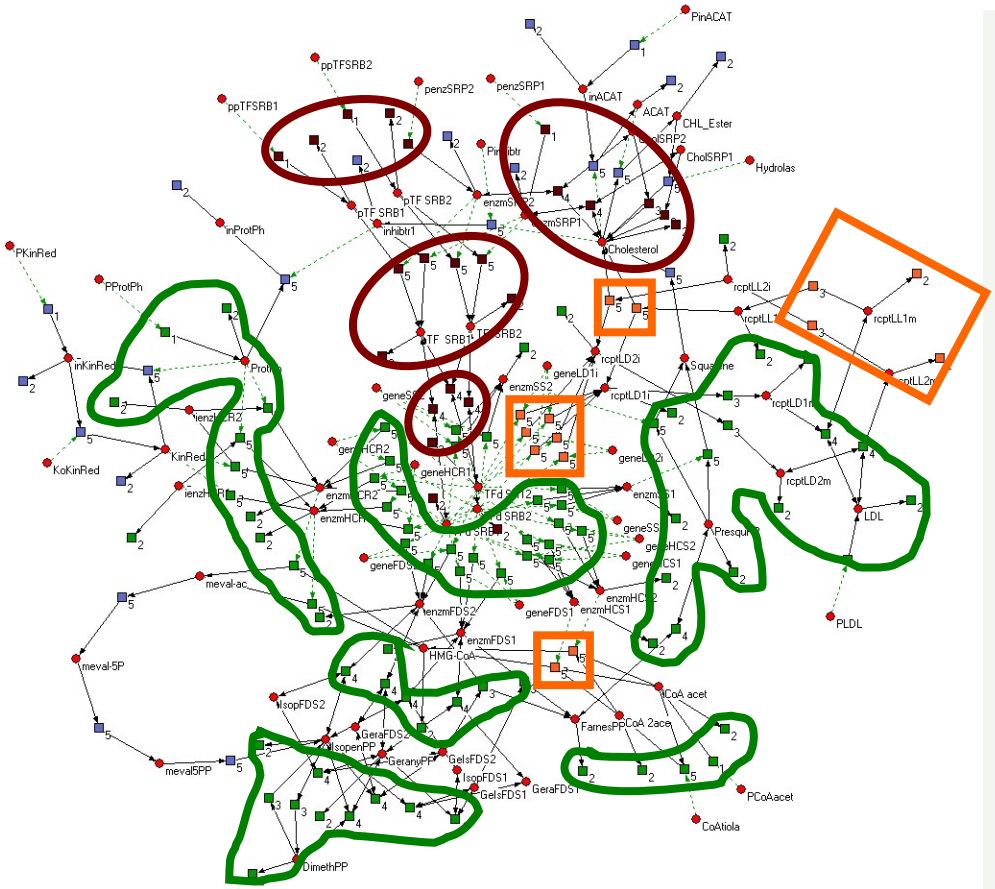
The effects of single mutations of varying intensities of all the model's parameters on the gene network regulating cholesterol biosynthesis in the cell were studied. Overall, ~2000 calculations were performed.

A mutational portrait of the gene network regulating cholesterol biosynthesis in the cell



Changes in free cholesterol stationary content in the cell upon mutational changes in parameters of the mathematical model of the gene network regulating cholesterol biosynthesis: the ordinate represents free cholesterol content in the cell, the abscissa, the number n from expression $Ki \cdot 2^n$, where Ki is the value ith parameter of the system in norm.
K_1 = Exchange constant of the enzyme SRP; K_2 = constant of the reverse reaction of SREBP dimerization; K_3 = Michaelis-Menten constant of the enzyme acetoacetyl-CoA thiolase; K_4 = exchange constant of the enzyme ACAT.

A mutational portrait of the gene network regulating cholesterol biosynthesis in the cell



Gene network with indicated sensitivities of the stationary free cholesterol content in the cell to mutational changes in parameters

- Changes in the rates of these processes affect considerably the stationary cholesterol concentration, changing it from 0 to over 200% of the norm;
- Changes in the stationary concentration of free cholesterol are below 35% of the norm;
- Changes in the stationary cholesterol concentration do not exceed 25% of the norm.

A mutational portrait of the gene network regulating cholesterol biosynthesis in the cell

Characterization of the mutational portrait of the gene network regulating cholesterol biosynthesis in the cell	Mutation type	The fraction of processes changed by mutations of the type indicated
	Impairing	~15%
	Weakly impairing	~45%
	Neutral	~40%

Reasons of low sensitivities of free cholesterol content in the cell to mutational changes in many parameters of the model:

I. Occurrence of nonlimiting stages in biochemical pathways of the gene network in question;
II. Occurrence of two processes responsible for supplementary cholesterol amounts in the cell, namely, (a) synthesis of cholesterol in the cell itself and (b) cholesterol transport from blood plasma into the cell via LDL receptors;
III. Possible shunting of certain biochemical reactions of the cholesterol biosynthesis pathways; and
IV. Regulation of intracellular cholesterol concentration according to a negative feedback mechanism.

Conclusion (1)

- The generalized chemical kinetic simulation approach allows the gene network dynamics to be described and studied at different levels of their organization taking into account their specific features, such as the ability to be structures, hierarchy, etc.
- This method was used to construct dynamic models of gene networks regulating cholesterol biosynthesis in the cell, inducing macrophage activation, and controlling erythrocyte maturation. The models were adapted to experimental data using a new technique of constructing scripts.
- The models were used to solve several theoretical problems. In particular, effects of certain actual and hypothetical mutations were studied as well as the mutational portrait of gene networks. This class of problems is of special interest when detecting optimal targets for pharmacological regulation.
- Analysis of effects of various mutations and detection of key processes in gene networks susceptible to regulation may form the background for correcting pathological states taking into account individual genotype-specific features.

Conclusion (2)

Application of the computer dynamic model of the gene network regulating cholesterol biosynthesis in the cell

Studying and understanding the pathological processes at the cell and organism level (various diseases, syndromes, inborn and acquired errors, mutations, etc);

The model of gene network regulating cholesterol biosynthesis in the cell is applicable to:

- Atherosclerosis;
- Coronary heart disease (CHD);
- Smith-Lemli-Opitz syndrome;
- Mevalonic aciduria;
- Desmosterolosis;
- CHILD syndrome;
- Different kinds of dysplasia;
- etc;

Identification of the genetic and biochemical defects and analysis of their effects on functions of gene networks;
Development of optimal methods for influencing systems for normalizing their functions (e.g. for medical treatment using therapeutics and so on); and Investigating basic biological problems, such as evolution, etc.

1.2. Computer analysis and modeling (continued)

1.2.7. Optimal pharmaceutical control normalizing gene networks functioning

1.2.7.1. In search of optimal control of gene network functions; new-generation pharmacology

1.2.7.2. Optimal control: normalization of functioning of a pathological gene network

1.2.7.3. GeneNet database: a fragment of the gene network controlling NO synthesis under macrophage activation induced by bacterial infection

1.2.7.4. Macrophage activation gene network: modeling of the mutation influence on the dynamics of NO synthesis and optimal pharmaceutical control normalizing the function of the mutant gene network

In search of optimal control of gene network functions; new-generation pharmacology (1)

Only limiting links of a gene network can be targets of optimal pharmacological control. They are few. Development of the mutational profile of a gene network is an obligatory stage of the identification of optimal pharmacological control.

In order to prevent transition to an uncontrollable state, a gene network should be brought back to normal through a sequence of successive stationary states.

In search of optimal control of gene network functions; new-generation pharmacology (2)

Obligatory components:

- Individual genotype-specific choice of drugs for correction of human disease.
- Identification of optimal strategies for correction of individual genetic defects by computer analysis and modeling of the function of impaired genetically controlled systems and processes.

Optimal control: normalization of functioning of a pathological gene network

Optimal pharmaceutical control is the normalization of a gene network functioning by shifting from the steady state of the pathological networks VP to the vicinity of the normal steady state VN

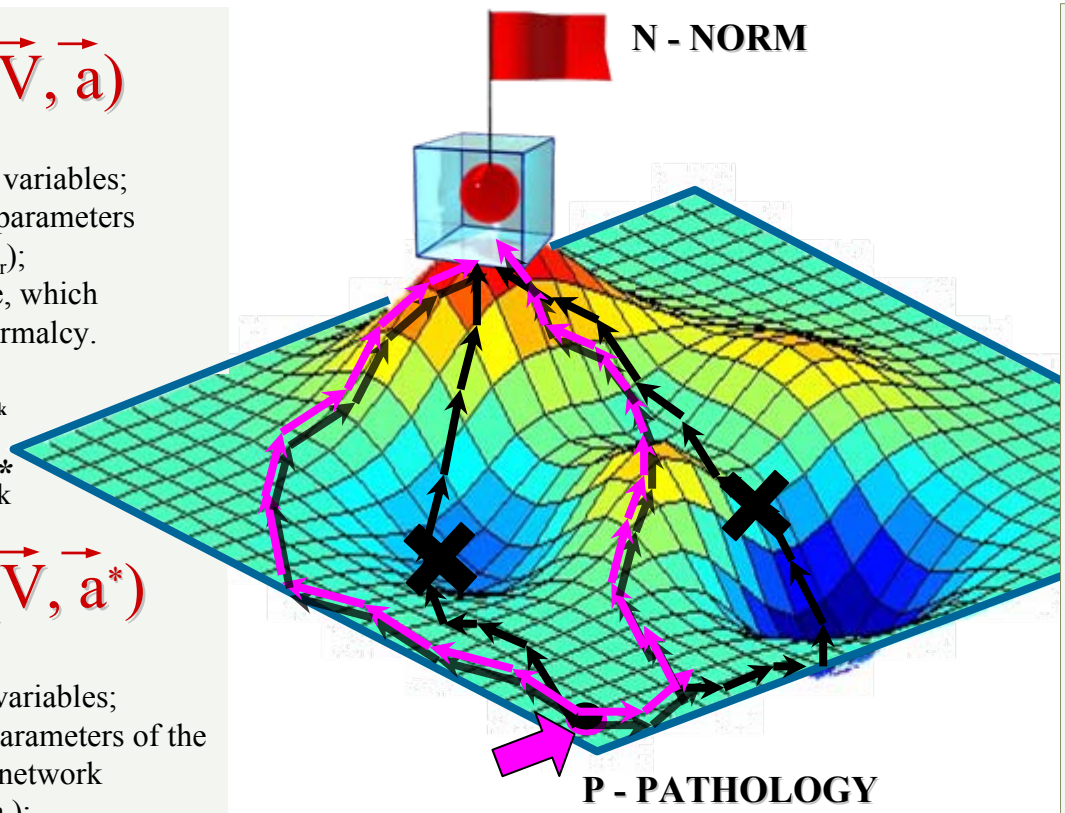
$$\frac{d\vec{V}}{dt} = \varphi(\vec{V}, \vec{a})$$

$\vec{V}$ is the vector of variables;
$\vec{a}$ is the vector of parameters $(a_1,\ldots,a_k,\ldots, a_r)$;
N is a steady state, which corresponds to normalcy.

$a_k \rightarrow a_k^*$

$$\frac{d\vec{V}}{dt} = \varphi(\vec{V}, \vec{a}^*)$$

$\vec{V}$ is the vector of variables;
$\vec{a}$ is the vector of parameters of the pathological gene network $(a_1,\ldots,a^*_k,\ldots, a_r)$;
N is a steady state, which corresponds to pathology.



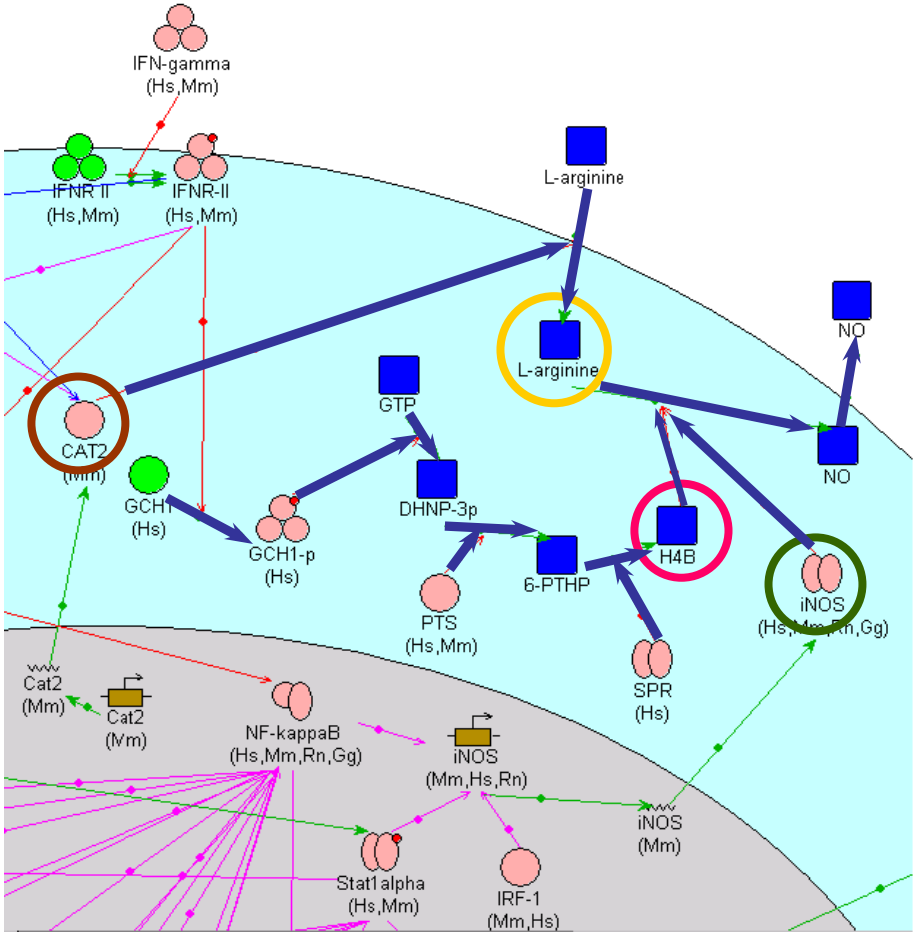
$$\frac{d\vec{V}}{dt} = \varphi(\vec{V}, \vec{a}^*, \vec{U})$$

$\vec{U}$ is the optimal control, which recovers the function of a gene network by shifting the stationary state of the pathological gene network to the vicinity of the normal steady state.

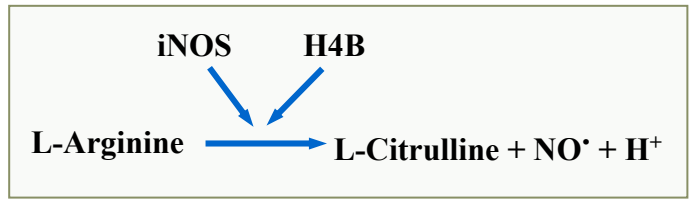
$\vec{U}$ is a class of piecewise linear control functions describing the change of the vector of parameters a in the process of pharmacological control)

Optimal control, which normalizes the critical variable of the gene network, should keep other important variables of this network within normalcy.

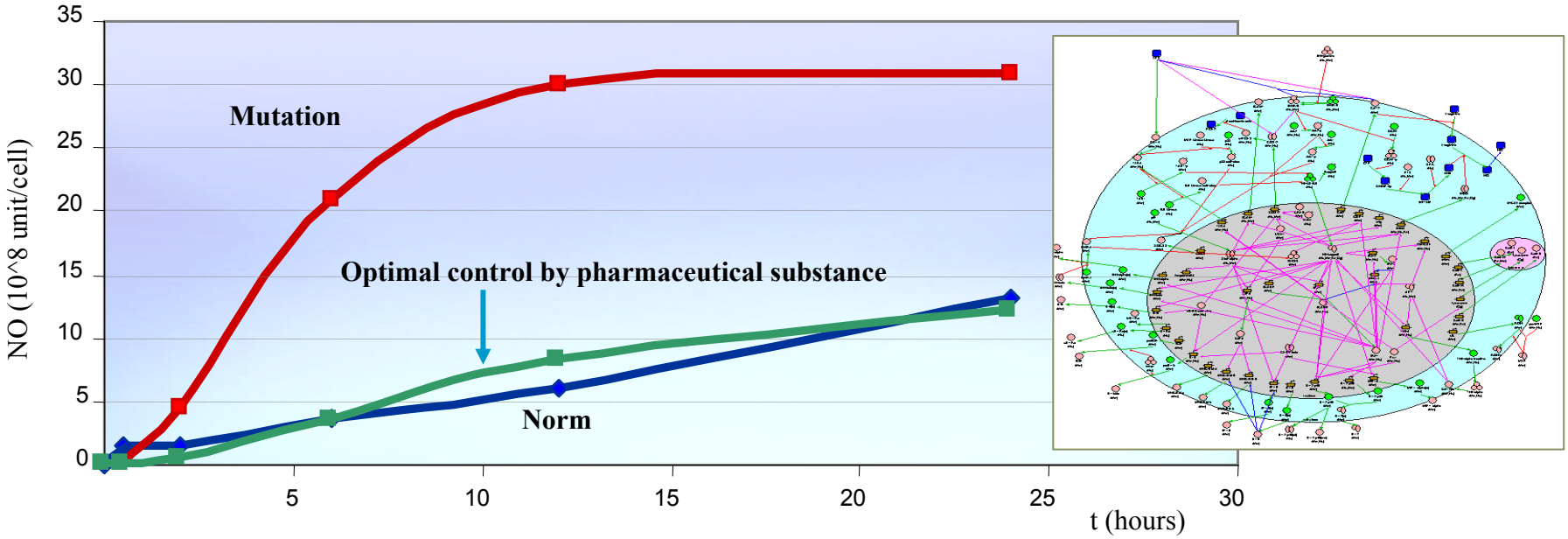
GeneNet database: a fragment of the gene network controlling NO synthesis under macrophage activation induced by bacterial infection



- iNOS (inducible nitric oxide synthase), an enzyme catalyzing NO synthesis;
- L-arginine, substrate;
- CAT2, L-arginine transporter
- H4B (tetrahydrobiopterin), cofactor.



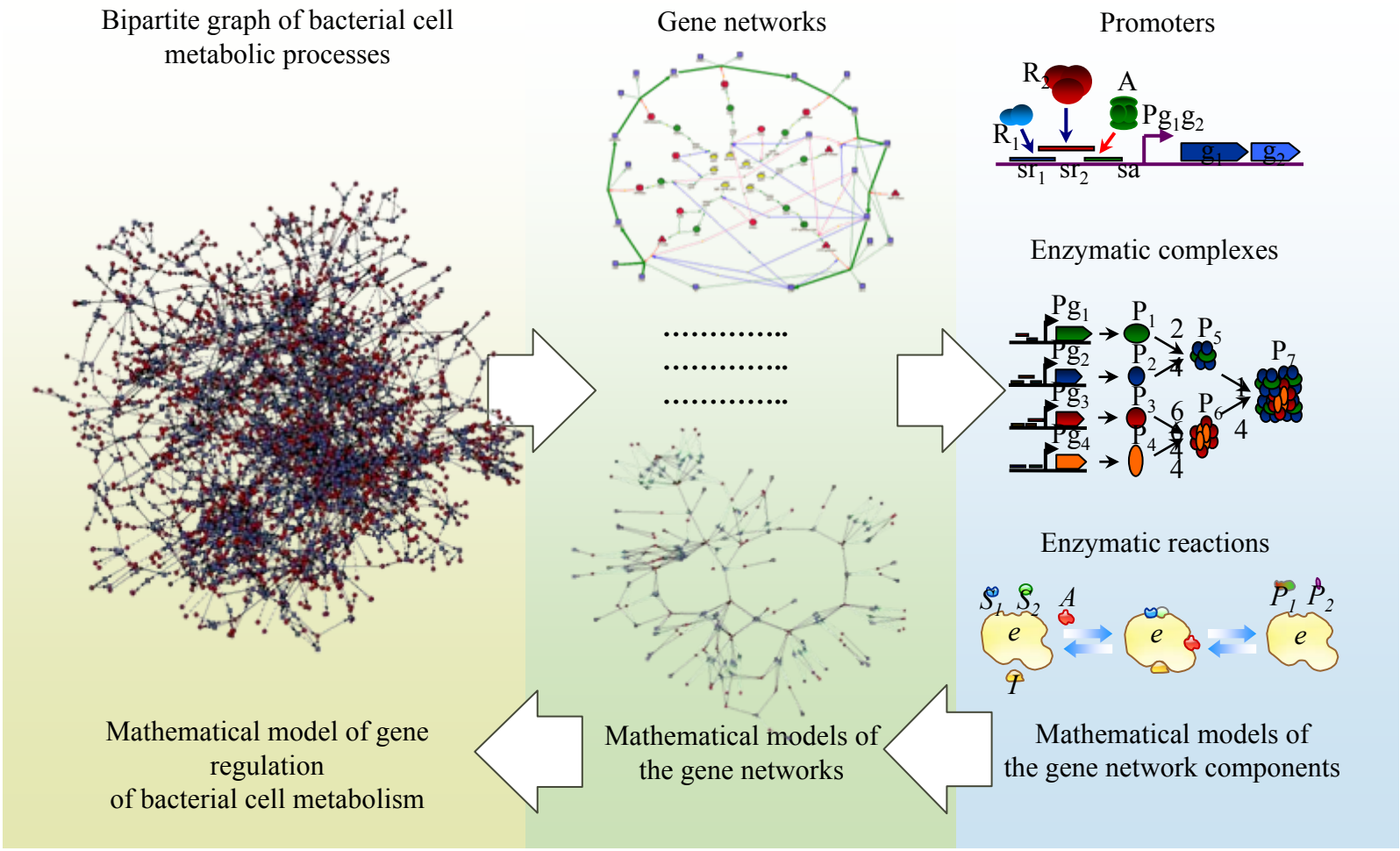
Macrophage activation gene network: modeling of the mutation influence on the dynamics of NO synthesis and optimal pharmaceutical control normalizing the function of the mutant gene network



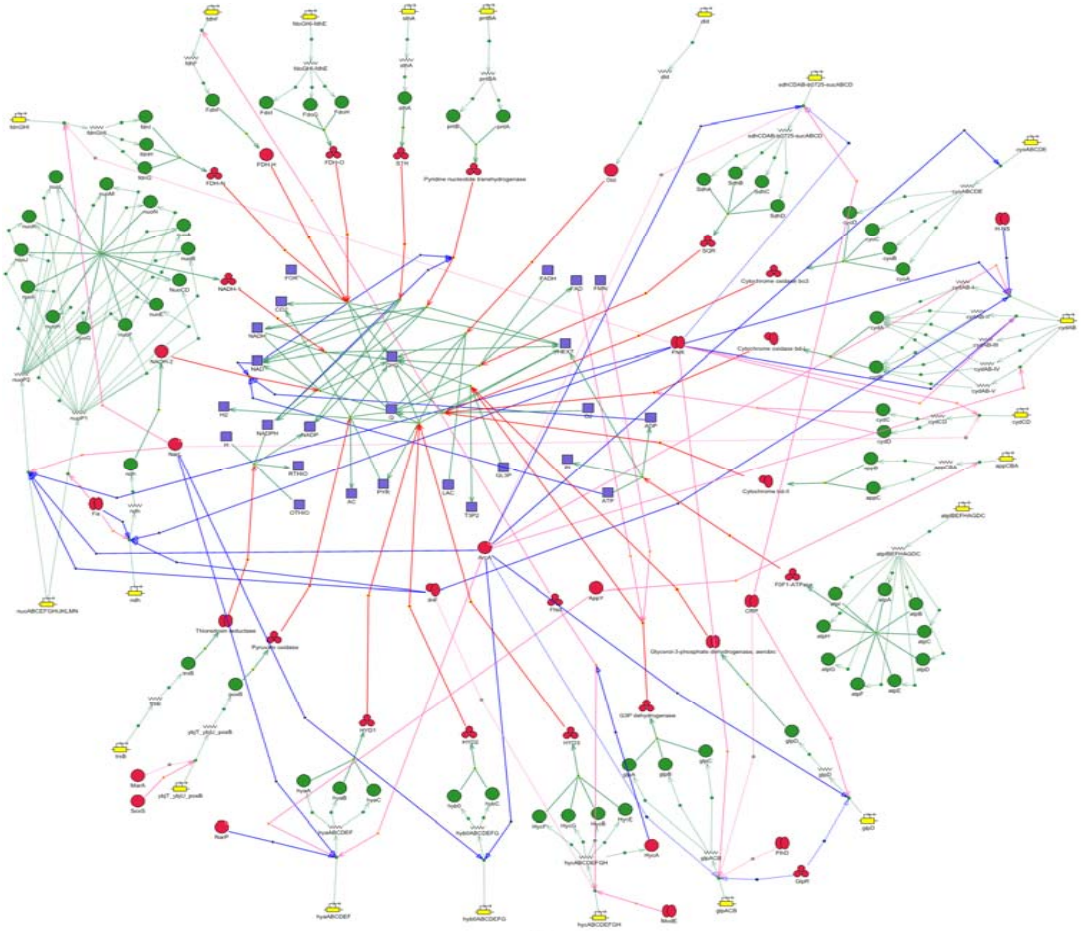
E. coli statistics

Cell size ~ 1×1×2 μm
Avarage weight ~ 10^{-12} g
Escherichia coli K12 genome - 4639221 n.p.
Number of protein coding genes - 4311
Number of operons ~ 2500
Number of promoters > 2500
Number of transcription regulatory proteins - 247
Number of enzymatic reactions ~ 1230
Number of proteins with catalytic activity - 1656
Number of metabolites ~ 1000
Avarage half-life of mRNA ~ 3-7 min
Cell division time ~ 20-40 min

Mathematical modeling of gene networks of *E.coli* metabolism. A principle scheme of the approach



Bacterial gene networks reconstruction: EC_GeneNet database

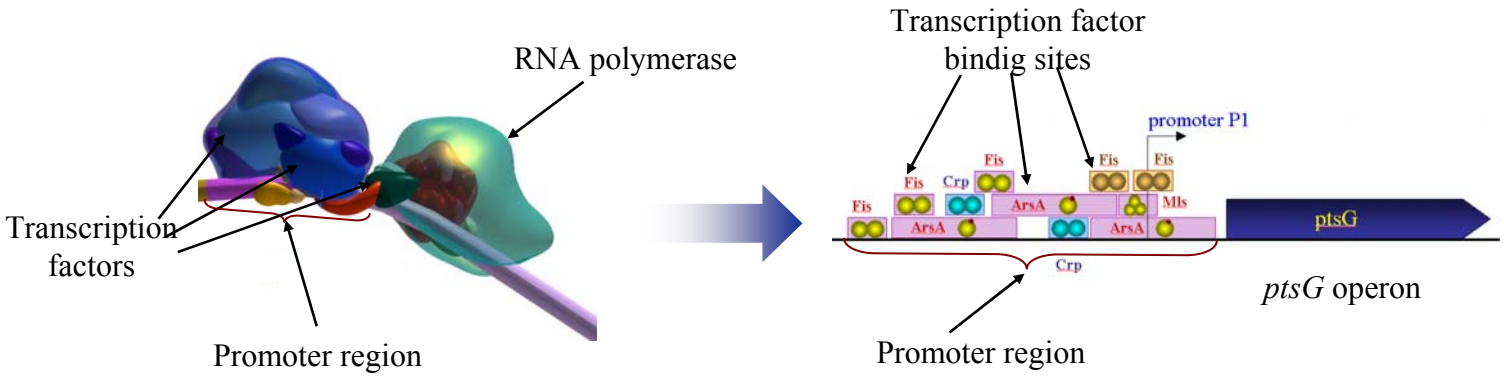


EC_GeneNet database (*E.coli* gene networks reconstruction).
Current version of EC_GeneNet contains descriptions more then 30 gene networks (biosynthesis of amino acids, nucleotides and cell respiratory processes)

Gene network controlling cell respiratory processes in *E. coli*

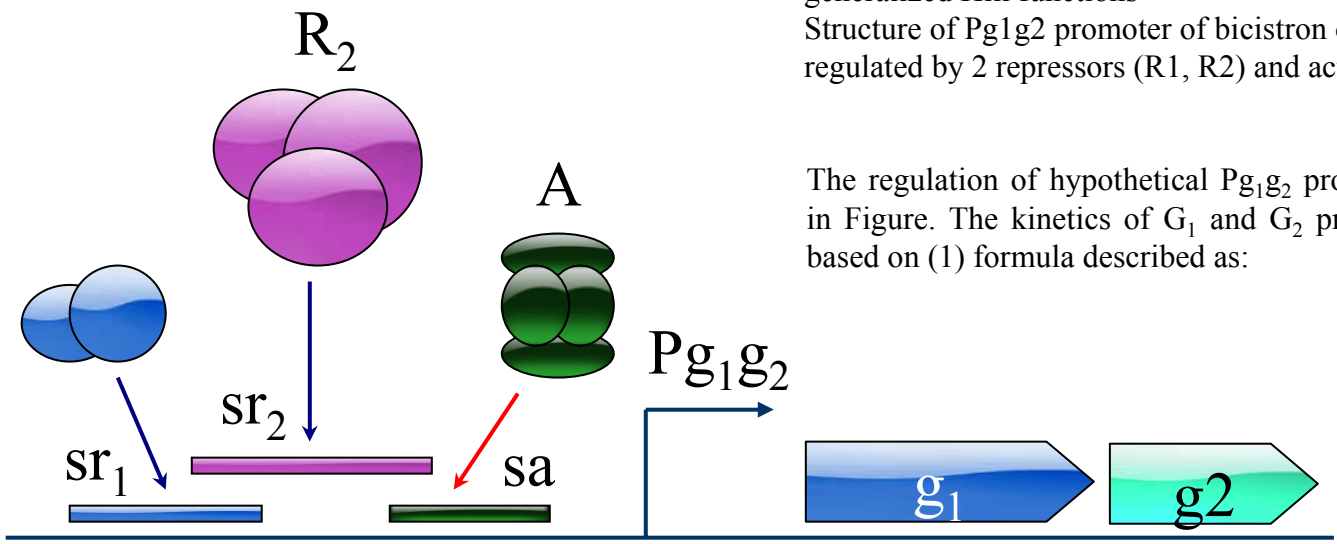
Description of bacterial gene transcription regulation in the TRRD database

EC_TRRD data base contains descriptions of 587 genes, 403 promoters, 1308 transcription factor binding sites and 3100 gene expression patterns.



EC_TRRD database

Methods and algorithms

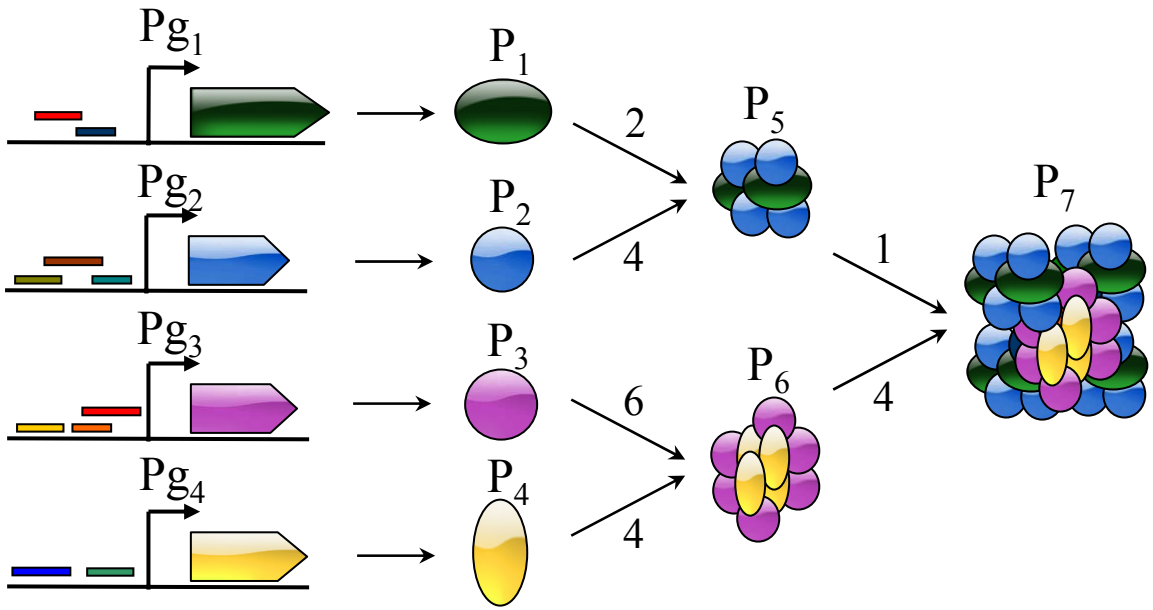


Modeling of genes expression regulation in terms of generalized Hill functions
Structure of Pg1g2 promoter of bicistron operon regulated by 2 repressors (R1, R2) and activator (A).

The regulation of hypothetical Pg_1g_2 promoter is shown in Figure. The kinetics of G_1 and G_2 proteins synthesis based on (1) formula described as:

$$\frac{dG_1}{dt} = \frac{dG_2}{dt} = k_{g12} \cdot \left(k + \left(\frac{A}{K_{sa}} \right)^{h_{sa}} \right) \Bigg/ \left(1 + \left(\frac{A}{K_{sa}} \right)^{h_{sa}} + \left(\frac{R_1}{K_{sr_1}} \right)^{h_{sr1}} + \left(\frac{R_2}{K_{sr_2}} \right)^{h_{sr2}} + \left(\frac{A}{K_{sa}} \right)^{h_{sa}} \cdot \left(\frac{R_2}{K_{sr_2}} \right)^{h_{sr2}} \right)$$

Methods and algorithms



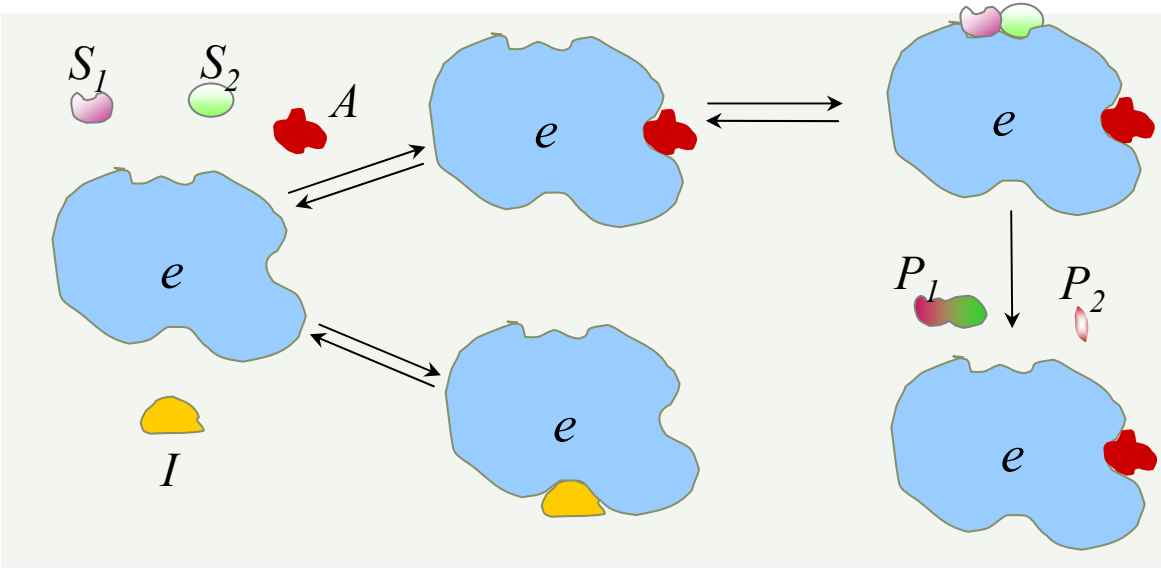
Modeling of enzyme complexes formation in terms of generalized Hill functions

Structure of stochiometric graph of the enzyme complex formation; graph arcs values correspond to stoichiometries of proteins or intermediate complexes entry in a complex.

Enzyme complexes concentrations are calculated on the basis of systems of reactions of intermediate complexes formation, which can be presented as stochiometric graphs (see an example in Figure) by using the iterative algorithm.

Methods and algorithms

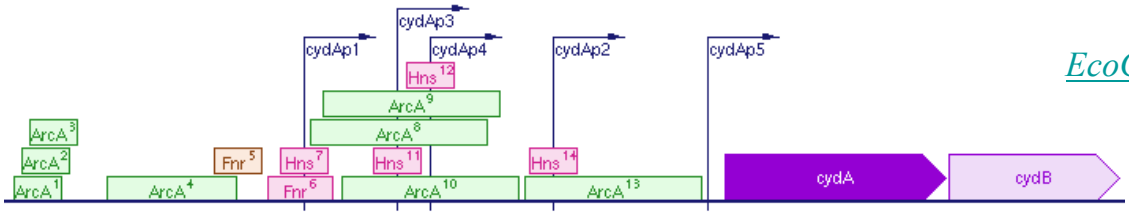
Modeling of kinetics of enzymatic reaction in terms of generalized Hill functions



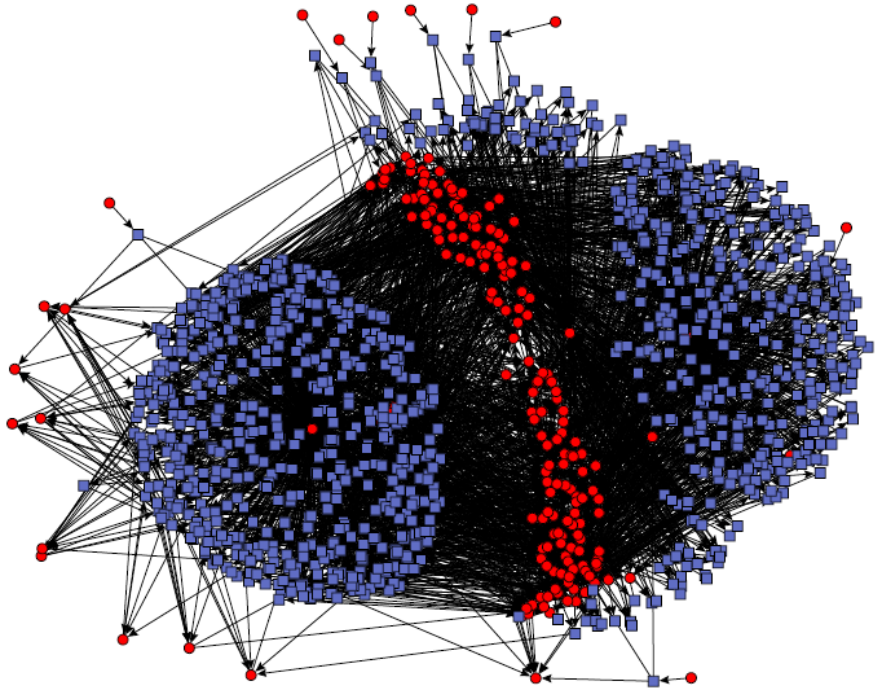
e – enzyme,
S1, S2 – substrates,
P1, P2 – products,
I – inhibitor,A – activator.

$$V = \frac{k_o \cdot e \dfrac{(S_1)^{h_{s1,1}} (S_2)^{h_{s2,1}} (A)^{h_{A,1}}}{(K_{s_1,s_2,A,1})^{h_{s1,1}+h_{s2,1}+h_{A,1}}}}{1 + \left(\frac{A}{K_{A,1}}\right)^{h_{A,1}} + \frac{(S_1)^{h_{s1,1}} (A)^{h_{A,1}}}{(K_{s_1,A,1})^{h_{s1,1}+h_{A,1}}} + \frac{(S_2)^{h_{s2,1}} (A)^{h_{A,1}}}{(K_{s_2,A,1})^{h_{s2,1}+h_{A,1}}} + \frac{(S_1)^{h_{s1,2}} (S_2)^{h_{s2,2}} (A)^{h_{A,2}}}{(K_{s_1,s_2,A,2})^{h_{s1,2}+h_{s2,2}+h_{A,2}}} + \left(\frac{R}{K_R}\right)^{h_R}}$$

Mathematical modeling of *cydAB* operon expression regulation in terms of elementary processes



EcoCyc database (www.ecocyc.org)



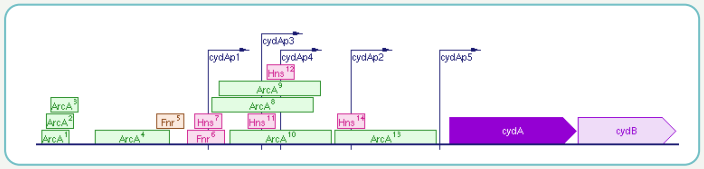
Fragment of system of differential equations:

$$\begin{cases} \dfrac{dcydAp}{dt} = -k_{11f}\cdot cydAp\cdot ArcA + k_{11r}\cdot cydAp_1ArcA \\ \qquad\qquad - k_{21f}\cdot cydAp\cdot Fnr + k_{21r}\cdot cydAp_1Fnr \\ \qquad\qquad - k_{31f}\cdot cydAp\cdot Hns + k_{31r}\cdot cydAp_1Hns + \ldots \\ \dfrac{dcydAp_1ArcA}{dt} = k_{11f}\cdot cydAp\cdot ArcA - k_{11r}\cdot cydAp_1ArcA \\ \dfrac{dcydAp_1Fnr}{dt} = k_{21f}\cdot cydAp\cdot Fnr - k_{21r}\cdot cydAp_1Fnr \\ \ldots \end{cases}$$

Bipartite graph of the mathematical model
5 transcription start regions;
10 binding sites of transcription factors.

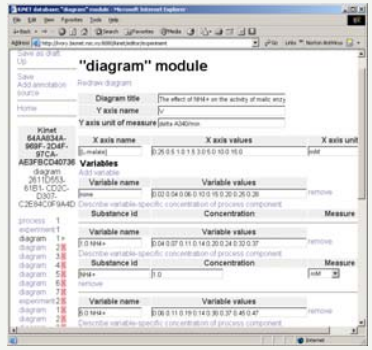
The mathematical model contains
197 dynamic variables;
510 processes.

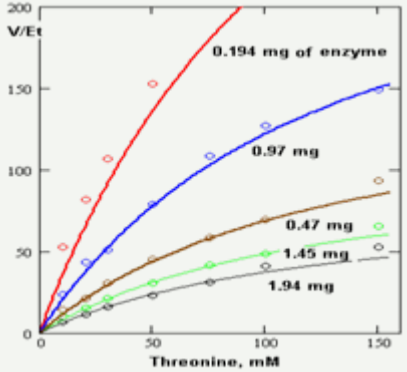
Mathematical modeling of *cydAB* operon expression regulation in terms of generalized Hill functions



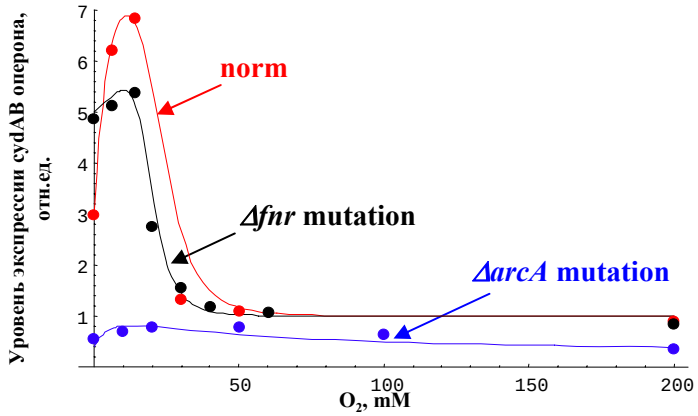
Hill function generation

EC_Kinet: database of kinetic and stationary dependences





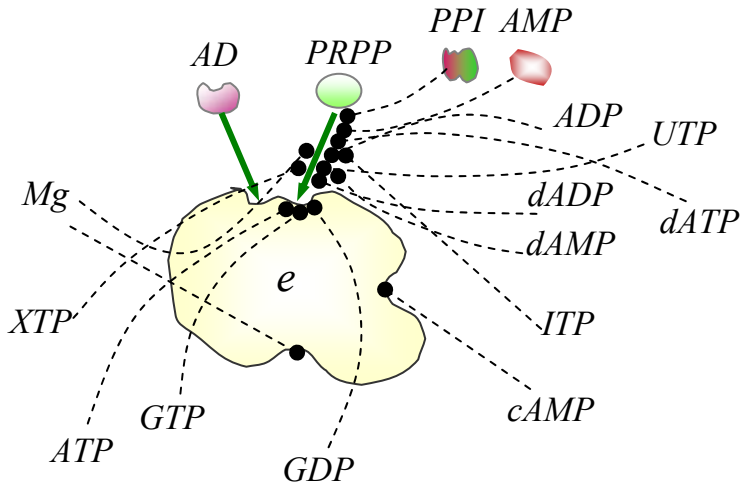
$$f_{cyd} = \frac{k_0 + \left(\frac{Fn}{k_F}\right)^{nF} + \left(\frac{A}{k_A}\right)^{nA} + \frac{Fn^{nFn_1} \cdot A^{nA_1}}{k1_{Fn,A}^{nFn_1 + nA_1}}}{1 + \left(\frac{A}{k_A}\right)^{nA} + \left(\frac{H}{k_H}\right)^{nH} + \left(\frac{Fn}{k_{Fn}}\right)^{nFn} + \frac{Fn^{nFn_1} \cdot A^{nA_1}}{k2_{Fn,A}^{nFn_1 + nA_1}} + \frac{A^{nA_2} \cdot H^{nH_1}}{k_{A,H}^{nA_2 + nH_1}} + \frac{H^{nH_2} \cdot Fn^{nFn_2}}{k_{H,Fn}^{nH_2 + nFn_2}} + \frac{Fn^{nFn_3} \cdot H^{nH_3} \cdot A^{nA_3}}{k_{Fn,H,A}^{nFn_3 + nH_3 + nA_3}}}$$



Dots on the plot are experimental data from (Tseng et al., J Bacteriol. 1996).
Curves are results of numerical calculation using mathematical model.

Modeling of enzymatic reaction regulation

AD + PRPP -> PPI + AMP



e – Adenine phosphoryltransferase,
S_1 – AD; S_2 – PRPP;
P_1 –PPI; P_2 – AMP.

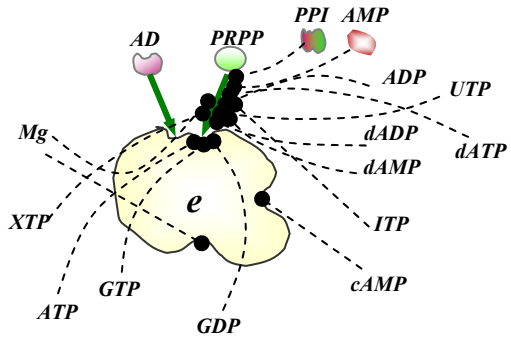
Regulators:
R_1 – ADP; R_2 – dADP;
R_3 – ATP; R_4 – dATP;
R_5 – dAMP; R_6 – GTP;
R_7 – ITP; R_8 – XTP;
R_9 – UTP; R_{10} – GDP;
R_{11} – Mg; R_{12} - cAMP

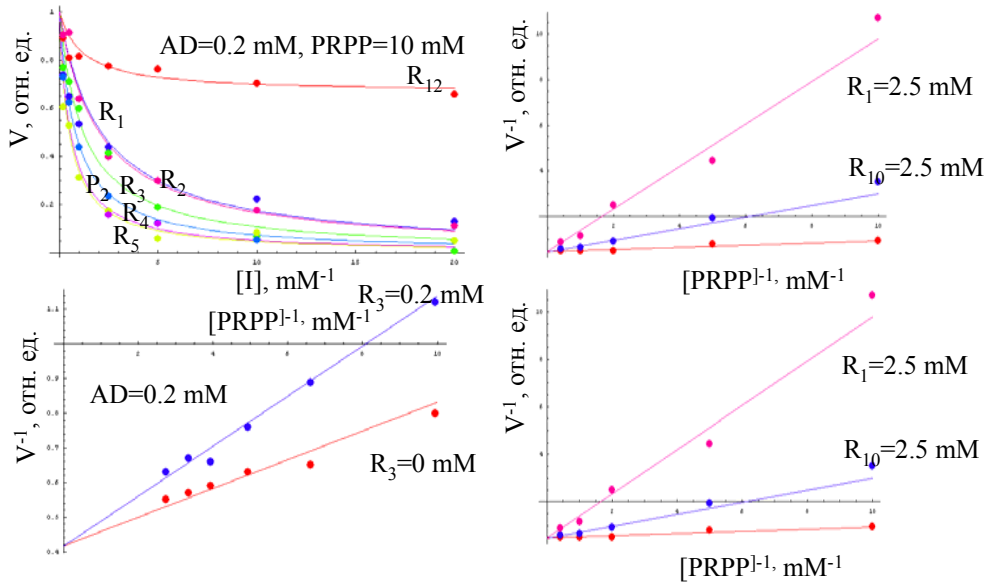
Fragment of system of differential equations:

$$\begin{cases} \frac{de}{dt} = -k_{1f}\cdot e\cdot AD + k_{1r}\cdot eAD + k_{fr}\cdot eAMP - k_{1freg}\cdot e\cdot cAMP + k_{1rreg}\cdot ecAMP - k_{2freg}\cdot e\cdot Mg + k_{2rreg}\cdot eMg \\ \frac{deAD}{dt} = k_{2f}\cdot e\cdot AD - k_{1r}\cdot eAD - k_{2f}\cdot eAD\cdot PRPP + k_{2r}\cdot eADPRPP - k_{3f}\cdot eAD\cdot ADP + k_{3r}\cdot eADADP - k_{4f}\cdot eAD\cdot AMP + k_{4r}\cdot eADAMP \\ \qquad - k_{5f}\cdot eAD\cdot UTP + k_{5r}\cdot eADUTP - k_{6f}\cdot eAD\cdot ITP + k_{6r}\cdot eADITP - k_{7f}\cdot eAD\cdot PPI + k_{7r}\cdot eADPPI - k_{8f}\cdot eAD\cdot PPI + k_{8r}\cdot eADPPI + \ldots \\ \frac{deADPRPP}{dt} = k_{2f}\cdot eAD\cdot PRPP - k_{2r}\cdot eADPRPP - k_{fp1}\cdot eADPRPP \\ \ldots \end{cases}$$

Modeling of enzymatic reaction regulation in terms of generalized Hill functions

AD + PRPP -> PPI + AMP





Dots on the plots are experimental data from (Hochstadt-Ozer & Stadtman, J. Biol. Chem., 1971).
Curves are results of numerical calculation using mathematical model.

Steady-state rate equation:

$$V = \frac{k_f \cdot e_0 \cdot \frac{S_1}{K_{m,S_1}} \cdot \frac{S_2}{K_{m,S_2}}}{\left(1+\frac{S_1}{K_{m,S_1}}\right)\cdot\left(1+\frac{S_2}{K_{m,S_2}}+\frac{P_1}{k_{i,P_1}}+\frac{P_2}{k_{i,P_2}}+\sum_{j=1}^{5}\frac{R_j}{k_{i,R_i,S_2}}+\sum_{j=6}^{10}\left(\frac{R_j}{k_{i,R_i,S_2}}\right)^2+\frac{R_{11}}{k_{i,R_{11},S_2}}\right)} \cdot \frac{1}{1+kI_{R_{12}}\cdot\frac{R_{12}}{k_{R_{12}}+R_{12}}}$$

1.3. Computer bacterial cell: approaches, results, perspectives (continued)

1.3.5. Mathematical modeling of the anaerobic and aerobic catabolic pathways in *E. coli*

1.3.5.1. Schematic presentation of major routes of the anaerobic and aerobic catabolic pathways in E. coli

1.3.5.2. Generalized flux model of the effect of the oxygen supply rate on the in vivo TCA cycle activity

1.3.5.3. Generalized flux model of the effect of the oxygen supply rate on on the formation rate of acetate

1.3.5.4. Generalized flux models of the effect of the oxygen supply rate on activities of major routes of catabolic pathways in E.coli

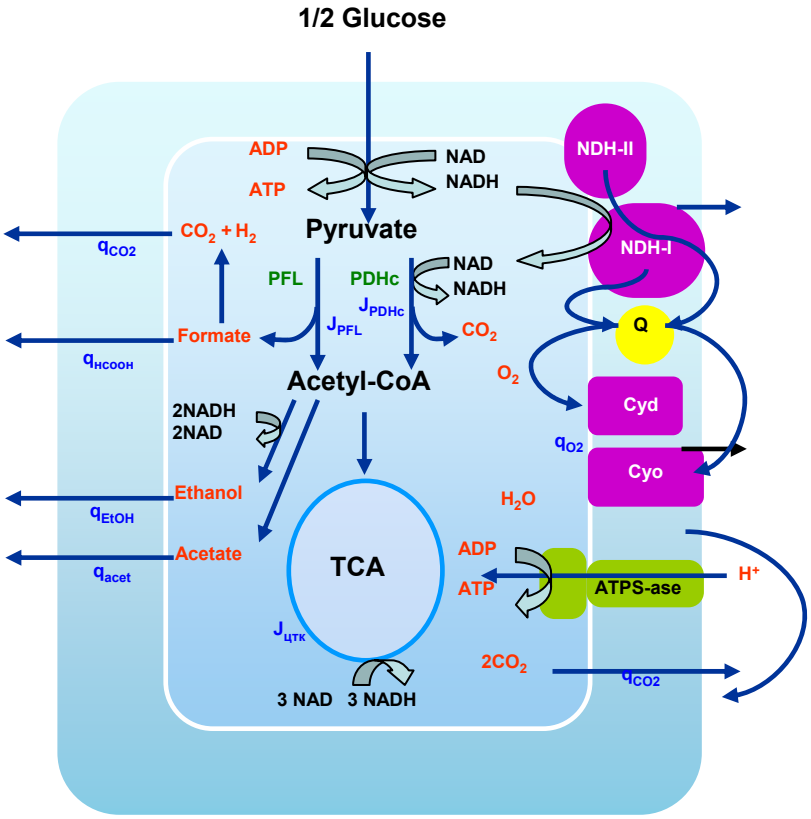
1.3.5.5. Overall flow model of oxygen consumption rate by the cell

1.3.5.6. Mathematical modeling of the effect of oxygen concentration on the rate of its consumption by cytocrome bo and bd oxydases of wild type and ΔarcA mutant

1.3.5.8. Prediction based on the models of functioning of cytochrome oxidase bd and cytochrome oxidase bo of Δfnr mutant

1.3.6. Conclusions

Major routes of the anaerobic and aerobic catabolic pathways in *E. coli*



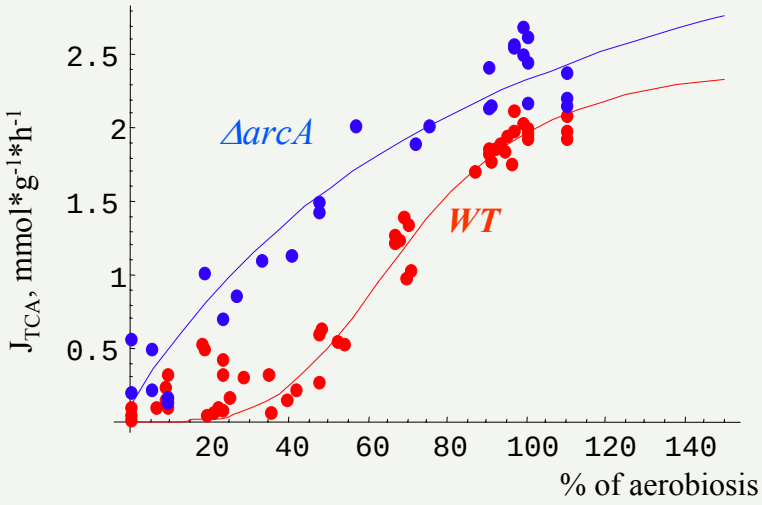
Goal: to construct mathematical models of catabolic pathways and investigate oxygen consumption rate by the cell using the models.

q_{CO2} - total formation rate of carbon dioxide
q_{HCOOH} –formation rate of formate
q_{EtOH} - formation rate of ethanol
q_{acet} - formation rate of acetate
q_{O2} - total oxygen consumption rate
J_{TCA} - carbon flux through TCA cycle
J_{PDHc} - carbon flux through PDHc (pyruvate dehydrogenase complex)
J_{PFL}- carbon flux through PFL (pyruvate formate-lyase)

Cyd – cytochrome oxydase bd,
Cyo – cytochrome oxydase bo,
PFL – pyruvate formate-lyase,
PDHc – pyruvate dehydrogenase complex,
NDH-I – NADH dehydrogenase I,
NDH-II – NADH dehydrogenase II

Alexeeva et al., J Bacteriol. 2000

Effect of the oxygen supply rate on the *in vivo* TCA cycle activity. Modeling in terms of generalized Hill functions



Dots on the plot are experimental data from (Alexeeva et al., J Bacteriol. 2003).
Curves are results of numerical calculation using mathematical models.

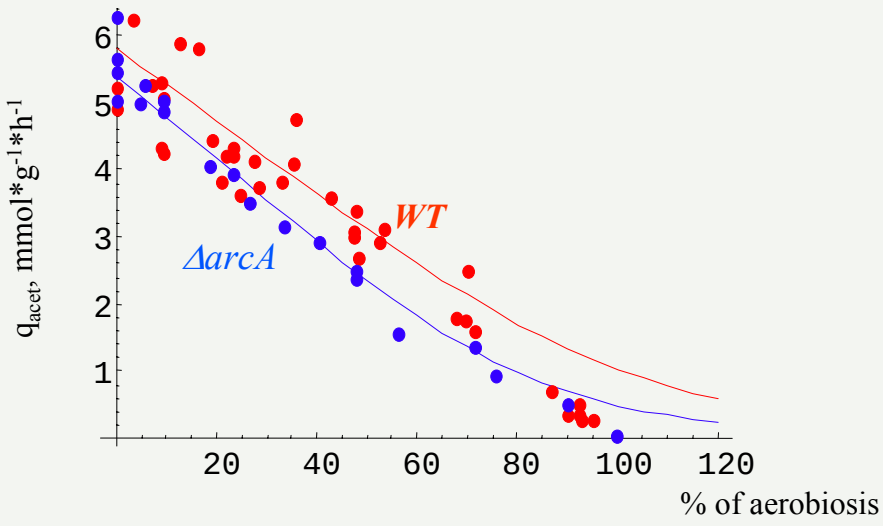
$$J_{(WT)} = \frac{k_0 + \left(\frac{O_2}{k_1}\right)^{h_1}}{1 + \left(\frac{O_2}{k_2}\right)^{h_2}}$$

$k_0=0;\ k_1=56;$
$k_2=70;\ h_1=4;\ h_2=4$

$$J_{(\Delta arcA)} = \frac{ka_0 + \left(\frac{O_2}{ka_1}\right)^{ha_1}}{1 + \left(\frac{O_2}{ka_2}\right)^{ha_2}}$$

$ka_0=0.1;\ ka_1=22;\ ka_2=100;$
$ha_1=1;\ ha_2=1.0$

Effect of the oxygen supply rate on the formation rate of acetate. Modeling in terms of generalized Hill functions



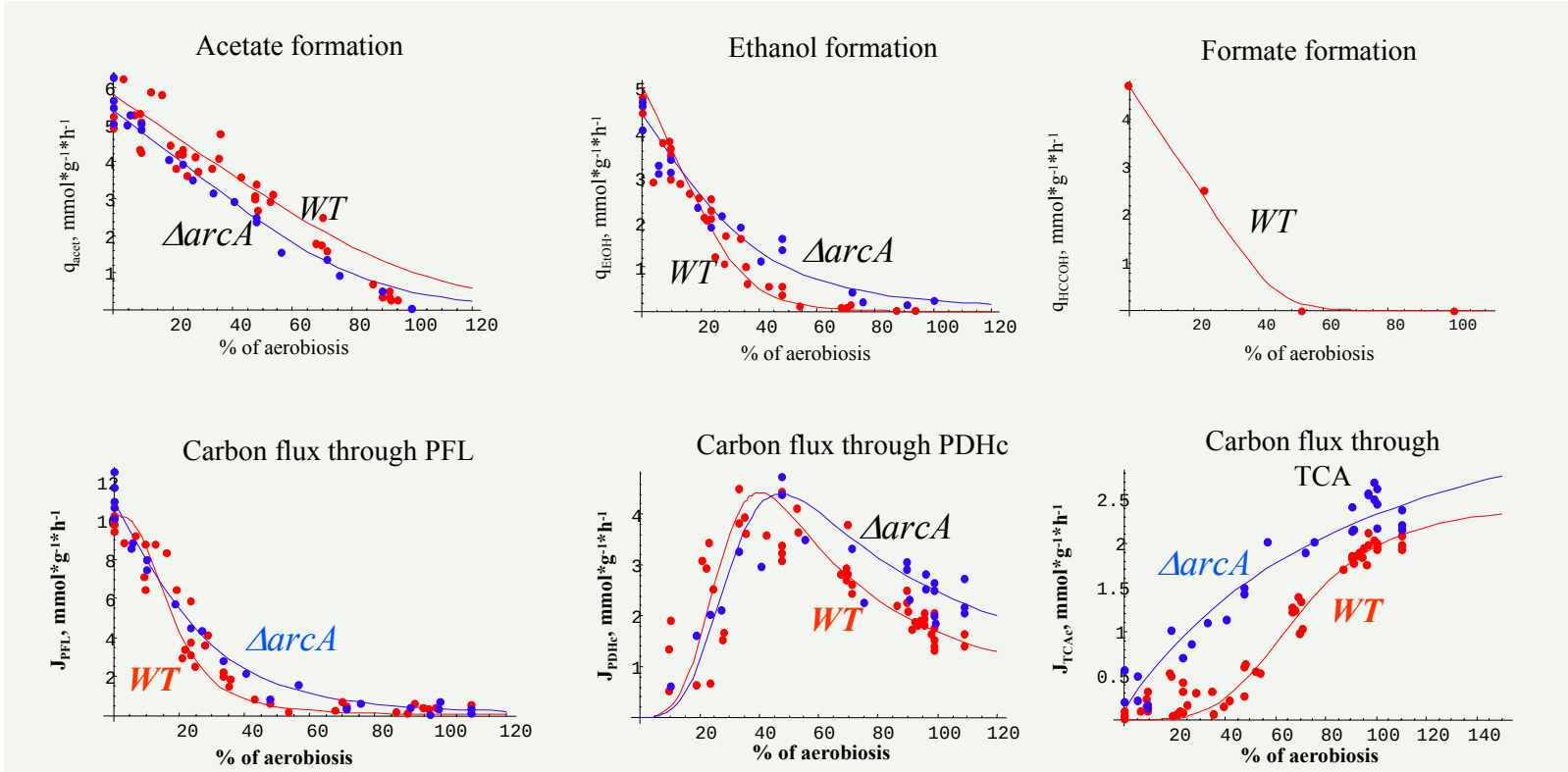
Dots on the plot are experimental data from (Alexeeva et al., J Bacteriol. 2003).
Curves are results of numerical calculation using mathematical models.

$$q_{acetate(WT)} = \frac{k_0}{1+\left(\frac{k_1\cdot O_2}{k_0}\right)+\left(\frac{k_1\cdot O_2}{k_0}\right)^2+\left(\frac{k_1\cdot O_2}{k_0}\right)^3+\left(\frac{k_1\cdot O_2}{k_0}\right)^4+\left(\frac{k_1\cdot O_2}{k_0}\right)^5+\left(\frac{k_1\cdot O_2}{k_0}\right)^6} \qquad k_0=5.8;\ k_1=0.054$$

$$q_{acetate(\Delta arcA)} = \frac{ka_0}{1+\left(\frac{ka_1\cdot O_2}{ka_0}\right)+\left(\frac{ka_1\cdot O_2}{ka_0}\right)^2+\left(\frac{ka_1\cdot O_2}{ka_0}\right)^3+\left(\frac{ka_1\cdot O_2}{ka_0}\right)^4+\left(\frac{ka_1\cdot O_2}{ka_0}\right)^5+\left(\frac{ka_1\cdot O_2}{ka_0}\right)^6} \qquad ka_0=5.4;\ ka_1=0.062$$

Effect of the oxygen supply rate on activities of major routes of catabolic pathways in *E.coli*. Modeling in terms of generalized Hill functions

Dots on the plot are experimental data from (Alexeeva et al., J Bacteriol. 2003).
Curves are results of numerical calculation using mathematical models.

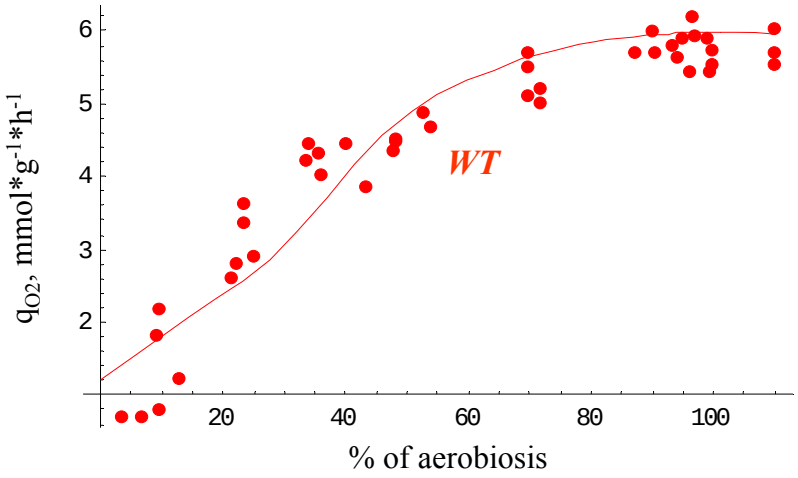


Overall flow model of oxygen consumption rate by the cell. Modeling in terms of generalized Hill functions

$J_{PDHc} = 2\,J_{glucose} - 3.808\,J_{PFL}$
$q_{CO2} = -0.667\,J_{glucose} + 0.667\,q_{O2} + 2.8\,J_{PFL}$
$q_{ац} = 3.333\,J_{glucose} - 0.333\,q_{O2} - 8.482\,J_{PFL}$
Delgado et al. 2000

$J_{TCA} = 1/2\,q_{CO2(TCA)}$
$q_{CO2(TCA)} = 2/3\,(q_{CO2} + q_{HCOOH} - q_{EtOH} - q_{acet})$
$q_{CO2} = 3 + q_{EtOH} + q_{acet} + q_{HCOOH}$
Alexeeva et al. 2000

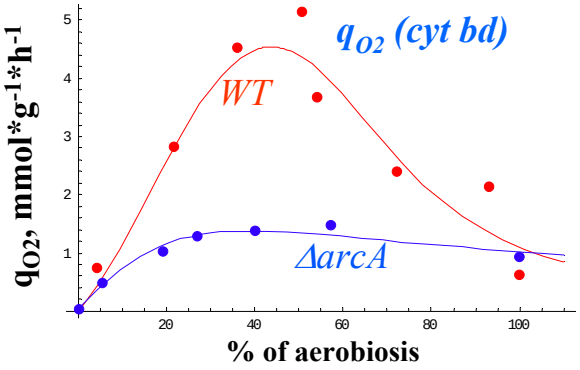
$$q_{O2} = 3.08\,(q_{эт} + q_{ац} + q_{HCOOH}) + 0.18\,q_{ацетат} - 1.67\,J_{ПДК}$$



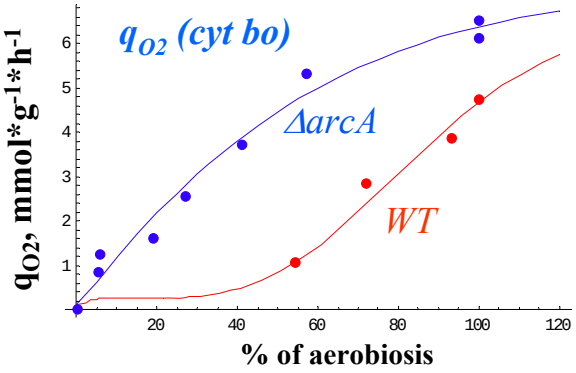
Dots on the plot are experimental data from (Alexeeva et al., J Bacteriol. 2003).
Curves are results of numerical calculation using mathematical models.

The effect of oxygen concentration on the rate of its consumption. Modeling in terms of generalized Hill functions

Dots on the plot are experimental data from (Alexeeva et al., J Bacteriol. 2003).
Curves are results of numerical calculation using mathematical models.

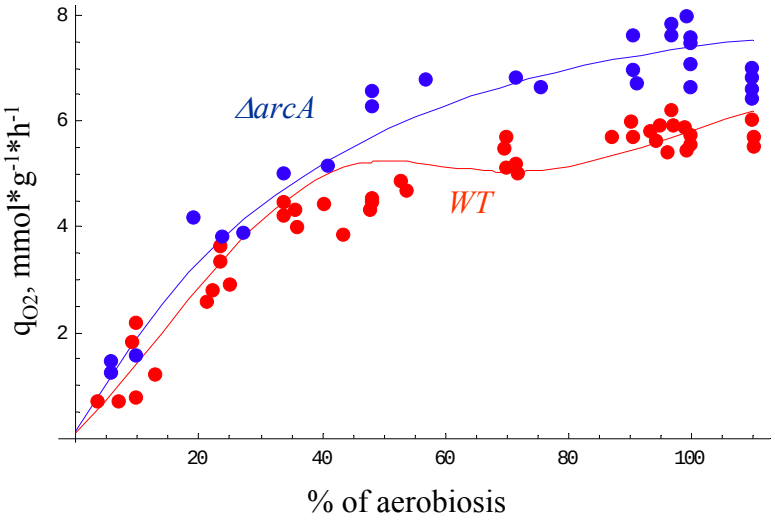


The effect of increase in oxygen concentration in the medium on the rate of its conversion by cytochrome oxidase bd in the cells of wild type and mutants (knockout strain in arcA gene)



The effect of increase in oxygen concentration in the medium on the rate of its conversion by cytochrome oxidase bo in the cells of wild type and mutants (knockout strain in arcA gene)

The rates of oxygen consumption by the cells of wild type and ΔarcA mutant. Modeling in terms of generalized Hill functions

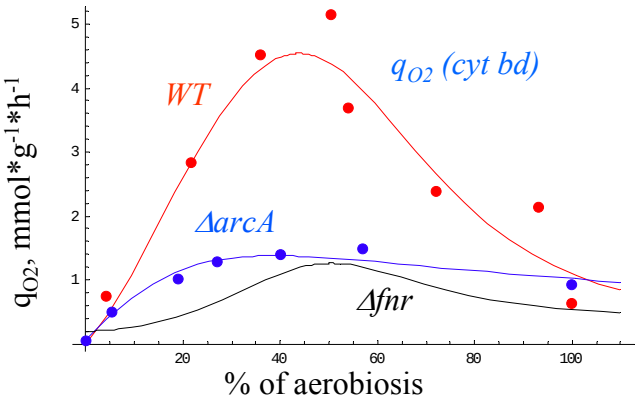


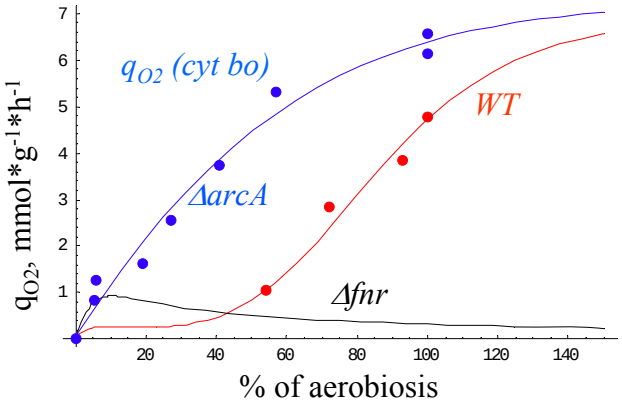
Dots on the plot are experimental data from (Alexeeva et al., J Bacteriol. 2000).
Curves are results of numerical calculation using mathematical models.

$$q_{WT}[O_2] = q_{WT(cyt\ bo)}[O_2] + q_{WT(cyt\ bd)}[O_2] \qquad q_{\Delta arcA}[O_2] = q_{\Delta arcA(cytbo)}[O_2] + q_{\Delta arcA(cytbd)}[O_2]$$

The effect of oxygen concentration on the rate of its consumption. Modeling in terms of generalized Hill functions.

Dots on the plot are experimental data from (Alexeeva et al., J Bacteriol. 2000).
Curves are results of numerical calculation using mathematical models.





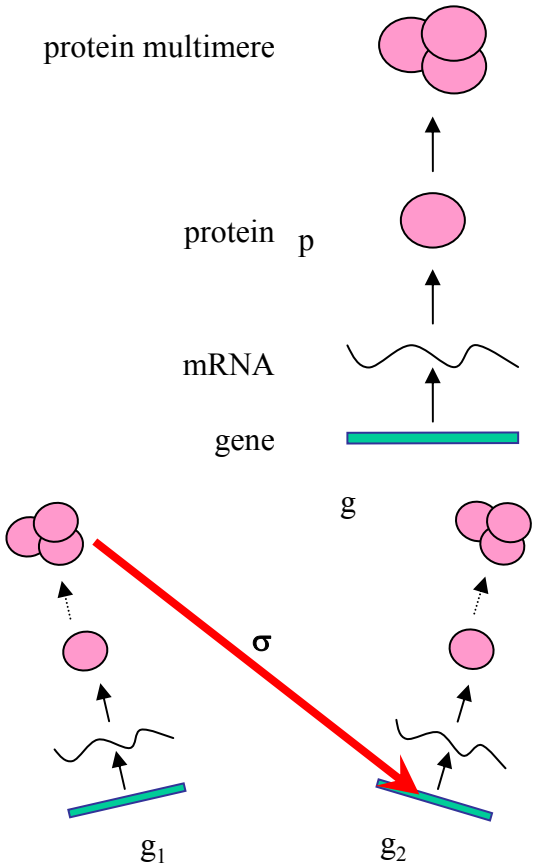
Prediction based on the models of functioning of cytochrome oxidase bd and cytochrome oxidase bo of Δfnr mutant

Conclusions

- The gene network of respiration regulation of the E. coli cell is reconstructed basing on 150 experimental papers. The network includes information about expression regulation of 20 operons (62 genes), functioning of 16 protein regulators, and formation of 20 enzymatic complexes involved in 13 metabolic reactions.
- Mathematical models are constructed that describe in terms of Hill functions the effect of oxygen on expression regulation of four operons (nuoA-N, ndh, cydAB, and cyoABCDE), encoding the main enzymatic complexes of the cell respiration system---NADH dehydrogenases (I and II) and terminal cytochrome oxidases (bd and bo types). The models were numerically adapted to experimental data on functioning of the cells of wild type and ΔarcA and Δfnr mutants.
- The flux models are constructed that describe (A) main catabolic pathways in E. coli cell depending on the saturation of the medium with oxygen, (B) cytochrome oxidases in cells of wild type and mutants depending on the saturation of the medium with oxygen, and (C) respiration of the wild type cell depending on the functioning of main cell catabolic pathways under changes in oxygen concentration in the medium.
- Structural stability of the model of expression of terminal oxidases in the cells of wild type and mutants during verification to match various experimental data was demonstrated as well as compliance of the predicted oxygen consumption rate by cells of various strains with experimental data.
- Using the mathematical model, the rates of oxygen consumption by cytochrome oxidases bd and bo were theoretically predicted for Δfnr mutant.
- The modeling approach presented is a way to construct a minimal model of cell functioning depending on certain environmental characteristics. Construction of such models is very promising for solving problems of optimization of biotechnological processes and etc.

1.4 Hypothetical gene networks:computer analysis and modeling

Elements of hypothetical gene networks



Genetical element (g) is an elementary structural unit of HGN
Functioning of genetic element is a protein synthesis
Activity of genetic element is the rate of protein synthesis

Product (p) is a protein encoded by **g**

Regulatory relation (σ) is an elementary unit of hypothetical gene network, by which
activity of one genetic element (g2) is repressed by another genetic element (g1)

Description of hypothetical gene networks dynamics

$$\frac{dp_i}{dt} = \alpha_i\left(1 + \sum_{j \in D_i} \beta_{i,j} p_j^{\gamma_{i,j}}\right) - \beta_i p_i \ , \ i = \overline{1,n} \ .$$

n is the number of genetic elements (g_i, i=1,…,n)

p_i is concentration of a protein encoded by i-th genetic element g_i

$$D_i = \left\{ j_1, \ldots, j_{k_i} \,\middle|\, \beta_{i,j_l}\gamma_{i,j_l} \neq 0,\ l = 1,..,k_i \right\}$$

is a set of numbers of genetic elements which are regulators of g_i;

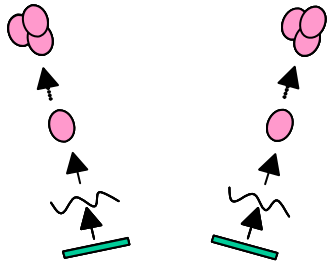
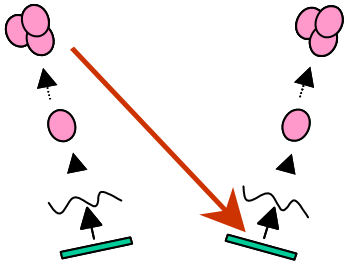
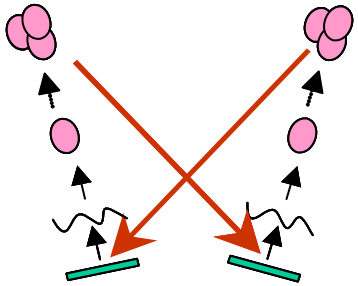
β_i are the constants of the rates of processes decreasing concentration of p_i (degradation, transport from compartment, etc.);

α_i, $\beta_{i,j}$, are coefficients regulating synthesis of the protein p_i by regulators p_j;

$\gamma_{i,j}$ is a measure of non-linear influence of p_j on activity g_i.

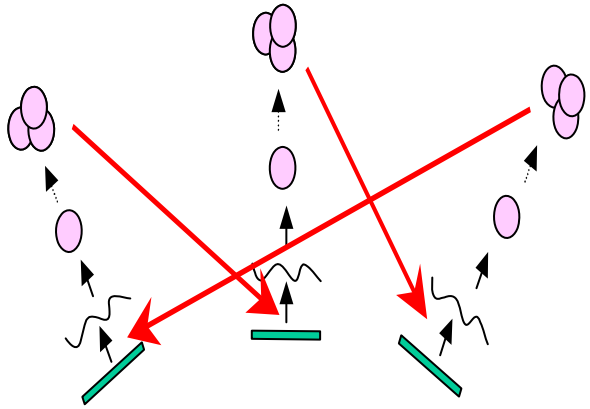
In the simplest case, $\gamma_{i,j}$ has the sense of dimensionality (measured as the number of subunits) of a molecule-regulator. In general case, it characterizes the complexity of regulatory process and it may be expressed by not-integer. From biological viewpoint, every β_i, α_i, $\beta_{i,j}$, $\gamma_{i,j}$, is not a negative value.

Hypothetical gene networks composed by two genetic elements

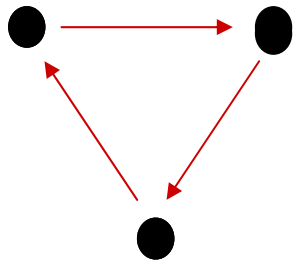
Diagramma of HGN	Structural graph	Mathematical model	Functioningregimes
		$dp_1/dt=a-p_1$ $dp_2/dt=a-p_2$	One stationary point
		$dp_1/dt=a-p_1$ $dp_2/dt=a/(1+p_1^g)-p_2$	One stationary point
		$dp_2/dt=a/(1+p_1^g)-p_2$ $dp_2/dt=a/(1+p_1^g)-p_2$	Two stationary points

The hypothetical gene networks compose by three genetic elements and three regulatory relations

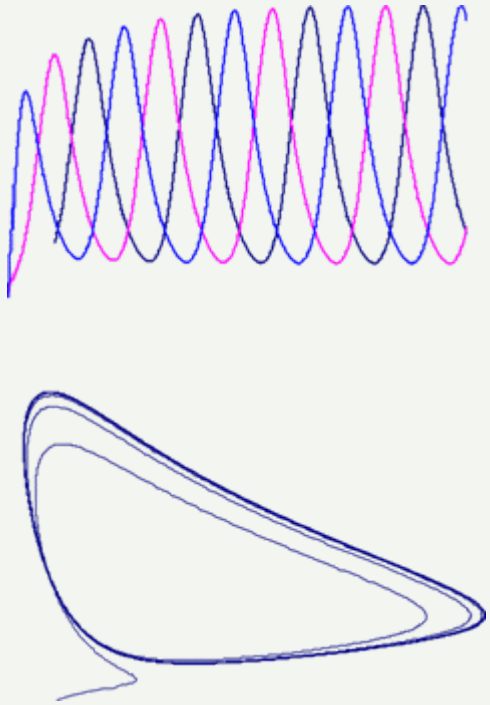
Diagramma of HGN



Structural graph



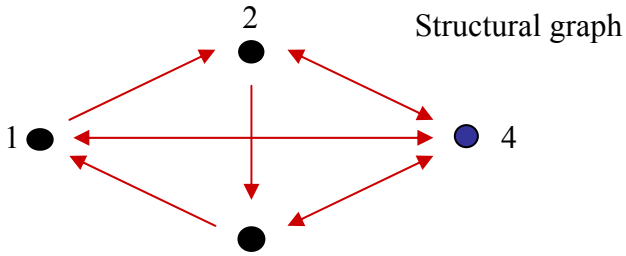
The stable limit cycle

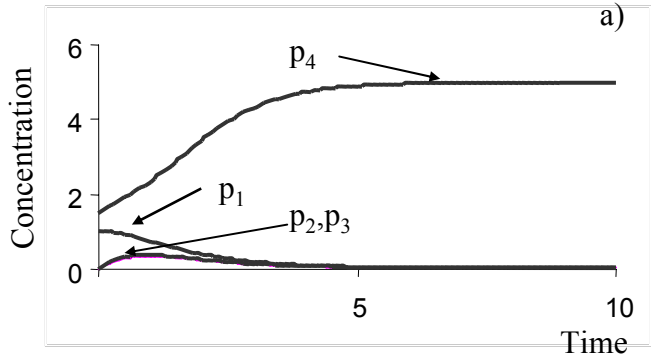


Mathematical model

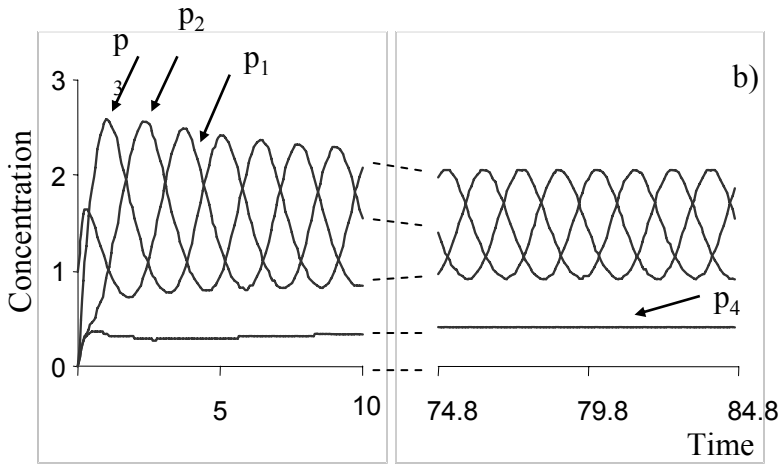
$$dp_1/dt = \alpha/(1+p_3^{\gamma}) - p_1$$
$$dp_2/dt = \alpha/(1+p_1^{\gamma}) - p_2$$
$$dp_3/dt = \alpha/(1+p_2^{\gamma}) - p_3$$

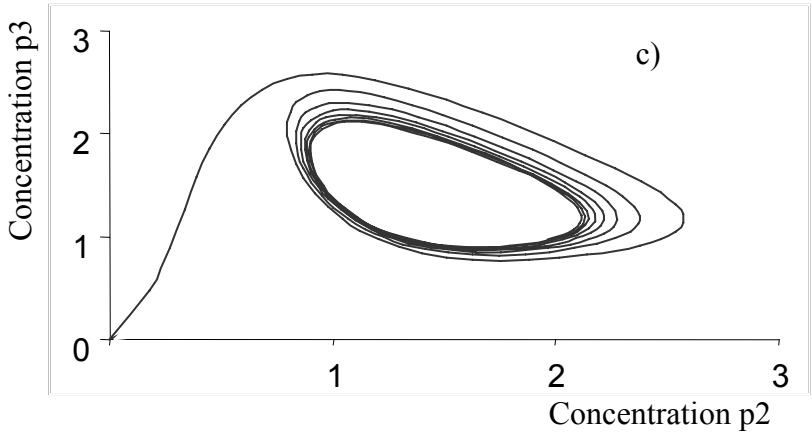
Hypothetical gene networks with four genetic elements, which have two qualitative different limit regimes





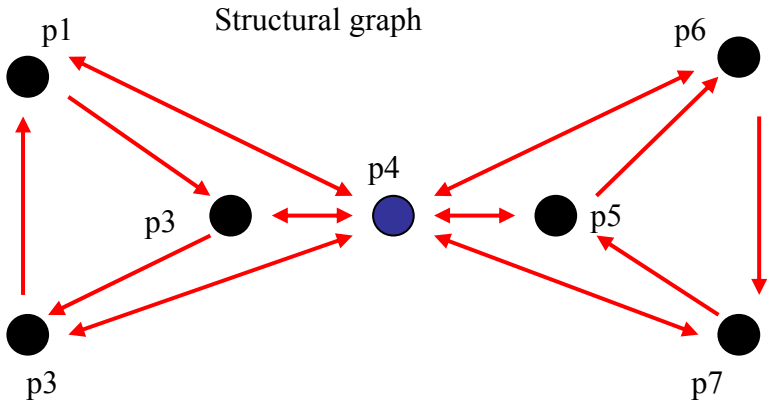
a) The stationary point is attracted by the initial data p1=1, p2=p3=0, p4=1.5

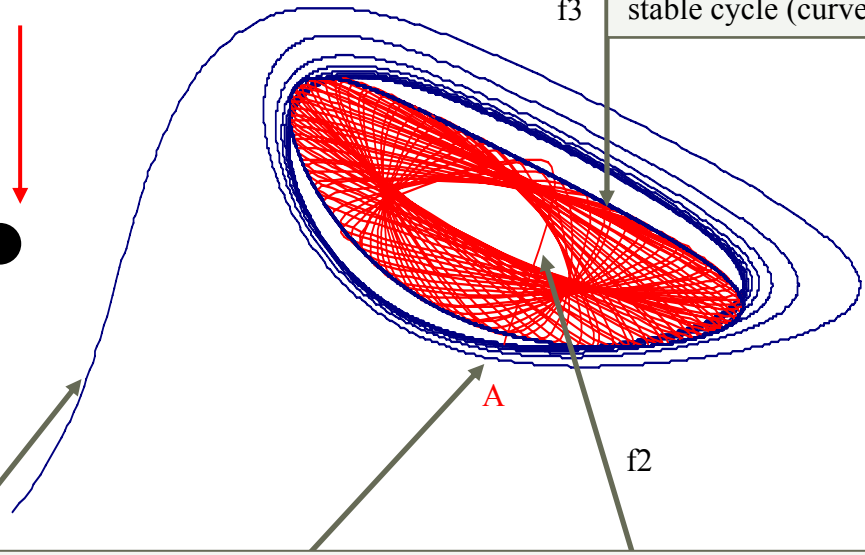




b,c) The oscillating behavior is attracted by the initial data p1=1, p2=p3=p4=0

An example of hypothetical gene networks consisting of 7 genetic elements, which have three stable cycles, one stable point and five unstable cycles.





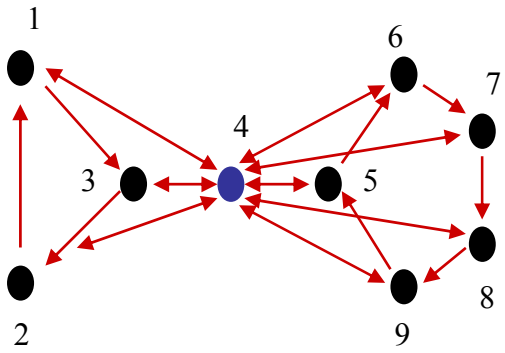
After stimuli A the HGN is attracted by the novel global stable cycle (curve f3).

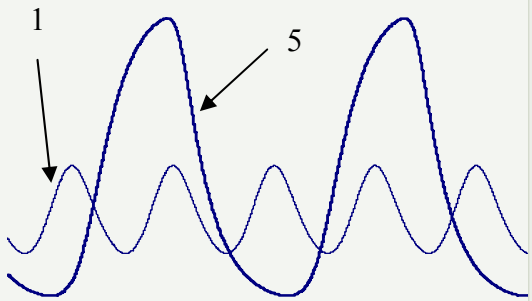
At the time=0, the HGN (curve f1) starts in the hyperspace p1=p5, p2=p6, p3=p7 and stays there, being attracted to the hyperspace stable cycle.

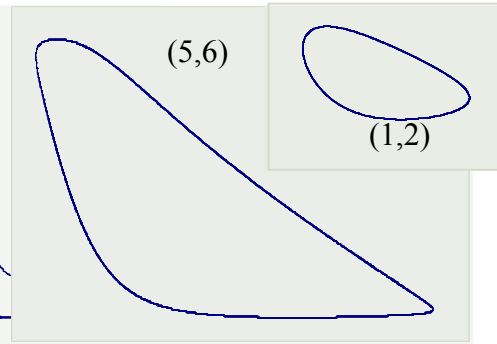
At the point A, we make external stimuli, which turned the HGN (curve f2) out of the hyperspace p1=p5, p2=p6, p3=p7.

An example of hypothetical gene networks consisting of 9 genetic elements, which have the stable quazi-periodical attractor

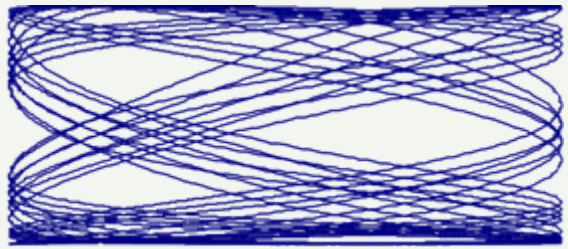
Structural graph



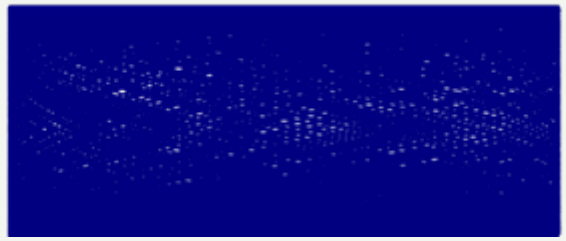




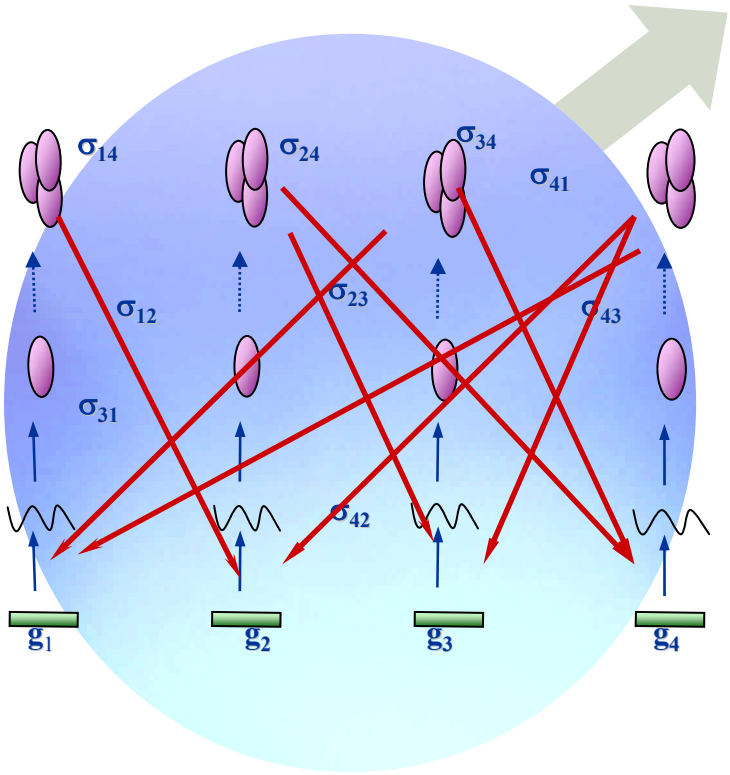
Phase curve (1,5)
10000 points,
The step 10^{-2}

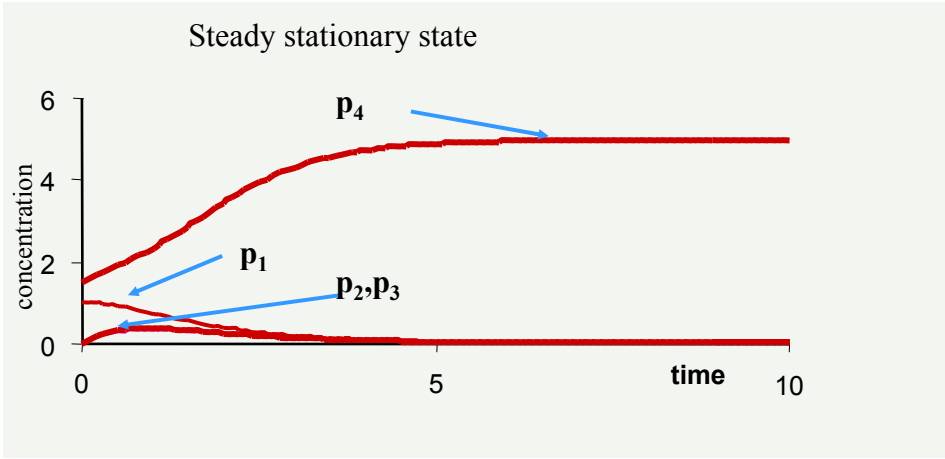


Phase curve (1,5)
10000 points,
The step 10^{-1}



A new regulatory association in the hypothetical gene network can dramatically affect its functioning



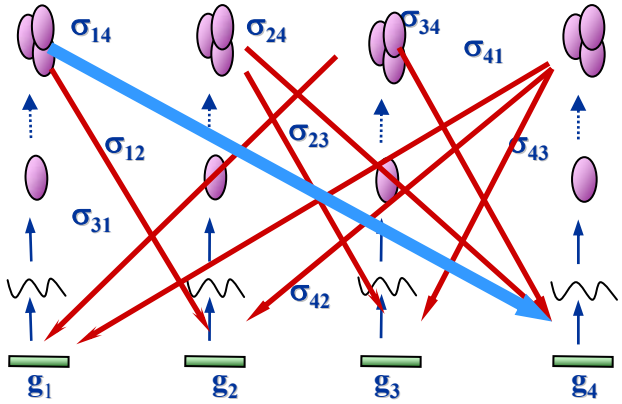


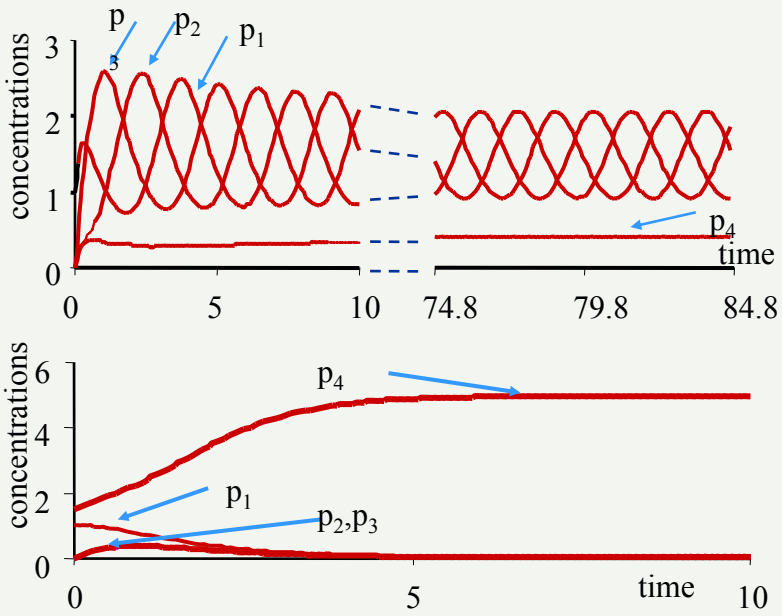
Initial variant of the gene network graph: 7 inhibiting regulatory associations.
FUNCTIONING IS THE ONLY STEADY STATIONARY STATE OF THE GENE NETWORK

As a new regulatory association appears (or an existing one disappears), gene network dynamics may dramatically change.

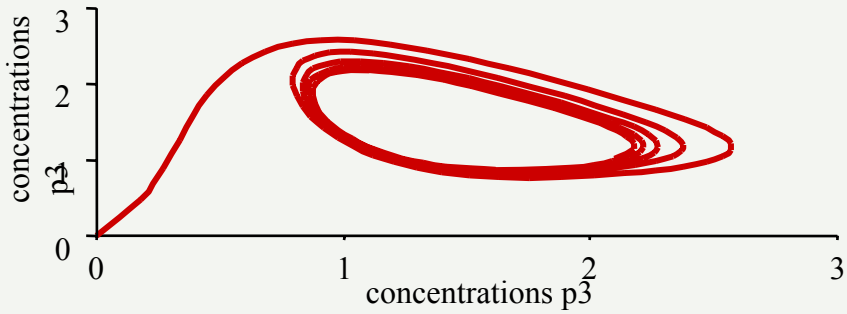
In the given case a gene network has two modes of dynamics: a stable stationary point or a stable cycle depending on the initial data of functioning

Regimes of functioning of initial gene network





Under initial conditions
p1=1, p2=p3=p4=0 and limitations
m≥3 and α≥5 , the gene network is characterized by cyclic pattern of functioning

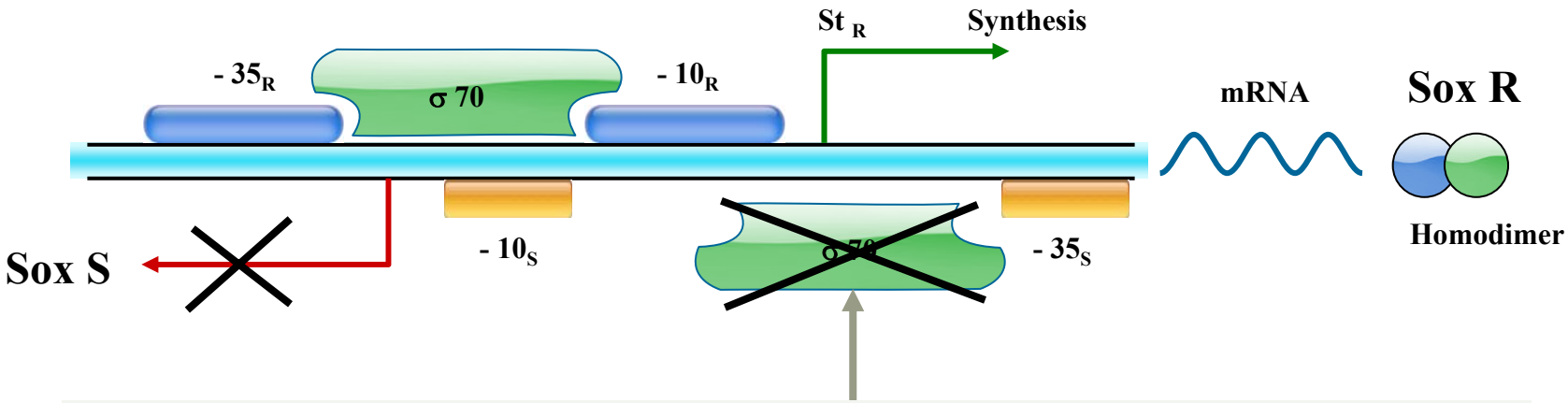


Under initial conditions
$p_1=1$, $p_2=p_3=0$, $p_4=1{,}5$, the gene network is characterized by stable stationary point (all variables are constant)

Artificial gene networks: genosensors for detection of biologically active components and stress factors

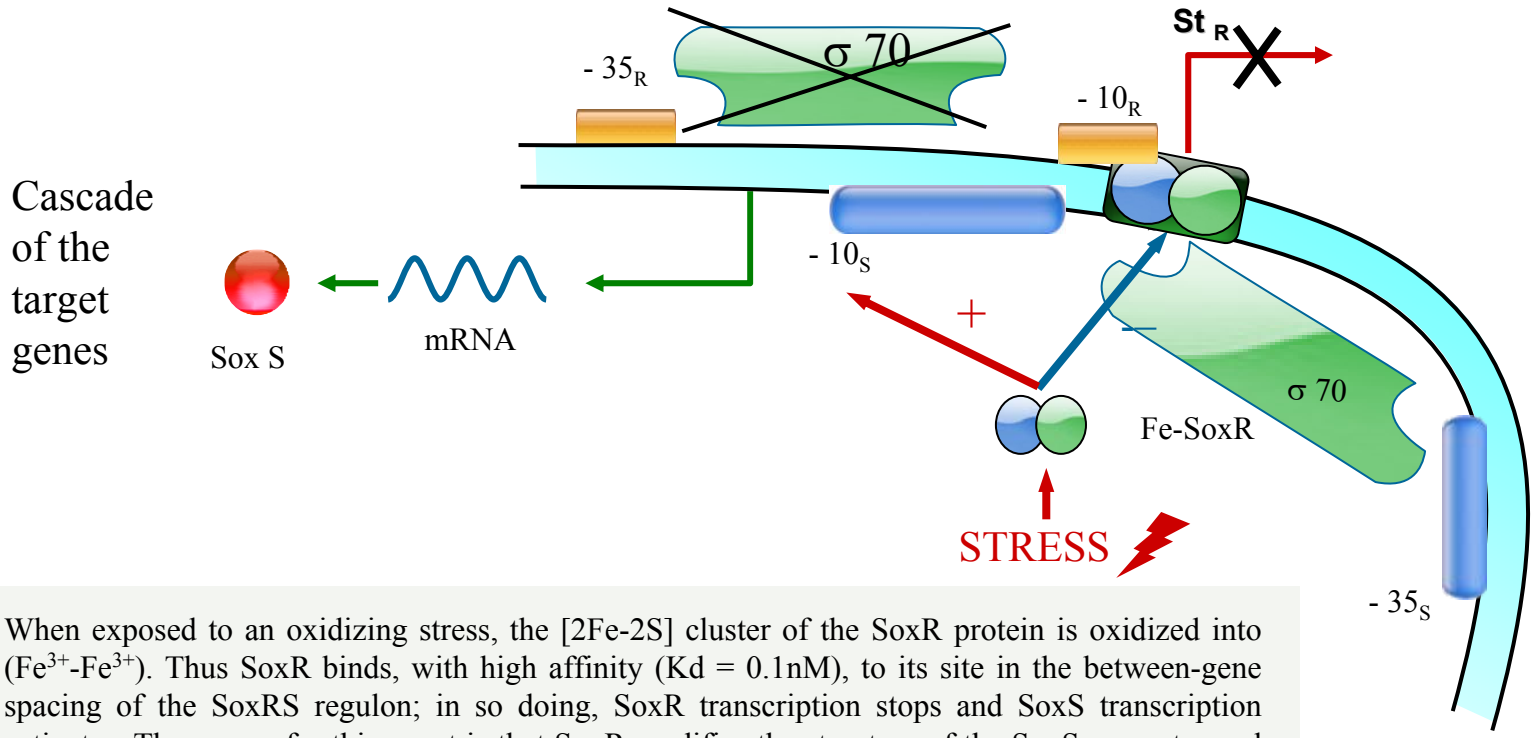
The genosensor is a complex biological structure, which allows bioactive media to be tested. As far as humans are concerned, testing water, air and foods for ecological safety is a matter of primary importance. And so is testing therapeutic drugs and biologically active food additives for their ability to adversely affect the normal functioning of the cell's metabolic system. The genosensor is what it is and does what it does because the cell is able to activate the expression of genes, whose products protect the cell from various damaging effects. The promoter of one of such genes is just the central object of the structure. This promoter is associated with the reporter gene, whose product appears as the system is activated. The genes encoding the structure of luminescent and fluorescent proteins (luxAB of the bacteria Vibrio fischeri or Photorhabdus luminescens, lucFF of the firefly, gfp of the jellyfish Aequorea victoria and dsred of Dictiosoma), which are easy to detect, are used as effector genes. The last component of the system is the sensor, which is a bacterial cell, into which the structure was introduced. All the components required for activation of the genosensor in response to a stressing action should be there. If the system is incomplete, the genes that encode the structure of the regulatory proteins required for the activation are forced into the system. In the next slide, a schematic of the SoxS gene promoter is presented. Its activation suggests that the medium being tested contains oxidizing agents, which damage all the main cell components: DNA, RNA, proteins and membranes.

Sox r/s regulon:a general schematic



The SoxR and SoxS genes have a common promoter; however, they are transcribed in opposite directions. When cell growth is regular, constitutive transcription of SoxR is observed and is then followed by synthesis of the SoxR protein. SoxS, however, tells a different story: its transcription is suppressed because the spacer between blocks –10 and –35 of the RNA-polymerase binding site on the SoxS promoter is two nucleotides longer than is required for RNA-polymerase. As a result, RNA-polymerase fails to efficiently and effectively read out information and only the SoxR protein is present in the cell. This protein is a dimer, each monomer containing the [2Fe-2S] cluster, and thus is inactive.

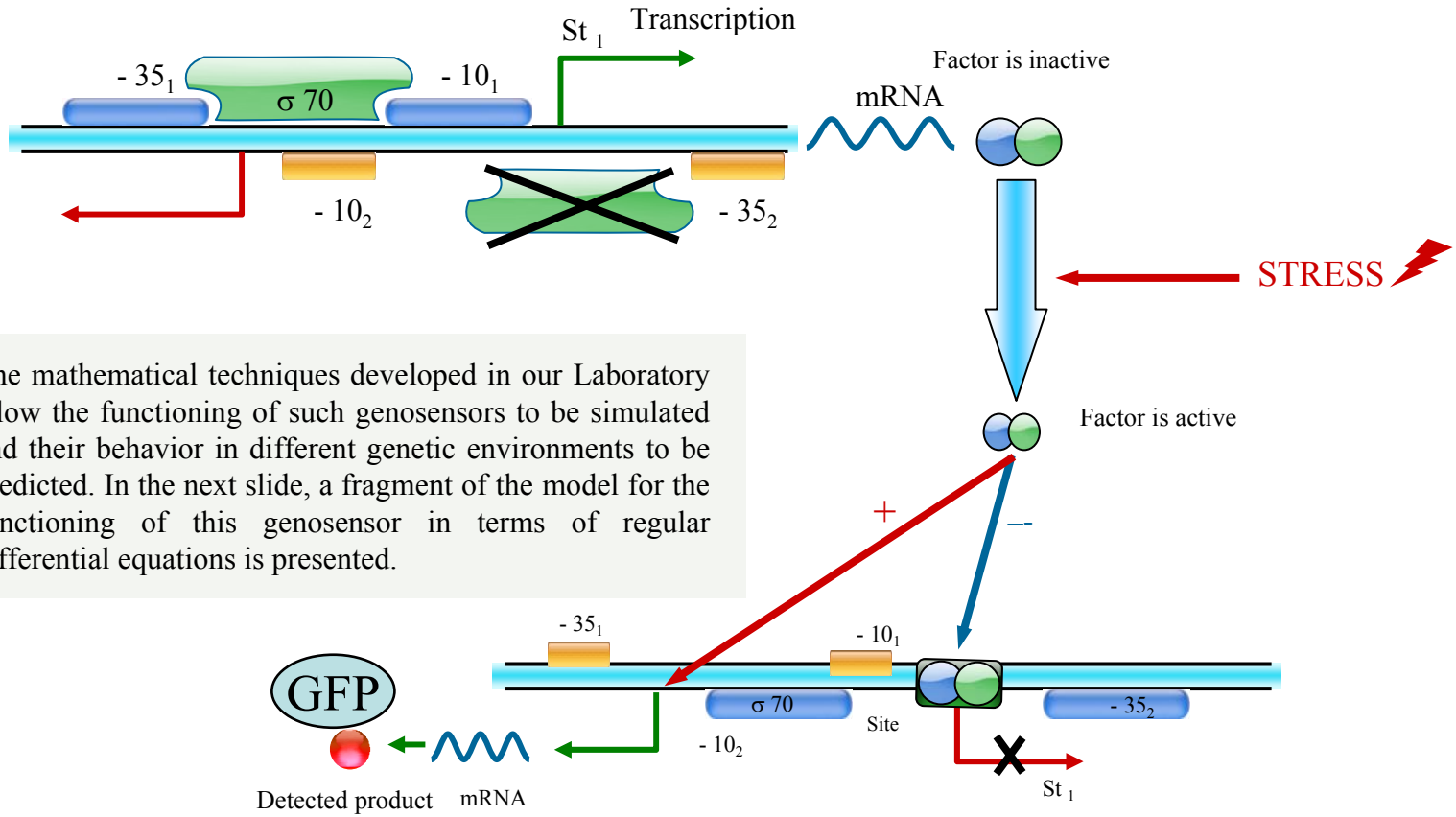
Sox R/S Regulon: activation of SoxS gene transcription in response to stress



When exposed to an oxidizing stress, the [2Fe-2S] cluster of the SoxR protein is oxidized into (Fe^{3+}-Fe^{3+}). Thus SoxR binds, with high affinity (Kd = 0.1nM), to its site in the between-gene spacing of the SoxRS regulon; in so doing, SoxR transcription stops and SoxS transcription activates. The reason for this event is that SoxR modifies the structure of the SoxS promoter and the polymerase can activate its transcription. SoxS mRNA appears in the cell two minutes after exposure to stress and 10 minutes later its level reaches its maximum. After the stressing agent is removed, SoxS mRNA disappears from the cell within 20 min.

A high level of induction of the SoxS promoter, a high sensitivity to a damaging agent (2-100 mkMol) and quick recovery from stress are the properties that make it very useful as part of a genosensor. In the next slide, a general schematic of this structure is presented.

The principal scheme of a genosensor activated by stress is based around elements of the SoxRS regulon. A stress reaction is detected by green fluorescent protein (GFP) fused with the promoter of SoxRS.



The mathematical techniques developed in our Laboratory allow the functioning of such genosensors to be simulated and their behavior in different genetic environments to be predicted. In the next slide, a fragment of the model for the functioning of this genosensor in terms of regular differential equations is presented.

A fragment of the mathematical model describing the functioning of the genosensor

...

$$dDBB/dt = -k_{s,BB} D \cdot BB + (k_{d,BB} - k_{s,RNA-pol}[RNA - pol])DBB + (k_{d,RNA-pol} + k_{ini,transcr})DBBR$$

$$dA/dt = k_t E_A + k_{v,B} B - (f_{\mathrm{mod},A} + k_{\deg r,A})A - 2k_{\dim,AA} A^2 + 2k_{dis,AA} AA - k_{\dim,AB} A \cdot B + k_{dis,AB} AB$$

$$dB/dt = -(k_{v,B} + k_{\deg r,B})B + f_{\mathrm{mod},A} A - k_{\dim,AB} A \cdot B + k_{dis,AB} AB - 2k_{\dim,BB} B^2 + k_{dis,BB} BB$$

$$dAA/dt = -(f_{\mathrm{mod},AA} + k_{\deg r,AA})AA + k_{v,AB} AB + k_{\dim,AA} A^2 - k_{dis,AA} AA$$

$$dAB/dt = -(f_{\mathrm{mod},AB} + k_{\deg r,AB} + k_{v,AB})AB + k_{v,BB} BB + k_{\dim,AB} A \cdot B - k_{dis,AB} AB + f_{\mathrm{mod},AA} AA$$

$$dBB/dt = -(k_{\deg r,BB} + k_{v,BB})BB + k_{\dim,BB} B^2 - k_{dis,BB} BB + f_{\mathrm{mod},AB} AB$$

...

Dynamic variables

A – protein A; B – modified form of protein A; AA – homodimer of protein A; AB – heterodimer of proteins A and B; BB – homodimer of protein B; EA – ribosome during protein A-encoding mRNA translation termination.
D – DNA-site for binding dimer BB; DBB – “dimer BB – DNA-site D” complex
[RNA-pol] – RNA polymerase

Parameters

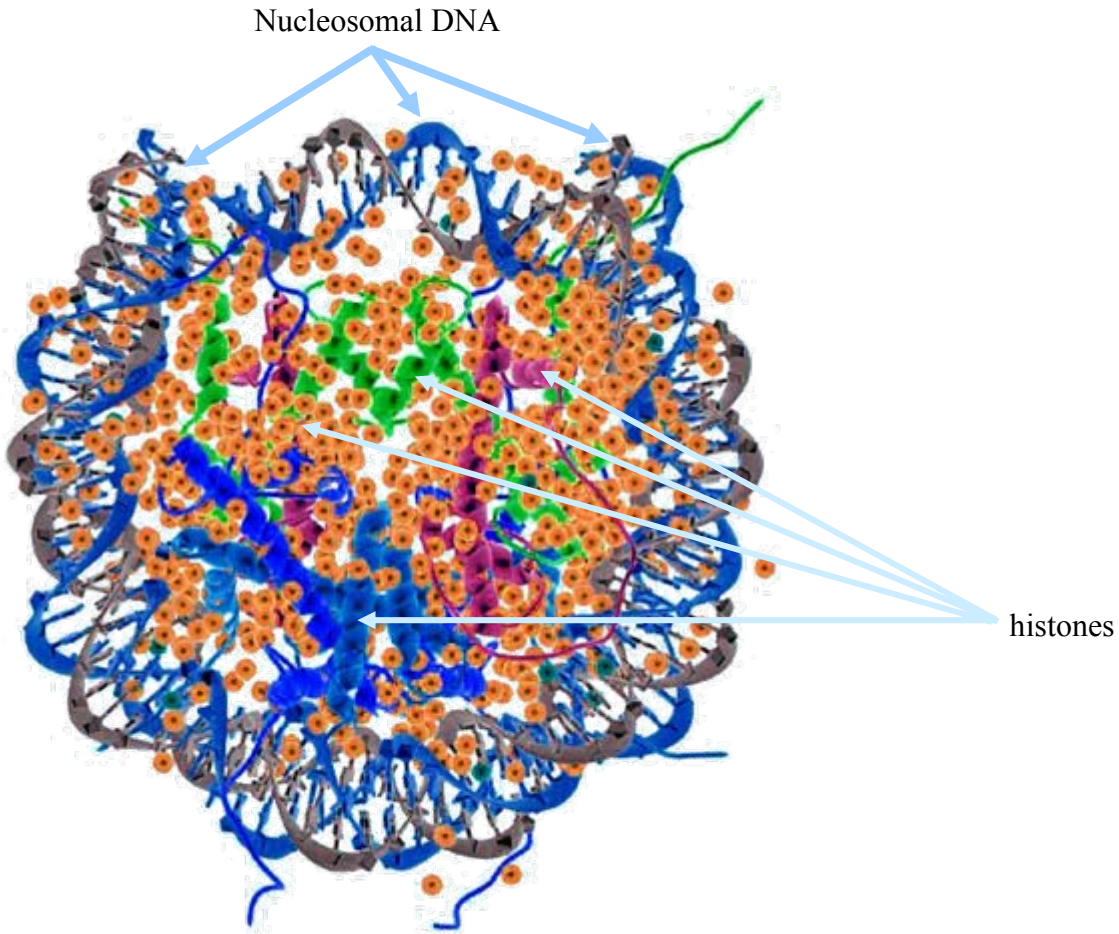
k_t – constant of translation termination for mRNA encoding protein A
Kdegr,A, kdegr,B, kdegr,AA, kdegr,AB, kdegr,BB – constants of degradation for proteins A, B, AA, AB and BB
Kdim,AA kdim,BB – constants of dimerization for proteins A and B, kdis,AA, kdis,BB – constant of dissociation into subunits for dimers AA and BB
Kdim,AB – constant of association for AB complex, kdis,AA – constant of dissociation into subunits for AB complex
fmod,A, fmod,AA, fmod,AB – external parameters of modification control for protein A as monomer, homodimer and heterodimer
Kv,B, kv,AB, kv,BB– constants of recovery of protein B in its modified form to initial A, as part of monomer, heterodimer and homodimer

ksBB – dimer BB and DNA site binding rate constant, kdBB – dimer BB and DNA site dissociation rate constant
ks,RNA-pol, kd,RNA-pol – ribosome and DMM complex association and dissociation rate constants
kini,transcr – transcription initiation rate constant

Chapter 2
COMPUTATIONAL GENOMICS

2.1. An integrated computer system for analysis of nucleosomal DNA organization

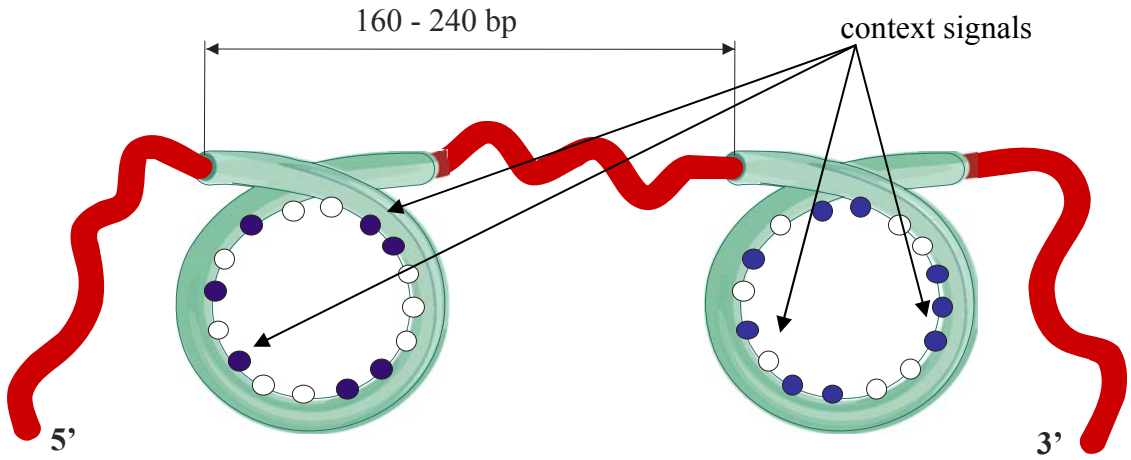
Nucleosome is the basic unit of chromatin packaging



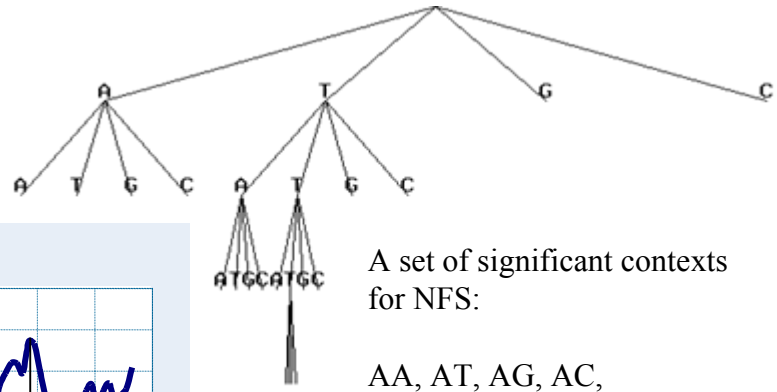
Levitskii V.G. (1999) *Dokl Akad Nauk*., V.64(2), p. 255-259

Nucleosome positioning code: common features

1. the code is extremely degenerated: differing DNA sequences can be recognized and are able to interact with histone octamer;
2. weakness of context signals of interaction with histone octamer;
3. varying disposition of positioning signals within the region of interaction with histone octamer;
4. obligate signals are absent in this code;
5. positioning of core octamer at the concrete DNA site is performed on the base of specific subset of signals located in particular set of positions (taken out of potentially large variability of such positions).

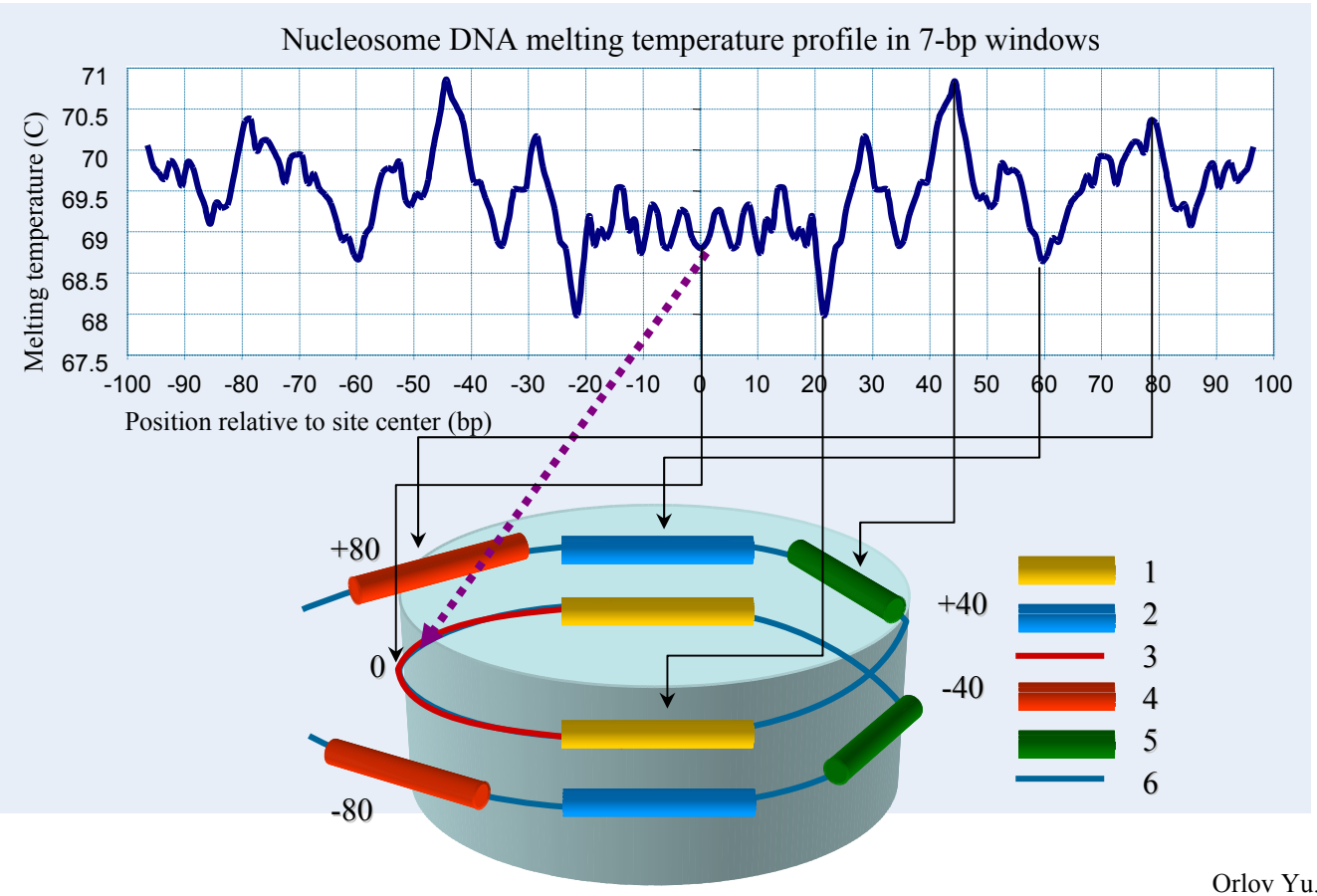


Contextual and Conformational Properties of DNA nucleosome formation sites



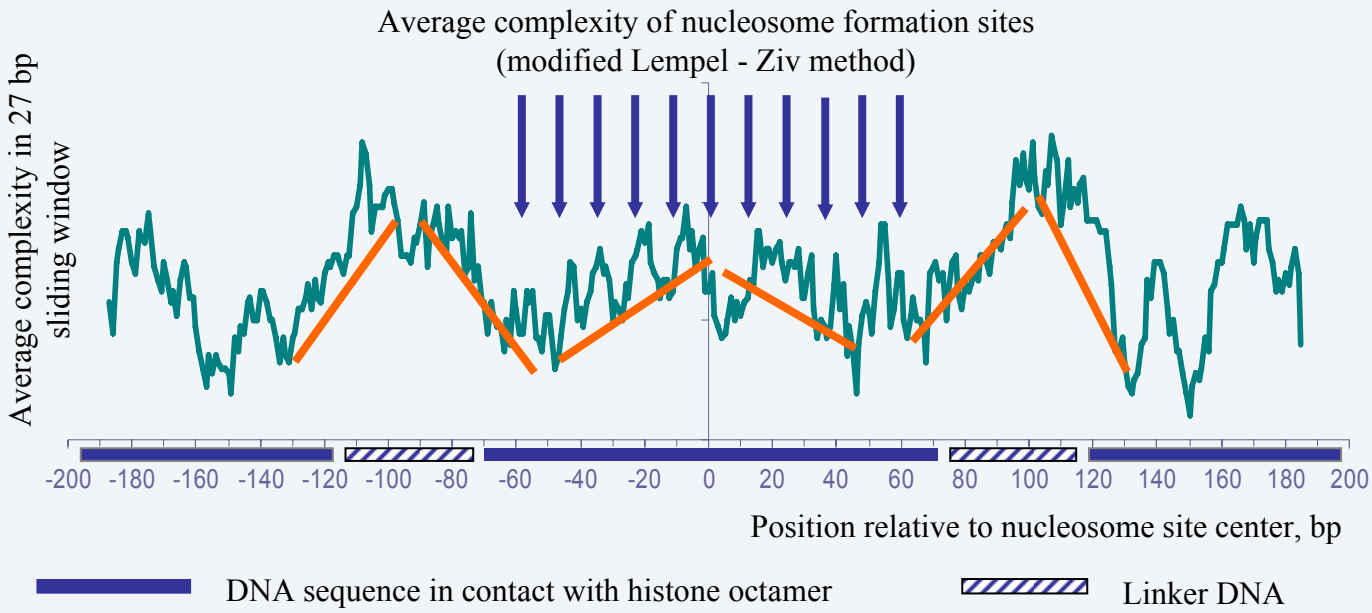
A set of significant contexts for NFS:

AA, AT, AG, AC,
GT, CT,
{A,T,G,C}AT,
{A,T,G,C}TT,
{A,T,G,C}TTT



Orlov Yu.L. et al. (2002) *In Silico Biology*, V.2(3), p. 257-262.

Complexity of nucleosome formation sites



Average complexity profile for phased DNA sequences containing nucleosome formation sites [-200;+200].
Profile trends are indicated by red lines Arrows pinpoint to periodically located (10 – 11 bp) local complexity minima

NPRD: Nucleosome Positioning Region Database

Experimental data on locations and characteristics of nucleosome formation sites

Levitsky V.G, Katokhin A.V., Podkolodnaya O.A., Furman D.P., Kolchanov N.A.
NPRD: Nucleosome Positioning Region Database. Nucl. Acids. Res., 2005, 33, 67-70.

This entry is from: NUCLEOSOME

Save | Link | Printer Friendly

NUCLEOSOME:N00093

AccessionNumber
N00093
Number
-2
Annotator
Levitsky V.G.; 10.11.03
Method
in vivo
Species
Saccharomyces cerevisiae
Taxon
Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
DNABankLink
GenBank; YSCTOPI; K03077;
Description
DNA topoisomerase I gene
TOP1 gene, MAK1 gene, YOL006C gene
Gene
gene
Region
5' region;
Function
promoter
varBounds
K03077; SR:500; -352 to -204; -339 to -178; -331 to -164;();();();
Comments
Nucleosomal core particles have been mapped at nucleotide level over the promoter region, encompassing the presumptive TATA sequence (-93) and the two RNA initiation sites (-64, -52); the most upstream nucleosome particle forms on to a 29 bp-long poly(dA-dT) element (-262 to -234). This simple organization remains constant throughout both the logarithmic and the linear phase of growth, with the exception of an increased accessibility to micrococcal nuclease of the nucleosome covering the TATA box and the RNA initiation sites during the diauxic shift (nuc -1) the switching from the fermentative to the respiratory metabolism) in parallel with an increase of the DNA topoisomerase I mRNA
Sequence
tatatatttatacataacttcaatgtcttgaacacgtaaatttttagtgtgagaaaccttt
ttcactccgggtaatacctgctgtagtcttcaaaaaaaaaaaaaaaaaaaaaaaaaaaaacc
agccatggaaagctctttatttttttacttttacggctttttttccctttttcttatatgatc
gatgcacgt
posEvidence
W303-1a (RS188) strain: MNase digestion, restriction enzymes cleavage;
keyWords

Taxonomy

Link to EMBL/GenBank

List of gene name

Gene region

Nucleosome location

Comments

Sequence

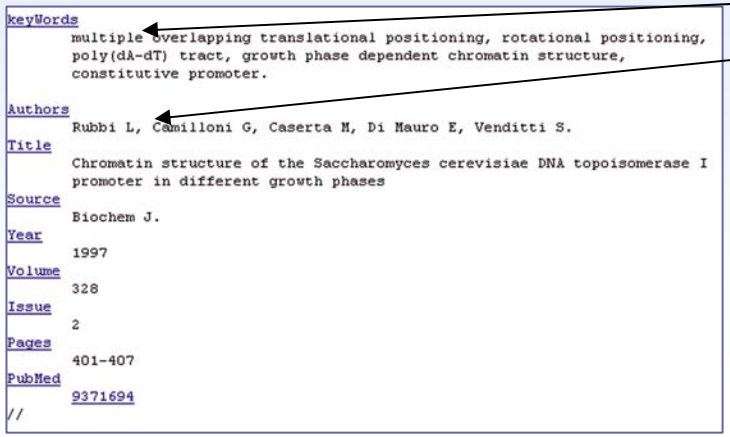
Experimental evidence of nucleosome positioning

http://srs6.bionet.nsc.ru/srs6/, start, Nucleosome databases, NUCLEOSOME

NPRD: Nucleosome Positioning Region Database

Experimental data on locations and characteristics of nucleosome formation sites

http://srs6.bionet.nsc.ru/srs6/, start, Nucleosome databases, NUCLEOSOME



Keywords

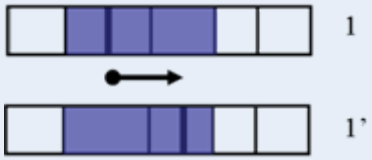
Bibliographic reference to paper: Authors, Title, Source, Year, Volume, Issue, Pages, PubMed idetifier (PMID)

When annotating nucleosome formation sites experimental mapping we pay a special attention to several characteristics: (i) the location of nucleosome relative to functional components of the genome—within genes (5'- or 3'-regions, enhancers, etc.) or outside genes (repetitive DNA: satellite, centromeric, etc.); (ii) type of gene activity related to nucleosome position, (iii) influence of nonhistone proteins, (iv) occurrence of translational or rotational nucleosome positioning, (v) characteristics of tissue types and states of cell activity, (vi) detailed characterization of experimental methods used and accuracy of determining the nucleosome position, and (vii) results of applying theoretical and computer methods to analysis of contextual and conformational DNA properties.

Recon method - nucleosome formation potential calculation

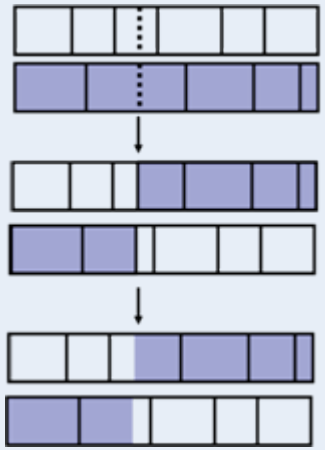
Genetic Algorithm:
Search of partition into local fragments

Mutation



coherent joining and splitting

Recombination

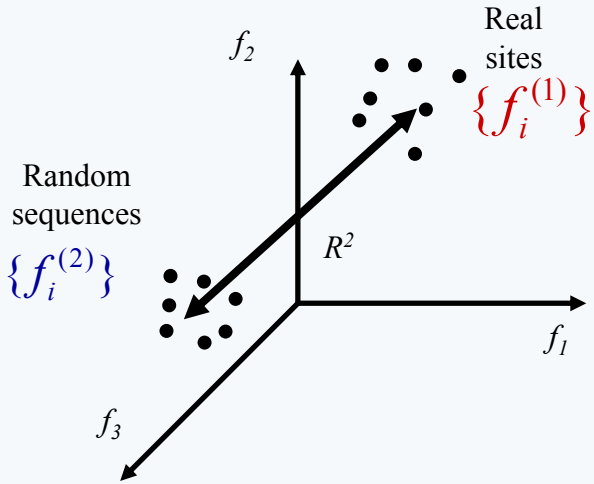


introduction of a 'break' (dotted line) into initial partitions 1 and 2

partitions after the exchange

formation of the final partitions 1' and 2'

Discriminant analysis of dinucleotide frequencies for partition fragment



The fitness of a subset of N local dinucleotide frequencies $\{f_i^{(1)}\}$ is estimated by a Mahalanobis distance R^2:

$$R^2 = \sum_{i=1}^{N}\sum_{k=1}^{N}([f_i^{(1)} - f_i^{(2)}] \times S_{i,k}^{-1} \times [f_k^{(1)} - f_k^{(2)}])$$

Here $S_{n,k}^{-1}$ is an element of the matrix S^{-1}, inverse to the matrix $S = S^{(1)} + S^{(2)}$
These matrices are the covariance matrices of the vectors of dinucleotide frequencies $\{f_i^{(2)}\}$ and $\{f_i^{(1)}\}$

WWW interface of the program Recon

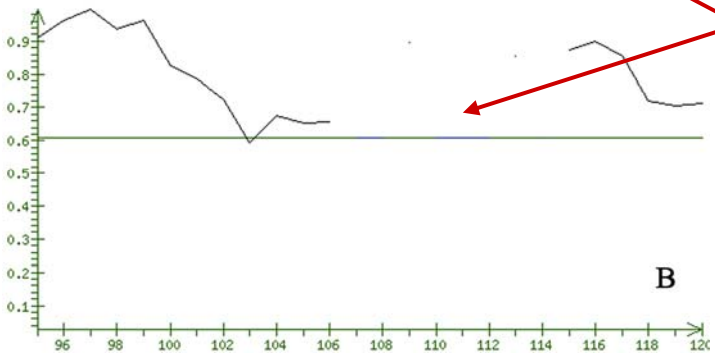
http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/

V. G. Levitsky RECON: a program for prediction of nucleosome formation potential. *Nucl. Acids. Res.*, 2004, 32, W346-W349.

Output data of the Recon program

```
105     0.674046
106     0.652641
107     0.658818
108     *
109     *
110     0.893687
111     *
112     *
113     *
114     0.854674
115     *
116     0.872270
117     0.897377
118     0.852726
```

A



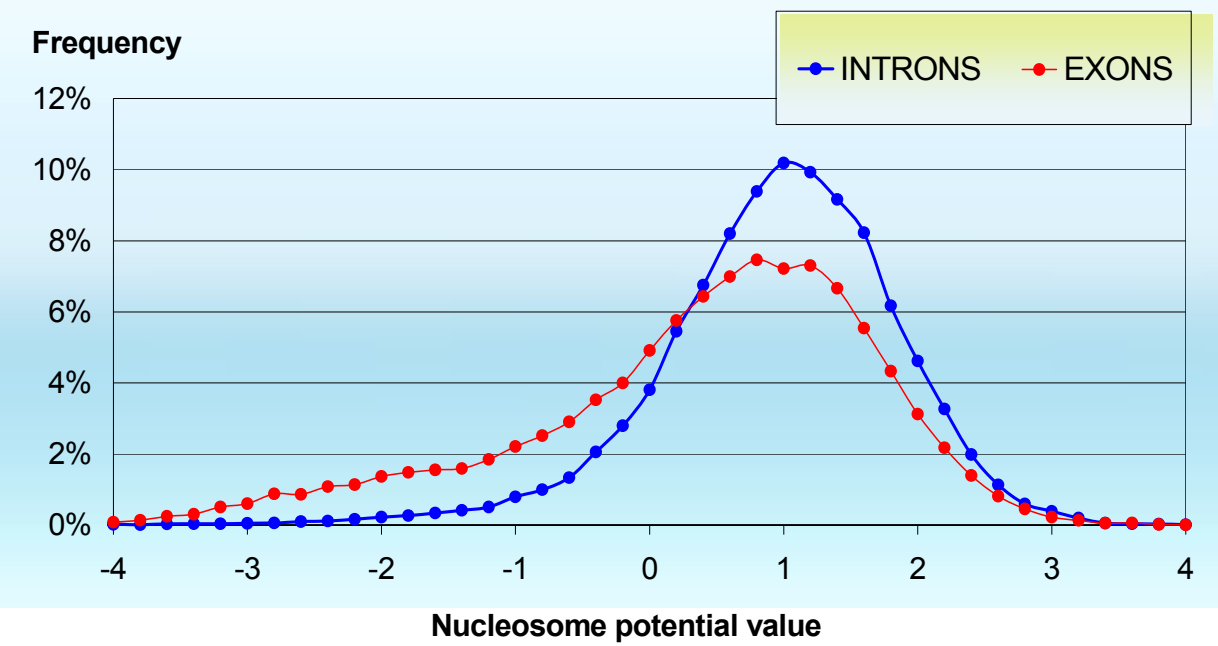
B

Text mode

Region of abnormal dinucleotide composition, for example AAAAAATTTTTT

Graphic mode

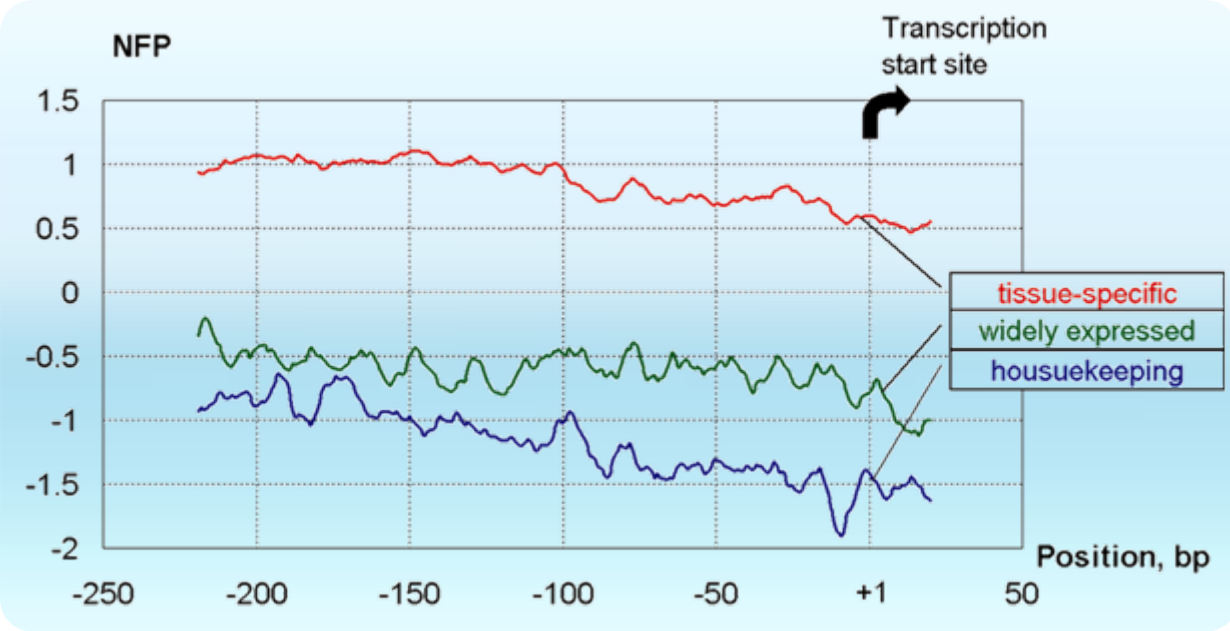
Nucleosome formation potential of exons and introns



The distribution for exons is shifted leftward relative to that of introns and has the pronounced tail at its left flank.

Levitsky,V.G., Podkolodnaya,O.A., Kolchanov,N.A., Podkolodny,N.L. Nucleosome formation potential of exons, introns, and Alu repeats. Bioinformatics, 17(11), 2001, 1062-1064.

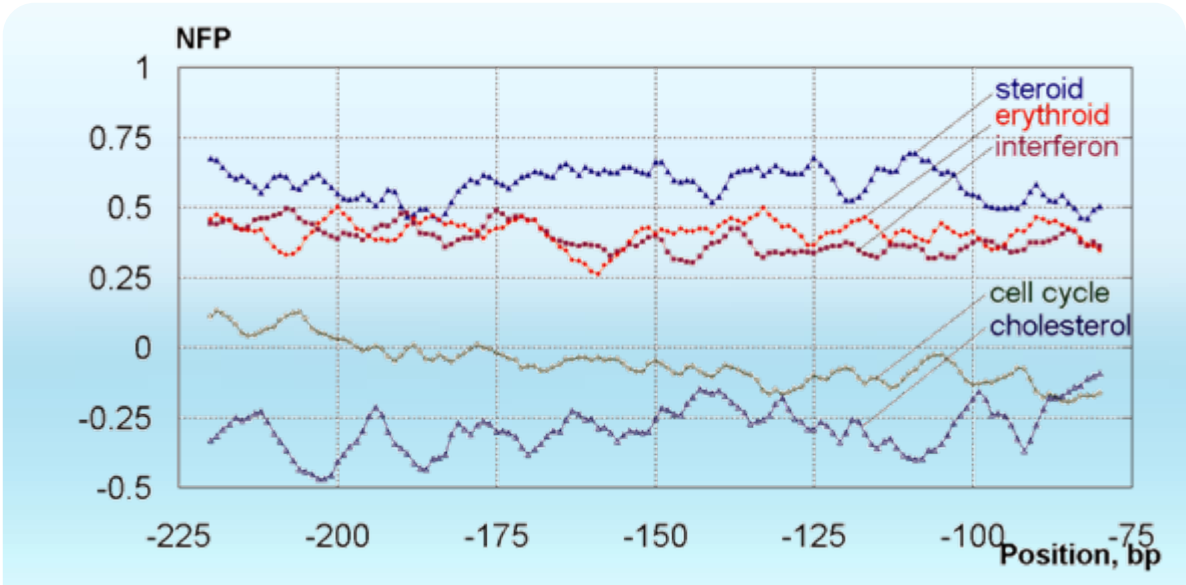
Nucleosome formation potential: promoter analysis



It was found that in promoters of tissue-specific genes, the nucleosome formation potential was essentially higher than in genes expressed in many tissues, or housekeeping genes. Hence, capability of nucleosome positioning in promoter region may serve as a factor regulating gene expression.

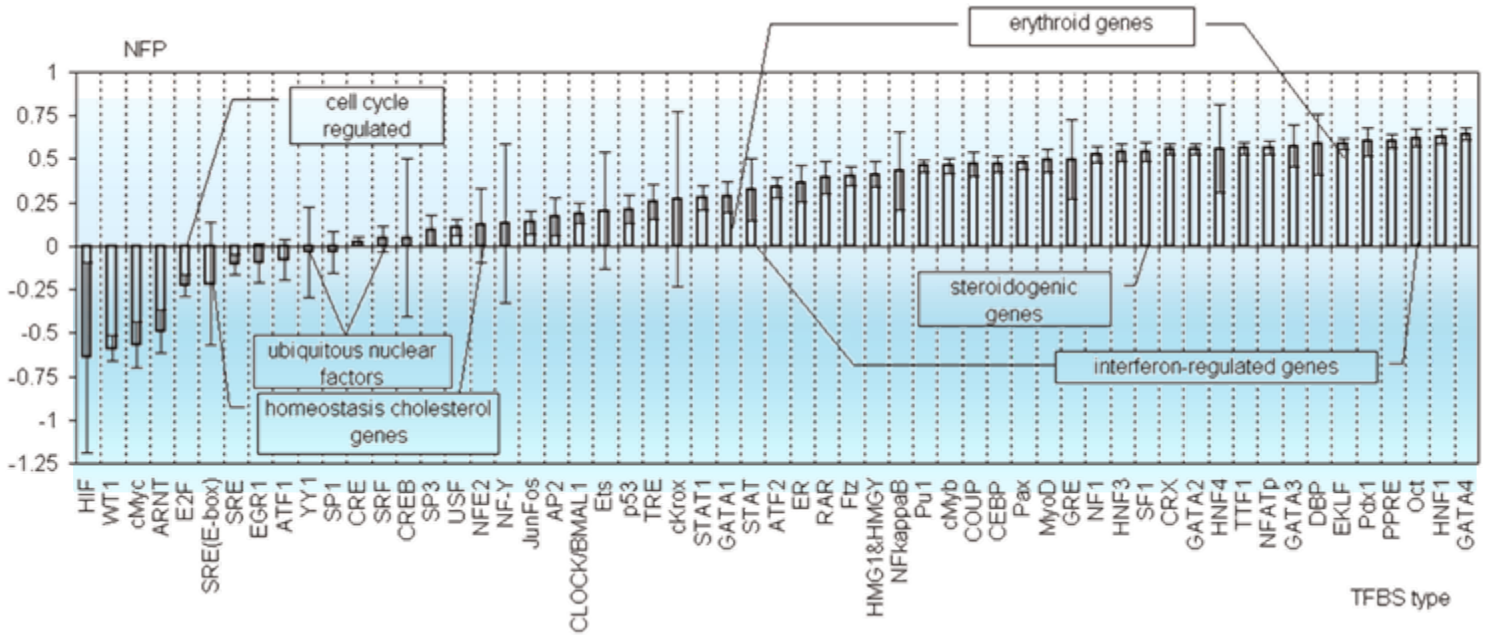
Levitsky,V.G., Podkolodnaya,O.A., Kolchanov,N.A., Podkolodny,N.L. Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis. Bioinformatics, 2001, 17(11), 998-1010.

Nucleosome formation potential: promoters of genes with different patterns of expression



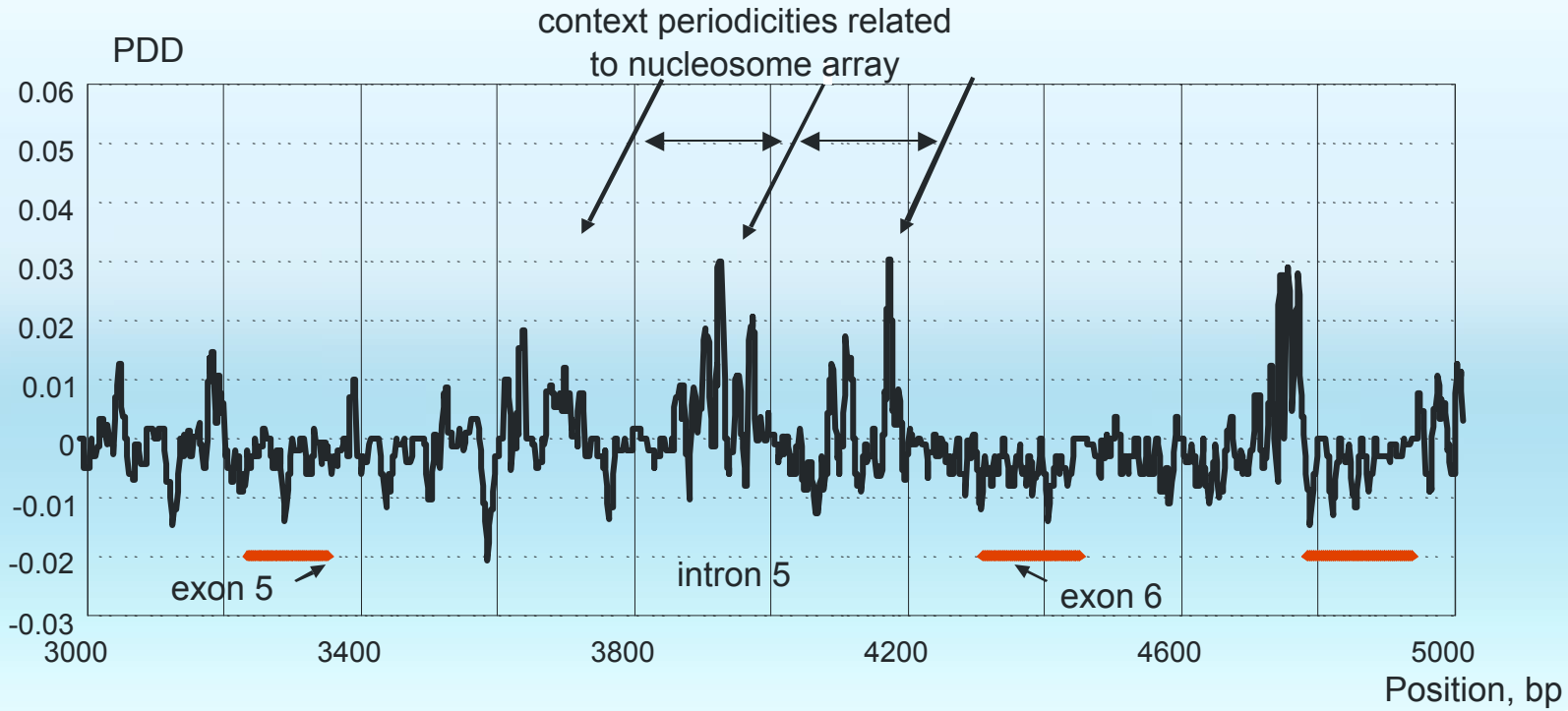
Low NFP values →	genes regulating intracellular cholesterol level, cell cycle regulated genes →	rapid transcription initiation, high expression level, possible easy access of transcription factors to the promoters
High NFP values →	erythroid specific regulated genes, genes controlling biosynthesis of steroid hormones, interferon regulated genes →	tissue-specific genes, inducible genes access of transcription factors to the promoters is restricted

Average nucleosome formation potential values and respective confidence intervals for the sample of DNA fragments containing TFBS (*Transcription Factor Binding Sites)* at the central position



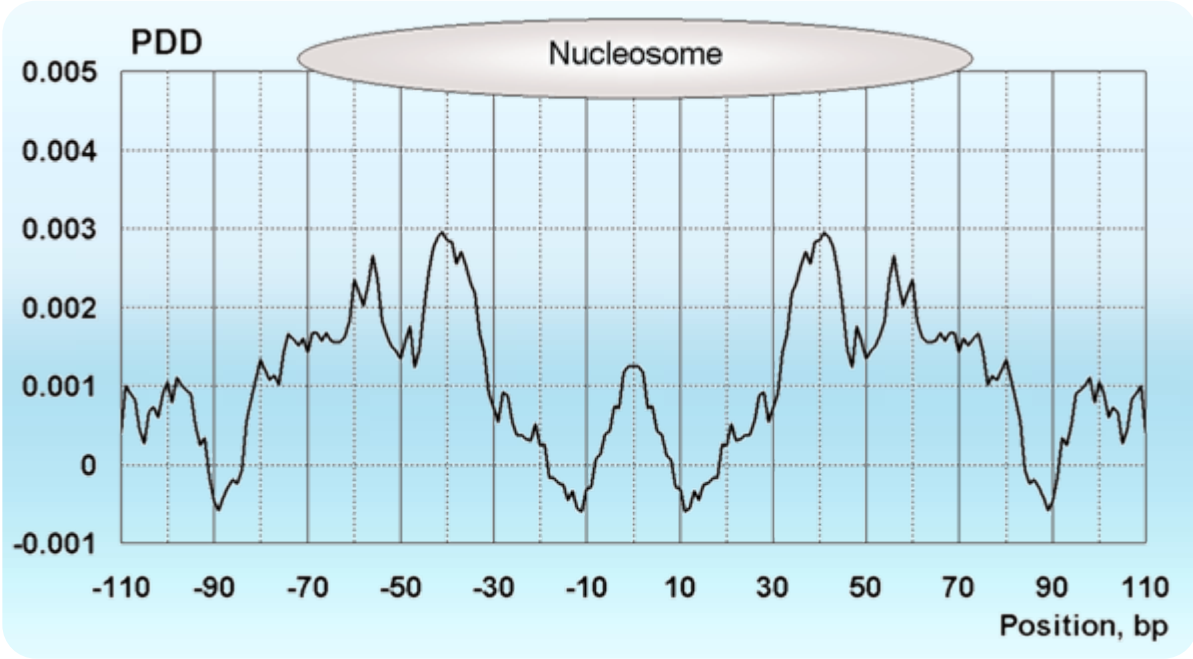
The distribution is consistent the reference of these site to a particular expression patterns of genes. For example, the lowest NFP values were for the TFBS occurring in the cell cycle regulated genes (-0.23 E2F, -0.57 *cMyc*), the genes regulating lipid metabolism (-0.10 SRE), and also ubiquitous TFBS (-0.04 YY1, -0.03 SP1). The highest NFP values were for the TFBS most frequently occurring in the tissue-specific and inducible genes (0.62 Oct, 0.54 SF1).

Periodic dinucleotide density (PDD) for the chicken ovalbumin gene



Discrete alignment signals appear to be present in the gene in two of the large introns.Strong nucleosome alignment signal exists in intron 5.
Lauderdale J.D., Stein A. Introns of the chicken ovalbumin gene promote nucleosome alignment in vitro. *Nucleic Acids Res.,* 1992, 20, 6589-6596
This signal is defined by a 200 bp periodicity of phased AA and TT dinucleotide pairs

Periodic dinucleotide density (PDD) profile for the sample of sequences containing nucleosome formation sites in the center position



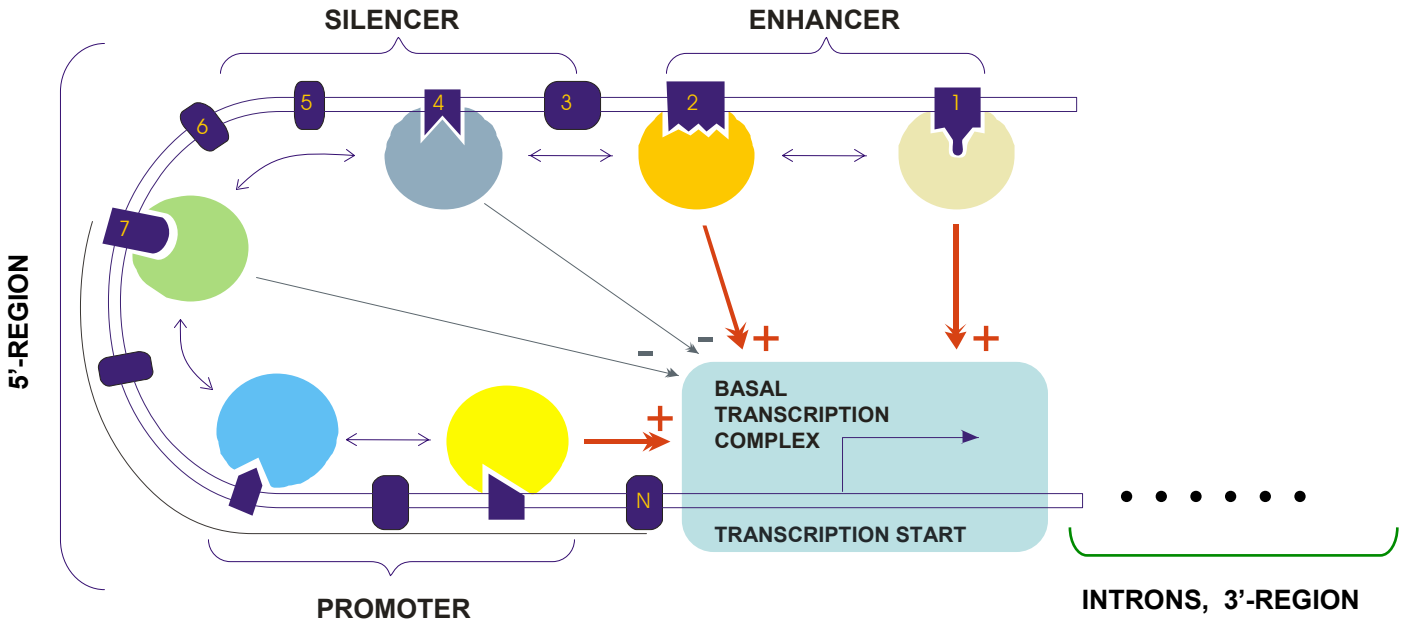
The presence of a phased dinucleotide at a distance of one or two DNA helix turns determines an increase in the ability of DNA to bend regularly.

Nucleosome DNA contains two more pronouncedly bent regions with a length of 40–50 bp at positions [–73; –25] and [+25; +73] relative to NFS center. These two regions are separated with a less bent region of 30–50 bp located at [–25; +25]

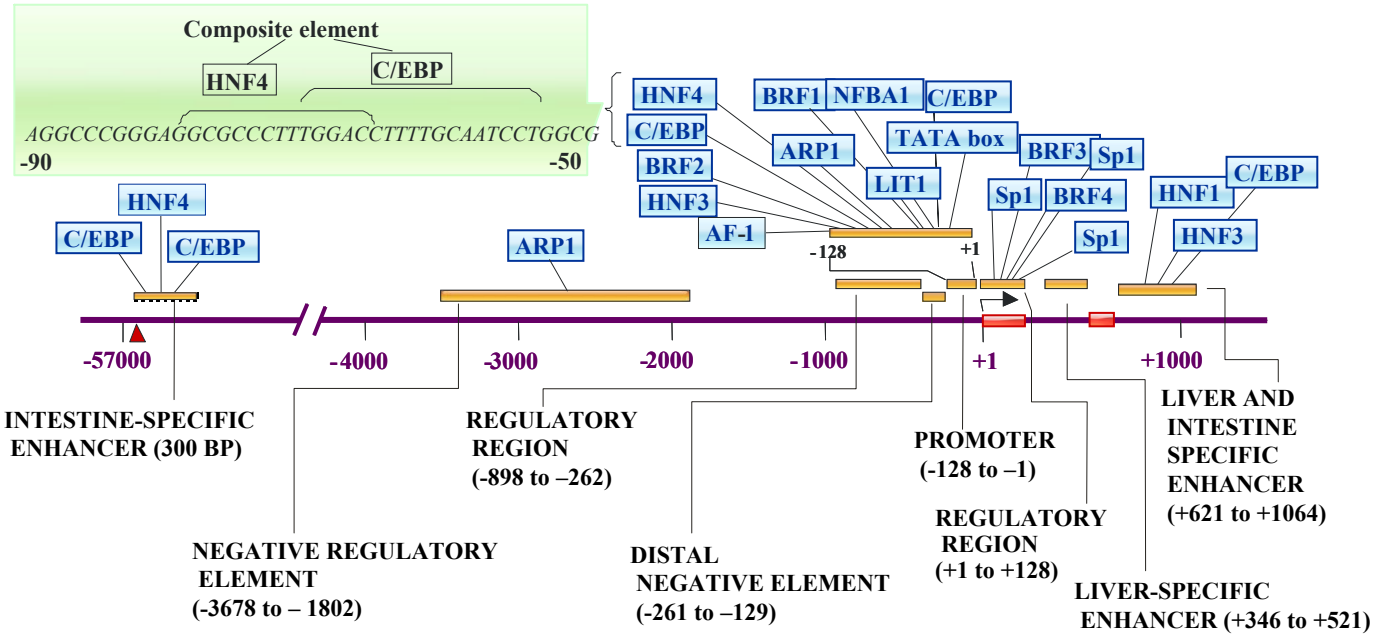
2.2. Transcription regulatory regions database (TRRD): its status in 2005

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

General model of eukaryotic gene transcription regulation



Organization of the regulatory regions controlling transcription of the gene for human apolipoprotein B.



8 regulatory units
23 transcription factor binding sites

TRRD accumulates:

1) THE DATA ON STRUCTURAL ORGANIZATION AND FUNCTIONAL CHARACTERISTICS OF:

TRANSCRIPTION FACTOR BINDING SITES

REGULATORY UNITS, (PROMOTERS, ENHANCERS, AND SILENCERS)

TRANSCRIPTION INITIATION STARTS OF GENES

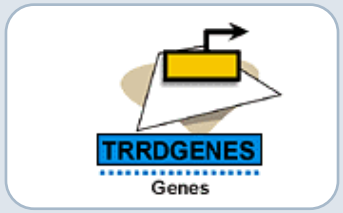
TRRSTARTS

LOCUS CONTROL REGIONS

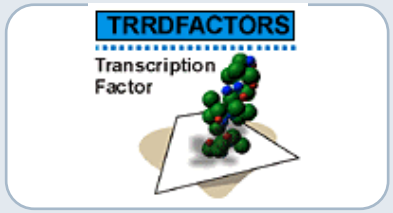
TRRDLCR

2) THE DATA ON PATTERNS OF GENE EXPRESSION

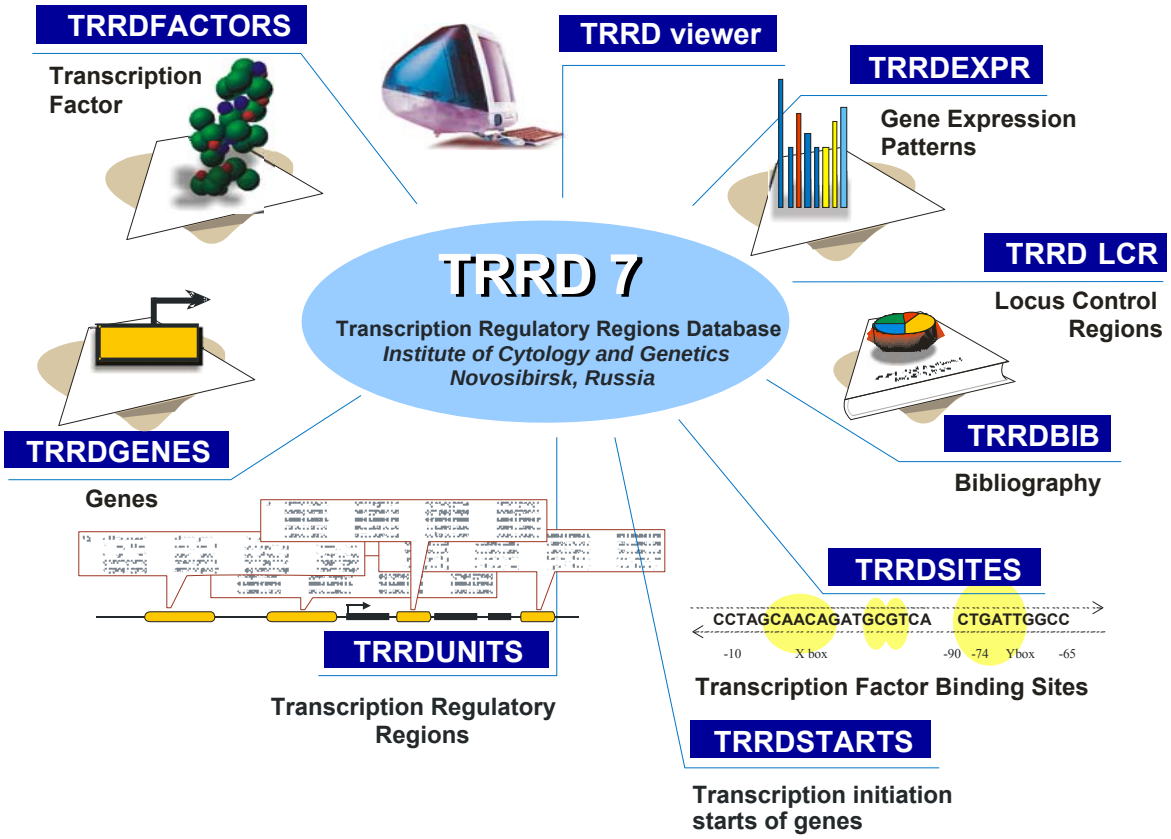
4) GENERAL INFORMATION ON GENES TOGETHER WITH HIERARCHICALLY ORGANIZED PRESENTATION OF ALL THE REGULATORY ELEMENTS



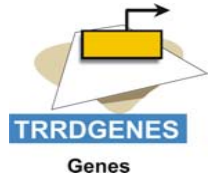
3) THE DATA ON TRANSCRIPTION FACTORS



TRRD is an informational resource comprising a family of databases



The table TRRDGENES: general description of the gene



GeneID Hs:AAP
Links: Binding sites Regulatory units Transcription factors
Gene expression regulation Bibliography
GeneAC A00596
Species human, Homo sapiens
GeneName_Brief AAP
GeneName_Full Alzheimer's disease amyloid A4 precursor protein
GeneSynonym amyloid beta protein precursor gene, amyloid precursor protein gene, APP, PAD
DNABankLink EMBL; HSPADP; X12751
EMBL; HSAPPB01; M24546
DataBankLink SWISS-PROT; A4_HUMAN; P05067(Expasy server)
CleanEx; HGNC:620; APP Ensembl; ENSG00000142192;
GenAtlas; APP; GeneCards; APP; GeneLynx; APP;
HGNC; HGNC:620; APP HOVERGEN; P05067; MIM; 104760;
SOURCE; APP; Hs EntrezGene; APP; 351 GDB; GDB:119692; APP
KeyWords heat shock-induced, adhesion protein, TATA-less promoter,
pathogenesis-related protein, multiple transcription initiation sites
Chromosome 21; 21q21.3
RegRegion 5'region
RegUnitAC REGULATORY UNIT: P00760
RegUnit promoter; ST; ; S2907, S3931, S3932, S3933, S3934, S3935
PromotTisSp 0
ExperimentCodes HeLa: 6.1.1, 6.8 [Quitschke W.W. et al., 1996]
//

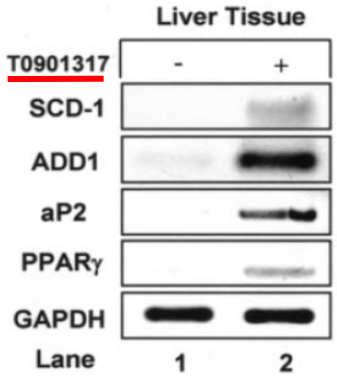
The table TRRDEXP: expression patterns of the gene

TRRDEXPR Gene Expression Patterns

ExpressionPatternAC A01381.008
GeneID Mm:PPARG
ExpressionDetectionDevice mRNA
Organ epididymis
Tissue white fat
ExpressionLevel present
RegUnitLink P01981
Reference [Seo J.B. et al., 2004]

ExpressionPatternAC A01381.018
GeneID Mm:PPARG
ExpressionDetectionDevice mRNA
Organ liver
ExpressionLevel present
IndReprName T0901317
Influence induction
Reference [Seo J.B. et al., 2004]

Northern blot analysis

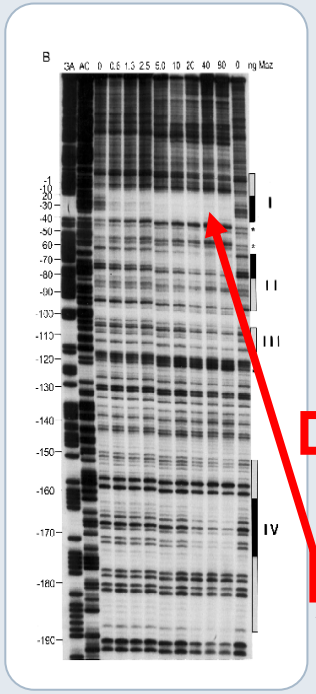


RT-PCR analysis

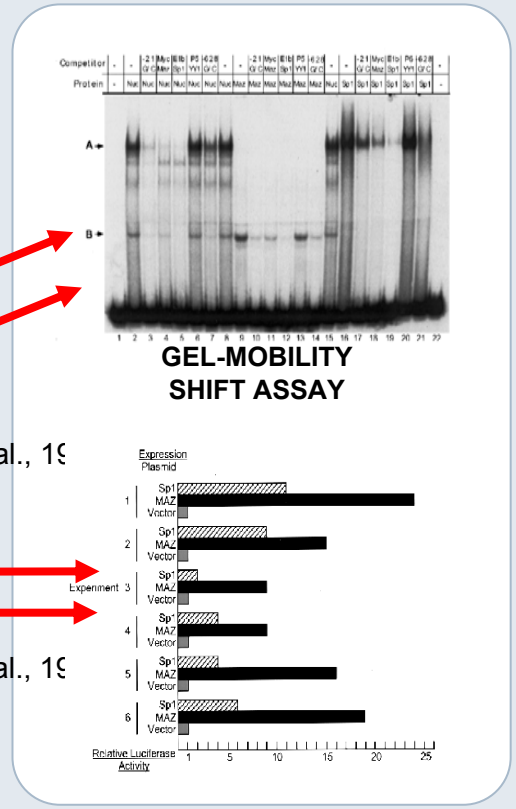
The table TRRDSITES: description of the HNF-4 binding site in the human ApoB gene

```
-88 5′ cccgggaggCGCCCTTTGGACCTtttg 3′ -62
   3′ gggccctccGCGGGAAACCAGGAaaac 5′
```

SiteAC S1160
GeneID Hs:APOB
RegUnitAC P00670
SiteName HNF-4;
PreferredName HNF-4
SiteNameSynonym AF-1 binding site
SiteNameSynonym BA1 binding site
SiteIndex 1
FactorName HNF-4; hepatic nuclear factor 4
FactorInfluence increase
Sequence cccgggaggCGCCCTTTGGACCTtttg
SequencePosition -88 to -62
FootprintSequencePosition -82 to -62
DNA_BankLink M15053: 68
ImportantPos
---------CG--CTTTGGACCT--t-; HNF4 [Metzger S. et al., 19
ExperimentCodes
3.5 (HNF4) [Metzger S. et al., 1993]
rat liver cells: 3.6 (HNF4) [Metzger S. et al., 1993]
human HepG2 cells: 6.2 [Metzger S. et al., 1993]
HeLa cells: 6.2, 6.6 (HNF4) [Metzger S. et al., 1993]
1.1.5 (HNF4), 3.3, 3.5 (HNF4), 4.2 (HNF4) [Ladias J.A. et al., 19
//



DNASE I FOOTPRINTING



TRANSIENT EXPRESSION ANALYSIS

Examples of assays providing the information about transcription factor binding sites inputted into TRRD

Type of experiment	Assay code in TRRD
Detection of transcription factor binding sites	
DNase I footprinting with nuclear extract	1.1.1
DNase I footprinting with purified or recombinant protein	1.1.5
Genomic footprinting	1.5
Methylation protection assay	4.1
Methylation interference assay	4.2
Electrophoretic mobility shift assay (EMSA) with nuclear extract	3.1
EMSA performed in the presence of competitive oligonucleotides	3.2
EMSA performed with mutant probes or competitors	3.3
Identification of DNA-binding proteins	
DNase I footprinting with purified or recombinant protein	1.1.5
DNase I footprinting with nuclear extract and specific antibodies	1.1.6
EMSA with purified or recombinant protein	3.5
EMSA with nuclear extract and specific antibodies	3.6
Confirming the functional importance of the site	
Insertion of isolated site 5' of homologous or heterologous promoter	6.3.2
Comprehensive mutant analysis	6.2
Trans-activation of a reporter gene by overexpression of a distinct transcription factor	6.6
Genomic footprinting	1.5

The complete list of experimental assays providing the data on transcription factor binding sites inputted in TRRD is available at http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/digcodes.shtml

The table TRRDSITES

TRRDSITES4:S5743
SiteAC
S5743
GeneID
Gene: Hs:APOA1
RegUnitAC
REGULATORY UNIT: P00051
SiteName
T3R/RXR bs; T3R/RXR alpha binding site
SiteIndex
2
FactorName
T3R beta/RXR alpha;
FactorInfluence
decrease
Sequence
ACTGAACCCTTGACCCCTGCCCT
SequencePosition
-214 to -192
DNA_BankLink
J04066:1858, M20656:2264, J00098:257
ImportantPos
--TG--CC-TTGACCC-------; T3R beta/RXR alpha; [Tzameli I. and Zannis V.I., 1996]
SeqContradiction
ACTGAACCCTTGACCCCTGCAAA
PosContradiction
-218 to -196
ExperimentCodes
3.5 (RXR alpha), 3.5 (T3R beta/RXR alpha) [Tzameli I. and Zannis V.I., 1996]

COS-1 cells: 3.1, 3.2.2, 3.3, 3.6 (T3R beta), 4.2 [Tzameli I. and Zannis V.I., 1996]

HepG2 cells: 6.5 (T3R beta), 6.5 (9-cis RA), 6.5 (all-trans-RA), 6.6.1.1 (RXR alpha/T3R beta), 6.6.1.1 (T3R beta), 6.6.1.1 (RXR alpha) [Tzameli I. and Zannis V.I., 1996]

1. Important Positions

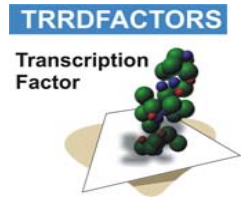
- Methylation interference assay
- Electrophoretic mobility shift assay performed with wild type and mutant probes or competitors
- Effect of mutations in binding sites on gene promoter activity in the transient transfection assay

2. Sequence Contradiction

3. Positions Contradiction

Discrepancies in the site sequence or its positions between the paper annotated and the corresponding data from embl/genbank

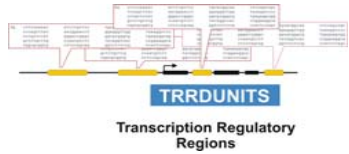
Table TRRDFACTORS



Description of transcription factor

Identifier F5743.1
GeneID Hs:APOA1
SiteAC Site: S5743
FactorName T3R beta/RXR alpha; T3R beta / retinoic X receptor alpha heterodimer
FactorSubunitName T3R beta;
FactorSource recombinant
Cells COS-1
Reference [Tzameli I. and Zannis V.I., 1996]
FactorSubunitName RXR alpha; retinoic X receptor alpha
FactorSource recombinant
Cells COS-1
Reference [Tzameli I. and Zannis V.I., 1996]

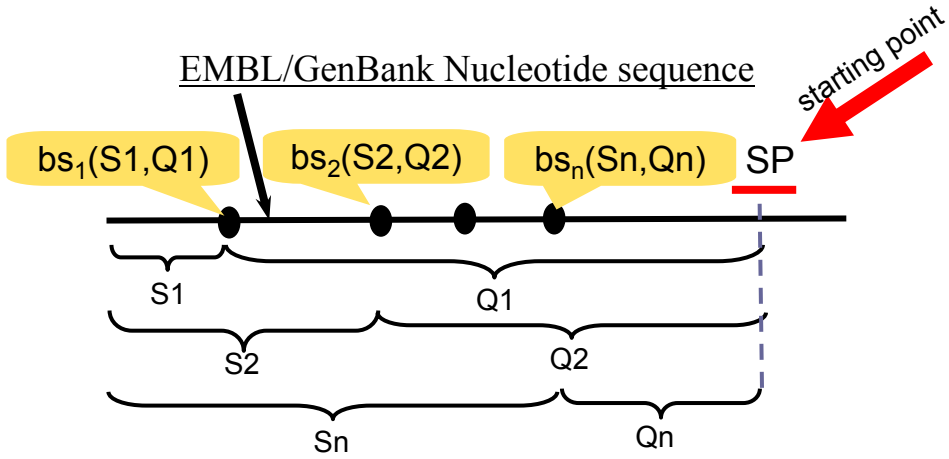
TRRDUNITS: description of transcription regulatory units (promoters, enhancers, silencers)



RegUnitAC	P00562
GeneID	Rn:D2
RegRegion	5'region
RegUnit	Promoter; ST; -150 to +1; S79, S80
DNA BankLink	EMBL; RND2RPR; X77137; 704 to 855
LeftTrunc	0
RightTrunc	0
SeqLength	152
Sequence	cccaggcccc acagtgcaga gatagttctg gggccctggg tgggtggggc ctctgtacaa ggggcggggt tcccgggcgc ctcgtggcca gggtgacccc gccccctcct cctgcgcagc gctctgattc cgcggagctg tccagcctca gt
PromotTisSp	0
PromotInd	1
ExperimentCodes	6.1.1, 6.8 [Minowa T. et al., 1992]

Extraction of regulatory units DNA sequences from EMBL/GENBANK

Step 1: determining the position of starting point (SP) in EMBL/GenBank entry



Step 2: localization and extraction of nucleotide sequence corresponding to the regulatory unit

EMBL/GenBank :

SQ

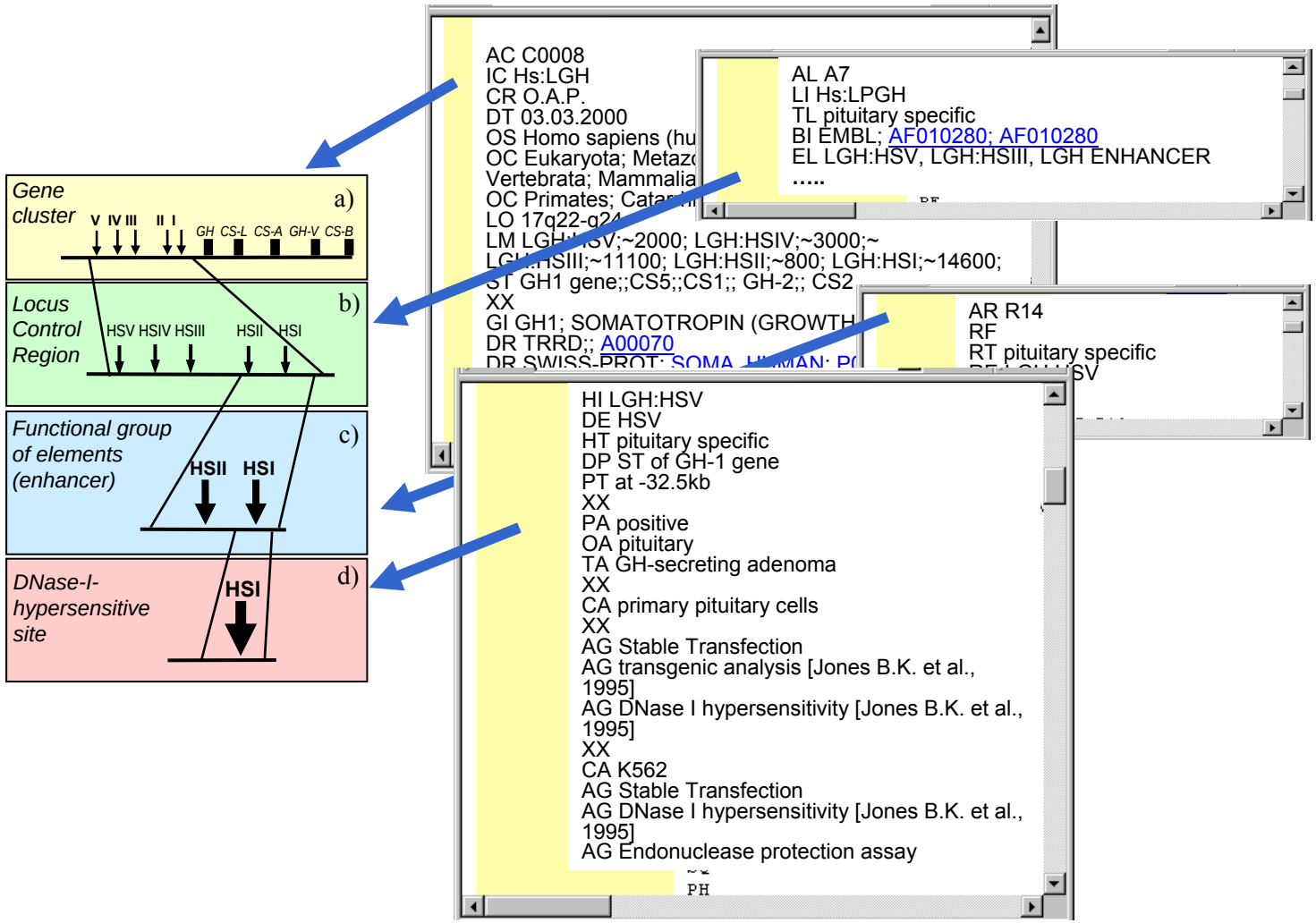
```
aagcttatcgatgataagcggtca
aacatgagaattcgcgatagggat
cacctcgatccagccttatctagg
ccctcccccctgtcaaacaccctt
gtcctttgttaccagaacaggcca
cctttgcacctgctgttccctcct
ccaaggggggtgggggttatccttc
cagagagttcctgcctacttcagg
aaatagctaaaccctctcccatgg
cctgtttcagtcggccgcagtggg
ggtacatatttgaccctcactcag
gagactggaaatcaga
```

TRRDUNITS :

"Sequence"

```
ggtcaaacatgagaattcgc
gatagggatcacctcgatcc
agccttatctaggcccctcc
ccctgtcaaacacccttgtc
ctttgttaccagaacaggcc
acctttgcacctgctgttcc
ctcctccaaggggggtggggg
ttatc
```

TRRDLCR: locus control region description



AN EXAMPLE : human IL-8

Human IL-8 It is one of the hottest genes of molecular biology investigations. Data on it's expression regulation is essential for solving biomedical problems.

IL-8 is a member of chemokine gene family, plays an important role during physiologic cell chemotaxis as well as during pathologic immune system responses. IL-8 is the proinflammatory cytokine secreted by a variety of cell types, including T cells, and macrophage-foam cells of atherosclerotic lesions. IL-8 is also known to be an autocrine growth factor. It plays mitogenic and morphogenic activity and regulates angiogenesis in tumors.

IL-8 production is rapidly induced by a very wide range of stimuli encompassing proinflammatory cytokines such as tumor necrosis factor (TNF) or IL-1, bacterial o viral products and cellular stress. Remarkably, some stimuli, such as IL1 or TNF, up-regulate IL-8 by more than 100-fold.

AN EXAMPLE : Human IL-8

TRRD entry A00038 - Hs:IL8

365 EXPRESSION PATTERNS:

IL8 is expressed in
12 cell types
84 cell lines

IL8 is regulated
by ~ 130
inductros
or repressors

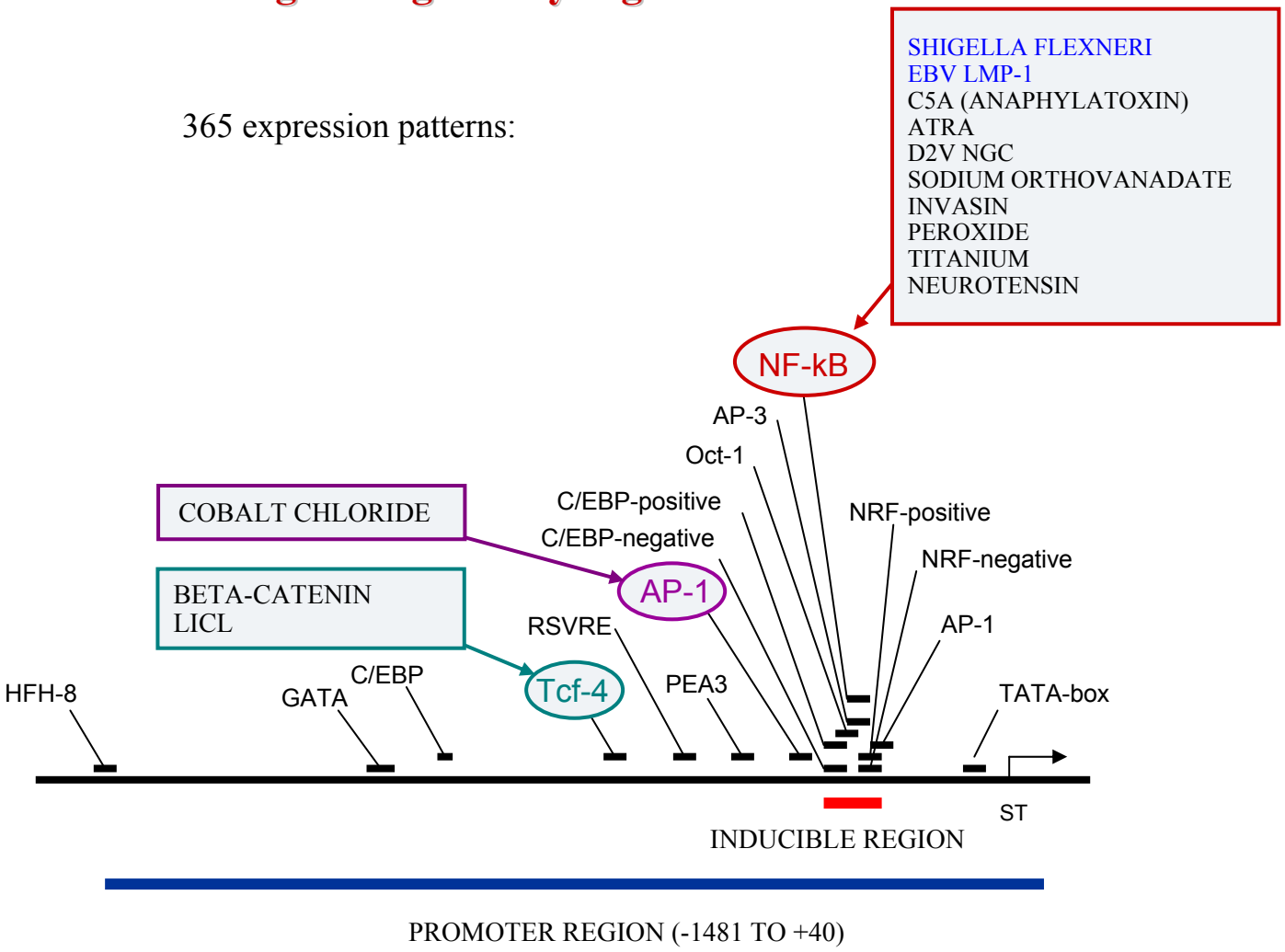
IL8 promoter region includes
16 transcription factor binding sites

PGA1, PGA2, PGD2,PGJ2, 3-O-C12-HSL, ATRA, Ad7, BSO, anaphylatoxin, CMV IE1, Cryptosporidium parvum, D2V NGC, DMSO, DTT, EBV LMP-1, EGCG, E. coli, FVIIa, H2O2, HCV NS5A, HIV, HPV, H. pylori, IFN-alpha, IFN-beta, IFN-gamma, IL-1, IL-15, IL-2, L-NAME, LPS, LiCl, MEKK1, MG132, MKK73E, Micrococcus luteus, NIK, Ox-PAPC, PD98059,PD98059, PD98059, PDTC, PDTC, PGPC, PHA, PMA, PNR, POVPC, PTX, Pseudomonas aeruginosa, R-59949, RSV, SIN-1, SN50, SNAP, SNOG, Shigella flexneri, TNF-alpha, Tcf4, U0126, U0126, UTI, actinomycin D, alpha-toxin, aminoguanidine, anoxia, beta-catenin, beta-mercaptoethanol cadmium chloride, calcium ionophore A 23187, ceramide, ciglitazone, clarithromycin, cobalt chloride, curcumin, cycloheximide, dexamethasone, erythromycin, fibrin, forskolin, geldanamycin, hepatitis B virus X protein, hypoxia, invasin, isohelenin, lactacystin, mAb Apo-1, neuropeptide substance P, neurotensin, nickel subsulfide, nicotine, okadaic acid, acidosis, pRB, paclitaxel, parthenolide, peroxide, peroxynitrite, peroxisome proliferator Wy-14.643, proteasome inhibitor Nleu, respiratory syncytial virus, rotavirus, sPGN, sodium arsenite, sodium butyrate, sodium orthovanadate, t-BOH, titanium, trichostatin A, troglitazone, vanadyl sulfate, Wortmannin, SB 203580, rebamipide

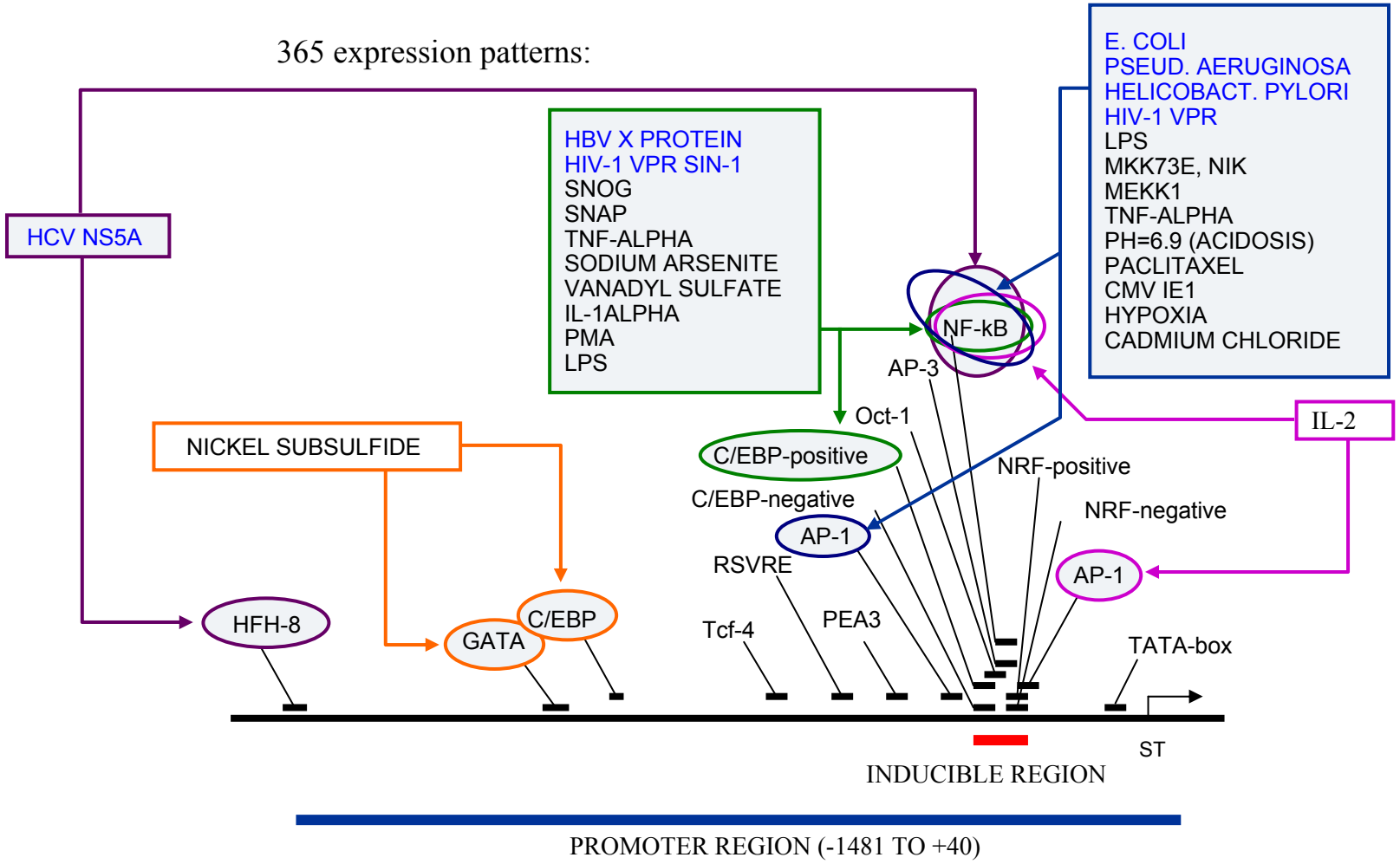
The total number of annotated articles - 152

The scheme of IL-8 gene regulatory regions

365 expression patterns:

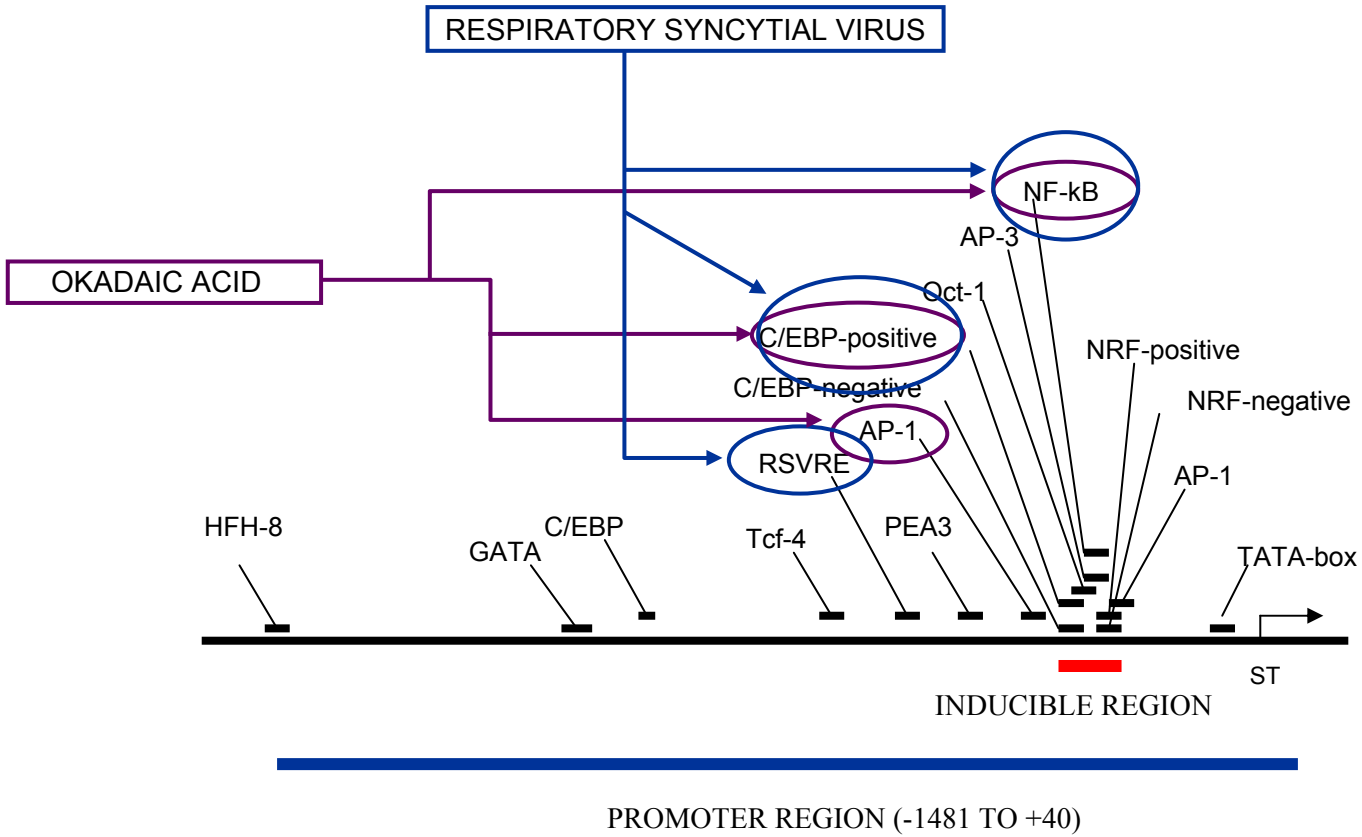


The scheme of IL-8 gene regulatory regions

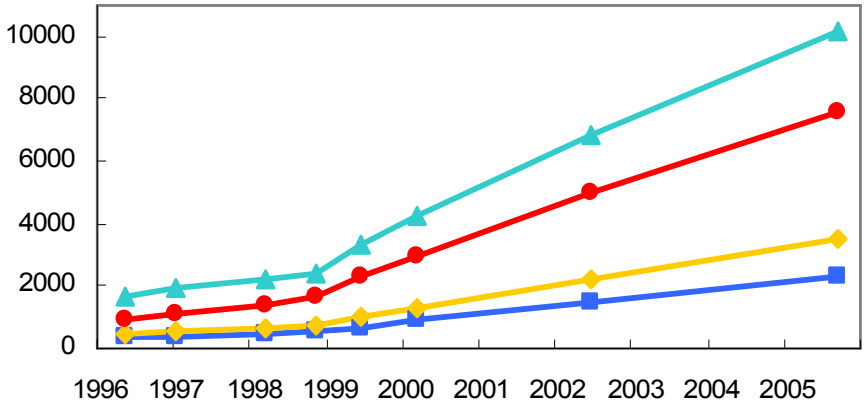


The scheme of IL-8 gene regulatory regions

365 expression patterns:



The TRRD progress since 1996 till 2005



10135 TRANSCRIPTION FACTOR BINDING SITES

7609 SCIENTIFIC PUBLICATIONS

3400 REGULATORY UNITS

2344 GENES

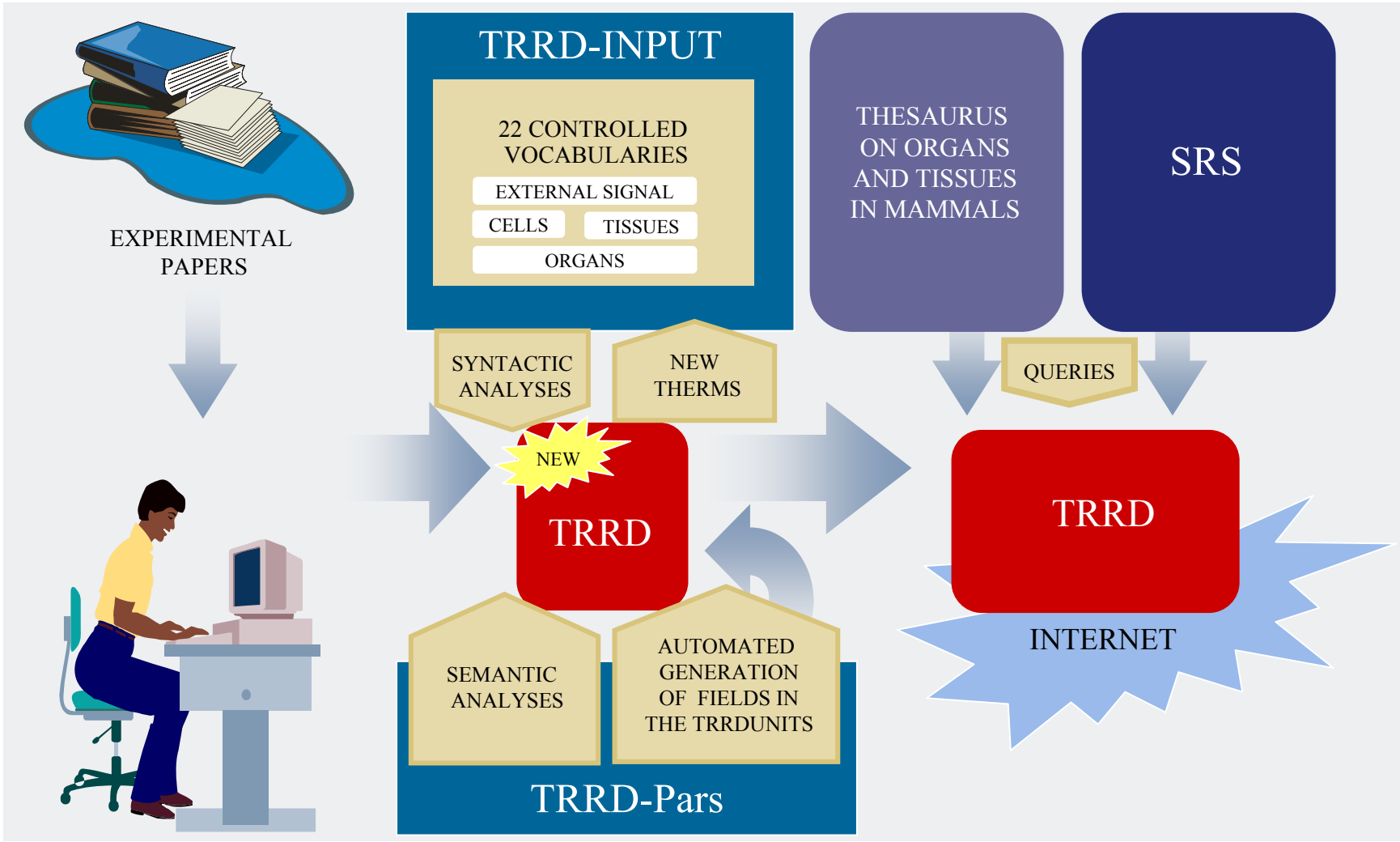
•TRRD is one of the largest among the world information resources on the structure–function organization of regulatory regions in eukaryotic genes that is filled in by manual annotation.

•The total number of genome sequences (binding sites and regulatory units) accumulated in TRRD is more then 13 000 !!!

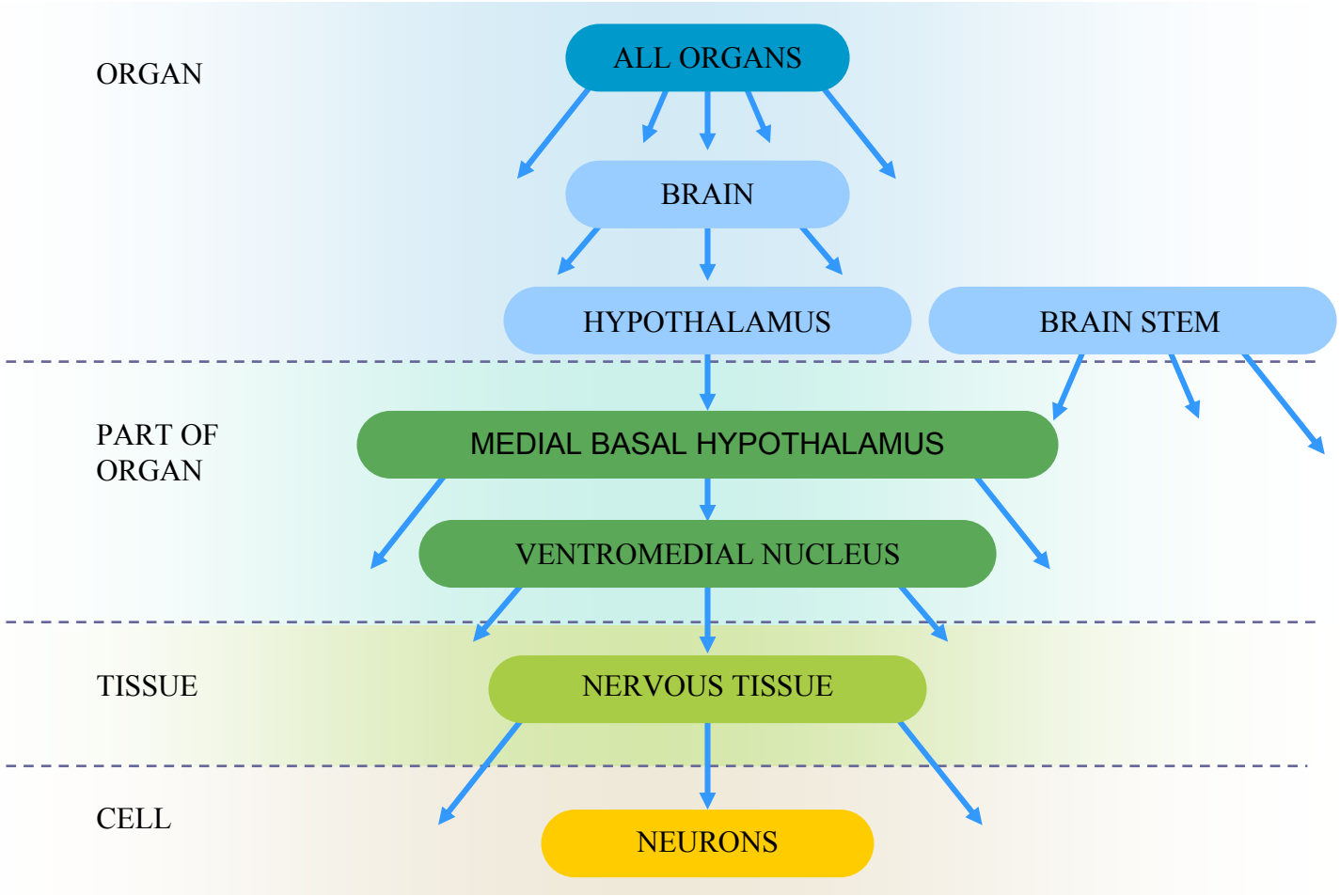
	Totally in TRRD.	Of the below species (%)			
		Human	Mouse	Rat	Other species
Genes	2344	32%	22%	15%	31%
Regulatory units	3400	36%	19%	14%	31%
Transcription factor binding sites	10 135	36%	18%	14%	32%

The largest amount of information accumulated in TRRD pertains to human, mouse, and rat genes

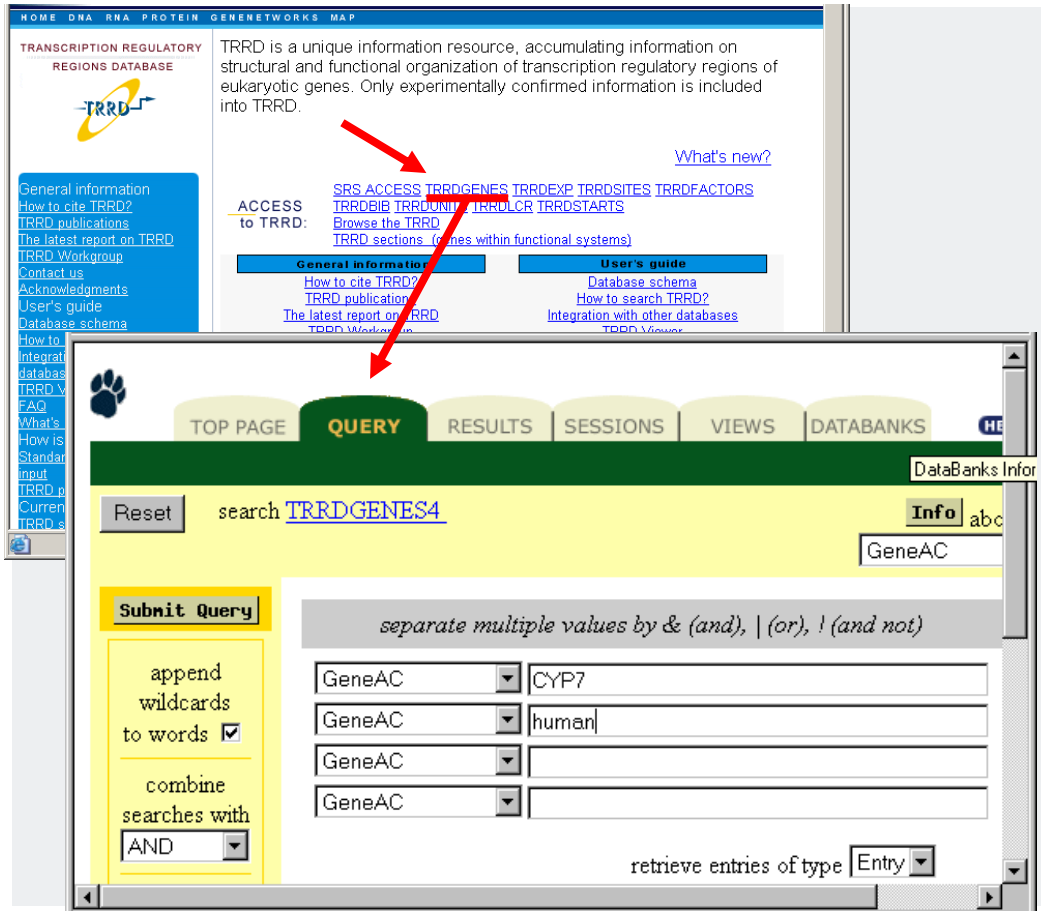
TRRD: data input, standardization and processing



Hierarchical organization of controlled vocabularies of morphological terms in the TRRD database

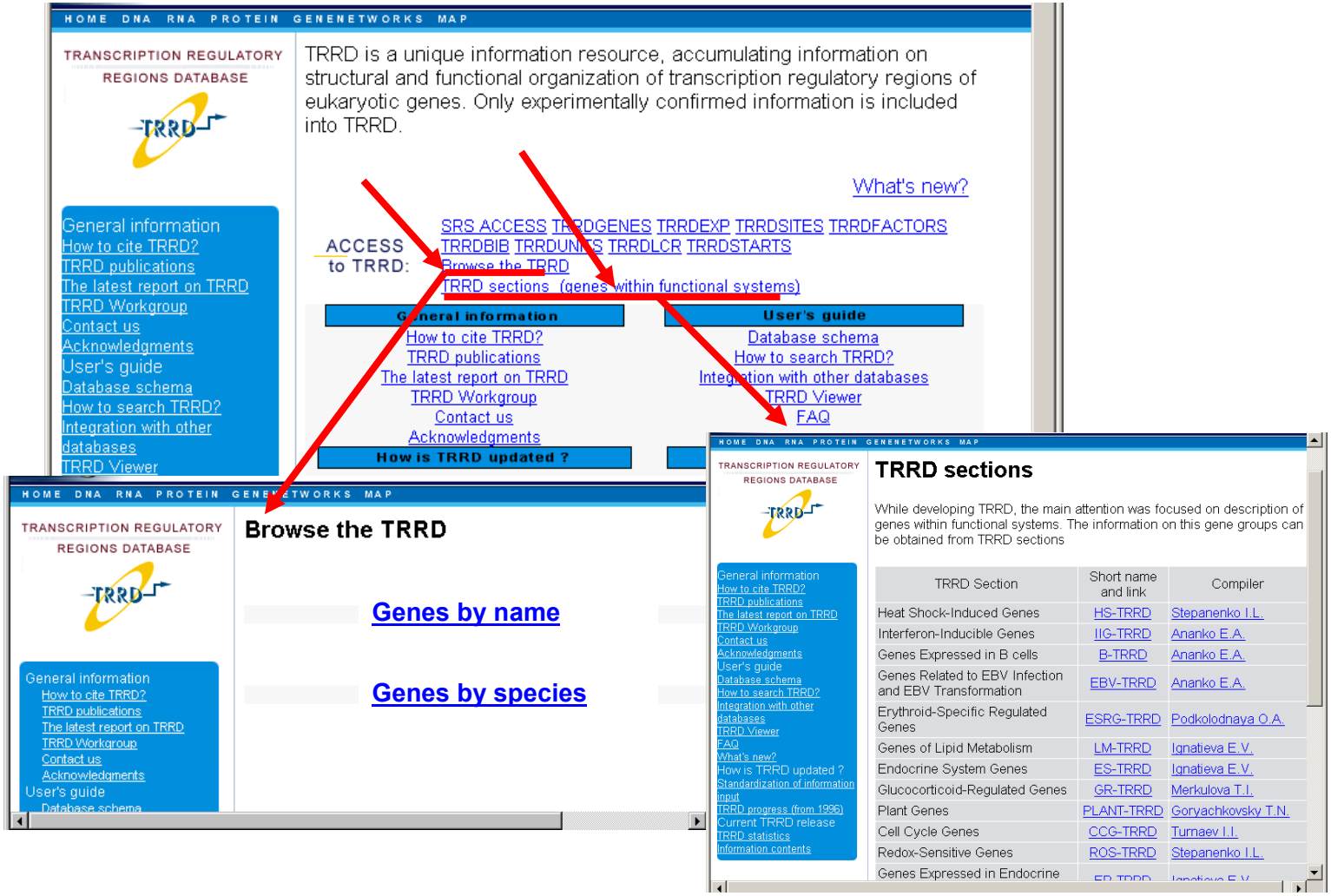


Data search in TRRD: SRS



DATABASE	NUMBER OF INDEXED FIELDS IN RELEASE 6.01
TRRDGENES	24
TRRDUNITS	11
TRRDEXP	17
TRRDSITES	16
TRRDFACTORS	14
TRRDLCR	40
TRRDBIB	9
TOTAL NUMBER	131

Data search in TRRD: browsers and TRRD sections

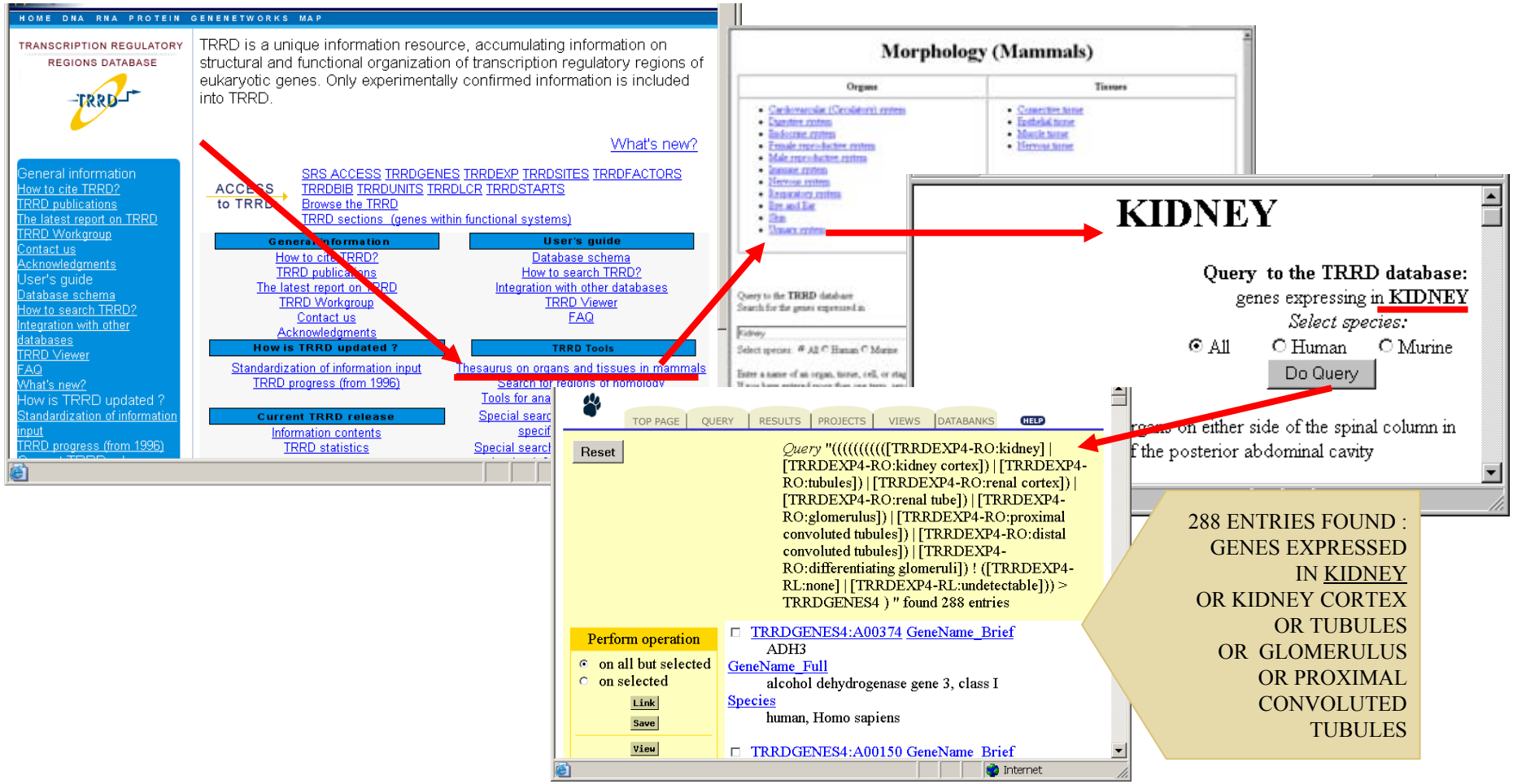


The TRRD sections http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/sections1.shtml

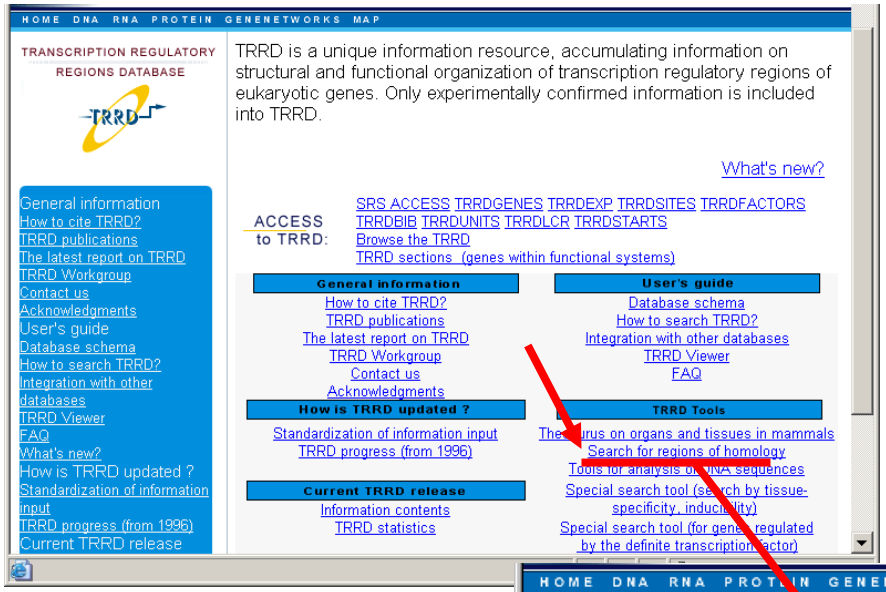
Within TRRD, the following topic sections are developed, uniting genes according to their functional characteristics:

TRRD Section	Short name and link	Compiler
Heat Shock-Induced Genes	HS-TRRD	Stepanenko I.L.
Interferon-Inducible Genes	IIG-TRRD	Ananko E.A.
Genes Expressed in B cells	B-TRRD	Ananko E.A.
Genes Related to EBV Infection and EBV Transformation	EBV-TRRD	Ananko E.A.
Erythroid-Specific Regulated Genes	ESRG-TRRD	Podkolodnaya O.A.
Genes of Lipid Metabolism	LM-TRRD	Ignatieva E.V.
Endocrine System Genes	ES-TRRD	Ignatieva E.V.
Glucocorticoid-Regulated Genes	GR-TRRD	Merkulova T.I.
Plant Genes	PLANT-TRRD	Goryachkovsky T.N.
Cell Cycle Genes	CCG-TRRD	Turnaev I.I.
Redox-Sensitive Genes	ROS-TRRD	Stepanenko I.L.
Genes Expressed in Endocrine Pancreas	EP-TRRD	Ignatieva E.V.
Macrophage-Expressed Genes	MG-TRRD	Ananko E.A.
Genes, controlling blood coagulation and fibrinolysis	BCF-TRRD	Khlebodarova T.M., Podkolodnaya O.A.
Apoptosis Genes	Apoptosis-TRRD	Stepanenko I.L.
Hepatitis C virus-induced Genes	HCV-TRRD	Stepanenko I.L.
Genes, controlling circadian rhythm, and genes with circadian expression	CLOCK-TRRD	Khlebodarova T.M.
Genes encoding proteins involved in the Fe metabolism	FM-TRRD	Mischenko E.L. , Podkolodnaya O.A.

Data search in TRRD: querying based on thesaurus on organs and tissues in mammals



Data search in TRRD: BLAST



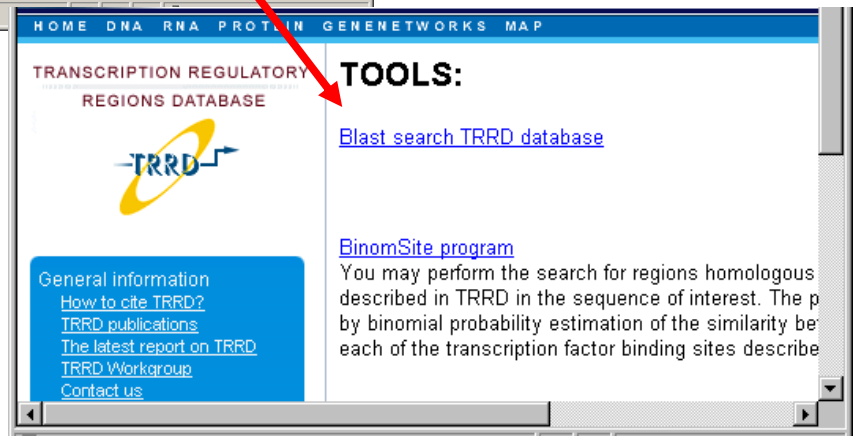
SEQUENCE UNDER INVESTIGATION:

`gtgtgaagaggagcgtacttttgtgtgtgaac`

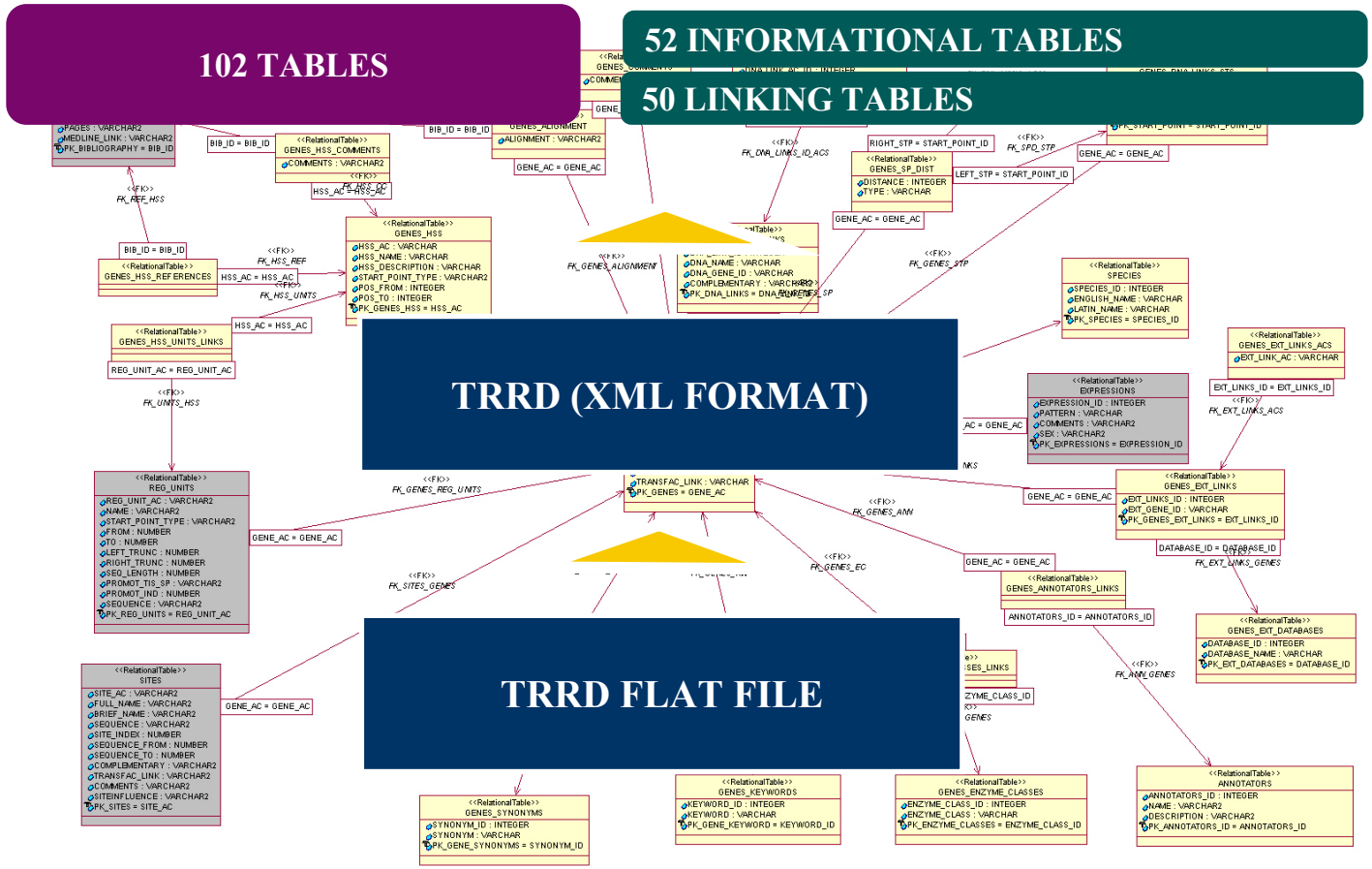
BLAST search TRRDUNITS

QUERY RESULTS:

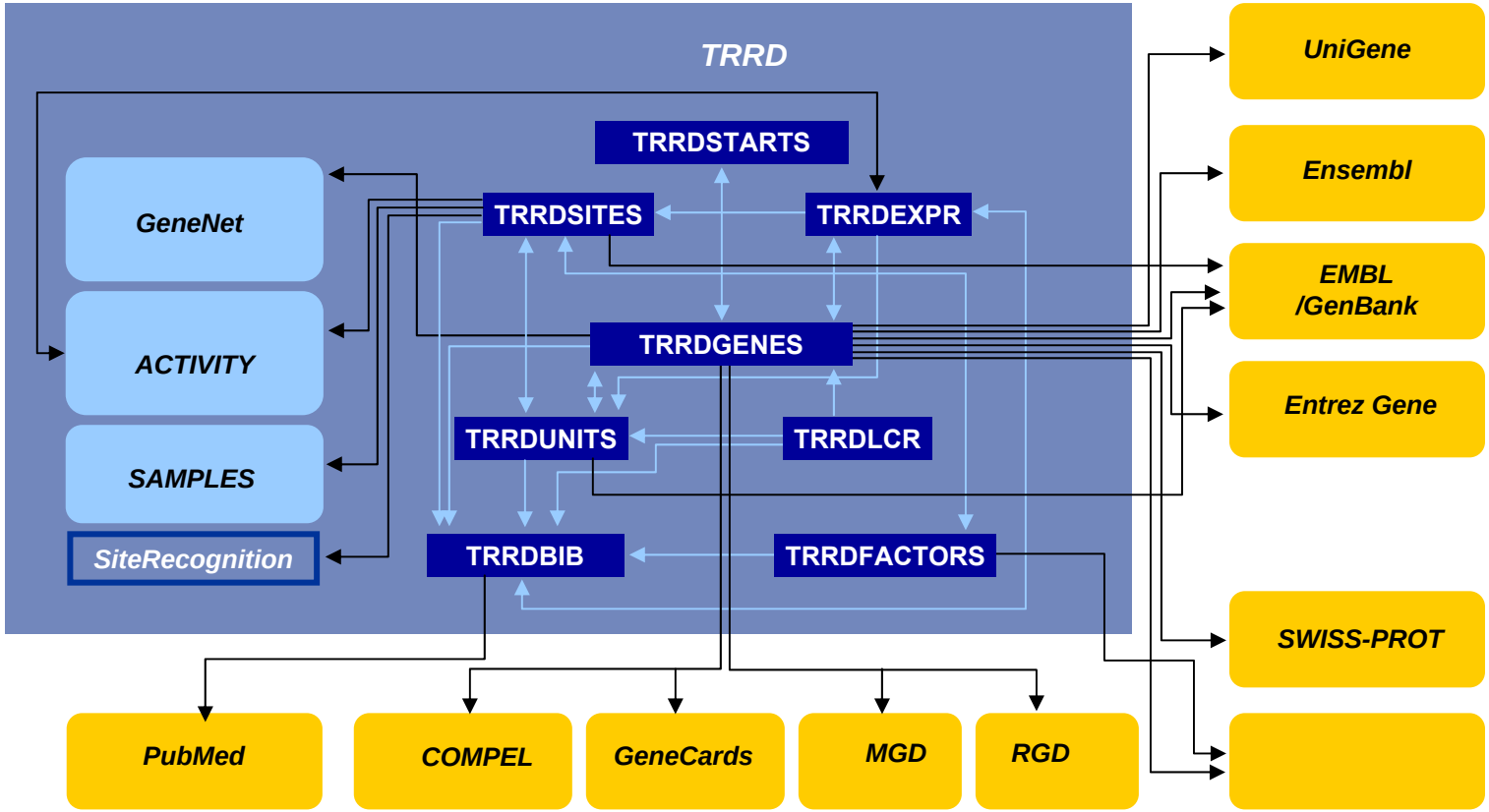
```
Query:gtgtgaagaggagcgtacttttgtgtgtga
                  ||||||||||||
Subject:      113 agcgtactttt 103
```



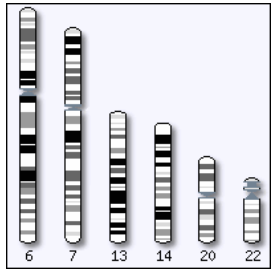
Relational version of the TRRD



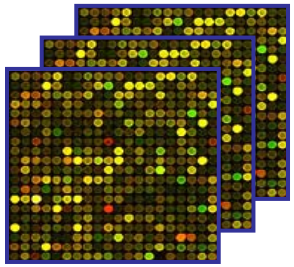
The links between TRRD tables and links from TRRD to external databases



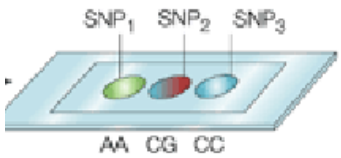
Applications of the TRRD



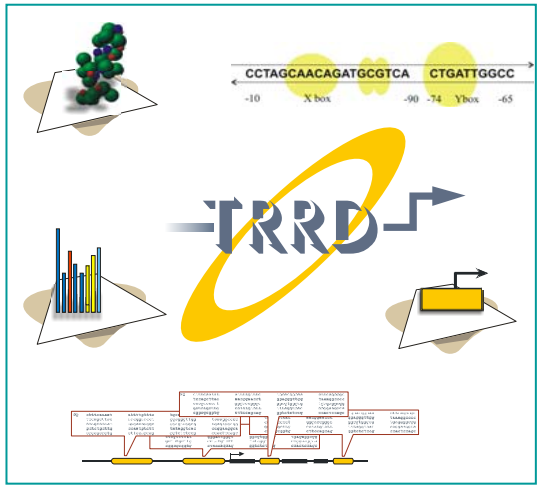
Annotation of the genomic sequences

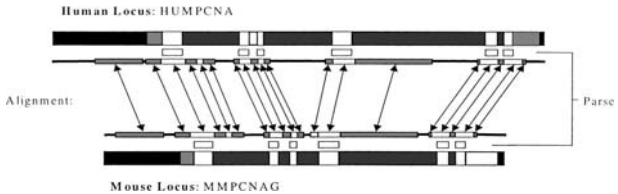


Interpreting Microarray data

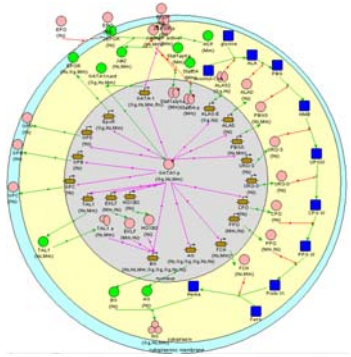


SNP analysis

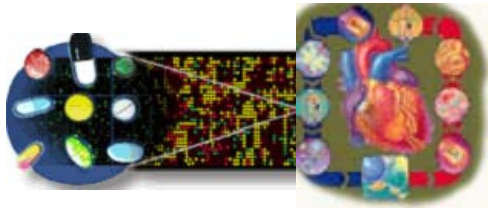




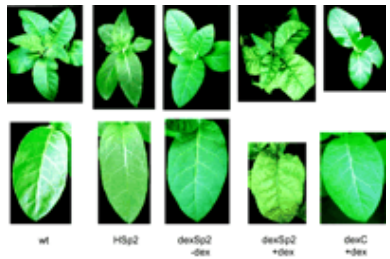
Comparative Genomics



Reconstruction of gene networks



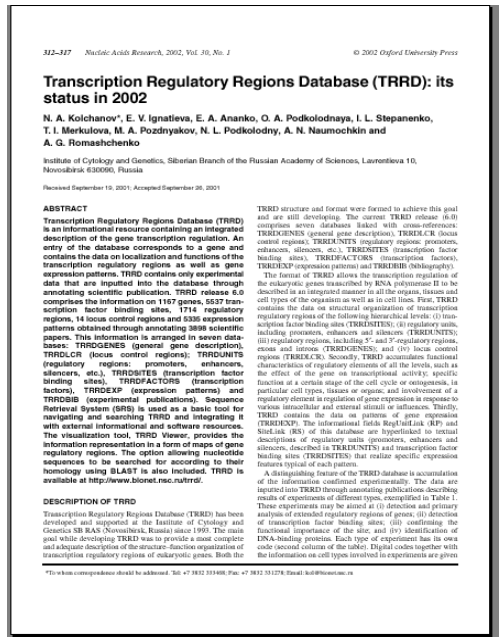
Gene therapy and molecular diagnostics



Transgenesis

The publications on the TRRD database

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 // Nucleic Acids Research, 2002, 30 (1), pp. 312-317.



312–317 Nucleic Acids Research, 2002, Vol. 30, No. 1 © 2002 Oxford University Press

Transcription Regulatory Regions Database (TRRD): its status in 2002

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

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Lavrentieva 10, Novosibirsk 630090, Russia

Received September 19, 2001; Accepted September 26, 2001

ABSTRACT

Transcription Regulatory Regions Database (TRRD) is an informational resource containing an integrated description of the gene transcription regulation. An entry of the database corresponds to a gene and contains the data on localization and functions of the transcription regulatory regions as well as gene expression patterns. TRRD contains only experimental data that are inputted into the database through annotating scientific publication. TRRD release 6.0 comprises the information on 1167 genes, 5537 transcription factor binding sites, 1714 regulatory regions, 14 locus control regions and 5335 expression patterns obtained through annotating 3898 scientific papers. This information is arranged in seven databases: TRRDGENES (general gene description), TRRDLCR (locus control regions); TRRDUNITS (regulatory regions: promoters, enhancers, silencers, etc.), TRRDSITES (transcription factor binding sites), TRRDFACTORS (transcription factors), TRRDEXP (expression patterns) and TRRDBIB (experimental publications). Sequence Retrieval System (SRS) is used as a basic tool for navigating and searching TRRD and integrating it with external informational and software resources. The visualization tool, TRRD Viewer, provides the information representation in a form of maps of gene regulatory regions. The option allowing nucleotide sequences to be searched for according to their homology using BLAST is also included. TRRD is available at http://www.bionet.nsc.ru/trrd/.

DESCRIPTION OF TRRD

Transcription Regulatory Regions Database (TRRD) has been developed and supported at the Institute of Cytology and Genetics SB RAS (Novosibirsk, Russia) since 1995. The main goal while developing TRRD was to provide a most complete and adequate description of the structure–function organization of transcription regulatory regions of eukaryotic genes. Both the TRRD structure and format were formed to achieve this goal and are still developing. The current TRRD release (6.0) comprises seven databases linked with cross-references: TRRDGENES (general gene description), TRRDLCR (locus control regions); TRRDUNITS (regulatory regions: promoters, enhancers, silencers, etc.), TRRDSITES (transcription factor binding sites), TRRDFACTORS (transcription factors), TRRDEXP (expression patterns) and TRRDBIB (bibliography).

The format of TRRD allows the transcription regulation of the eukaryotic genes transcribed by RNA polymerase II to be described in an integrated manner in all the organs, tissues and cell types of the organism as well as in cell lines. First, TRRD contains the data on structural organization of transcription regulatory regions of the following hierarchical levels: (i) transcription factor binding sites (TRRDSITES); (ii) regulatory units, including promoters, enhancers and silencers (TRRDUNITS); (iii) regulatory regions, including 5′- and 3′-regulatory regions, exons and introns (TRRDGENES); and (iv) locus control regions (TRRDLCR). Secondly, TRRD accumulates functional characteristics of regulatory elements of all the levels, such as the effect of the gene on transcriptional activity; specific function at a certain stage of the cell cycle or ontogenesis, in particular cell types, tissues or organs; and involvement of a regulatory element in regulation of gene expression in response to various intracellular and external stimuli or influences. Thirdly, TRRD contains the data on patterns of gene expression (TRRDEXP). The informational fields RegUnitLink (RP) and SiteLink (RS) of this database are hyperlinked to textual descriptions of regulatory units (promoters, enhancers and silencers, described in TRRDUNITS) and transcription factor binding sites (TRRDSITES) that realize specific expression features typical of each pattern.

A distinguishing feature of the TRRD database is accumulation of the information confirmed experimentally. The data are inputted into TRRD through annotating publications describing results of experiments of different types, exemplified in Table 1. These experiments may be aimed at (i) detection and primary analysis of extended regulatory regions of genes; (ii) detection of transcription factor binding sites; (iii) confirming the functional importance of the site; and (iv) identification of DNA-binding proteins. Each type of experiment has its own code (second column of the table). Digital codes together with the information on cell types involved in experiments are given

*To whom correspondence should be addressed. Tel: +7 3832 333498; Fax: +7 3832 331278; Email: kol@bionet.nsc.ru

E.V. Ignatieva, E.A. Ananko, O.A. Podkolodnaya, I.L. Stepanenko, T.M. Khlebodarova, T.I. Merkulova, M.A. Pozdnyakov, A.L. Proscura, D.A. Grigorovich, N.L. Podkolodny, A.N. Naumochkin, A.G. Romashchenko, N.A. Kolchanov Transcription Regulatory Regions Database (TRRD): description of transcription regulation and the main capabilities of the database. In: Bioinformatics of Genome Regulation And Structure. Ed. By N.kolchanov and R. Hofestaedt, Kluwer Academic Publishers, Boston/Dordrecht/London, 2004, pp.81-92.

N. Kolchanov, E. Ignatieva, O. Podkolodnaya, E. Ananko, I. Stepanenko, T. Merkulova, T. Khlebodarova, V. Merkulov, N. Podkolodny, D. Grigorovich, A. Poplavsky, A. Romashchenko Transcription regulatory regions database (TRRD): a source of experimentally confirmed data on transcription regulatory regions of eukaryotic genes . In: Bioinformatics of Genome Regulation and Structure II. (Eds. N.Kolchanov and R. Hofestaedt and L.Milanesi) Springer Science+Business Media, Inc. 2006, pp. 43-53.

2.3. The ArtSite database

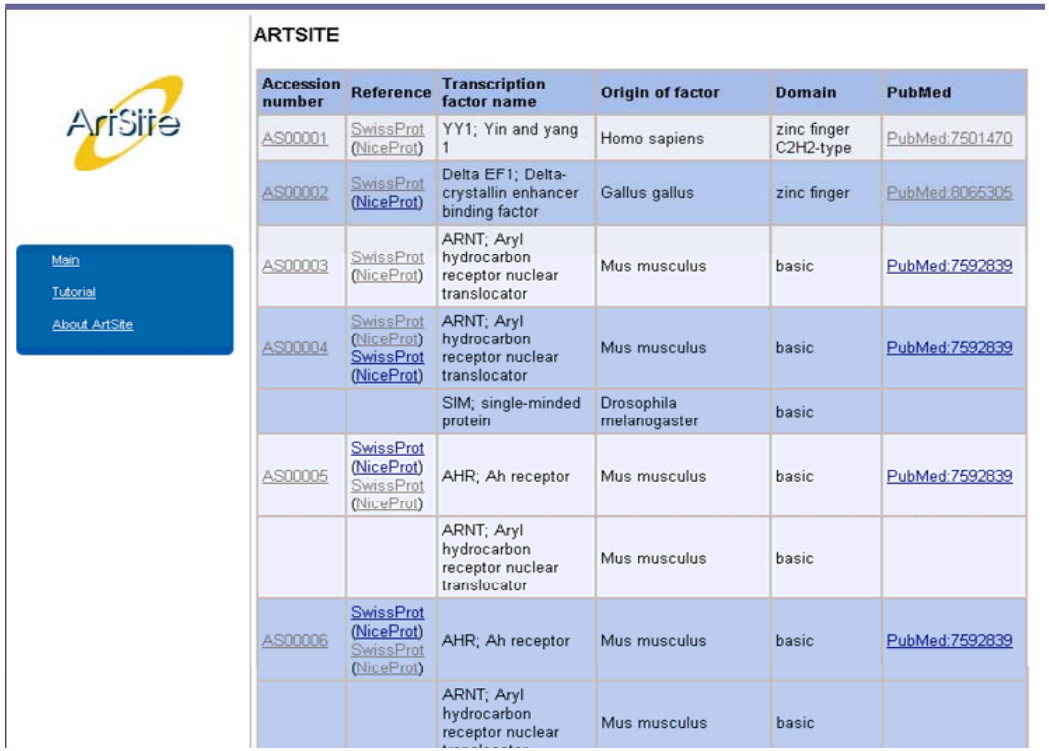
Recently, development of new technologies, in particular, SELEX, (Systematic Evolution of Ligands by EXponential enrichment), SAAB (Selected And Amplified Binding site imprint assay), REPSA (Restriction Endonuclease Protection Selection and Amplification), CASTing (Cyclical Amplification and Selection of Targets) and other *in vitro* selection procedures, yielded numerous data on the structures of binding sites for various transcription factors, both eukaryotic and prokaryotic. However, the questions on whether these data reflect the genuine structures of natural binding sites and what are the potential of applying these data to search for and prediction of natural sites are yet to be answered. We developed the database ArtSite, whose contents allowed us to make a comparative analysis of structures of natural and artificial binding sites.

ArtSite is a database accumulating information about the structures of sequences that specifically interact with DNA binding domains of transcription factors (TF) in pro- and eukaryotes. The characteristics of these sequences are described in the ArtSite database by means of frequency matrices, which are constructed on the basis of alignment of representative samples of TF binding sites. The samples are compiled by both genomic and synthesized *in vitro* DNA sequences binding TF in a specific manner, which are described in literature and revealed by different methods of selection.

T.M. Khlebodarova, O.A. Podkolodnaya, D.Y. Oshchepkov, D.S. Miginsky, E.A. Ananko, E.V. Ignatieva, I.L. Stepanenko. ARTSITE DATABASE: Structures of natural and *in vitro* selected transcription factor binding sites. In: Bioinformatics of Genome Regulation and Structure II. Ed. By N. Kolchanov and R. Hofestaedt, Springer Science+Business Media, Inc. 2005, pp. 55-65.

The description of ArtSite database

The ArtSite Web interface allows to make various queries (by name of TF, its synonyms, structure of DNA-binding domain, by origin of factor, and by literature source) and get a list of corresponding entries.



Accession number	Reference	Transcription factor name	Origin of factor	Domain	PubMed
AS00001	SwissProt (NiceProt)	YY1; Yin and yang 1	Homo sapiens	zinc finger C2H2-type	PubMed:7501470
AS00002	SwissProt (NiceProt)	Delta EF1; Delta-crystallin enhancer binding factor	Gallus gallus	zinc finger	PubMed:8065305
AS00003	SwissProt (NiceProt)	ARNT; Aryl hydrocarbon receptor nuclear translocator	Mus musculus	basic	PubMed:7592839
AS00004	SwissProt (NiceProt) SwissProt (NiceProt)	ARNT; Aryl hydrocarbon receptor nuclear translocator	Mus musculus	basic	PubMed:7592839
		SIM; single-minded protein	Drosophila melanogaster	basic	
AS00005	SwissProt (NiceProt) SwissProt (NiceProt)	AHR; Ah receptor	Mus musculus	basic	PubMed:7592839
		ARNT; Aryl hydrocarbon receptor nuclear translocator	Mus musculus	basic	
AS00006	SwissProt (NiceProt) SwissProt (NiceProt)	AHR; Ah receptor	Mus musculus	basic	PubMed:7592839
		ARNT; Aryl hydrocarbon receptor nuclear	Mus musculus	basic	

Fig.1. ArtSite WEB interface: search results view

The format of ArtSite database

An entry of ArtSite database corresponds to one selection experiment where in a matrix describing a binding site for a TF, or one of its domains, provided the factor interacts with DNA in a specific manner, is obtained. This format also allows for describing binding sites for heterodimeric proteins and intricate complexes of transcription factors. Description of such an entry is shown in Figure 2. An entry comprises 32 fields; of them, 21 fields are obligatory for filling in.

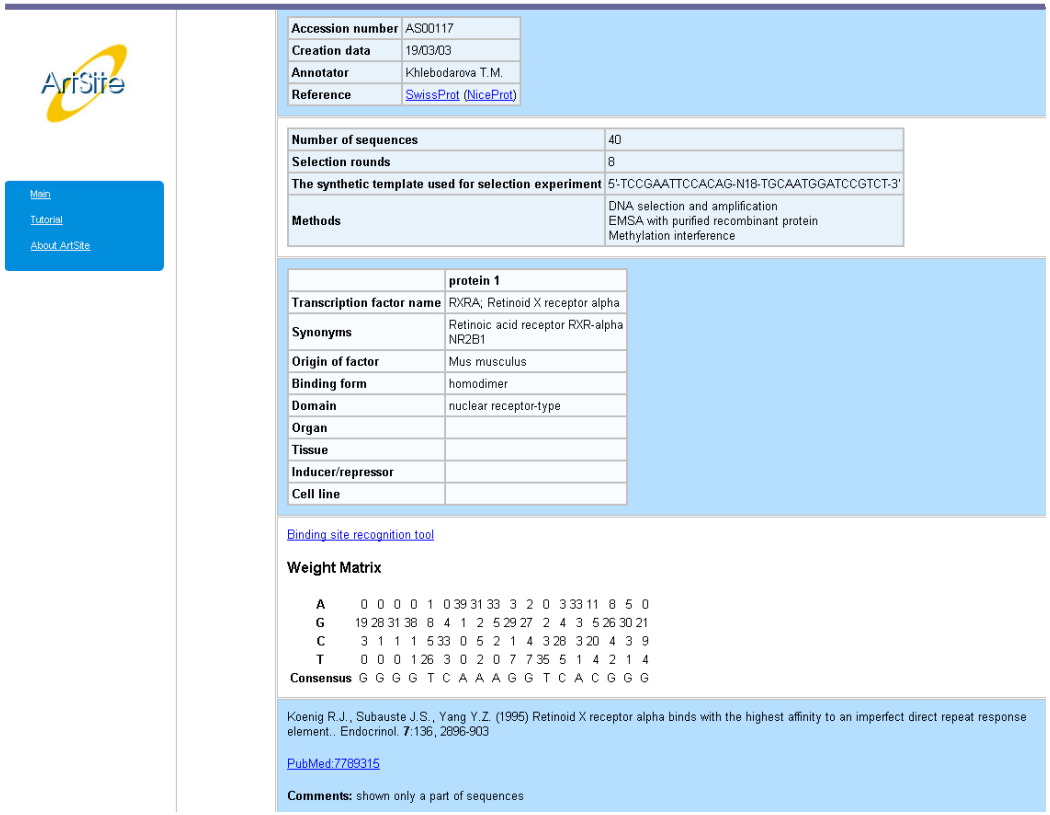


Fig.2. An example of the entry of ArtSite database describing sites for binding of the transcription factor RXRA to DNA detected *in vitro* experiments.

The content of ARTSITE database

The ARTSITE database is a natural extension of the database TRRD. The current release of the former database contains 560 matrices describing the binding sites for 356 transcription factors and their DNA-binding domains. Of them, 474 matrices were constructed basing on alignment of more than 15000 sequences detected using various variants for selecting transcription factor binding sites described in 215 original publications and 86 matrices describing natural, functional binding sites for 80 transcription factors. The latter 86 matrices were constructed basing on alignment of 2196 sequences extracted from TRRD. The data on species origin of transcription factors used for selection of binding sites by *in vitro* selection technologies is given in *Table* 1. The data given in *Table* 2 illustrate distribution of matrices by the structure of DNA-binding domains.

Table 1.

Organism	Number of matrices
Bacteria	12
Yeast	16
Fungi	1
C. elegans	1
Plants	28
Insect	20
(Drosophila)	434
Vertebrate	6
clawed frog	21
chicken	334
mammals	2
bovine	5
dog	183
human	116
mouse	1
rabbit	27
rat	17
Mammals virus	529
Total	

Table 2.

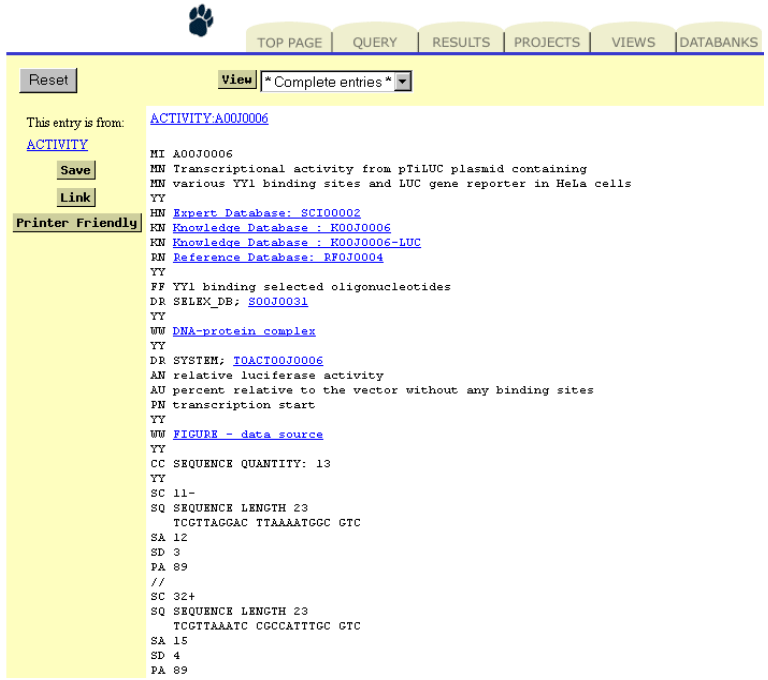
DNA-binding domain	Number of matrices (selected in vitro)	Number of matrices (natural sites)
Basic domain	103	30
CUT repeat	11	-
Ets-domain	19	4
Homeodomain	60	8
HMG box	14	6
Fork-head domain	10	1
MADS	18	1
Myb domain	20	1
Nuclear receptor type	30	6
Paired domain	7	-
POU domain	16	-
Tryptophan pentad repeat	5	-
Zinc finger	104	12
p53	2	3
helix-turn-helix	5	2
Other	18	15
Total	440	89

2.4. Computer system *Activity*

Analysis of context-dependent
conformational and and physico-chemical features (codes) of
DNA regulatory regions

Introduction

ACTIVITY is a database on DNA/RNA site sequences with known activity magnitudes, measurement systems, sequence-activity relationships under fixed experimental conditions, and procedures to adapt these relationships from one measurement system to another. This database deposits the information on DNA/RNA affinities to proteins and cell nuclear extracts, cutting efficiencies, gene transcription activity, mRNA translation efficiencies, mutability, and other biological activities of natural sites occurring within promoters, mRNA leaders, and other regulatory regions in pro- and eukaryotic genomes, their mutant forms and synthetic analogues. Since activity magnitudes are heavily system-dependent, ACTIVITY is supplemented by three sub-databases: (i) SYSTEM, measurement systems; (ii) KNOWLEDGE, sequence-activity relationships under fixed experimental conditions; and (iii) CROSS_TEST, procedures adapting a relationship from one measurement system to another. These databases are useful in molecular biology, pharmacogenetics, metabolic engineering, drug design, and biotechnology. ACTIVITY is available through the Web, http://wwwmgs.bionet.nsc.ru/systems/Activity/.

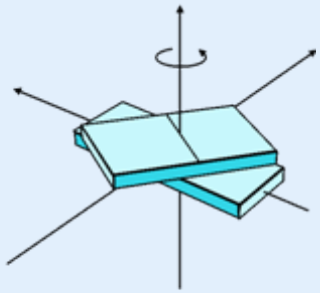


Description of the context dependent DNA double helix feature – *helical twist angle* in the database on conformational and physico-chemical properties of DNA

http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/

- An approach for predicting site activity based on its primary nucleotide sequence has been developed. The approach is realized in computer system ACTIVITY containing the databases on site activity and on conformational and physical-chemical DNA/RNA parameters. The computer system ACTIVITY is intended for generating programs with which to predict the activity of functional sites by nucleotide sequences. ACTIVITY analyzes a basis set of nucleotide sequences with known activity.
- The novelty of this approach is that Zadeh's fuzzy logic and decision making theory have been employed for determining the best "sequence→activity" regression. The best one thus determined is then automatically transformed into the "C" language source code of a computer-applicable program with which the activity for any nucleotide sequence is to be predicted.

MT PROPERTY COMPILATION "ACTIVITY"
//
MN Conformational
MD B-DNA
ML dinucleotide step
//
RN [1]
RA Suzuki M, Yagi N, Finch JT
RT Role of base-backbone and base-base interactions in
RT alternatingRT DNA conformations.
RJ FEBS Lett (1996) 379: 148-152
//
PN Helical twist
//
PU degrees
AA 35.6
AT 29.3
AG 31.9
AC 31.1**
TA 39.5**
TT 35.6
TG 36.0
TC 35.9
GA 35.9
GT 31.1**
GG 33.3
GC 34.6
CA 35.9
CT 31.9
CG 34.9
CC 33.3



Twist (degrees)

The sequence of the site S can be characterized by the mean value of the q-th conformational or physical-chemical property of DNA in the region (a, b):

$$X_{q,a,b}(S) = \frac{\sum_{i=a}^{b-1} P_q(s_i\, s_{i+1})}{b - a}$$

CODES OF FUNCTIONAL SITES ACTIVITY

Linear additive model for prediction of site activity

$$F(S) = F_0(S) + \sum_{k=1}^{K} F_k \times X_k(S)$$

$F(S)$ - the value of the activity of a site with the nucleotide sequence S;

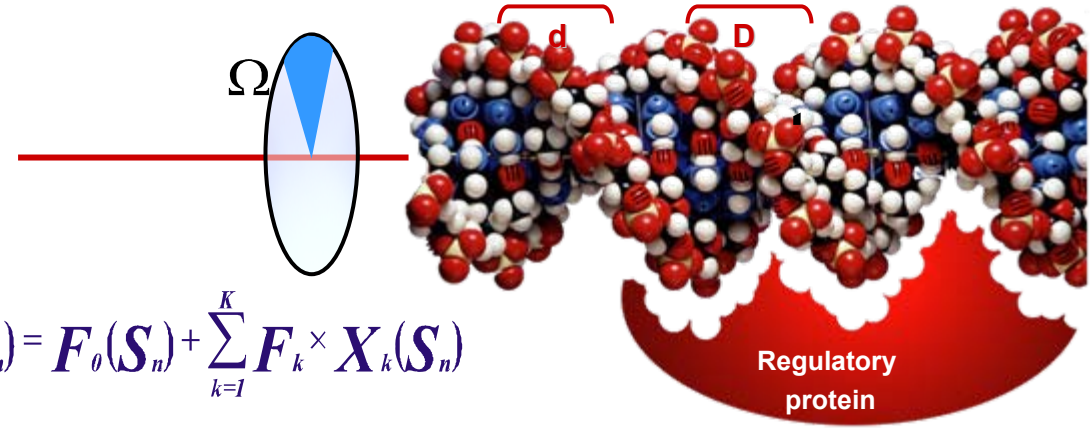
$F_0(S)$ - the basal activity level of the sites of the given type;

F_K - the contribution of the facultative feature X_K to site activity;

$X_K(S)$ - the value of the facultative feature X_K for the site sequence S of the given type.

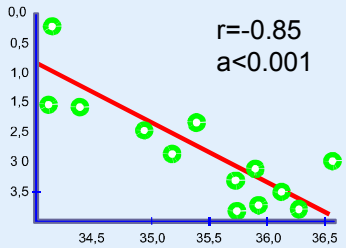
Codes of functional sites activity

Affinity of regulatory proteins to their binding sites is determined by DNA conformational properties and can be strongly changed by single nucleotide substitutions



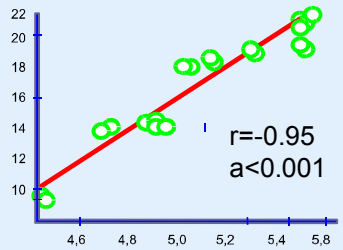
$$F(S_n) = F_0(S_n) + \sum_{k=1}^{K} F_k \times X_k(S_n)$$

Affinity USF/DNA



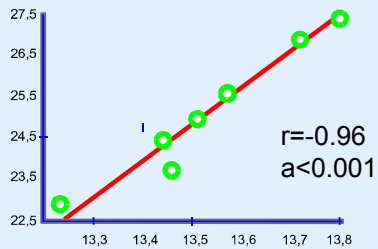
Ω ' Helical twist angle

Affinity TBP/DNA



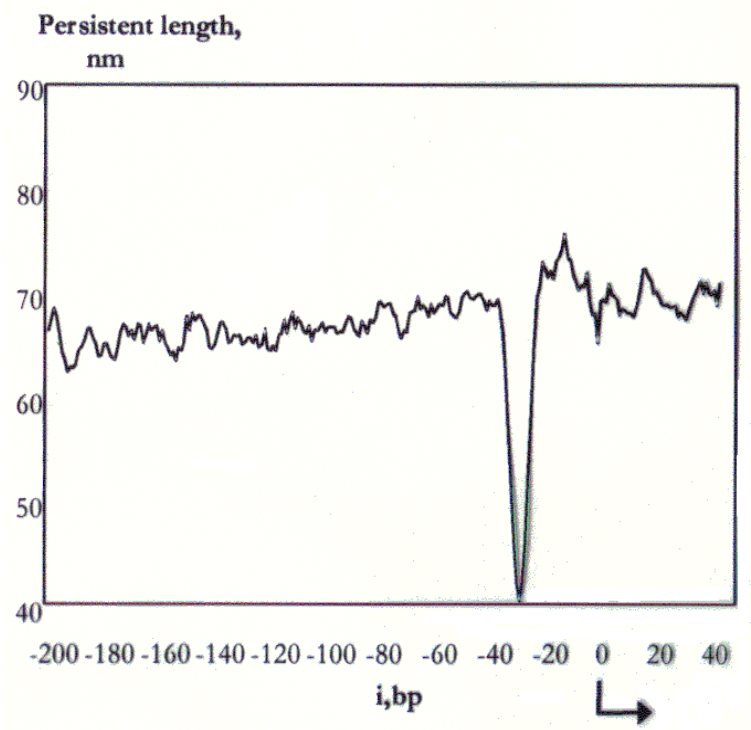
d , Small groove width

Affinity CRO/DNA

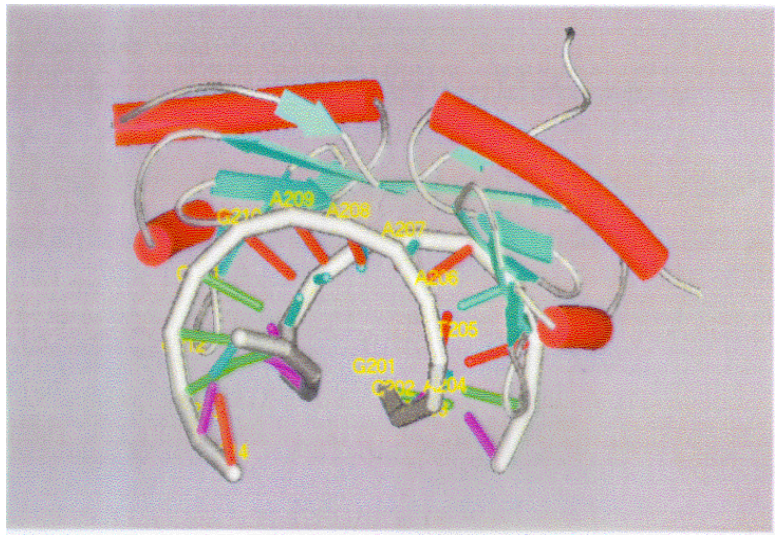


D , Large groove width

Context-dependent conformational and physico-chemical features (codes) of DNA regulatory regions



Calculated profile of bending stiffness for TATA-containing promoters of verterbrates



x-ray structure of tata-box binding polypeptide (TBP)-DNA complex (index NDB PDT009)

2.5. Transcription factor binding sites computer analysis and recognition

Computer-based experimental approach to study transcription factor binding sites

Laboratory of Theoretical Genetics, Laboratory of Gene Expression Regulation;
Sector of Molecular Genetic Mechanisms of Protein-Nucleic Interactions;

Prediction quality assessment, data interpretation

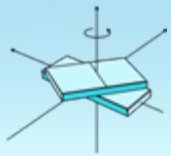
Computer analysis
- Development of TFBS databases
- Identification of significant properties of TFBS
- Development of methods for identification of TFBS

Experiment set-up to verify predicted transcription factor binding sites

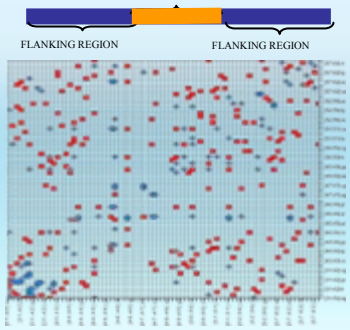
Experimental Research
- Synthesis of oligonucleotides corresponding to predicted sites
- Gel-shift analysis to study binding of oligonucleotdes to nuclear proteins isolated from tissues that express the corresponding factors
- Identification of binding proteins using antibodies

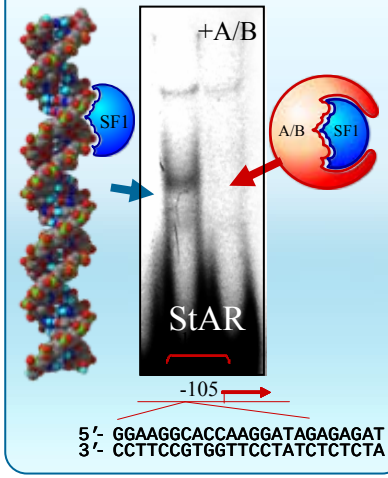
SITECON
by D.Y. Oshchepkov

A method for site recognition by context-dependable conformational and physico-chemical properties of DNA

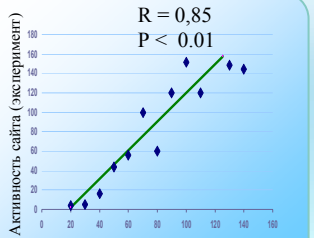


SiteGA
by V.G. Levitsky
A method for recognition of transcription factor SF1 binding sites





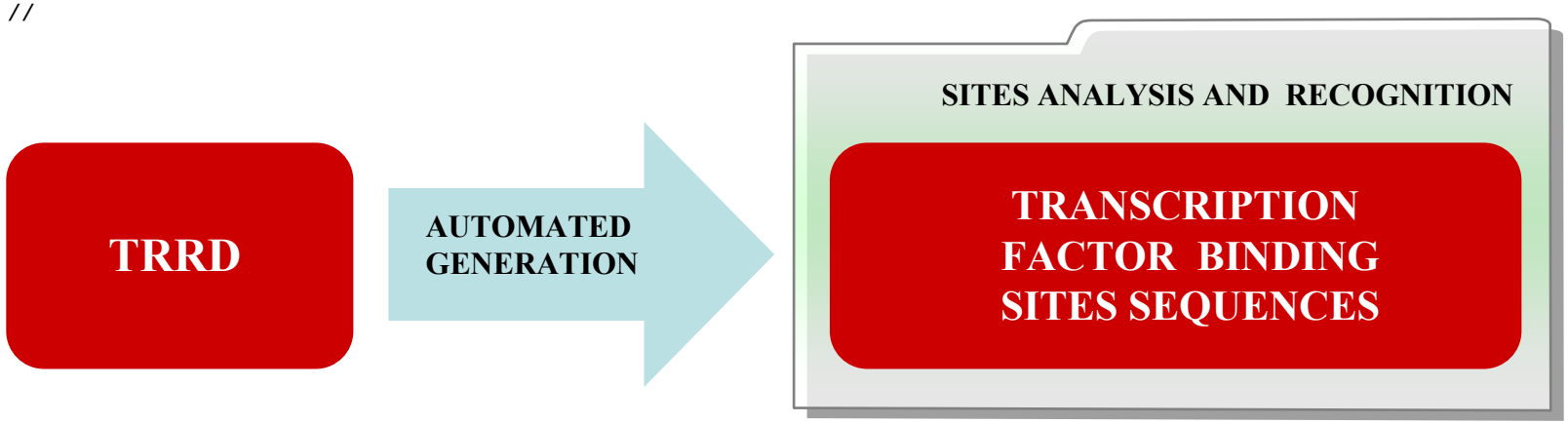
Analysis and recognition of TATA-boxes in eukaryotic promoters



Sequence data: TRRD/ ARTSITE based samples of transcription factor binding sites

```
ID   es_250_1; DNA
AC   es_250_1
CC   DE   adrenodoxin gene
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata;
OC   Vertebrata; Mammalia; Eutheria; Primates;
OC   Catarrhini; Hominidae; Homo.
DR   EMBL; M23665; HSADRDO01; ; join(133..382)
CC   ST (EMBL/GENBANK) 333
DR   TRRDGENES; A00860; Hs:ADX; 4.2;
FT   {0,0} [1;250]; EXP
SQ   ctttcaaaat attttgtttc tgcacggcaa cttcagccgc
     tccagcttac aacggaacct ggagggttgg taaaggcccc
     cccgccccat gggaccgggc ggcgtgggcg tgagaggcgg
     gctctgcttg ccaatgtctt tataggtcac ccggaaggca
     cggcgcggtg cttccagcag ggtctctccg ccactccagc
//
```

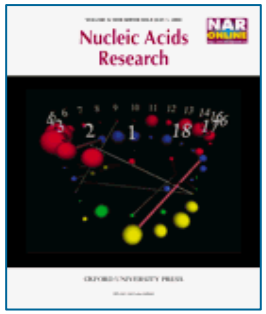
```
SQ   ctttcaaaat attttgtttc tgcacggcaa cttcagccgc
     tccagcttac aacggaacct ggagggttgg taaaggcccc
//
SQ   cccgccccat gggaccgggc ggcgtgggcg tgagaggcgg
     gctctgcttg ccaatgtctt tataggtcac ccggaaggca
//
SQ   cggcgcggtg cttccagcag ggtctctccg ccactccagc
     attttgtttc tgcacggcaa cttcagccgc gctaagttgc
//
```



2.5.1. SITECON:

a tool for detecting conservative conformational and physicochemical properties in transcription factor binding sites and for sites recognition

The method SITECON helps to make analysis of conserved context-dependable conformational and physico-chemical properties of DNA



Nucleic Acids Research 2004, V.32, W208-W212.

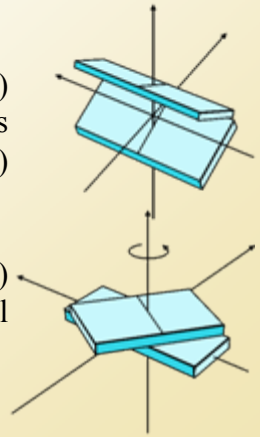
Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A., Ignatieva E.V., Khlebodarova T.M.

SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition.

SITECON uses information from the database on conformational and physico-chemical properties of DNA (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/)

Roll (ρ)
Angle of opening of the bases
(across the short axis)

Twist (ω)
angle of twist of DNA spiral



Dinucleotide	Roll, (degrees)	Twist, (degrees)
AA	0.3	35.3
AT	–0.8	31.2 *
AG	4.5	31.2 *
AC	0.5	32.6
TA	2.8	40.5 **
TT	0.3	35.3
TG	0.5	32.6
TC	–1.3	40.3
GA	–1.3	40.3
GT	0.5	32.6
GG	6.0 **	33.3
GC	–6.2 *	37.3
CA	0.5	39.2
CT	4.5	31.2 *
CG	–6.2 *	36.6
CC	6.0 **	33.3

SITECON: a method for study of the conserved conformational and physicochemical properties in short regions of DNA sites

A set of *N* aligned (phased) functional DNA sequences

```
tcaatccctg ggtttgccca …
acagctagaa ttgtctccta …
cttccagatt cctgagaggc …
tgccttccta tcactgaata …
```

Physicochemical property F_i is ascribed to each dinucleotide
("PROPERTY" database, 38 properties)

```
1.2, 1.4, 2.1, 3.1, 1.6, 1.8, 2.5, …
2.3, 3.1, 3.7, 3.2, 1.8, 1.2, 2.1, …
1.3, 3.1, 2.6, 2.4, 1.6, 2.3, 1.3, …
2.8, 1.4, 2.6, 1.8, 3.2, 1.5, 2.1, …
```

The mean value $\overline{F}_i$ and variance σ_{F_i} of each property in each position *k* of alignment is calculated

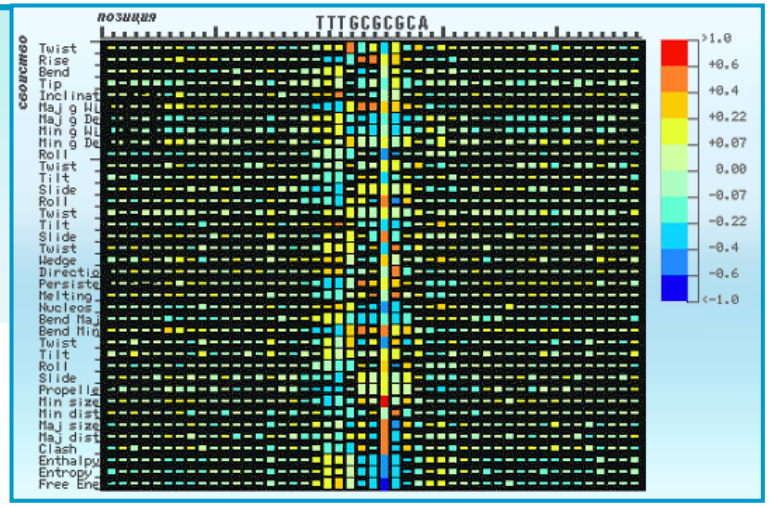
$$\overline{F}_i = \frac{1}{N}\sum_k F_{ik} \quad , \quad \overline{\sigma}_{F_i} = \frac{1}{N-1}\sum_k (F_{ik} - \overline{F}_i)^2$$

The significance of σ_{F_i} for each property *i* in each position *k* is estimated, comparing it with a set of random sequences
Using χ^2 test

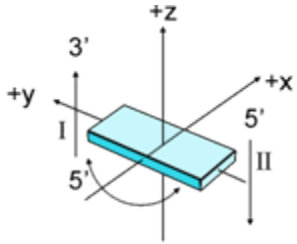
Matrix analysis, recognition rules design

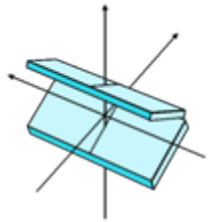
On the diagram block size indicates the significance level. The deviations from the means for the random sequences are in different colors.

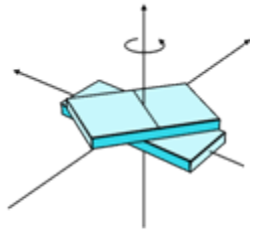
$$\overline{F}_{show} = \overline{F}_{il} - \overline{F}_{rand} \Big/ \sigma_{rand}$$



Context-dependent conformational properties of the B-form DNA




Roll (ρ)


Twist (ω)

Dinucleotide	Roll, degree	Twist, degree
AA	0.3	35.3
AT	–0.8	31.2 *
AG	4.5	31.2 *
AC	0.5	32.6
TA	2.8	40.5 **
TT	0.3	35.3
TG	0.5	32.6
TC	–1.3	40.3
GA	–1.3	40.3
GT	0.5	32.6
GG	6.0 **	33.3
GC	–6.2 *	37.3
CA	0.5	39.2
CT	4.5	31.2 *
CG	–6.2 *	36.6
CC	6.0 **	33.3

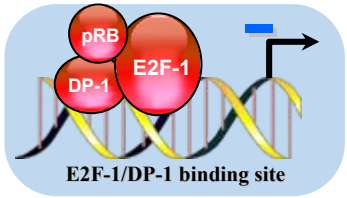
DB "PROPERTY"
(http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/)
Ponomarenko J.V., Ponomarenko M.P., Frolov A.S., Vorobyev D.G., Overton G.C., Kolchanov N.A. Conformational and physicochemical DNA features specific for transcription factor binding sites. Bioinformatics 1999; 15:654–68.

2.5.2. Computer analysis of e2f/dp transcription factor binding site using SITECON method

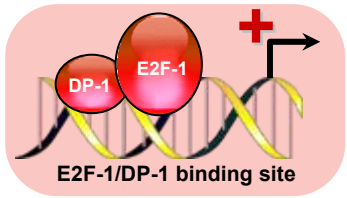
Cell cycle machinery of a eukaryotic cell: alternative repression and activation of large gene groups involving transcription factor e2f-1

(reconstruction using GeneNet database)

Transcription suppression

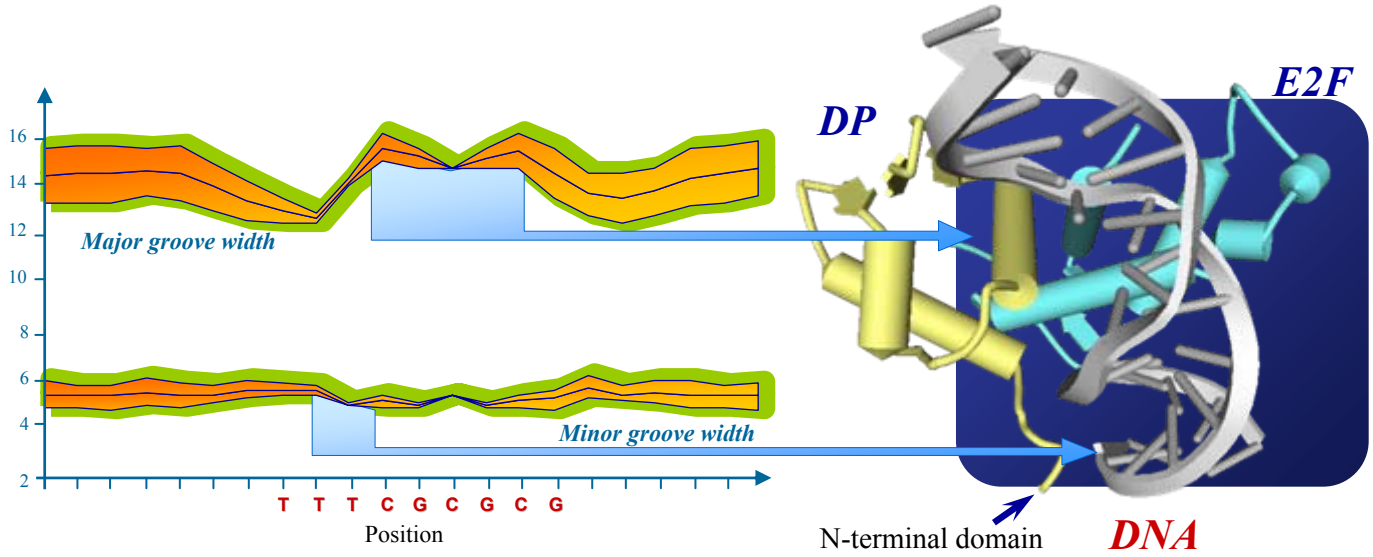


Transcription enhancement



Ananko E.A., Podkolodnaya O.A., Turnaev I.I. (Laboratory of Theoretical Genetics)

SITECON: analysis of e2f/dp binding sites



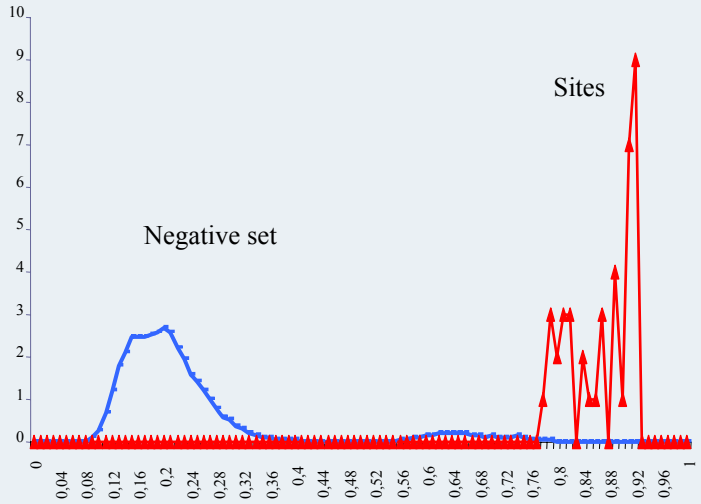
Conformational similarity value

$$P_{\Sigma} = \sum_{i=0,l=0}^{I,L} \delta_{il} \frac{1}{\sqrt{2\pi}\sigma_{F_{il}}} \exp\left(-\left(\left(\overline{F_{il}} - F_{il}\right)/\sigma_{F_{il}}\right)^2\right) \Big/ \sum_{i=0,l=0}^{I,L} \delta_{il}$$

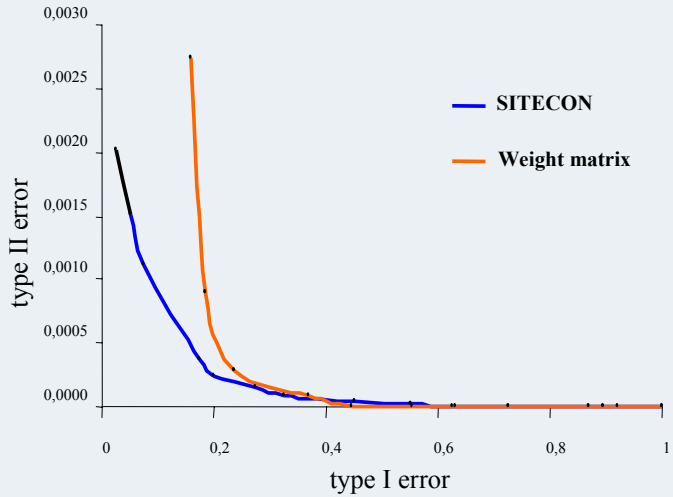
 http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/ mailto:diman@bionet.nsc.ru

SITECON: recognition of e2f/dp binding sites

distributions of the values of the conformational similarity score for the positive and negative sets

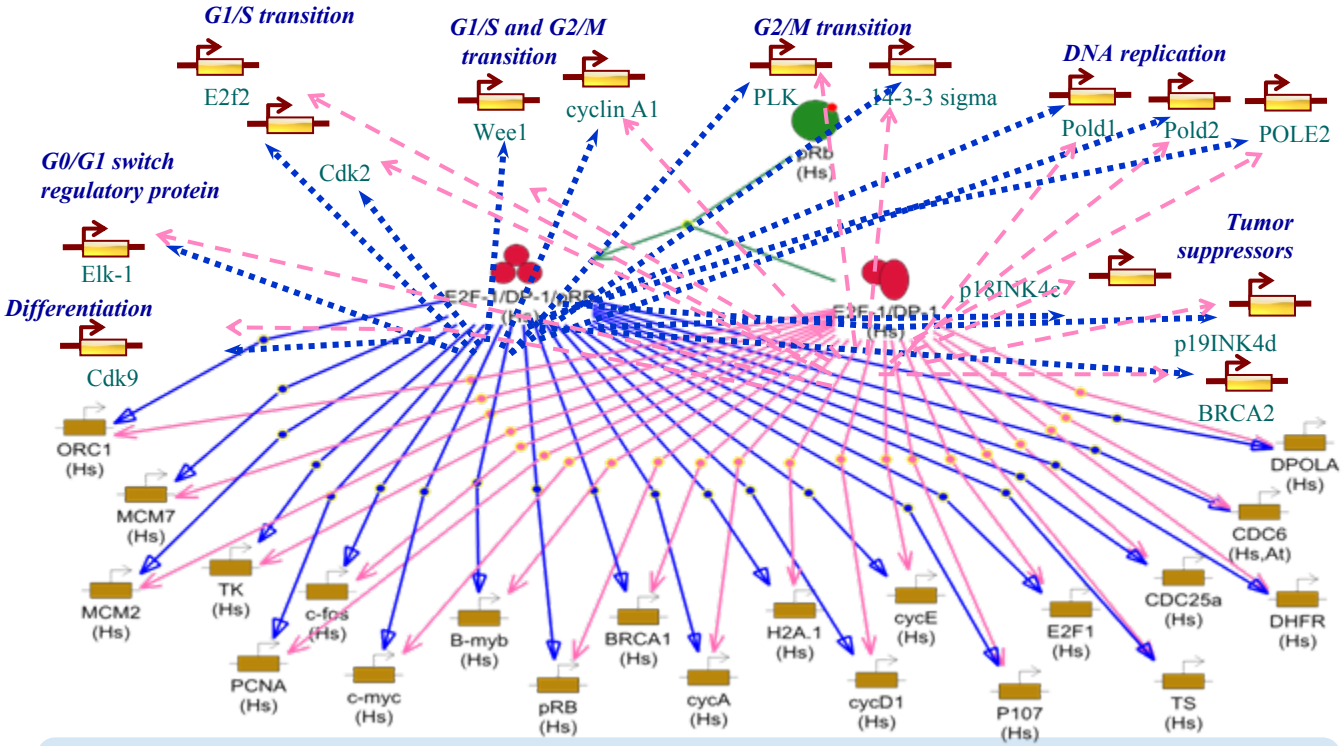


recognition quality comparison



Conformational similarity	83%	85%	87%
Type I error (false negatives)	0.11	0.15	0.18
Type II error (false positives)	$3.4*10^{-4}$ (1/2900)	$1.6*10^{-4}$ (1/6300)	$5.4*10^{-5}$ (1/18500)

Adding potential target genes of transcription factor e2f-1/DP-1 to the cell cycle regulation gene network after prediction of its binding sites. Potential target genes for transcription factors interaction are dotted lined.

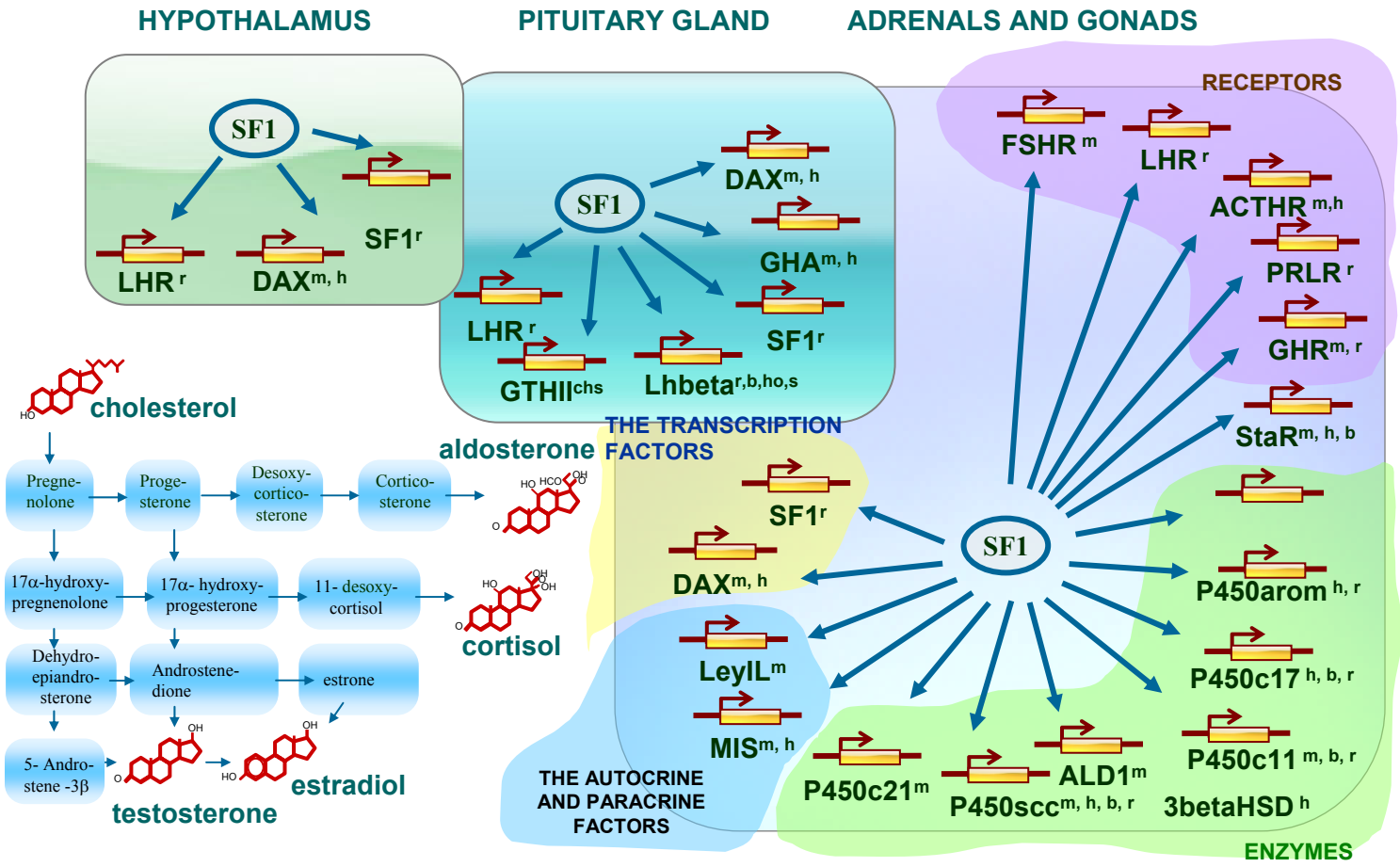


New transcription factor E2F-1 binding sites in 14 human genes were predicted

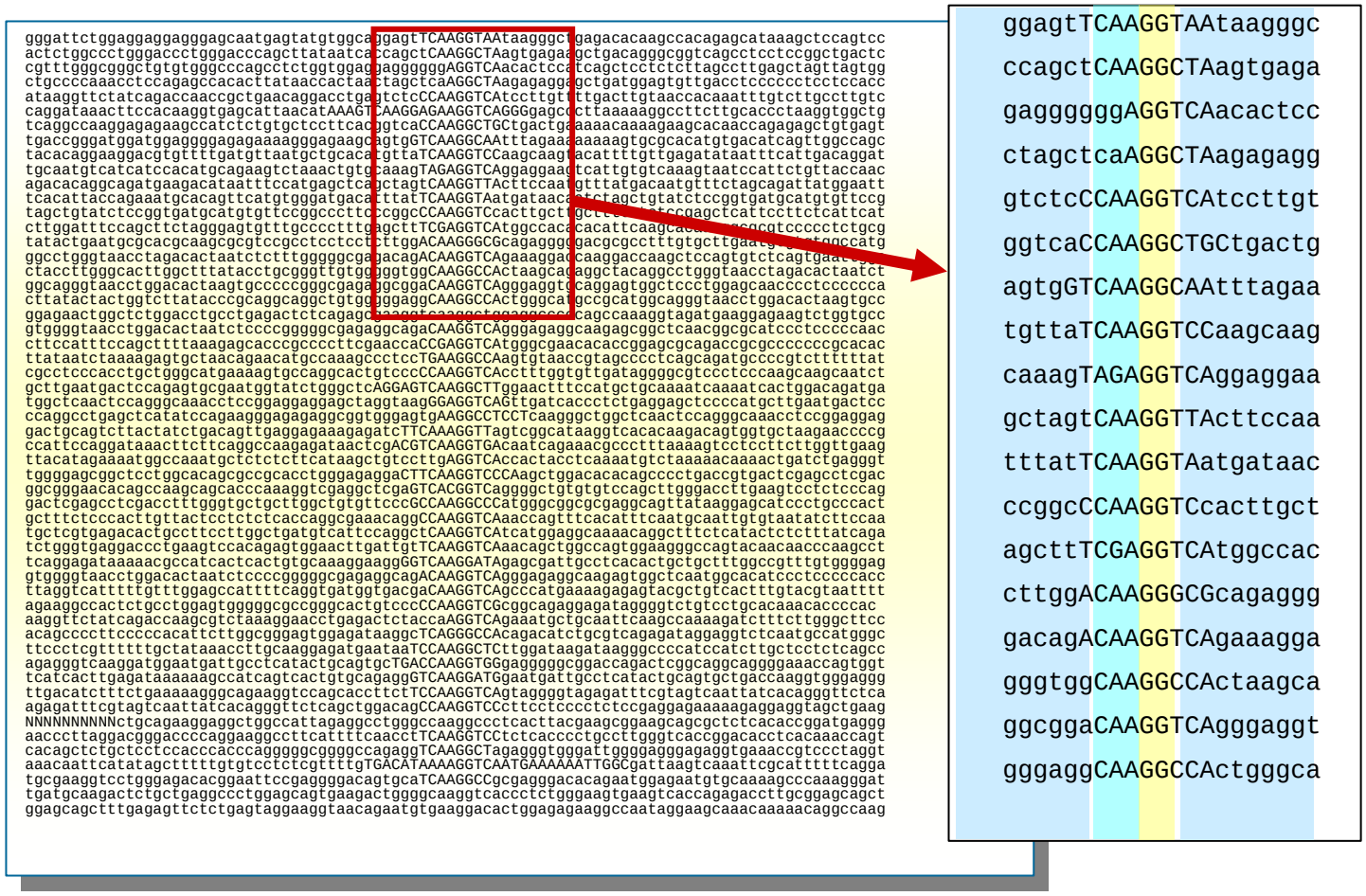
Turnaev I.I., Oshchepkov D.Y., Ananko E.A. (Laboratory of Theoretical Genetics)

2.5.3. Computer assisted experimental studies of SF-1 transcription factor binding site using SITECON method

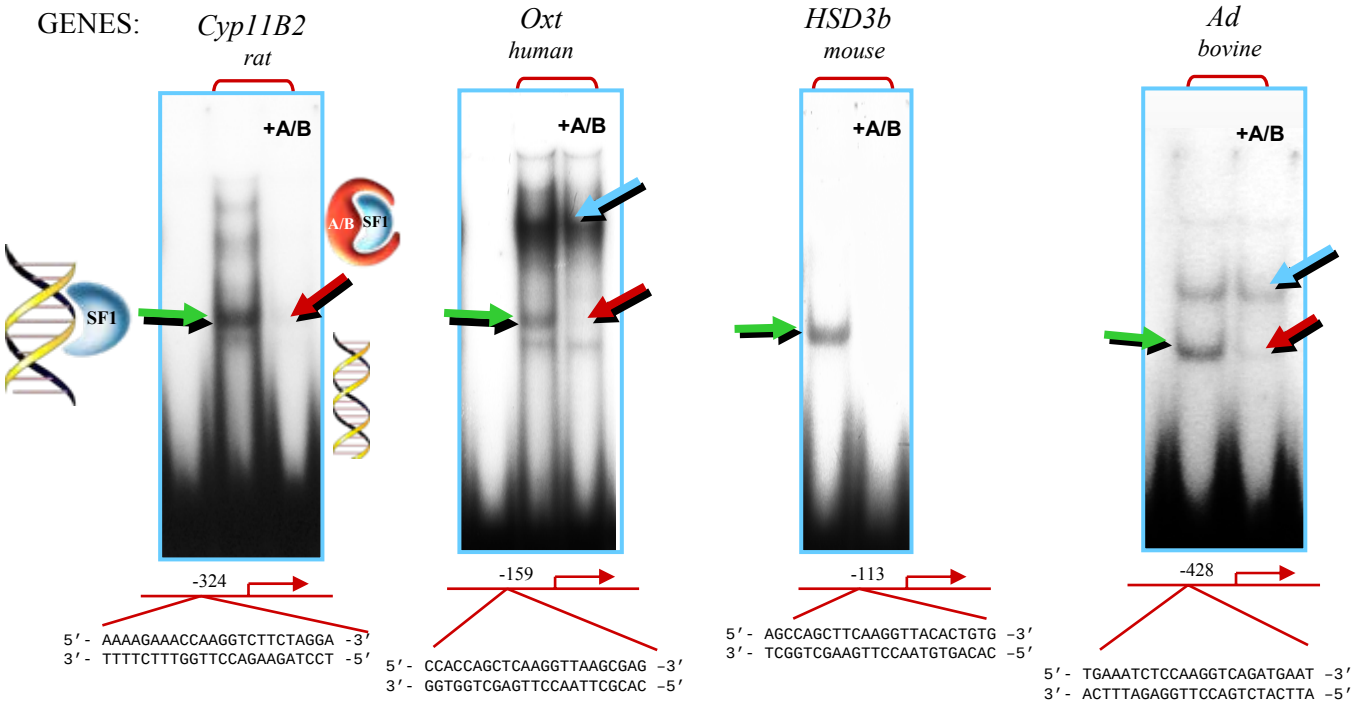
SF-1 (steroidogenic factor 1) coordinates functions of different hierarchical levels of the steroidogenesis gene network



SF-1 site sequences extracted from the TRRD database

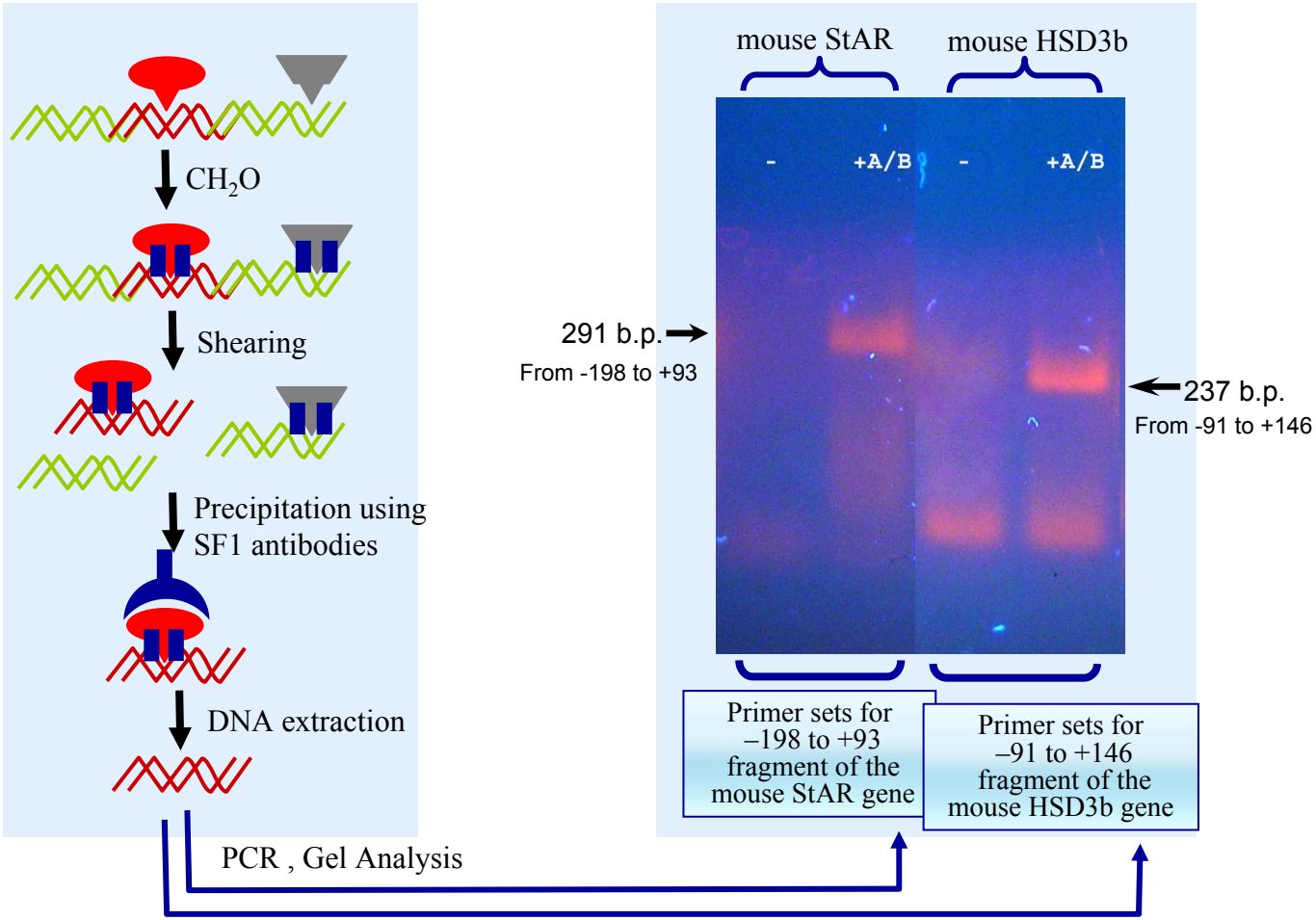


Experimental support for the functional activity of the SF-1 sites predicted by SITECON, by electrophoretic mobility shift assay (EMSA) with addition of antibodies against SF-1



- The shifted SF1/DNA complex
- Disappearance (or weakening) of the bands due to antibody against SF-1 (A/B) in the right lane supports SF-1 binding to the site.
- The complexes formed by the other proteins

In vivo Chromatin Immunoprecipitation assay using anti-SF1 antibodies



SITECON

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

Selected publications :

D. Y. Oshchepkov, E. E. Vityaev, D. A. Grigorovich, E. V. Ignatieva and T. M. Khlebodarova (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucl.Acids Res*., Jul 1;32(Web Server issue):W208-12.

Oshchepkov,D.Y., Turnaev,I.I., Pozdnyakov,M.A., Milanesi,L., Vityaev,E.E. and Kolchanov,N.A. (2004) SITECON—a tool for analysis of DNA physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition. In Kolchanov,N. and Hofestaedt,R. (eds), Bioinformatics of Genome Regulation and Structure. Kluwer Academic Publishers, Boston/ Dordrecht/London, pp. 93–102.

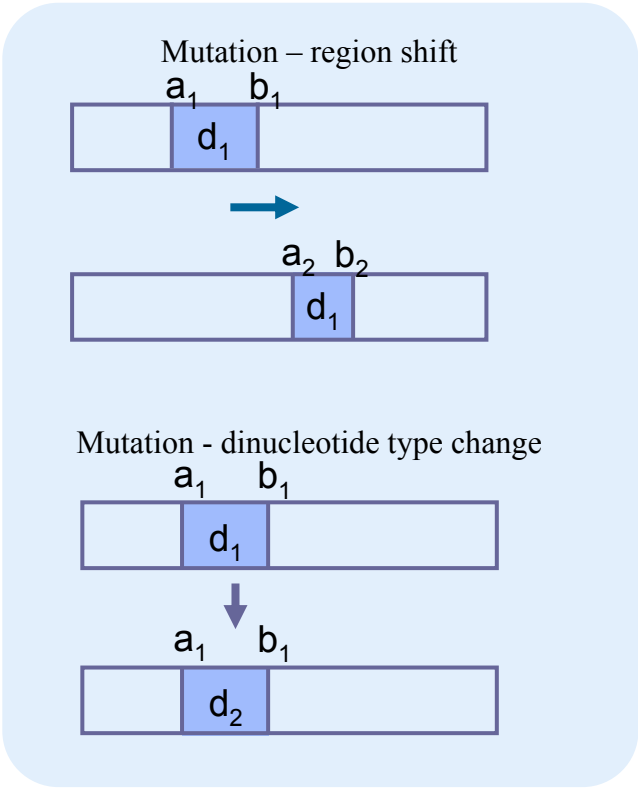
2.5.4. SiteGA :

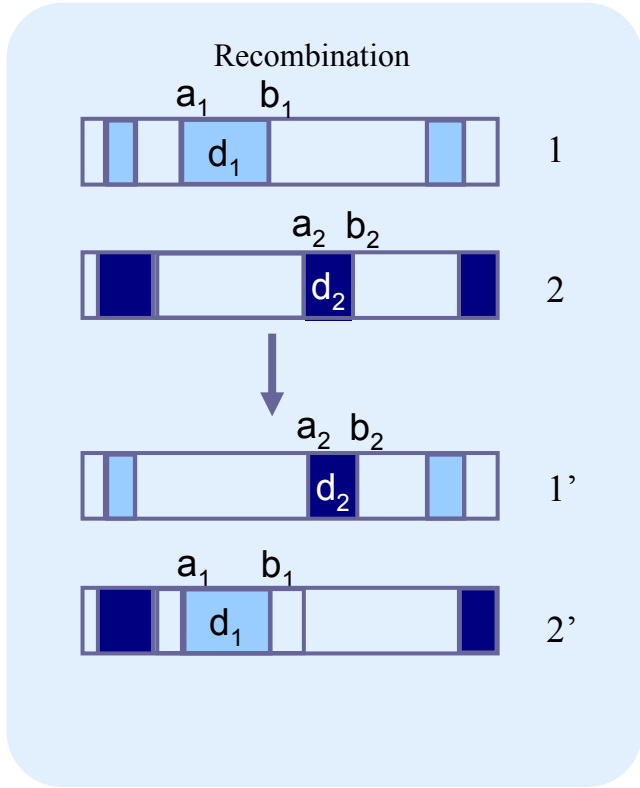
a tool for transcription factor binding sites analysis and recognition based on the genetic algorithm

We developed and implemented the SiteGA method for transcription-factor binding-sites recognition

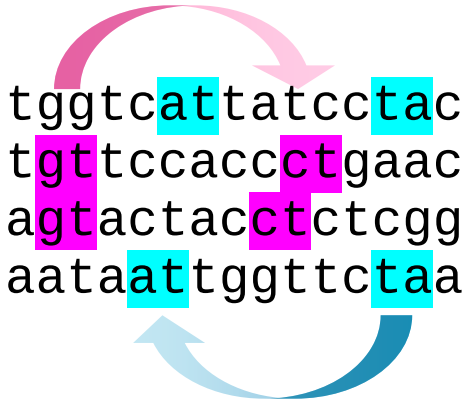
SiteGA method: Genetic Algorithm

Search for a set of locally positioned dinucleotides (LPD),
each LPD defined by location $[a_1,b_1]$ and dinucleotide type d_1





SiteGA method novelty – consideration of dependencies between different site position. Does the additivity assumption fit the biological data perfectly?



In most cases additivity assumption provides a very good approximation of the true nature of the specific protein-DNA interactions.

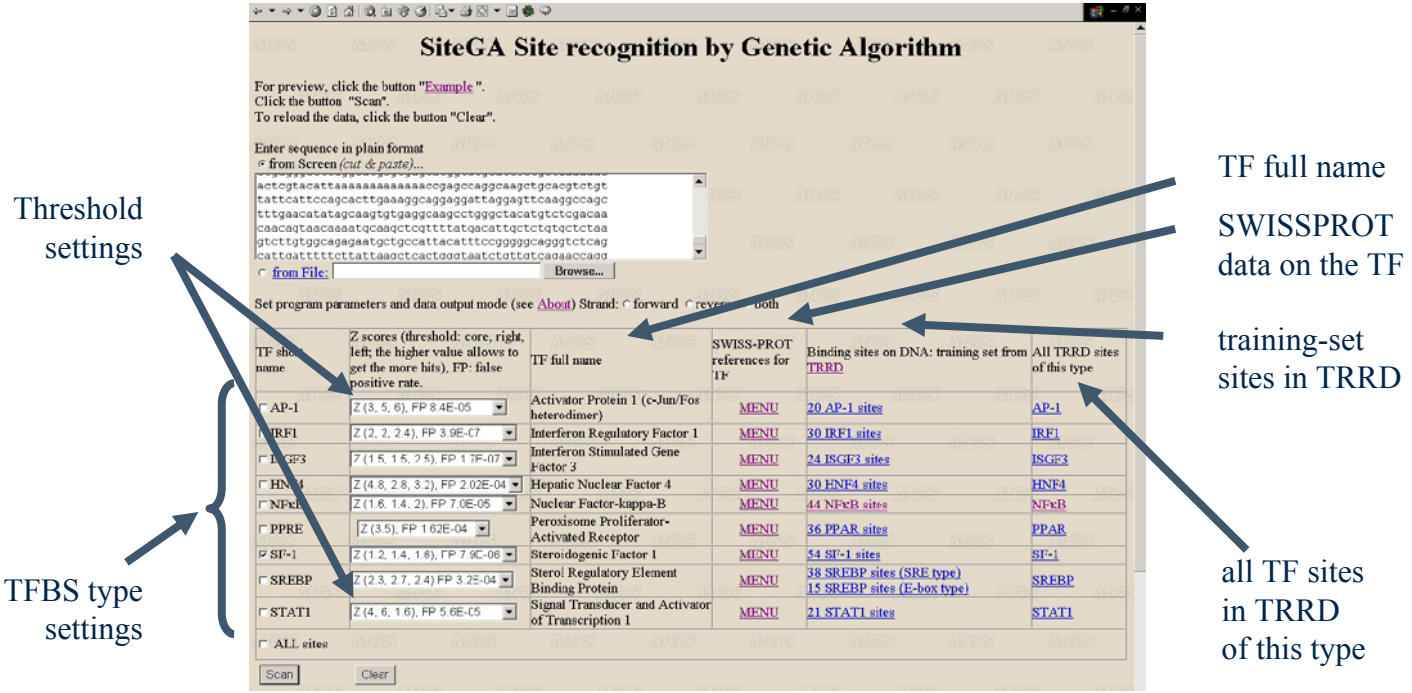
Benos, P.V., Bulyk, M.L., Stormo, G.D. 2002, Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res. 30(20):4442-4451.

Nevertheless the WM representation is severely limited by the assumption that positions in a site contribute independently to the total score. As a result, the major drawback of using WM for genome-wide TFBSs prediction is its high false-positive rate.

- *Barash,Y., Elidan,G., Friedman,N., Kaplan,T. (2003) Modeling Dependencies in Protein-DNA Binding Sites. In Vingron,M., Istrail,S., Pevzner,P. and Waterman,M. (eds), Proceedings of the Seventh Annual International Conference on Computational Molecular Biology. ACM, New York , 28–37.*
- *Zhou,Q., Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics, 20, 909-916.*
- *Pudimat R, Schukat-Talamazzini EG, Backofen R. A multiple-feature framework for modelling and predicting transcription factor binding sites.Bioinformatics.2005, 21(14),3082-8.*
- *King OD, Roth FP. A non-parametric model for transcription factor binding sites.Nucleic Acids Res. 2003, 31(19):e116.*

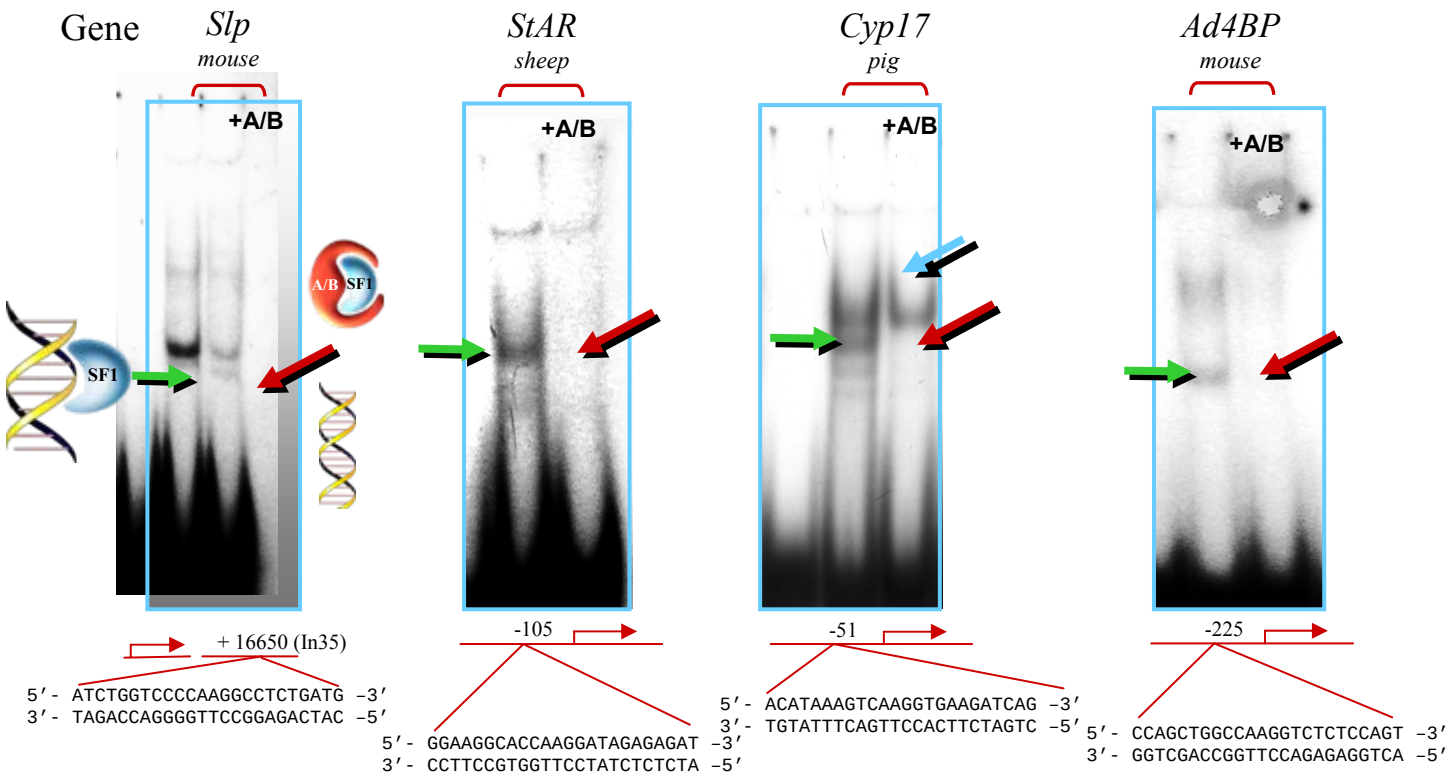
WWW interface of the SiteGA program

http://wwwmgs2.bionet.nsc.ru/mgs/programs/sitega/



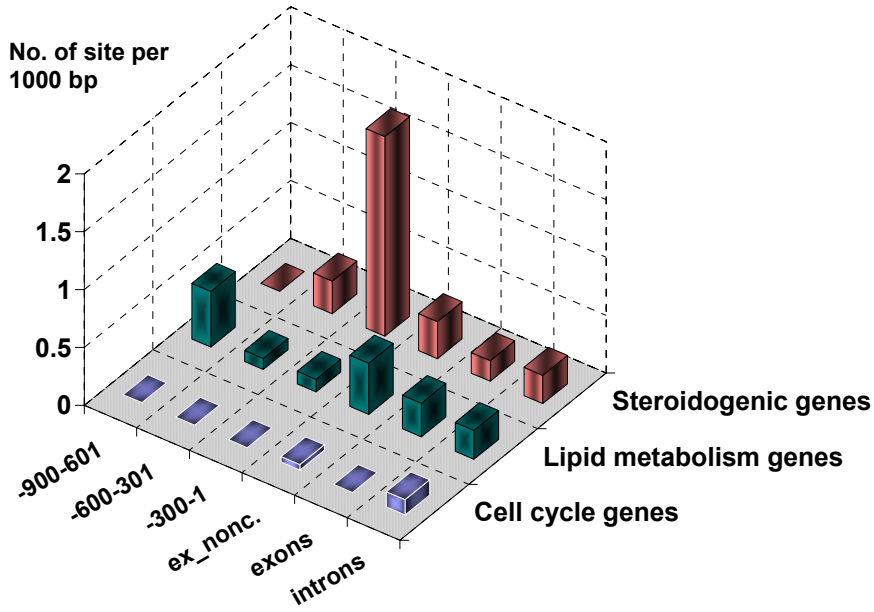
The SiteGA web interface allows the user to select a subset of recognition methods, so that the output provides results for every TFBS in turn. The recognition method use the Z-score settings as a thresholds. The interface also provides hyperlinks to the SWISSPROT data on the TF, the training-set sites and all TF sites of the same type in TRRD

Experimental support for the functional activity of the SF-1 sites predicted by SiteGA, by gel shift analysis (EMSA) with addition of antibodies against SF-1



- The shifted SF1/DNA complex
- Disappearance (or weakening) of the bands due to antibody against SF-1 (A/B) in the right lane supports SF-1 binding to the site.
- The complexes formed by the other proteins

SiteGA method results: SF-1 sites prediction in the 5'-flanking regions, exons and introns



The steroidogenic genes contained the highest number of predicted sites in the –300/-1 region. The lipid metabolism genes showed the densest SF-1 sites in the –900/-600 region, but, their number was more than 3 times smaller than in the –300/-1 region of the steroidogenic genes.

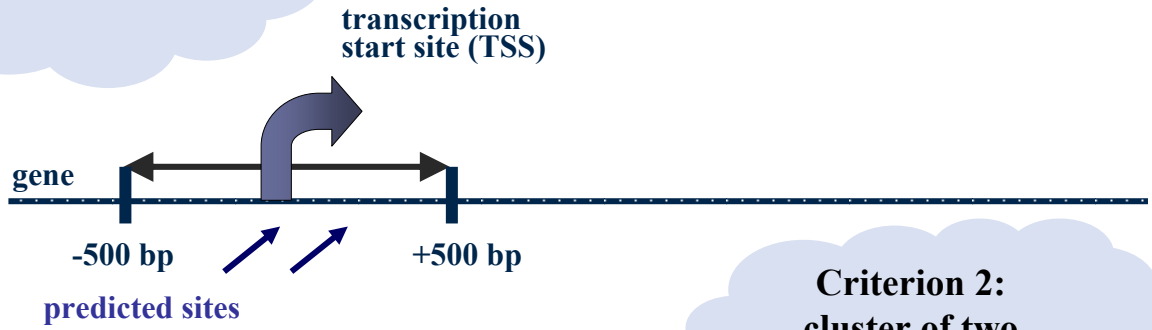
The content of the SF-1 sites in exons and introns of the steroidogenic and lipid metabolism genes was, on average, four times smaller than in the –300/-1 region of the steroidogenic genes

The comparatively high prediction level for the SF-1 sites in the lipid metabolism genes is probably due to the associated identification of a number of sites for LRH1 (a close homolog of SF-1)

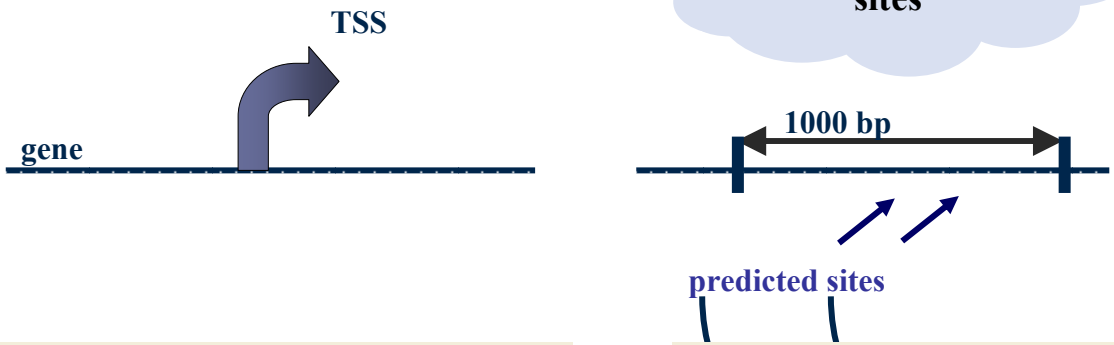
The extremely low level of predicted SF-1 sites was observed for non-coding exons and introns of the cell cycle genes. These genes contain no SF-1 sites both along the entire analyzed 5'regions (-900/-1) and in coding exons

Criteria for SF-1 regulated gene recognition in the human genome

Criterion 1: promoter



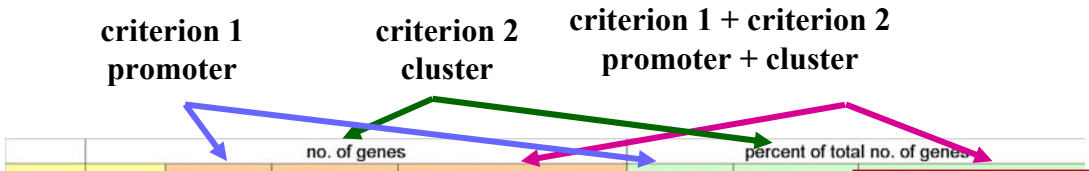
Criterion 2: cluster of two sites



Large window size (1000 bp) allows:
- partially set off the effect of bad TSS annotation
- take in consideration multiple TSS

Thus, we may obtain an upper estimate for the number of SF-1 regulated genes in human genome

Recognition of for SF-1 regulated genes in the human genome by both criteria



	no. of genes				percent of total no. of genes		
chromo-some	total analyzed	sites found in the region [-500;+500] relative to TSS	cluster of 2 sites anywhere, distance < 1000 bp	sites found in the region [-500;+500] relative to TSS + cluster of 2 sites anywhere, distance < 1000 bp	sites found in the region [-500;+500] relative to TSS	cluster of 2 sites anywhere, distance < 1000 bp	sites found in the region [-500;+500] relative to TSS + cluster of 2 sites anywhere, distance < 1000 bp
1	2578	314	977	216	12.2	37.9	8.4
2	1742	178	699	105	10.2	40.1	6.0
3	1360	141	596	99	10.4	43.8	7.3
4	1009	68	378	43	6.7	37.5	4.3
5	1176	82	434	55	7.0	36.9	4.7
6	1375	104	482	66	7.6	35.1	4.8
7	1364	137	545	106	10.0	40.0	7.8
8	914	86	354	53	9.4	38.7	5.8
9	989	86	400	60	8.7	40.4	6.1
10	956	92	433	63	9.6	45.3	6.6
11	1690	182	529	107	10.8	31.3	6.3
12	1247	109	508	72	8.7	40.7	5.8
14	985	60	288	46	6.1	29.2	4.7
15	873	104	364	73	11.9	41.7	8.4
16	1031	111	421	76	10.8	40.8	7.4
17	1354	167	493	111	12.3	36.4	8.2
18	397	32	169	22	8.1	42.6	5.5
19	1573	127	576	72	8.1	36.6	4.6
20	710	85	274	54	12.0	38.6	7.6
21	334	25	106	19	7.5	31.7	5.7
22	678	85	292	62	12.5	43.1	9.1
X	1085	65	311	37	6.0	28.7	3.4
Y	250	15	45	7	6.0	18.0	2.8

Total number of predicted SF-1 regulated genes in human genome ~1500 (5%)

2.6. ARGO:

A web system for the detection
of degenerate motifs and large-scale recognition of eukaryotic promoters

Reliable recognition of the promoters in the eukaryotic genomes remains an open issue. This is largely due to the poor understanding of the features of the structural–functional organization of the eukaryotic promoters essential for their function and recognition.

However, it was demonstrated that detection of ensembles of regulatory signals characteristic of specific promoter groups allows to increase the accuracy of promoter recognition and prediction of specific expression features of the queried genes.

The ARGO_Motifs package was developed for analysis of functional nucleotide sequences. It allows the recognition of oligonucleotide motifs with the following properties: (1) degeneracy, i.e. the use of the extended IUPAC code (A,T,G,C, R=G/A, Y=T/C, M=A/C, K=G/T, W=A/T, S=G/C, B=T/G/C, V=A/G/C, H=A/T/C, D=A/T/G, N=A/T/G/C); (2) region-specificity, i.e. the preferential occurrence in a certain region of a functional sequence; (3) quasi-invariance, i.e. the occurrence in certain sequence subgroups only; (4) contrast, i.e. much more frequent occurrence in functional than random sequences.

The ARGO_Viewer package was developed for recognition of tissue-specific gene promoters basing on the presence and distribution of oligonucleotide motifs obtained by the ARGO_Motifs program.

ARGO_Motifs description

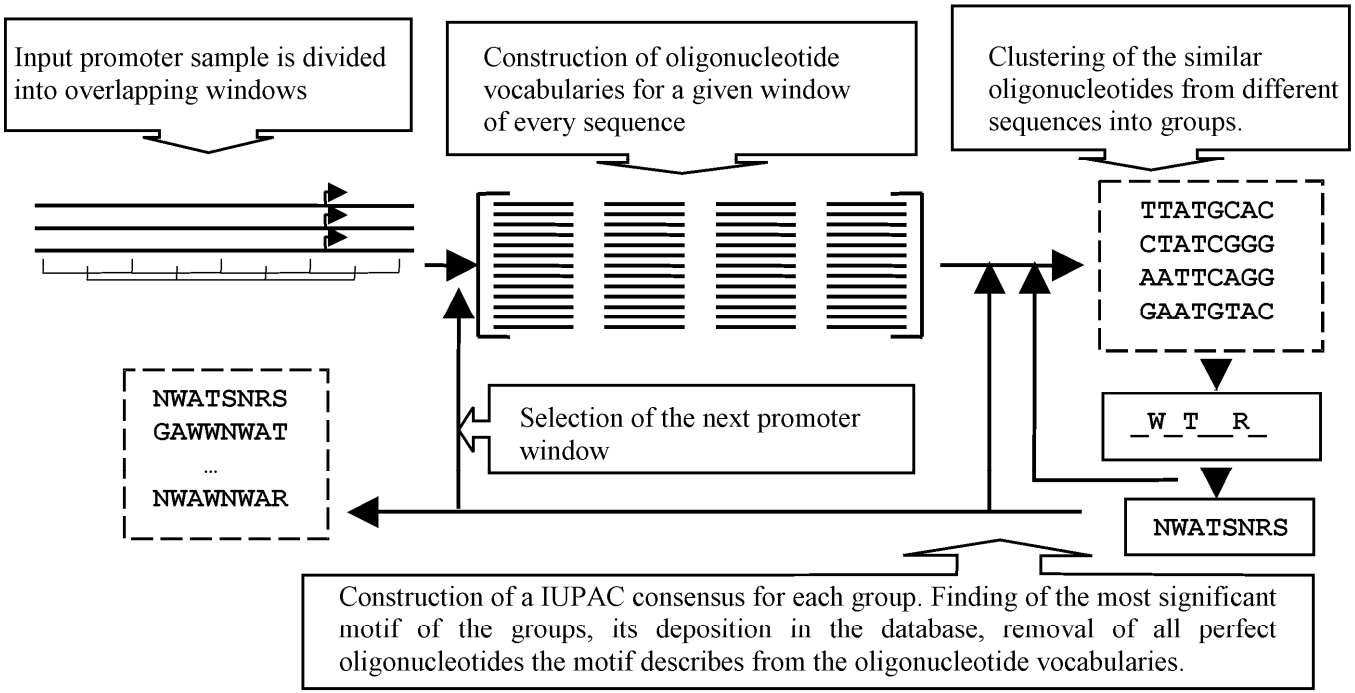
Search for degenerate motifs in a sample of functional sequences using the ARGO_Motifs program is implemented by grouping of similar perfect oligonucleotides from the oligonucleotide vocabularies corresponding to different sequences. Each oligonucleotide of the sequence vocabularies is considered, and group for each oligonucleotide is formed. A group consists of oligonucleotides belonging to the vocabularies of other sequences differing from it by not more than R positions (R<r0, where ro is the threshold similarity value). Then, the consensus in an extended IUPAC code is constructed for each oligonucleotide group using an iteration procedure. Each position of the consensus is occupied by the most significant of the 15 possible letters, their significance is estimated independently of each other using the binomial criterion. The obtained oligonucleotide motifs are regarded as significant, if they meet requirement. The significant motif that has the smallest probability to occur by chance is deposited in the databank, while all the perfect oligonucleotides it describes are removed from the vocabularies of the oligonucleotide sequences. The procedure for the detection of the motif ranking next in significance is applied in the same way to the modified vocabularies. The procedure is iterated until the detection of common degenerate motifs that satisfy condition is still feasible.

The degenerate oligonucleotide motif obtained using this procedure is considered significant, if it meets the following criteria:

$$\begin{cases} a) F > f_0 \\ b) P(n,N) < p_0 \\ c) Q < q_0 \end{cases}$$

Here, *F* is the proportion of promoters containing the motif in the window under analysis; *f0* is the threshold level of the motif occurrence in the promoter sample; *P(n, N)* is the probability of the accidental occurrence of the motif in the analyzed window in not less than *n* sequences of *N*; *p0* is the threshold probability level (see the estimation method below); *Q* denotes the proportion of sequences of the negative sample containing the motif, and *q0* is the threshold level of the motif occurrence in the negative sample. A set of 1000 randomly generated sequences of the length *L* is used as the negative sample. Thus, an oligonucleotide motif is accepted as significant, if (i) it occurs frequently in a promoter sample, (ii) infrequently in a sample of random sequences, and (iii) its occurrence probability by chance in a sample of promoter sequences is significantly low.

Layout of the algorithm for the recognition of degenerate oligonucleotide motifs in a promoter sample



ARGO_Viewer description

Tissue-specific promoters are recognized by the ARGO_Viewer program, in a scanning window sliding with a specified step along the genomic sequence analyzed. In every window, the corresponding region-specific oligonucleotide motifs obtained by the ARGO_Motifs are detected. Then, the similarity between the distributions of the motifs found in this window and in promoters of the groups studied is assessed. As a measure of similarity between the *j*th promoter and the sequence studied, the value

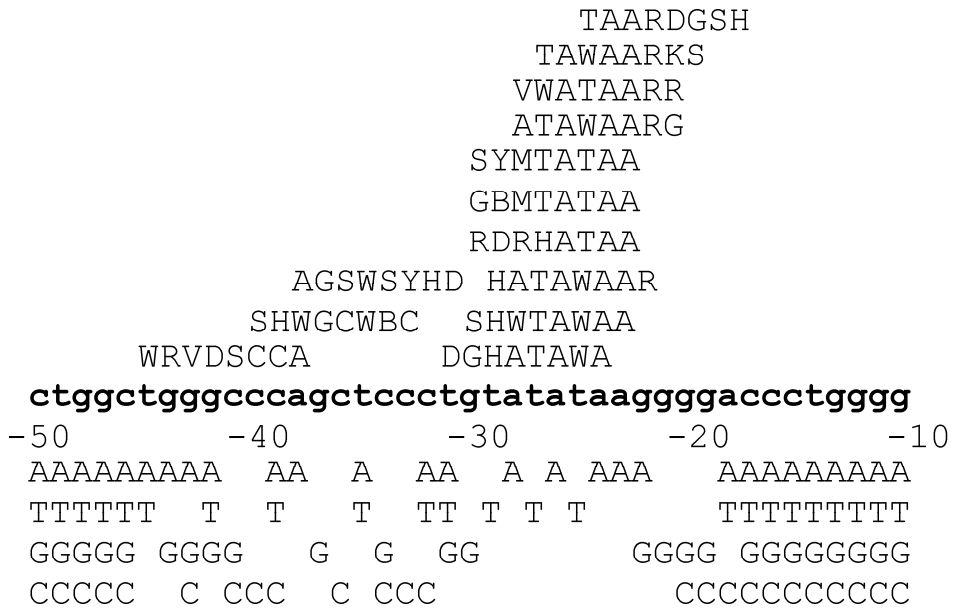
$$P_j = -\sum_{k=1}^{L} \log\ p_k / L$$

is used, where L is the size of the window analyzed and *pk* is the product of nucleotide frequencies consistent with the motifs covering the *k*th position.

The greater is *Pj*, the lower is the probability of chance occurrence of the motif set characteristic of the *j*th promoter in the sequence.

Thus, the promoter displaying the maximum value of the similarity function is found. If this value exceeds a certain threshold value, it is thought that the promoter of the considered group is identified in the window.

Example of determination of the set of permissible nucleotides for each position of the [–50; -10] region of an erythroid-specific promoter



Location of region-specific oligonucleotide motifs on the sequence

Sequence of the promoter region under examination

Set of allowed nucleotides

Description of Web-interface of the ARGO_Motifs program

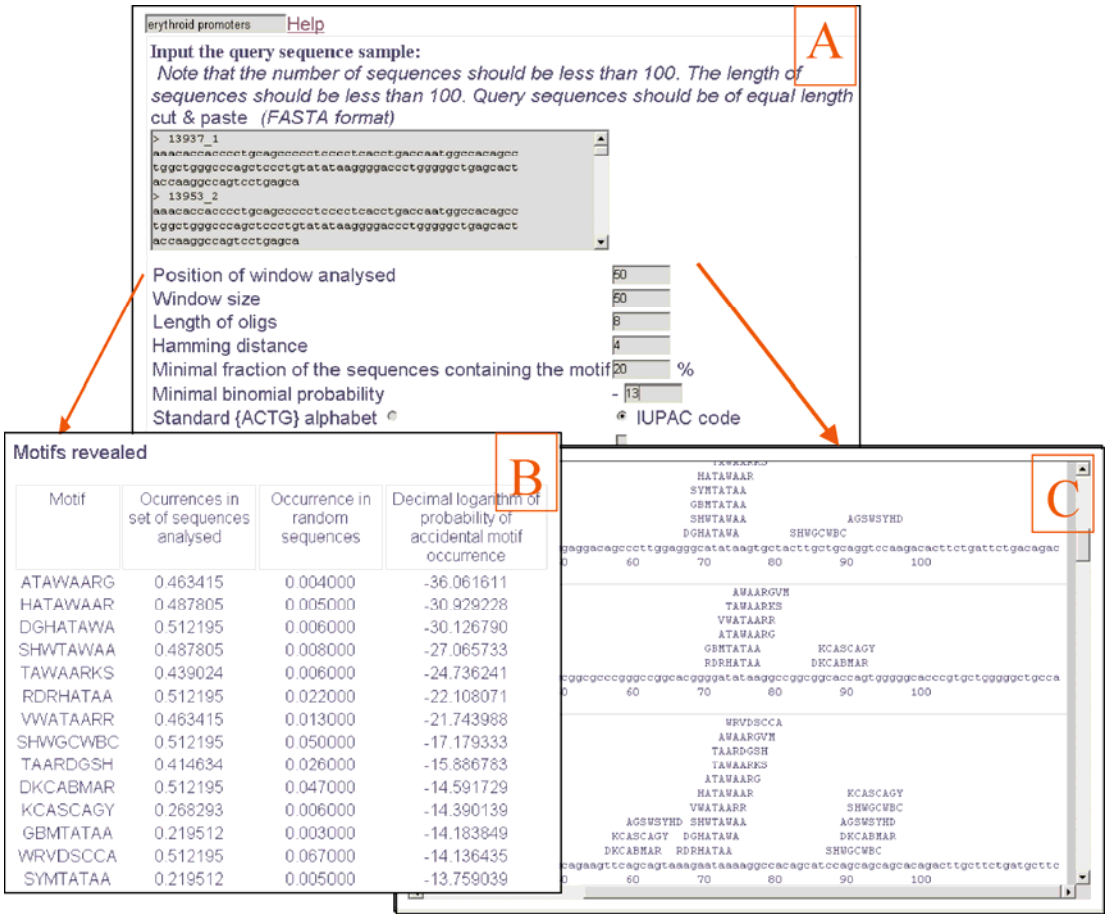
The public version of ARGO_Motifs (Figure 3a) is available at http://wwwmgs2.bionet.nsc.ru/argo/ and http://emj-pc.ics.uci.edu/argo/.

The user can paste a set of analyzed sequences of equal lengths in FASTA format via the sequence input. All the parameters needed for analysis are specified in the lower part of the window. The program was designed to search for region-specific motifs. Therefore, once the sample of DNA sequences is input, the user can analyze consecutively the regions of interest. In addition, the length of the motifs detected and the Hamming's distance, the degree of similarity between the perfect oligonucleotides clustered in a motif are indicated. The user can search for both perfect oligonucleotide motifs in the 4 single letter-based (A, T, G, and C) code and degenerate motifs in the 15 single letter-based IUPAC code. The program allows the motifs meeting the significance criteria to be found in both DNA strands. It is possible to specify for the motifs detected both the boundary value of binomial probability of their random occurrence in the examined sample and the threshold occurrence rate (%) of a motif, i.e., the fraction of analyzed sequences containing of the motif.

The results of the sequence analysis are displayed as a table containing the motifs detected and their characteristics. As an example, Figure shows the motifs found in the [–50; +1] region of promoters of erythroid-specific genes. The motifs of length l = 8 meeting the below parameters of condition (1) were considered significant: P(n, N) < 10–13; f0 = 20%; and q0 = 100%. As an example, let us consider the first oligonucleotide listed in Figure : ATAWAARG = (A)(T)(A)(A/T)(A)(A)(A/G)(G), found in the [–50: +1] region relative to the transcription start. This motif was found in 19 promoters of 41 (46%), exceeding the threshold (20%) approximately twofold. The random occurrence probability of this motif in 19 or more of the 41 promoters is 10–36. In the negative sample, this motif occurred in the queried region only in 4 random sequences of 1000 (0.4%). Hence, this motif meets the significance criteria.

In addition to the table output mode, the user can get a distribution pattern of the found motifs in the selected window of the sample analyzed (Figure). This representation may be useful for detection of ensembles of mutually present motifs and subgrouping of the sequences of the total sample.

Example of ARGO_Motifs input and output windows



Motif	Ocurrences in set of sequences analysed	Occurrence in random sequences	Decimal logarithm of probability of accidental motif occurrence
ATAWAARG	0.463415	0.004000	-36.061611
HATAWAAR	0.487805	0.005000	-30.929228
DGHATAWA	0.512195	0.006000	-30.126790
SHWTAWAA	0.487805	0.008000	-27.065733
TAWAARKS	0.439024	0.006000	-24.736241
RDRHATAA	0.512195	0.022000	-22.108071
VWATAARR	0.463415	0.013000	-21.743988
SHWGCWBC	0.512195	0.050000	-17.179333
TAARDGSH	0.414634	0.026000	-15.886783
DKCABMAR	0.512195	0.047000	-14.591729
KCASCAGY	0.268293	0.006000	-14.390139
GBMTATAA	0.219512	0.003000	-14.183849
WRVDSCCA	0.512195	0.067000	-14.136435
SYMTATAA	0.219512	0.005000	-13.759039

A. Input window. The region [–50; +1] of promoters of erythroid-specific genes is analyzed.

B. A table containing the motifs detected and their characteristics.

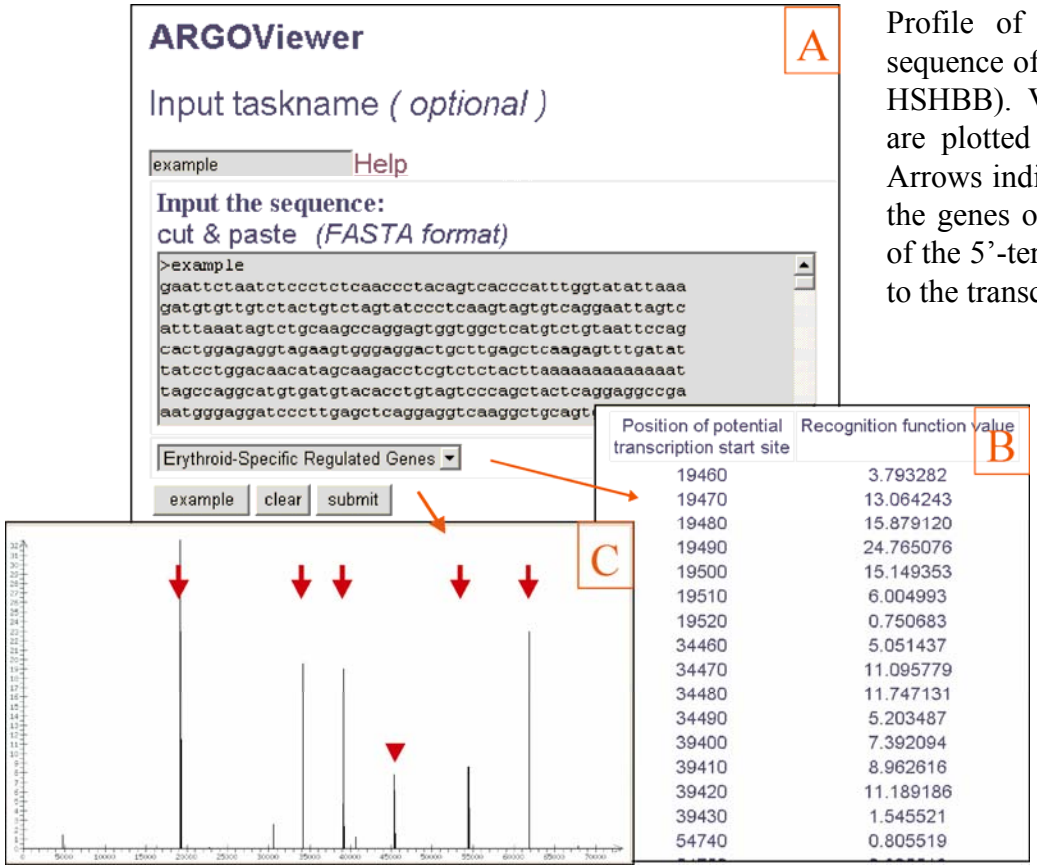
C. A distribution pattern of the found motifs.

Web-interface of the ARGO_Viewer program

The ARGO_Viewer package was developed for recognition of tissue-specific gene promoters on the basis of the presence and distribution of oligonucleotide motifs obtained by the ARGO_Motifs program. The public version of the ARGO_Viewer is available at http://wwwmgs2.bionet.nsc.ru/argo/ and http://emj-pc.ics.uci.edu/argo/.

The user can paste the genomic sequence analyzed in FASTA format into the sequence input box. The class of promoters to be searched for is specified at the bottom of the window. The program provides the search for promoters in both the direct and complementary DNA strands. Furthermore, two modes of output recognition results are provided. In the case of text mode, the user gets a list of positions of potential transcription starts. In graphic mode, the program constructs the profile of recognition function. The program implementation is illustrated by the example of human β-globin region (ID HSHBB), of 73308 bp in length, mapped on chromosome 11. This sequence contains five experimentally detected transcription start sites at positions 19487, 34478, 39414, 54740, and 62137 together with the promoter region of a pseudogene in the vicinity of position 45557.

Predicted positions of the transcription starts in five genes of this cluster differed from the real starts by not more than 20 bp. Therefore the proposed procedure provides high efficiency of promoter recognition.



Profile of the promoter recognition function for the sequence of the human β-globin gene clusters (EMBL ID: HSHBB). Values of the recognition function (ordinate) are plotted versus positions of the sequence (abscissa). Arrows indicate the positions of the transcription starts of the genes of this cluster. The triangle shows the position of the 5'-terminal region of the pseudogene corresponding to the transcription start point.

Characteristics of the motifs, specific for transcription factor SF1 binding sites

Set of 22 sequences, containing the consensus of SF1 binding site in direct orientation

```
5′... ggcaggagttCAAGGTaataagggctgaga ...3′
3′... ccgtcctcaaGTTCCAttattcccgactgt ...5′

5′... ctgagtctccCAAGGTcatccttgttttga ...3′
3′... gactcagaggGTTCCAgtaggaacaaaagt ...5′
                    ...
5′... gacatttattCAAGGTaatgataacaatct ...3′
3′... ctgtaaataaGTTCCAttactattgttaga ...5′

5′... cttcccggccCAAGGTccacttgcttgctt ...3′
3′... gaagggccggGTTCCAggtgaacgaacgaa ...5′
```

37 motifs revealed

Motif	Region of the site	Presence	P(n,N)
WYTNYCAS	-45: -25	0.36	10^{-9}
CNGSMNCT	-30: -10	0.36	10^{-9}
NYCAAGGY	-10: +10	0.68	10^{-31}
RAGGTCMN	-10: +10	0.68	10^{-31}
CAWGGYNM	-10: +10	0.45	10^{-15}
AAGGTCNN	-5: +15	0.45	10^{-18}

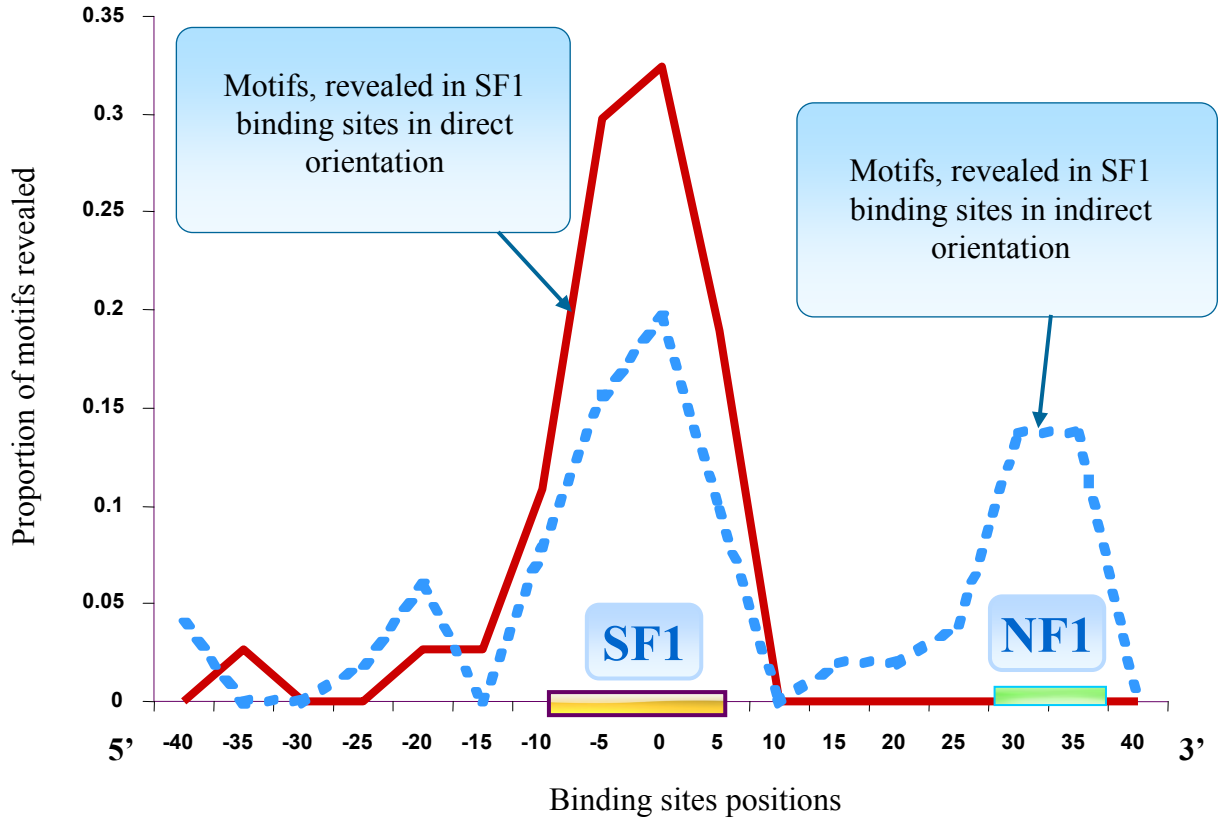
Set of 23 sequences, containing the consensus of SF1 binding site in antisense chain in direct orientation

```
5′... ttctcacttaGCCTTGagctggtgattata ...3′
3′... aagagtgaatCGGAACtcgaccactaatat ...5′

5′... tcctctcttaGCCTTGagctagttagtggt ...3′
3′... aggagagaatCGGAACtcgatcaatcacca ...5′
                    ...
5′... tttctaaattGCCTTGaccactgcttctcc ...3′
3′... aaagatttaaCGGAACtggtgacgaagagg ...5′

5′... attggaagtaACCTTGactagctgagctca ...3′
3′... taaccttcatTGGAACtgatcgactcgagt ...5′
```

51 motifs revealed

Motif	Region of the site	Presence	P(n,N)
GGNGGAGG	-50: -30	0.21	10^{-8}
KKKGNGAG	-50: -30	0.30	10^{-8}
NRDCCTTG	-10: +10	0.69	10^{-30}
CCTTGWCN	-10: +10	0.52	10^{-22}
YCYRGRKN	+20: +40	0.47	10^{-11}
RYYCWGGN	+25;+45	0.34	10^{-9}

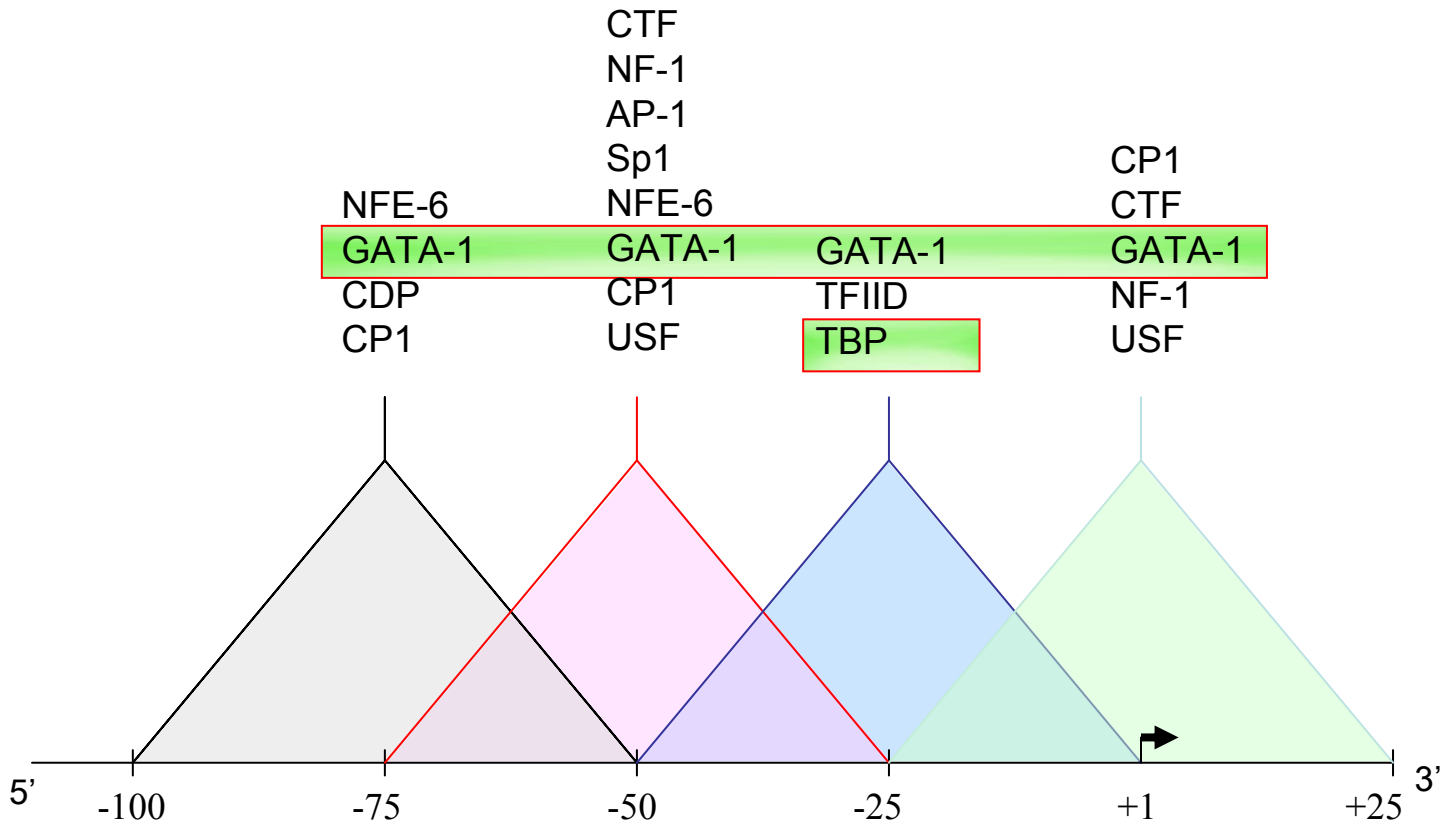
Distribution of the number of motifs revealed along the transcription factor SF1 binding sites



Characteristics of the motifs, revealed in the erythroid – specific promoters

Motif	Region of promoter	Presence in promoters	Presence in random sequences	P(n,N)
RCCAATND	-100: -50	0.59	0.02	$10^{-29.4}$
CCAAT-box, usual for erythroid – specific promoters				
TGACCAAT	-100: -50	0.35	0.00	$10^{-33.45}$
NTCASCAK	-75: -25	0.21	0.00	$10^{-8.77}$
CAGCMNDD	-75: -25	0.59	0.05	$10^{-16.6}$
GRSSNCAG	-75: -25	0.51	0.03	$10^{-15.9}$
ATAWAARG	-50: +1	0.48	0.00	$10^{-35.5}$
TATA-box, binding site of TBP				
DGNATAWA	-50: +1	0.59	0.01	$10^{-30.7}$
CTTCTGRN	-25:+20	0.40	0.00	$10^{-25.9}$
AAGGCCAN	-25:+20	0.24	0.00	$10^{-15.07}$

Examples of transcription factors binding sites, similar to the region – specific motifs, revealed in erythroid – specific promoters



Recognition of promoters of tissue–specific genes

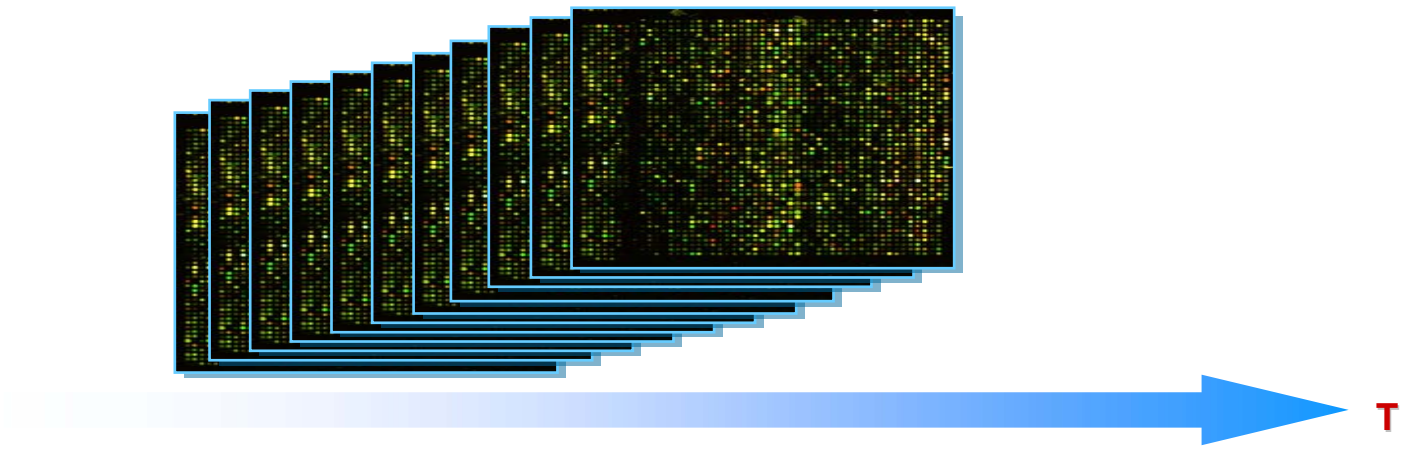
Program	False negative	False positive
ARGO (http://wwwmgs2.bionet.nsc.ru:8080/argo/)	0	$<1.4*10^{-5}$
TSSW (http://genomic.sanger.ac.uk/gf/gf.shtml)	0	$8.1*10^{-5}$
TSSG (http://genomic.sanger.ac.uk/gf/gf.shtml)	0.4	$1.4*10^{-4}$
NNPP (http://www.fruitfly.org/seq_tools/promoter.html)	0	$1.2*10^{-3}$
Proscan (http://biosci.umn.edu/software/proscan/promoterscan.htm)	0.4	$1.6*10^{-4}$

Set of promoters	Number of seqs	Number of motifs	False negative	False positive
Endocrine system genes	78	814	0.05	$<10^{-5}$ 1 per 100000 nuc.
Heat - shock genes	34	45	0.09	$2.3*10^{-4}$ 1 per 4348 nuc.
Interferon - inducible genes	41	131	0.07	$<10^{-5}$ 1 per 100000 nuc.
Genes of lipid metabolism	50	281	0.04	$<10^{-5}$ 1 per 100000 nuc.
Erythroid – specific genes	26	78	0.08	$\sim10^{-5}$ 1 per 100000 nuc.

Chapter 3
COMPUTATIONAL TRANSCRIPTOMICS

3.1. Large scale analysis of gene expression profiles by means of DNA microarrays

Large scale analysis of gene expression profiles by means of DNA microarrays



Identification of cell cycle genes on the basis of statistical analysis of the profiles of expression of human genes

The cells HeLa S3 were synchronized by arresting of the cell cycle at the S-stage by using double timidine block or at the stage of mitosis, by using timidin-nocalasole block.
Expression profiles for ~ 13 000 genes (42 000 clones) were estimated by DNA-arrays and accumulated in the database SMD (Stanford Microarray Database, http://genome-www5.stanford.edu/MicroArray/SMD/helpindex.html)
Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Whitfield M.L. et al., Mol. Biol. Cell. 2002 Jun;13(6):1977-2000.

Tools for statistical analysis of expression profiles of human genes:

monotonous increase/decrease

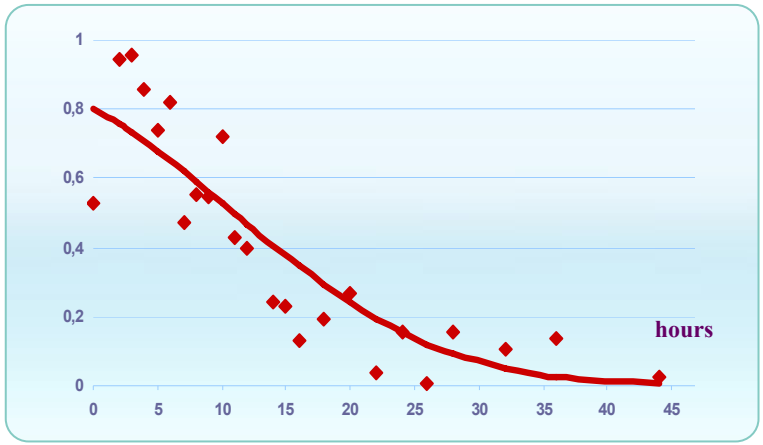
By statistical methods developed by us we have determined the groups of genes with the similar profiles of expression (recurrence, increase, decrease, etc.). Re-normalizing of initial profiles Z by PROBIT-rearrangement

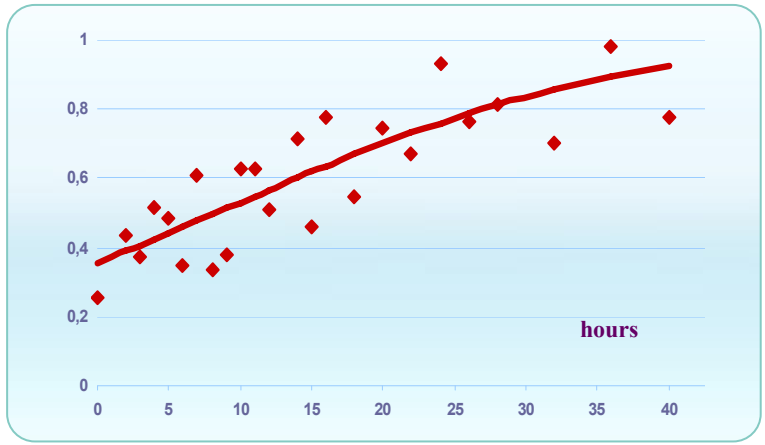
$$\text{PROBIT}= \Phi^{-1}(Z) \quad \text{where} \quad \Phi(x) := (2\pi)^{-1/2} \int_{-\infty}^{x} \exp(-r^2/2)\, dr,$$

Enabled to bring them to the linear form $\text{PROBIT}=a0 + b0 \times t + \varepsilon_t$

Among 42,000 profiles with high significance level (p-value<0.0001), we have found 864 monotonous curves.

Example of genes with monotonously varying profiles of expression





Tools for statistical analysis of the profiles of human gene expression:

Cyclically expressed genes

$$Y(t) = m0 + A \times \cos(2\pi \times t / T + \varphi) + \varepsilon t .$$
$$Y(t) = m0 + A1 \times \cos(2\pi \times t / T) + A2 \times \sin(2\pi \times t / T) + \varepsilon t ,)$$

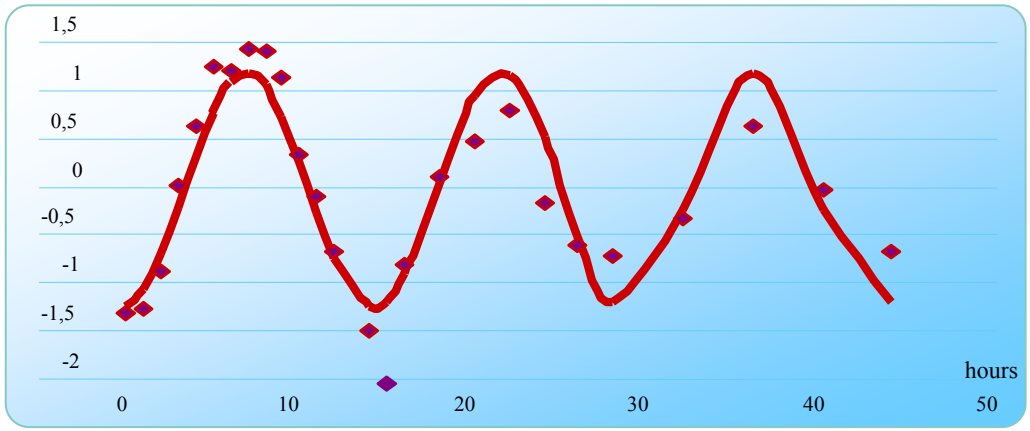
where

$A1=\cos \varphi$, $A2=-\sin \varphi.$;

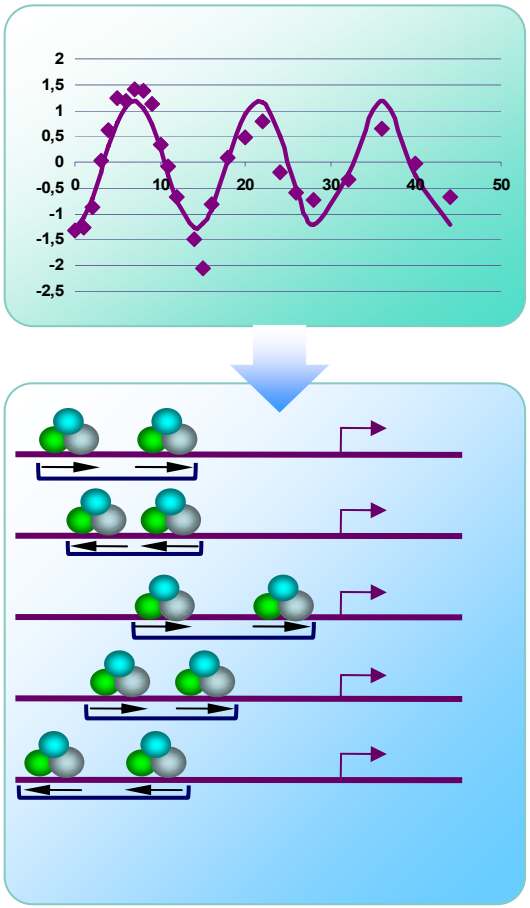
$A = (A12 + A22)–1/2$, $\varphi = \text{arctg}\ (–A2 / A1)$.

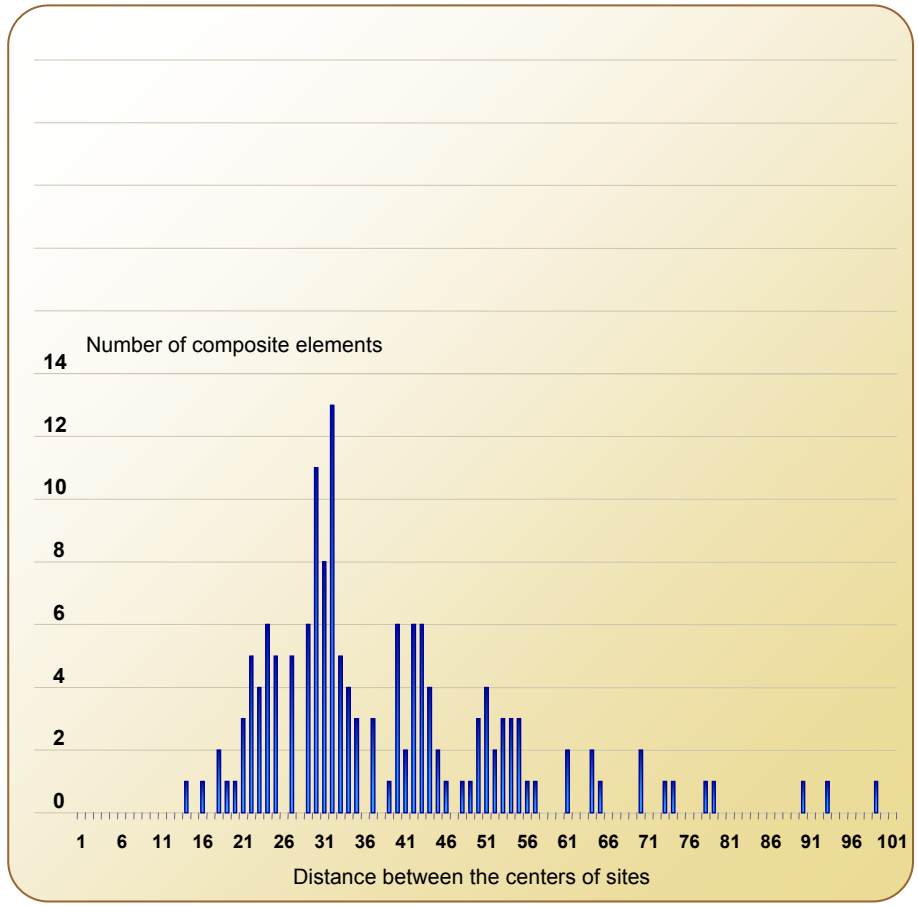
Among 42,000 profiles with high level of significance (p-value<0.0001), we have revealed 4485 periodical curves corresponding to ~ **2000 genes**.

AN EXAMPLE OF A GENE WITH CYCLICALLY VARYING LEVEL OF EXPRESSION

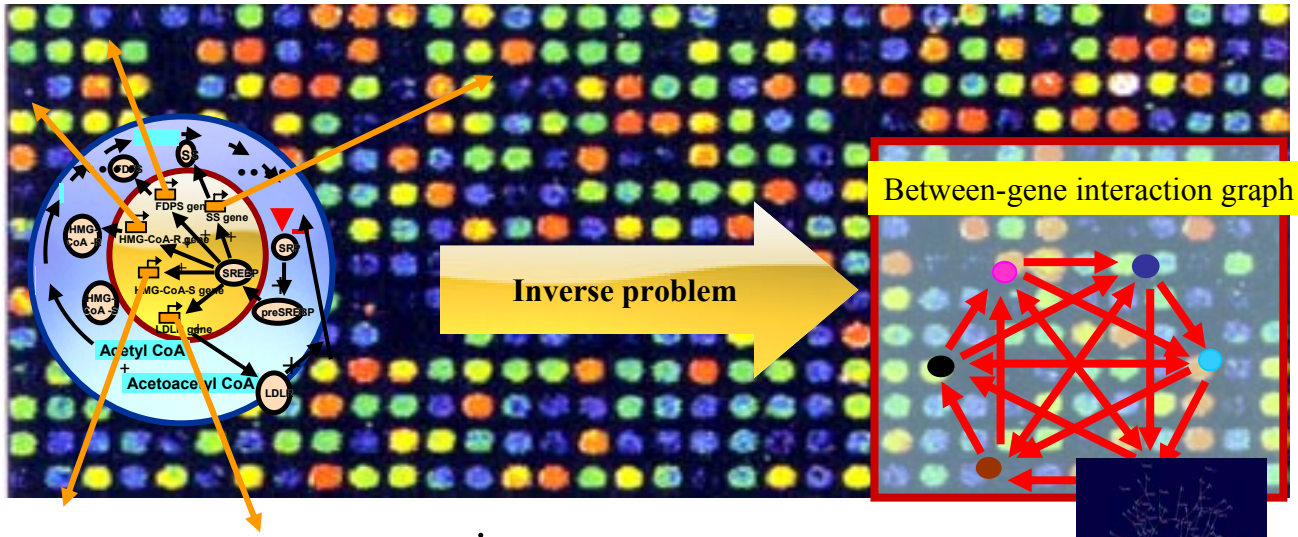


In 5`-regions of 96 cyclically expressed human genes, we have revealed potential composite element NF-Y/NF-Y



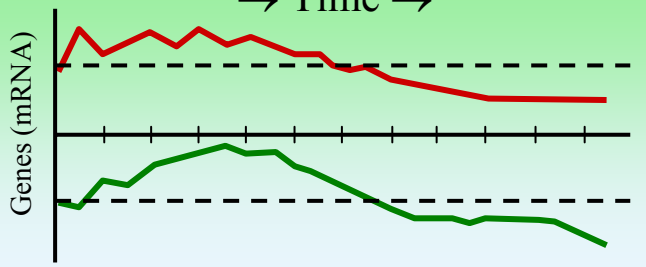


Computer-based technologies for gene network graph reconstruction using Microarray Analysis data

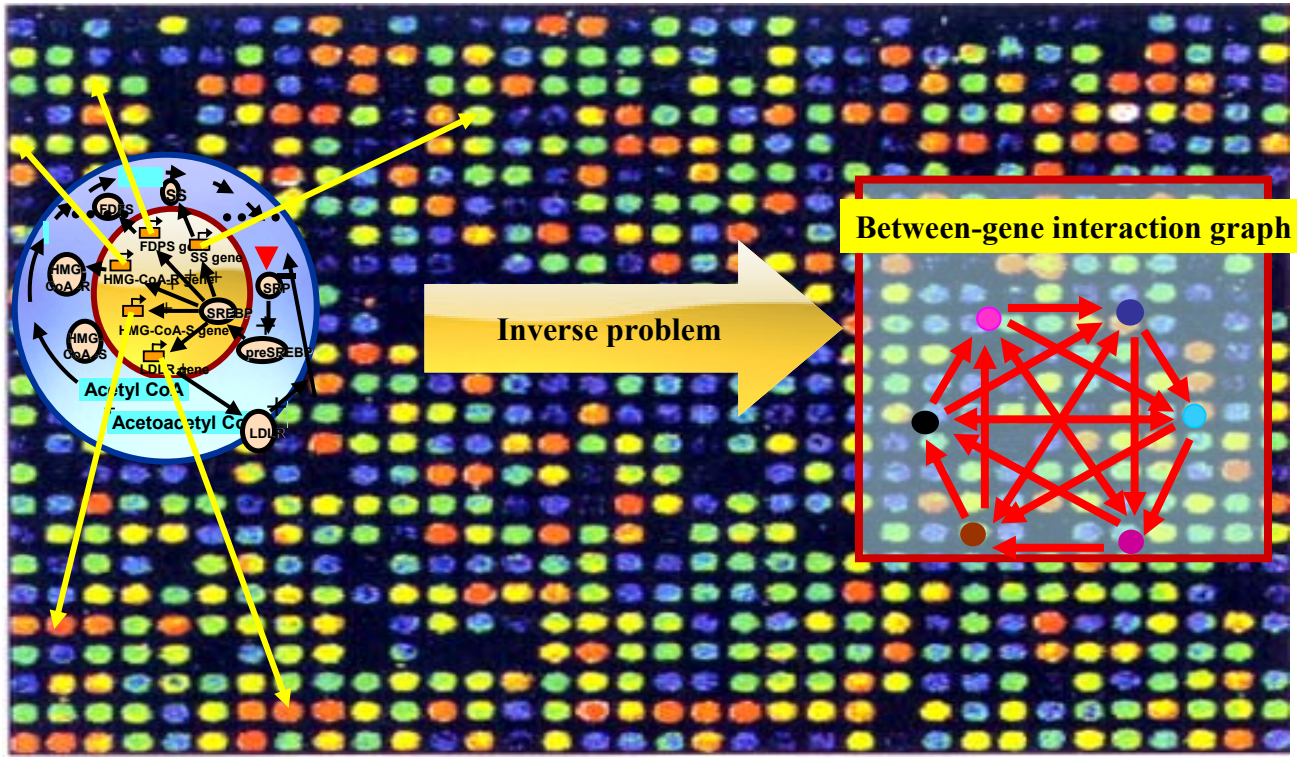


Initial data: x_i^j $\Leftrightarrow$ i-th gene concentration $\Leftrightarrow$ at j-th time point

$$X_{N\times M} := \begin{pmatrix} \rightarrow \text{Time} \rightarrow \\ \chi_1^1 & \chi_1^2 & \cdots & \chi_1^M \\ \chi_2^1 & \chi_2^2 & \cdots & \chi_2^M \\ \vdots & \vdots & \ddots & \vdots \\ \chi_N^1 & \chi_N^2 & \cdots & \chi_N^M \end{pmatrix}$$



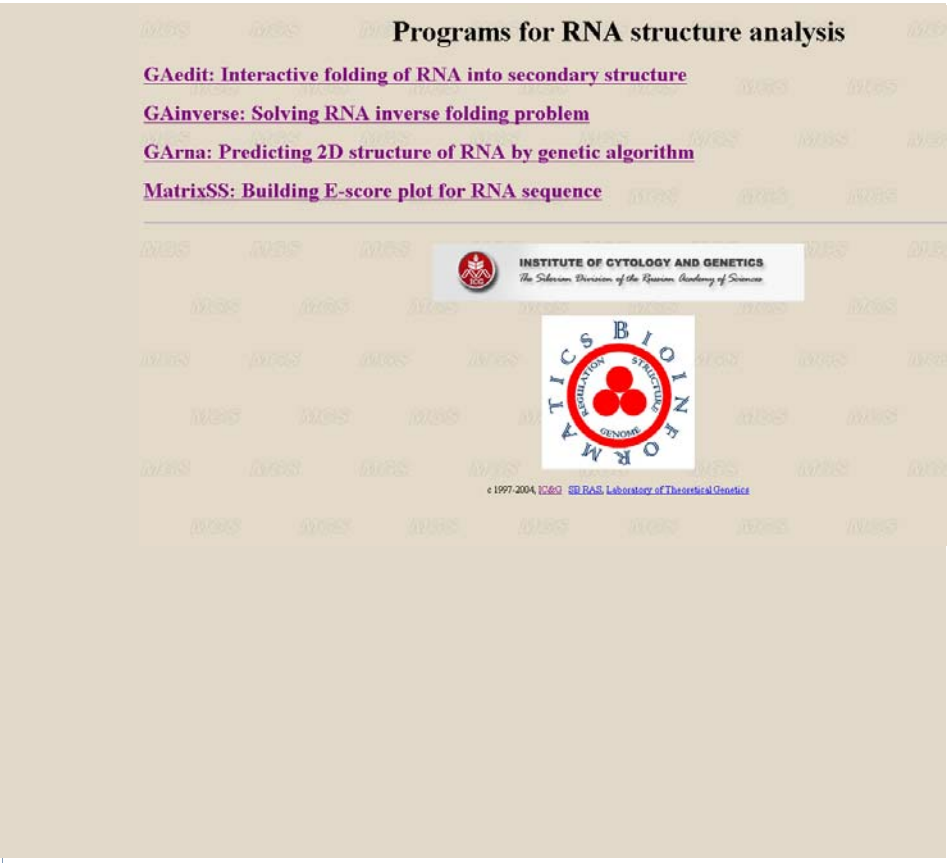
In silico design of high-selectivity DNA microchips to diagnose multifactorial diseases



3.2. Web-server GArna for RNA structure analysis: it's state in 2004

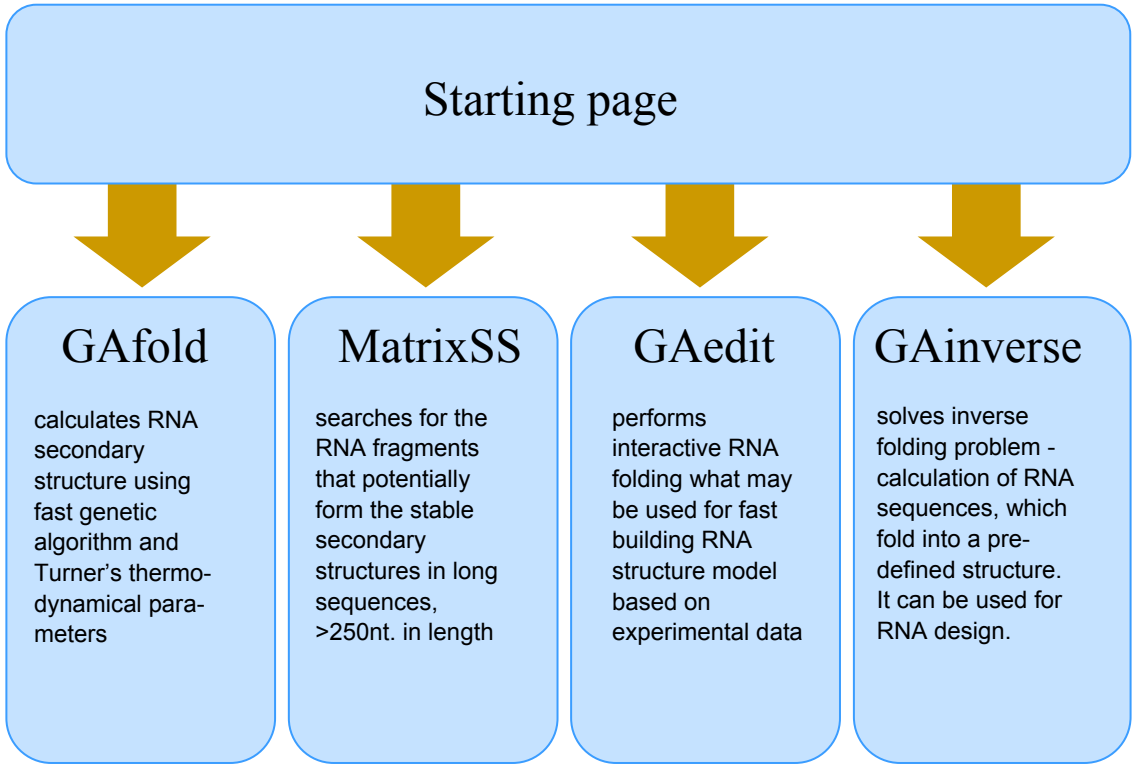
http://wwwmgs2.bionet.nsc.ru/mgs/systems/garna

Starting page with links to our programs



Introduction

A number of practical problems of biotechnology (search for non-coding RNAs, RNA design, oligonucleotide hybridization efficiency and RNA structure prediction, etc.) address RNA secondary structure analysis. Some of the computational tools solving these problems can be reached at the mfold and Vienna sites (Zuker, 2003; Hofacker, 2003), both based on dynamic calculations. Our programs for RNA secondary structure analysis use fast genetic algorithm (Titov et al., 2002). Besides the common features such as RNA structure calculation and drawing, our site allows to search for secondary structure in long RNA sequences and some other unique features. The GArna server provides access to four programs: GAfold, MatrixSS, GAinverse and GAedit. Here we describe these programs. All of them can be accessed from a starting page, where each program has a short description and a user manual.



Motivation: Fast web-servers based on RNA secondary structure prediction programs can be very helpful for solving the contemporary biotechnological problems.

Results: We present a new version of GArna server where our programs for RNA secondary structure analysis can be accessed via Internet. Currently the server provides the following capabilities: (i) solution of direct and inverse folding problems, prediction of oligonucleotide hybridization; (ii) analysis of secondary structure characteristics and (iii) RNA secondary structure editor for interactive folding with constraints.

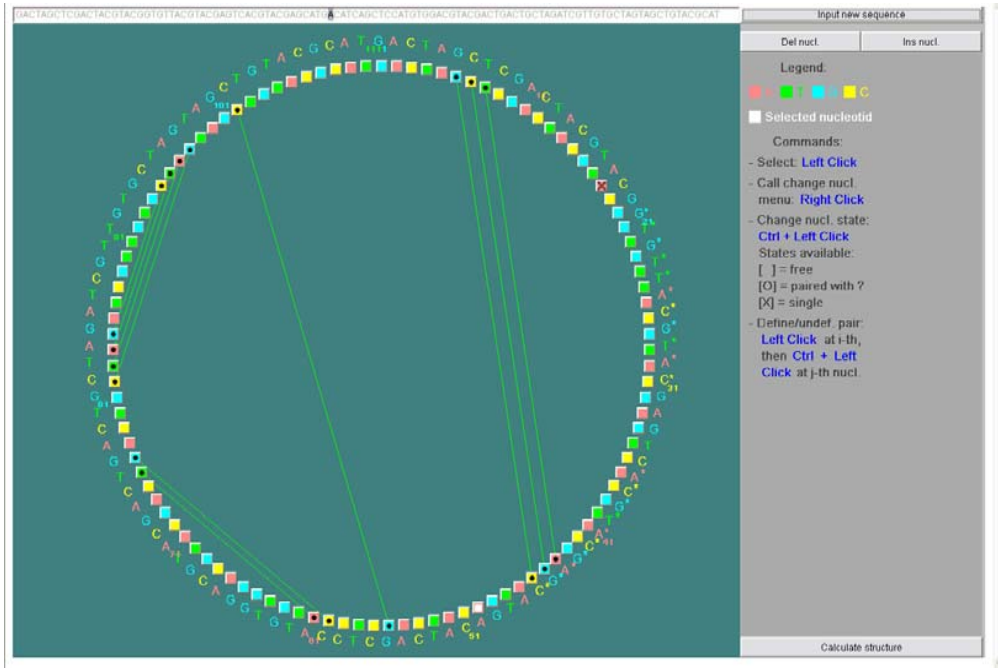
GAedit: Editor Mode

GAedit

This program performs interactive RNA folding what may be used for fast building RNA structure model based on experimental data. The program presents point-and-click graphic interface for input of restrictions on secondary structure.

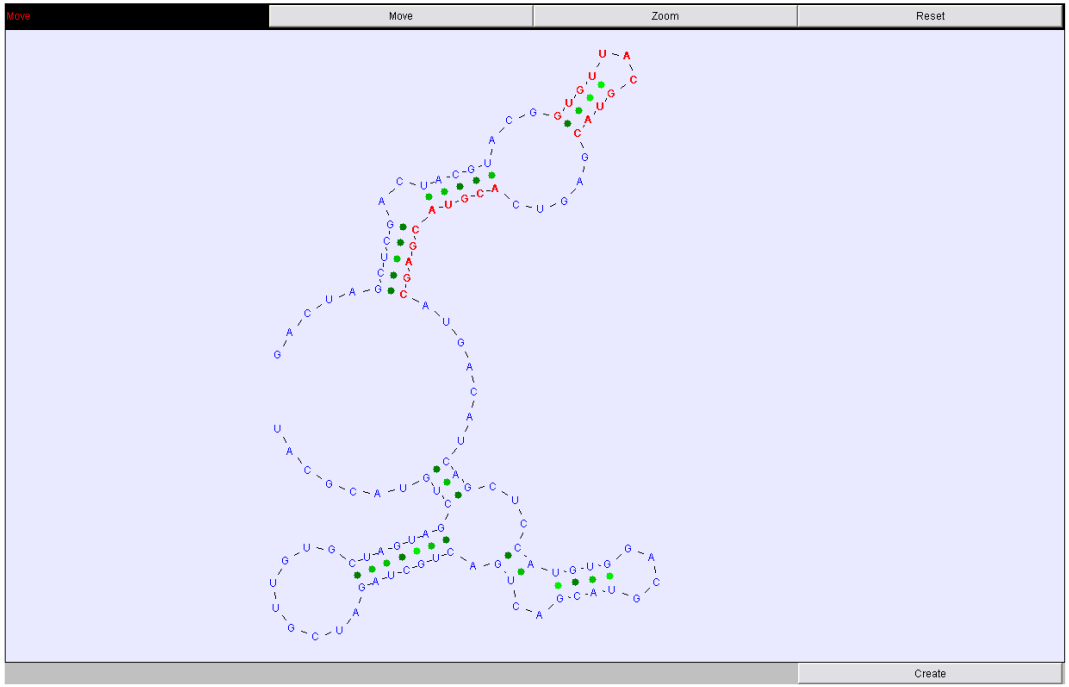
These constraints are of three different types:

1. Pairing of complementary nucleotides i and j.
2. Pairing of a certain nucleotide with unknown one.
3. The requirement, that a given nucleotide should be single-stranded.

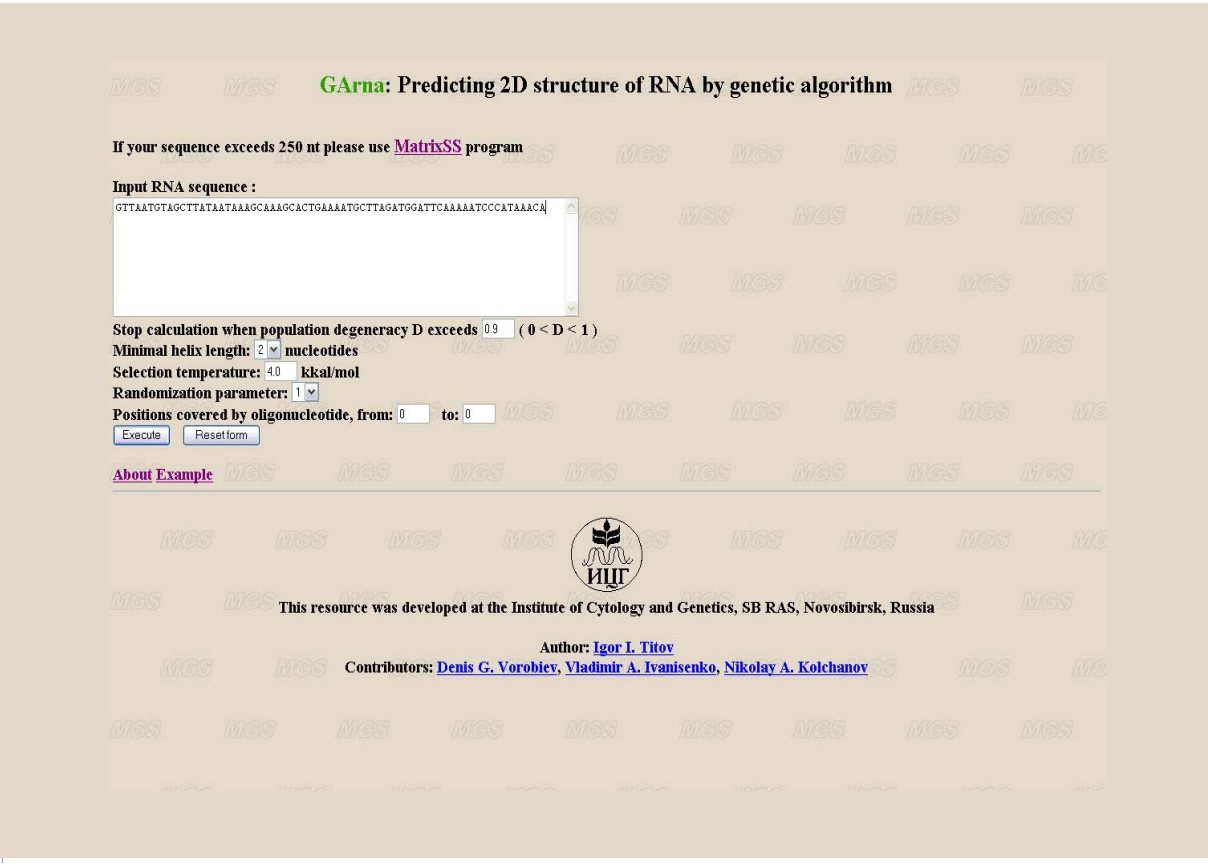


Editor's window includes the area for input of RNA sequence, a set of control items and RNA nucleotides, located along the circle Pressing button "Calculate structure" runs calculation of RNA secondary structure and then displays it in the applet window. Pressing button "Create" returns to the editor mode.

GAedit: Secondary structure display



GAfold input window



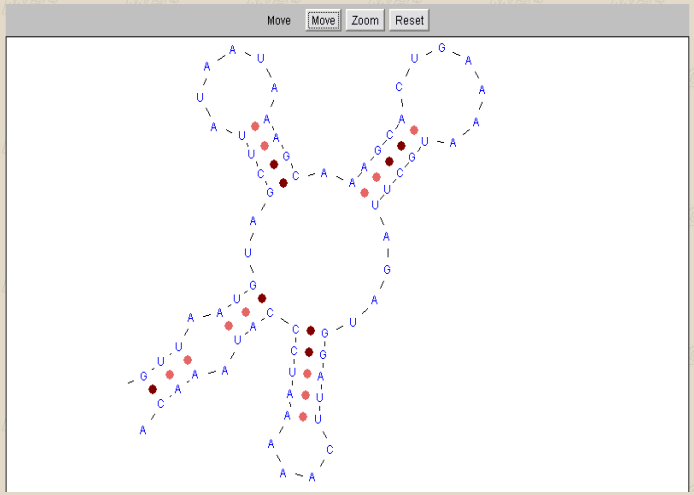
GAfold

The basic module of the server is GAfold – the program for calculation of RNA secondary structure using fast genetic algorithm (Titov et al., 2002) and Turner's thermodynamical parameters at physiological conditions (Jaeger et al., 1989). The programs GAinverse and GAedit also use this unit. GAfold calculates one of low-energy secondary structures for RNA sequences up to 250 nucleotides in length. The alternative structures can be found by changing the parameters of calculation. The program output is RNA secondary structure either in the graphical or textual representations.

http://wwwmgs.bionet.nsc.ru/mgs/programs/2dstructrna/

GAfold results

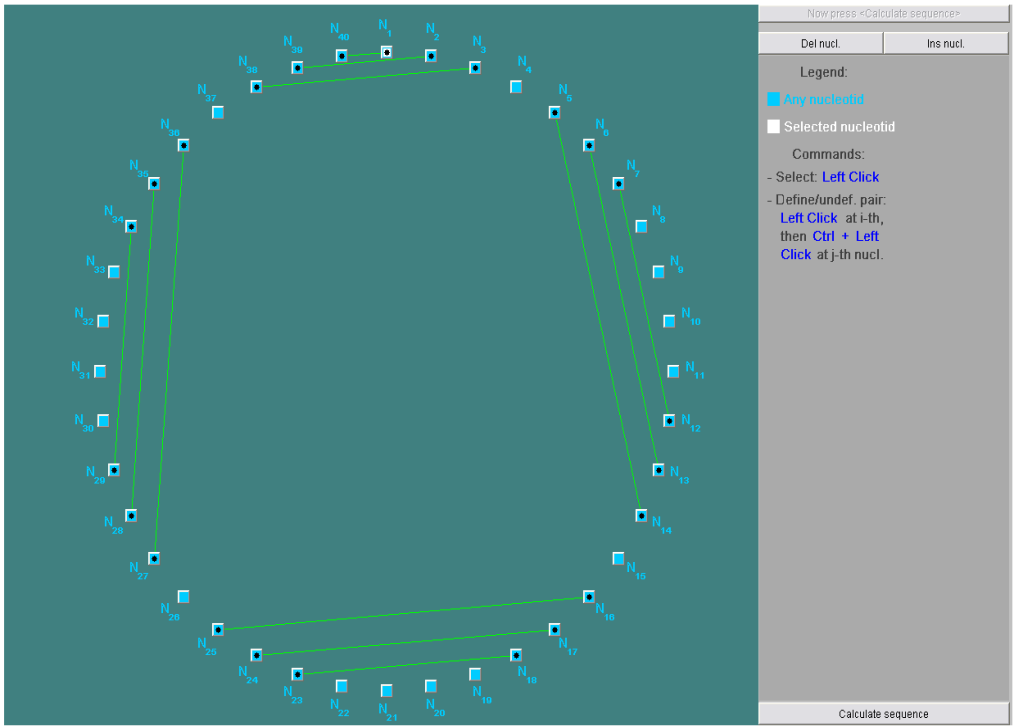
```
Your sequence has the optimal structure with

Energy = -8.4 kcal/mol

with a z-score of -2.94, which means INCREASED stability compared
to random sequences of the same length and nucleotide composition
(expected Energy = -3.1 kcal/mol, standard deviation = 1.8 kcal/mol)

Secondary structure in MFOLD .ct file format:

 67  ENERGY = -8.4  GA-generated
  1 G       0    2   66    1
  2 U       1    3   65    2
  3 U       2    4   64    3
  4 A       3    5    0    4
  5 A       4    6   62    5
```



Besides the structure itself, the GAfold program outputs a sequence potential of secondary structure formation. This potential, Z-score of the sequence structure energy, is characterized by comparing a given sequence with RNAs containing the same nucleotides, but going in random order. Z-score removes the compositional effect, when G/C-content of RNA strongly affects the energy of secondary structure. By observation, the average Z-score of tRNAs is about –2 (Rivas and Eddy, 2000; Titov et al., 2002). The built-in distributions of Z-scores of random sequences were calculated earlier (Titov et al., 2002).

GAfold allows the simple restrictions for secondary structure, prohibiting pairing of continuous fragment of a sequence. This feature can be used for calculation of secondary structure of RNA interacting with short oligonucleotide. Other types of restrictions are available in GAedit, described in the end of this work.

GAinverse: Editor Mode

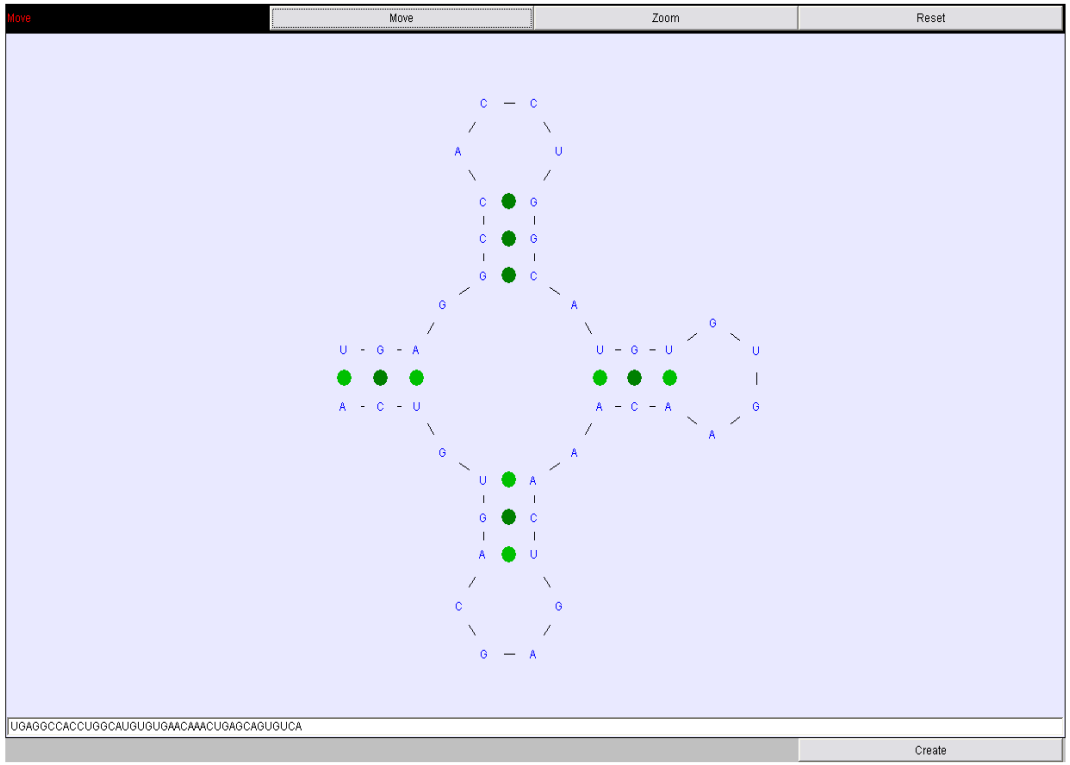


GAinverse

Specificity of complementary interactions makes RNA a perspective material for nanobiotechnology. GAinverse allows calculation of RNA sequences, which fold into a predefined structure, and can be used for RNA design. It is based on genetic algorithm and operates with a population of RNA sequences.

First, in the editor mode the length of yet unknown sequence and its structure constraints should be defined. Pressing the button “Calculate sequence” runs the inverse folding program, which calculates a desired nucleotide sequence. Then the program, predicting RNA secondary structure, displays the structure of the found sequence. This sequence is displayed in the same window.

GAinverse : secondary structure display



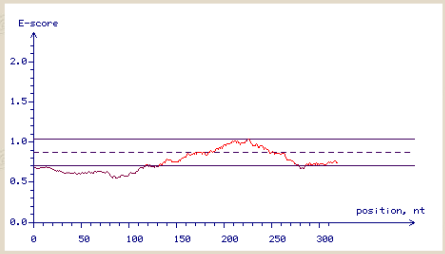
MatrixSS input window

MatrixSS

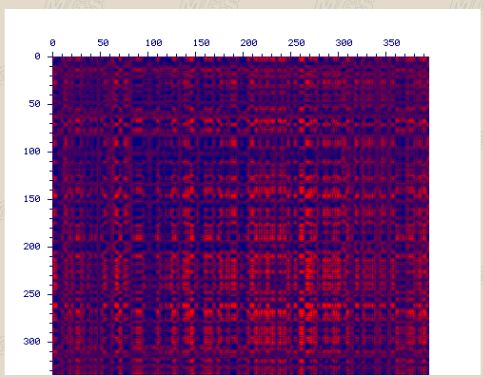
Due to significant progress of secondary structure predicting algorithms and computer processors, calculation of secondary structure of short RNAs can be now performed via Internet in real time. Long sequences still require a lot of time. The MatrixSS program performs fast simple search of the fragments that potentially form the stable secondary structures in the sequences of more than 250 nucleotides in length. MatrixSS calculates E-score – a nucleotide characteristic, which correlates with energy of secondary structure much better, than (G+C), (G-C)/(G+C) or other simple scores (Titov et al).

MatrixSS results

E-score profile



E-score matrix



The sequence input is analogous to GAfold. On output MatrixSS generates a symmetric dot-like matrix of potential complementarity. Summing over columns gives E-score profile – the potential of involvement of the sequence fragments to secondary structure. After finding a perspective fragment in a long sequence its secondary structure can be calculated by GAfold.

List of publications

Titov I.I., Vorobiev D.G., Ivanisenko V.A., Kolchanov N.A. (2002). *Fast genetic algorithm for RNA secondary structure analysis.* Russ. Chem. Bull. **51** (12) 1135-1144.

Titov I., Vorobiev D., Palyanov A. (2006) A toolbox for analysis of RNA secondary structure based on genetic algorithm. In: Bioinformatics of Genome Regulation and Structure II (N. Kolchanov, R. Hofestaedt, L. Milanesi, eds) Springer. pp. 105-110.

3.3. Translation initiation and termination

Analysis of the specific contextual features of translation initiation and termination sites in *Saccharomyces cerevisiae*

Investigation of mRNA sequence organization is of importance to reveal the features influencing translation efficiency and specificity.

We performed statistical analysis of translation initiation and termination sites of well-studied eukaryotic organism Saccharomyces cerevisiae.

Yeast mRNAs were analyzed using trinucleotide weight matrices and vocabularies of significant oligonucleotide motifs.

A statistically significant difference of nucleotide contexts between high- and low-expressed mRNAs was found.

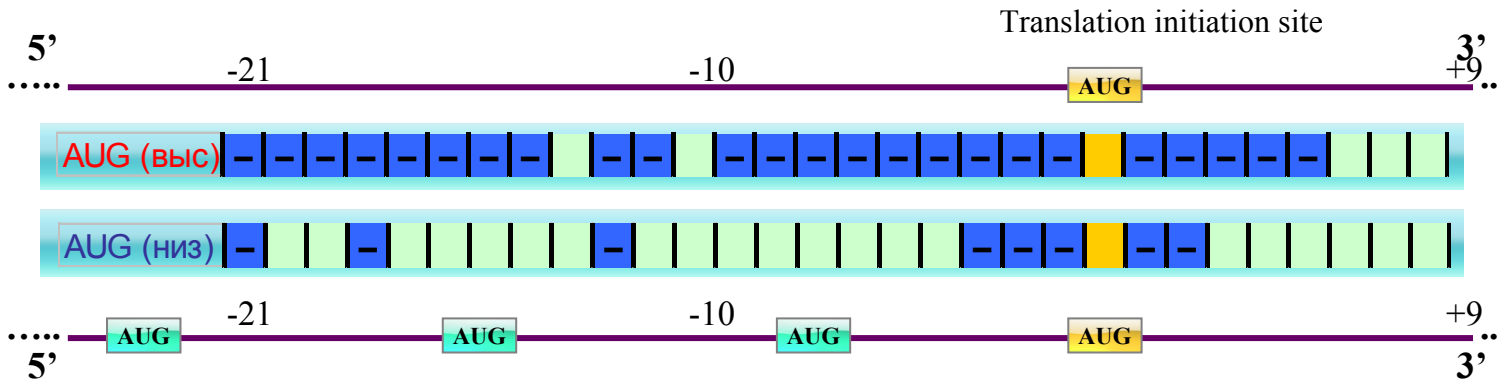
Computer simulation of evolution using genetic algorithm demonstrated that the rate-limiting stage model could explain this phenomenon.

Position weights of AUG trinucleotide in the region of translation initiation site of high- and low- expressed mRNA

$$w_{b,k} = \log\left(\frac{f_{b,k}}{e_{b,k}}\right)$$

$f_{b,k}$ – frequency of the presence of trinucleotide b in position k of 5'- nontranslated region of mRNA;
$e_{b,k}$ – frequency of its presence in random sequences

5'-context of high-expressed mRNA has not false AUG, as usual



5'-context of low-expressed mRNA has a lot of false AUG

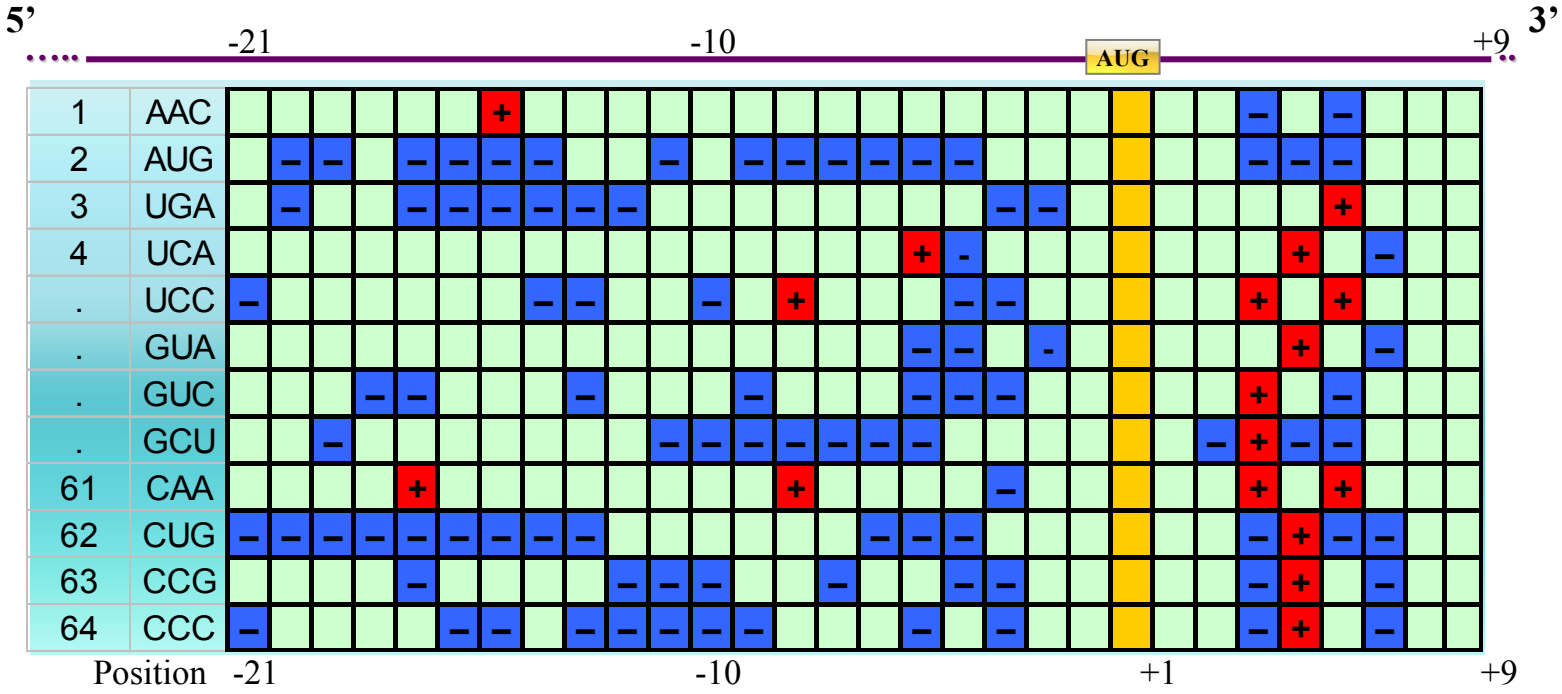
Position trinucleotide weights of the contrast weight matrix in translation initiation site.

$$w_{b,k} = \log\left(\frac{f_{b,k}}{e_{b,k}}\right)$$

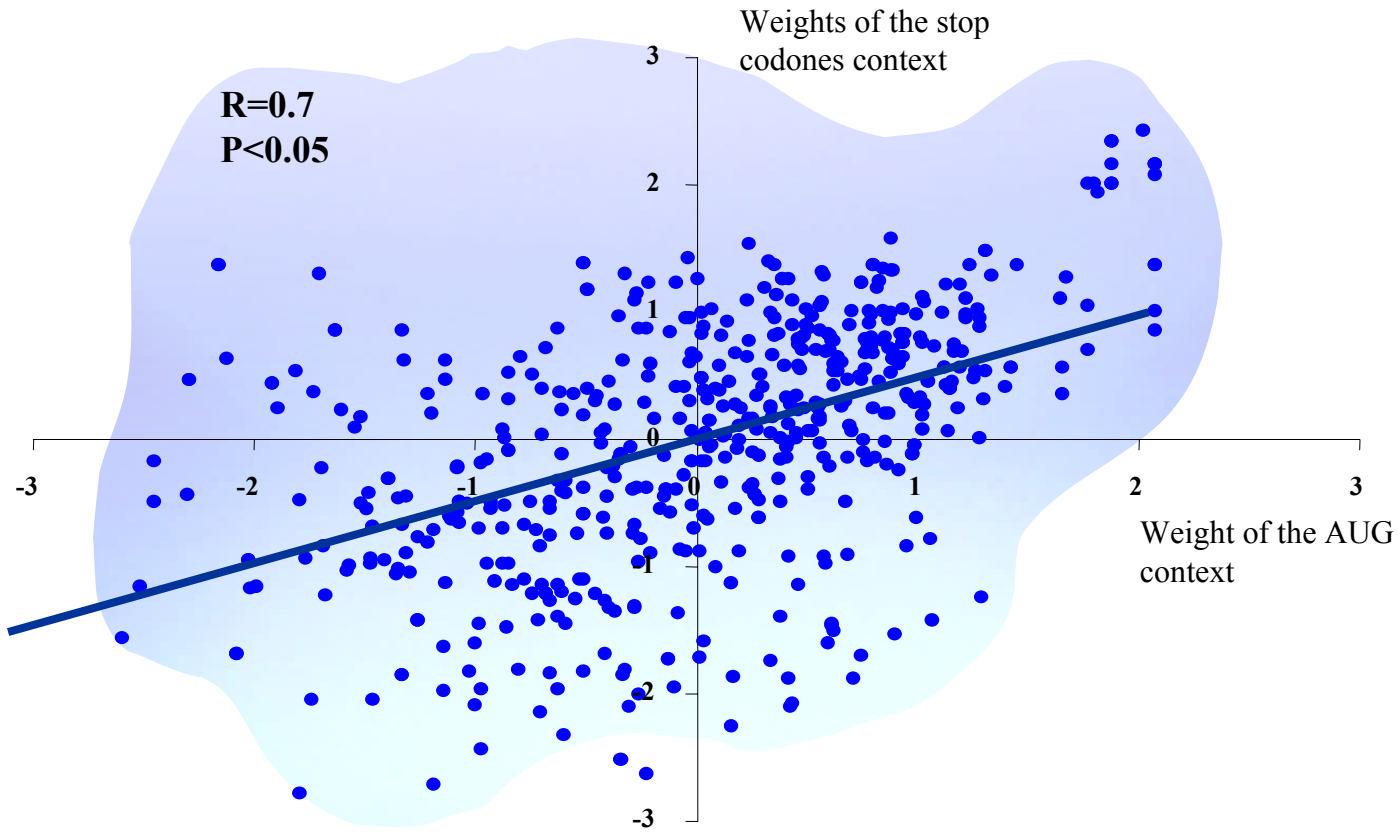
$f_{b,k}$ – frequency of the presence of trinucleotide b in position k of high-expressed mRNA;
$e_{b,k}$ – frequency of the presence of trinucleotide b in position k of low-expressed mRNA

$$S_{5'\text{-}} = \sum_{k=1\ldots L} w_{b,k}$$

Weight of the sequence analyzed – characterizes its similarity level with trinucleotide weight matrix

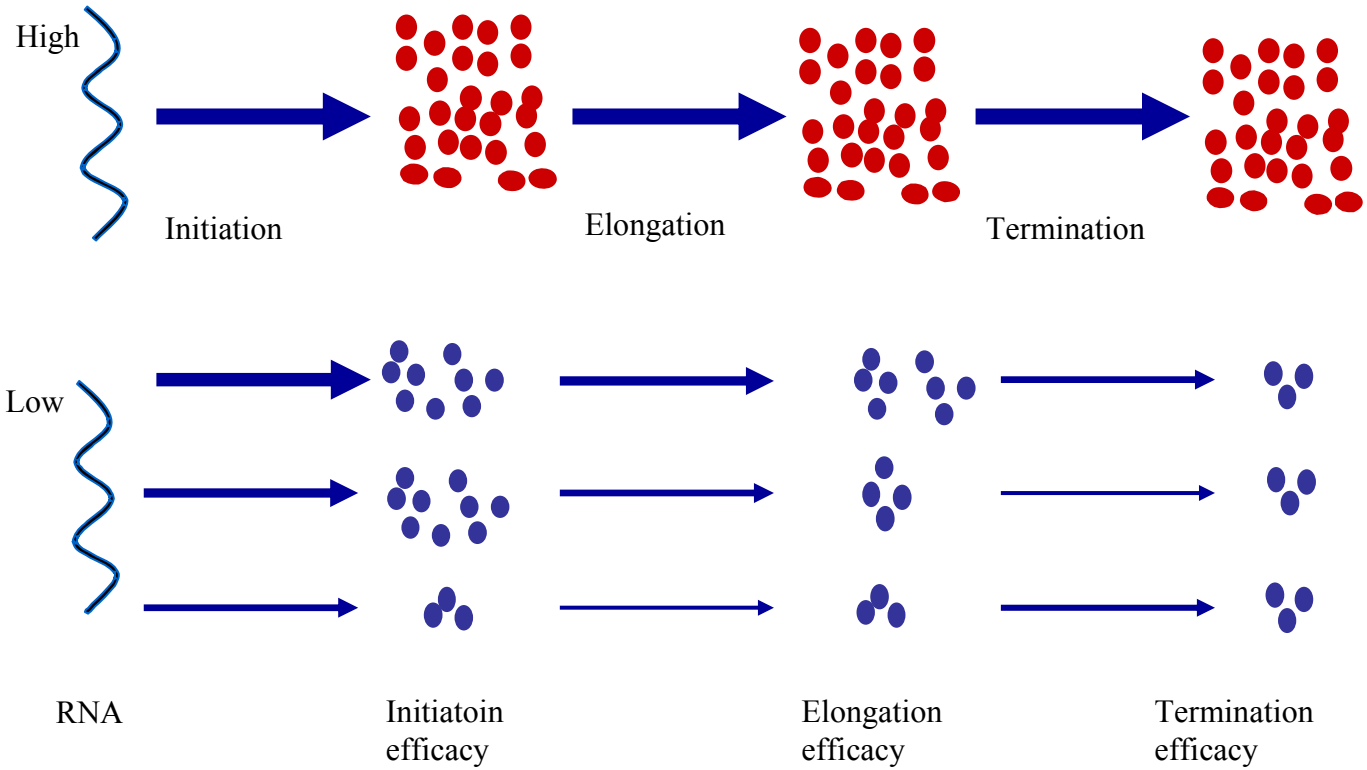


Dependence between context of AUG codon (X axis) and the context of termination codones (Y axis) for high-expressed mRNA (CAI>0.3)

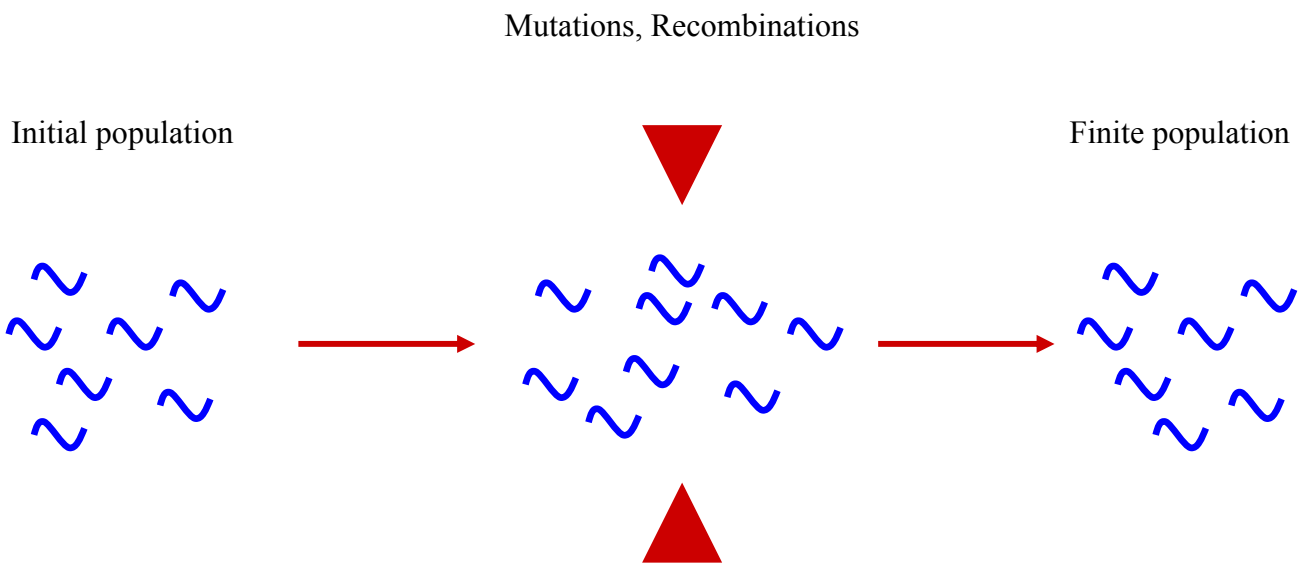


The low level of polypeptide production may be caused by lim iting any stage of translation

The high level of polypeptide production should meet the demand of high efficiency at every stage of expression



Genetic algorithm: general description



Selection directed to increasing translation rate *F* according to the limiting stage model

Fitness calculation

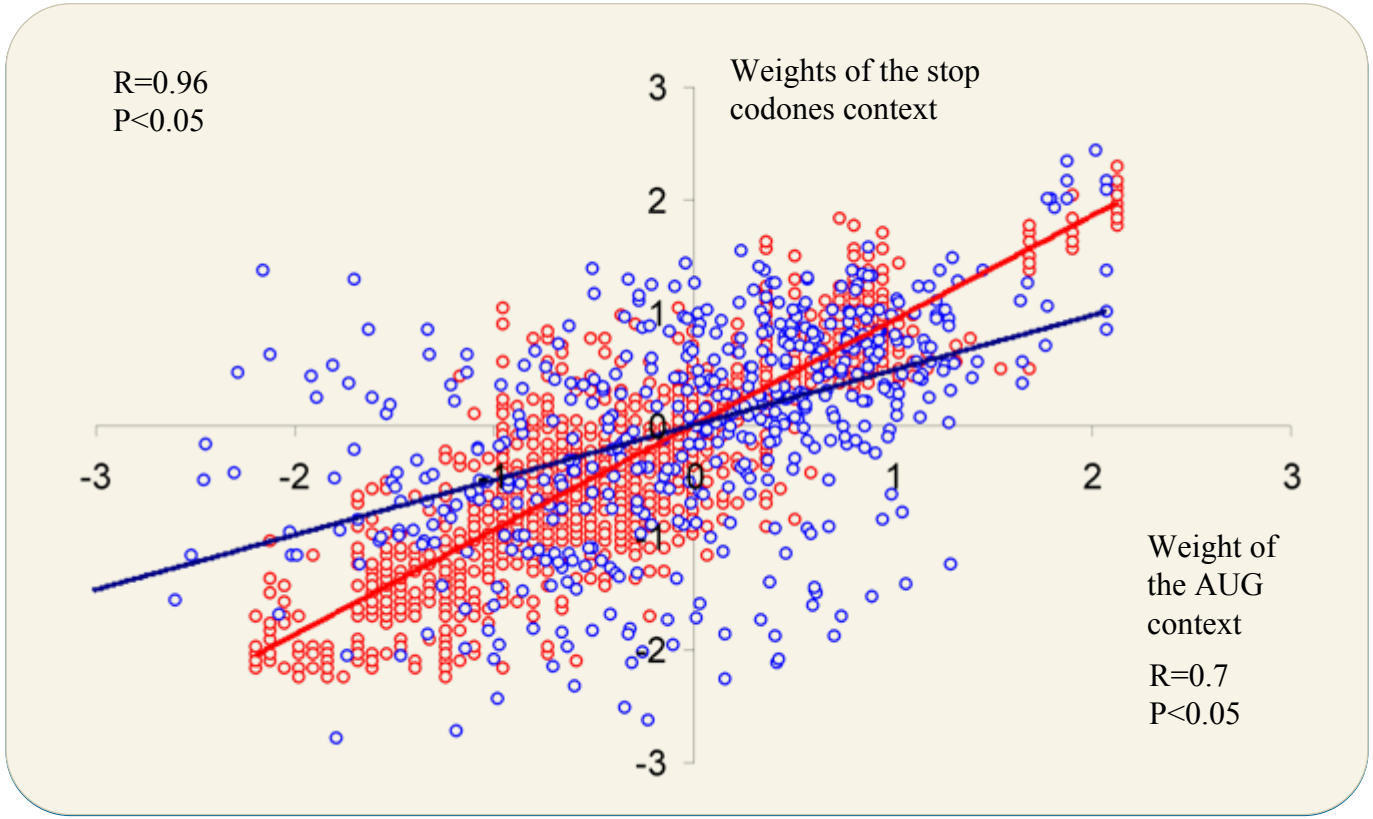
$$F = \min\begin{cases} Score(5'-region) \\ CAI(coding\_region) \\ Score(3'-region) \end{cases}$$

Score(5'-region) – the score of the 5'-region calculated from the corresponding weight matrices

Score(3'-region) – the score of the 3'-region calculated from the corresponding weight matrices

CAI(coding_region) - codon adaptation index of coding region.

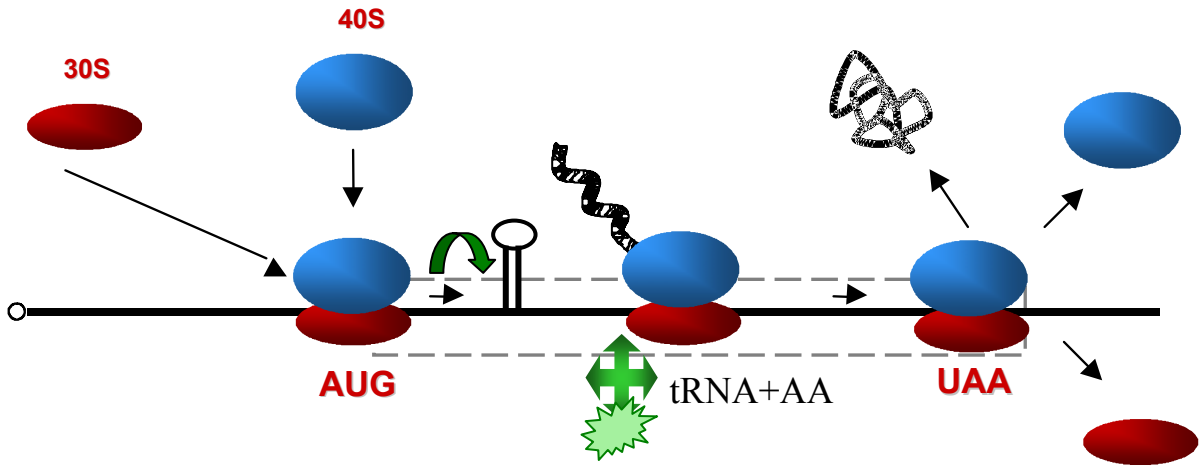
Comparison of the dependence between context of AUG codon (X axis) and the context of termination codones (Y axis) for high-expressed mRNA (blue rings) and computer simulated by genetic algorithm sequences (red rings).



3.4. Prediction of translation efficiency

Schematic model of translation of prokaryotic mRNA

(mRNA, start of translation, hairpin, protein, and stop-codon)



We analyzed the interrelation between the efficiency of gene expression and the nucleotide composition of all protein-coding sequences in 240 unicellular organisms.

We demonstrated that frequency analysis of gene-codon composition fails to reflect adequately the gene expression efficiency of all these organisms.

We constructed a measure, the elongation-efficiency index, that considers simultaneously the information on codon frequencies and the degree of mRNA local self-complementarity.

According to our analysis, these 240 species fall into five groups differentiated by the process that makes the key contribution to the elongation rate.

The elongation efficiency index comprises two addends: the former depends of the codon usage frequency; the latter, on local mRNA hairpins.

The quality of nucleotide composition of a particular (*i*th) mRNA is estimated according to the value of elongation efficiency index EEI(i), which has the meaning of average elongation time of all the accountable codons in a gene:

$$EEI(i) = u_1 T_a(i) + u_2 T_e(i),$$

where $u_1 = 0$ or 1; $u_2 = 0$ or 1 are weight coefficients determining the contribution of each term into the value of this index.

Elongation time depends on the codons used: frequent/rare <—>quick/slow

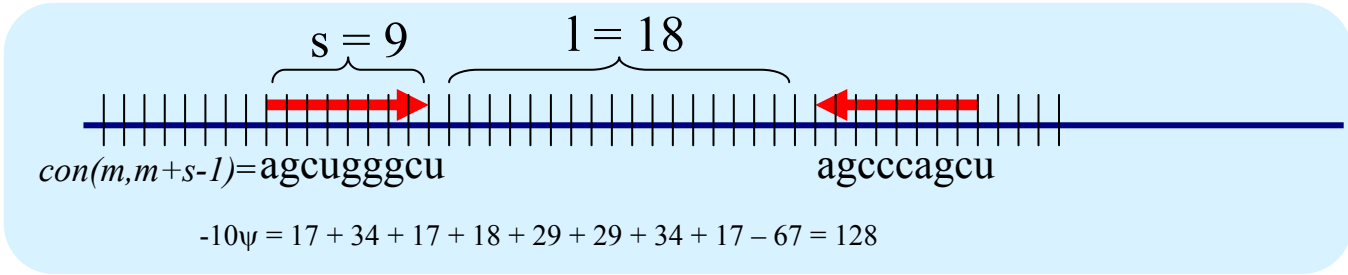
The first term T_a estimates the average time required for isoacceptor aminoacyl-tRNA to be placed in the ribosome R site from the codon composition and is calculated according to the below equation

$$T_a(i)=\sum_{j=1}^{n_i}\beta_{\delta(i,j)}\Big/n_i\,,\ \beta_\delta=\frac{\sum_{m=1}^{C}\sqrt{\alpha_m}}{\sqrt{\alpha_\delta}}\,,$$

where the value $1/\beta_{\delta(i,j)}$ is interpreted as the optimal relative concentration of aminoacyl-tRNA complementary to the *j*th accountable codon; $\alpha_{\delta(i,j)}$ and α_m have meanings of usage frequencies of the codons δ(i,j) and m in a certain mRNA subset.

Local complementarity index is the measure of the number of local hairpins and their energies

$$LCI(i,j) = \sum_{\substack{m:\ m\le j\le m+s-1\ \ or \\ m+s+l-1\le j\le 2m+2s+l-2}} 10\left\{\sum_{s=s_{\min}}^{s_{\max}}\left[\sum_{l=l_{\min}}^{l_{\max}} -\psi\left(con(m,\ m+s-1), \overline{con(m+s+l-1, 2m+2s+l-2)}\right)\right]\right\}$$



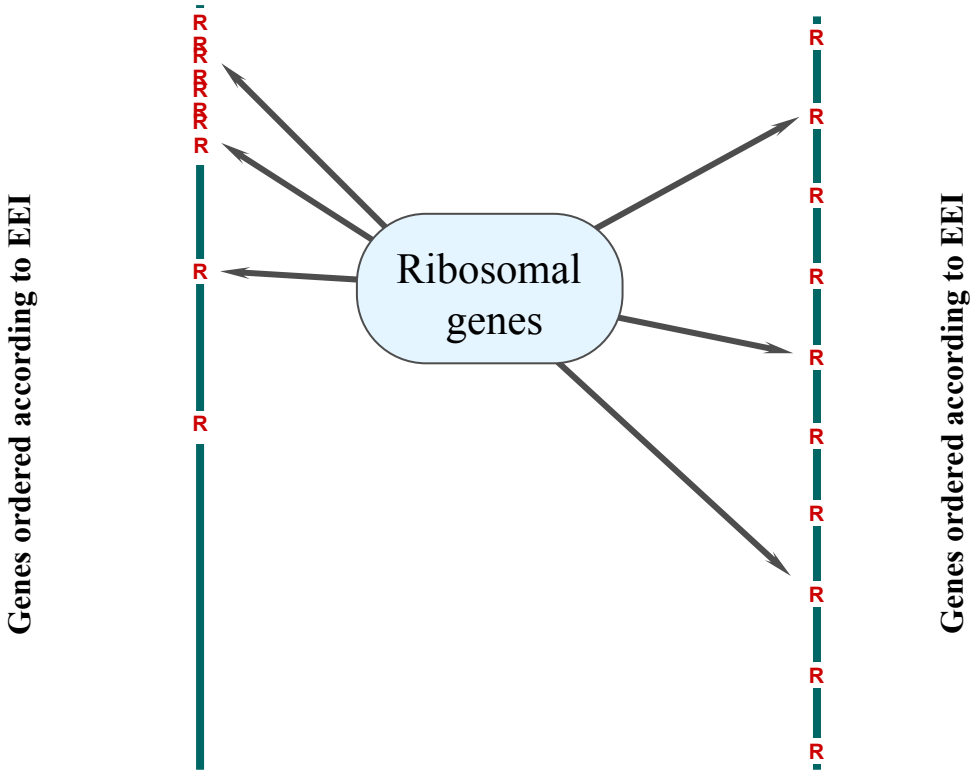
aa, au, ua, ca, cu, ga, gu, cg, gc, gg
uu, au, ua, ug, ag, uc, ac, cg, gc, cc
9 9 11 18 17 23 21 20 34 29

Destabilizing weights of loops

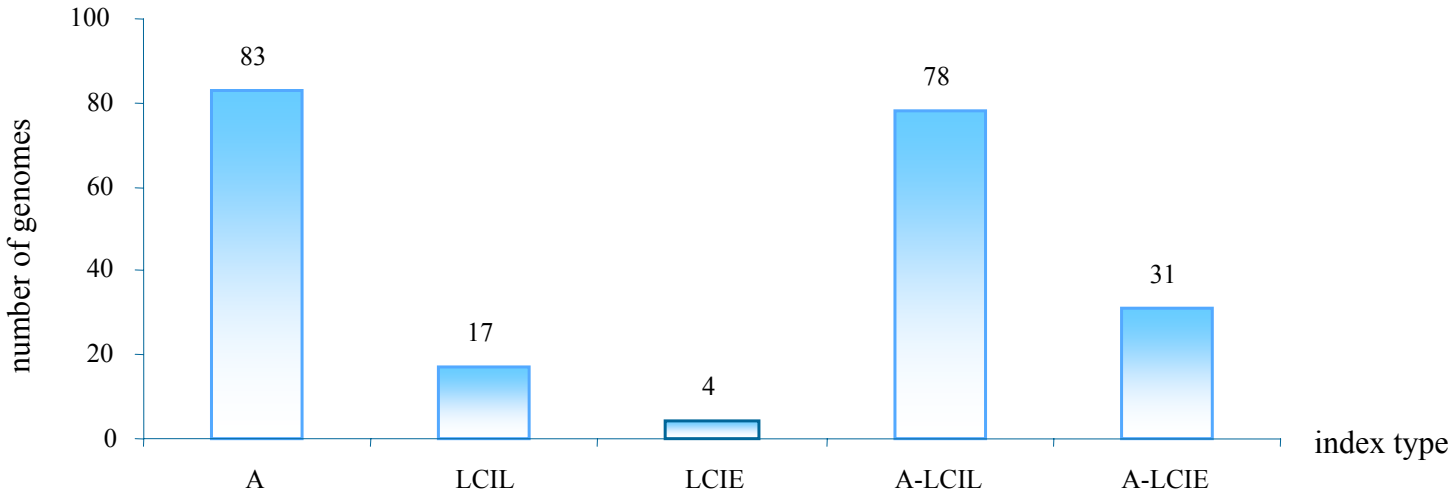
-1000 -1000 -74 -59 -44 -43 -41 -41 -42 -43
-46 -49 -53 -56 -59 -61 -64 -67 -69 -71
-73 -75 -77 -79 -81 -83 -85 -87 -88 -89
-90 -91 -92 -93 -94 -95 -96 -97 -98 -99
-99 -99 -100 -100 -100 -100 -101 -101 -101 -101

D.N.Turner, N.Sugimoto (1988) Ann.Rev. Biophys. Biophys.Chem.17, 167-192

Ribosomal genes are considered highly expressed
To the left, the index works adequately; to the right, inadequately.

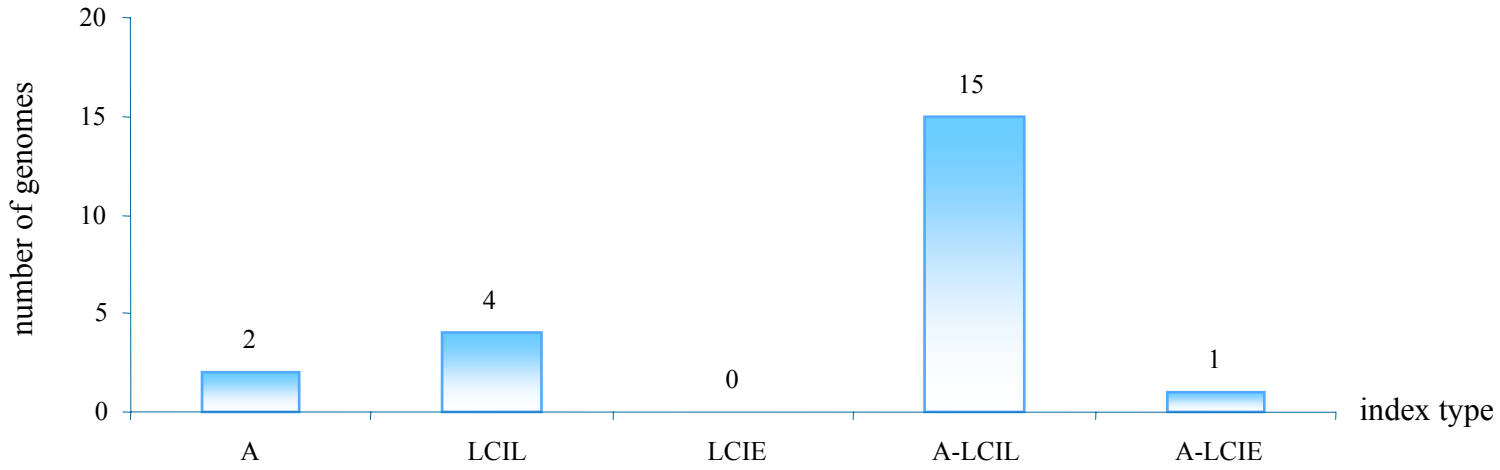


Distribution of 213 bacterial genomes by 5 groups of translational optimization



A – group with preferential codon usage, LCIL – secondary structures with account of length, LCIE - secondary structures with account of energy, A-LCIL – codon usage with secondary structures with respect to length, A-LCIE – codon usage with secondary structures with respect to energy.

Distribution of 22 Archaea genomes by 5 groups of translational optimization



A – group with preferential codon usage, LCIL – secondary structures with account of length, LCIE - secondary structures with account of energy, A-LCIL – codon usage with secondary structures with respect to length, A-LCIE – codon usage with secondary structures with respect to energy.

Hypotheses

We may hypothesize that in the case of group ζ and Aζ organisms, encounter of a ribosome with a hindrance triggers a mechanism that spends a predetermined batch of resources (time or energy) to remove all the hindrances within mRNA region of a certain length, independently of their energies and number.

On the contrary, the corresponding mechanism of group ψ and Aψ organisms is somehow capable of estimating the hindrance "capacity" and spends the proportional time (or, possibly, energy) for its removal.

Hypotheses

The loss of the elongation stage sensitivity to the codon composition, observed in the case of organisms belonging to ζ and ψ groups, may also suggest the following explanations: either (a) the placement of isoacceptor aminoacyl-tRNA in the ribosome A site of is approximately efficient for all the codons or (b) or proceeds in parallel with the process removing the hindrances ahead of the moving ribosome (preparatory stage of translocation) and this process is slower. Thus, the latter process shields in a sense the former process, thereby providing the evolutionary neutrality of the codon mutations.

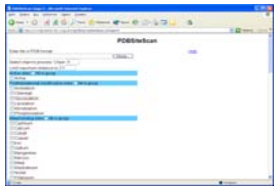
Chapter 4
COMPUTATIONAL PROTEOMICS

4.1. Computer-assisted approaches facilitating search of targets for drugs, drug design, and evaluation of molecular toxicity

Computer-assisted approaches to medicinal and biotechnological issues based on computer proteomics have been developed at the Institute of Cytology and Genetics SB RAS

PDBSite: a functional site database

PDBSiteScan: a program for functional site recognition

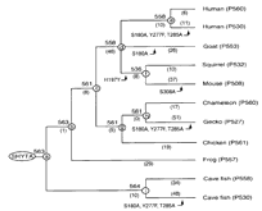


PDBSiteComplex: a program for molecular complex reconstruction

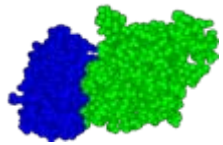
WebProAnalyst: a program for quantitative structure-activity relationships analysis in protein families

Protein function annotation

Molecular evolution

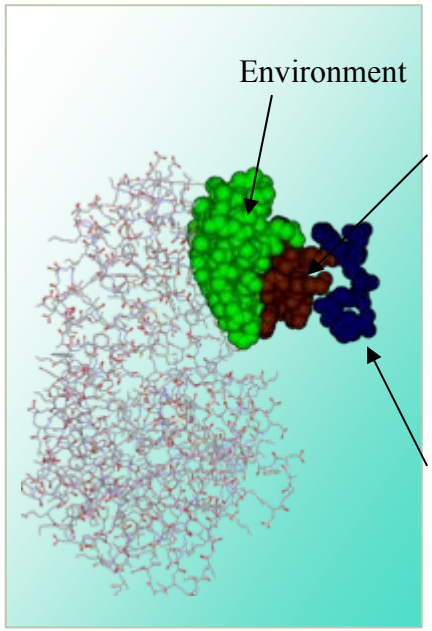


Search for drug targets



Data on spatial structure and main features of functional sites of proteins and their ligands are accumulating in the PDBSite and PDBLigand databases

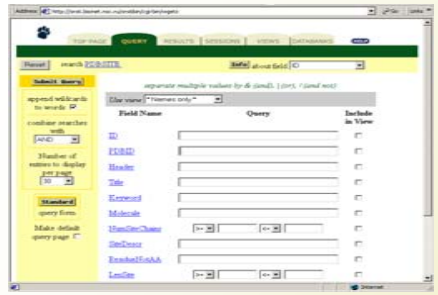
http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/

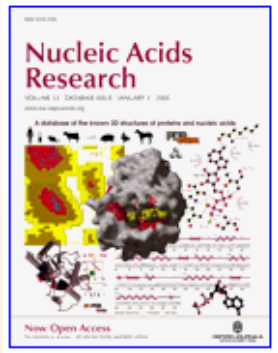


Annotated proteins from PDB are data sources

PDBSite database

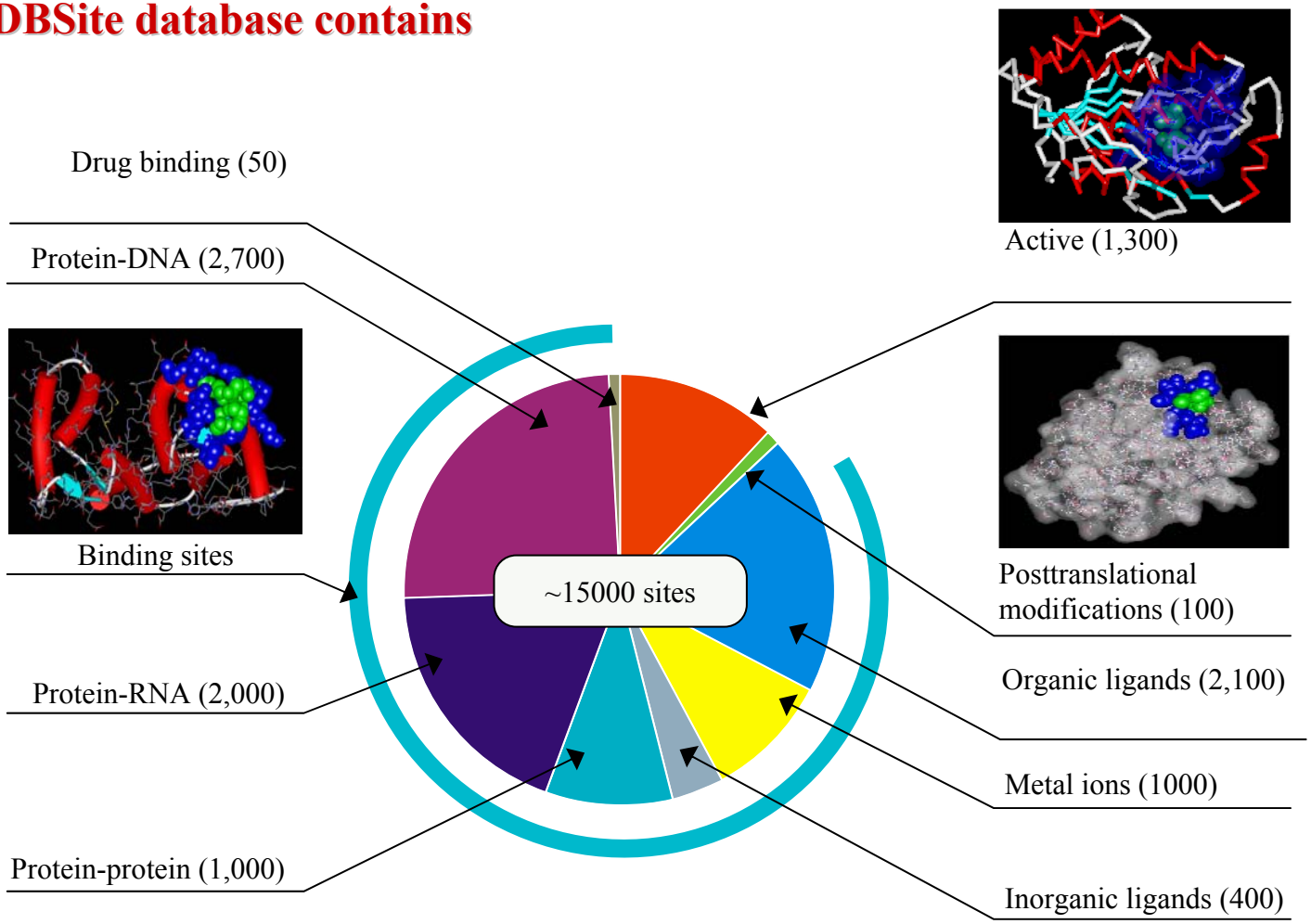
PDBLigand database

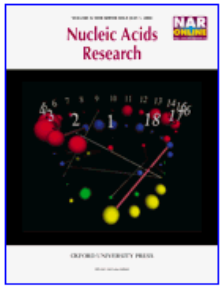




Ivanisenko et al, (2005) Nucleic Acids Res.

The PDBSite database contains

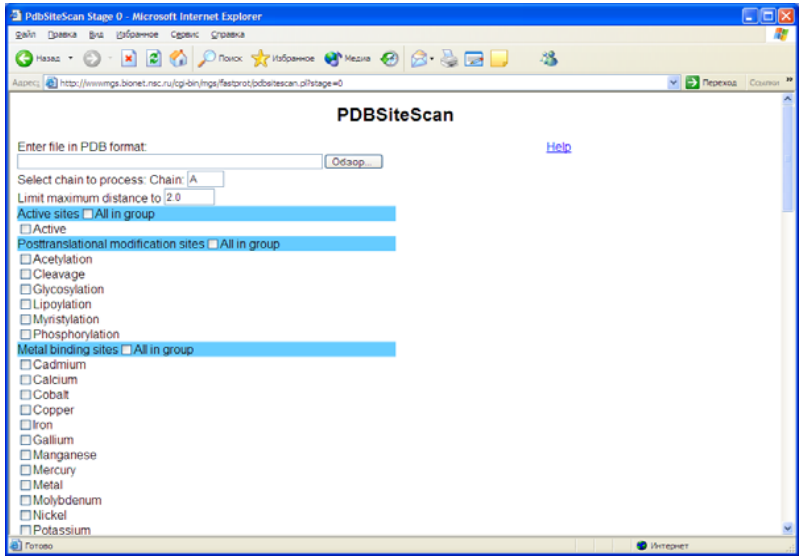




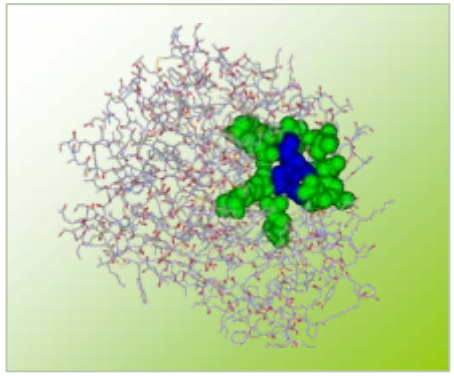
Ivanisenko et al, (2004)
Nucleic Acids Res.

PDBSiteScan: a program for the recognition of functional site

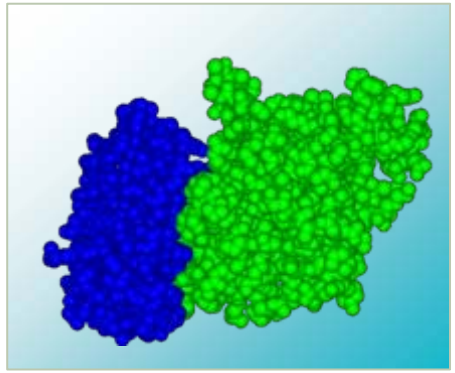
http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html



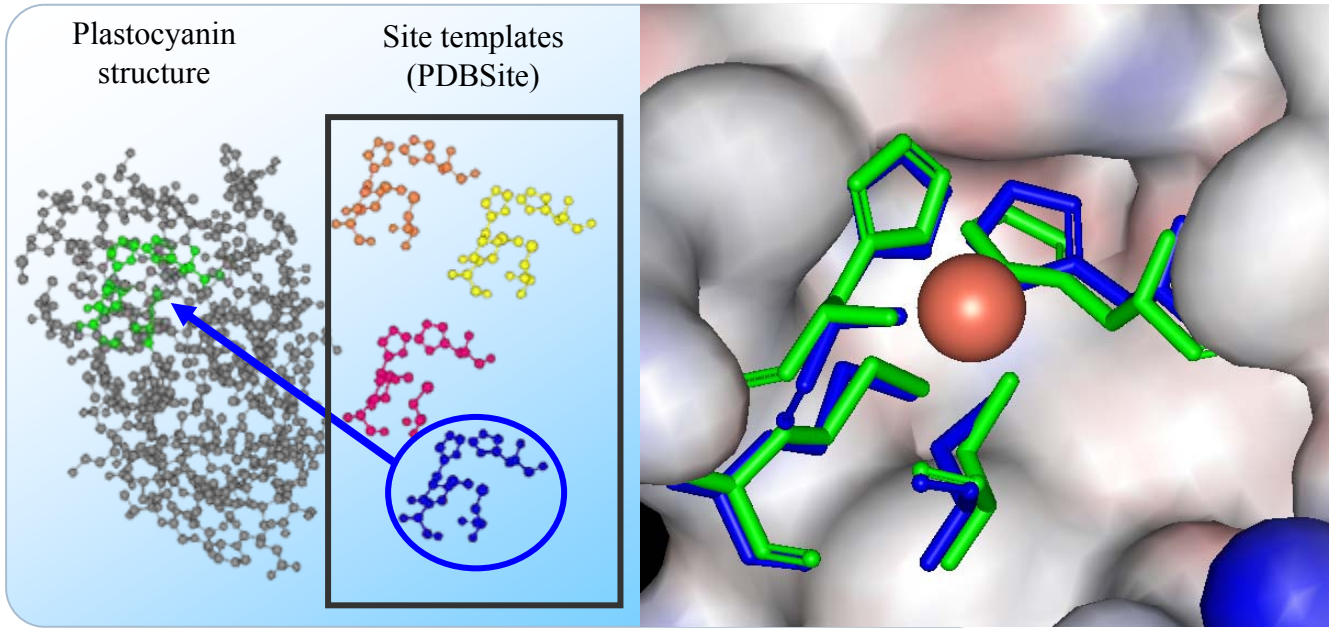
Functional site recognition



Reconstructio n of site-ligand complexes

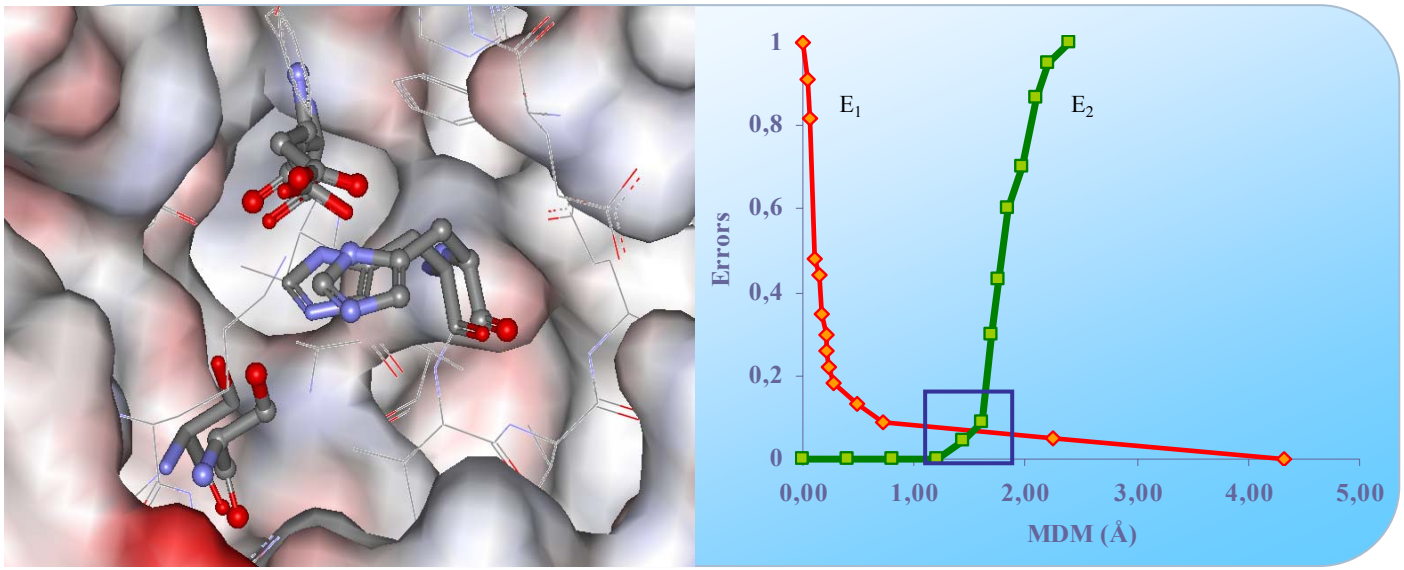


Example. Search for copper binding site in plastocyanin (PDB ID 1BXU)



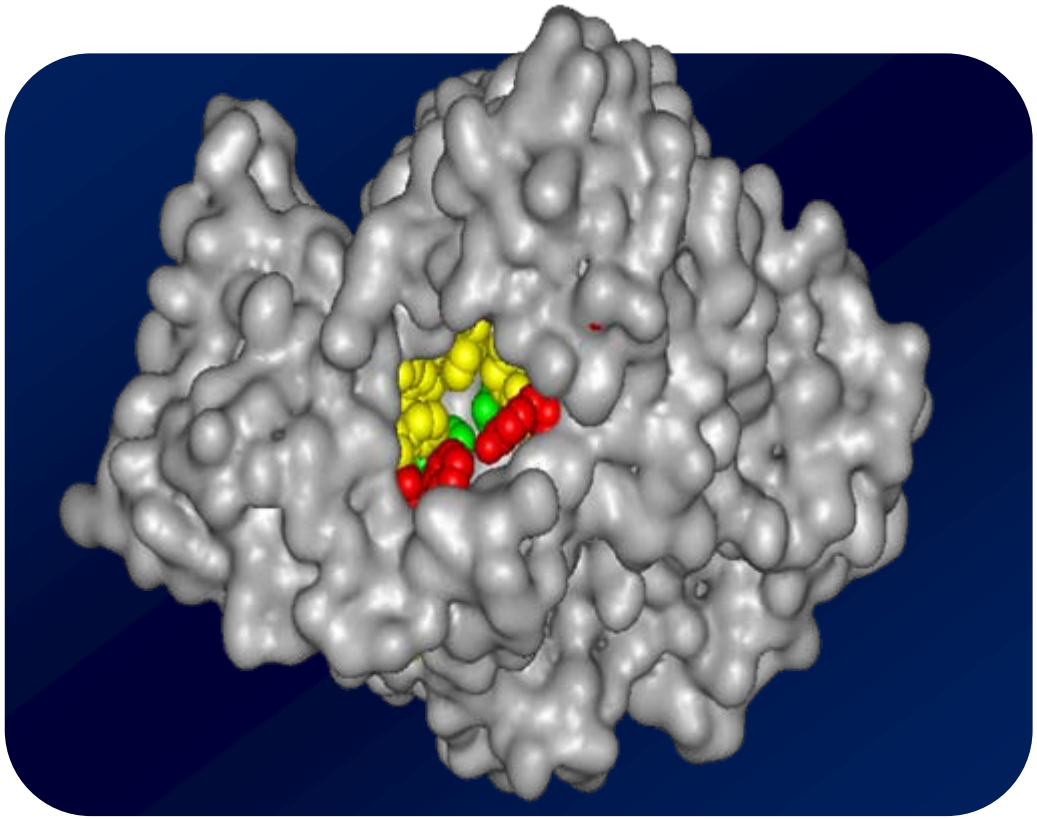
The residues of the recognized site in plastocyanin are in green, those of template site from the PDBSite database (ID 1B3ICU) are in blue. Orange ball highlights copper ion.

Accuracy estimation for catalytic center recognition in hydrolase superfamily



MDM – maximum distance mismatch between site template and protein fragment.
E_1 – type I error (underprediction)
E_2 – type II error (overprediction)

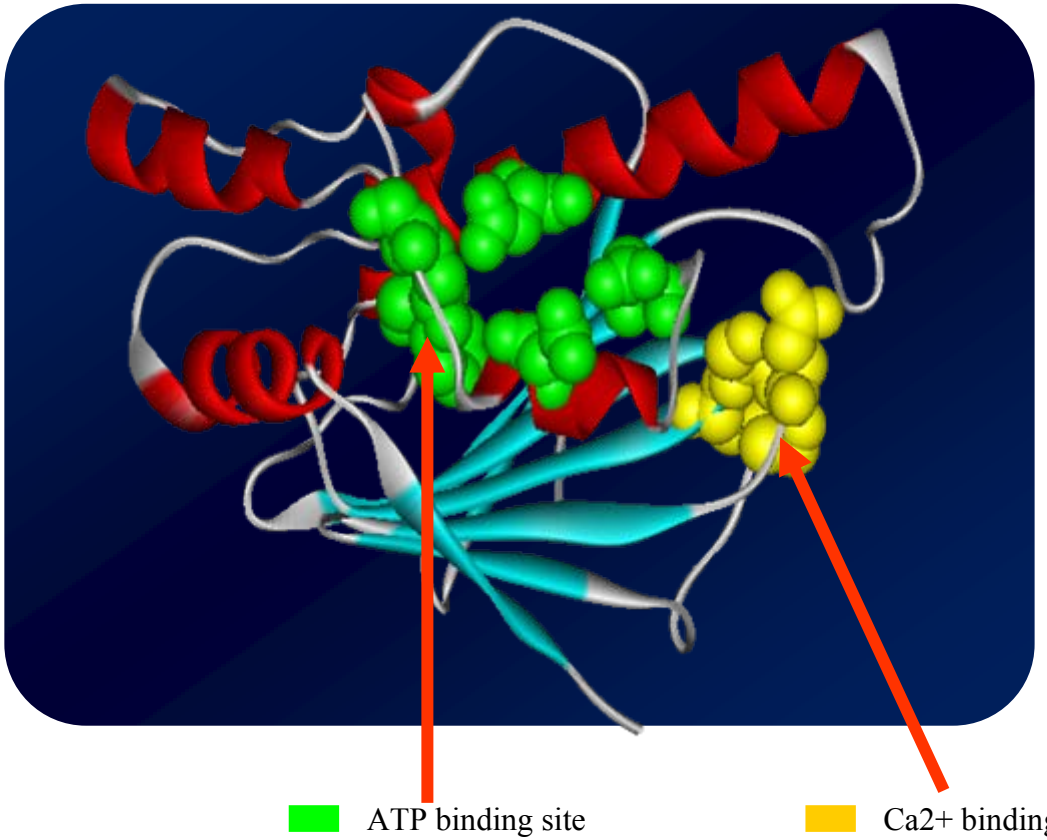
Analysis of 3-D structure of Acetylcholinesterase by PDBSiteScan program: putative functional sites recognition



inhibitor binding site FBQ binding site catalytic active site

- **FUNCTION**: rapidly hydrolyzes choline released into the synapse. It may be involved in cell-cell interactions.
- **CATALYTIC ACTIVITY**: Acetylcholine + H_2O = choline + acetate.
- **SUBCELLULAR LOCATION**: the h form is attached to the membrane by a GPI-anchor.
- **SIMILARITY**: belongs to the type-B carboxylesterase/lipase family.

Analysis of 3-D structure of heat shock protein HSP82 by PDBSiteScan program: putative functional sites recognition

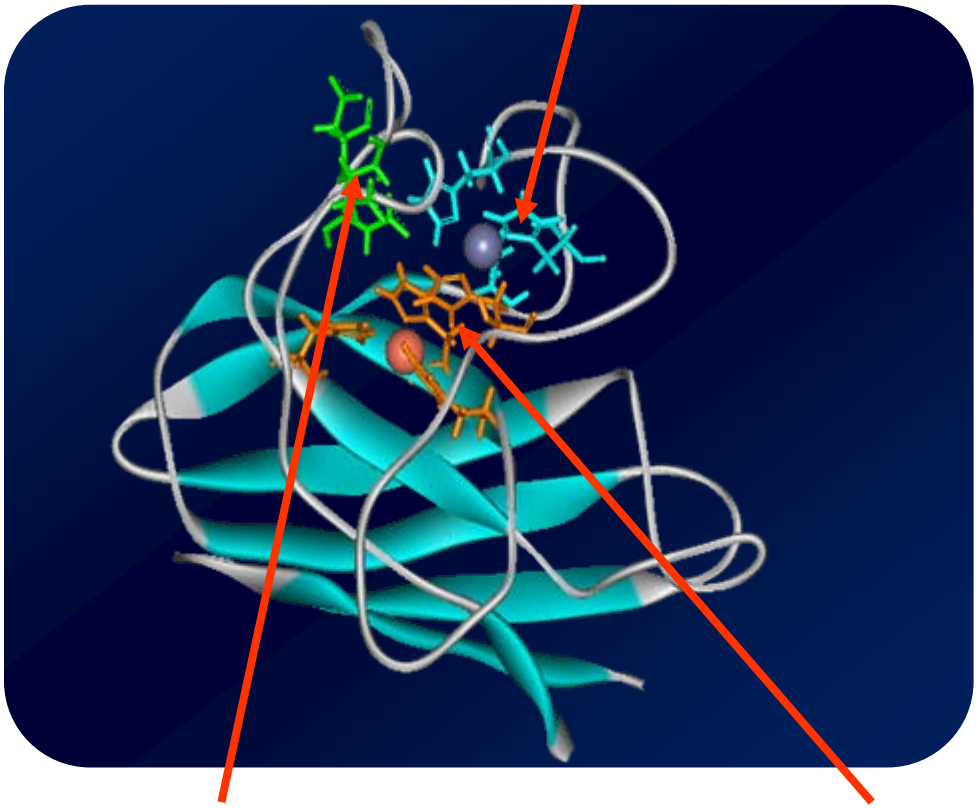


• **FUNCTION:** HSP82 is an essential protein that is required by cells in higher concentrations for growth at higher temperatures. Molecular chaperone. Has ATPASE activity.

• **SUBCELLULAR LOCATION**: Cytoplasmic.

• **SIMILARITY**: belongs to the heat shock protein 90 family.

Analysis of 3-D structure of Superoxide dismutase [Cu-Zn] by PDBSiteScan program: putative functional sites recognition



•**CATALYTIC ACTIVITY**:
2 superoxide + 2 H^+ = O_2 + H_2O_2.

•**COFACTOR**: Copper and zinc.

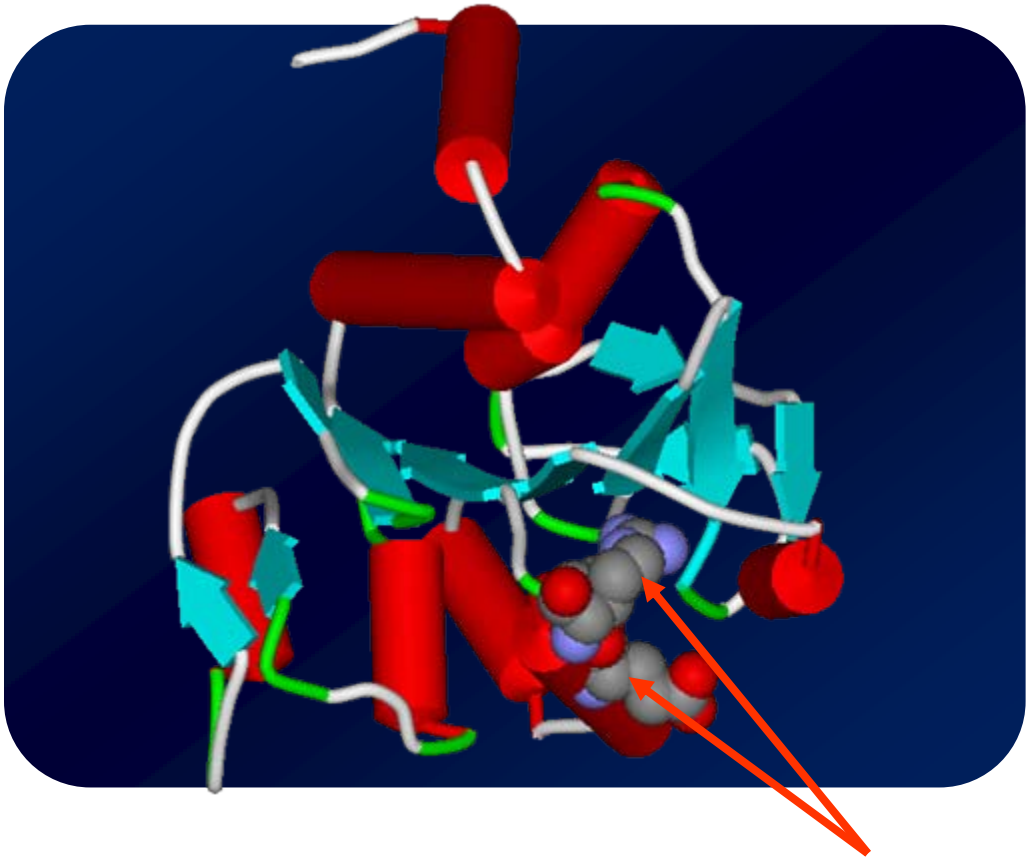
•**SUBCELLULAR LOCATION**: Cytoplasmic.

•**DISEASE**: Defects in sod1 are the cause of amyotrophic lateral sclerosis (ALS), a degenerative disorder of motorneurons in the cortex, brainstem and spinal cord. ALS is characterized with muscular weakness and atrophy beginning in the hands and spreading to the forearms and legs.

•**FUNCTION**: Destroys radicals which are normally produced within the cells and which are toxic to biological systems.

•**SIMILARITY**: Belongs to the Cu-Zn Superoxide Dismutase Family.

Analysis of 3-D structure of Human DJ-1 by PDBSiteScan program: putative functional sites recognition

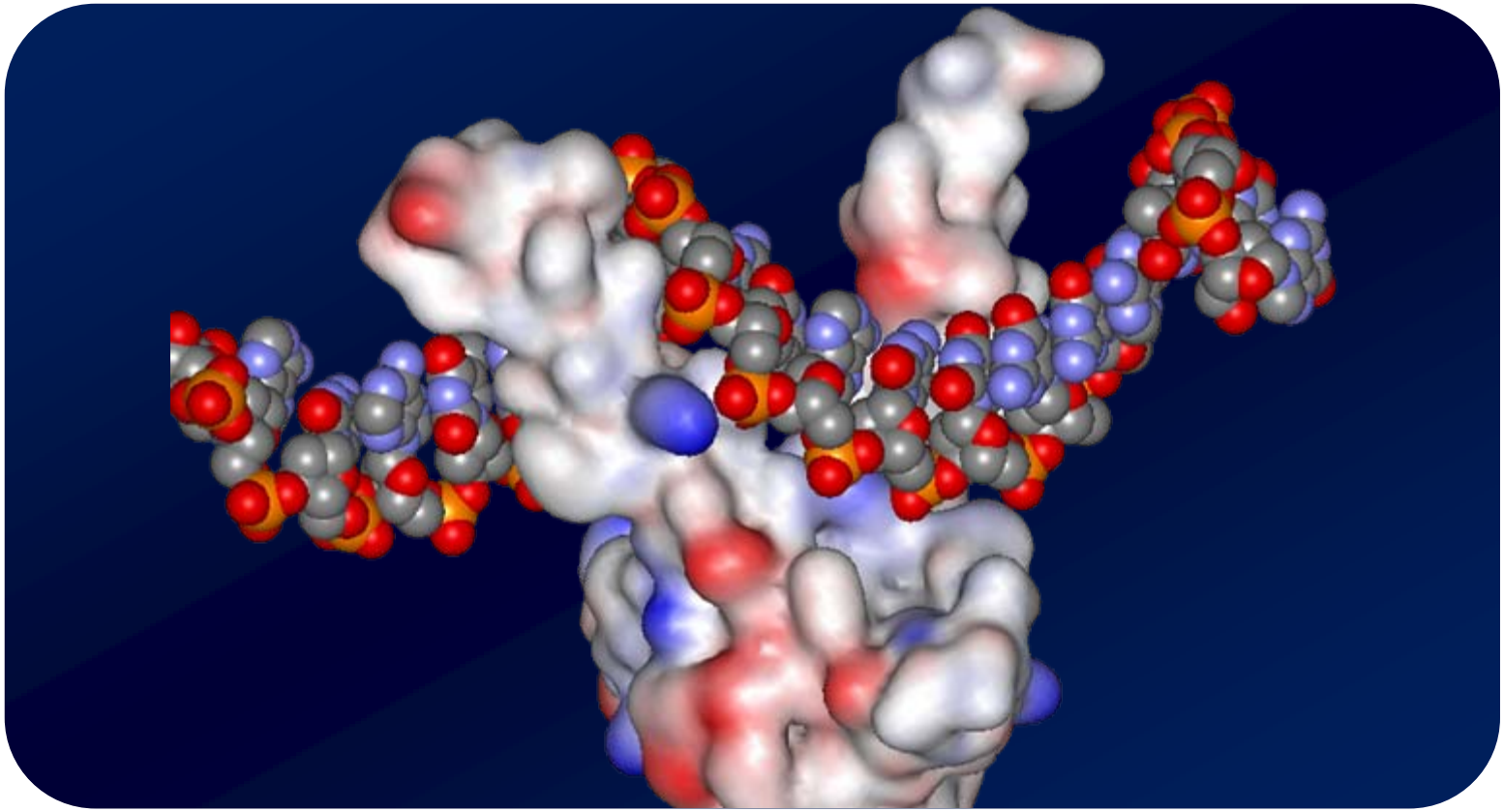


Identified Na-binding site is shown as balls.

• **FUNCTION**: **UNKNOWN**. A putative intracellular cysteine protease. Furthermore, DJ-1 was identified as a regulatory subunit (RS) of an RNA-binding protein (RBP) complex and inhibits the RNA-binding activity of RBP.

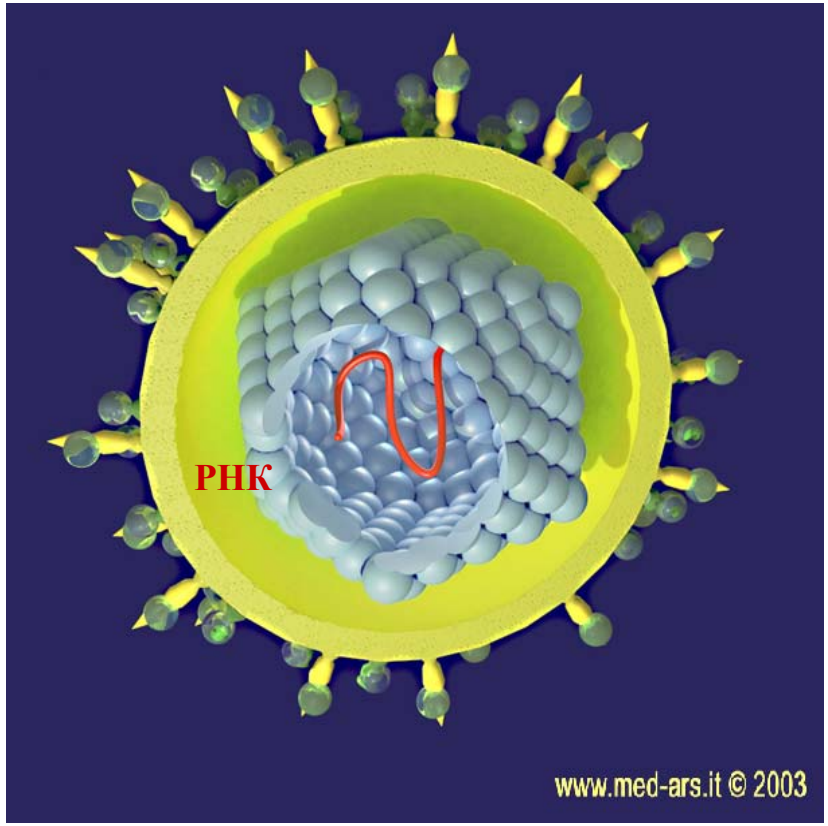
• **DISEASE**: DJ-1 is a protein involved in multiple physiological processes, including cancer, Parkinson's disease, and male fertility.

The structure of the DNA-protein complex of transcription factor SREBP-1A and its binding site (reconstruction using PDB-site and PDB-siteScan tools).



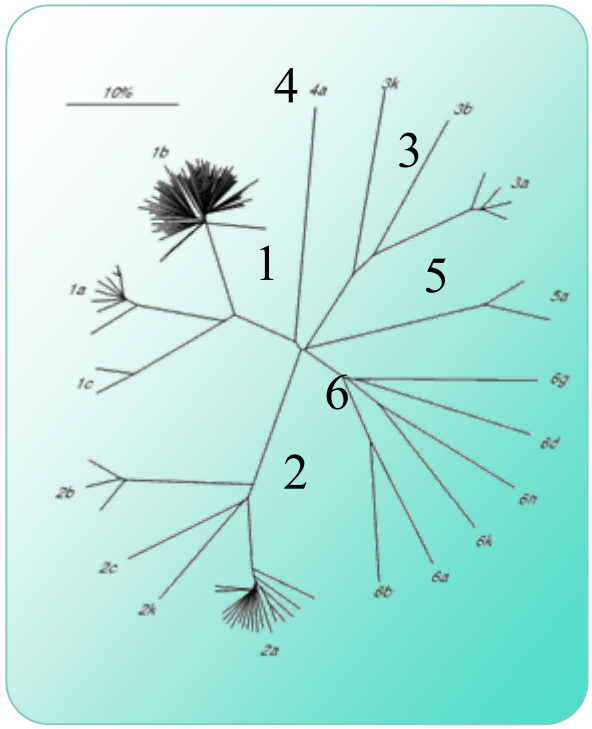
4.2. Search for potential antiviral drug targets

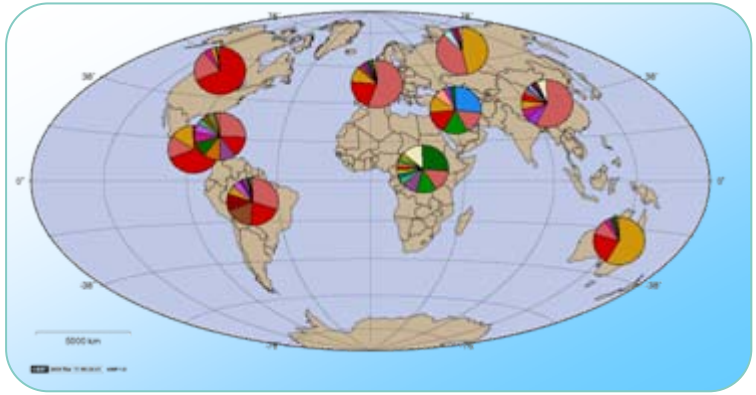
Hepatitis C virus

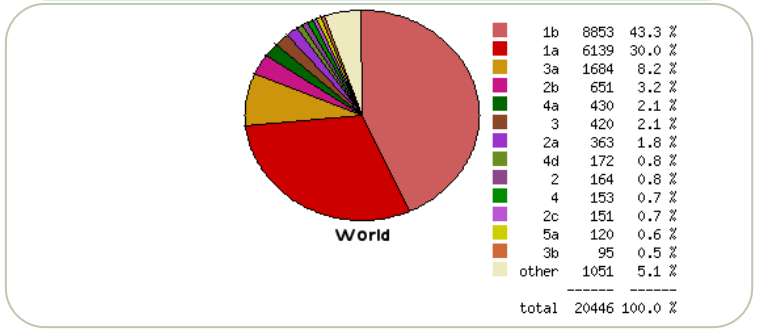


Hepatitis C virus is a member of flavivirus family. The chronic infection is protracted for 10-15 years, causes liver cirrhosis, provokes cancer, suppresses immune system.

Main genotypes of hepatitis C virus

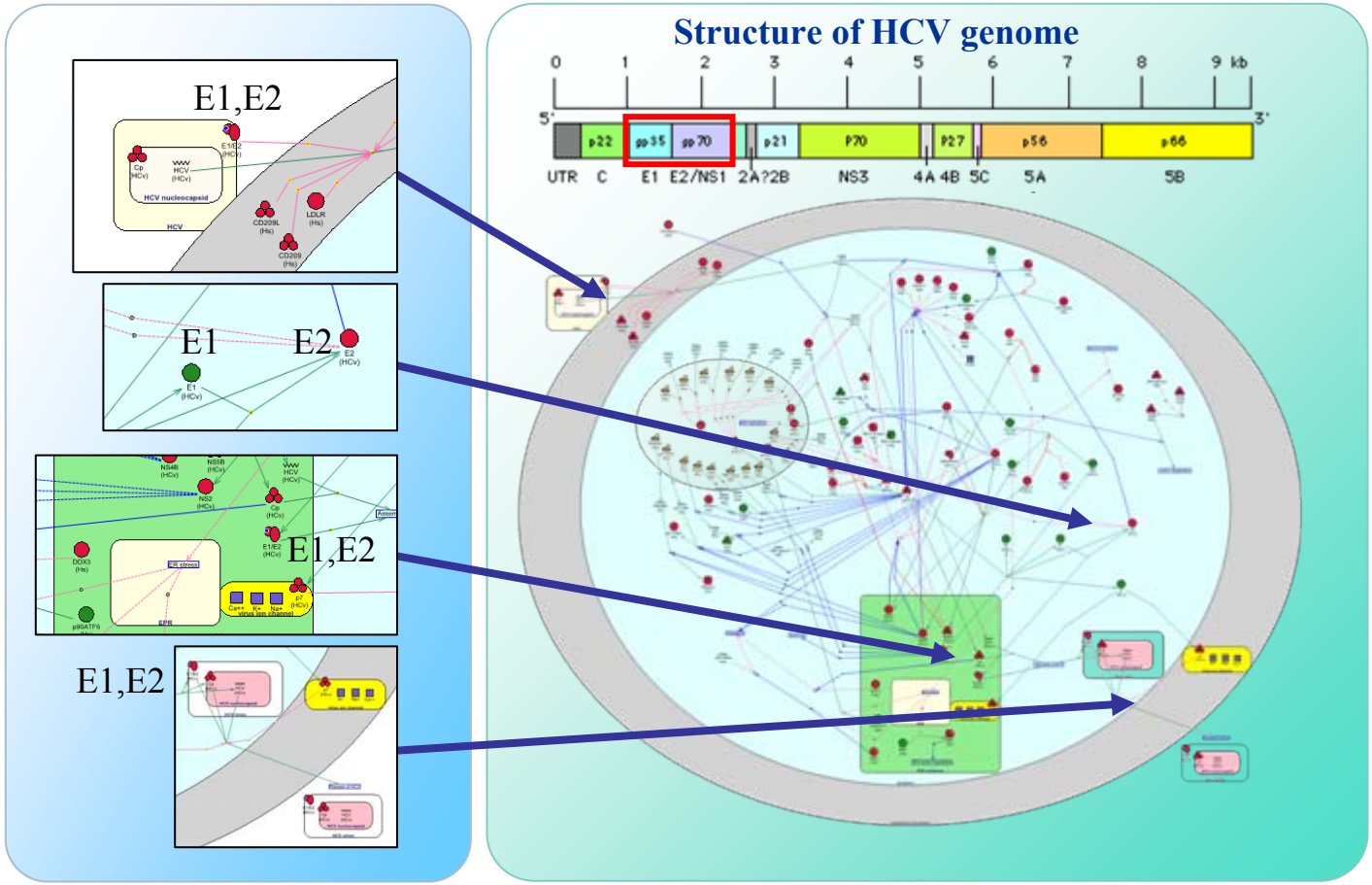






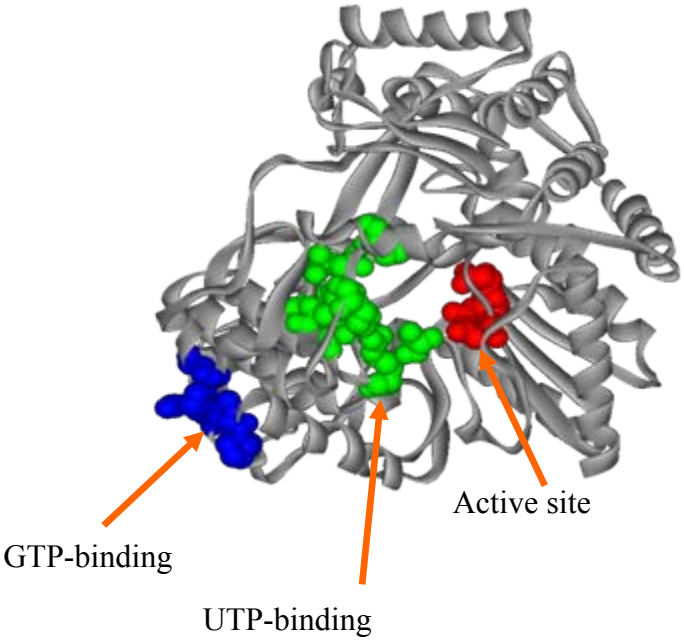
There are 6 main genotypes of hepatitis C virus and a number of subtypes. The most frequent are 1b(40%), 1a (30%), 2a, 2b, 3a.

A gene network for the life cycle of hepatitis C virus

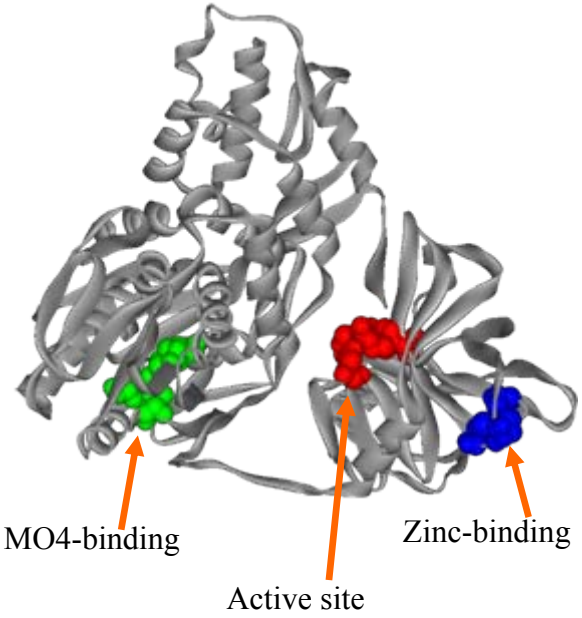


Potential targets for antiHCV drugs

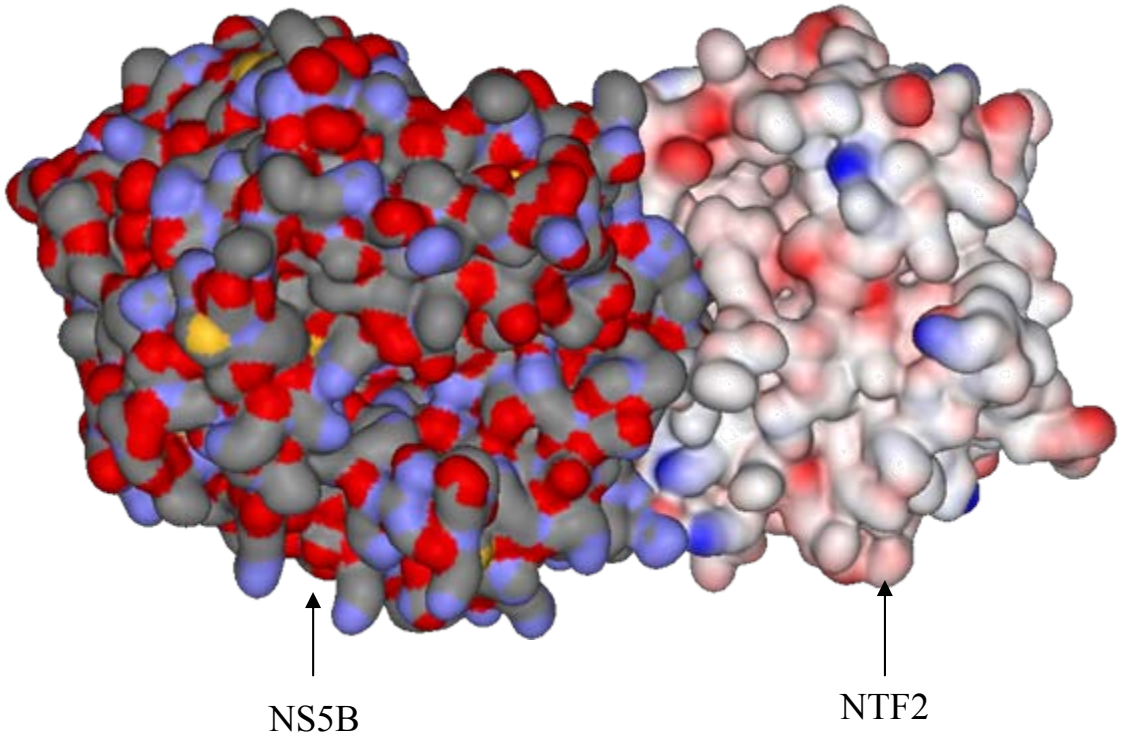
HCV NS5B transferase



HCV NS3 hydrolase

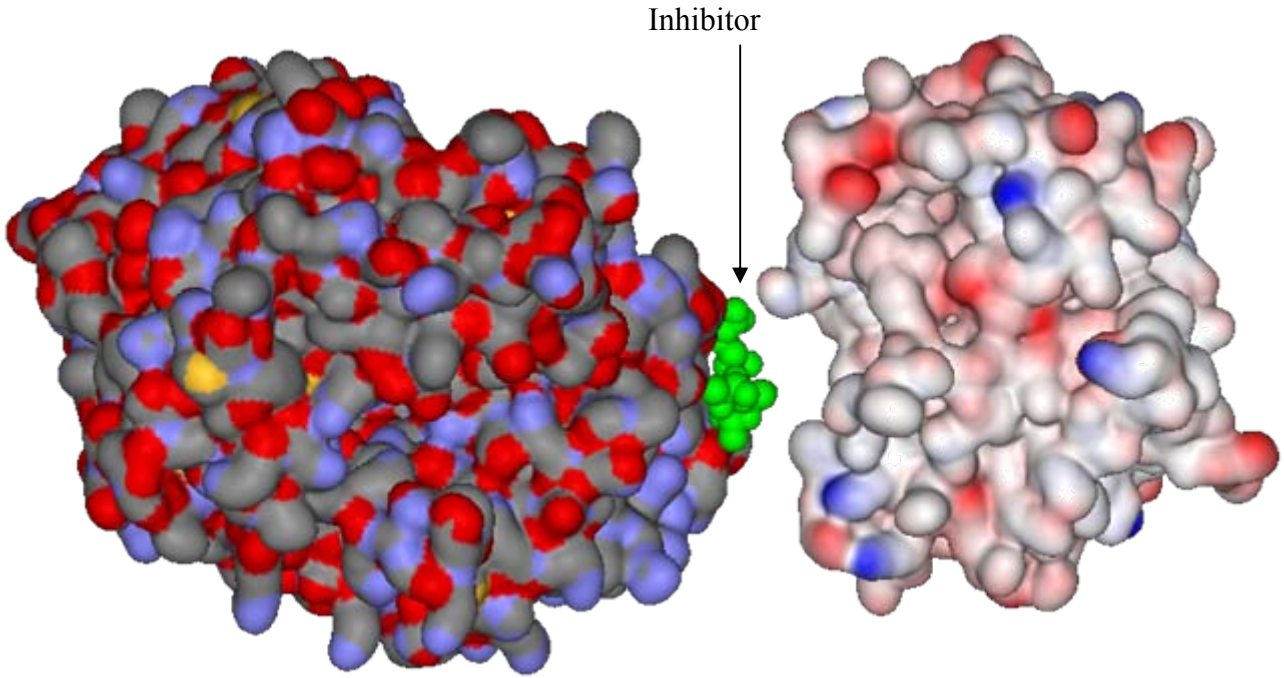


Potential targets for antiHCV drugs



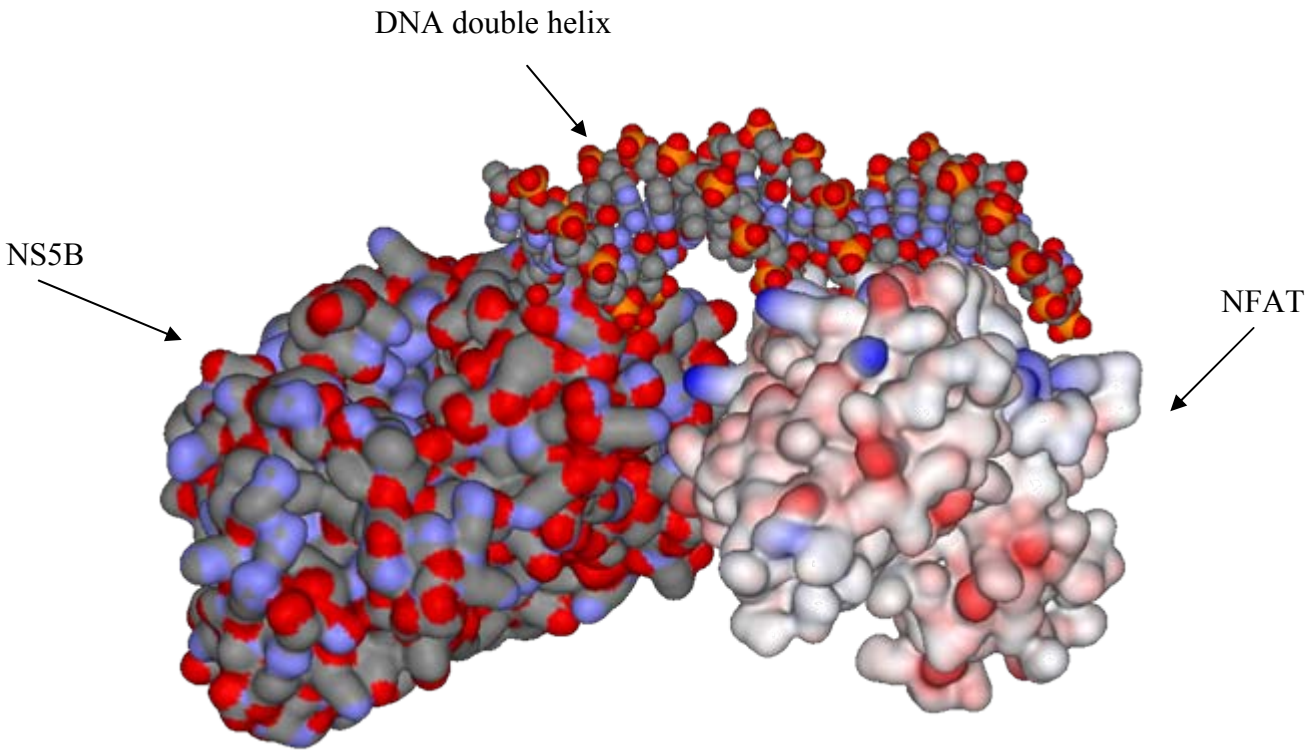
Predicted complex formed by HCV NS5B protein with human NTF2 protein providing transport of proteins to cell nucleus

Potential targets for antiHCV drugs



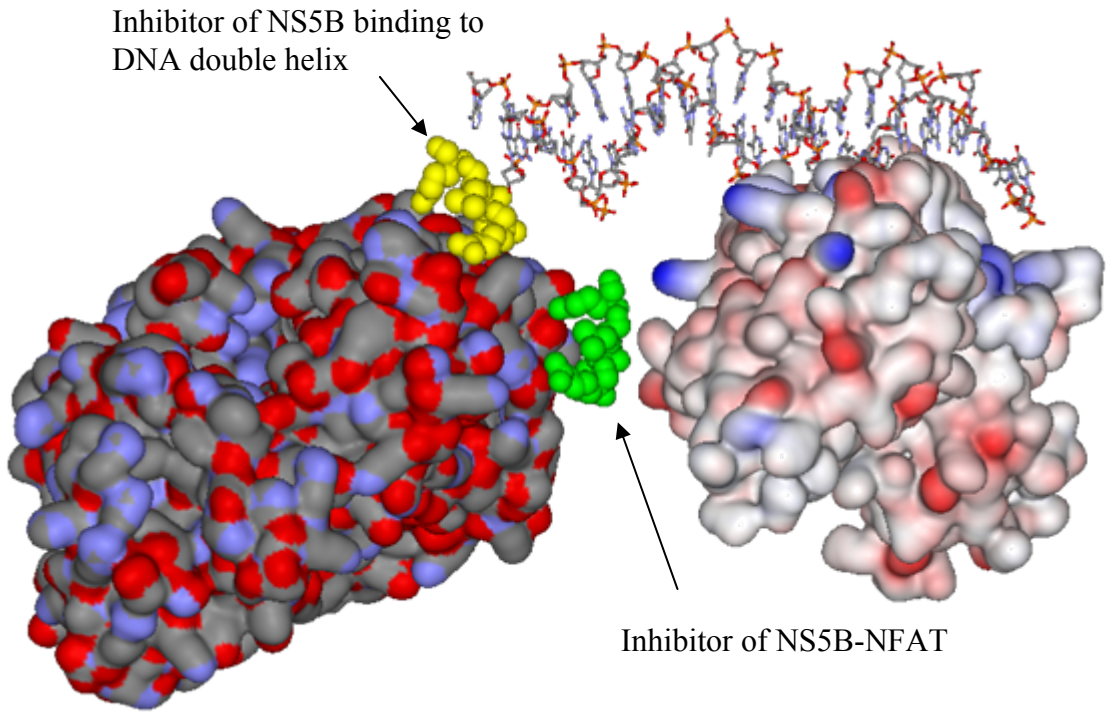
Inhibitor of the formation of NS5B-NTF2 complex prevents NS5B transport into cell nucleus.

Potential targets for antiHCV drugs



Predicted complex formed by HCV NS5B protein with human transcriptional factor NFAT and DNA double helix.

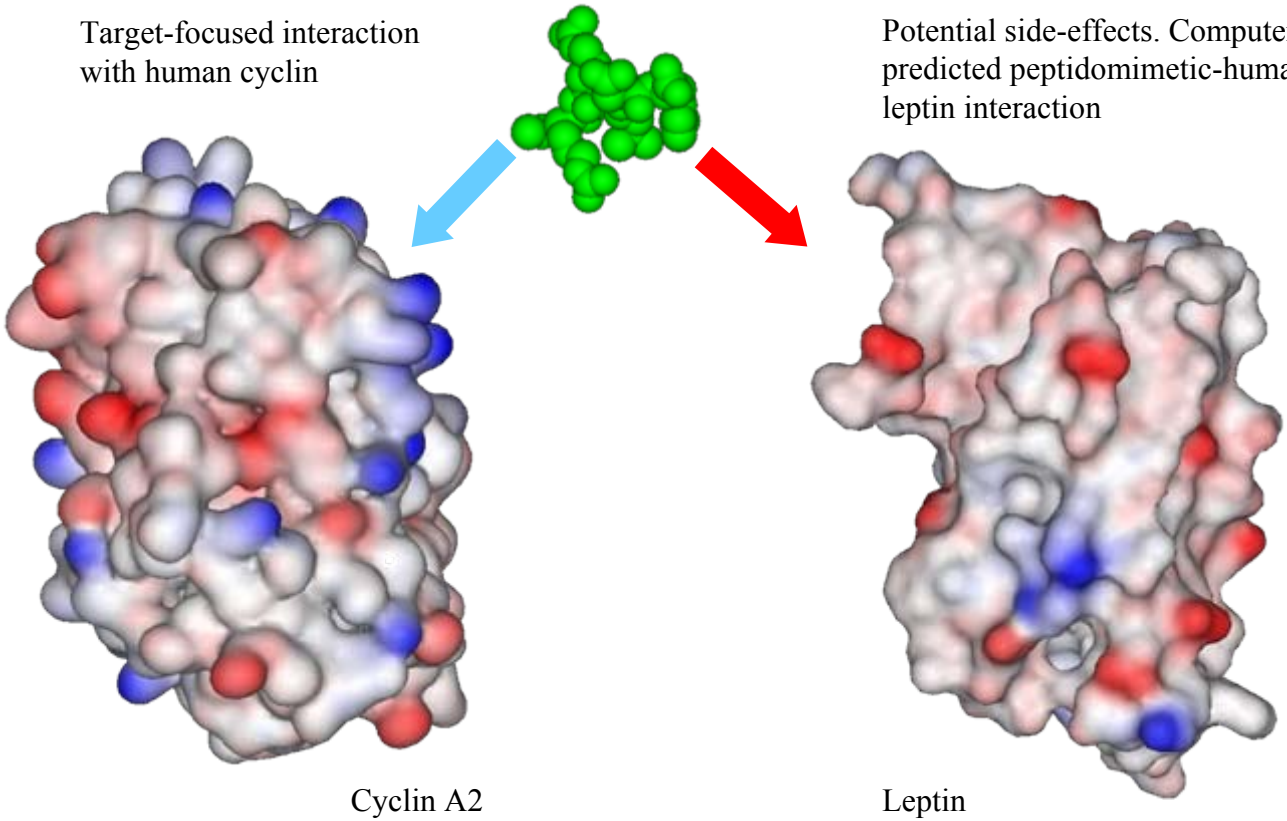
Potential targets for antiHCV drugs



4.3. Search of promising targets for drug action in diseases caused by mutations in the human genome

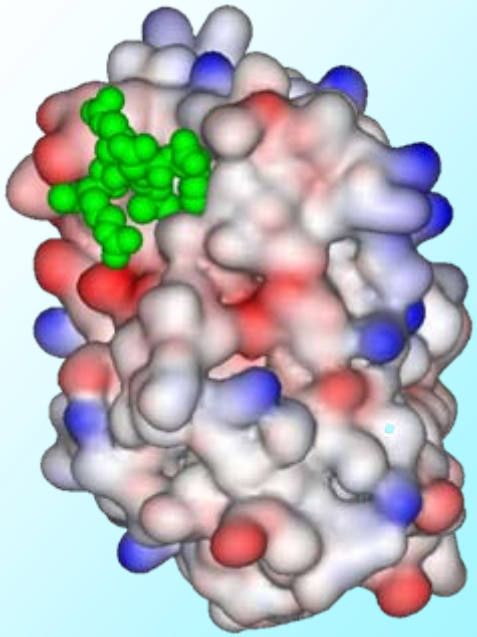
4.3.1. Search for new targets for the drugs causing potential drug side effects

Computation in drug discovery: a peptidomimetic was developed to inhibit cyclin function. It is used as a model to design antitumor drugs.



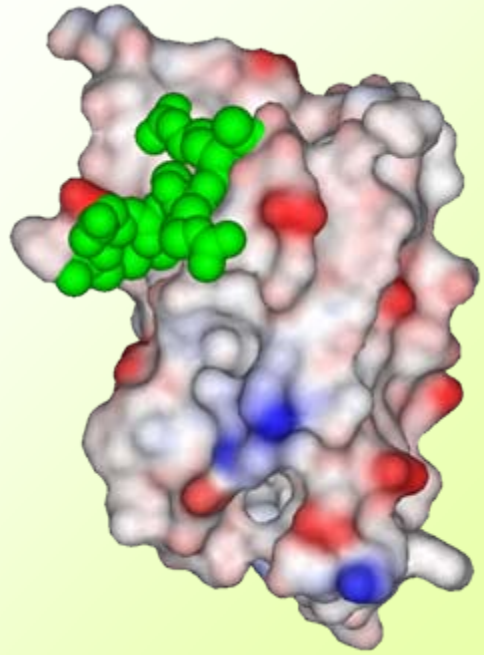
Side-effects of peptidomimetic derived drugs may be associated with their capacity to inhibit leptin function

Target-focused interaction with human cyclin



Cyclin A2

Potential side-effects. Computer predicted peptidomimetic-human leptin interaction

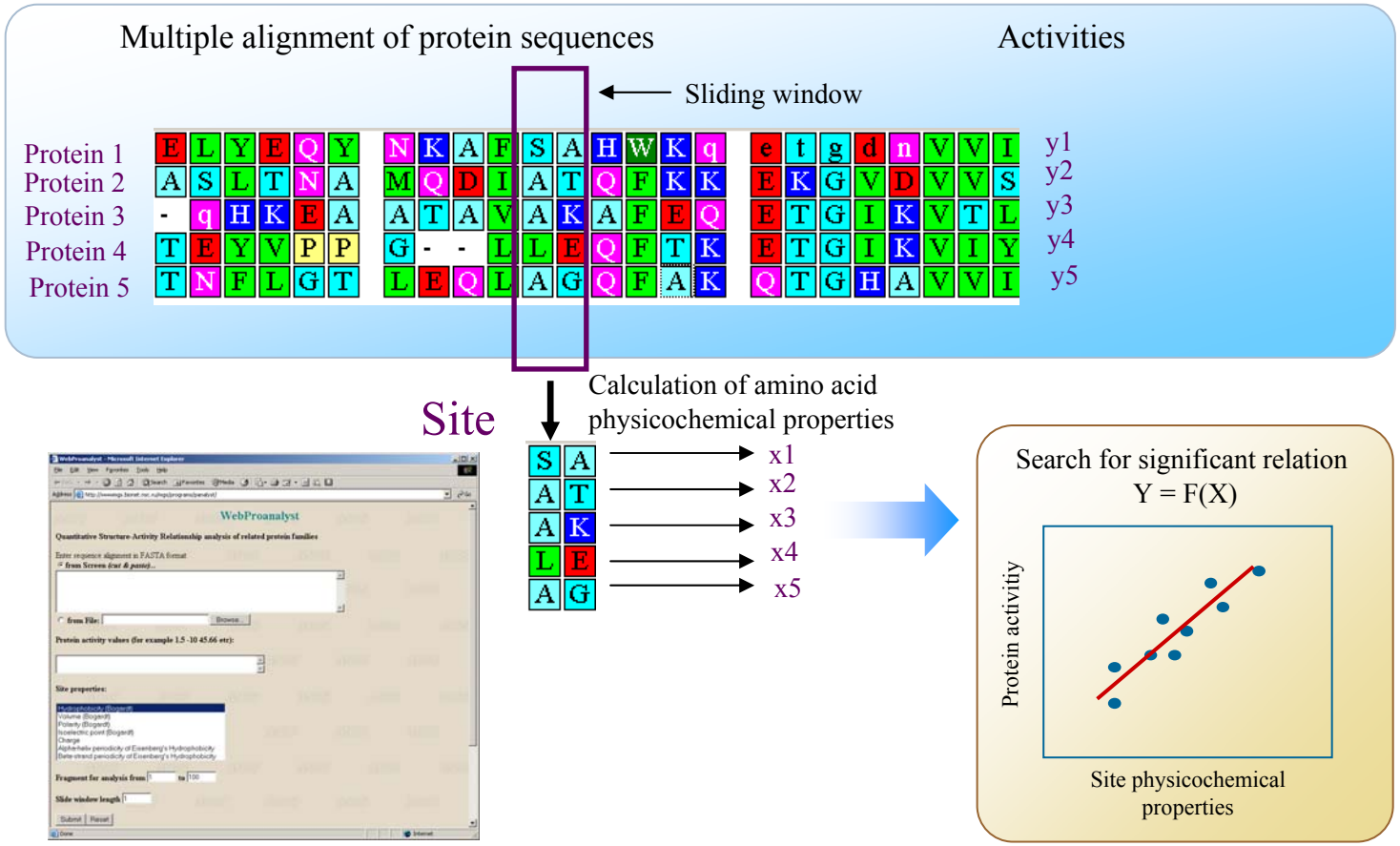


Leptin

4.3.2. Computer design of proteins with improved biomedical properties: promising candidates for medicinal preparations

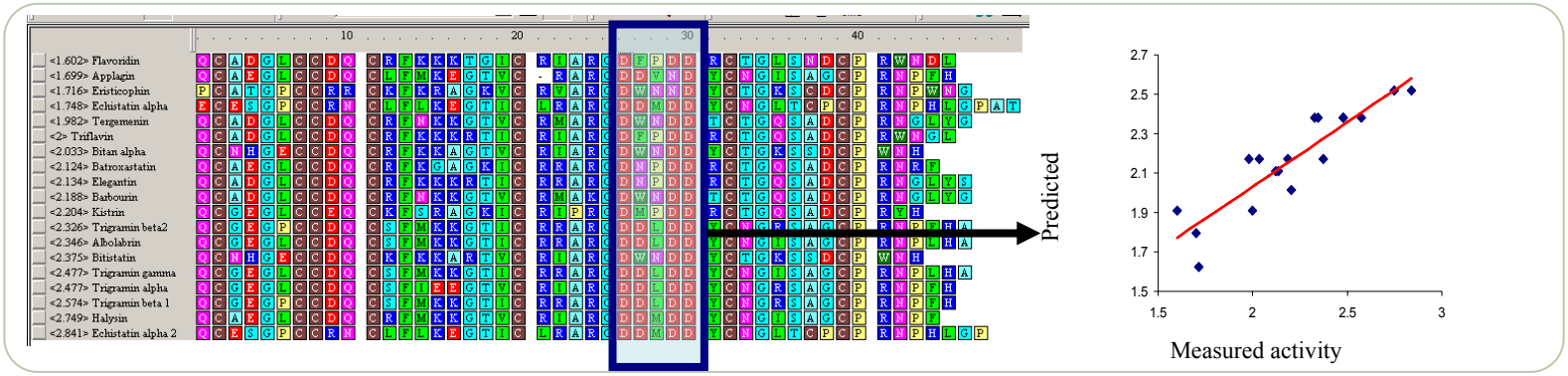
WebProAnalyst: a program for analysis of quantitative structure-activity relationships in homologous protein families ссылка20Ch6-7

http://wwwmgs.bionet.nsc.ru/mgs/programs/panalyst/

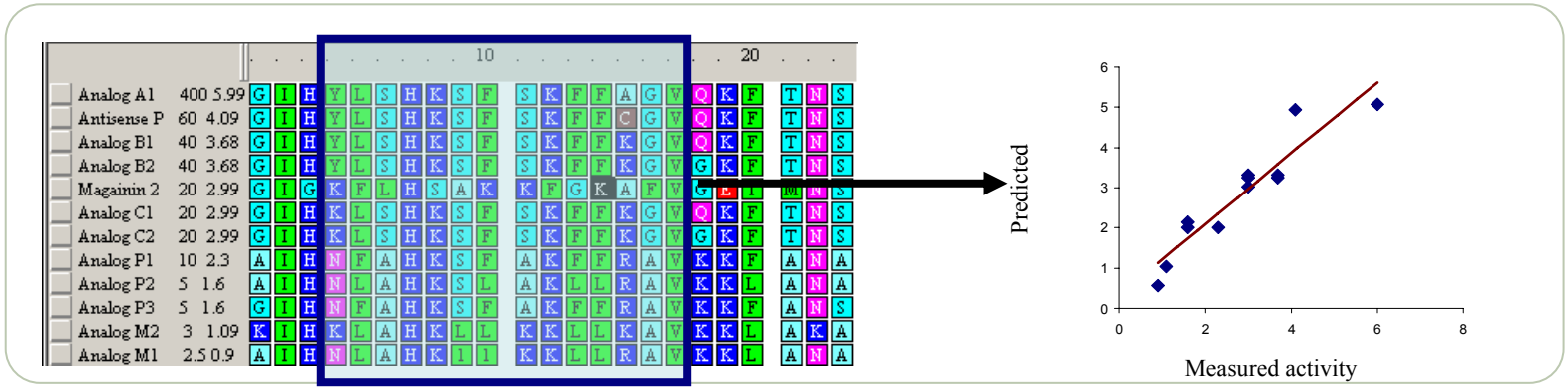


Examples. Quantitative structure-activity relationships in protein families

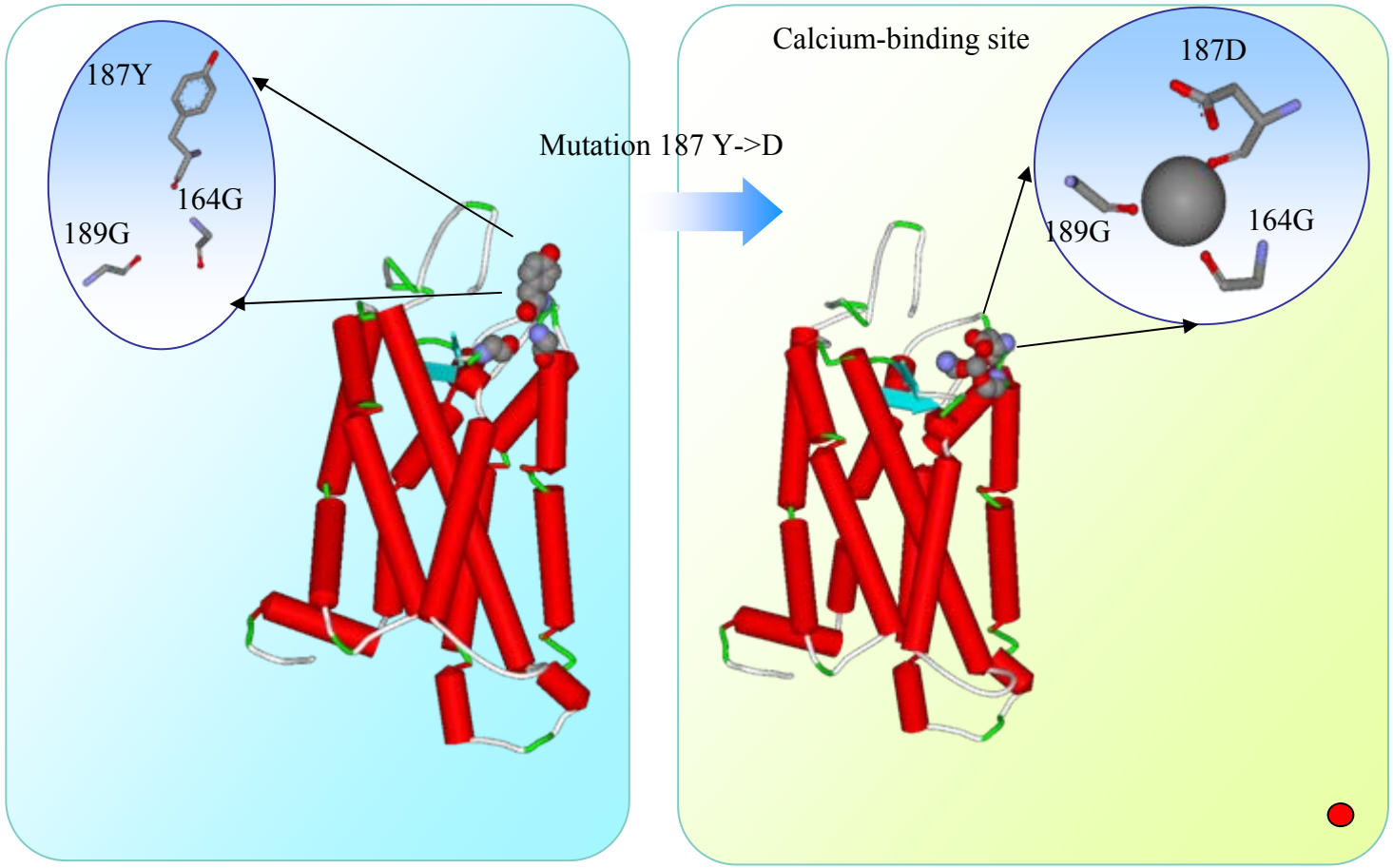
Relation between disintegrin capacity to inhibit platelet aggregation and site properties (charge and hydrophobic moment).



Relation between peptide antimicrobial activity and site hydrophobic moment.



Computer redesign of novel protein functional sites using point mutations



A computer-assisted high-throughout target search strategy was developed on the basis of recognition and analysis of protein functional sites

- Databases accumulating information about spatial structures and physicochemical features of functional sites
- Programs designed to recognize functional sites in protein spatial structures
- Programs designed to reconstruct molecular protein-ligand complexes
- Programs designed to quantitative analysis of sequence-activity relationships in protein families

The offered computer-assisted high-throughput target discovery facilitates and simplifies the resolution of the following issues

- Functional protein annotation
- Uncovering of the molecular mechanisms of impaired protein function
- Planning mutations that directionally affect protein activity
- Identification of potential drug targets

Future implications

- A better understanding of how drugs work
- Search of promising targets for drug action in diseases caused by mutations in the human genome
- Business aspects of lower-risk cheaper drugs

Current strategy: Finding drugs for targets.
Inverted cheaper strategy: finding targets for drugs. Better targets make better drugs.

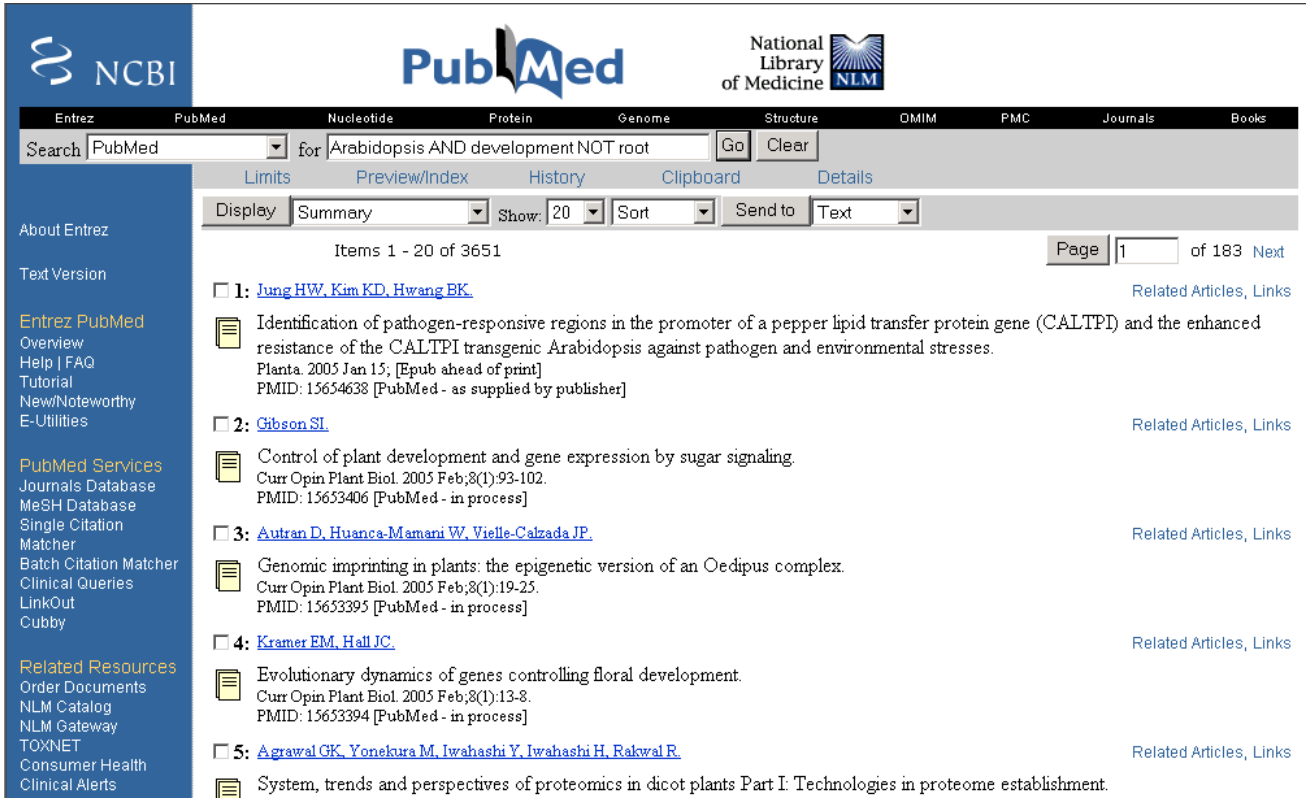
Publications

1. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res.,* 2005, V33, D183–D187
2. Ivanisenko VA, Eroshkin AM, Kolchanov NA. (2005) WebProAnalyst: an interactive tool for analysis of quantitative structure-activity relationships in protein families. Nucleic Acids Res. V. 33, W99-W104.
3. Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res.*, 32, W549-W554
4. Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A., Ivanisenko, L.N., Debelov, V.A., Matsokin, A.M. PDBSITESCAN: a program searching for functional sites in protein 3D structures. In: Bioinformatics of genome regulation and structure. Ed. By N. Kolchanov and R. Hofestaedt, *Kluwer Academic Publishers*, Boston/Dordrecht/London, 2004, pp. 185-192.
5. Afonnikov DA, Kolchanov NA. (2004) CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences, *Nucleic Acids Res*, V.32 , W64-W68.
6. Ivanisenko V.A., Pintus S.S., Krestyanova M.A., Demenkov P.S., Znobisheva E.K., Ivanov E.E., Grigorovich D.A. PDBSITE, PDBLIGAND and PDBSITESCAN: a computational workbench for the recognition of the structural and functional determinants in protein tertiary structures combined with protein draft docking. Proc. of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure. 2004. V.1. P. 269-273.
7. Pintus S.S., Ivanisenko V.A. A molecular mechanism for the structure-functional alterations in mutant forms of human p53 protein. Proc. of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure. 2004. V.1. P. 338-342.

Chapter 5
PLANT DEVELOPMENT: COMPUTER ANALYSIS AND MODELING

5.1. AGNS: Arabidopsis GeneNet Supplementary database

The number of publications related to phenotypic abnormalities and expression of the arabidopsis genes during development of above organs greatly increased



NCBI

PubMed

National Library of Medicine NLM

Entrez PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for Arabidopsis AND development NOT root Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show: 20 Sort Send to Text

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts

Items 1 - 20 of 3651 Page 1 of 183 Next

1: Jung HW, Kim KD, Hwang BK. Related Articles, Links

Identification of pathogen-responsive regions in the promoter of a pepper lipid transfer protein gene (CALTPI) and the enhanced resistance of the CALTPI transgenic Arabidopsis against pathogen and environmental stresses.
Planta. 2005 Jan 15; [Epub ahead of print]
PMID: 15654638 [PubMed - as supplied by publisher]

2: Gibson SI. Related Articles, Links

Control of plant development and gene expression by sugar signaling.
Curr Opin Plant Biol. 2005 Feb;8(1):93-102.
PMID: 15653406 [PubMed - in process]

3: Autran D, Huanca-Mamani W, Vielle-Calzada JP. Related Articles, Links

Genomic imprinting in plants: the epigenetic version of an Oedipus complex.
Curr Opin Plant Biol. 2005 Feb;8(1):19-25.
PMID: 15653395 [PubMed - in process]

4: Kramer EM, Hall JC. Related Articles, Links

Evolutionary dynamics of genes controlling floral development.
Curr Opin Plant Biol. 2005 Feb;8(1):13-8.
PMID: 15653394 [PubMed - in process]

5: Agrawal GK, Yonekura M, Iwahashi Y, Iwahashi H, Rakwal R. Related Articles, Links

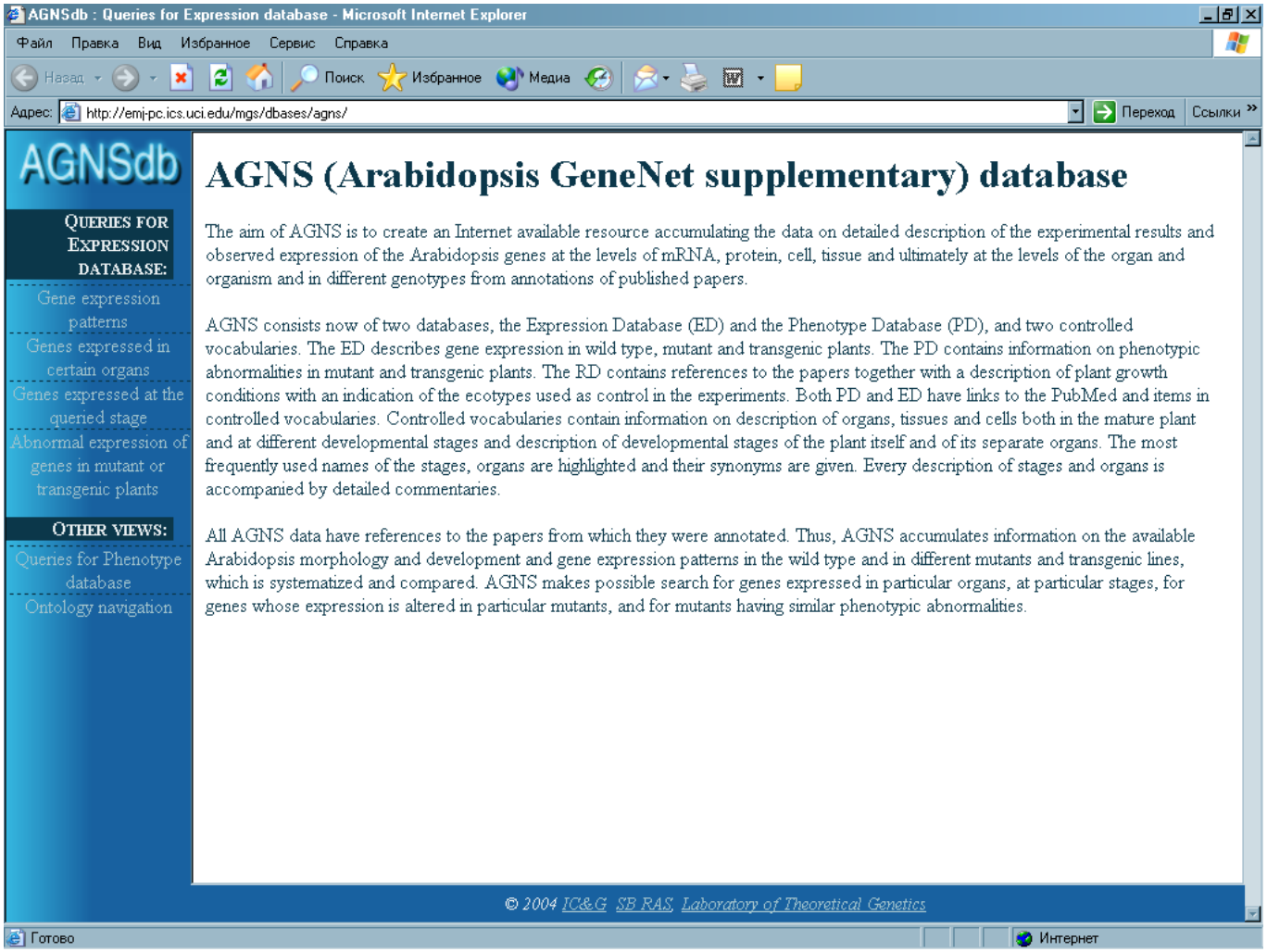
System, trends and perspectives of proteomics in dicot plants Part I: Technologies in proteome establishment.

Our goal is the continued development and maintenance of the AGNS, an Internet available resource for accumulation of experimental data on gene expression during development of above organs in Arabidopsis.

By integrating data from the published papers and providing appropriate user tools, AGNS allows the user to examine patterns of gene expression in different genetic backgrounds and explore the genetic programs that underlie normal development, and dysregulations of development leading to phenotype abnormalities.

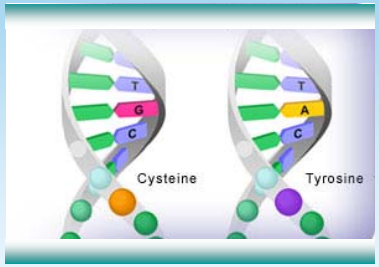
As data accumulate, AGNS provides increasingly complete and refined information. A comprehensive hierarchy of organ structure and developmental stages provides a standard for describing the time and location of gene expression or formation of phenotype abnormalities .

http://wwwmgs2.bionet.nsc.ru/agns/

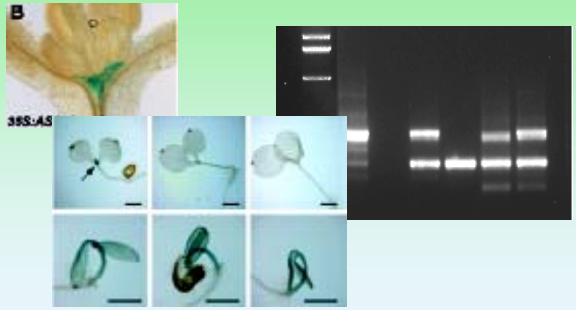

AGNSdb : Queries for Expression database - Microsoft Internet Explorer
Файл Правка Вид Избранное Сервис Справка
Назад Поиск Избранное Медиа
Адрес: http://emj-pc.ics.uci.edu/mgs/dbases/agns/
Переход Ссылки
AGNSdb
QUERIES FOR EXPRESSION DATABASE:
Gene expression patterns
Genes expressed in certain organs
Genes expressed at the queried stage
Abnormal expression of genes in mutant or transgenic plants
OTHER VIEWS:
Queries for Phenotype database
Ontology navigation
AGNS (Arabidopsis GeneNet supplementary) database
The aim of AGNS is to create an Internet available resource accumulating the data on detailed description of the experimental results and observed expression of the Arabidopsis genes at the levels of mRNA, protein, cell, tissue and ultimately at the levels of the organ and organism and in different genotypes from annotations of published papers.
AGNS consists now of two databases, the Expression Database (ED) and the Phenotype Database (PD), and two controlled vocabularies. The ED describes gene expression in wild type, mutant and transgenic plants. The PD contains information on phenotypic abnormalities in mutant and transgenic plants. The RD contains references to the papers together with a description of plant growth conditions with an indication of the ecotypes used as control in the experiments. Both PD and ED have links to the PubMed and items in controlled vocabularies. Controlled vocabularies contain information on description of organs, tissues and cells both in the mature plant and at different developmental stages and description of developmental stages of the plant itself and of its separate organs. The most frequently used names of the stages, organs are highlighted and their synonyms are given. Every description of stages and organs is accompanied by detailed commentaries.
All AGNS data have references to the papers from which they were annotated. Thus, AGNS accumulates information on the available Arabidopsis morphology and development and gene expression patterns in the wild type and in different mutants and transgenic lines, which is systematized and compared. AGNS makes possible search for genes expressed in particular organs, at particular stages, for genes whose expression is altered in particular mutants, and for mutants having similar phenotypic abnormalities.
© 2004 IC&G SB RAS, Laboratory of Theoretical Genetics
Готово
Интернет

Arabidopsis GeneNet Supplementary database –AGNS

Sequence database

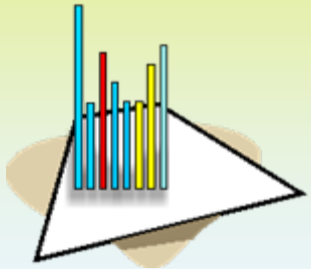


Expression database

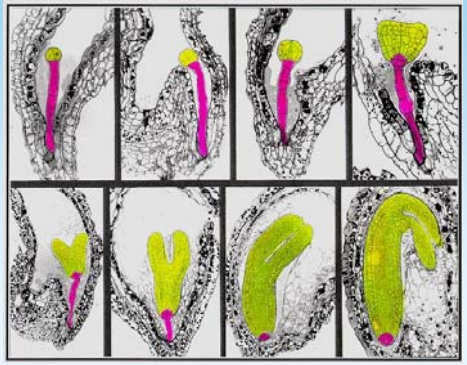


Phenotype database

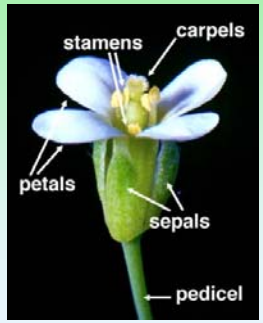
Reference database



Developmental stages: vocabulary



Organs, tissues, cells: vocabulary



AGNS: the sequence database, an example of the card

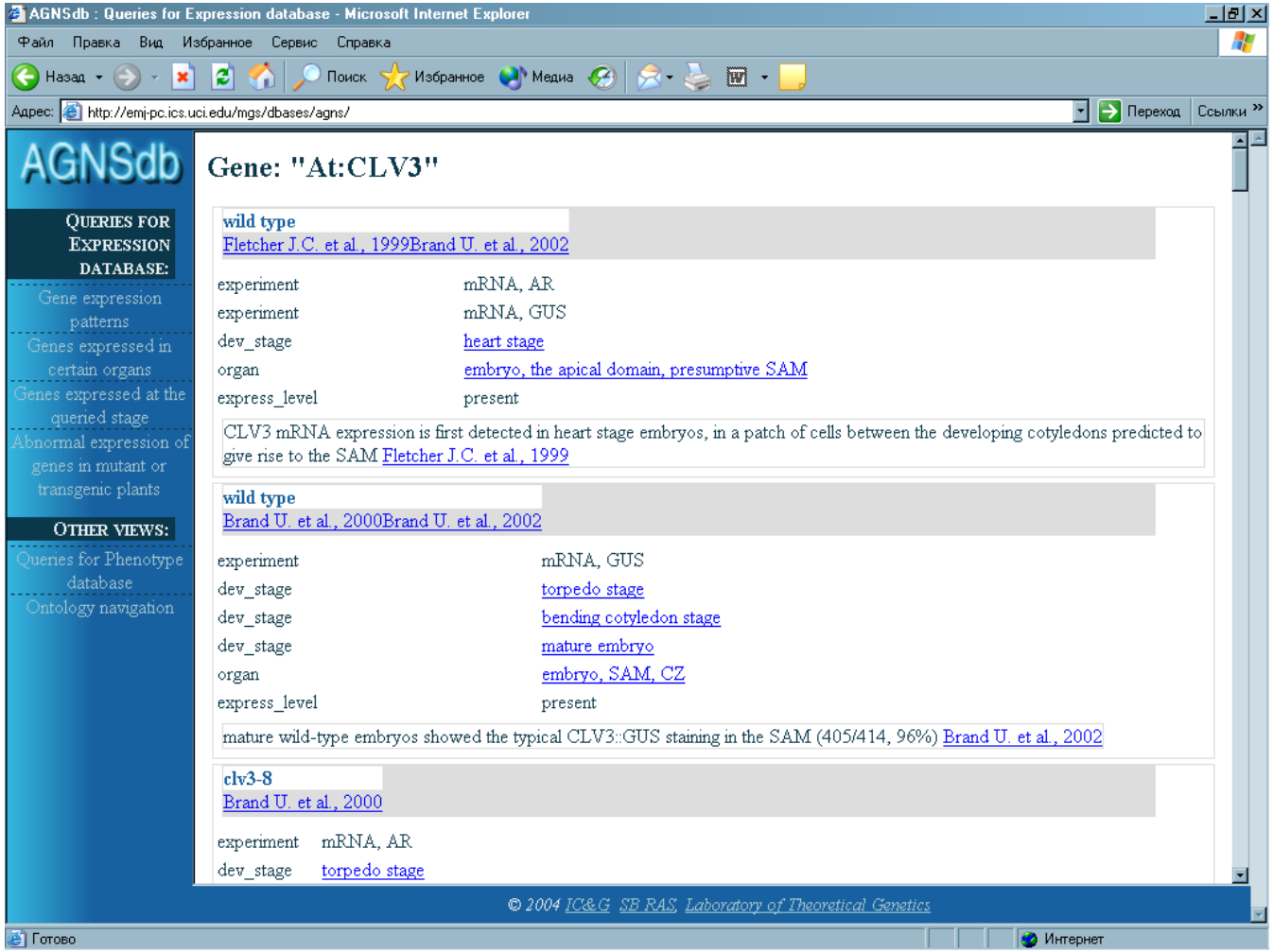
ID At:CLV3
XX
OS mouse-ear cress, Arabidopsis thaliana
NG CLAVATA3
SY
BI GenBank; NM_128283; SR: 299
BI GenBank; NM_201812; SR: 138
BI GenBank; AF126009; SR: 58
AI At2g27250
PF CLE family of plant polypeptides
XX

MA clv3-1 [Fletcher J. C. et al., 1999]
BG Ler
AS I
MR NM_128283; exon, transition G-A, SR; +263
MR NM_201812; intron, transition G-A, SR; +257
MR AF126009; exon 3, transition G-A, SR; +537
RC the independently derived *clv3-1* and *clv3-5* have intermediate phenotypes and contain a G to A transition at position +266 relative to the translation initiation site [Fletcher J. C. et al., 1999]
XX
MA clv3-2 [Fletcher J. C. et al., 1999]
BG Ler
AS S
RC clv3-2 and clv3-4 both contain breakpoints occurring between the Mfe I and Dra I restriction sites flanking the third exon [Fletcher J. C. et al., 1999]
XX
MA clv3-3 [Fletcher J. C. et al., 1999]
BG Ws-2
AS W
RC The *clv3-3* allele is caused by T-DNA integration and confers a weak *clv3* phenotype [Feldman K.A. and Marks M.D., 1987]. The *clv3-3* T-DNA insertion site lies 175 base pairs (bp) downstream of the polyadenylate [poly(A)] addition site, potentially disrupting an enhancer element. The *En-1* element in *clv3-7* inserted in the second intron close to the intron–exon 3 boundary [Fletcher J. C. et al., 1999]
XX

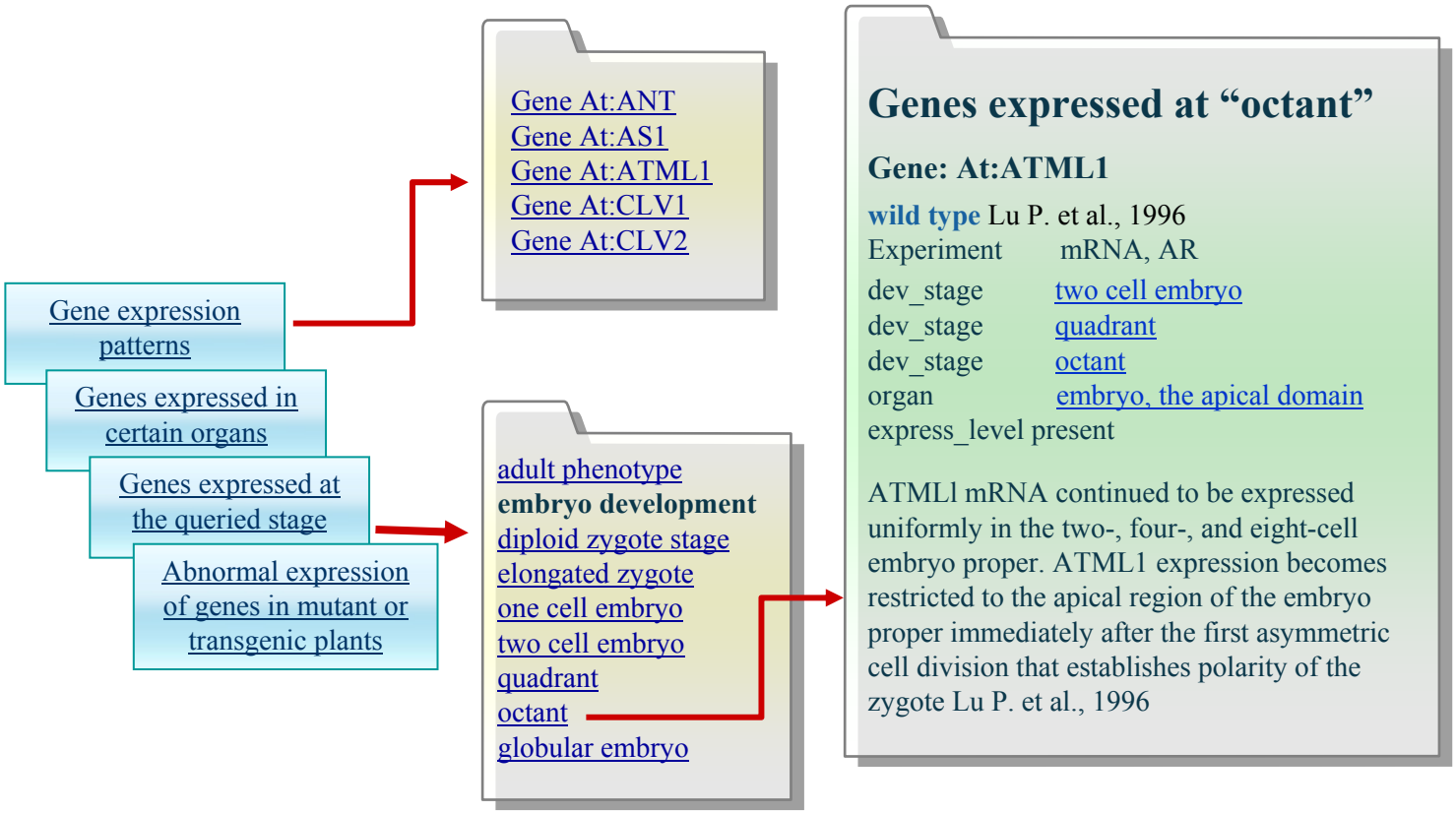
AGNS: the sequence database, an example of the card

ID At::AGO1
XX
MA wild type [Lynn K. et al., 1999]
RT mRNA, AR
RD globular embryo
RO embryo
RL high
RL present
XX

MA wild type [Lynn K. et al., 1999]
RT mRNA, AR
RD globular embryo
RO suspensor
RL present
XX
MA wild type [Lynn K. et al., 1999]
RT mRNA, AR
RD torpedo stage
RO embryo
RL present
RC AGO1 mRNA does not accumulate differentially in the adaxial region of the cotyledons unlike ZLL mRNA [Lynn K. et al., 1999]
XX
MA wild type [Bohmert K. et al., 1998]
RT mRNA, blot
RD seedling
RO seedling
RL present
RC during 10 days of seedling development from the cotyledon stage (8 days) until the development of 6–10 secondary leaves (19 days), no major changes in the expression of AGO1 were detected [Bohmert K. et al., 1998
XX
//


AGNSdb : Queries for Expression database - Microsoft Internet Explorer
Файл Правка Вид Избранное Сервис Справка
Назад Поиск Избранное Медиа
Адрес: http://emj-pc.ics.uci.edu/mgs/dbases/agns/
Переход Ссылки
AGNSdb
QUERIES FOR EXPRESSION DATABASE:
Gene expression patterns
Genes expressed in certain organs
Genes expressed at the queried stage
Abnormal expression of genes in mutant or transgenic plants
OTHER VIEWS:
Queries for Phenotype database
Ontology navigation
Gene: "At:CLV3"
wild type
Fletcher J.C. et al., 1999Brand U. et al., 2002
experiment mRNA, AR
experiment mRNA, GUS
dev_stage heart stage
organ embryo, the apical domain, presumptive SAM
express_level present
CLV3 mRNA expression is first detected in heart stage embryos, in a patch of cells between the developing cotyledons predicted to give rise to the SAM Fletcher J.C. et al., 1999
wild type
Brand U. et al., 2000Brand U. et al., 2002
experiment mRNA, GUS
dev_stage torpedo stage
dev_stage bending cotyledon stage
dev_stage mature embryo
organ embryo, SAM, CZ
express_level present
mature wild-type embryos showed the typical CLV3::GUS staining in the SAM (405/414, 96%) Brand U. et al., 2002
clv3-8
Brand U. et al., 2000
experiment mRNA, AR
dev_stage torpedo stage
© 2004 IC&G SB RAS, Laboratory of Theoretical Genetics
Готово
Интернет

Queries for the Expression database




AGNSdb : Queries for Expression database - Microsoft Internet Explorer
Файл Правка Вид Избранное Сервис Справка
Назад Поиск Избранное Медиа
Адрес: http://emj-pc.ics.uci.edu/mgs/dbases/agns/
Переход Ссылки
AGNSdb
QUERIES FOR EXPRESSION DATABASE:
Gene expression patterns
Genes expressed in certain organs
Genes expressed at the queried stage
Abnormal expression of genes in mutant or transgenic plants
OTHER VIEWS:
Queries for Phenotype database
Ontology navigation
Genes with abnormal expression in "clv3-1"
Gene: At:CLV1
clv1-10
Trotochaud A.E. et al., 1999
clv2-2
Jeong S. et al., 1999
clv2-2
Jeong S. et al., 1999
clv3-1
Trotochaud A.E. et al., 1999
clv3-2
Trotochaud A.E. et al., 1999
experiment protein, blot
experiment protein, ELISA
dev_stage flowering
organ shoot apex, 450 kD protein complex
express_level none
anomaly absent
clv1-10, clv3-1, clv3-2, CLV1 protein was detected exclusively in the 185 kD complex Trotochaud A.E. et al., 1999
clv2-2, clv2-3, CLV1 protein is present in a novel, higher molecular mass complex of ~600 kD Jeong S. et al., 1999
Gene: At:CLV3
© 2004 IC&G SB RAS, Laboratory of Theoretical Genetics
Интернет

AGNS: the phenotype database, an example of the card (main fields)

MA cuc2 [Aida M. et al., 1997]
MA cuc2 cuc1/+ [Aida M. et al., 1997]
MA cuc2 cuc1 [Aida M. et al., 1997]
MA stm-1 [Barton M.K. and Poethig R.S., 1993] [Clark S.E. et al., 1996] [Byrne M.E. et al., 2002]
MA stm-1 as-1 knat1-bp [Byrne M.E. et al., 2002]
MA stm-1 bop-1 [Ma C.M. et al., 2003]
MA stm-1 clv3-2 [Clark S.E. et al., 1996]
MA stm-2 blr [Byrne M. E. et al., 2003]
MA stm-2 mgo-1 [Laufs P. et al., 1998, D]
MA stm-2 spy-5 [Hay A. et al., 2002]
MA stm-2 wus-1 [Endrizzi K. et al., 1996]
MA stm-2 zll-3 [Endrizzi K. et al., 1996]
MA stm-5 [Endrizzi K. et al., 1996] [Laufs P. et al., 1998, D]
MA stm-5/stm-2 [Endrizzi K. et al., 1996]
MA stm-5 mgo1 [Laufs P. et al., 1998, D]
MA stm-5 wus-1 [Endrizzi K. et al., 1996]
MA stm-5 zll-3 [Endrizzi K. et al., 1996]
MA stm-6 [Endrizzi K. et al., 1996]
MA stm-6 wus-1 [Endrizzi K. et al., 1996]
MA TCP4::mTCP4 [Palatnik J.F. et al., 2003]
MA 35S::mTCP4 [Palatnik J.F. et al., 2003]
MA 35S::mTCP4 jaw-D [Palatnik J.F. et al., 2003]
MA tpl-1 [Long J.A. et al., 2002]
MA zll [Moussian B. et al., 1998]
MA zll-pnh-2 [Lynn K. et al., 1999]
MA zll-pnh-4 [Lynn K. et al., 1999]
MA zll-pnh-8 [Lynn K. et al., 1999]
MA zll-pnh-9 [Lynn K. et al., 1999]
MA zll-pnh-11 [Lynn K. et al., 1999]
MA zll-3 wus-1 [Moussian B. et al., 2003]
RD mature embryo
RD seedling
RO primary SAM
FL absent or very strongly reduced and not restored at germination

AGNS: phenotype database, an example of the card (comments)

RC cuc2, in 0.08% of plants, all these seedlings have cotyledons fused along one side [Aida M. et al., 1997]
RC cuc2 cuc1, in the mature embryo both sides of the cup-shape cotyledons directly met at their bases, cells in this region are vacuolating, and the relative sizes of nuclei were the same as the nuclei in cells around, no dead cells were observed; in seedlings large, highly vacuolated cells are the SAM position [Aida M. et al., 1997]
RC cuc2 cuc1 stm-1, in seedlings large, highly vacuolated cells are the SAM position [Aida M. et al., 1997]
RC stm-1, bases of cotyledons meet at an acute angle, cells at junction have storage grains thus belong to cotyledons, the corresponding position is occupied by cells indistinguishable from surrounding cells in the hypocotyls and cotyledons, although they may be somewhat smaller in size, no tissues resembling meristems were found in dissected 16 day-old plants [Barton M.K. and Poethig R.S., 1993] [Clark S.E. et al., 1996] [Byrne M.E. et al., 2002]
RC stm-1 as-1 bp, shown in 8 day-old seedlings [Byrne M.E. et al., 2002]
RC stm-1 clv3-2, in some plants even in 8 –day-old seedlings there was no evidence of small, densely staining cells that might indicate SAM [Clark S.E. et al., 1996]
RC stm-2, the corresponding position is occupied by a variable number of meristem-like densely staining cells [Clark S.E. et al., 1996], an average number of SAM cells is 11.5 with about 4 cells in L1 [Endrizzi K. et al., 1996]
RC stm-2 clv3-2, some small densely staining cells were often present above the junction of the vascular elements [Clark S.E. et al., 1996]
RC stm-2 mgo1, in 82% of plants on 6 dag, in 28-30% - on 12-19 dag [Laufs P. et al., 1998, D]
RC stm-2 wus-1, stm-6 wus-1, the number of small densily staining cells appeared more reduced than in either single parental nutant in the mature embryo, none of these seedlings initiated a primary SAM and formed leaves [Endrizzi K. et al., 1996]
RC stm-2 zll-3, lacked a group of small densely staining cells [Endrizzi K. et al., 1996]
RC stm-5, no SAM is visible in the apex enclosed by fused petioles, up to 10 dag in most seedlings no small densely stained cells or leaf primordia were observed. The cells of apex were typically larger and cells are more vacuolated than in the wild type but smaller than differentiated cells, most seedling did not produce any further organs and eventually senesced [Endrizzi K. et al., 1996] stm-5, in 89-93% of plants on 6-12 dag, in 79% - on 19 dag [Laufs P. et al., 1998, D]
RC stm-5 zll-3, indistinguishable from the stm-5 [Endrizzi K. et al., 1996]
RC tpl-1, SAM is absent in all embryos, in 3% of plants grown at 24° and in 64% of plants grown at 28° roots developed at the place of SAM [Long J.A. et al., 2002]
RC wus-1, the corresponding position is occupied by a few cells, that were slightly larger, more vacuolated and lacked prominent nuclei compared to the wild-type SAM. In seedlings the cells in the apex were slightly larger and stained less intensely than cells of a wild-type SAM, but were smaller and less vacuolated relative to differentiated cortex or epidermal cells [Laux T. et al., 1996]
RC wus-1 zll-3, similar to wus-1 and zll-3 [Endrizzi K. et al., 1996]
RC zll-3, SAM reduced in size and flat and seedlings without shoot meristem activity [Endrizzi K. et al., 1996]
RC zll, in about 30% of embryos, percentage depends on allelle [Moussian B. et al., 1998]
RC 35S::mTCP4, in 81% plants[Palatnik J.F. et al., 2003]
RC TCP4::mTCP4, in 61% plants [Palatnik J.F. et al., 2003]
XX

Queries for the Phenotype database

Mutations resulted in phenotypic abnormalities of the selected organs

primary SAM
cauline leaf axillary SAM
floral meristem
floral primordium
flower
flower whorl l, sepals

Mutations resulted in phenotypic abnormalities in "cauline leaf axillary SAM"

ATML1::CLV3 ATML1::CLV1 Lenhard M. and Laux T., 2003
ATML1::CLV3 ATML1::clv1-4 Lenhard M. and Laux T., 2003

dev_stage	adult phenotype
organ	cauline leaf axillary SAM
anomaly	not developed

Phenotypes of the mutants

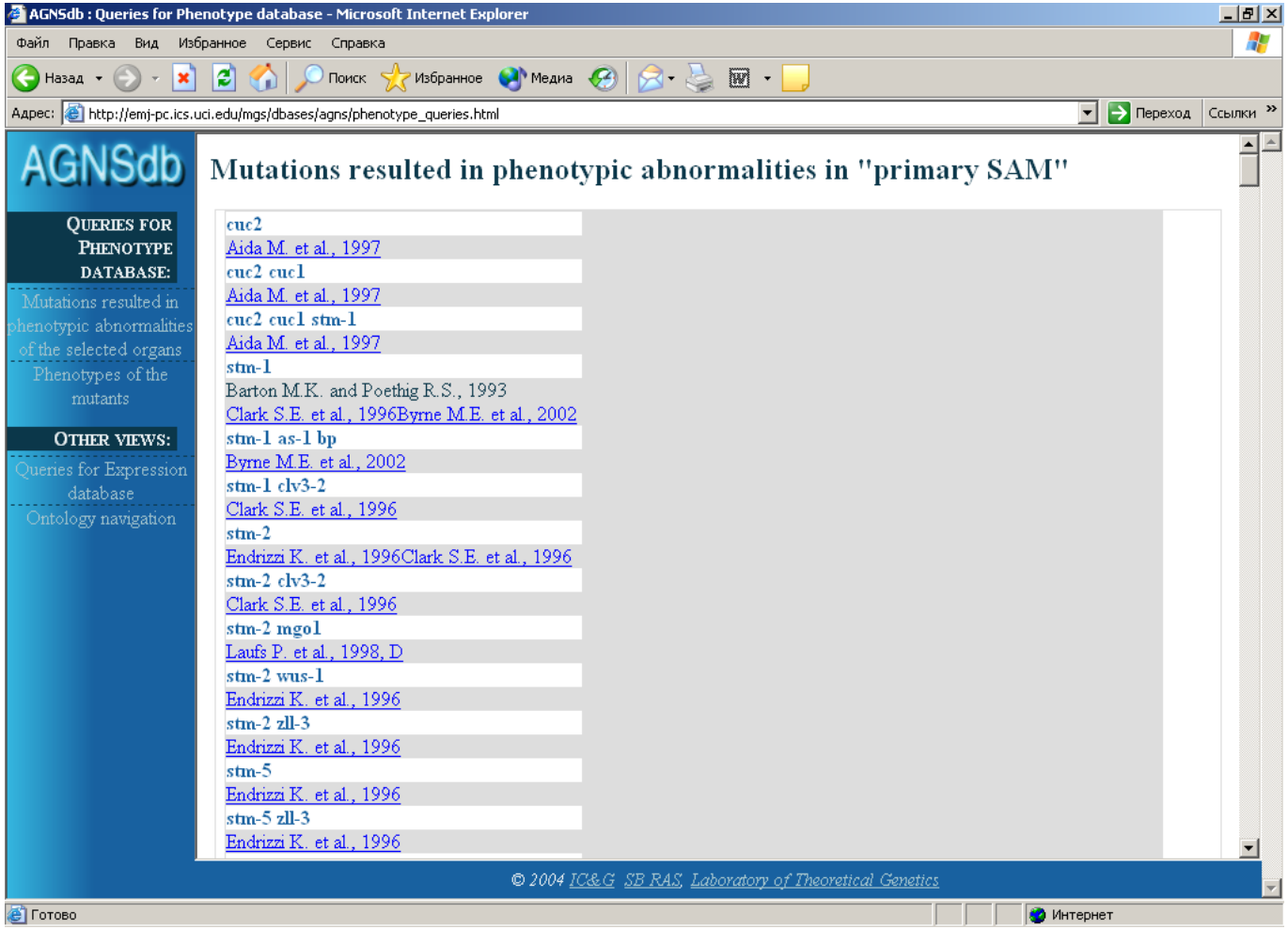
Single mutants amp1-pt ago1-7 ago1-8 ap2-1 ap3-1 ap3-4 35S::AP3 35S::AP3 ap3-3 35S::AS AtCAP-E1 clf clv1 clv1-1 clv1-2 clv1-3 clv1-4 clv1-5 clv1-6 clv1-7 clv2 clv2-1 clv2-2 clv2-3 clv2-4 clv3-1

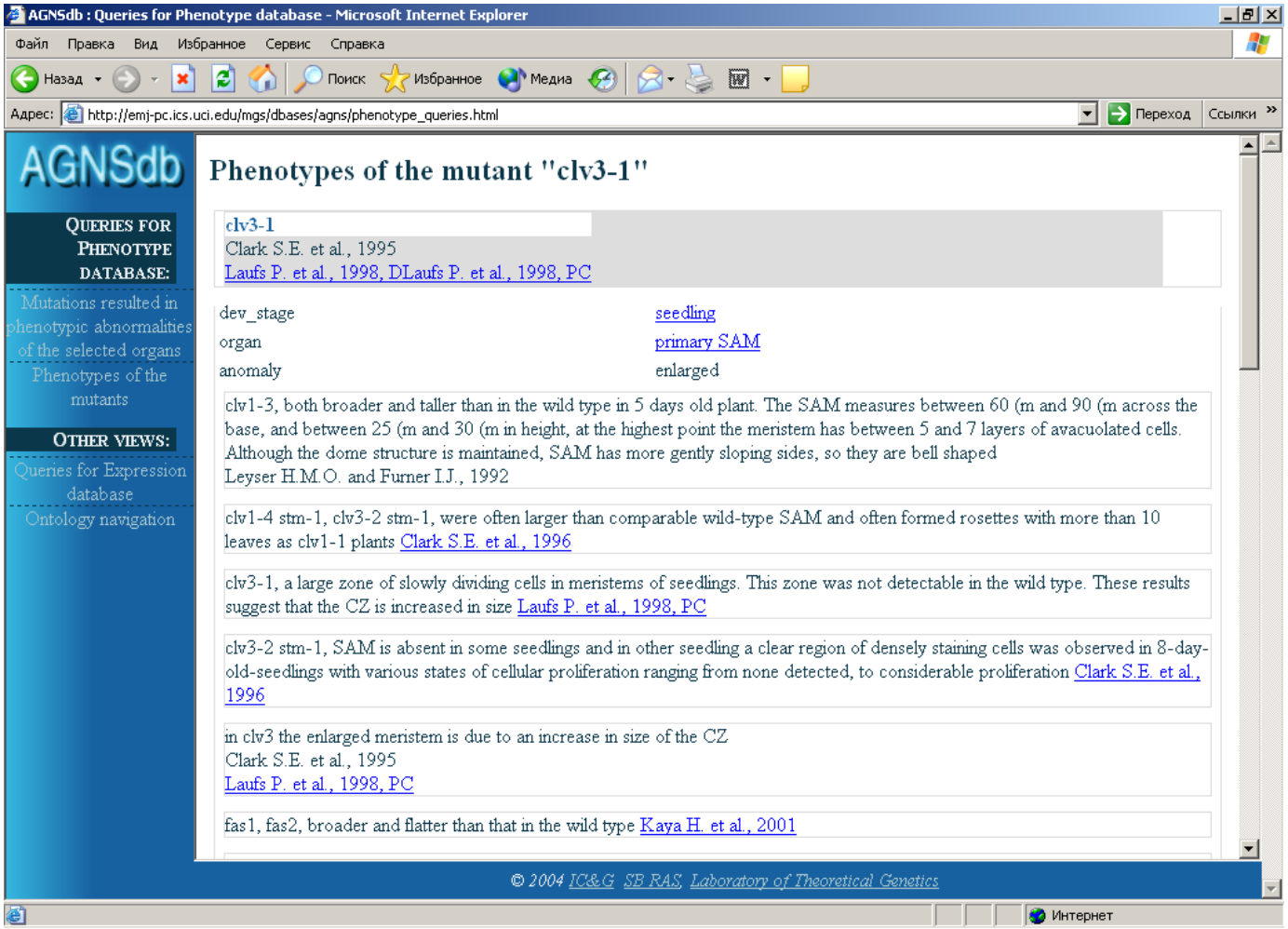
Phenotype of the mutant "ap2-1"

ap2-1 Bowman J.L. et al., 1989

dev_stage	adult phenotype
organ	flower whorl 2
anomaly	whorl 3

AP3::WUS, in intermediate AP3::WUS lines, whorls two and three were occupied by supernumerary stamens and carpelloid stamens. In strong AP3::WUS lines, organ number in whorls two and three was further increased, and all organs in these whorls consisted of carpelloid stamens Lohnman J.U. et al., 2001

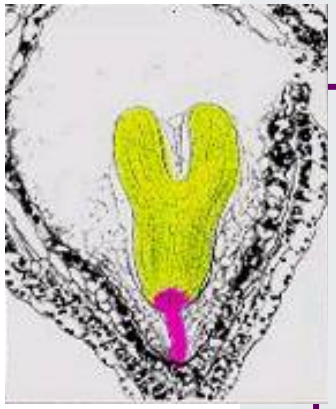

AGNSdb : Queries for Phenotype database - Microsoft Internet Explorer
Файл Правка Вид Избранное Сервис Справка
Назад Поиск Избранное Медиа
Адрес: http://emj-pc.ics.uci.edu/mgs/dbases/agns/phenotype_queries.html
Переход Ссылки
AGNSdb
QUERIES FOR PHENOTYPE DATABASE:
Mutations resulted in phenotypic abnormalities of the selected organs
Phenotypes of the mutants
OTHER VIEWS:
Queries for Expression database
Ontology navigation
Mutations resulted in phenotypic abnormalities in "primary SAM"
cuc2
Aida M. et al., 1997
cuc2 cuc1
Aida M. et al., 1997
cuc2 cuc1 stm-1
Aida M. et al., 1997
stm-1
Barton M.K. and Poethig R.S., 1993
Clark S.E. et al., 1996Byrne M.E. et al., 2002
stm-1 as-1 bp
Byrne M.E. et al., 2002
stm-1 clv3-2
Clark S.E. et al., 1996
stm-2
Endrizzi K. et al., 1996Clark S.E. et al., 1996
stm-2 clv3-2
Clark S.E. et al., 1996
stm-2 mgo1
Laufs P. et al., 1998, D
stm-2 wus-1
Endrizzi K. et al., 1996
stm-2 zll-3
Endrizzi K. et al., 1996
stm-5
Endrizzi K. et al., 1996
stm-5 zll-3
Endrizzi K. et al., 1996
© 2004 IC&G SB RAS, Laboratory of Theoretical Genetics
Готово
Интернет


AGNSdb : Queries for Phenotype database - Microsoft Internet Explorer
Файл Правка Вид Избранное Сервис Справка
Назад Поиск Избранное Медиа
Адрес: http://emj-pc.ics.uci.edu/mgs/dbases/agns/phenotype_queries.html
Переход Ссылки
AGNSdb
QUERIES FOR PHENOTYPE DATABASE:
Mutations resulted in phenotypic abnormalities of the selected organs
Phenotypes of the mutants
OTHER VIEWS:
Queries for Expression database
Ontology navigation
Phenotypes of the mutant "clv3-1"
clv3-1
Clark S.E. et al., 1995
Laufs P. et al., 1998, DLaufs P. et al., 1998, PC
dev_stage seedling
organ primary SAM
anomaly enlarged
clv1-3, both broader and taller than in the wild type in 5 days old plant. The SAM measures between 60 (m and 90 (m across the base, and between 25 (m and 30 (m in height, at the highest point the meristem has between 5 and 7 layers of avacuolated cells. Although the dome structure is maintained, SAM has more gently sloping sides, so they are bell shaped
Leyser H.M.O. and Furner I.J., 1992
clv1-4 stm-1, clv3-2 stm-1, were often larger than comparable wild-type SAM and often formed rosettes with more than 10 leaves as clv1-1 plants Clark S.E. et al., 1996
clv3-1, a large zone of slowly dividing cells in meristems of seedlings. This zone was not detectable in the wild type. These results suggest that the CZ is increased in size Laufs P. et al., 1998, PC
clv3-2 stm-1, SAM is absent in some seedlings and in other seedling a clear region of densely staining cells was observed in 8-day-old-seedlings with various states of cellular proliferation ranging from none detected, to considerable proliferation Clark S.E. et al., 1996
in clv3 the enlarged meristem is due to an increase in size of the CZ
Clark S.E. et al., 1995
Laufs P. et al., 1998, PC
fas1, fas2, broader and flatter than that in the wild type Kaya H. et al., 2001
© 2004 IC&G SB RAS, Laboratory of Theoretical Genetics
Интернет

AGNS: the vocabulary on organs, tissues, cells

Index	RO (Organ, tissue, cell)	Synonyms	Details	Complete definition in hierarchical clustering
2.1.1.2.	embryo, the apical domain, the inner cells	the upper inner cells; the upper central cells; the inner cells above O'line; the apical domain, subepidermal cells; the apical domain, subepidermal layer; the apical domain, hypodermal layer; the apical domain, subprotodermal laye; the apical layer of inner cells	Cells of the apical domain not included into protoderm. Characterized by periclinal (transverse) divisions from late globular to mid torpedo stage resulting in the late heart stage in a three-layered apical domain [Barton M.K. and Poethig R.S., 1993].	Embryo, the apical region, the apical domain, the inner cells

The hierarchical form of organ descriptions



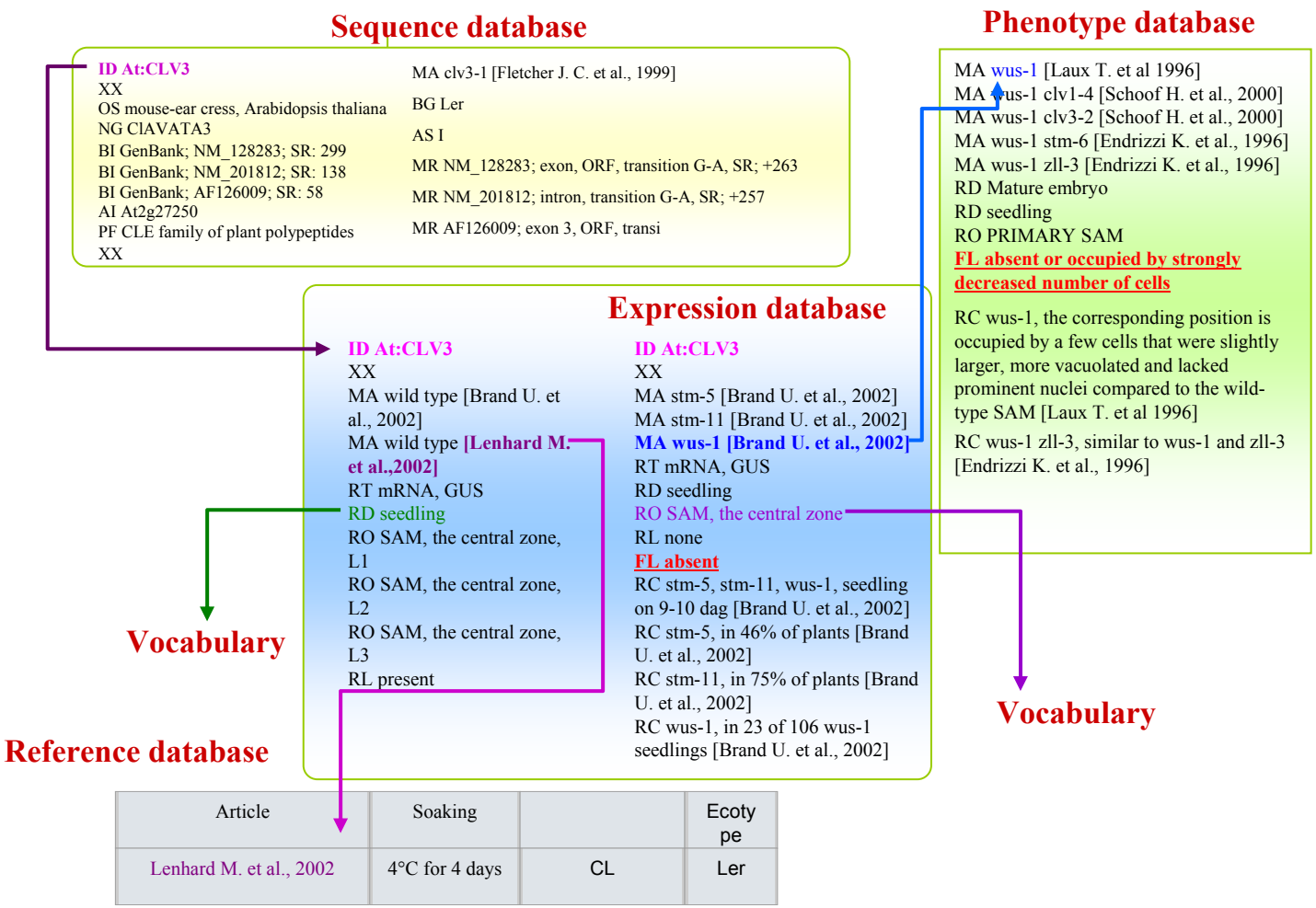
2.1 the apical region
- 2.1.1 Apical domain
  - 2.1.1.1 Presumptive SAM
    - 2.1.1.1.1 ECR
      - L1 2.1.1.1.1.1
      - L2 2.1.1.1.1.2
      - corpus 2.1.1.1.1.3
    - 2.1.1.1.2 EPR
      - L1 2.1.1.1.2.1
      - L2 2.1.1.1.2.2
      - corpus 2.1.1.1.2.3
  - 2.1.1.2 The inner cells
  - 2.1.1.3 SAM
    - 2.1.1.3.1 CZ
      - L1 2.1.1.3.1.1
      - L2 2.1.1.3.1.2
    - 2.1.1.3.2 PZ
      - L1 2.1.1.3.2.1
      - L2 2.1.1.3.2.2
- 2.1.2 Central domain O'line → … 2.1.2.1 · 2.1.2.3
- 2.1.3 Protoderm
- 2.1.4 Cotyledon primordia Cotyledon
- 2.1.5
  - 2.1.5.1 L1
  - 2.1.5.2 L2
  - 2.1.5.3 L3
- 2.1.6 → 2.1.6.1 · 2.1.6.3

2.2 the basal region
- 2.2.1 Hypophysis
- 2.2.2 suspensor

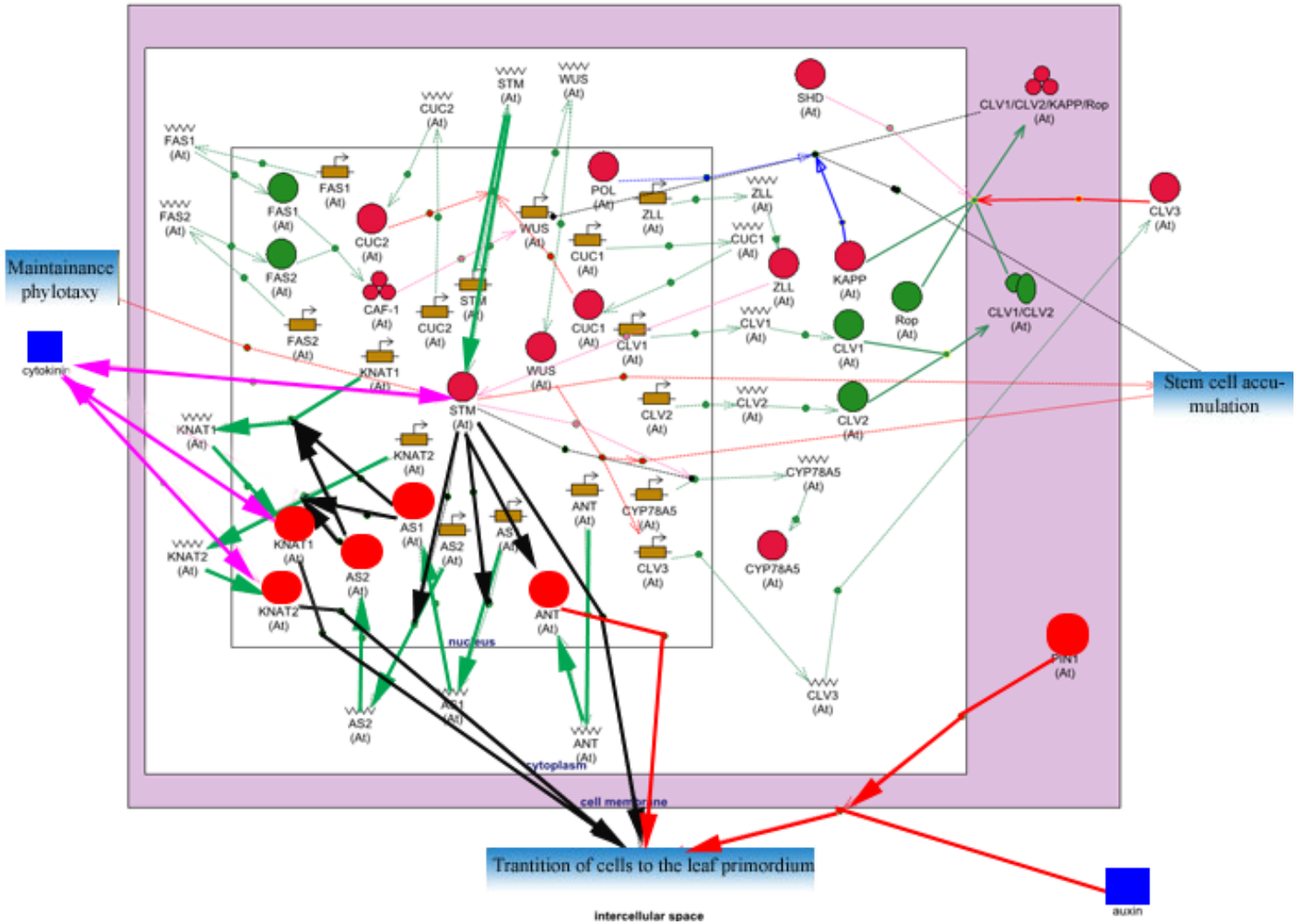
AGNS: the vocabulary on developmental stages

index	RD (Stage title)	Synonyms	Details	The stages of parts or other parts of the organ occurred at the same time
5.2.7.	FDS13	Anthesis, B3, ovule maturity stage, before fertilization	**Whorl 1**. The stage begins when the sepals open. **Whorl 2**. The petals can be seen between the sepals and continue to elongate rapidly. **Whorl 3.** Anthesis occurs. Stamen filaments extend faster. **Whorl 4.** The gynoecium possesses well-differentiated stigmatic papillae, style, and ovaries. The densely packed stigmatic papillae are 20 μm to 35 μm long. The stigma is already receptive at this stage. **Whorl 4, ovule.** Ovule maturity stage. The outer integument completely overgrows the inner integument, forming the micropylar opening. The nucellus degenerates, and the embryo sac is appressed to a newly differentiated cell layer of the inner integument, the endothelium. Within the embryo sac, megagametogenesis is completed with the formation of a seven-celled, eight-nucleate female gametophyte. No obvious morphological changes of gynoecium and unfertilized ovules occurred after stage 13. Duration of the stage is 6 hr. FDS13 occurs on the daa [Smyth D.R. et al., 1990], [Bowman J. L. et al., 1991; PC], [Muller A., 1961].	ODS 4-1 - ODS 4-IV EDS I- EDS III, SCDS 1

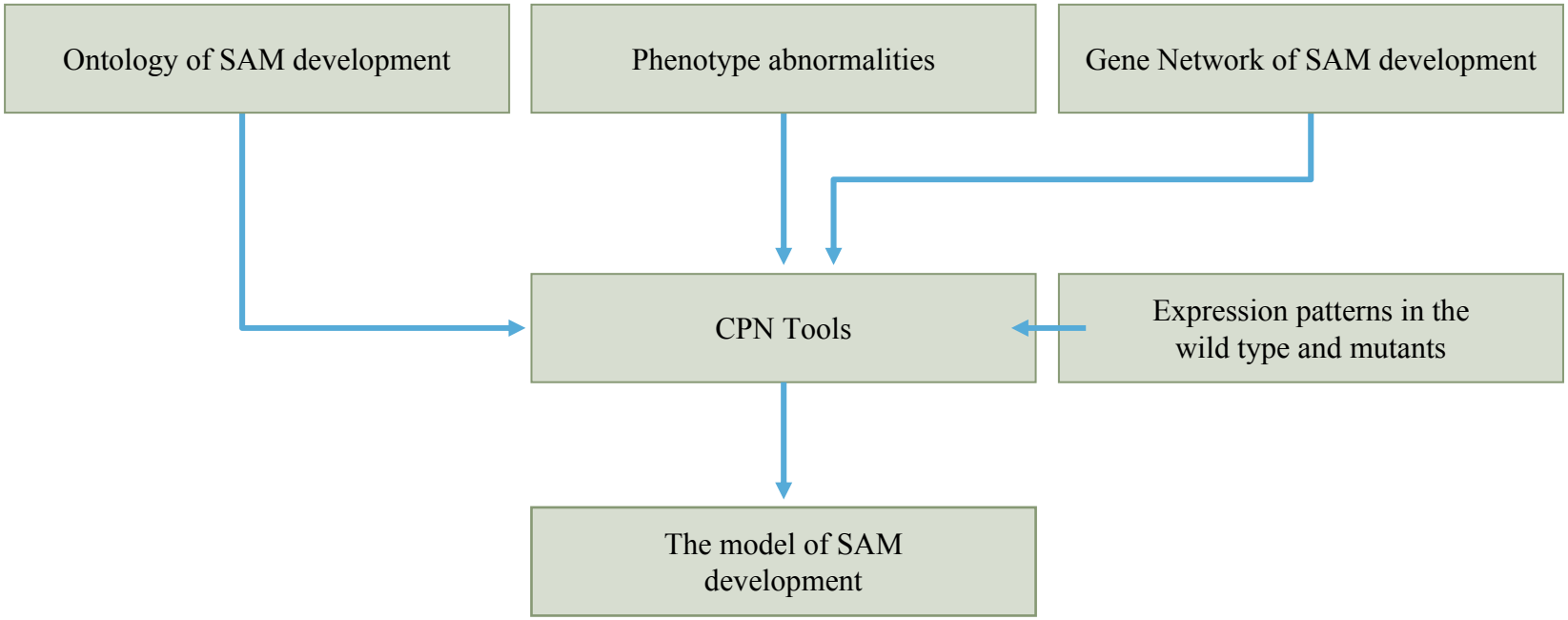
Correlation between the databases of AGNS



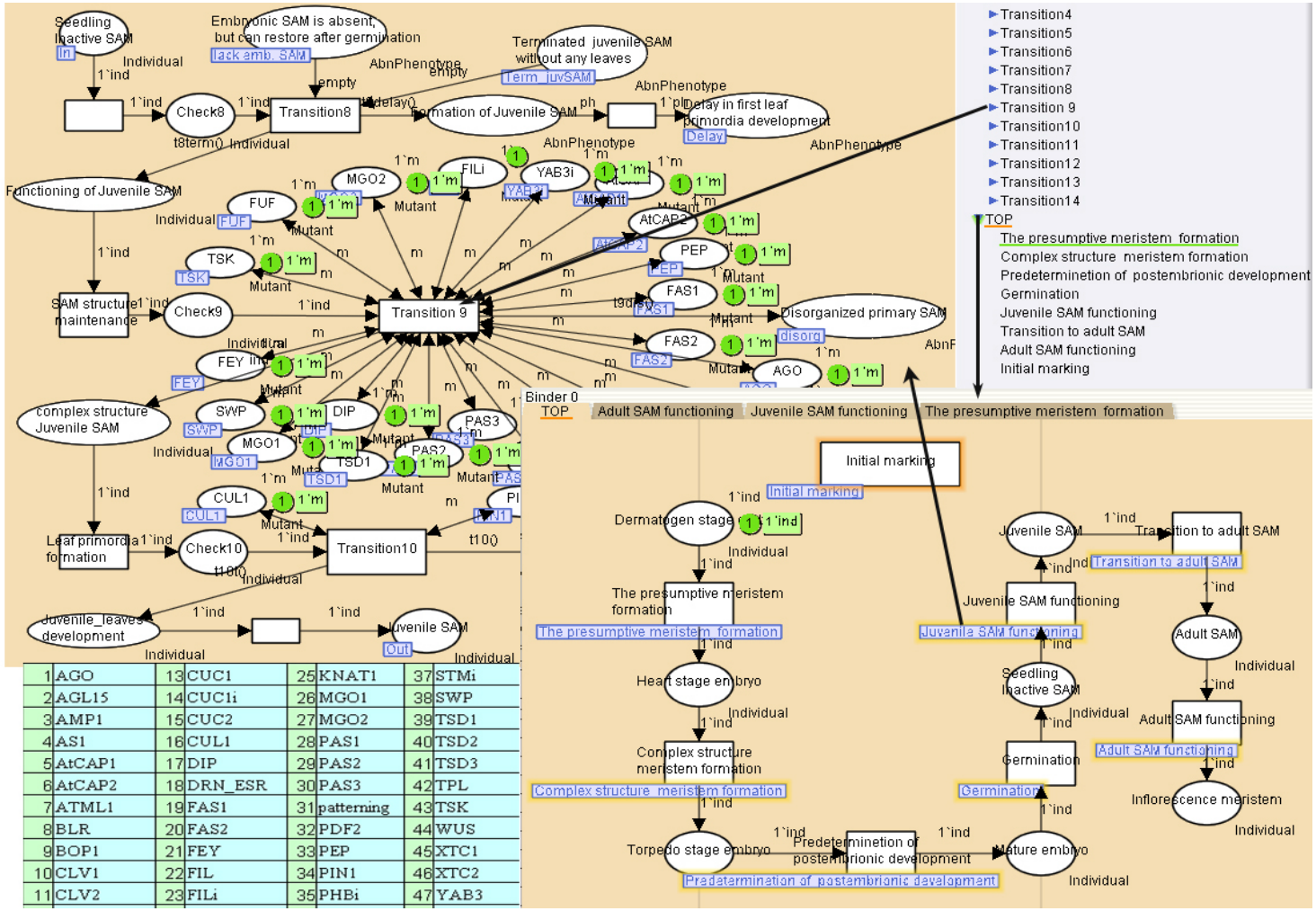
Article	Soaking		Ecotype
Lenhard M. et al., 2002	4°C for 4 days	CL	Ler


Maintainance phylotaxy
cytokinin
Stem cell accumulation
Trantition of cells to the leaf primordium
auxin
intercellular space
cytoplasm
cell membrane
nucleus
STM (At)
WUS (At)
CUC2 (At)
FAS1 (At)
FAS2 (At)
CAF-1 (At)
KNAT1 (At)
KNAT2 (At)
AS1 (At)
AS2 (At)
ANT (At)
POL (At)
ZLL (At)
CUC1 (At)
SHD (At)
KAPP (At)
Rop (At)
CLV1 (At)
CLV2 (At)
CLV3 (At)
CLV1/CLV2/KAPP/Rop (At)
CLV1/CLV2 (At)
CYP78A5 (At)
PIN1 (At)

AGNS data used for modeling



The model of Arabidopsis SAM development



1	AGO	13	CUC1	25	KNAT1	37	STMi
2	AGL15	14	CUC1i	26	MGO1	38	SWP
3	AMP1	15	CUC2	27	MGO2	39	TSD1
4	AS1	16	CUL1	28	PAS1	40	TSD2
5	AtCAP1	17	DIP	29	PAS2	41	TSD3
6	AtCAP2	18	DRN_ESR	30	PAS3	42	TPL
7	ATML1	19	FAS1	31	patterning	43	TSK
8	BLR	20	FAS2	32	PDF2	44	WUS
9	BOP1	21	FEY	33	PEP	45	XTC1
10	CLV1	22	FIL	34	PIN1	46	XTC2
11	CLV2	23	FILi	35	PHBi	47	YAB3

5.2. “Transgenesis”: informational resources to design experiments in the plant molecular biology & biotechnology fields

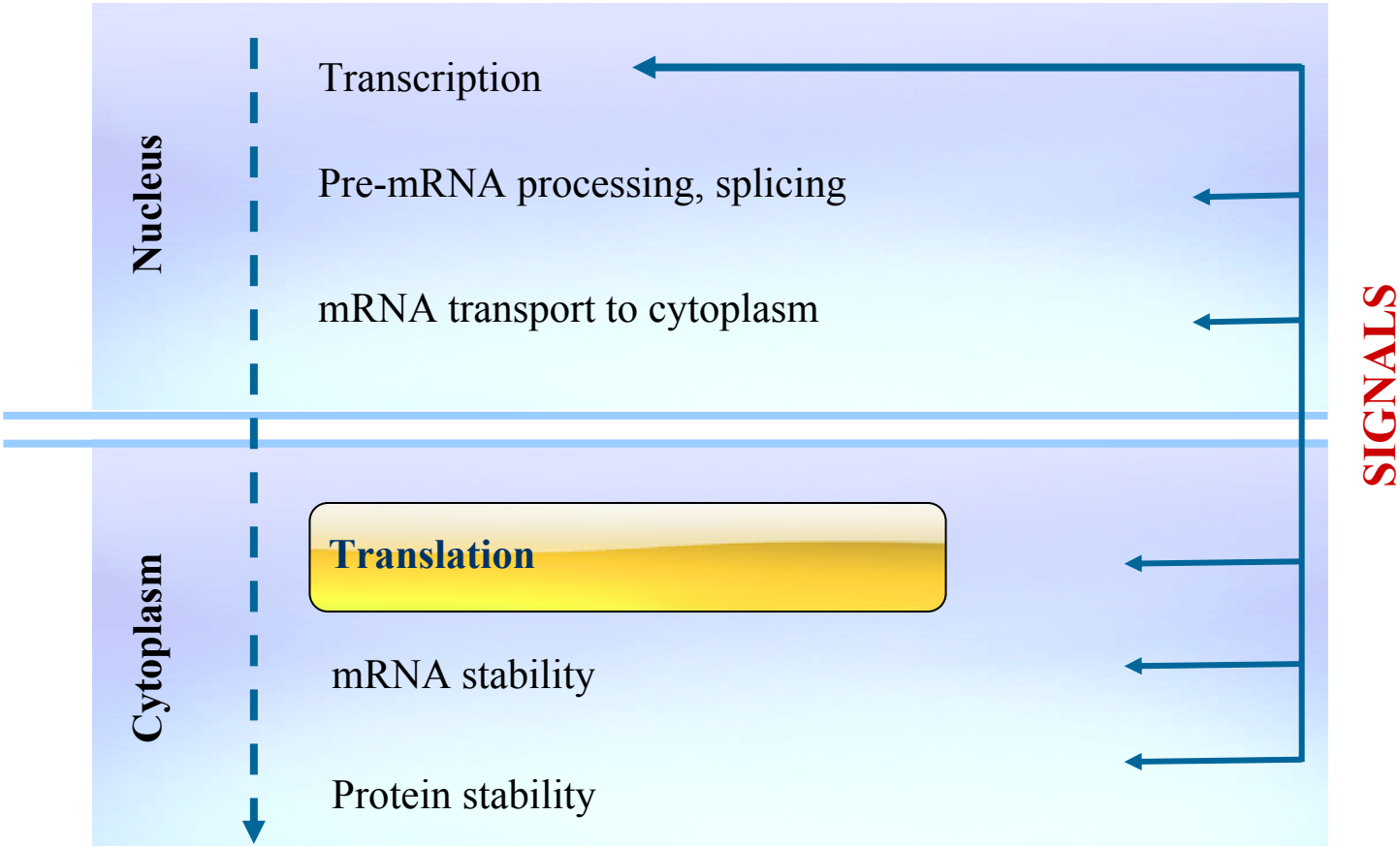
"Transgenesis"

•This multicomponent system is targeted to provide computational support for experiments in plant molecular biology, biotechnology, and molecular genetics. We plan to include modules providing user with numerous opportunities including:

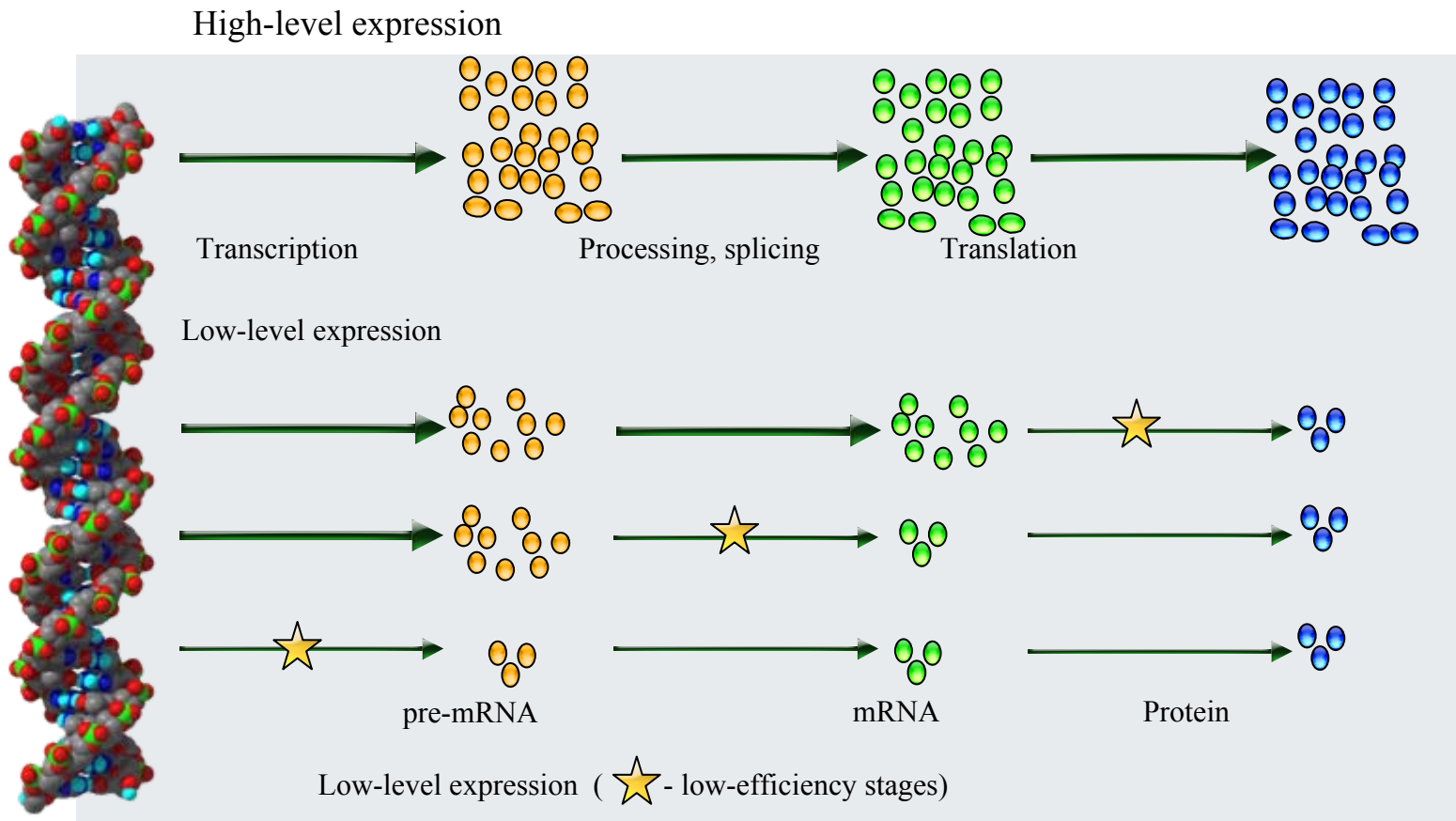
- selection of potential gene targets to make the desirable effect on plant morphology, physiology, biochemistry, etc.
- selection of an appropriate regulatory signals (promoter, translational enhancers, etc.).
- design of a transgene (modification of codon content, translation initiation and termination signals, elimination of potential misfunctional splicing and poly(A)-signals)

•Current version of "Transgenesis" includes some databases (TRRD, TRSIG, AGNS) and pilot versions of the prediction programs (Transgene, Leader_RNA)

Transgenesis: where *gene expression intensity* is controlled in *plants cells*



Transgenesis: the “limiting link” concept applied to gene expression processes in plants

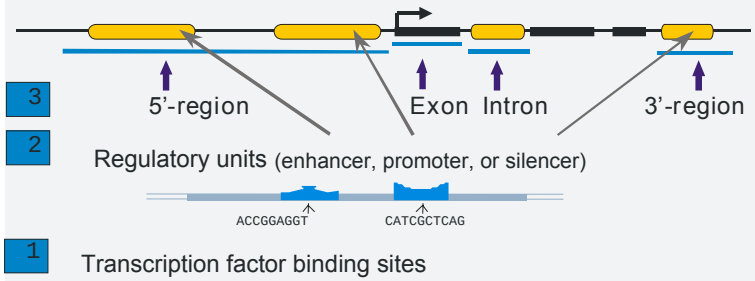


«Transgenesis» modules

Informational resources on expression signals & regulatory regions	TRRD (Transcription Regulatory Regions Database) TRSIG (TRanslation SIGnal)
Informational resources on regulatory networks	TRSIG (TRanslation SIGnal) AGNS (Arabidopsis GeneNet Supplementary DataBase) GeneNet (Gene Networks)
Prediction programs	Transgene (codon content): modeling and analysis Leader_RNA (translation initiation rate)

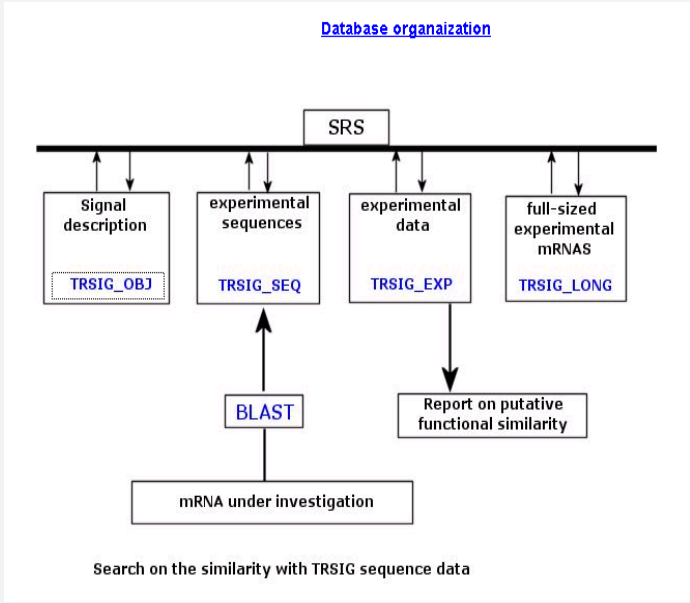
Plant Transgenesis optimization: Selection of appropriate promoters and post-transcriptional signals

TRRD: promoter database



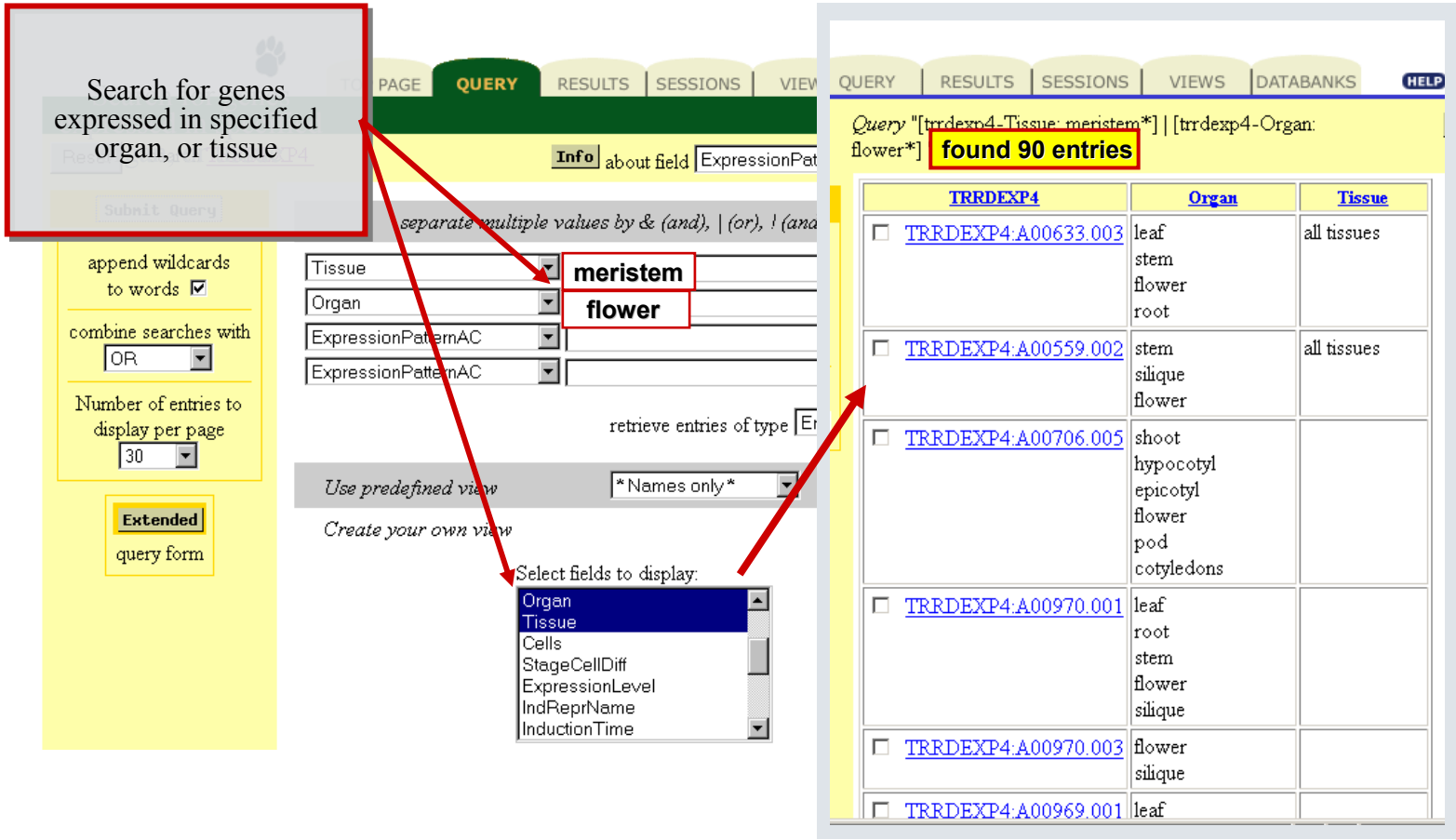
TRSIG: translational enhancers and mRNA stability signals

Database on Translation Signals (TRSIG)

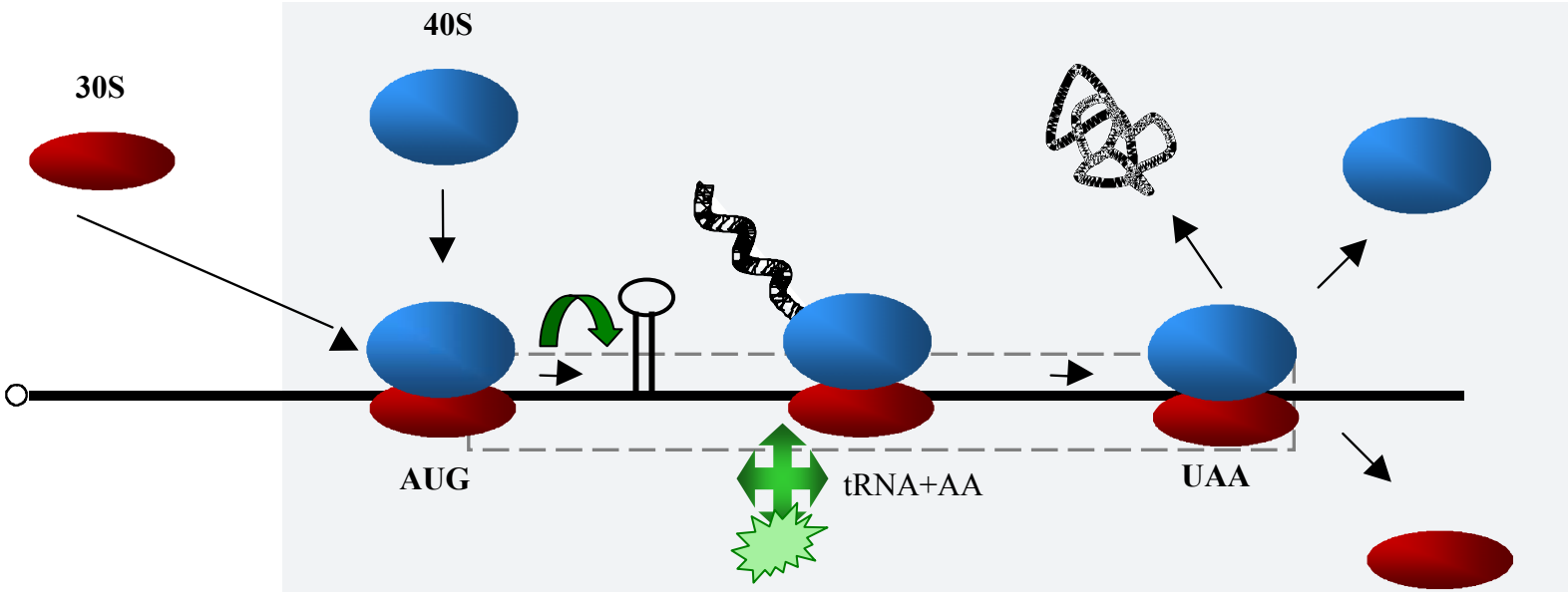


http://wwwmgs2.bionet.nsc.ru/mgs/dbases/trsig/

Plant Transgenesis optimization: SRS access to TRRD to find the promoters for transgenesis

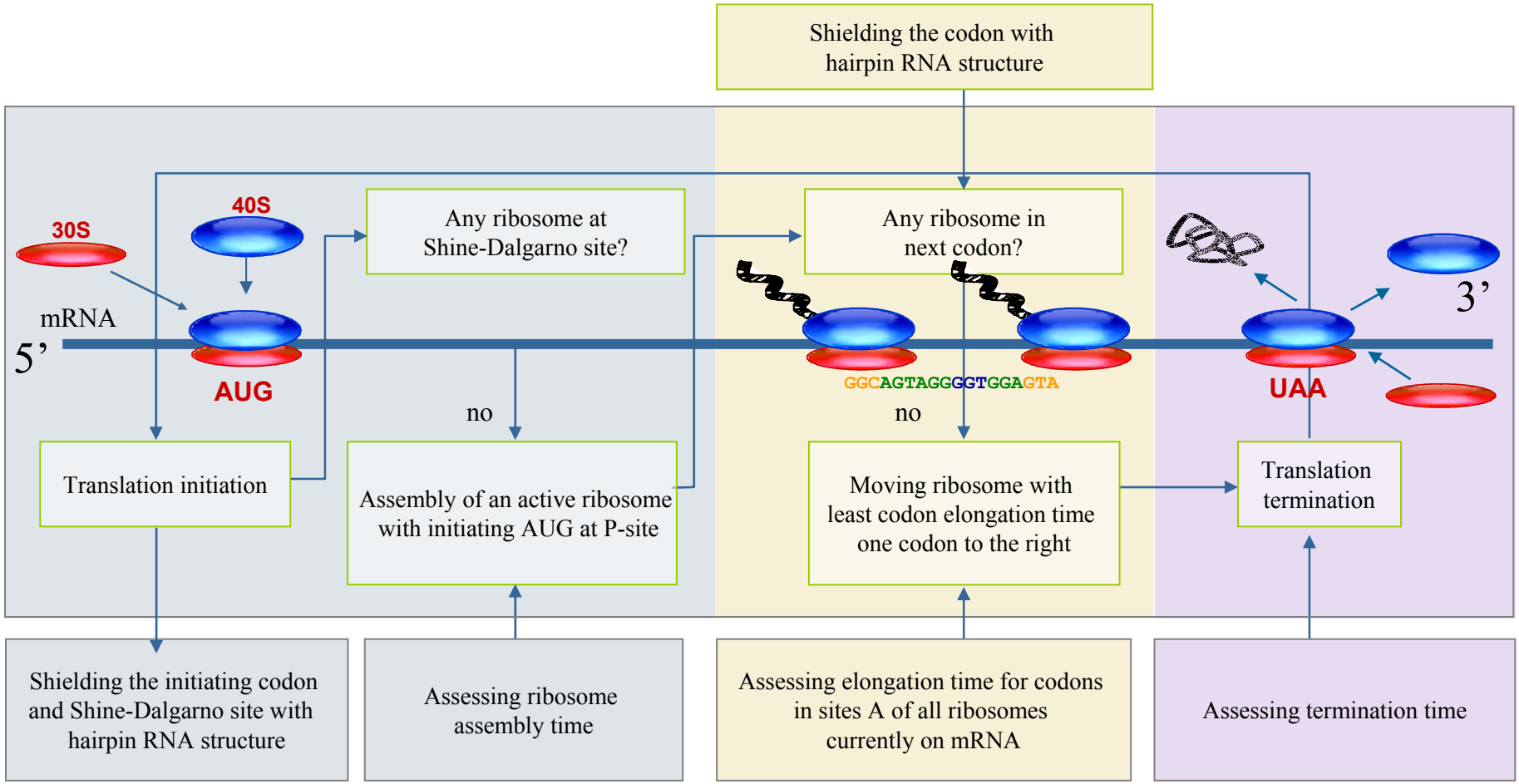


Schematic model of translation of prokaryotic mRNA

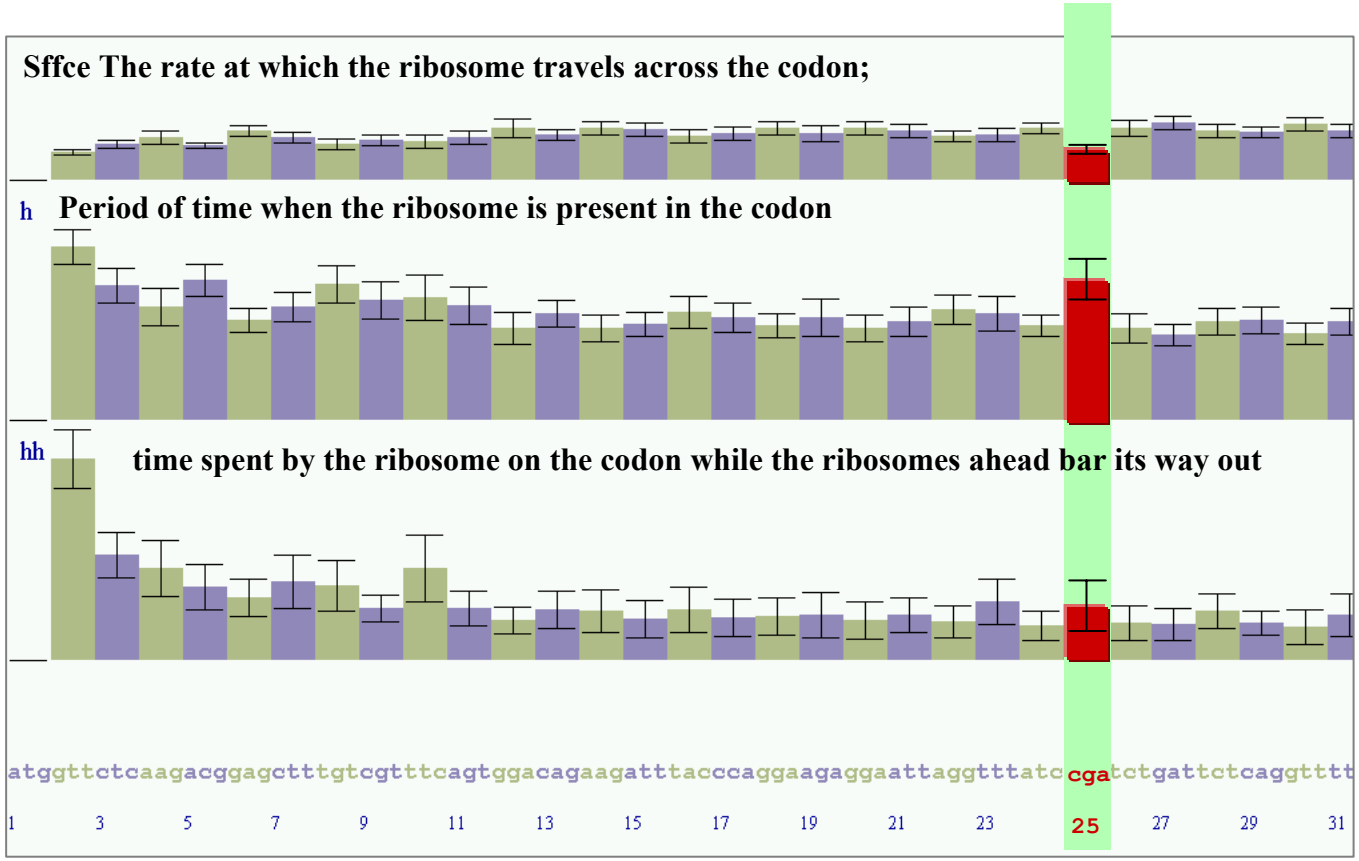


mRNA, start of translation, hairpins, stop-codon

Plant Transgene optimization: stochastic modeling for translation

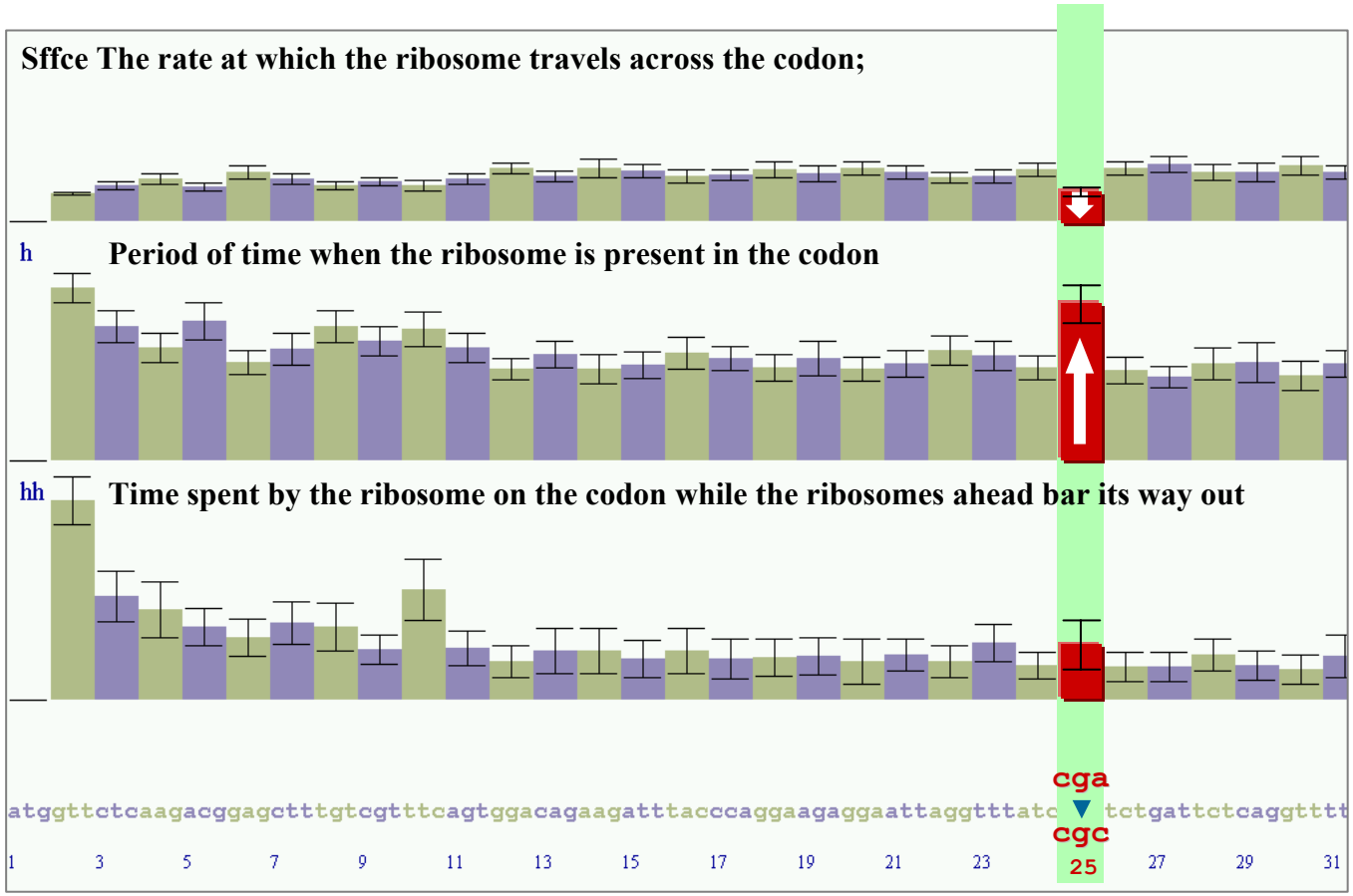


Plant Transgenesis optimization: stochastic modeling of the Arabidopsis thaliana At3g53020 mRNA translation process in a bacterial system.



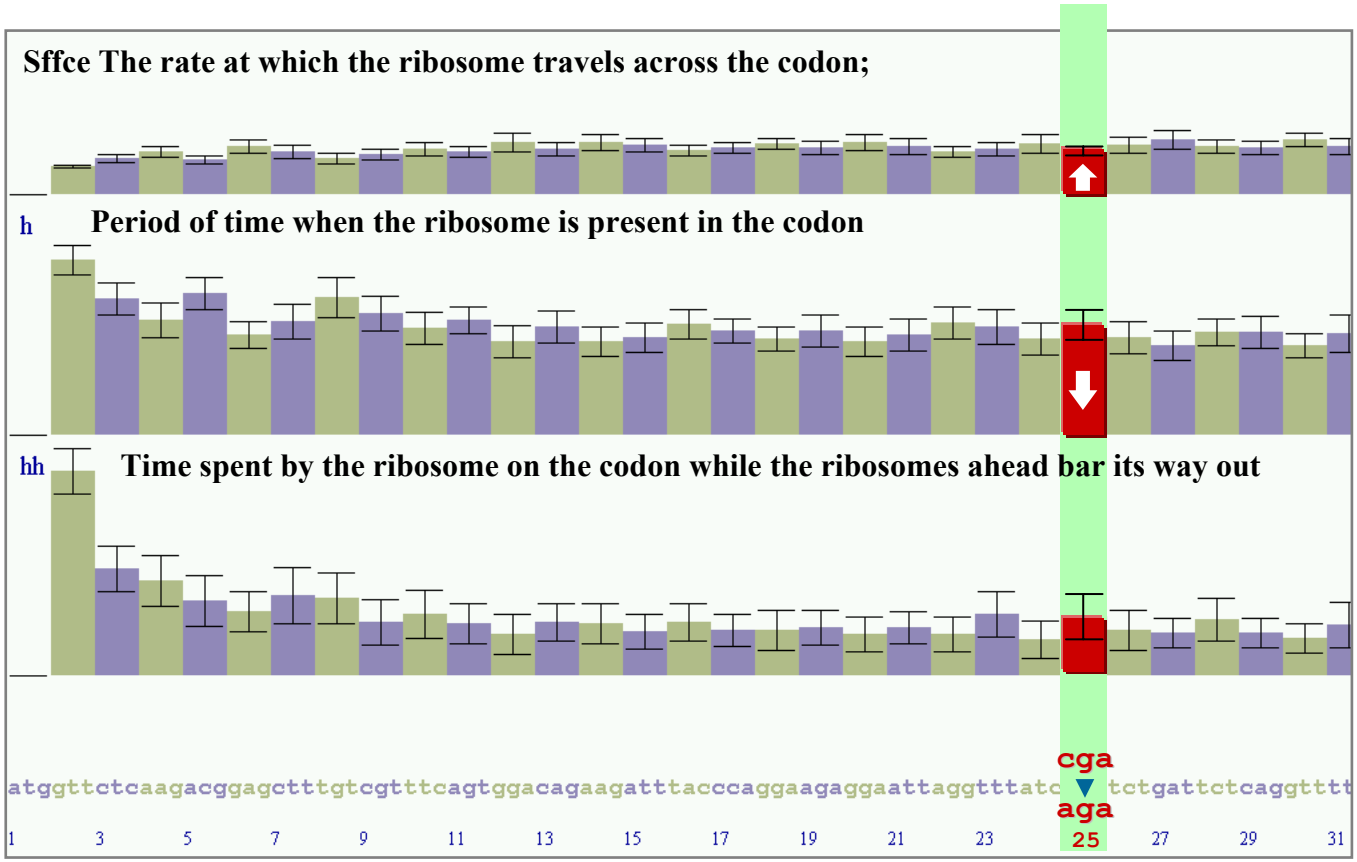
Ribosomes travel slowly across codon 25 CGA (coding for arginine).

Plant Transgenesis optimization: Stochastic modeling of the Arabidopsis thaliana At3g53020 mRNA translation process in a bacterial system



Codon 25 CGA (coding for arginine) is replaced by the "slowest" synonymous codon CGC. Ribosome travel rate on the codon reduced dramatically.

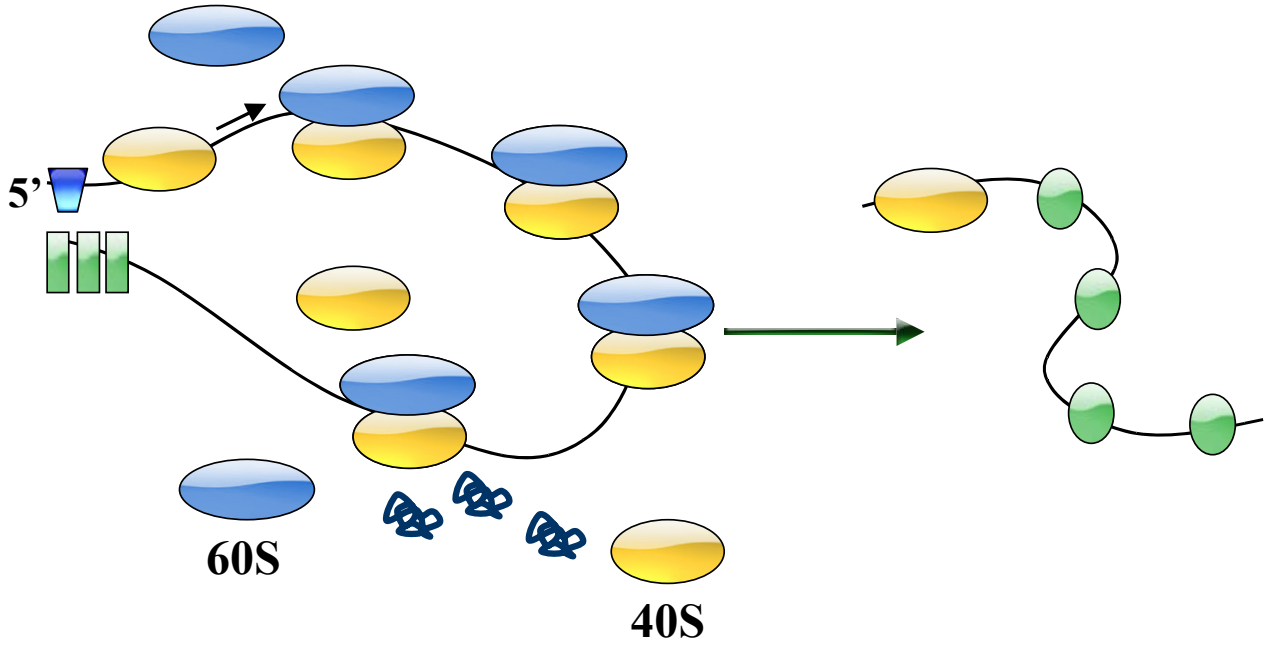
Plant Transgenesis optimization: Stochastic modeling of the Arabidopsis thaliana At3g53020 mRNA translation process in a bacterial system



Codon 25 CGA (coding for arginine) is replaced by the "fastest" synonymous codon AGA Ribosome travel rate on the codon increased.

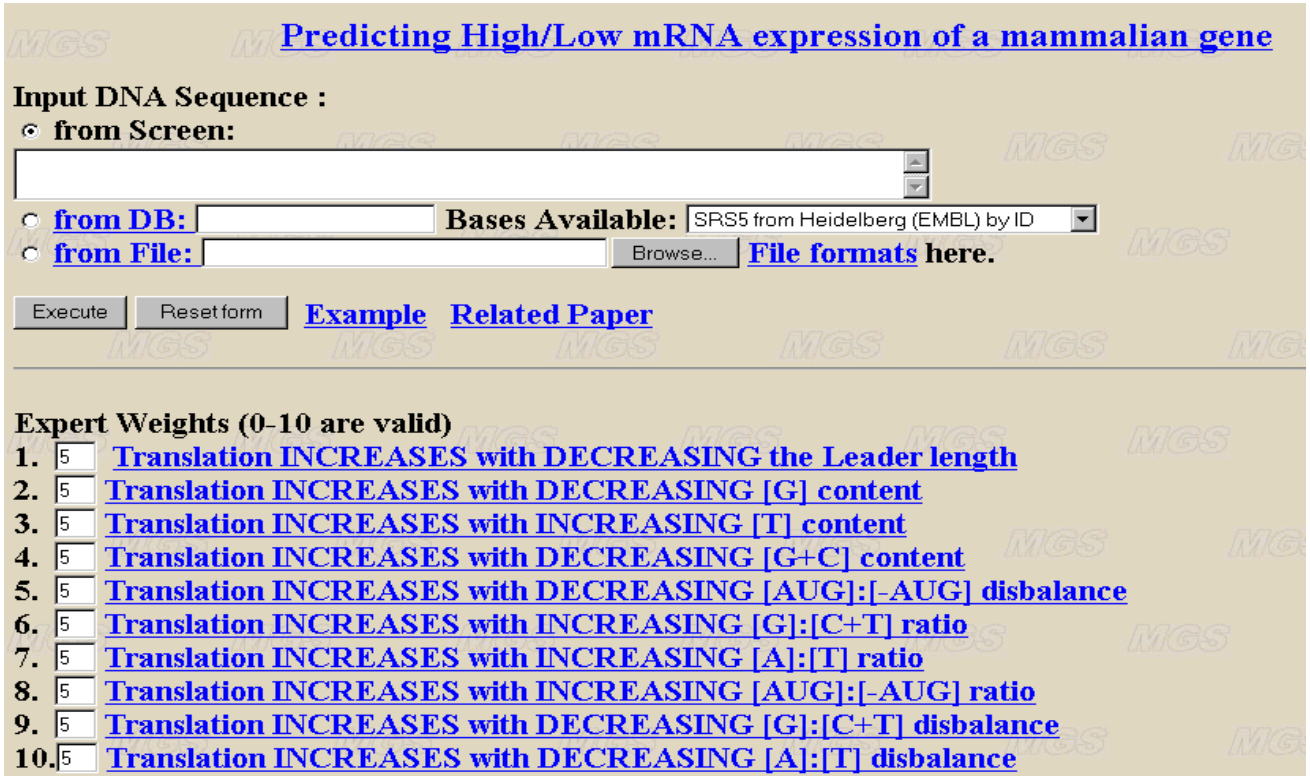
Plant Transgenesis optimization: There are stress-, stage- and tissue-specific regulators of translation and cytoplasmic stability of mRNA

•Translation of most mRNA stops when heat-shock, hypoxia or tissue damage take place.



•Intensive mRNA translation of certain genes is still under way, stress notwithstanding

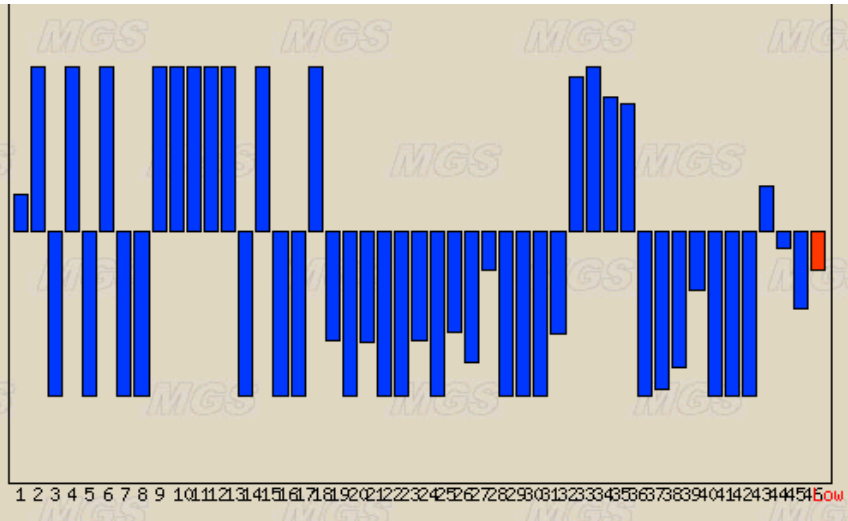
Plant Transgenesis optimization: Prediction of mRNA



translation initiation rate on the basis of 5'-UTR features: computer system Leader_RNA (input page)

http://wwwmgs.bionet.nsc.ru/mgs/programs/leadermrna/ma_mrna.html

Plant Transgenesis optimization: : Prediction of mRNA translation initiation rate on the basis of 5'-UTR features: computer system Leader_RNA (output page)

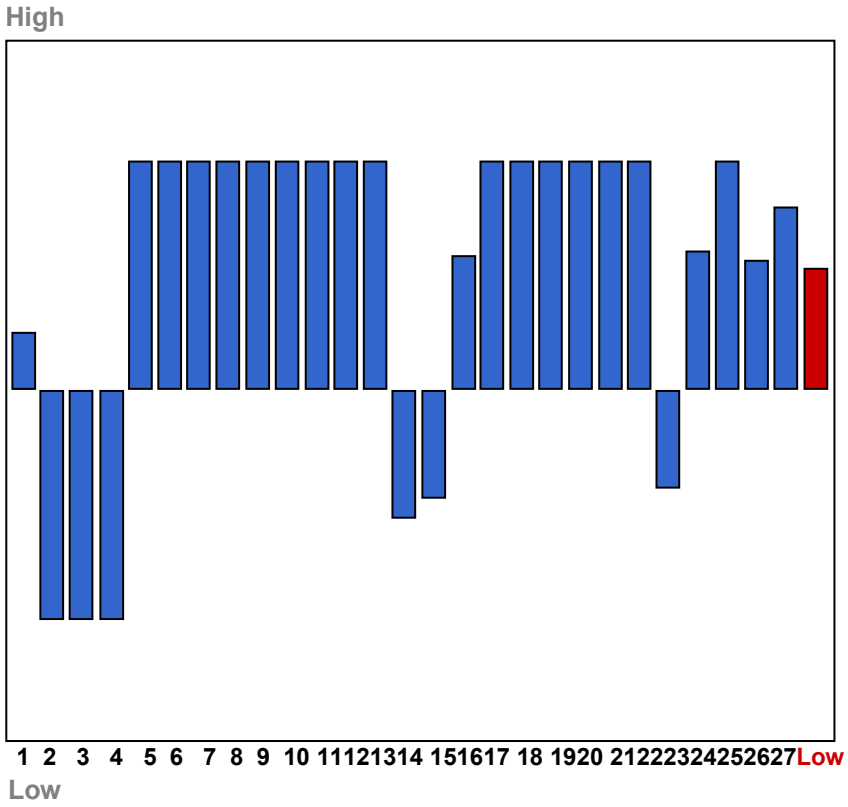


Data is:
CAACACAGAC ACACACTCTA CTACACGCAT CAGCAGCAAA CCCCCCCCCC
CCACACACTA CCAGCAGCAC CCCAACCGGT AAAAAAAAAC CCCCTATATA
CACACTACGA GGGGCCAAAA CCCCAAATTT TAAGCAA

Prediction is mean of the Expert decisions = *-0.23761* **that means** *Low.*

http://wwwmgs.bionet.nsc.ru/mgs/programs/leadermrna/ma_mrna.html

Control prediction of translation activity of high-level expression mRNA
(*embl ac at30srs13*)



Data is:
CGTTTGCCTT ATCCGTTCAG CTCATCTTCT TCTTCTTCTT CGTCACCTCT
GAATTAGTTT CCCAGAATCC GAAATTCCTA GGAAGAGAAC ATAACA

Prediction is mean of the Expert decisions = 0.532137 that means High.

*Expert [Estimate*Weight=Decision]'s are following:*

1. *Translation INCREASES with DECREASING the Leader length 0.253257 * 5 = 1.26629*
2. *Translation INCREASES with DECREASING [T] content -1 * 5 = -5*
3. *Translation INCREASES with DECREASING [AUG]:[-AUG] disbalance -1 * 5 = -5*
4. *Translation INCREASES with INCREASING [A]:[T] ratio -1 * 5 = -5*
5. *Translation INCREASES with INCREASING [AUG]:[-AUG] ratio 1 * 5 = 5*
6. *Translation INCREASES with DECREASING [A]:[T] disbalace 1 * 5 = 5*
7. *Translation INCREASES with DECREASING [AUG] content 1 * 5 = 5*
8. *Translation INCREASES with DECREASING [AUG] framed 1 * 5 = 5*
9. *Translation INCREASES with DECREASING [AUG] optimized 1 * 5 = 5*
10. *Translation INCREASES depends on the "-3 position" rule 1 * 5 = 5*
11. *Translation INCREASES with DECREASING [AUG] "-3"-ruled 1 * 5 = 5*
12. *Translation INCREASES with DECREASING [K] content of [-17;-1] 1 * 5 = 5*
13. *Translation INCREASES with DECREASING [KB] content of [-17;-1] 1 * 5 = 5*
14. *Translation INCREASES with INCREASING High-consensus matches -0.54879 * 5 = -2.74395*
15. *Translation INCREASES with DECREASING Low-consensus matches -0.455469 * 5 = -2.27735*
16. *Translation INCREASES with INCREASING High-ShortFreqMatr 0.583073 * 5 = 2.91537*
17. *Translation INCREASES with INCREASING High/Low 1bp-FreqRatio 1 * 5 = 5*
18. *Translation INCREASES with INCREASING High/Low 1bp(KM)-FreqRatio 1 * 5 = 5*
19. *Translation INCREASES with INCREASING High/Low 2bp(KM)-FreqRatio 1 * 5 = 5*
20. *Translation INCREASES with INCREASING High/Low 3bp(KxM)-FreqRatio 1 * 5 = 5*
21. *Translation INCREASES with INCREASING High/Low 5bp(KM)-FreqRatio 1 * 5 = 5*
22. *Translation INCREASES with INCREASING High/Low 6bp(KM)-FreqRatio 1 * 5 = 5*
23. *Translation INCREASES with INCREASING High/Low 3bp(ATGCx)FreqRatio -0.421578 * 5 = -2.10789*
24. *Translation INCREASES with INCREASING High/Low 5bp(ATGCx)FreqRatio 0.608727 * 5 = 3.04363*
25. *Translation INCREASES with INCREASING High/Low 3bp(KxM)-FreqRatio 1 * 5 = 5*
26. *Translation INCREASES with INCREASING High/Low 5bp(KxM)-FreqRatio 0.557566 * 5 = 2.78783*
27. *Translation INCREASES with INCREASING High/Low 7bp(KxM)-FreqRatio 0.79092 * 5 = 3.9546*

Plant Transgene optimization: selection of appropriate synonymous codon content

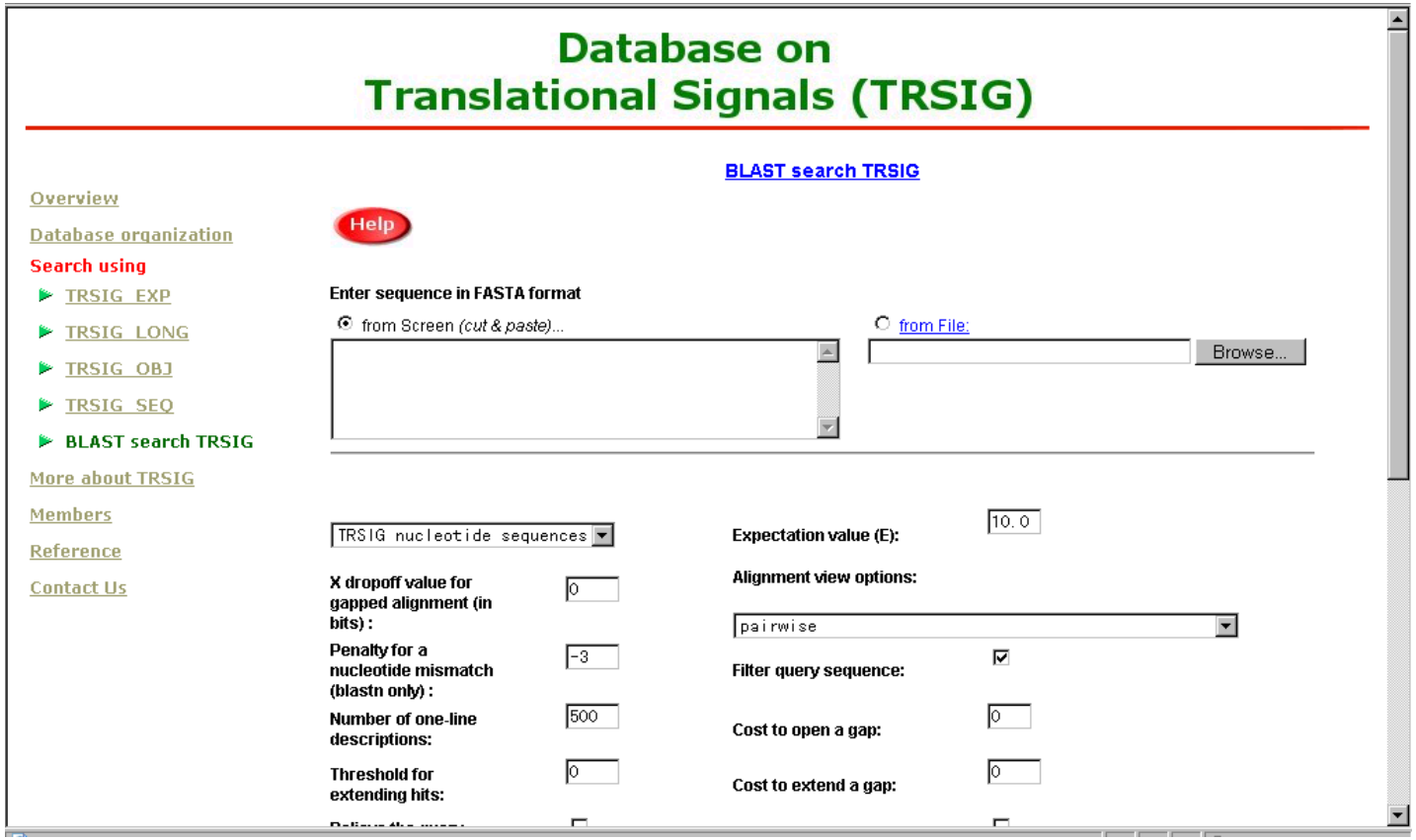
Synonymous codons	Average frequencies	Select variant for replacement
Ala		all best
GCA	21.9	GCA
GCC	11.4	GCA / GCC / GCG / GCT
GCG	4.6	
GCT	30.3	
Arg		all best
AGA	15.1	AGA
AGG	11.0	AGG
CGA	5.5	CGA
CGC	3.5	CGC
CGG	**2.9**	CGG
CGT	8.3	CGT
Asn		all best
AAC	16.7	AAC
AAT	29.4	AAT
Asp		all best
GAC	15.5	GAC
GAT	39.5	GAT
Cys		all best
TGC	6.8	TGC
TGT	10.6	TGT
Gln		all best
CAA	21.1	CAA

Organism: Solanum tuberosum
Length of CDS: 1377
Changed synonymous codons are marked with red

```
Met Gly Ser Pro Ala Lys Tyr Leu Ser Val His Glu Thr Gln
ATG GGG TCT CCA GCA AAA TAC TTG TCT GTG CAC GAA ACT CAG
ATG GGT TCT CCA GCT AAA TAC TTG TCT GTG CAC GAA ACT CAA

Gly Glu Ile Met Trp Asn Thr Ser Glu Ser Ala Glu Lys Thr
GGA GAG ATA ATG TGG AAC ACG TCC GAA TCG GCG GAA AAA ACA
GGA GAG ATT ATG TGG AAC ACG TCC GAA TCT GCG GAA AAA ACA

Ile Val Arg Thr Val Thr Gly Cys Leu Leu Ser Leu Leu Ile
ATA GTT CGG ACG GTG ACT GGT TGC TTG CTT TCT CTT CTC ATC
ATT GTT AGA ACG GTG ACT GGT TGT TTG CTT TCT CTT CTT ATC

Asn Ile Leu Val Cys Ser Ala Val Leu Lys Phe Arg His Leu
AAC ATT CTG GTA TGC TCT GCG GTC CTC AAG TTC AGG CAC TTG
AAC ATT CTT GTA TGT TCT GCG GTC CTT AAG TTC AGG CAC TTG

Ile Phe Ile Val Ser Leu Ala Val Ser Asp Leu Phe Val Ala
ATT TTC ATT GTG TCT CTG GCT GTT TCT GAC CTG TTT GTT GCT
ATT TTC ATT GTG TCT CTT GCT GTT TCT GAT CTT TTT GTT GCT

Lys Ala Val Ala Glu Val Ala Gly Tyr Trp Pro Phe Gly Pro
AAA GCA GTG GCC GAG GTT GCG GGA TAT TGG CCA TTC GGA CCC
AAA GCT GTG GCC GAG GTT GCG GGA TAT TGG CCA TTC GGA CCA

Ala Phe Asp Ile Met Cys Ser Thr Ala Ser Ile Leu Asn Leu
GCT TTC GAC ATT ATG TGC TCC ACG GCG TCC ATC CTC AAT CTC
GCT TTC GAT ATT ATG TGT TCC ACG GCG TCC ATC CTT AAT CTT
```

http://wwwmgs2.bionet.nsc.ru/mgs/systems/transgene/

Plant Transgenesis optimization: Use BLAST search to find homology between mRNA of interest and TRSIG stored translational signals

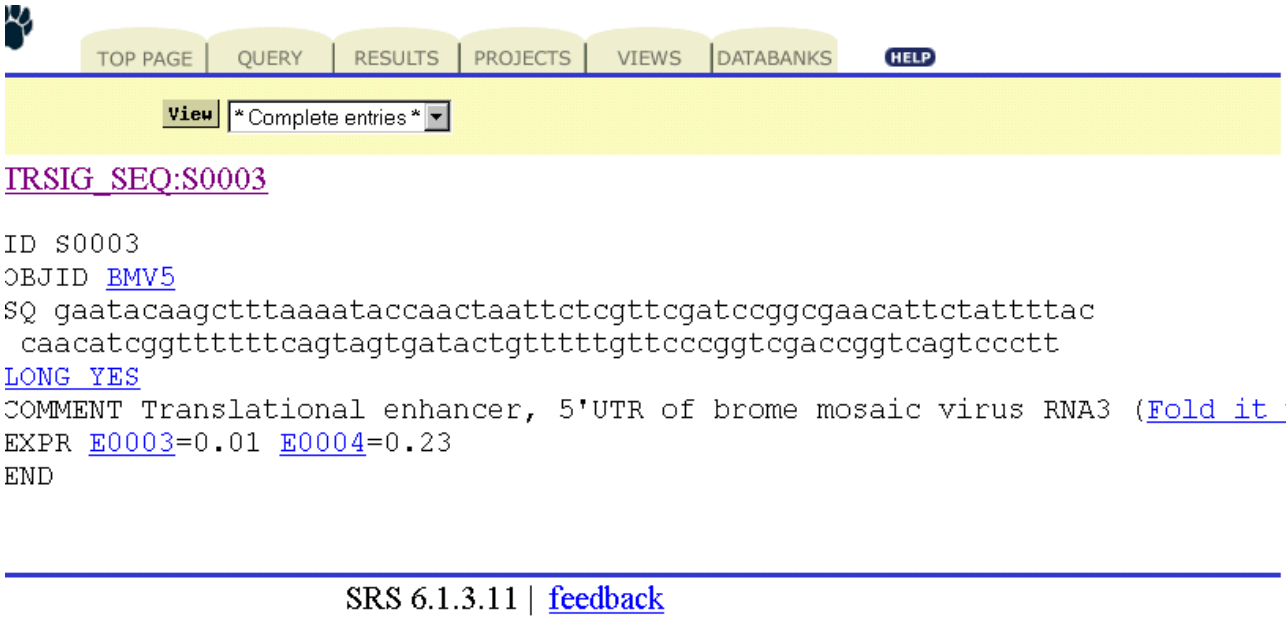


http://wwwmgs2.bionet.nsc.ru/mgs/dbases/trsig/

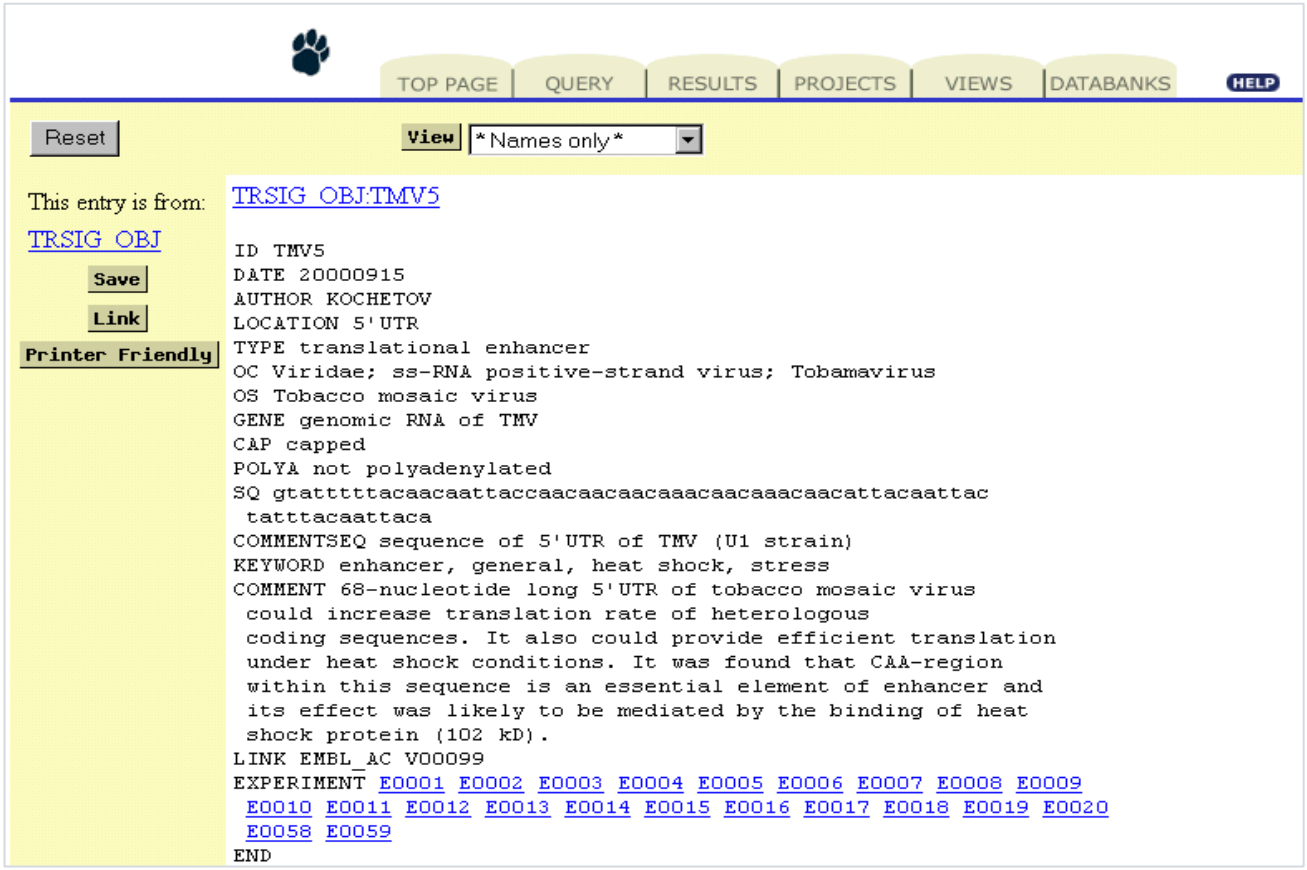
Plant Transgenesis optimization: TRSIG BLAST search output

```
Sequences producing significant alignments:                    (bits) Value

gnl|TRSIG_SEQ|S0010 5'UTR of Rouse Sarcoma Virus                 26   5e-04
gnl|TRSIG_SEQ|S0001 5'UTR of tobacco mosaic virus (omega)        26   5e-04

>gnl|TRSIG_SEQ|S0010 5'UTR of Rouse Sarcoma Virus
          Length = 101

 Score = 26.3 bits (13), Expect = 5e-04
 Identities = 13/13 (100%)
 Strand = Plus / Plus


Query: 1  caacaacaaacaa 13
          |||||||||||||
Sbjct: 37 caacaacaaacaa 49


>gnl|TRSIG_SEQ|S0001 5'UTR of tobacco mosaic virus (omega)
          Length = 98

 Score = 26.3 bits (13), Expect = 5e-04
 Identities = 13/13 (100%)
 Strand = Plus / Plus


Query: 1  caacaacaaacaa 13
          |||||||||||||
Sbjct: 34 caacaacaaacaa 46
```

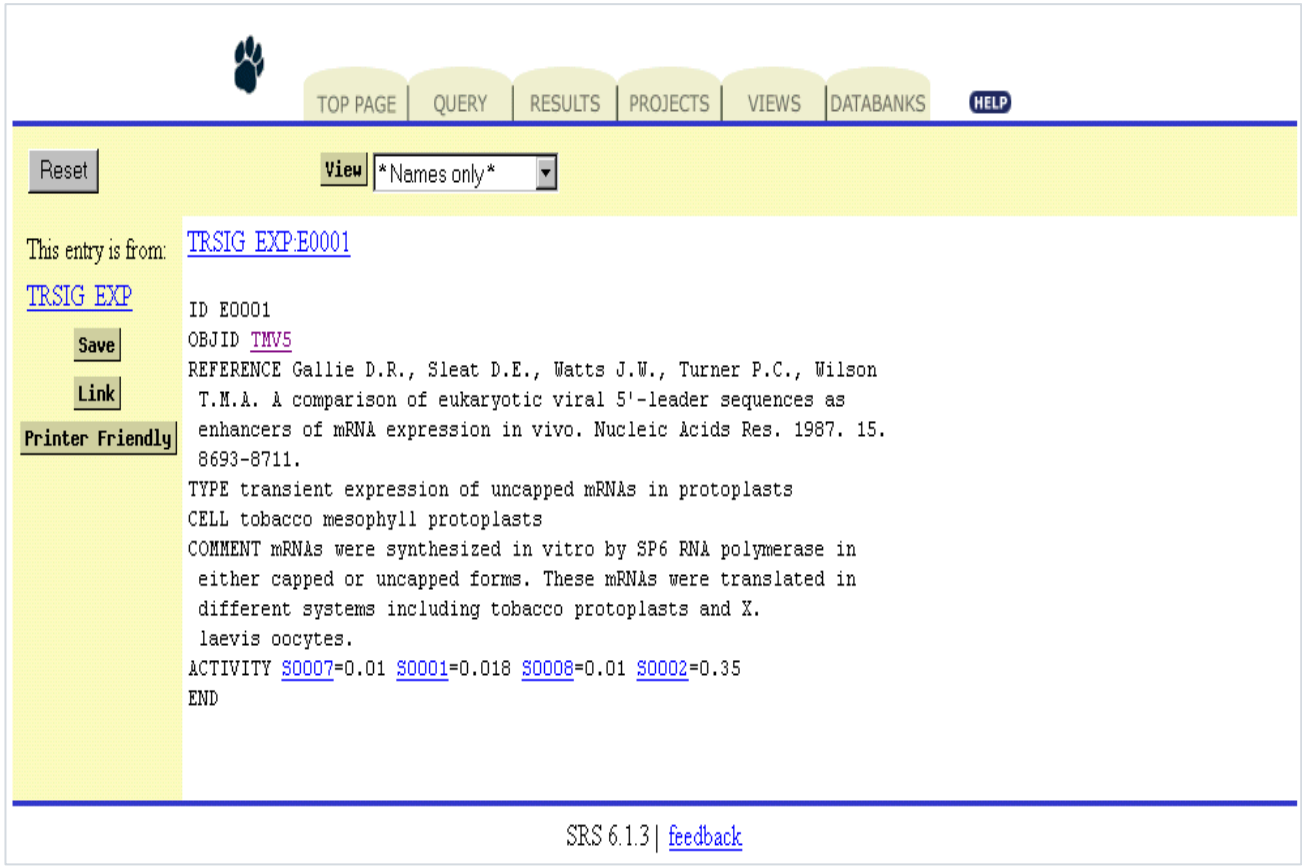
Plant Transgenesis optimization: Get sequences of homologous signals and use cross-references for additional information



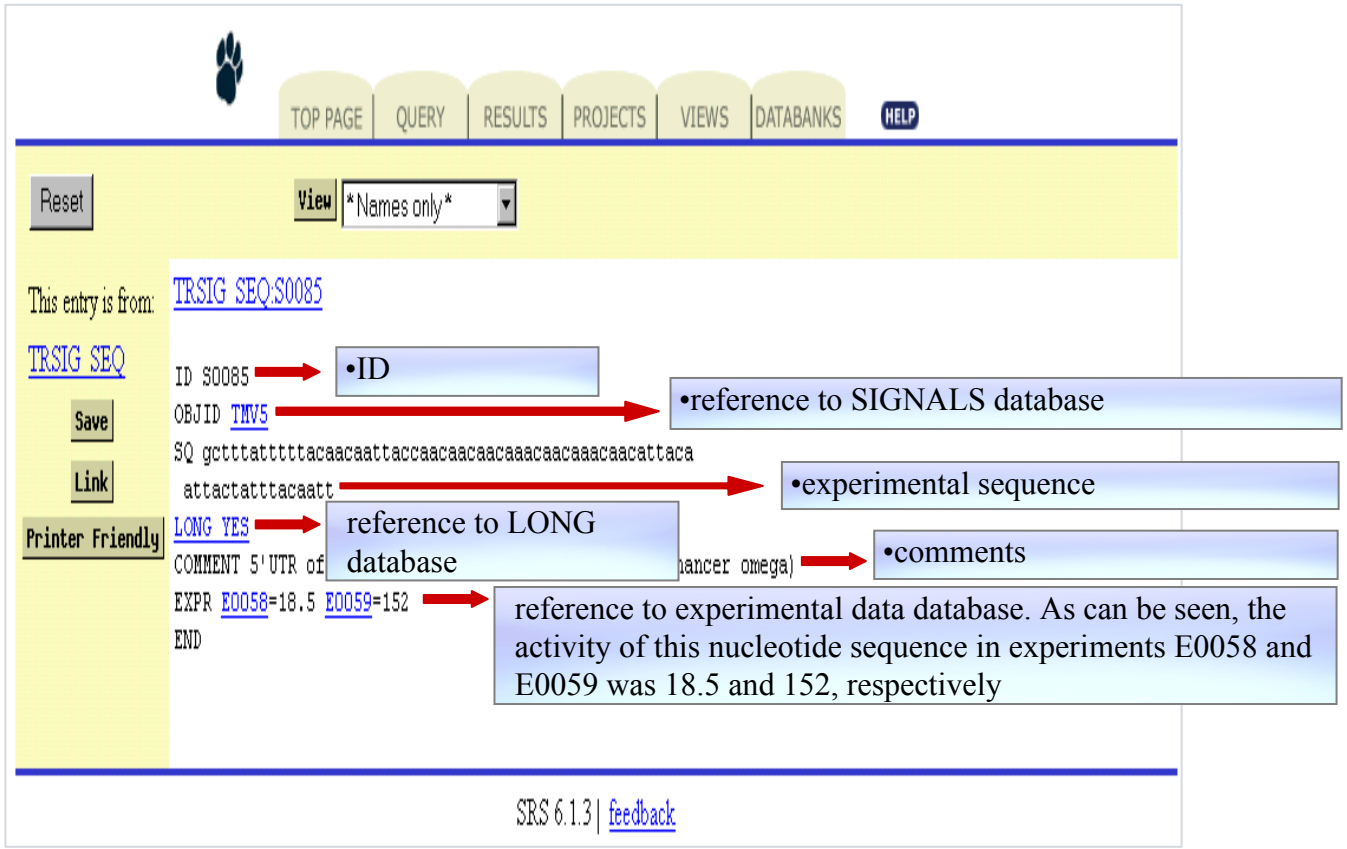
Plant Transgenesis optimization: Examine TRSIG OBJECT database entry by clicking on objid field



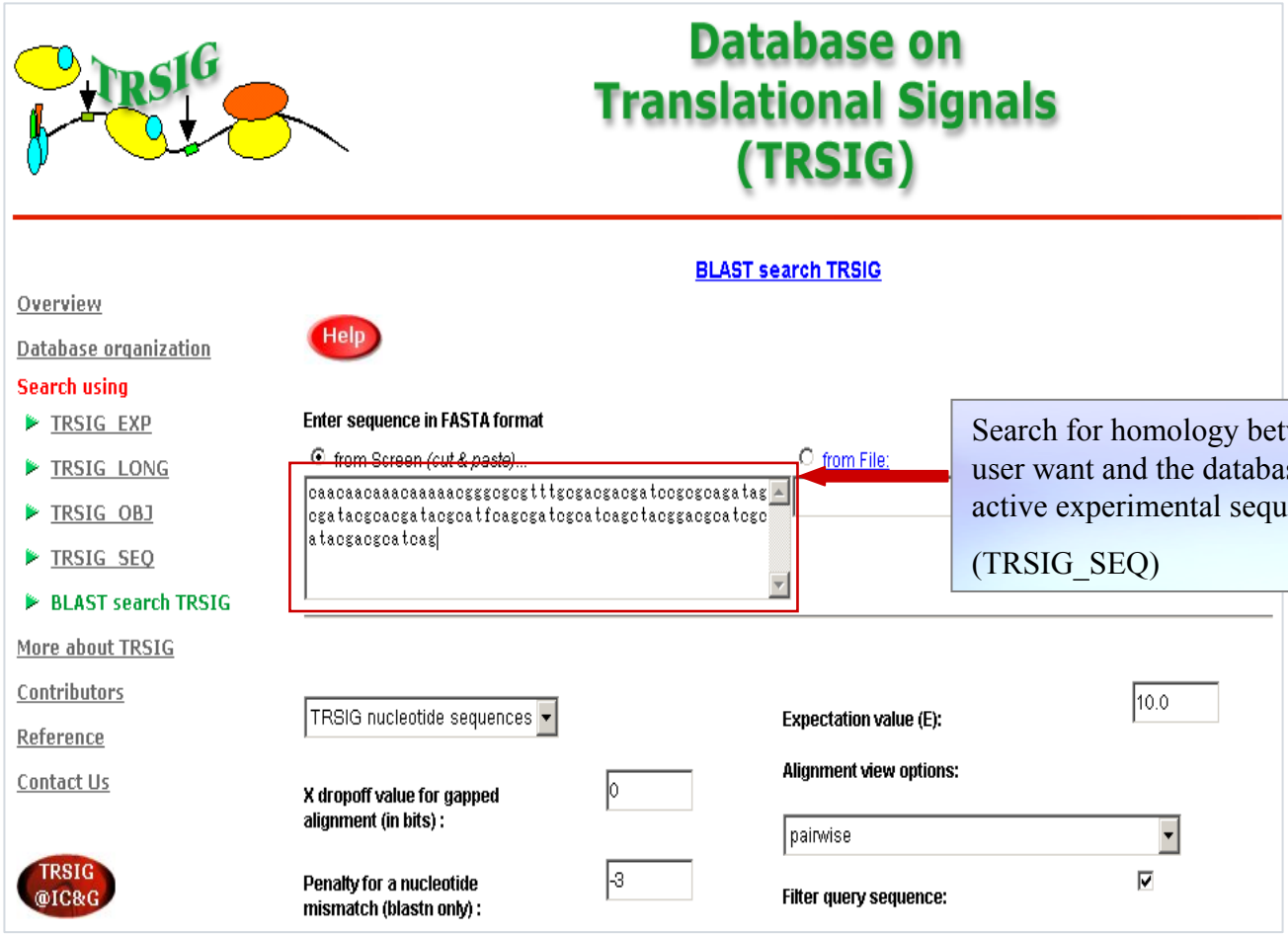
Plant Transgenesis optimization: Examine TRSIG experiment database entry by clicking on experiment field



Plant Transgenesis optimization: An example entry in TRSIG databse provides structured information on the translation enhancer of tobacco mosaic virus



Plant Transgenesis optimization: TRSIG Database can be searched for post-transcription control signals in mRNA the user want



Search for homology between mRNA the user want and the database for translation-active experimental sequences

(TRSIG_SEQ)

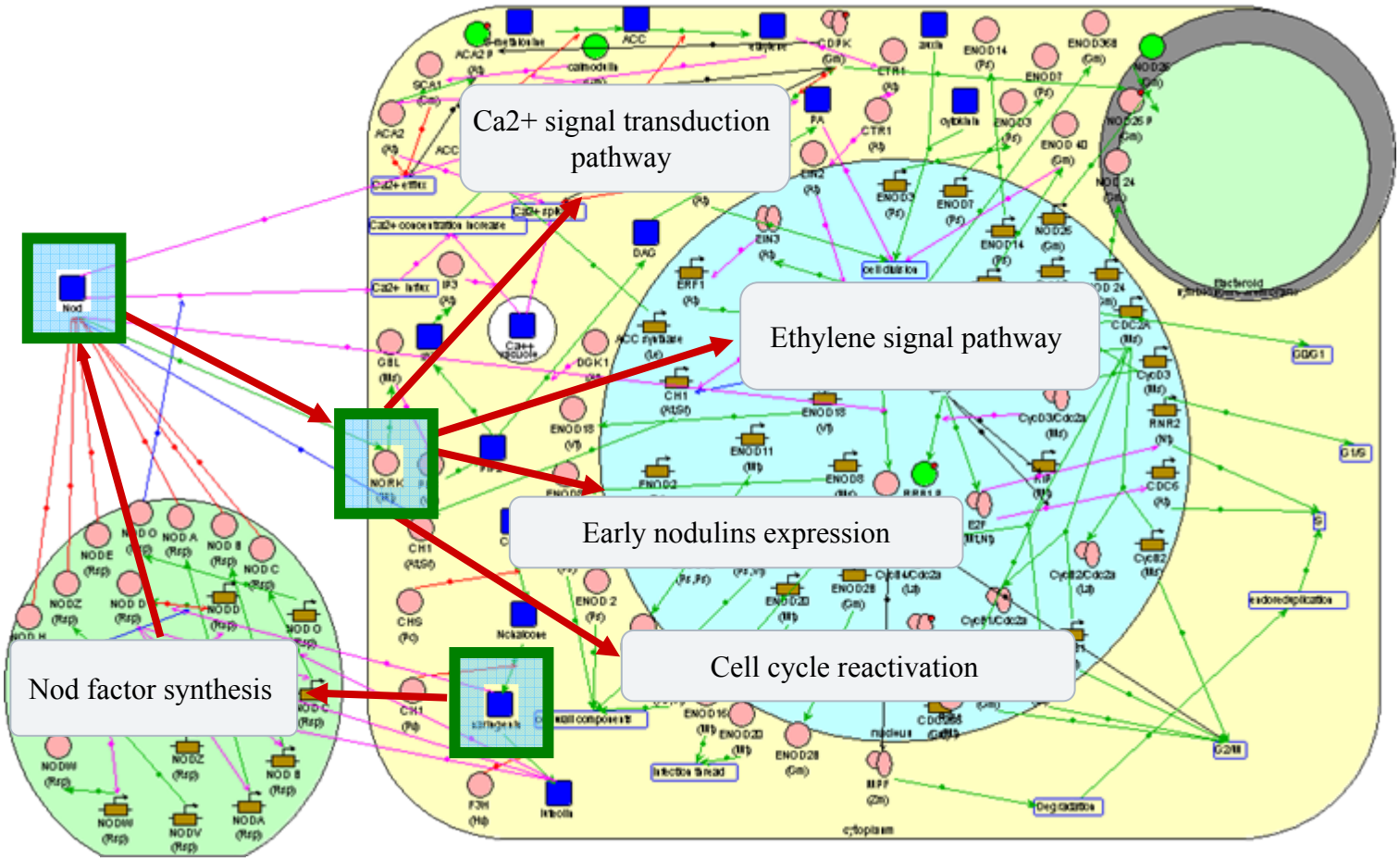
Plant Transgenesis optimization: An example of how to use TRSIG while searching for potential post-transcription expression control signals in mRNA

•glycoprotein P

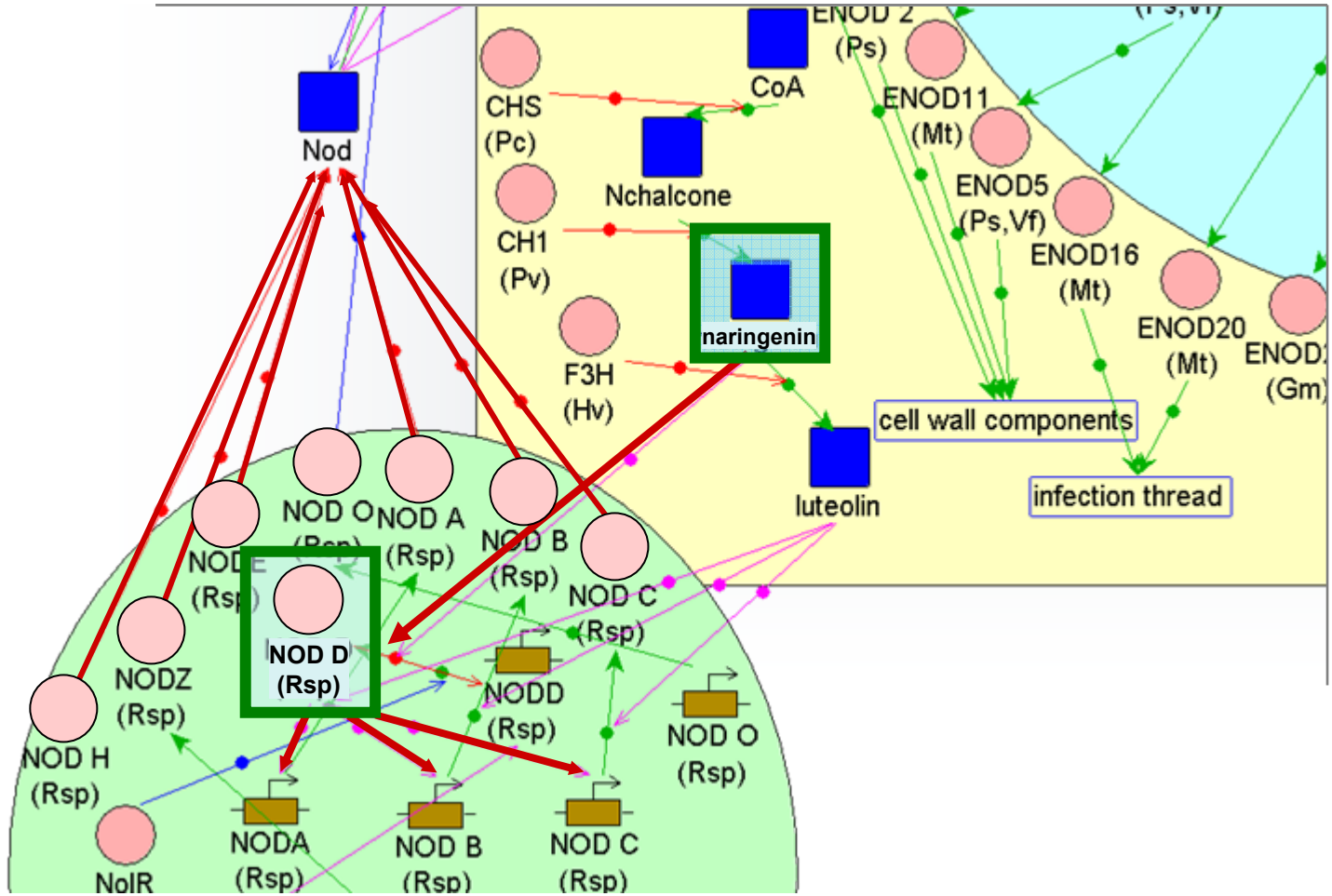
```
ATATPGP1 (145 nt)  ..cataaccaccaacaacaactcacgaagctccagagaaactcaccggaaATG
                                 ||||||||||||
TMV:               ..tacaacaattaccaacaacaacaaacaacaaacaacattacaattactatttaca
•ribosomal protein S1                        ||||||||||||
CLSORPS1G            ..cttatctgcgtatctcaacaaacaacaacataggaagaagatcaaagagtagc



TEV                    .cattctacttctattgcagcaatttaaatcatttcttttaaagcaaaagct
•cold-shock protein          ||||||||||
ATLTI78 (81 nt):  ..tttgattacttctattggaaagaaaaaaatctttggaaaATG
```

A search for homology between plant mRNA 5'-UTR and the nucleotide sequences database TRSIG_SEQ revealed that some mRNA contain fragments of translation RNA from tobacco mosaic virus (TMV) and tobacco etch virus (TEV)

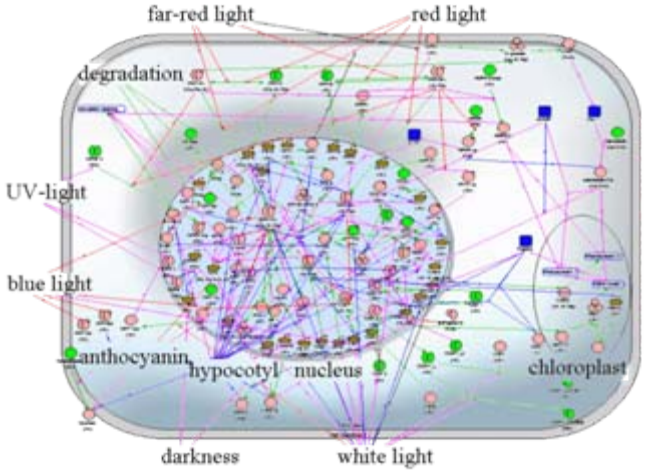
Plant Transgenesis optimization: A Hybrid Network of Nitrogen-Fixing Nodules: preinfection stage

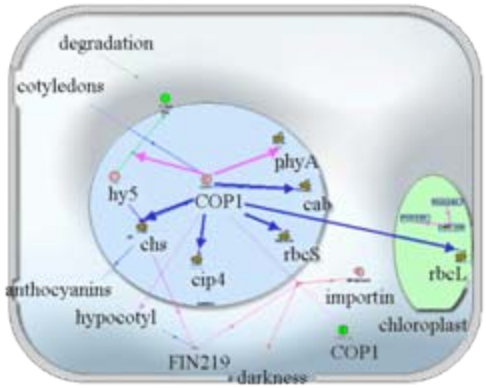


Plant Transgenesis optimization: Signal molecule of the bacteria - Nod factor synthesis

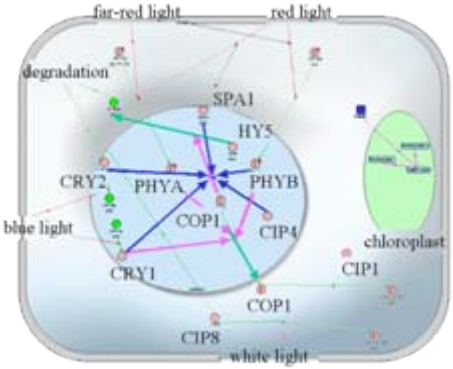


Plant Transgenesis optimization: Photomorphogenesis gene network

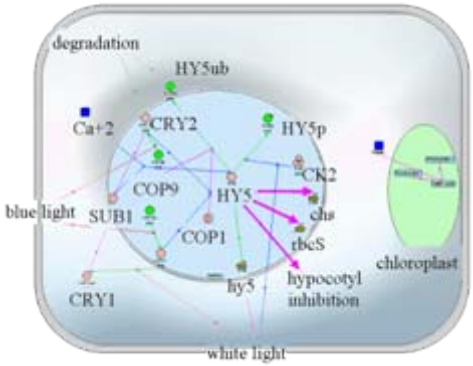




COP1 represses genes and promotes degradation of the transcription factor HY5 in darkness



Repression of COP1 activity by light is realized through photoreceptors and SPA and CIPs proteins.



Activation of HY5 transcription factor by light induces hypocotyl growth inhibition, activation transcription of genes.

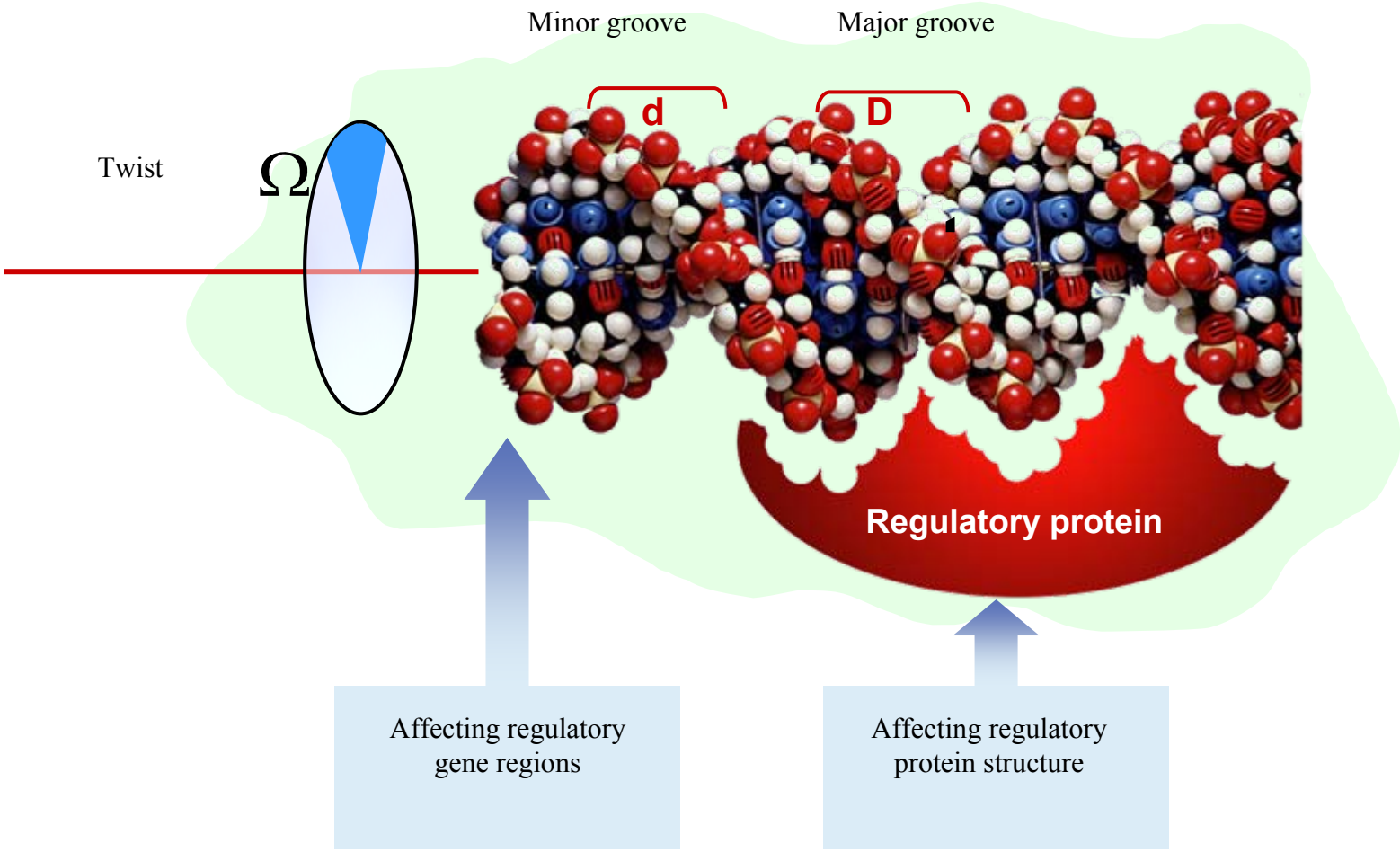
“Transgenesis”

Selected publications:

- Kochetov A.V., Ponomarenko M.P., Frolov A.S., Kisselev L.L., Kolchanov N.A. Prediction of eukaryotic mRNA translational properties // Bioinformatics. 1999. V.15. P.704-712.
- Kochetov AV, Grigorovich D, Kolchanov NA, Sarai A. Database on mRNA located eukaryotic expression signals influencing translation efficiency and specificity // Genome Informatics. Ser. 12. Edt: Matsuda et al. Universal Academy Press. Tokyo. Japan. 2001. P.492-493.
- Matushkin Yu.G., Likhoshvai V.A., Kochetov A.V. Local secondary structure may be a critical characteristic influencing translation of unicellular organisms mRNA. // In: Bioinformatics Of Genome Regulation And Structure. Ed. By N. Kolchanov and R. Hofestaedt, Kluwer Academic Publishers, Boston/Dordrecht/London, 2004, pp. 103-114.

Chapter 6
MOLECULAR PATHOLOGIES: COMPUTER ANALYSIS OF NUCLEOTIDE POLIMORPHISMS IN GENE REGULATORY REGIONS AND PROTEINS

Two types of regulatory mutations sensibly affecting regulatory functions of genes.



Computer-assisted experimental study of the influence of mutation G--->A on the binding of transcription factor YY1

Mutation within intron 6 of the triptophan 2,3 dioxygenase gene results in psychic pathologies such as

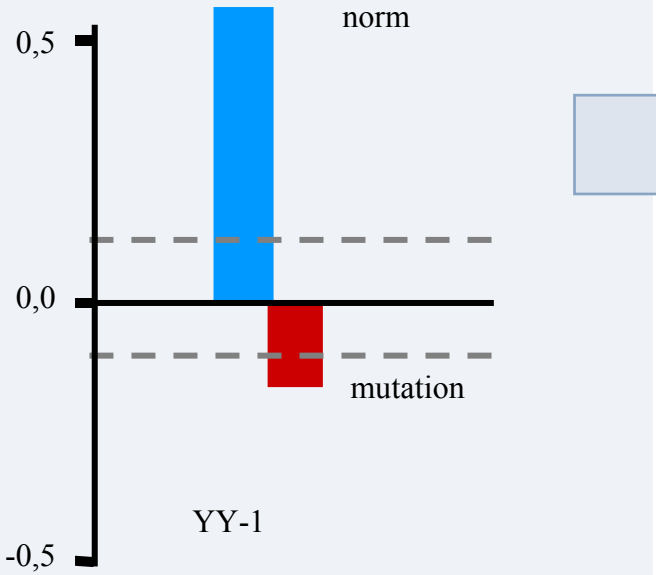
Tourette syndrome, attention deficit hyperactivity disorder and drug dependence etc.

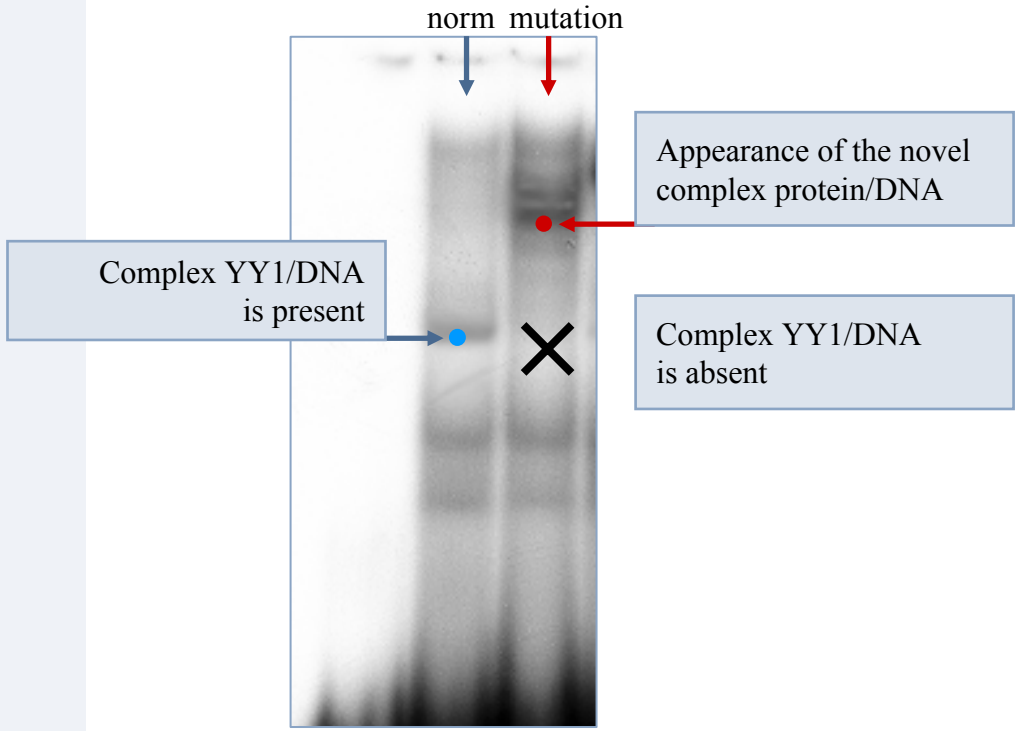
Wild type : 5'-cagtTGCCAAATAATG**G**CAGATAAGAATAGGGAG-3'

Mutation : 5'-cagtTGCCAAATAATG**A**CAGATAAGAATAGGGAG-3'

Vasiliev GV, Merkulov VM, Kobzev VF, Merkulova TI, Ponomarenko MP, Kolchanov NA. FEBS Lett. 1999 Nov 26;462(1-2):85-8.

Computer-assisted experimental study of the influence of mutation G--->A on the binding of transcription factor YY1

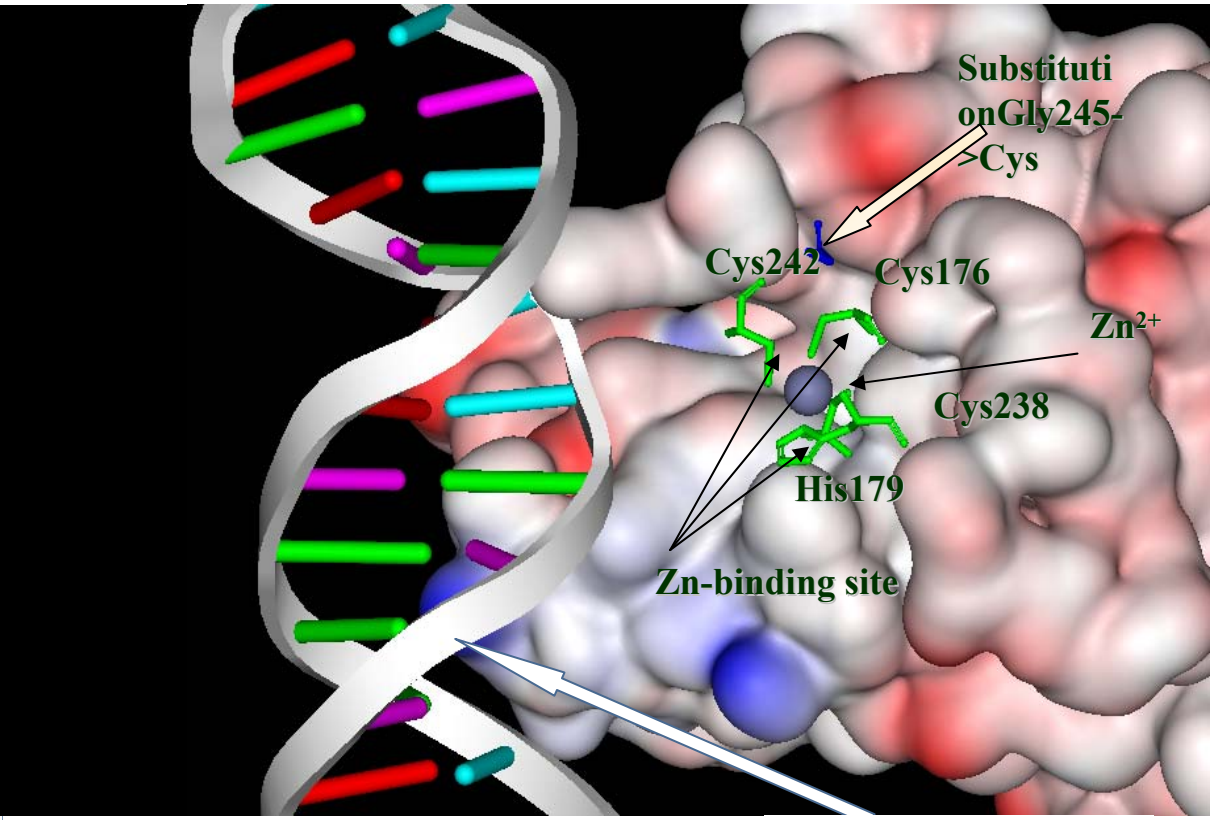




Computer analysis suggests that the intron of the gene being studied contains the transcription factor YY1 binding site, which is likely to be damaged by mutation. Further experimental studies confirmed this theoretical prediction.

Computational proteomics: PDB-site and PDB-siteScan tools.

Studies of Molecular mechanism for impaired function of mutant p53 protein (SubstitutionGly245->Cys) leading to tumor development

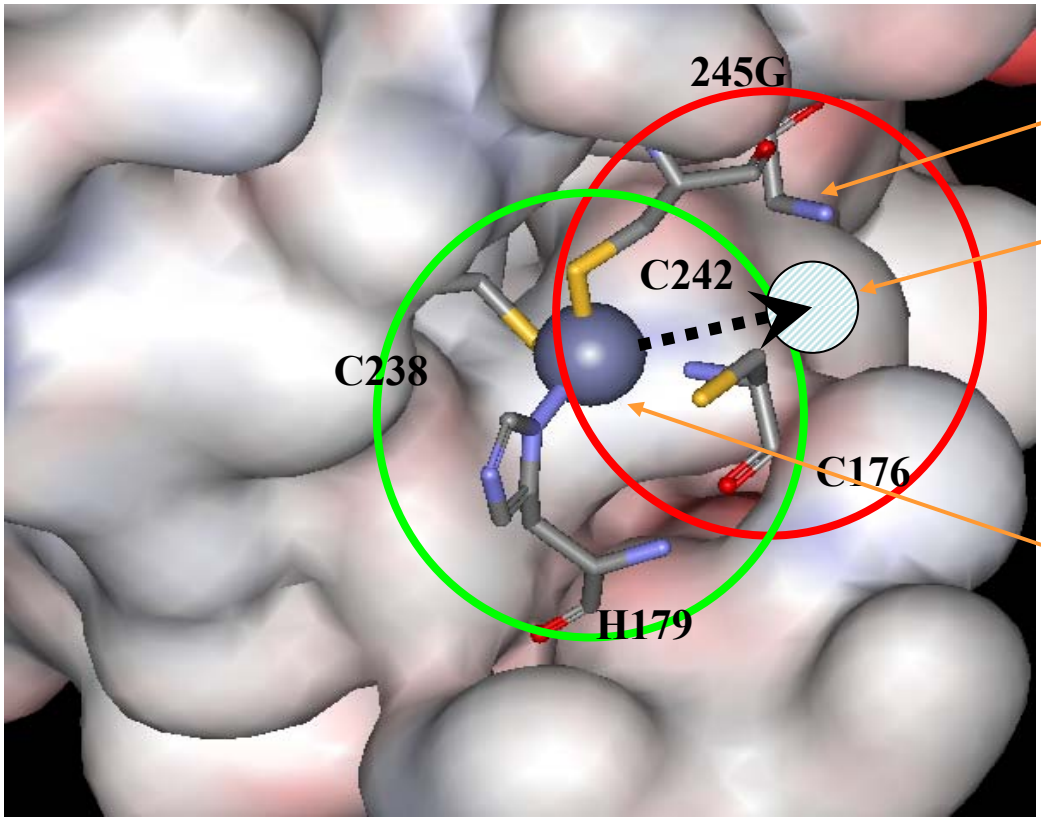


Normal DNA-binding

Structural analysis of Gly245→Cys mutation in DNA binding domain of p53 protein that causes inherited tumor predisposition (Li-Fraumeni syndrome).

Computational proteomics: PDB-site and PDB-siteScan tools.

Molecular mechanism for impaired function of mutant p53 protein (SubstitutionGly245->Cys) leading to tumor development



245 G->C mutation

Predicted shift of zinc ion

Zinc ion

Research suggests that mutation near the normal Zn2+ binding site gives birth to a putative Zn2+ binding site.

Residues of normal Zn^{2+} binding site

Residues of new putative Zn^{2+} binding site

Chapter 7
COMPUTATIONAL EVOLUTIONARY BIOLOGY

7.1. Special aspects of the molecular evolution of proteins

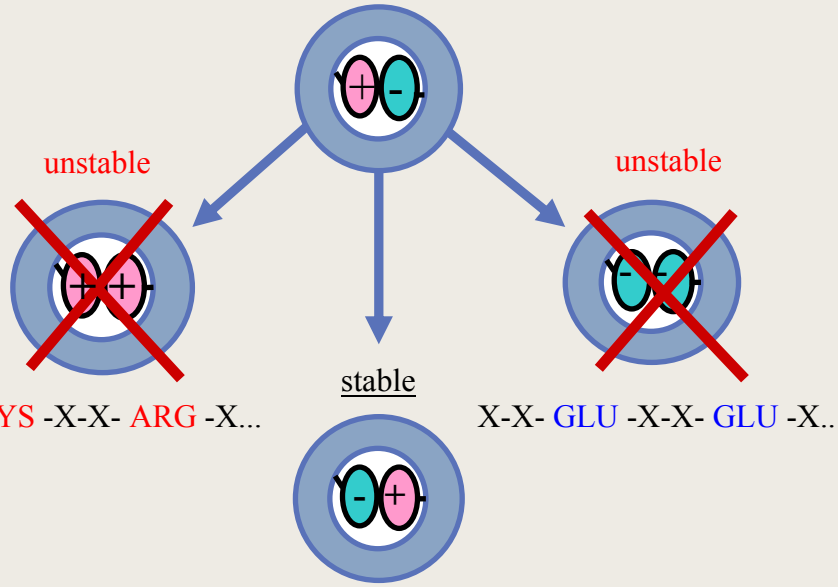
7.1.1 Coordinated substitutions in proteins

An example of compensatory (relatively to the charge sign) AMINO acid substitutions at a pair of protein positions

LIVKSMDGAL
STMECARLIT
GTSDNSHQLI
LIMKVVDGYA

Analysis of multiple alignment of sequences of a protein family

..X-X-X- LYS -X-X- ASP -X-X-X..



X-X- LYS -X-X- ARG -X...

X-X- GLU -X-X- GLU -X..

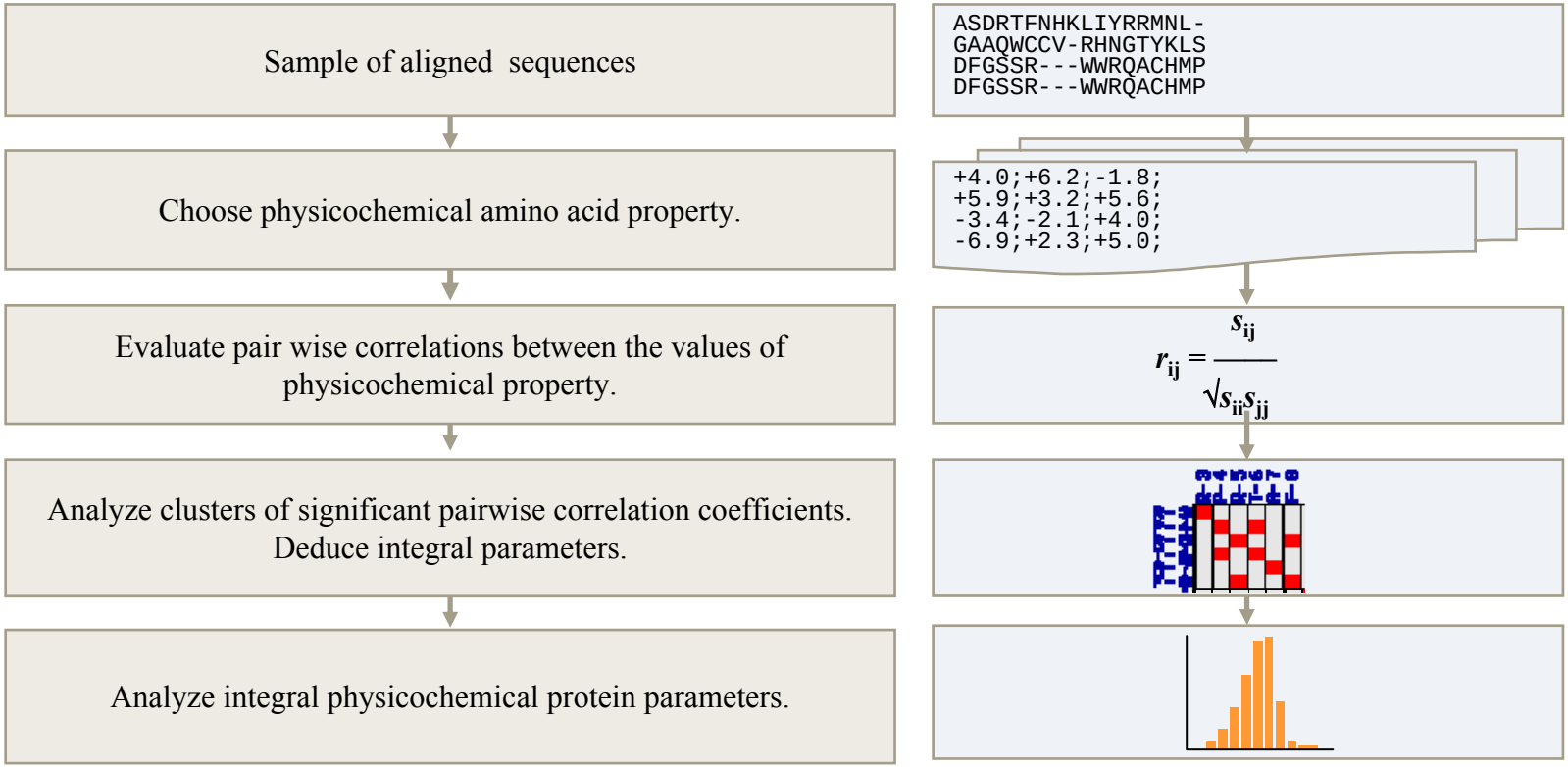
…X-X- GLU -X-X- ARG -X...

CRASP: software package for revealing protein position pairs with the co-adaptive evolutionary mode of substitutions

Availability:http://wwwmgs2.bionet.nsc.ru/mgs/programs/crasp/

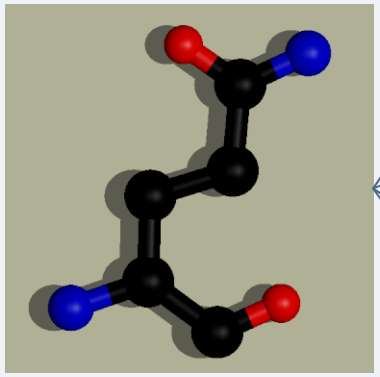
The main tasks of analysis:

- To reveal pairs (or clusters) of protein positions with the coordinated substitutions;
- To reveal integral physicochemical protein characteristics, the conservativeness of which is determined by coordinated residue substitutions.



More than 100 important physical and chemical amino acid characteristics are considered

The values of characteristics reflect specific interactions of residues:

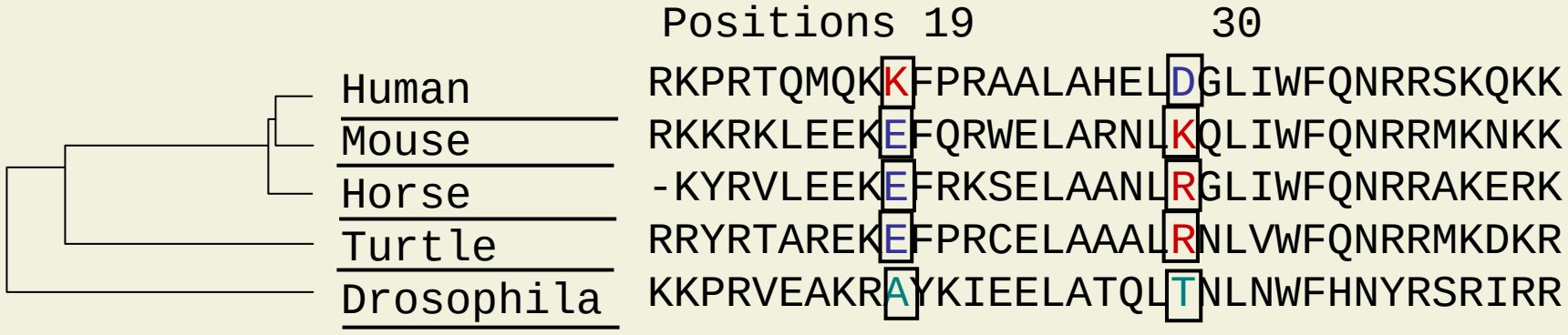


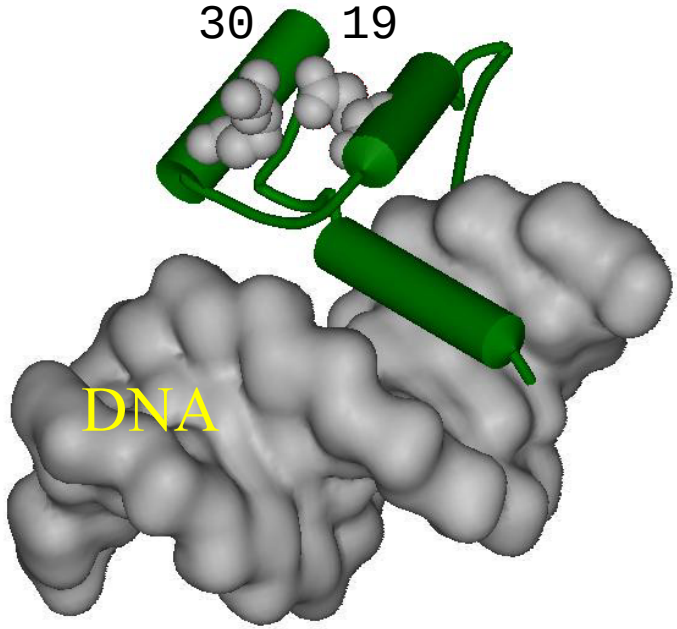
- HYDROPHOBICITY
- SECONDARY STRUCTURE PROPENSITY
- SIZE
- CHARGE

Main goals of the CRASP package analysis:

- detection of protein position pairs with co-adaptive residue substitutions;
- detection of integral conserved protein characteristics caused by the co-adaptive residue substitutions.

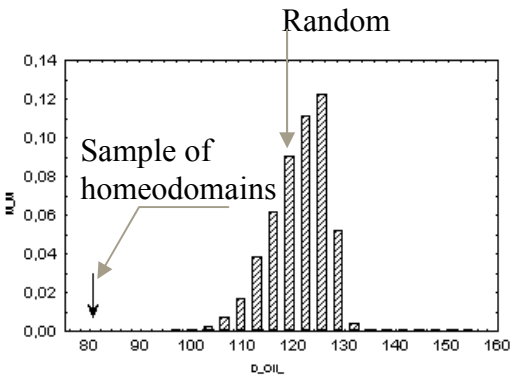
Sample compensatory substitutions in a DNA-binding domain in the "homeodomain" family

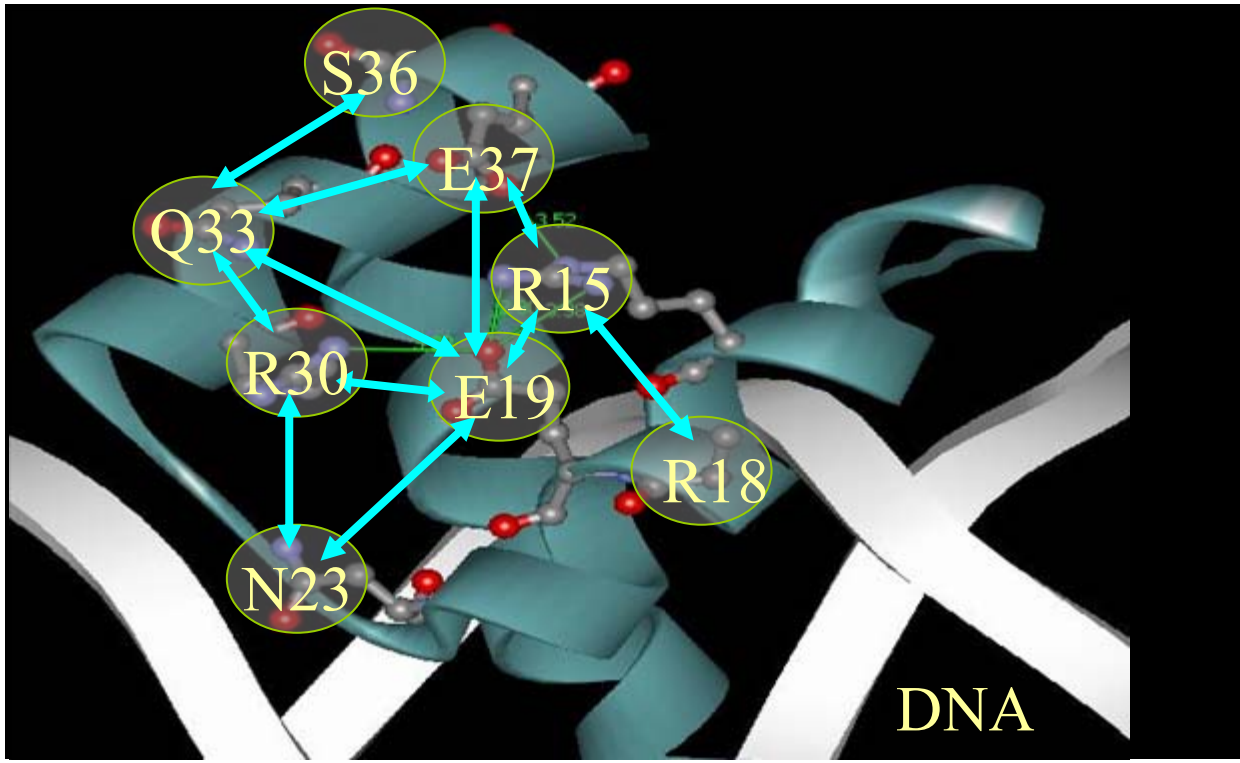




Cluster of correlating positions in the proteins of “homeodomain” family

Localization of residues in the spatial structure of the complex “homeodomain”-DNA. Negative dependencies are given by blue arrows.





Conservation of the summarized charge of a cluster

7.1. Special aspects of the molecular evolution of proteins

7.1.2 Evolution of the functional sites of proteins

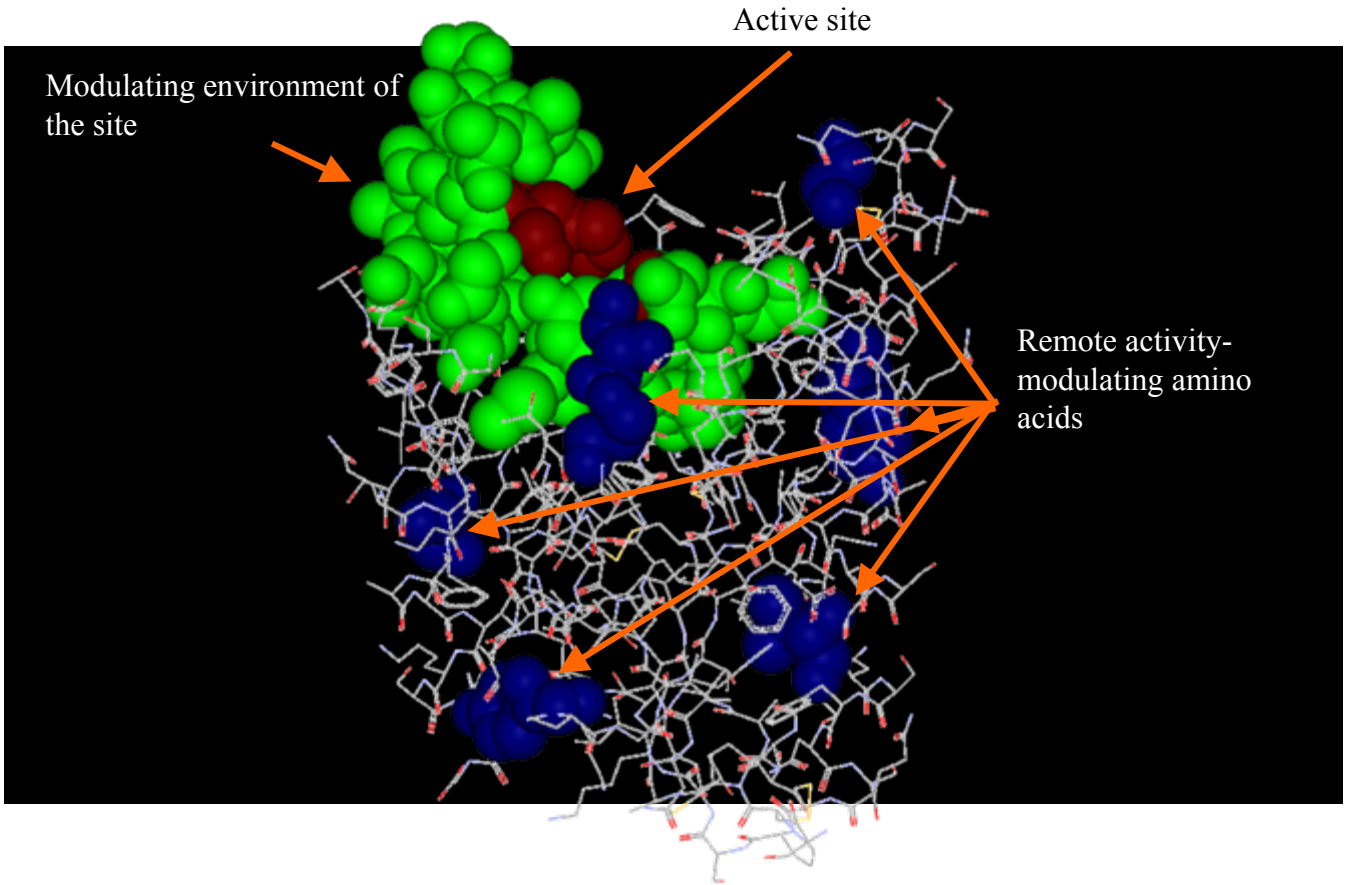
Analysis of quantitative structure-activity relationships in homologous protein families

- Identification of residues affecting protein activity – activity modulating centers
- Establishment of relations between physicochemical properties of activity modulating centers and protein activity

Tasks:

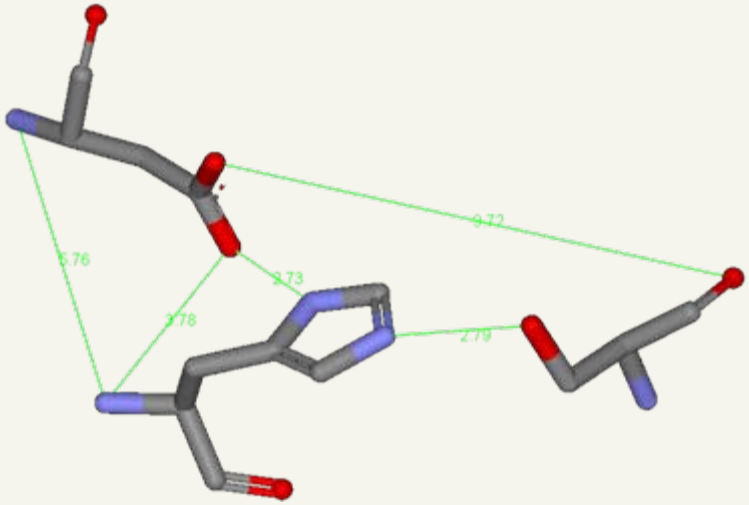
Design of protein engineering experiments,
Reconstruction of ancestral protein activities
Annotation of protein activities in databases of protein sequences
Annotation of artificial protein activities

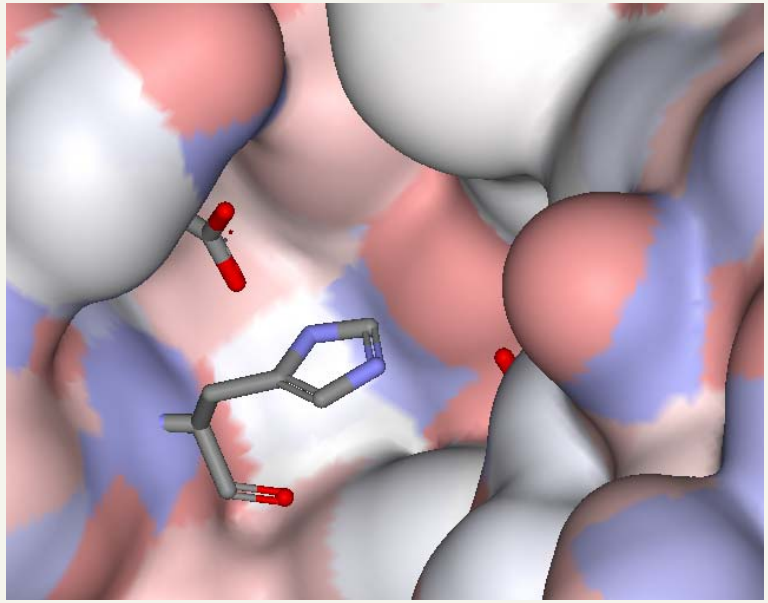
The concept of structural and functional organization of proteins



The functional centers of proteins are central objects in computational proteomics

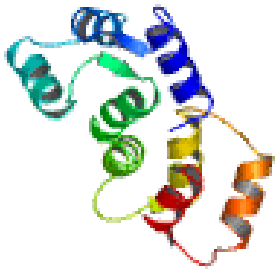
Physical and structural characteristics the specificity of functional sites depends upon





- Physical and chemical properties of the amino acids at the site
- Arrangement of amino acids in primary structure
- Specific features of conformation
- Secondary structure
- Accessibility to solvent
- Polarity and charge of the environment

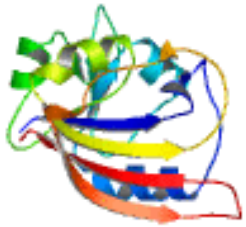
Function associated changes in protein conformation



Calmodulin

HIVRT

SERCA pump

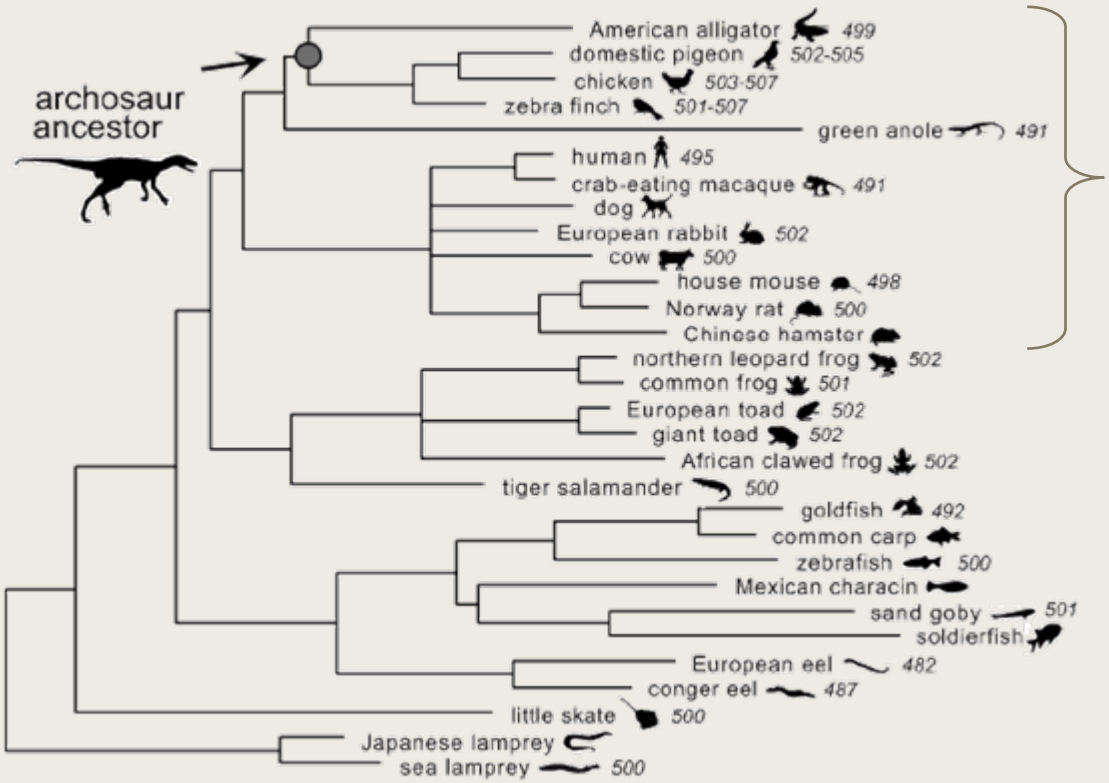


DHFR

HIV-protease

SOD

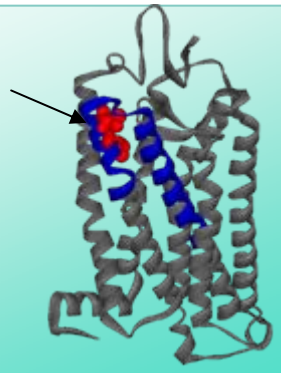
Prediction of the visible spectrum of archosaur sight (λmax for ancestral archosaur rodopsin)

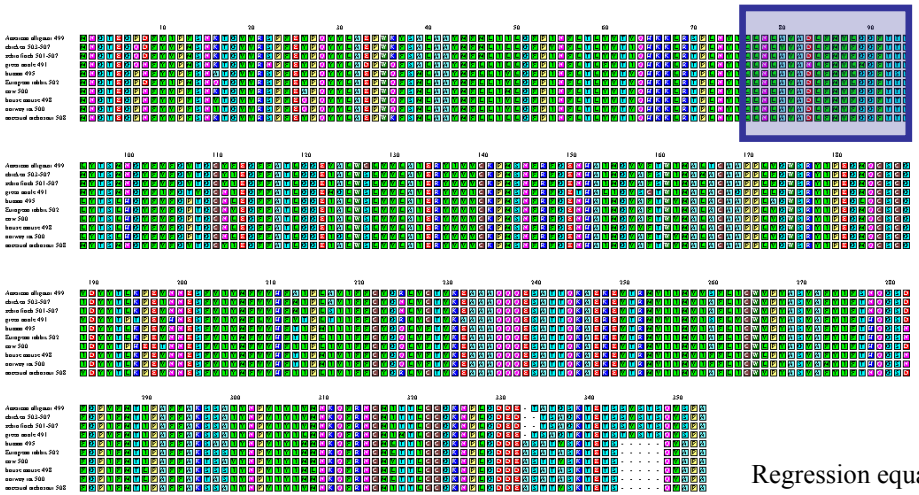


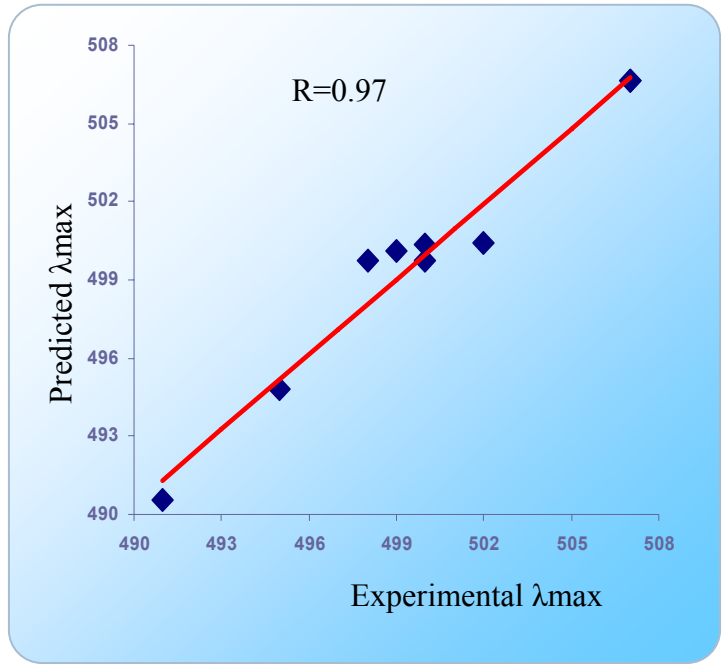
Phylogenyis from Chang et al., *Mol. Biol. Evol.* 19(9):1483–1489. 2002

λmax prediction for archosaur ancestral rodopsin by WebProAnalyst ссылка 28Ch4

Residues affecting shift of wavelength. The results agree with those of *Briscoe et al, Mol. Biol. Evol. (2001).*

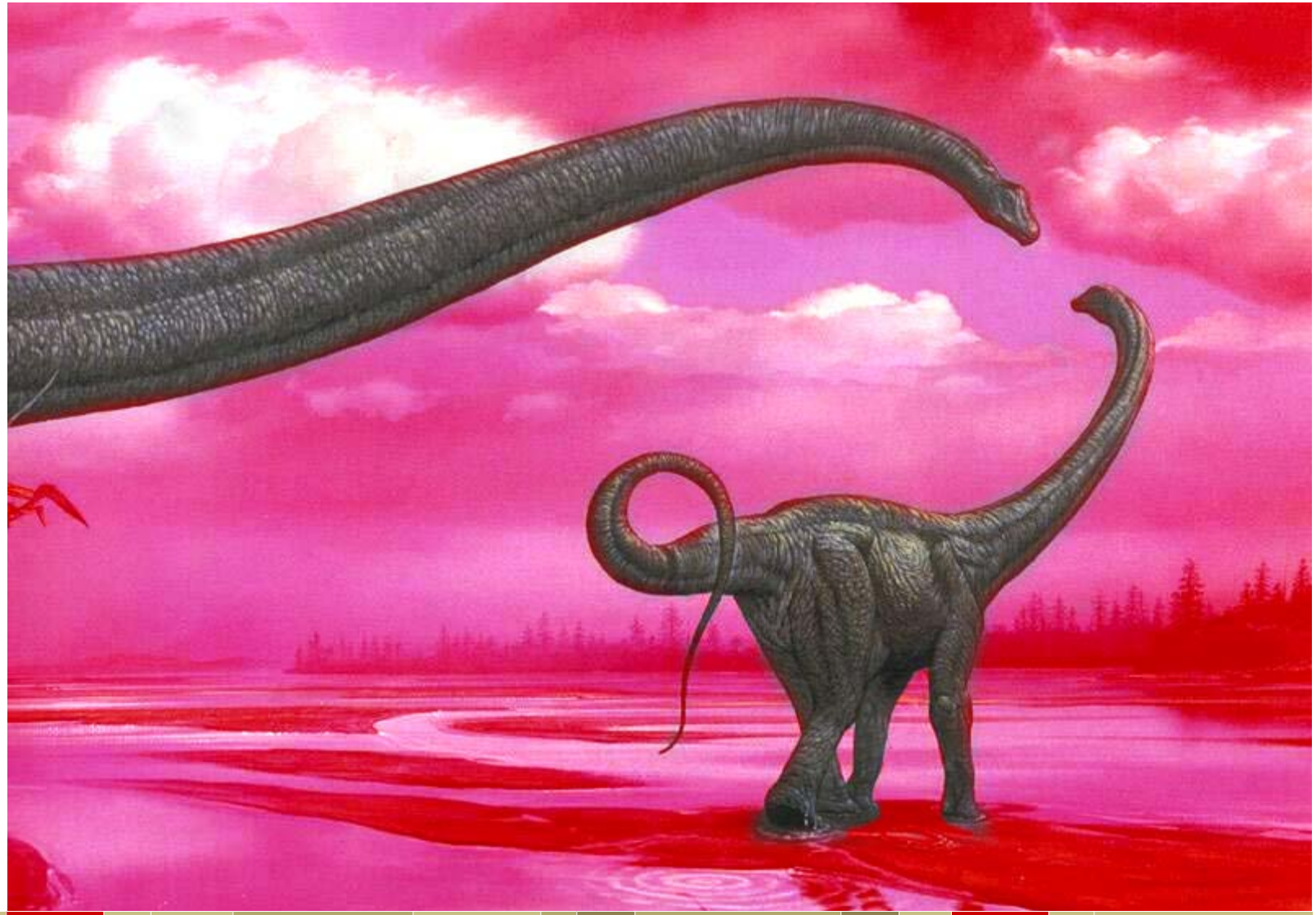






Predicted λmax: **Y= 506.7**
Measured λmax: **Y= 508**

Regression equation:
$Y=15.784*X_1-467.266*X_2-37.661$
X1 – mean for site isoelectric point (Bogard)
X2 – hydrophobic moment (Eisenberg)



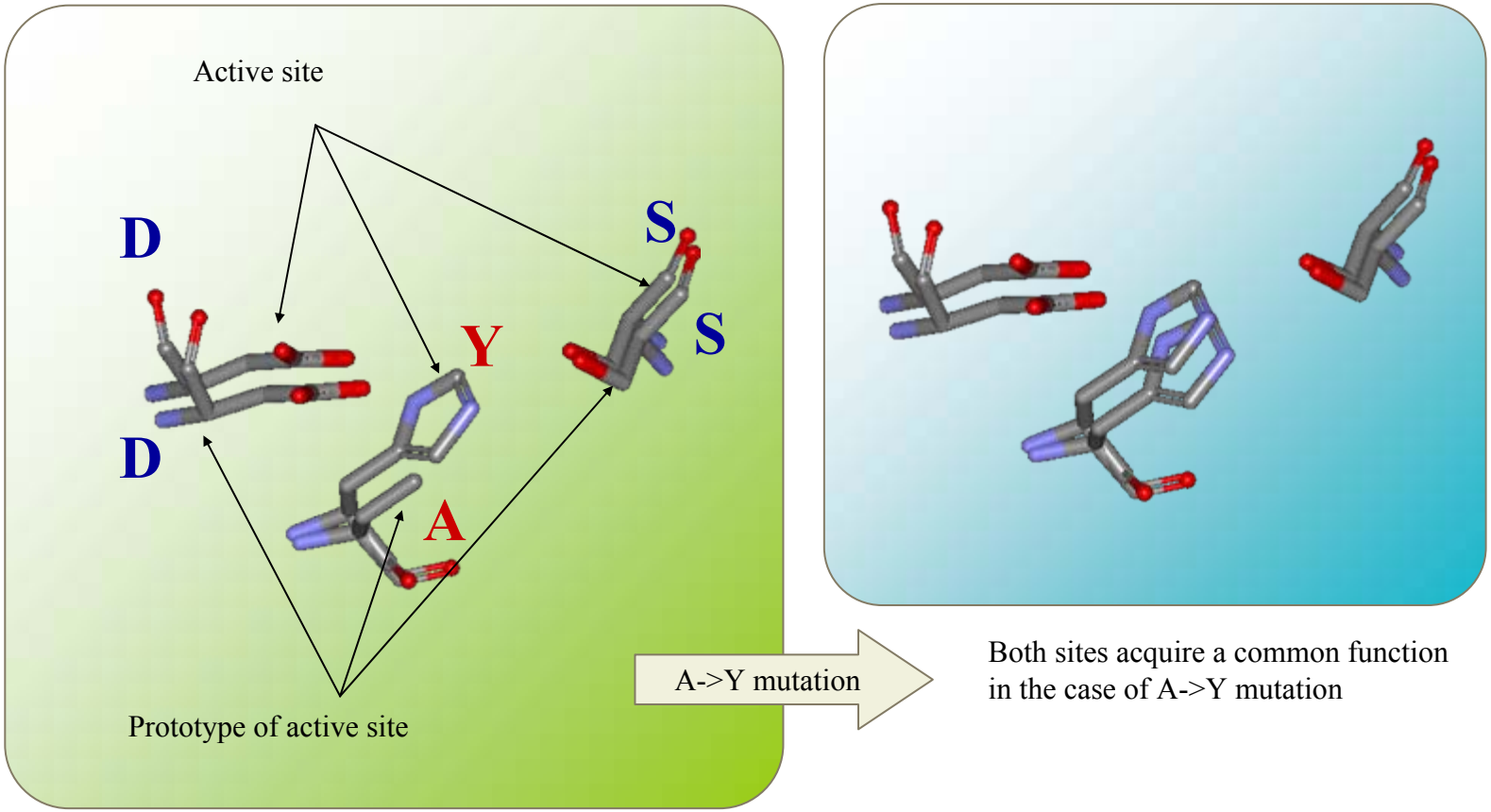
A new approach to molecular evolution of functional sites: "Paleontological excavations" of functional sites in protein tertiary structures

Proteins contain:

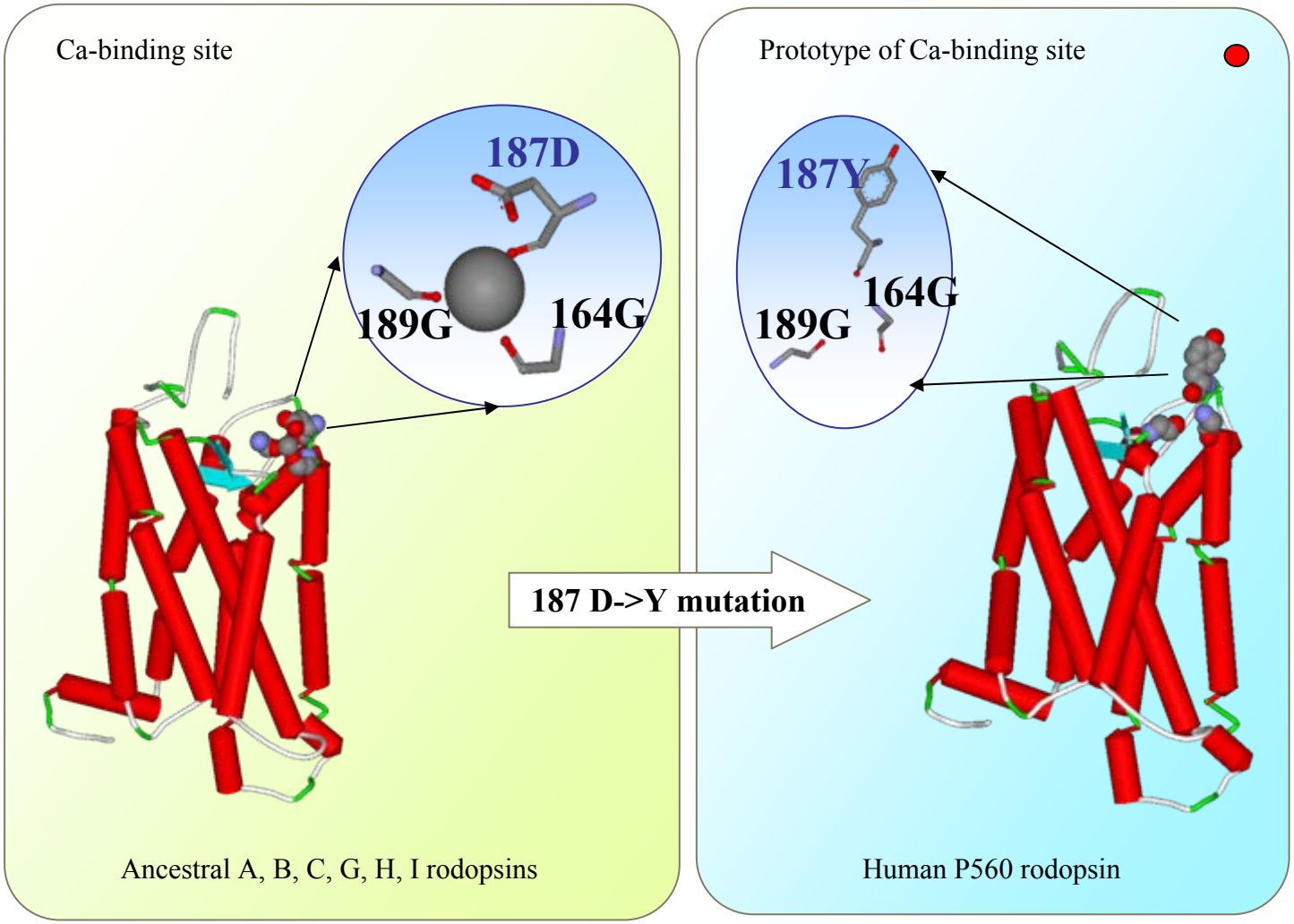
- Active functional sites
- Postsites, traces of earlier active functional sites effaced by mutation fixation
- Presites, regions of a protein structure that are potentially capable of forming new functional sites as a result of mutations

Prototypes of functional sites (post- and presites) have become detectable by analysis of protein tertiary structure

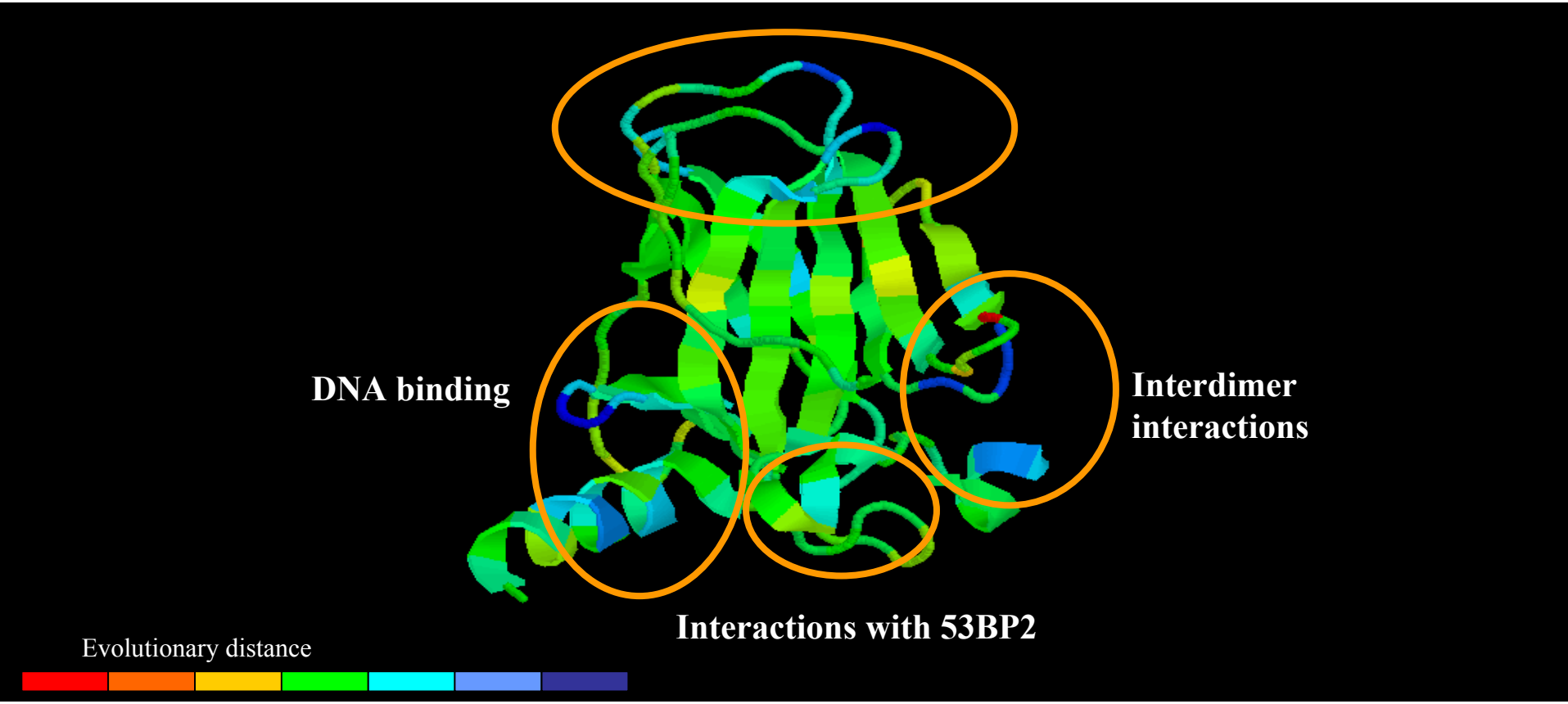
"Paleontological excavations" of functional sites in protein tertiary structures



Computer redesign of novel protein functional sites using point mutations



Distribution of site prototypes in spatial DNA-binding domain of p53 protein

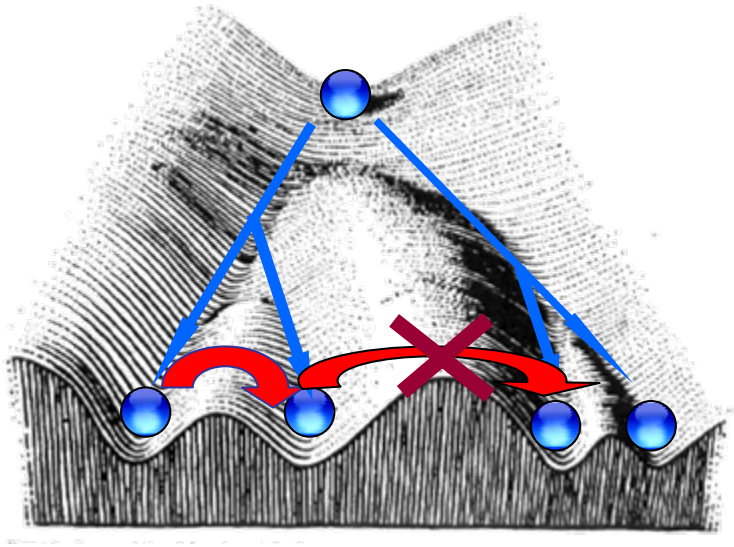


7.2. Molecular phylogeny and special aspects of the evolution of gene networks of morphogenesis

Conquesting an adative landscape during speciation

As evolution goes on, the organisms gradually become adaptive to particular environments by fixation of adaptive mutations, every time at a limited number of loci. Fixation of adaptive mutations causes the formation of specific morphological, physiological, biochemical and other systems giving the organisms the ability to survive in these environments and banning them from others.

In this situation, the evolution of a species switches from driving (adaptive) to stabilizing (as I.I. Shmalgauzen put it) mode. This switch-over results in the formation of powerful hierarchical systems with negative feedbacks, which stabilize the phenotype. This stabilization abrogates some evolutionary trends, canalizes others, favors the fixation of neutral mutations and the accumulation of damaging mutations under cover of regulatory circuits with negative feedbacks. This is what makes adaptive evolution stop. When stabilizing selection lasts long enough, the overall fitness of the species reduces.

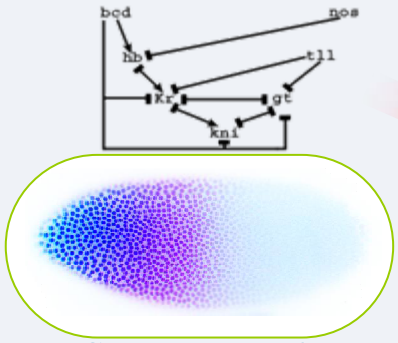


The movements of the ball portray the process of adaptation to a particular environment, alongside speciation.
Valleys stand for adaptive ecological niches.
Ridges stand for reproductive barriers.

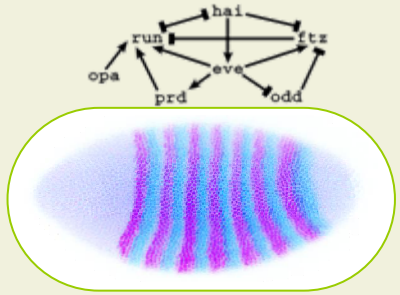
Module-based hierarchical system of gene network automata controlling embryogenesis in D. melanogaster

A hybrid gene network of initial stages of embryogenesis in Drosophila: the formation of the *bcd* and *hb* morphogen gradients along the anterior-posterioraxis of the embryo (blue and purple colors, respectively). These gradients set up initial conditions for the functioning of the gene network to come

Gene automaton 1

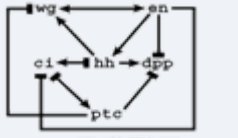


Gene automaton 2

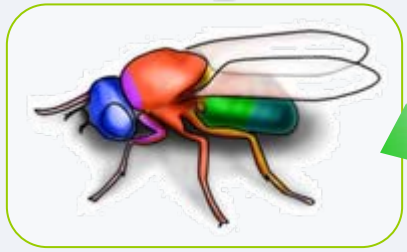


The gene network of the initial stages of segmentation in the Drosophila embryo. The figure below presents the patter of expression of the *eve* and *ftz genes* (purple and blue colors, respectively)

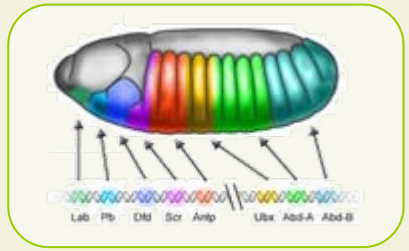
Gene automaton 3



A simplified scematic of the gene network of final stage of segmentation in Drosophila. The figure below presents the patter of expression of the *en* gene

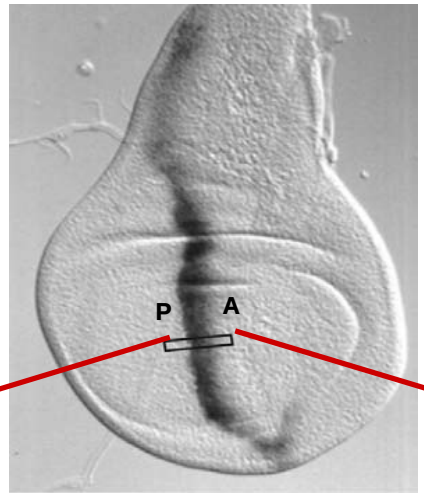


Gene automaton 4



Effects of *Hox-genes* on the specification of the developed body segments in Drosophila

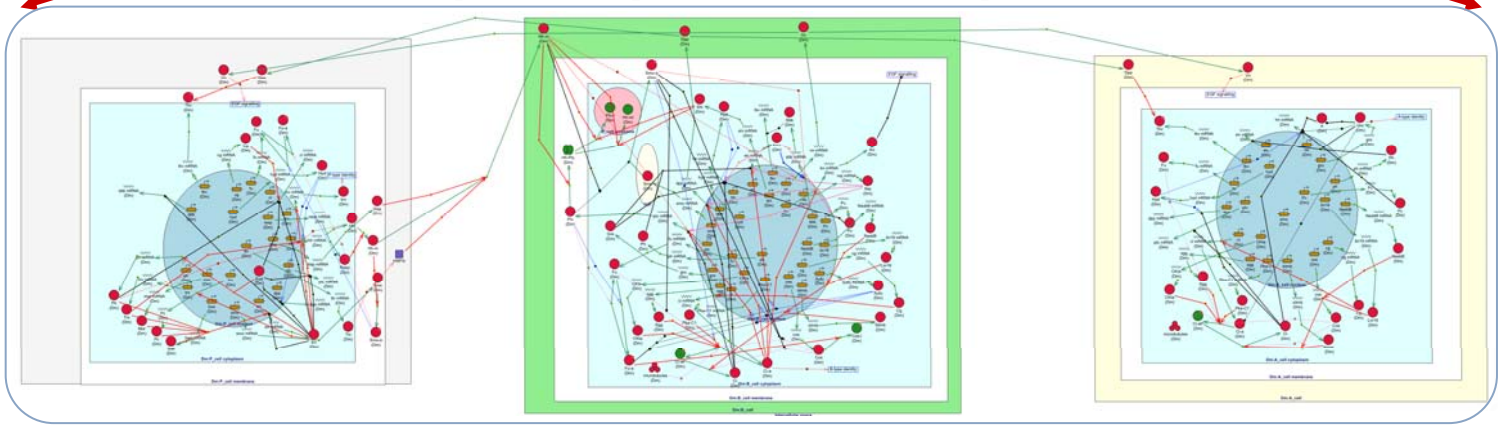
Gene network of the formation of the anterior-posterior border of wing compartments



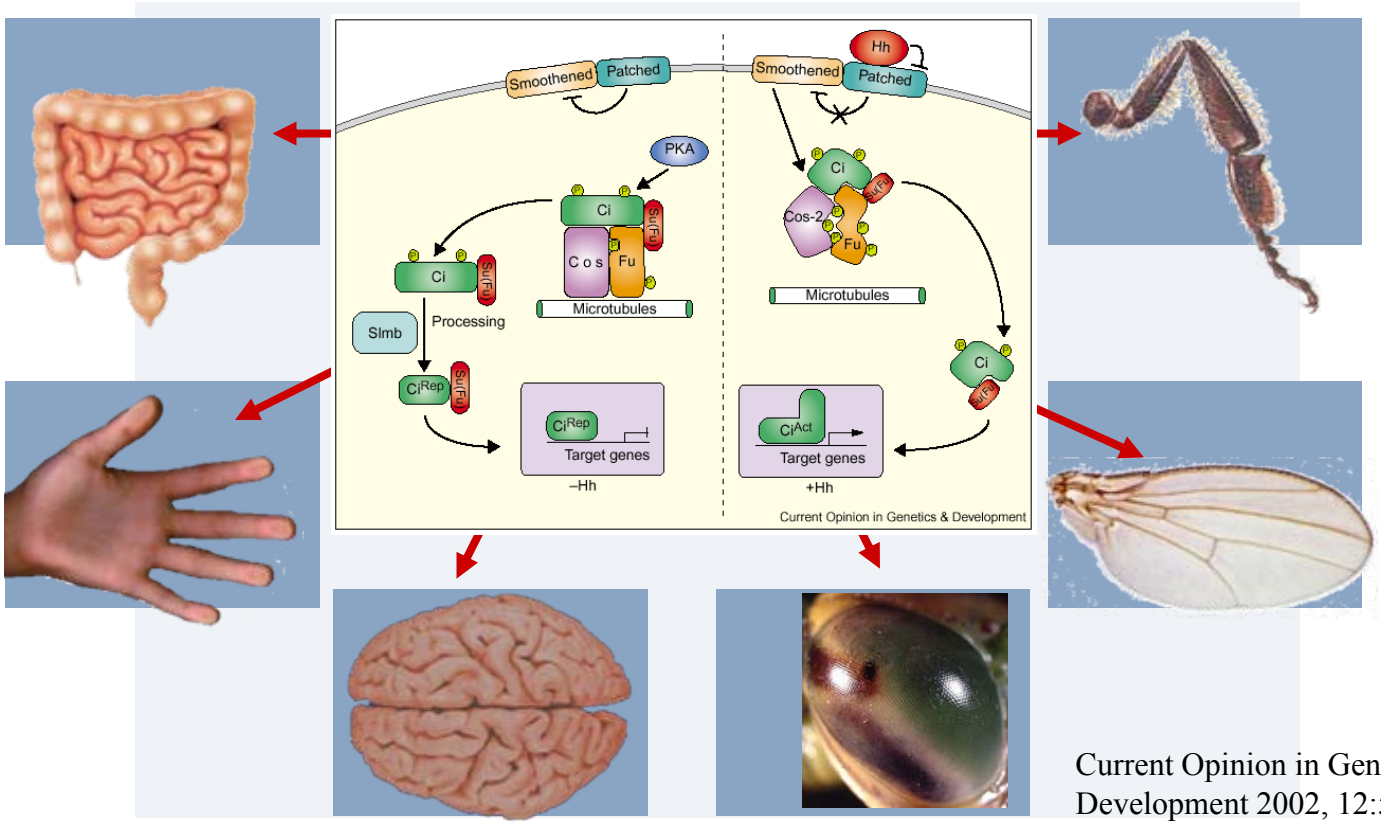
Gene network components:

Cell types	3;
genes	66;
RNA	65;
proteins	77;
reaction	250;
processes	4.

P – posterior compartment, A – anterior compartment.



Hh-cascade of signal transmission is a universal mechanism involved in the control of morphogenesis in many-celled organisms



Current Opinion in Genetics & Development 2002, 12:503–511

Requirement for adaptive evolution

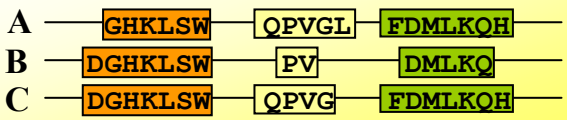
$$k_a / k_s > 1$$

k_a – adaptive mutation fixation rate,
k_s – neutral mutation fixation rate.

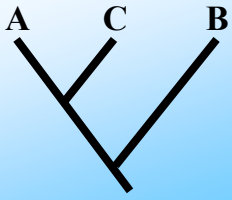
The basic property of the gene code is its degeneracy. *Synonymous* nucleotide substitutions in codons are the ones that do not modify amino acids they encode; *non-synonymous* are the ones that do.

The approach used to address the problem

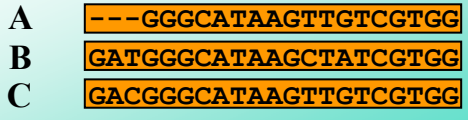
1 Aligning amino acid sequences



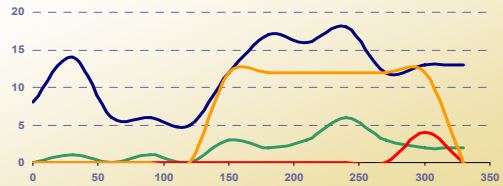
2 Building phylogenetic trees by amino acid sequences



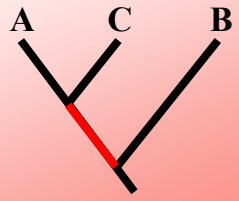
3 Aligning nucleotide sequences on the basis of amino acid sequence alignments



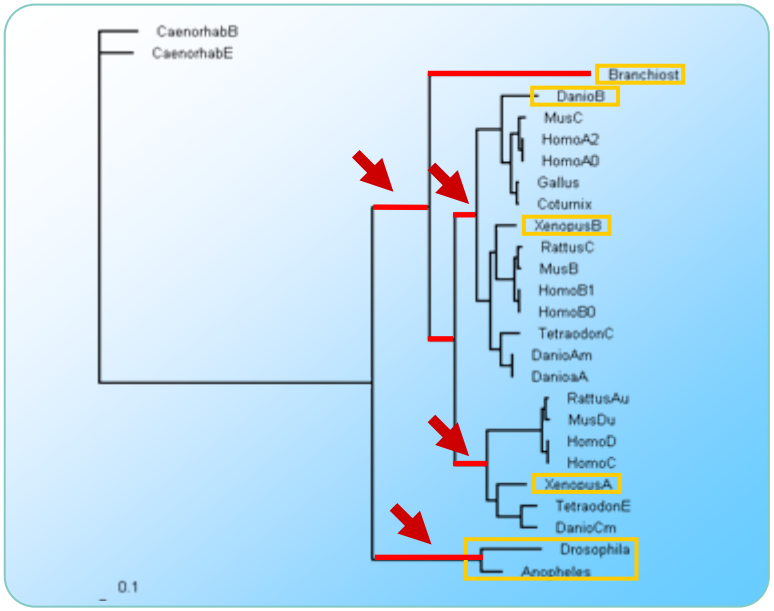
4 Identification of gene regions subject to adaptive evolution

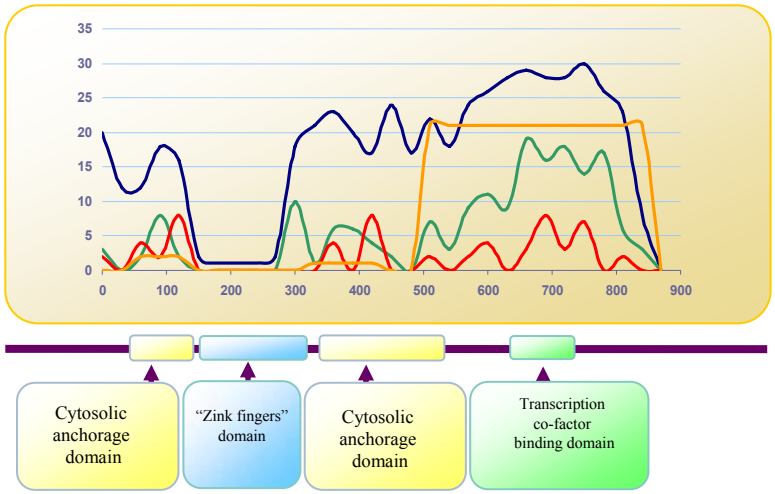


5 Identification of the phylogenetic tree branches, on which the adaptive evolution of the gene region revealed at the previous stage is under way

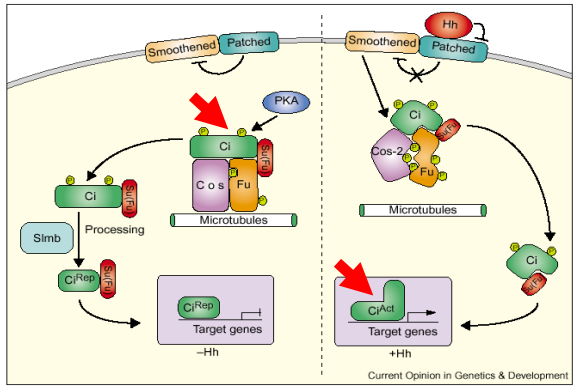


Adaptive evolution of the *cubitus interruptus* (*ci*) gene encoding the transcription factor

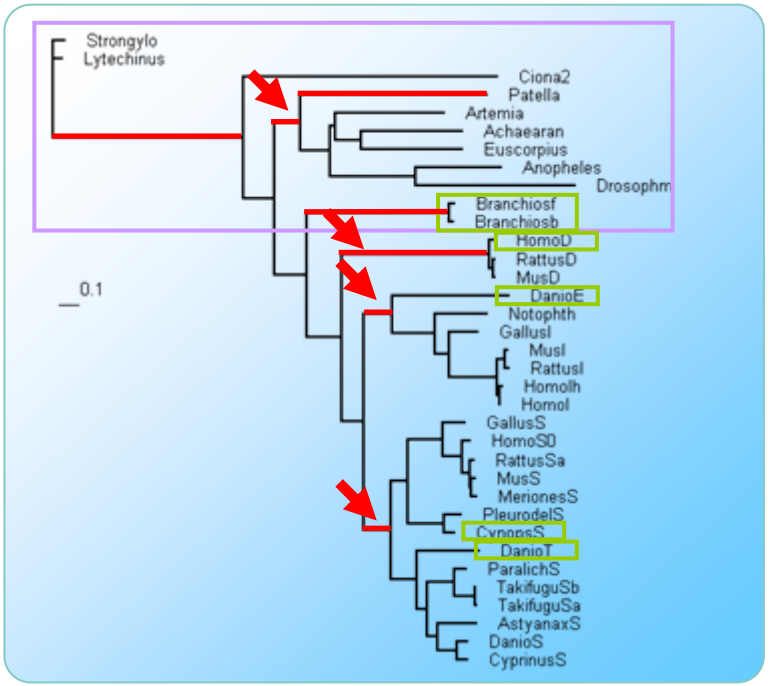


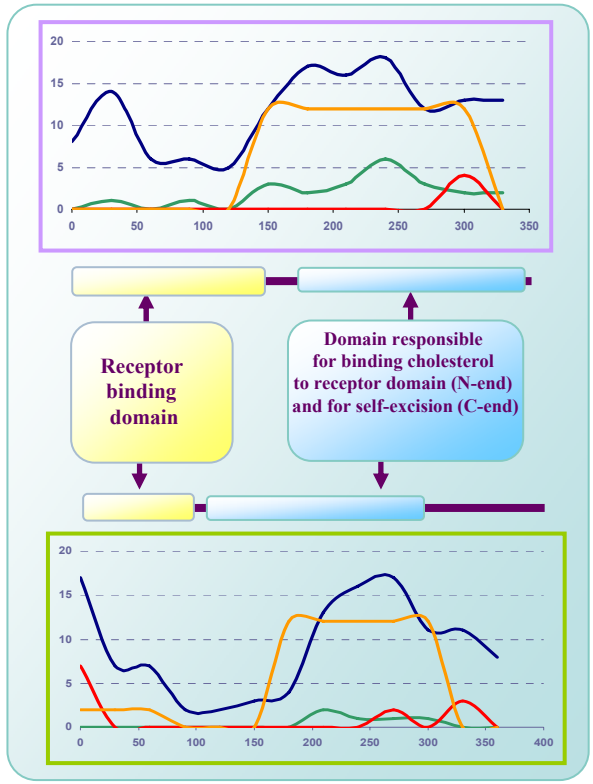


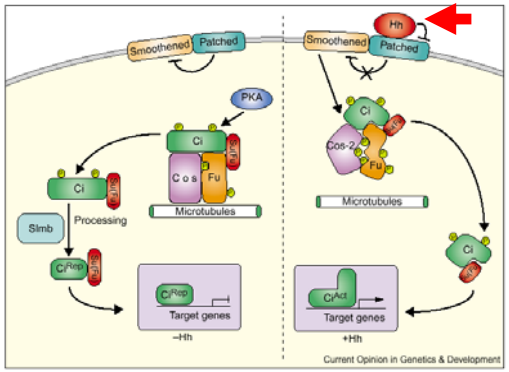
It has been revealed that the adaptive evolution of the *ci* gene:
(1) Correlates with the emergence of large taxa of bilateral organisms: arthropods and vertebrates;
(2) after duplication of the *ci* gene, its paralogs underwent adaptive evolution;
In protein CI, adaptive evolution is under way in domains responsible for keeping the protein in the cytoplasm and for binding to transcription co-factors.



Adaptive evolution of the Hedgehog gene (Hh) encoding the morphogene

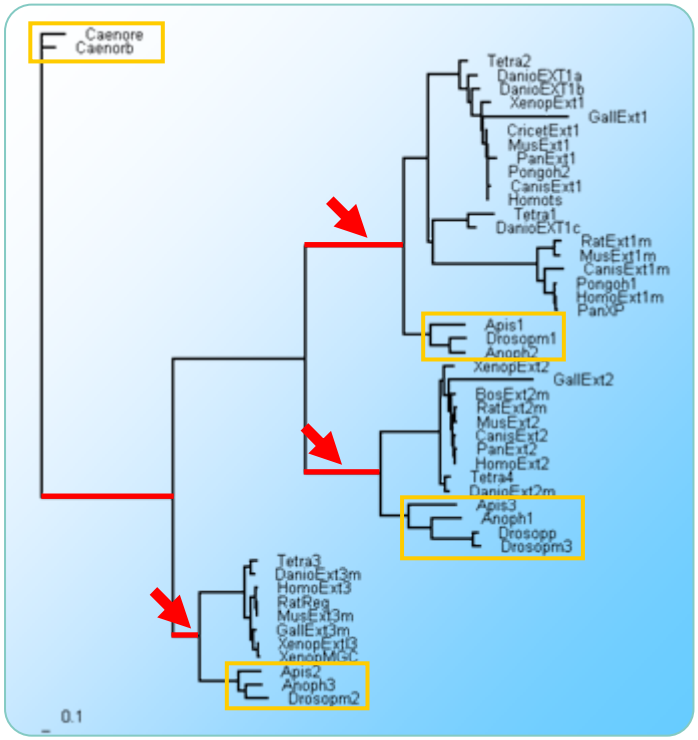


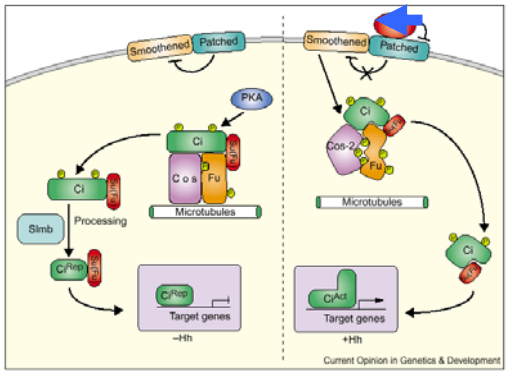




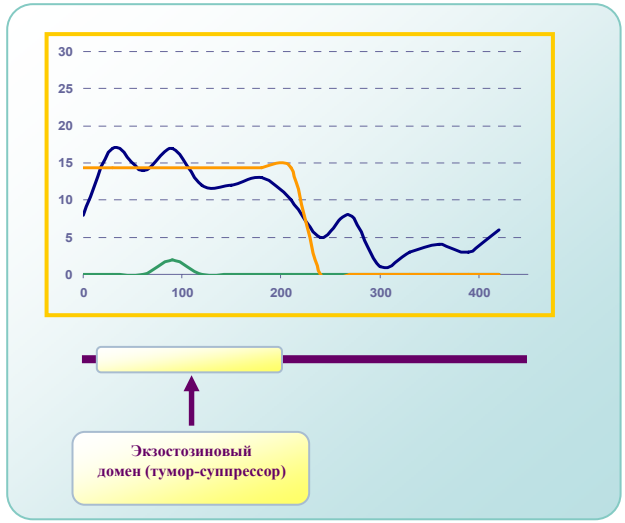
Adaptive evolution of the *Hedgehog* (*Hh*) gene encoding the morphogen

It has been revealed that the adaptive evolution of the *Hh* gene:
(1) correlates with the emergence of arthropods;
(2) after duplication of the *Hh* gene in vertebrates, its paralogs underwent adaptive evolution;
In protein Hh, adaptive evolution is confined to the intein domain responsible for self-excision.





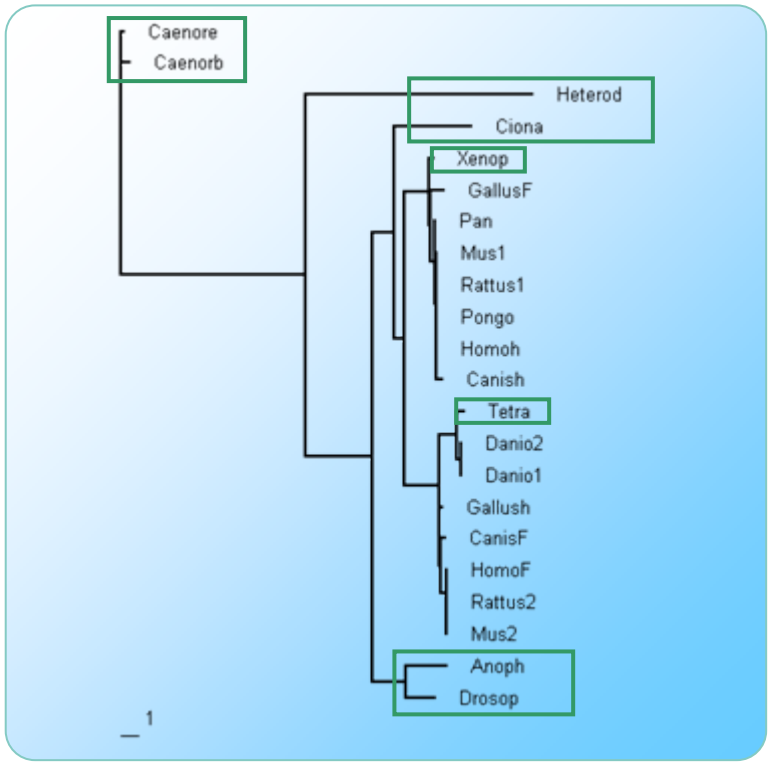
Adaptive evolution of the *Tout-velu* (*Ttv*), gene encoding nonspecific acetyl glucose aminotransferase, which modifies the Hh morphogen

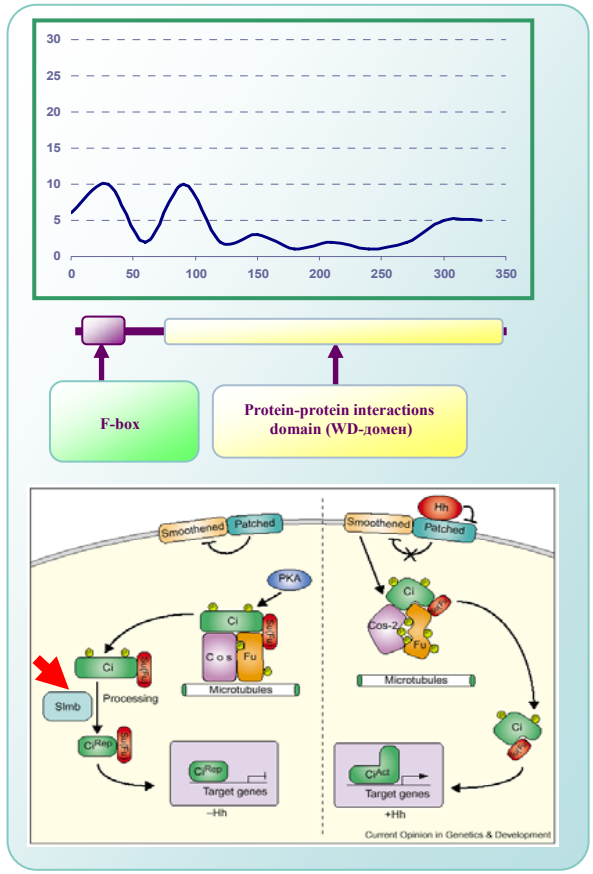


It has been revealed that the adaptive evolution of the *Ttv* gene doest not correlate with the emergence of large taxa of Bilateria, but does with the emergence of new prtein families.

It has been demonstrated that, as a new Ttv protein family emerges, adaptive evolution is under way across the functional protein domain entirely.

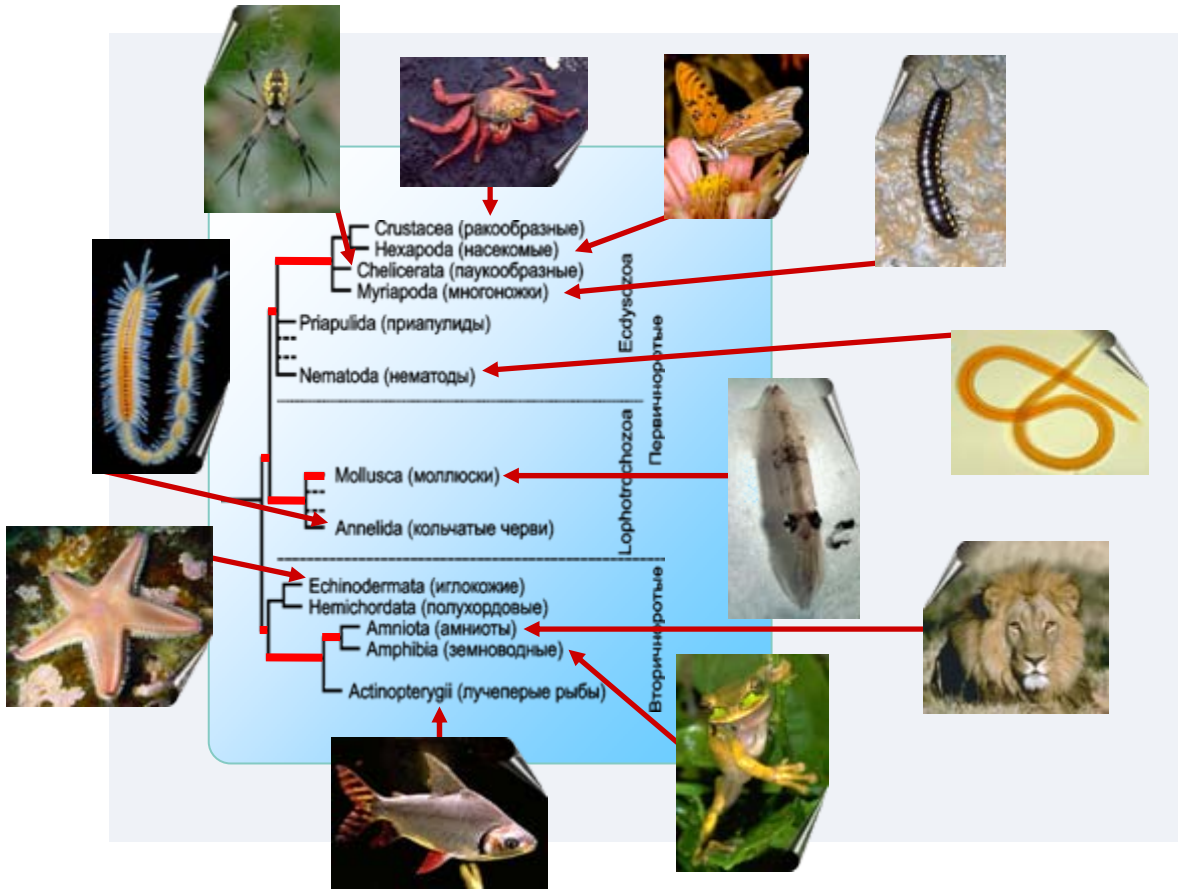
Adaptive evolution of the *Supernumerary limbs* (*Slmb*) gene encoding common ubiquitin-ligase, which initiates proteolysis of protein CI





- Adaptive evolution of the *Slimb* gene not revealed.
- Adaptively evolving domains of Slimb protein not revealed.

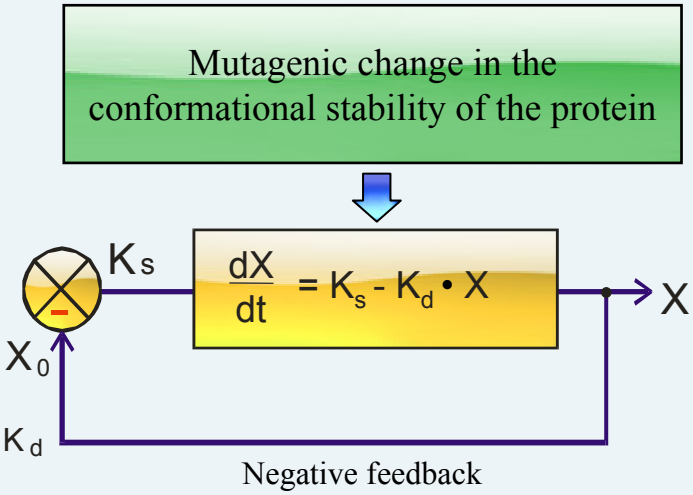
Projection of the events of adaptive evolution of Hh-signal cascade components on the phylogenetic tree of many-celled animals

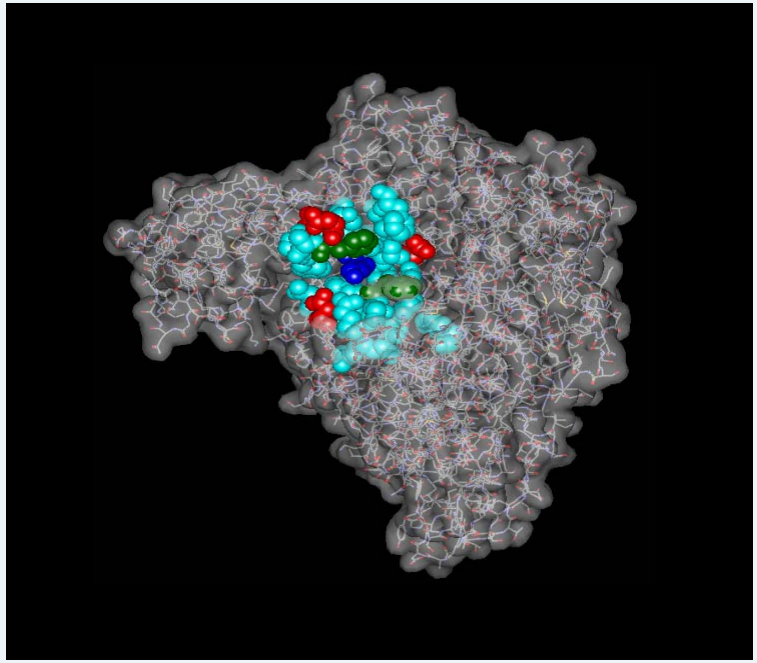


7.3. Theoretical modeling of evolution

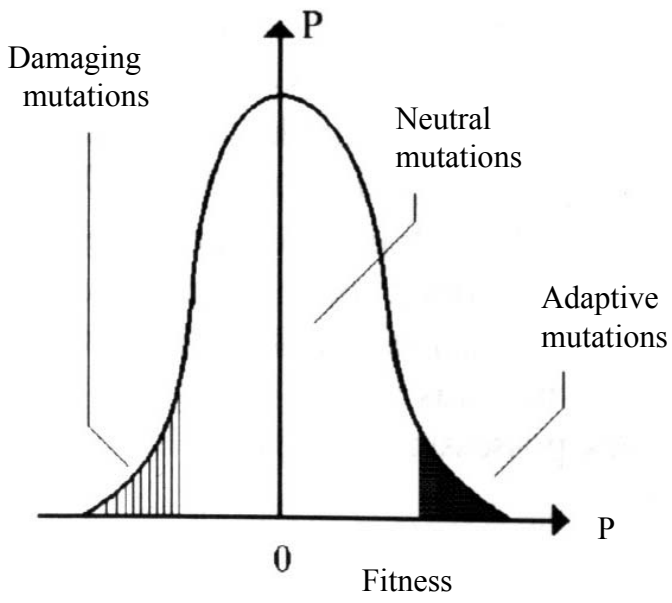
7.3.1 Regulatory circuits and evolution

The simplest feedback-enabled regulatory circuit controlling protein concentration in the cell

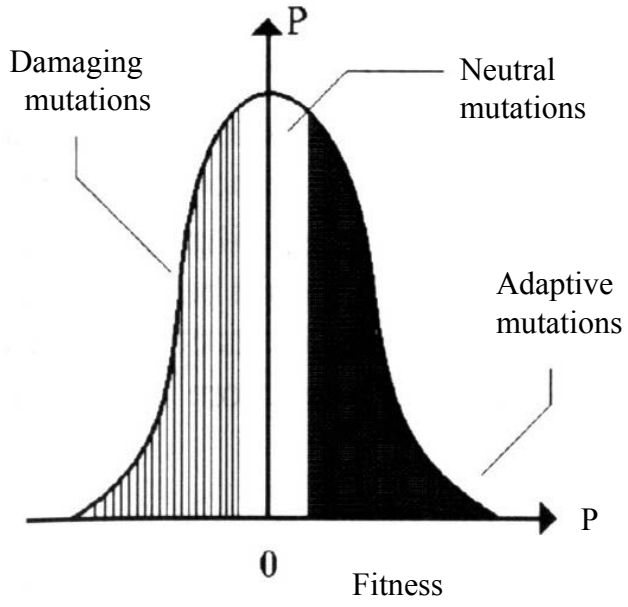




An illustration of how the mutation spectrum is "neutralized"



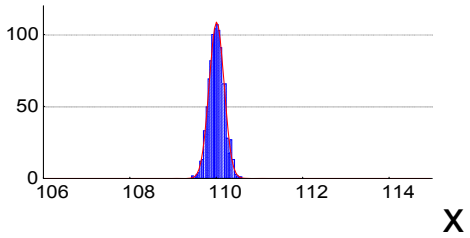
(a) Negative feedback-enabled system



b) Negative feedback-disabled system

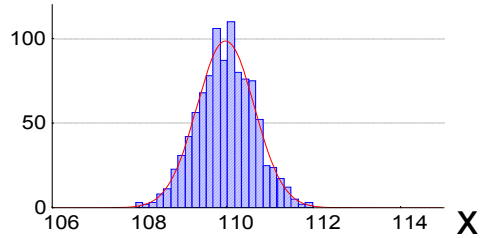
"Neutralization" of the mutation spectrum by negative feedback

As negative feedback grows stronger, the observed level of populational phenotypic variability reduces



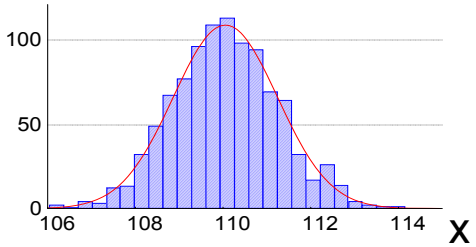
kf=1

Very strong negative feedback



kf=0.1

Moderate negative feedback



kf=0.03

Very weak negative feedback

$$W(X_i) = \sqrt{1/2\pi}\, e^{-\frac{1}{2}\left(\frac{X_i - X_0}{\sigma_X}\right)^2}$$

$$X_i = -\frac{C_2}{2} + \sqrt{\frac{C_2^2}{4} + \frac{C_5 C_2}{C_3}(E_i + C_4)}$$

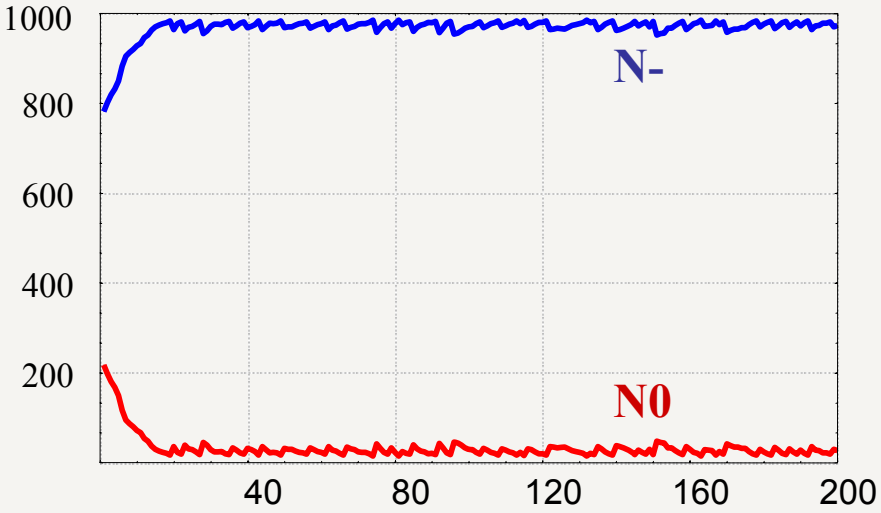
$$P(E_i) = \sqrt{1/2\pi}\, e^{-\frac{1}{2}\left(\frac{E_i - E_0}{\sigma E}\right)^2}$$

$$kf = 1/C2$$

A compensatory effect of negative feedback

- Any negative feedback minimizes (masks) the phenotypic manifestation of mutations by changing, a in a compensatory manner, the intensity of the processes it regulates;
- Negative feedbacks neutralize mutation spectra: most otherwise adaptive or damaging mutations become neutral in the presence of regulatory circuits;
- The compensatory effect of negative feedbacks is one of the main factors, due to which most mutations fixed in the course of evolution are neutral.

Competition between negative feedback-enabled (N-) and negative feedback-disabled (N0) individuals during population evolution under stabilizing selection



Population size N=const

Initial conditions:
50% individuals re negative feedback-enabled;
50% individuals re negative feedback-disabled.

Stabilizing selection,
X_0 – optimal protein concentration = const

$$P(E_i)=\sqrt{1/2\pi}\,e^{-\frac{1}{2}\left(\frac{E_i-E_0}{\sigma E}\right)^2}$$

Xi – protein concentration with modified stability DE_i

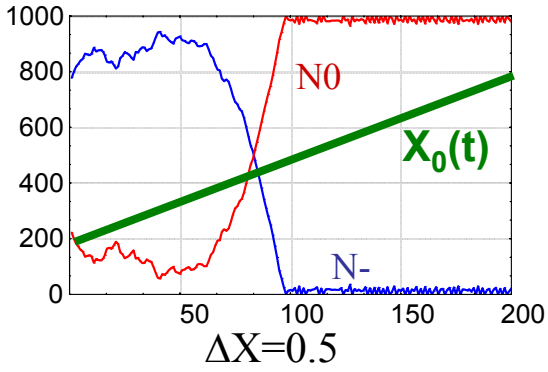
$$W(X_i)=\sqrt{1/2\pi}\,e^{-\frac{1}{2}\left(\frac{X_i-X_0}{\sigma_X}\right)^2}$$

Competition between negative feedback-enabled (N-) and negative feedback-disabled (N0) individuals during population evolution under driving selection

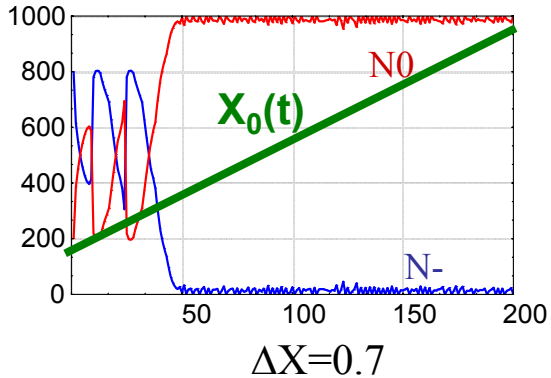
$$P(\Delta E) = \sqrt{1/2\pi}\, e^{-\frac{1}{2}\left(\frac{\Delta E}{\sigma_E}\right)^2}$$

$$W(X) = \sqrt{1/2\pi}\, e^{-\frac{1}{2}\left(\frac{X - X_0}{\sigma_X}\right)^2}$$

$$X_0(t_{k+1}) = X(t_k) + \Delta X$$



ΔX=0.5

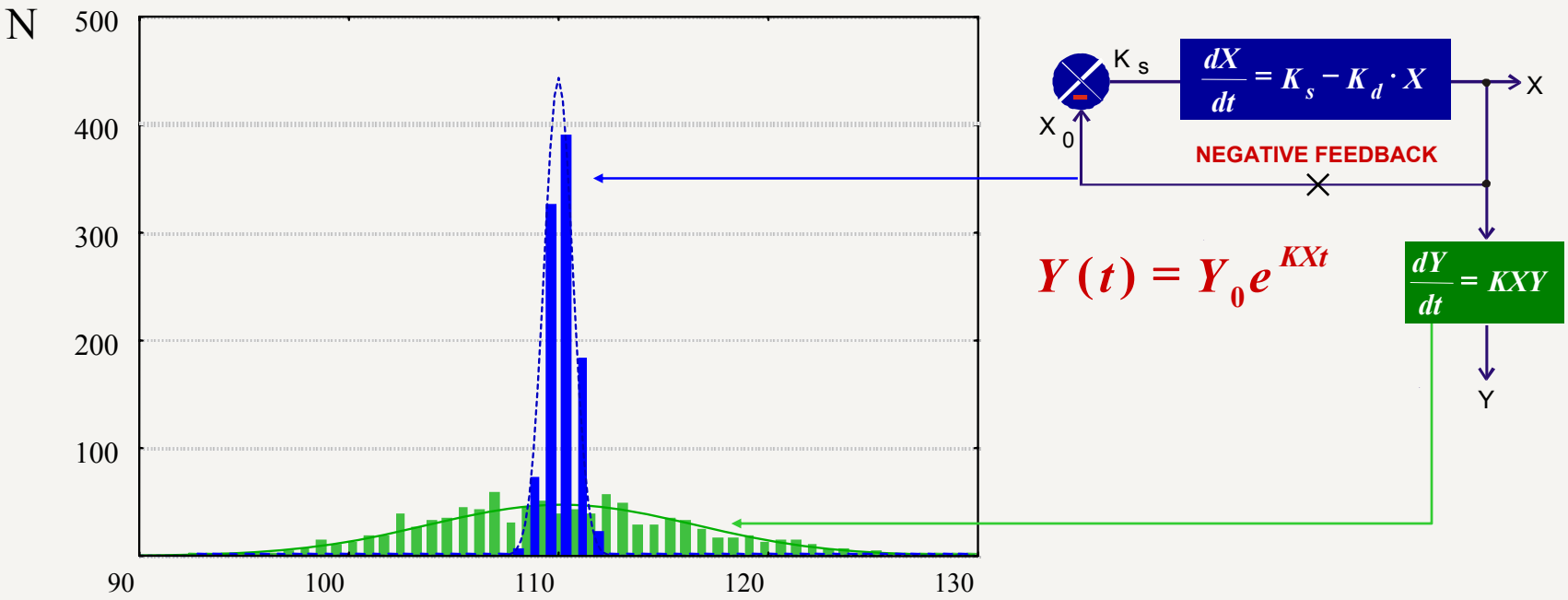


ΔX=0.7

Population size N=const

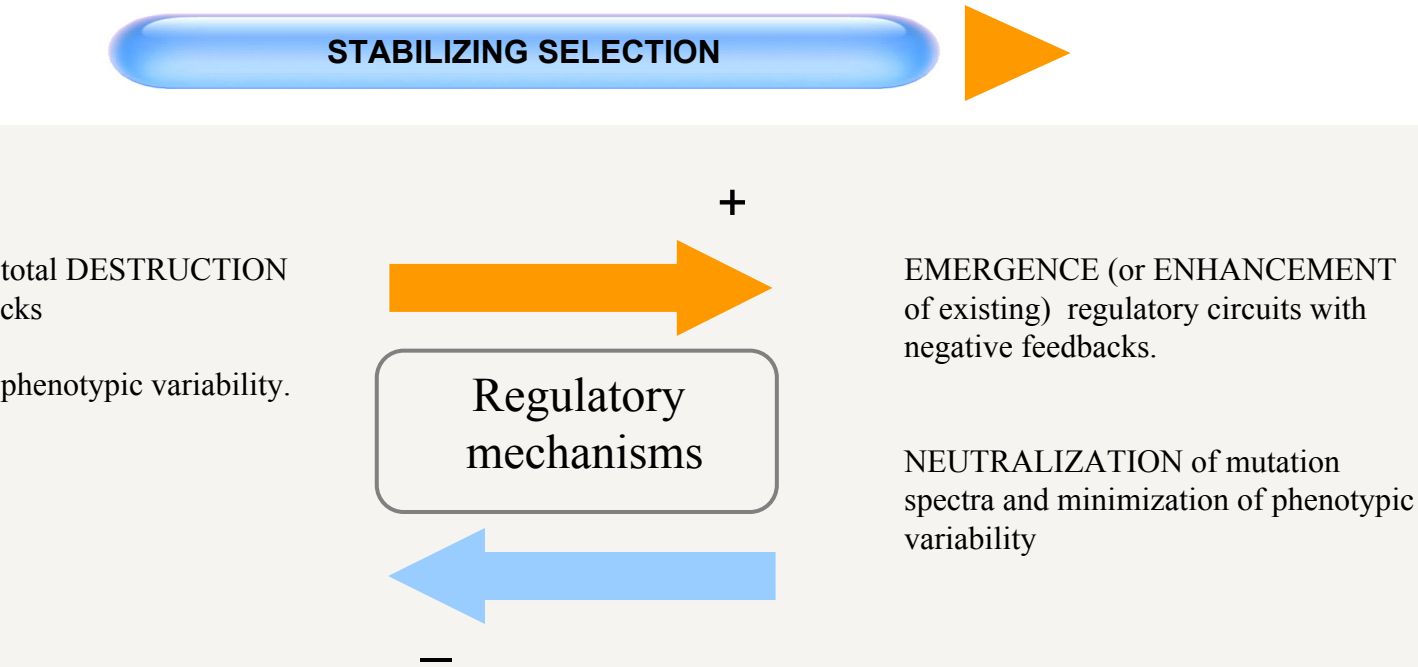
Optimal protein concentration X_0 changes, at each stage of evolution, by DX; $X_{i+1}=X_i+DX$

Hypermanifestation of the mutation spectrum in a hierarchically organized biological system



- Hierarchically subordinate parameter Y depends exponentially on higher-rank parameter X.
- Consequently, mutational changes of parameter X are exponentially enhanced at the hierarchically subordinate level and cause changes to parameter Y to be stronger expressed.
- This phenomenon is known as hypermanifestation of the mutation spectrum of the hierarchically subordinate relative to the mutation spectrum of higher-ranked parameter X.

Evolutionary swinging: alternation between stabilizing and driving selection



7.3. Theoretical modeling of evolution

7.3.2 Evolutionary Constructor is a software program for simulating the evolution of interacting populations by mutation and horizontal transfer

"Evolutionary constructor" is useful for

- Modeling the emergence and effects of mutations
- Modeling environmental effects on populations
- Modeling the co-evolution of populations
- Modeling the horizontal transfer of genetic material

Program architecture

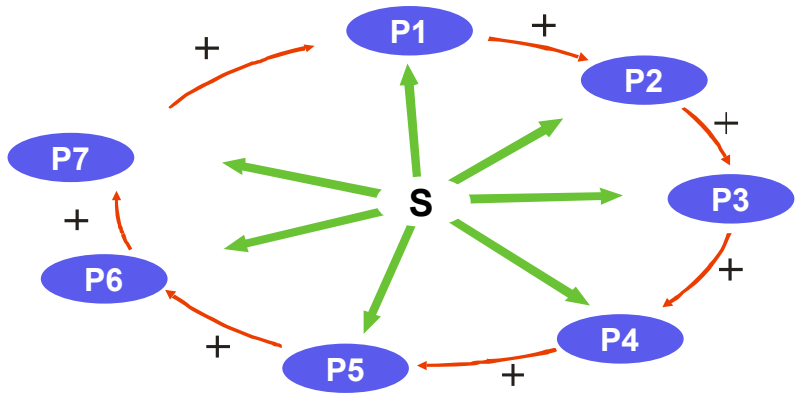
- Main objects
  Populations
  - Individuals
    - Genome
    - Metabolites and substrates consumed
    - Phenotype
- Environment
  - In-channel substrates
  - Other physical and chemical factors

- Main processes
  - Channel in the environment
  - Substrate exchange between populations and the environment
  - Calculating individual fitness
  - Change in population size

"Altruistic/selfish mutations" model

Each individual is characterized by three vectors:

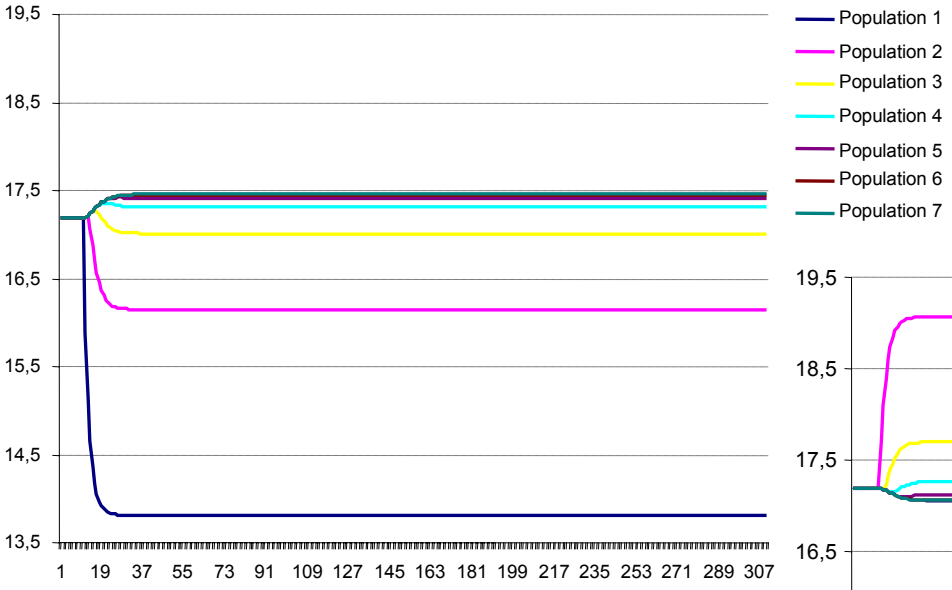
$C=(c_0,c_1,\ldots,c_N)$ – substrate consumption efficiency constant
$D=(d_0,d_1,\ldots,d_N)$ – product synthesis efficiency constant
$S=(S_0,S_1,\ldots,S_N)$ – concentration of substrates consumed
Population growth:



P – population size

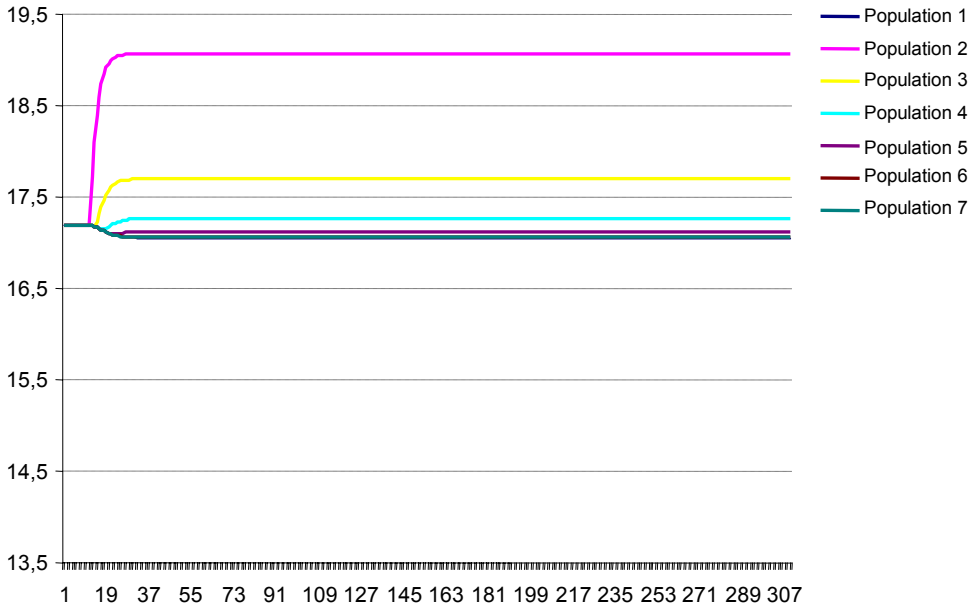
$$F(S,C,P)=\sqrt{c_0 s_0 \cdot \sum_{i=1}^{N} c_i s_i} - k_{death} \cdot P^2$$

Results of a series of numerical experiments “Altruistic mutations”



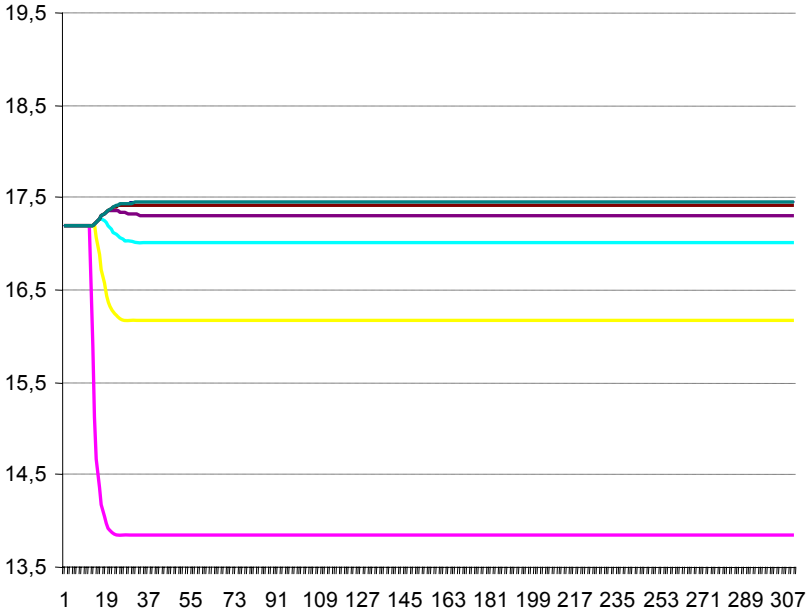
Population 1 increases product synthesis by 50%

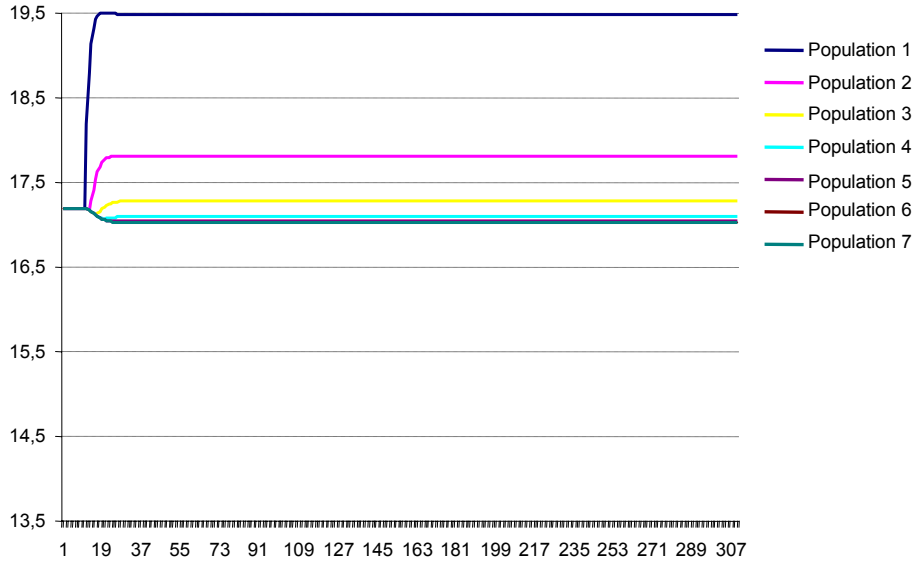
Population 1 reduces specific substrate consumption efficiency by 50%



Results of a series of numerical experiments “Selfish mutations”

Population 1 increases
specific substrate consumption
efficiency by 50% 50%





Population 1 reduces product
synthesis by 50%

Conclusions

- "Altruistic" mutations

  - Produce a positive effect on total population size
  - Produce a negative effect on the mutant population

- "Selfish" mutations
  - Produce a negative effect on total population size
  - Produce a positive effect on the mutant population

- Native populations behave in a more consolidated manner when affected by mutations that have effects on substrate consumption efficiency than when affected by mutation tha have effects on product synthesis
- The "trophic ring" system is quite resistible to mutations affecting substrate consumption efficiency and product synthesis

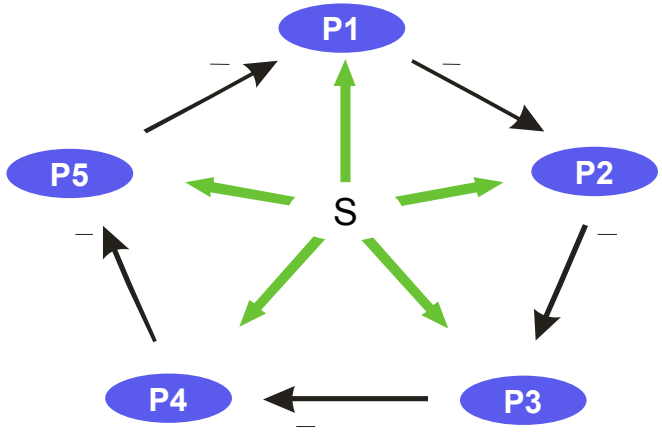
“Inhibiting populations” model

Individual:

$C=(c_0,c_1,\ldots,c_N)$ – substrate consumption efficiency constants

$D=(d_0,d_1,\ldots,d_N)$ – product synthesis efficiency constants

$S=(S_0,S_1,\ldots,S_N)$ – concentration of substrates consumed

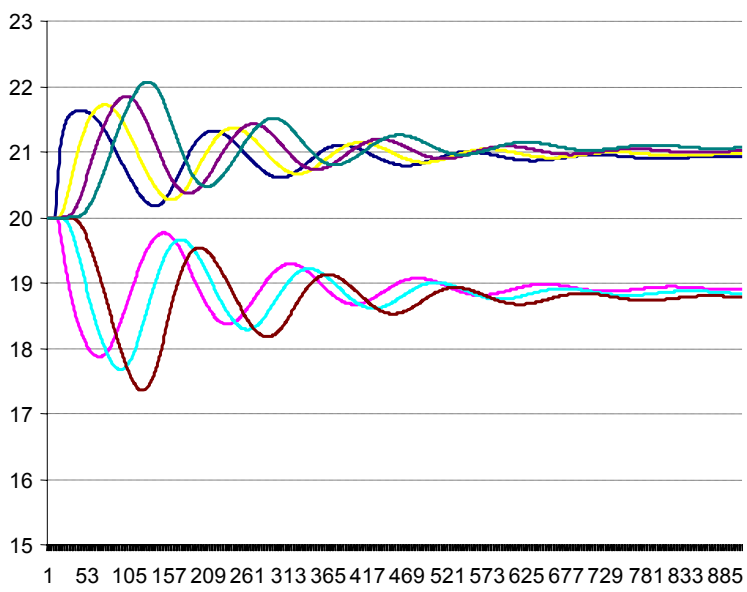


Population growth:

P – population size

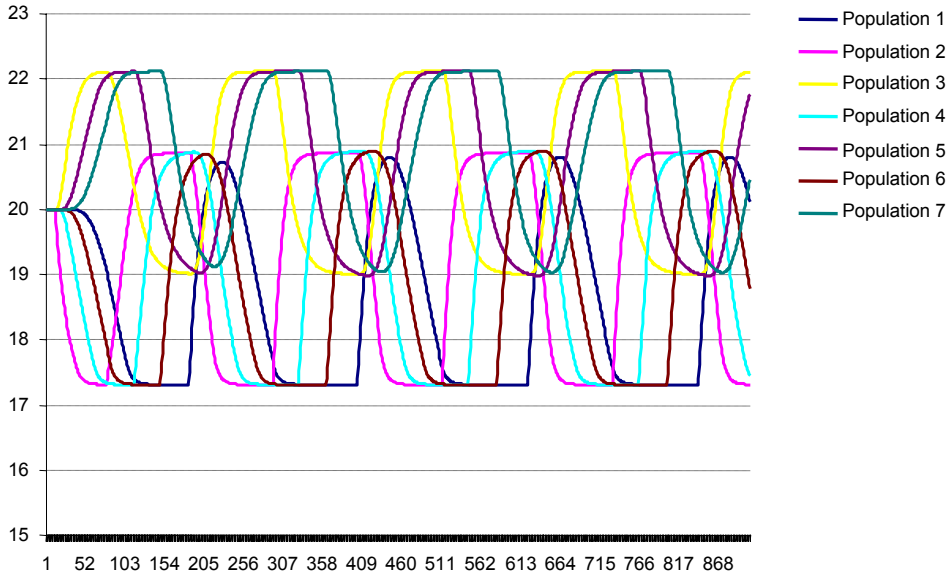
$$F_1(S,C,P)=a_{basal}\cdot P-\sqrt{\sum_{i=1}^{N}c_i s_i}-k_{death}\cdot P^2$$

Results of a series of numerical experiments “Inhibiting populations”



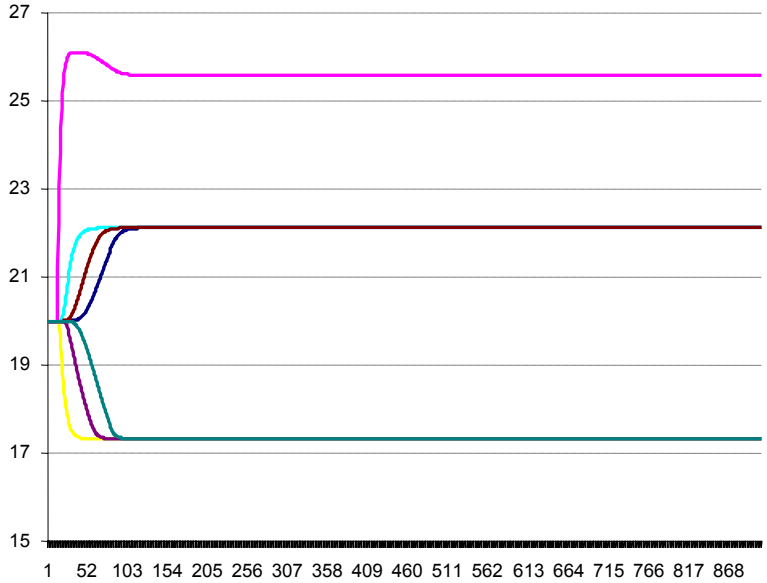
Population 1 increases product synthesis by 10%

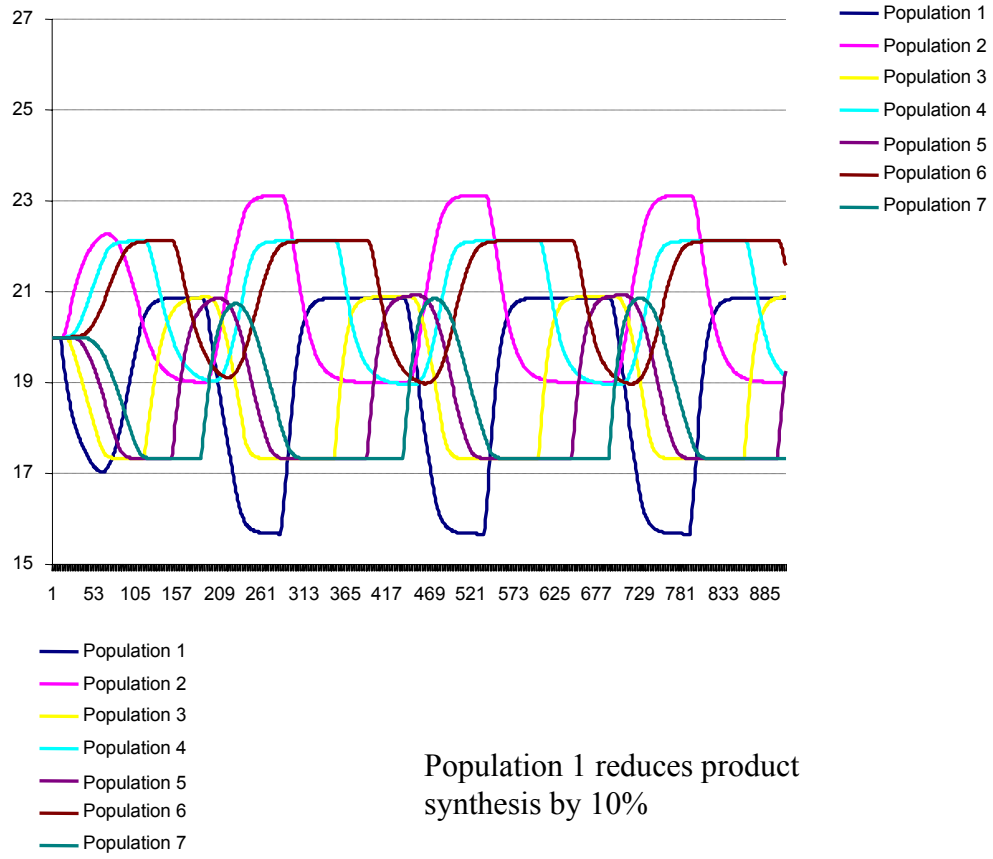
Population 1 reduces specific substrate consumption efficiency by 10%



Results of a series of numerical experiments “Inhibiting populations”

Population 1 increases specific substrate consumption efficiency by 10%



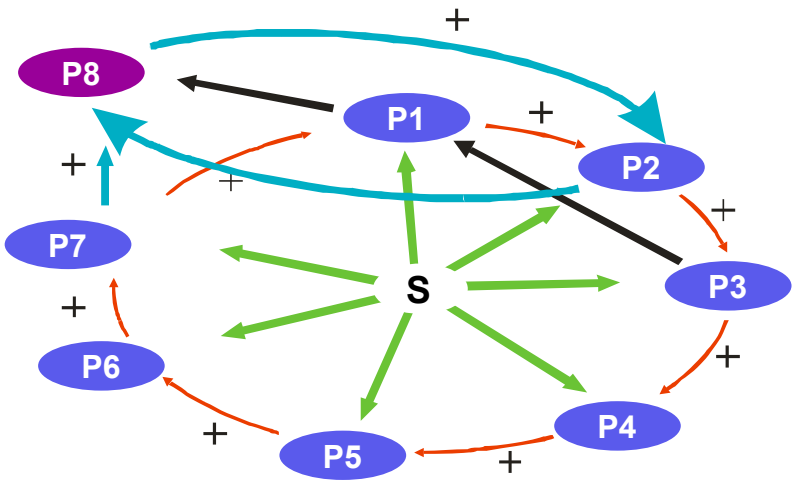


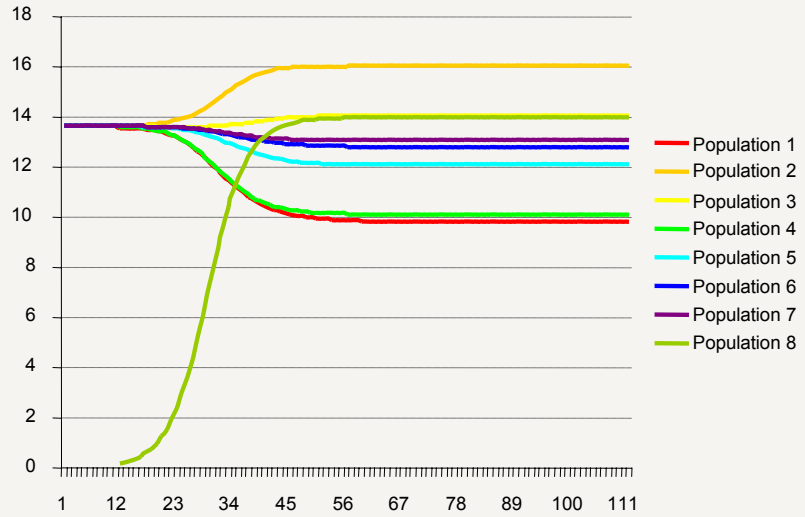
Population 1 reduces product synthesis by 10%

“Inhibiting populations” Conclusions

- The main difference is that, at particular values of the parameters, cyclic modes appear
- The populations are clearly clustered

Modeling horizontal transfer in a stepwise activation system.

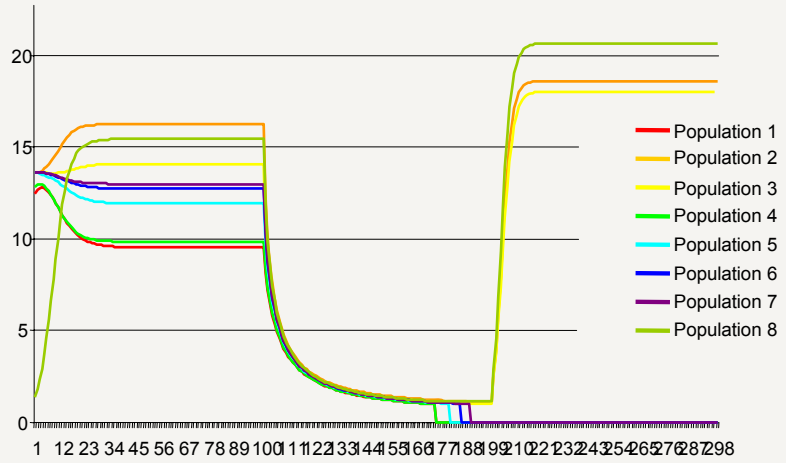




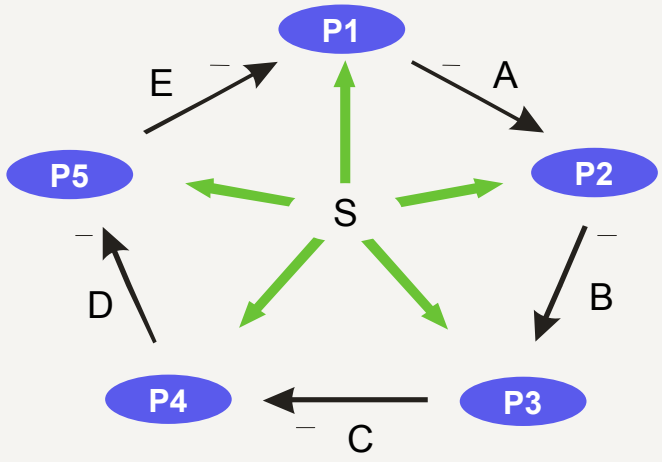
If the external conditions do not changes, population 8 reaches a plateau.

The gene responsible for consumption of specific substrate produced by population 2is transferred from population 3 to population 1. This gives birth to population 8, which can consume specific substrates produced by populations 2 and 7.

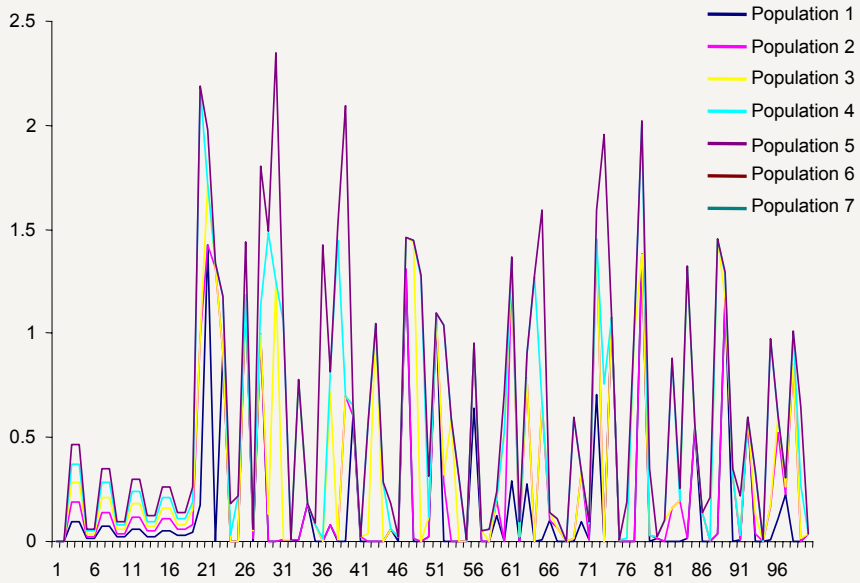
When the supply of non-specific substrate is temporarily curtailed, the fragment of the “ring”, which is light-green in the plot, dies out. Upon resumption of substrate supply at previous rates, only the “subring”, which resulted from horizontal transfer, survives



Modeling the evolution of a system of populations with their one-by-one inhibition in the ring: the pattern of evolution claims chaos.



Free-for-all in a ring-like system leads to chaos



Each population is affected by the closest neighbor's inhibitory action twice: at time t and at time t+dt.

Prospects

- Models will be genetically more complex:
  - Modeling polymorphism of the genes responsible for consumption and synthesis of specific substrates/products
  - Modeling polymorphism of the genes responsible for consumption of non-specific substrate
  - Modeling the processes of horizontal transfer of genetic material
  - Modeling competition between two or more “trophic rings”
  - Modeling of co-processes of inhibition/activation
  - Modeling the evolution of the “hypercycles” of genetic macromolecules
- Software developing
- Distributed modeling support

  Writing a language for evolutionary scenarios

List of publications

Ivanisenko V.A., Pintus S.S., Krestyanova M.A., Demenkov P.S., Znobisheva E.K., Ivanov E.E., Grigorovich D.A. PDBSITE, PDBLIGAND and PDBSITESCAN: a computational workbench for the recognition of the structural and functional determinants in protein tertiary structures combined with protein draft docking. Proc. of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure. 2004. V.1. P. 269-273.

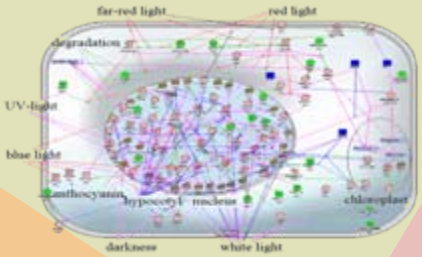
Pintus S.S., Ivanisenko V.A. A molecular mechanism for the structure-functional alterations in mutant forms of human p53 protein. Proc. of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure. 2004. V.1. P. 338-342.

Afonnikov D.A., Oshchepkov D.Y., Kolchanov N.A. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. Bioinformatics, 2001, 17(11):1035-46.

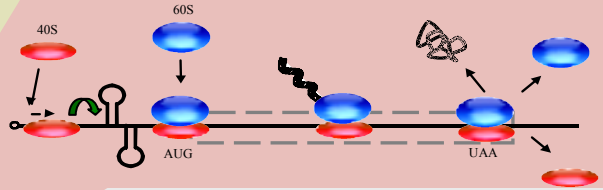
Matushkin Yu.G., Morozova I.N., Morozov P.S. Theoretical analysis of mutation spectra of cytochrome P450 superfamily. Mol Biol (Mosk), 1999, 33(4):696-9.

Shindyalov I.N., Kolchanov N.A., Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 1994, 7(3):349-58.

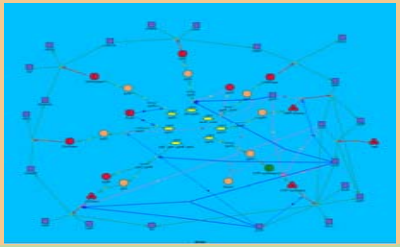
Transgenesis: designing transgens and transgenomes



Modeling molecular-genetic systems and processes

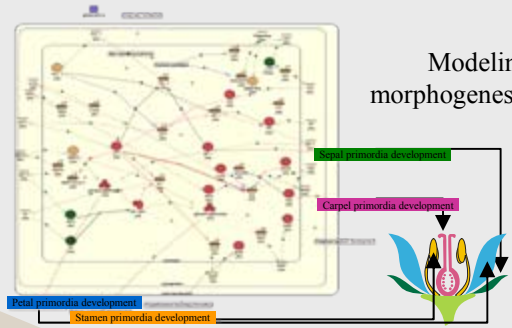


Biotechnology: designing superproducers

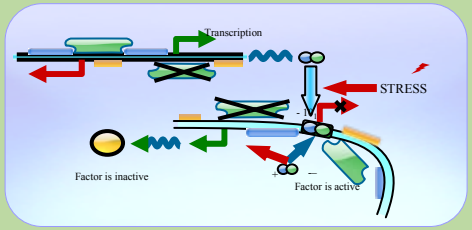


Computational systems biology: fundamental and applied aspects

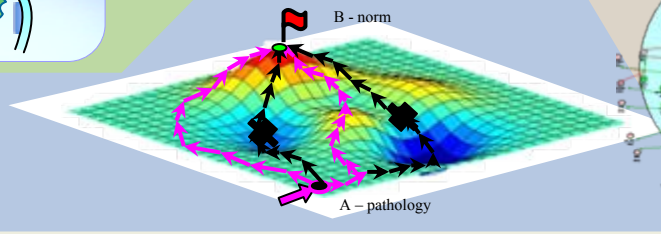
Modeling morphogenesis



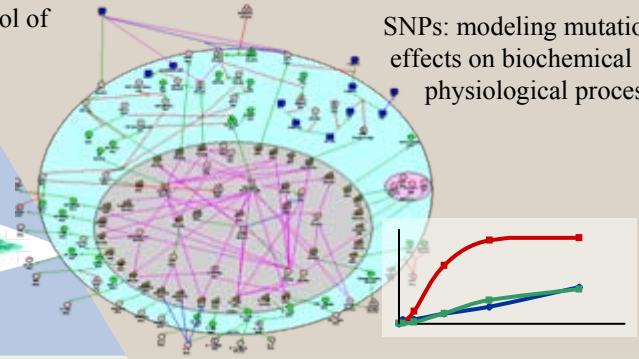
Computer-assisted design of genosensors



Modeling optimal pharmacological control of gene networks



SNPs: modeling mutational effects on biochemical and physiological processes



Conclusions

Thank you for attention granted to the work in the Laboratory of Theoretical Genetics of the IG&G, SB RAS.

We hope that this presentation has given you the idea about the fast-paced research activity in the Laboratory of Theoretical Genetics of the IG&G, SB RAS.

The Laboratory is open for contacts and seek for cooperation in all aspects highlighted in this presentation.

The conference, BGRS'2006, which the Laboratory will be holding in Novosibirsk, Russia, from July 16 to July 22, is an excellent opportunity for identification and discussion of prospects for cooperation.

The Organizers to BGRS'2006 can be contacted by e-mail: bgrs2006@bionet.nsc.ru

The Conference site is accessible via the Internet: http://www.bionet.nsc.ru/meeting/bgrs2006/

Welcome to Novosibirsk!