Table of Contents

Introduction. About GeneNetWorks™ 4

Part I. DNA Integration Level 6

Chapter 1. Transcription Regulatory Regions Database (TRRD) 6
Example 1. How to view the map of regulatory regions of a gene? 12
Example 2. Access to TRRD database 13
Example 3. Access to functionally important gene systems (TRRD sections) 13
Example 4. Searching for the genes of a particular species 14
Example 5. Searching for the genes by name 16
Example 6. Searching for the genes by name and species name 18
Example 7. Searching for the genes with particular chromosomal localisation 20
Example 8. Searching for the genes by KeyWords 21
Example 9. Search for the liver-specific hypersensitive to DNase I sites 23
Example 10. Search for the transcription factors binding sites by names 25
Example 11. How to search for a site by its sequence? 27
Example 12. How to search for sites by the factor name and the factor influence? 28
Example 13. How to search genes by inducer or repressor name? 29
Example 14. How to search expression patterns from genes that are expressed in particular organ? 29

Chapter 2. Site Recognition Module 31
1. B-DNA Site Video 31
1.1. B-DNA Site Video Databases 31
Example of SRS queries to the B-DNA Site Video Databases: 34
1.2. Software 35
1.2.1. bDNA Profiles DNA. Feature-Recognition Tools 35
Release 2003 35
Example 37
1.2.2. DNA Property Plot 38
Example 39

Chapter 3. SELEX System 41
1. SELEX Knowledge Base 41
SELEX_DB 42
SELEX_BIB 42
2. SELEX Profiles Program 43
Example 45

Chapter 4. RegScan Module For DNA Functional Site Recognition 46
1. RegScan Databases 46
Example of query: 49
2. Software For Transcription Factor Binding Site Recognition 51
2.1. Program BinomSite 51
Example 52
2.2. Program MMSite 53
Example: 55
2.3. Programs for DNA Functional Sites Recognition 57
Example: Recognising of USF binding site in the sequence of interest 58
2.4. RGSiteScan Program 60
Example 61

Example....61
2.5. RecGroup program....62
Example....63
2.6. ARGO system....63
Example....66
3. Programs for Promoter Recognition....68
3.1. Program to recognize eukaryotic promoters....68
Example....69
3.2. ARGO-Viewer....70
3.3. POLIIISCAN....74
4. Other Programs....76
4.1. NASCA program....76
Example....78
4.2. Program for estimation of stochastic complexity of genetic texts....80
Example....83

Chapter 5. ACTIVITY System....85
1. ACTIVITY Database....85
Examples of SRS queries to the ACTIVITY database....88
2. Predicting Activities of Functional Sites in DNA/RNA....91
Example....92

Chapter 6. DNA Nucleosomal Organisation....94
1. PROFILES Databases....94
Example of SRS queries to the PROFILE database....95
2. Software....99
2.1. Nucleosome binding site recognition....99
Example....100
2.2. Recognition tools....101
Example....102

Part II. RNA Integration Level. 104

Chapter 1. Leader RNA. 104
1. LEADER_RNA Knowledge Base. 104
2. Software. 107
2.1. Programs for mRNA Translatability Prediction....107
Examples 109
2.2. MatrixSS: Building of E-score plot for RNA sequence 113
Example....114
2.3. GArna Program....115
Example....116

Part III. Protein Integration Level....118

Chapter 1. Databases....118
1. EnPDB 118
Example 1....121
Example 2....124
2. PDBSite Database....127
Example 1....129
Example 2....131
Example 3....133

Example 4 134
3. PDBSiteScan 136
Example 137
4. Artificial Selected Peptides/Proteins Database (ASPD) 138
5. DCS - Database on residue coordination spheres 140
Example 1 142
Example 2 145
6. Database on local conformations of protein chains -"conformons" (ConfDB) ... 148
Chapter 2. Software 151
1. CRASP 151
1.1. Analysis of pairwise positional correlations 151
Example: 155
1.2. Analysis of protein integral physico-chemical characteristics 155
Example: 160
Part IV. Gene Networks Integration Level 161
Chapter 1. GeneNet database 161
Example 1 165
Example 2 168
Chapter 2. GeneNet Software 170
1. GeneNet Viewer. 170
Example 171
2. GeneNet Modelling 175
Example 1 177
Example 2 182
Example 3 183
3. Mathematical models of gene networks in SBML format 184
3.1 About mathematical models of gene networks in SBML format 184
3.2 List of mathematical models in SBML format of gene networks represented in GeneNet database 184

INTRODUCTION. ABOUT GENENETWORKS™

Integrated system **GeneNetWorks™** is designed for accumulation of experimental data, data navigation, data analysis, and analysis of dependencies in the field of gene expression regulation. It integrates the databases and programs for processing the data about structure and function of DNA, RNA, and proteins, together with the other information resources important for gene expression description. The unique property of above described system is that all the resources within the system **GeneNetWorks™** are divided according to the natural hierarchy of molecular genetic systems and has the following levels:

1) DNA;
2) RNA;
3) proteins; and
4) gene networks.

Each module contains:

1) experimental data represented as a database or some sample;
2) program for data analysis;
3) results of an automated data processing;
4) tools for the graphical representation of these data and the results of the data analyses.

The database system **GeneNetWorks™** is the product developed at the Laboratory of Theoretical Genetics of the Institute of Cytology and Genetics of Siberian Branch of Russian Academy of Science in Novosibirsk (Russia). Institute of Cytology and Genetics since early 80's has been developing the database system, which trade name now is **GeneNetWorks™**. The using of the Institute name in advertising and commercial purpose is allowed only after written agreement with the Institute.

Institute of Cytology and Genetics was founded in 1957 and has large experience of scientific study in fields of genetic, cytology, microbiology etc., is involved in couple of international research projects. Since early nineties, the Institute is active in commercial projects, had organized few joint ventures with commercial structures. Now it has organized internal structures for performing of commercial projects, which includes leaders of the groups, project managers and executors of project. At present time, the experience is enough to be involved into international cooperation and performance of international commercial projects.

The tasks that could be solved by the **GeneNetWorks™ system (GNW)**:

Task	Tool	Section
What is known about the regulation of the gene XX?	**TRRD, TRRD-Viewer**	Part I. Chapter 1.
Which known genes are expressed in the tissue NN, or only at embryonic stage of development, or only during the S-phase of the cell cycle?	**TRRD**	Part I. Chapter 1
Which structural and functional characteristics are typical for transcription factor XX binding site?	**TRRD, ACTIVITY**	Part I. Chapters 1, 5
Which transcription factors might bind to the sequence XX?	**TRRD, B-DNA Site Video, RegScan Programs**	Part I. Chapters 1, 2.1, 4.2

Which transcription factors might regulate gene XX (where no information is available in TRRD)?	**B-DNA Site Video, RegScan Programs**	Part I. Chapters 2.1, 4.2
Which conformational or physico-chemical properties are characteristic for the sequence XX and which functional DNA sites have the same properties?	**B-DNA Site Video, RegScan Programs, NASCA**	Part I. Chapters 2.1, 4.2, 4.4.1
Which nucleotides in the sequence of the functional DNA or RNA site are important for binding of a protein?	**TRRD, SELEX**	Part I. Chapters 1, 3
Whether the sequence XX might be a promoter?	**RegScan Programs**	Part I. Chapter 4.2
Which regions of the sequence XX might be bound to a nucleosome?	**Nucleosome Site Recognition**	Part I. Chapter 6.2.1
Genes X, Y, and Z are coregulated in an experiment E (but the genes A, B, and C are not). Which sequences might be involved in the regulation?	**RegScan Programs**	Part I. Chapter 4.2
Homology to which known extended regulatory region is typical for the sequence XX?	**TRRD, RegScan Programs**	Part I. Chapters 1, 4.2
How strongly might the transcription factors bind the sequence XX?	**ACTIVITY**	Part I. Chapter 5
How to evaluate translation efficiency of RNA sequence XX?	**LEADER_RNA**	Part II. Chapter 1
Is RNA XX highly expressed?	**LEADER_RNA, GArna**	Part II. Chapters 1, 2.3
How can I use **GNW** to predict the secondary structure of RNA?	**GArna, MatrixSS Programs**	Part II. Chapters 2.3, 2.2
From published Phage display experiments what are the interacting residues?	**ASPD**	Part III. Chapter 1.3
How do I use **GNW** to find sequences of proteins and peptides selected by phage display?	**ASPD**	Part III. Chapter 1.3
How can I use **GNW** to search for proteins according the data on protein function and structural characteristics of protein active sites?	**PDBSite, EnPDB**	Part III. Chapters 1.2, 1.1
How can I use **GNW** to search for proteins based on the data about sites subjected to biochemical modifications?	**PDBSite**	Part III. Chapter 1.2
Which physico-chemical properties of the protein XX are conserved and which are variable?	**CRASP**	Part III. Chapter 2.1
How the substitution of amino acid in a protein sequence might influence chemical and physical protein characteristics?	**CRASP**	Part III. Chapter 2.1
To which extent are independent point mutations that cause amino acid substitutions in a protein sequence?	**CRASP**	Part III. Chapter 2.1
Which proteins, genes and non-proteinaceous substances support the functioning of the gene network XX?	**GeneNet Database**	Part IV. Chapter 1
In which processes participates the protein XX?	**GeneNet Database, GeneNet Viewer**	Part IV. Chapters 1, 2.1
How to display the general visualisation of structure-functional organisation of a gene network described in **GNW**?	**GeneNet Viewer**	Part IV. Chapter 2.1
Which models are available in **GNW**? How can models be changed?	**GeneNet Modelling**	Part IV. Chapter 2.2

For questions, please contact scientific supervisor of the project: Prof. N. A. Kolchanov, telephone: +7-3832-333468; e-mail: kol@bionet.nsc.ru

PART I. DNA INTEGRATION LEVEL

CHAPTER 1. TRANSCRIPTION REGULATORY REGIONS DATABASE (TRRD)

Release 2003

Database description:

Transcription **R**egulatory **R**egions **D**atabase (**TRRD**) is a unique information resource, accumulating information on structural and functional organisation of transcription regulatory regions of eukaryotic genes. Only experimental information is included into **TRRD**.

Access to **TRRD**:

http://www.domain.com/mgs/gnw/trrd/

Database content:

SRS table	Description	Number of entries
TRRDGENES	General information about genes	1862
TRRDUNITS4	Descriptions of regulatory units (promoters, enhancers, silencers)	2832
TRRDEXPR	Gene expression patterns	10575
TRRDSITES4	Descriptions of transcription factor binding sites	8326
TRRDFACTORS4	Descriptions of transcription factors	6706
TRRDLCR	Descriptions of Locus Control Regions	14
TRRDBIB4	References	6233

List of biological tasks that could be solved by using **TRRD**:

- to extract the list of genes by gene name and/or by species name;
- to view the map of transcription regulatory regions reconstructed by **TRRD-Viewer**
- to obtain patterns of gene expression regulation;
- to obtain data on transcription regulatory regions and transcription regulatory units involved in regulation of expression of a particular gene;
- to obtain data on structural and functional characteristics of transcription factor binding sites;
- to obtain descriptions of transcription factors used in the experiments on either binding capacities or functional activities of the corresponding sites;
- to extract references to original publications containing data listed above;
- to compare novel DNA sequences with regulatory units stored in **TRRD** using BLAST tool and find sequence alignments of sequences.

SRS tables format:

TRRDGENES

Line code	Field name	Field description
AC	GeneAC	TRRD gene accession number
ID	GeneID	TRRD identifier
DT	Updated	Date of the last update
GN	TransfacLink	Link to the TRANSFAC database
OS	Species	English and Latin names of species
NA	GeneName	Full, short gene names and synonyms
SN	GeneName_Brief	Short gene name
NG	GeneName_Full	Full gene name
SY	GeneSynonym	Synonyms of short or full gene names
EC	EnzymeClass	Enzyme classification
KW	KeyWords	Key words
CH	Chromosome	Chromosomal localisation
RG	RegRegion	Regulatory region
AP	RegUnitAC	Name and localisation of the regulatory unit; site links
PR	RegUnit	Name and localisation of the regulatory unit; site links
AG	ExperimentCodes	Cells, assay codes, reference to the paper
CE	CompElement	Composite element description
AL	Alignment	Alignment of extended regulatory regions
MP	StartPoints	Distance between start points
BI	DNABankLink	Link to EMBL/GenBank
CC	Comments	Comments
HN	HSS_AC	TRRD accession number of the hypersensitive site
HS	HSS_Position	Name of the hypersensitive site and its location
HD	HSS_Descript	Functional characteristics of the hypersensitive site
HR	HSS_Reference	Bibliographic reference
HG	HSS_Genes	Genes regulated by the LCR
HC	HSS_Comments	Comments on hypersensitive site
DR	BankLink	Links to other databases

TRRDUNITS

Line code	Field name	Field description
ID	RegUnitAC	Accession number of a regulatory unit
GN	GeneID	Identifier of an entry in the TRRD database
RG	RegRegion	Transcription regulatory region
PR	RegUnit	The description of promoter, enhancer or silencer (name and location; start point name; the list of all site accession numbers located in this regulatory unit)
AQ	DNA_BankLink	Link to EMBL/GenBank, and the first and last nucleotide positions of the sequence according to the EMBL/GenBank
LQ	LeftTrunc	The left part of the sequences is truncated (nucleotide number)
RQ	RightTrunc	The right part of the sequences is truncated (nucleotide number)
SL	SeqLength	The length of the sequence indicated in the field SE
SE	Sequence	The nucleotide sequence of the unit
PT	PromotTisSp	The information about promoter tissue-specific characteristics
PI	PromotInd	The information about promoter induction pattern
AG	ExperimentCodes	The field includes name of cells (tissue or organ) under experiment, codes of experiments described in the articles, and reference to the original paper
SE	Sequence	DNA sequence

TRRDEXP

Line code	Field name	Field description
RE	ExpressionPatternAC	Pattern identifier
ID	GeneID	Identifier of an entry (gene) in the TRRDGENES4
RT	ExpressionDetectionDevice	Molecular product used to estimate the expression level (protein or mRNA)
RY	CellCycleStage	Cell cycle stage (from G0 to M)
RD	StageOrgDev	Developmental stage (embryo, fetus, etc.)
RO	Organ	Organ
RU	Tissue	Tissue
RN	Cells	Cell type
RF	StageCellDiff	Cell differentiation stage
RL	ExpressionLevel	Gene expression level
RI	IndReprName	Inductor or repressor name
RH	InductionTime	Duration of the inductor's or repressor's effect
FF	Influence	The effect produced by external signal
RP	RegUnitLink	Accession number(s) of the regulatory unit(s) involved in regulation
RS	SiteLink	Accession number(s) of the site(s) involved in regulation
RC	Comments	Comments on the expression pattern
RR	Reference	Reference to the paper
RM	ExpComparison	Comparison of the expression levels from different expression patterns
CC	TextComments	Comments
RX	Sex	Sex of an organism

TRRDSITES4

Line code	Field name	Field description
AN	SiteAC	Site accession number
NM	SiteName	Site name
GID	GeneID	Identifier of an entry (gene) in the TRRDGENES4
NP	PreferredName	Preferred site name
AP	RegUnitAC	Accession number of a regulatory unit which includes this site
AT	FactorInfluence	Alteration of transcription level caused by binding of the factor to the site
NY	SiteNameSynonym	Site name synonym
NS	TransFacSiteReference	TRANSFAC link
WW	PredictionProgramLink	The reference to the Internet-accessible program for recognition of the site
DR	DatabaseReference	Link to the external database
TF	FactorName	A name of the protein or protein complex interacting with the site
SQ	Sequence	Site sequence
NI	SiteIndex	TRRD site index
PQ	SequencePosition	Positions of the site
PF	FootprintSequencePosition	Footprint positions
SC	SeqContradiction	In the case when the annotator finds a discrepancy in the site sequence between the paper annotated and the corresponding data from EMBL/GenBank, the sequence from the paper annotated are given in the field SeqContradiction (SC), while the site sequence corresponding to EMBL/GenBank are indicated in the fields Sequence (SQ).
PC	PosContradiction	In the case when the annotator finds a discrepancy in the site positions between the paper annotated and the corresponding data from EMBL/GenBank, the positions from the paper annotated are given in the field PosContradiction (PC), while the site positions corresponding to EMBL/GenBank are indicated in the fields SequencePosition (PQ) of the database TRRDSITES.
IP	ImportantPos	The nucleotides within transcription factor binding sites that are important for interactions with the corresponding proteins.
KK	TextComments	Comments
EF	ComparativeFactorAffinity	Comments on comparative affinity of binding sites
HM	SiteHomology	Comments on homologous sites in other genes
BF	DNA_BankLink	Position of the site sequence first nucleotide according to the EMBL/GenBank
AG	ExperimentCodes	Cells, assay codes, reference to the paper

TRRDFACTORS

Line code	Field name	Field description
ID	Identifier	Identifier
GI	GeneID	Identifier of an entry (gene) in the TRRDGENES4
AN	SiteAC	Site accession number
TF	FactorName	The abbreviated and full name of the factor

FS	FactorSubunitName	Factor subunit name (for heteromerous proteins)
TY	FactorNameSynonym	Synonymous names
TS	FactorOrigin	Species specificity
NF	TRANSFAC_link	TRANSFAC link
TO	FactorSource	The source of the factor (in vitro synthesised, recombinant, etc.)
TG	Organ	Organs used for isolation of the factor
TE	StageOrgDev	Stage of an organism development
TU	Tissue	Tissues used for isolation of the factor
TC	Cells	Cells used for isolation of the factor
TD	IndReprName	External signal
TR	Reference	Bibliographic reference
CC	Comments	Comments on the protein

TRRDLCR

Line code	Field name	Field description
AC	gene_locus_AC	gene cluster accession number
ID	gene_locus_ID	gene cluster identifier
OS	species	species
OC	Biological_classification	Biological classification (taxonomy)
LO	chromosomal_cytological_location	chromosomal and/or cytological location of the gene cluster
LM	locus_map	The map of the gene cluster
GI	gene_name	Gene name
DR	external_database	Link to SWISS-PROT, TRRD
AL	Locus_Control_Region_AC	Accession number of the Locus Control Region in TRRDLCR
LI	Locus_Control_Region_ID	Locus Control Region identifier in TRRDLCR
TL	LCR_tissue_specificity	Tissue specificity of the LCR
EL	LCR_functional_group	The functional group of LCR elements
BI	EMBL_GenBank_ID_AC	EMBL/GenBank identifier and accession number
AR	functional_group_AC	Accession number of the functional group of LCR elements in LCRTRRD
RF	function	The function of the functional group of LCR elements
NM	Name_of_RE	Name of the regulatory element
NY	Synonim_of_RE_name	Synonym of the regulatory element
RT	funct_group_tissue_specificity	Tissue specificity of the functional group of LCR elements
RE	functional_group_elements	Functional group elements (hypersensitiv sites or regulatory elements)
CC	comments	Comments
HI	HS_RE_ID	Identifier of the hypersensitiv site or regulatory element in TRRDLCR
AP	TRRD_accession_number_of_RE	Accession number of the hypersensitiv site or regulatory element in TRRDLCR
DE	HS_RE_name	Name of the hypersensitiv site or regulatory element
HT	HS_RE_tissue_specificity	Tissue specificity of the hypersensitiv site or regulatory element
DP	zero_point_of_HS_RE_sequence	"Zero" point of the hypersensitiv site or regulatory element sequence
PT	map_localization_of_HS_RE	Approximate map localization of the hypersensitiv site or regulatory element
PS	The_1st_relatively_po	The 1st position of the hypersensitiv site or regulatory

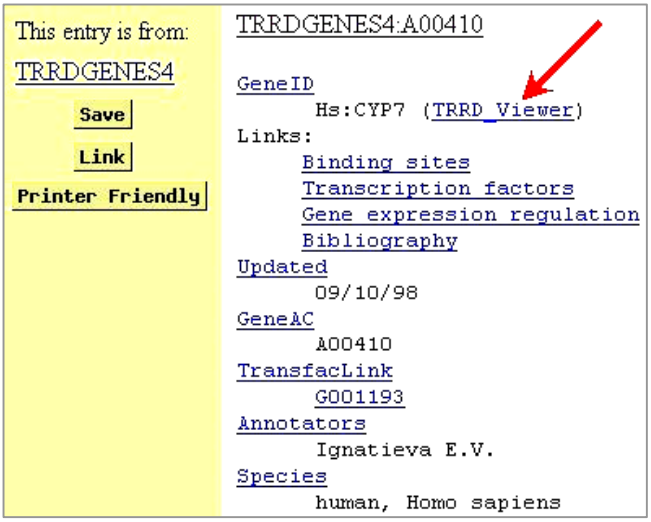
	sition_of_HS	element relatively to “zero” point
PE	last_position_of_HS_RE	The last position of the hypersensitiv site or regulatory element relatively to “zero” point
CD	HS_characteristic	The hypersensitiv site characteristic (strong, weak)
SQ	HS_RE_sequence	Nucleotide sequence of hypersensitiv site or regulatory element
PH	position_of_HS_RE_in_EMBL	Position of hypersensitiv site or regulatory element in EMBL/GenBank
PA	positive_negative_arguments	Positive or negative arguments
OA	organ	Organ
TA	tissue	Tissue
CA	cell_type	Cell_type
AG	arguments	Experiments for hypersensitiv site or regulatory element revealing and description
AU	authors	Authors
TI	title	Title
SO	source	Source
YR	year	Year
VL	volume	Volume
IS	issue	Issue
PGF	page_from	The first page
PGL	page_to	The lat page
ML	MEDLINE_index	MEDLINE index

TRRDBIB

Line code	Field name	Field description
ID	ReferenceId	Identifier of the reference
GID	GeneID	Identifier of an entry (gene) in the TRRDGENES4
AU	Authors	Authors of the paper
AI	ArticleInd	TRRD paper index
AD	ArticleDescr	Type of the annotated information
TI	Title	Title of the paper
SO	Journal	Journal name
VL	Volume	Volume number
YR	Year	Year
PG	Pages	Pages
ML	MEDLINE_link	MEDLINE link

TRRD Usage Examples:

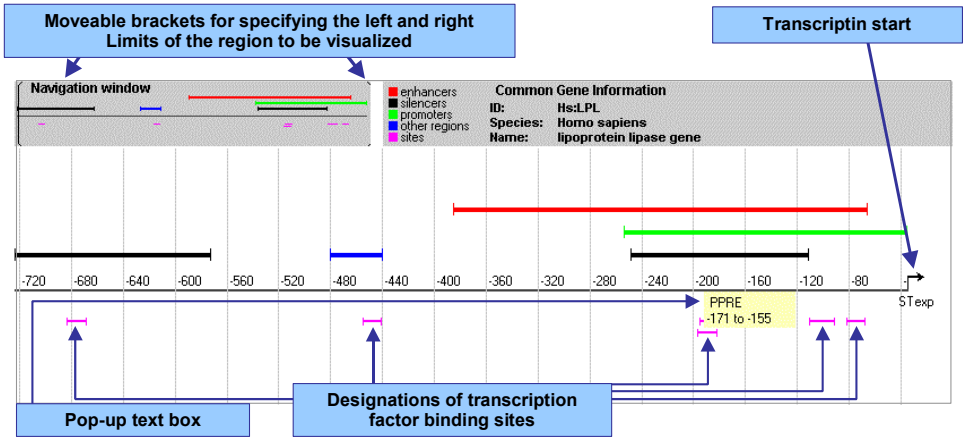
Example 1. How to view the map of regulatory regions of a gene?



The option to access the map of regulatory regions of a gene in question is provided by the table **TRRDGENES**. Each entry of the table has a special link TRRD Viewer. By clicking this hyperlink, you access to the map of regulatory regions of corresponding gene:

TRRD Viewer is Java applet and is developed using JDK 1.1.8. For the Viewer to work properly, either Internet Explorer 4.0 (or further releases) or Netscape Communicator 4.61 (or further releases) are required.

Visualisation of a gene regulatory map is exemplified below. Three windows are provided: (1) navigation window, (2) text box with the relevant information and designations, and (3) window with the map of gene regulatory regions. In this figure, visualisation of the map of gene regulatory regions by **TRRD Viewer** exemplified with human cholesterol 7-alpha-hydroxylase gene (CYP7) is displayed. A hooked black arrow indicates the transcription start of a gene. Shown on the axis is the distance from the reference point (here, the transcription start). A pop-up text box (yellow rectangle) presents the information on SRE binding site.



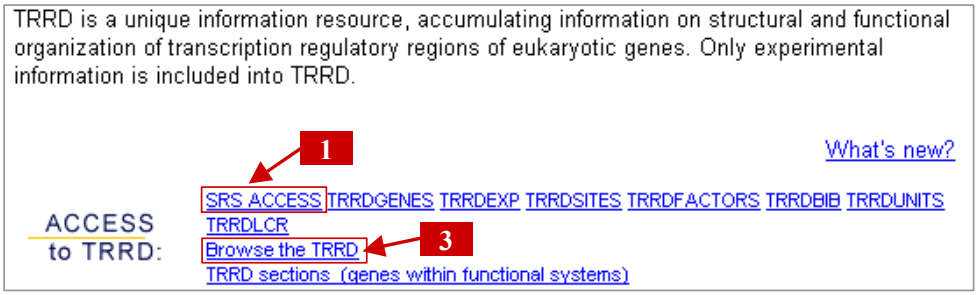
Options of **TRRD Viewer**:

1. Movable brackets for specifying the left and right limits of the fragment to be visualised in the window below;

2. Indicating a transcription factor binding site with the mouse cursor pops up the prompt with the site name and positions;
3. Indicating a regulatory unit (promoter, enhancer, or silencer) with the mouse cursor pops up the prompt with its name, reference point, and positions;
4. Clicking a regulatory unit (promoter, enhancer, or silencer) allows the binding sites contained in it to be shown, as upon clicking, all the corresponding sites will take the colour of this particular regulatory unit (green, blue, red, or black).

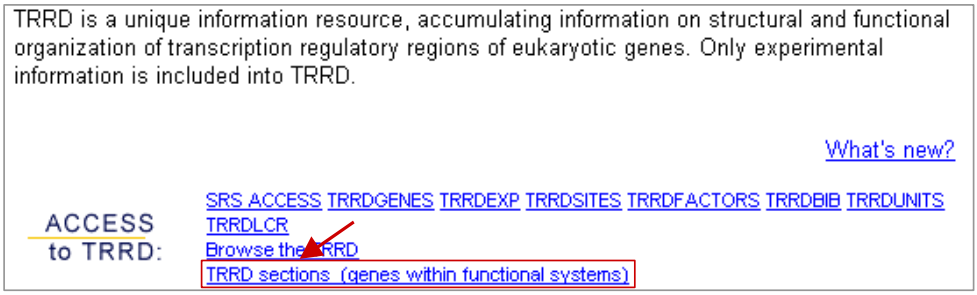
Example 2. Access to TRRD database

Access to **TRRD** by SRS (1), BLAST (2), Browser by species and Browser by gene name (3) is available at the **TRRD home page** via corresponding links:



Example 3. Access to functionally important gene systems (TRRD sections)

To get access to **TRRD sections**, click the link TRRD sections in the top menu of **TRRD homepage**:

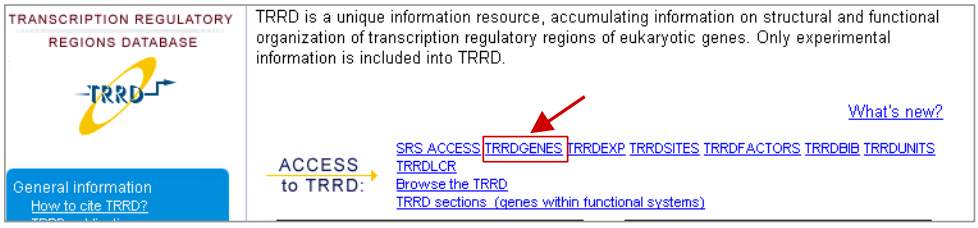


This will bring up a table with section names and hyperlinks. To get access to the section you are interested in, click the corresponding link (in the figure, **ESRG-TRRD** is chosen as an example):

TRRD Section	Short name and link	Compiler
Heat Shock-Induced Genes	HS-TRRD	Stepanenko I.L.
Interferon-Inducible Genes	IIG-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.
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.

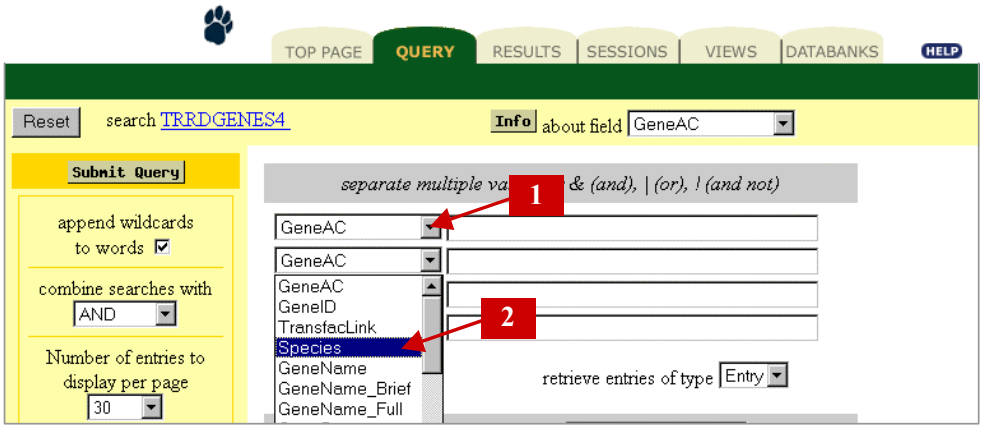
Example 4. Searching for the genes of a particular species

To display the list of genes of a particular species, use the standard (simple) Query Form of **TRRDGENES** table. Starting from **TRRD home page**, follow the hyperlink TRRDGENES:



On the 'Library Information' page of **TRRDGENES** then displayed, click 'Search' button:

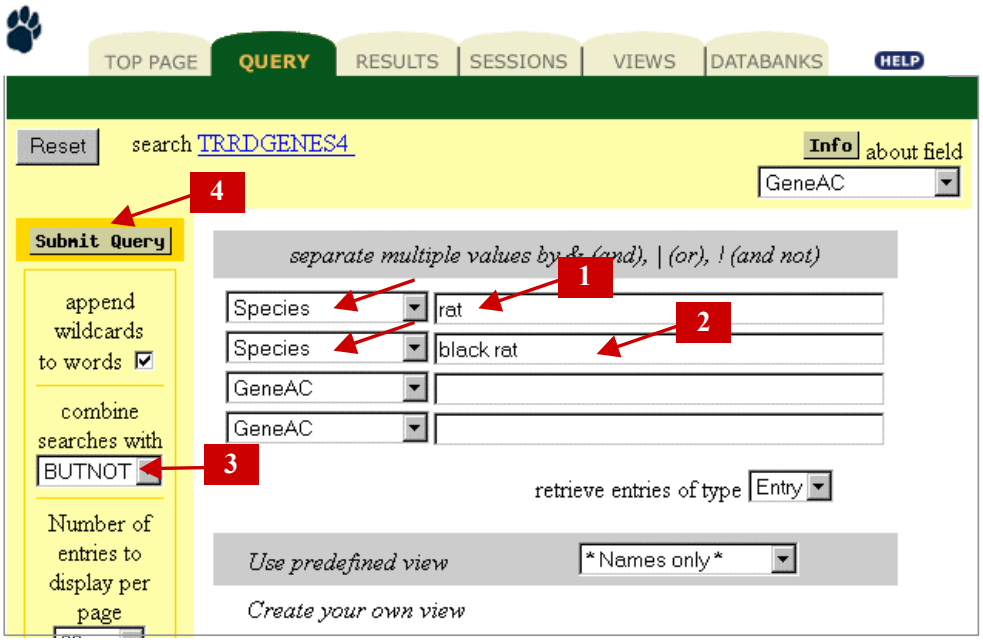
You will get the standard Query Form. Choose the data field 'Species' (arrow 2) by selecting the appropriate item in the drop-down list box 'GeneAC' (arrow 1 in the figure below):



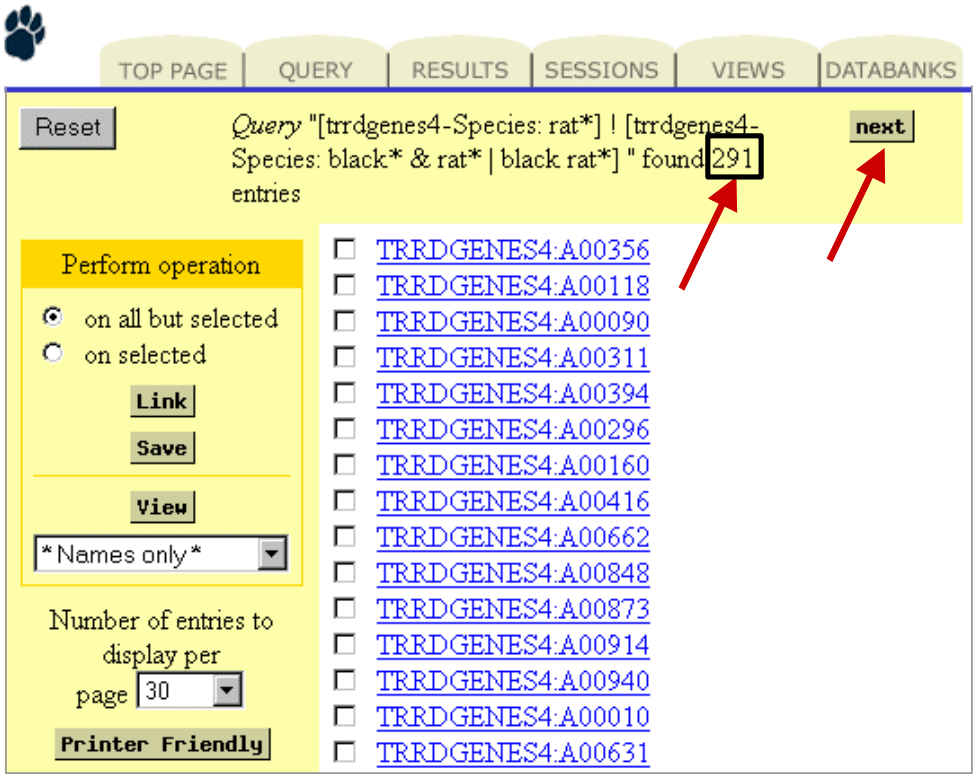
You may enter into the query text box the species name (or its part followed by wildcard symbol asterisk) in English or in Latin. Note that if you input a common species name in English, you may get the list of genes referring to several related species. For example, by entering 'mouse', you will get the list of genes of four murine species (*Mus caroli*, *Mus domesticus, Mus hortulans, Mus musculus*).

Example**:** search for the genes of grey (*Rattus norvegicus*), but not black (*Rattus rattus*) rat by the species name in English.

Enter the terms 'rat' (see arrow 1 in the figure below) and 'black rat' (arrow 2) into the appropriate consecutive 'Species' fields (unnumbered arrows). In the 'combine searches with' drop-down list select the 'BUTNOT' item (arrow 3), and then click the 'Submit Query' button (arrow 4):



The query brings up the resulting list of genes containing 291 entries. Use the vertical scrolling bar to fit the content of interest in the window. Click 'next' button to view the next portion of resulting data:



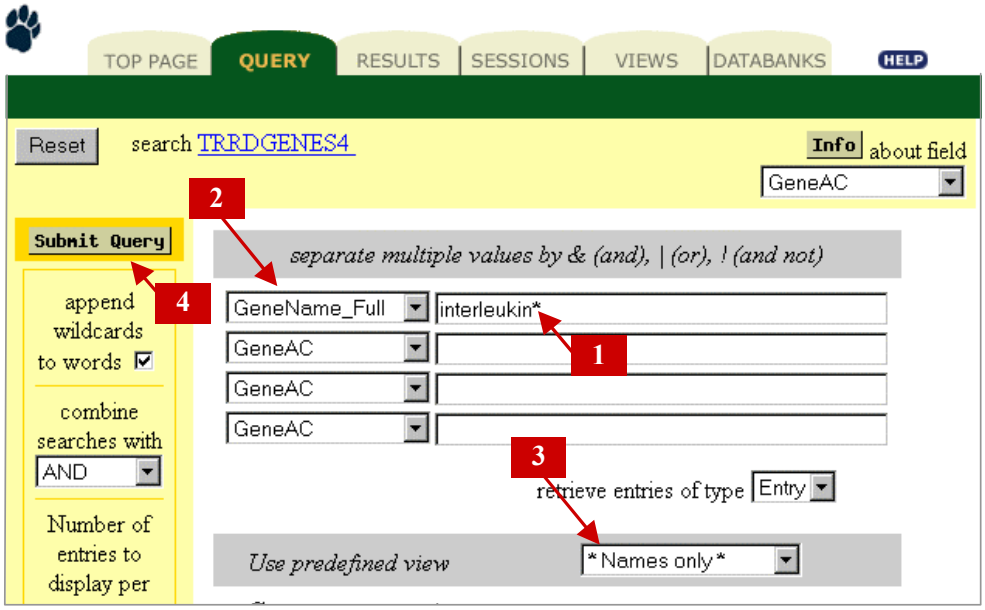
Example 5. Searching for the genes by name

To display the genes with the similar name, use again the standard Query Form for the **TRRDGENES** table.

Use the following data fields (by selecting appropriate items in the drop-down list boxes): 'GeneName_Full' and/or 'GeneSynonym', in case you know the complete gene name, or 'GeneName_Brief' and/or 'GeneSynonym' if you know the brief gene name.

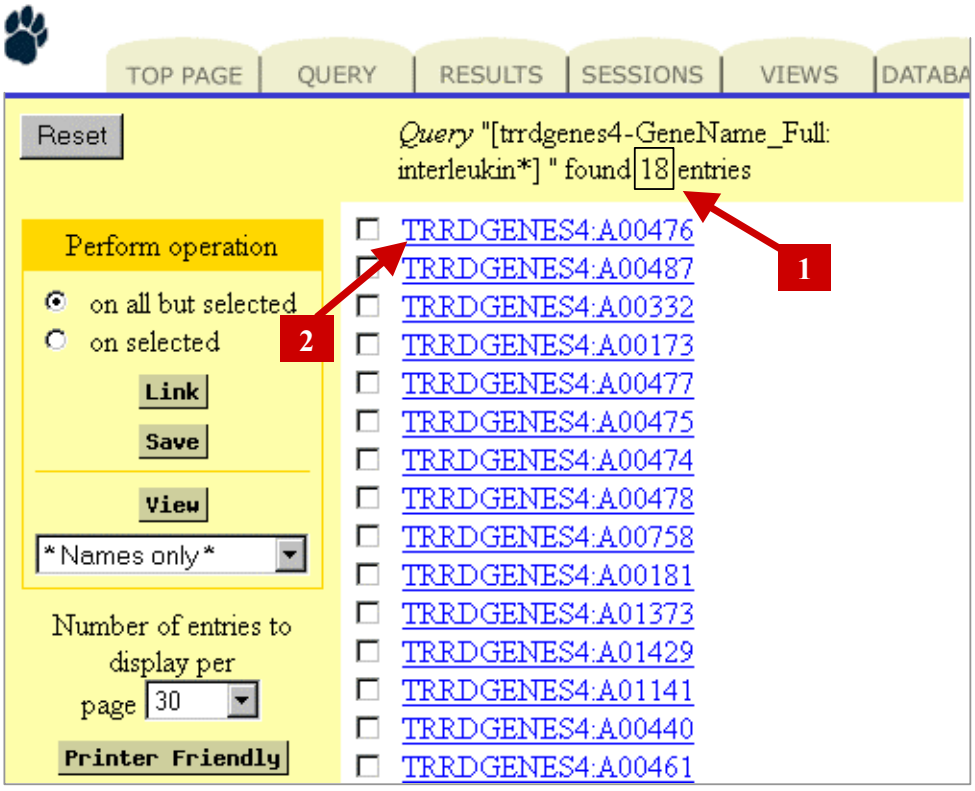
Example: querying the genes encoding interleukines.

Enter the term 'interleukin*' (see arrow 1 in the figure below) into the data field 'GeneName_Full' (arrow 2). Check the drop-down list box 'Use predefined view', the default option *Names only* should be selected, if not, select it (arrow 3). Click 'Submit Query' button (4):

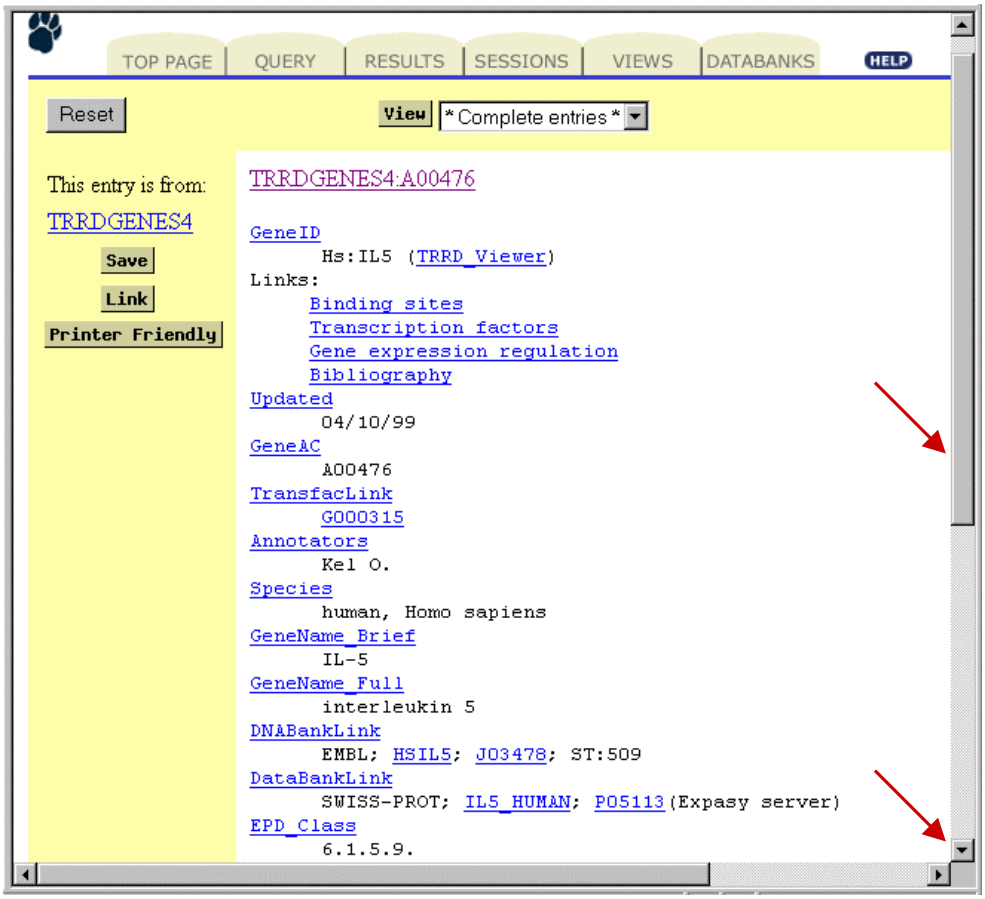


This will bring up the list of entries accumulated in **TRRD**. In total, 18 genes will be displayed (see arrow 1):

To receive more details click any hyperlink denoted in accordance with the gene index in **TRRD**, for example the first one on the top of the list (arrow 2).



The resulting detailed view is shown below. Its content exceeds the window dimensions; so use the vertical scrolling bar marked by arrows to fit it in:

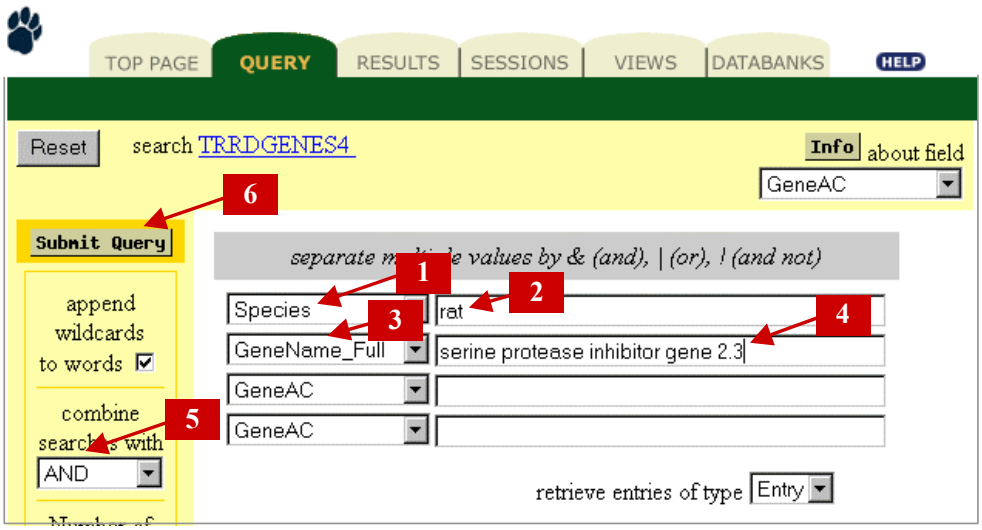


Example 6. Searching for the genes by name and species name

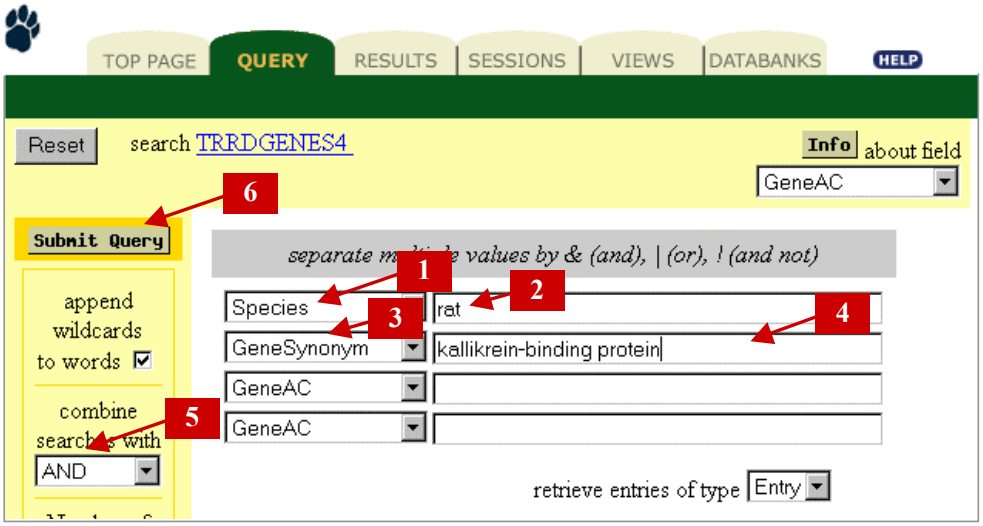
Use the standard Query Form for the **TRRDGENES** table in this and all below mentioned cases unless otherwise specified.

Example. Search for the gene encoding rat serine protease inhibitor 2.3.

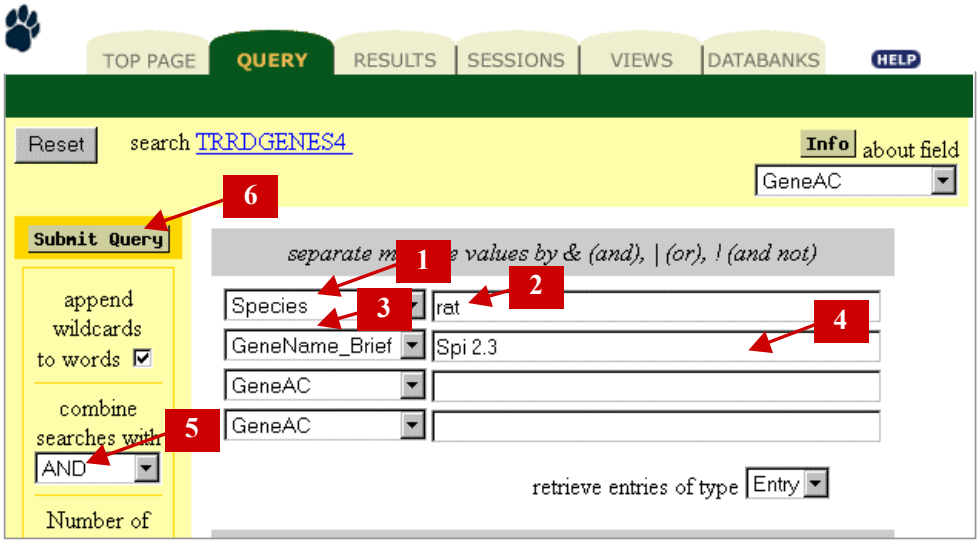
Enter the term 'rat' into the field 'Species' (steps 1 and 2 in the figures). Compose the query by one of the following three options:
a) You may use the complete gene name. In the data field 'GeneName_Full', enter the term 'serine protease inhibitor gene 2.3' (see steps 3 and 4 in the figure below):



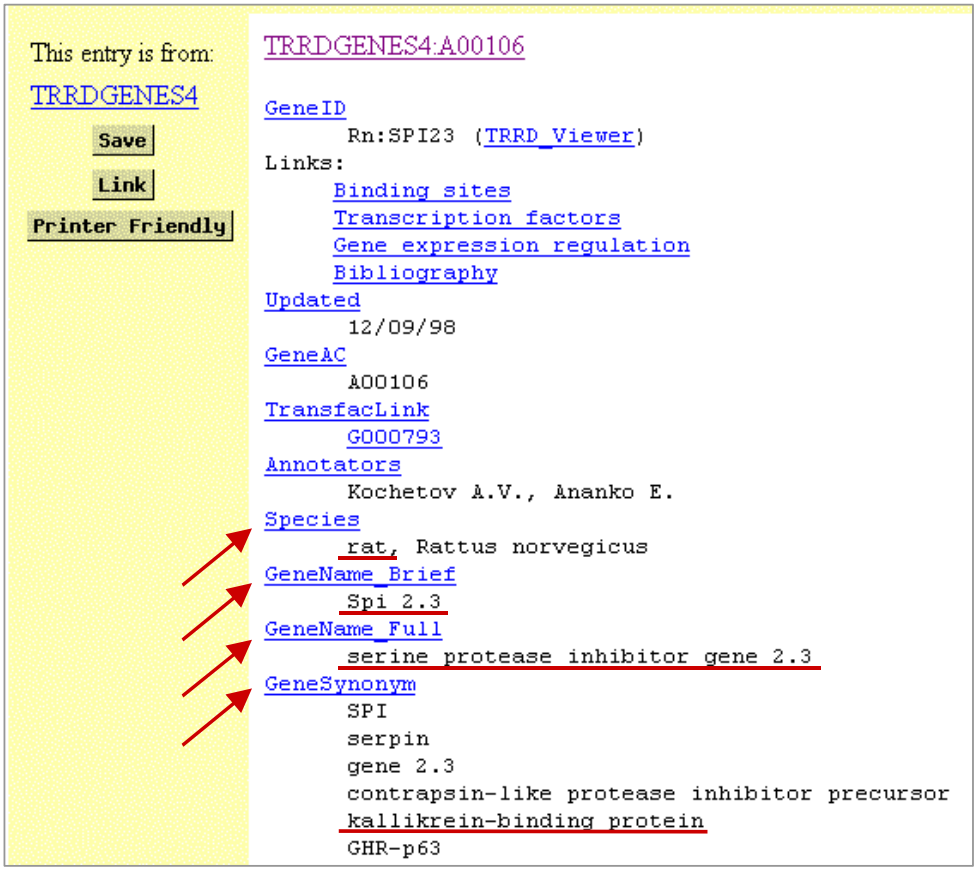
b) If you know the synonym, e.g., in our case 'kallikrein-binding protein', enter this term into the data field 'GeneSynonym':



c) If you know the standard gene abbreviation, for example, 'Spi 2.3', enter it into the data field 'GeneName_Brief'.
Combine the searches with the data field 'Species' and one of the fields mentioned above by 'AND', then submit the query (steps 5 and 6 in the figures):



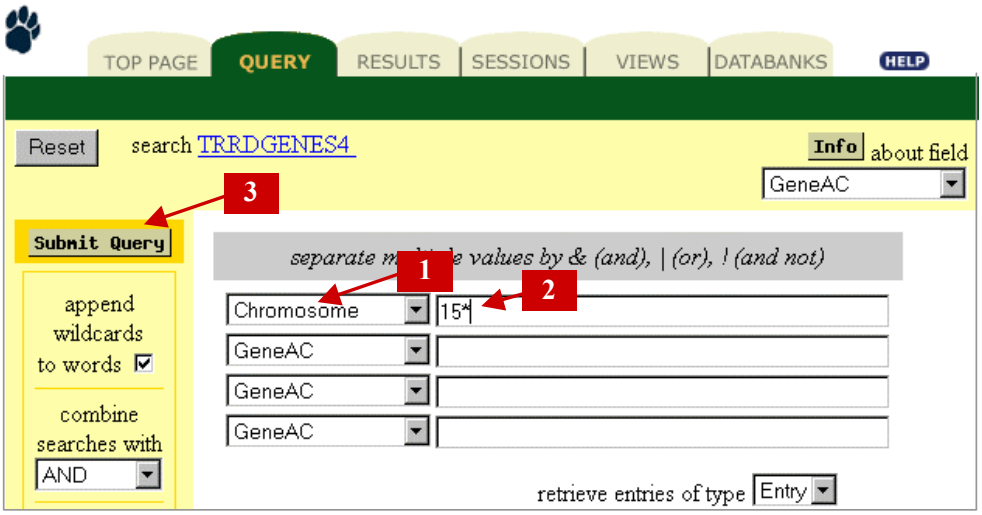
The query brings up the following result. Data items of interest are underlined by red:



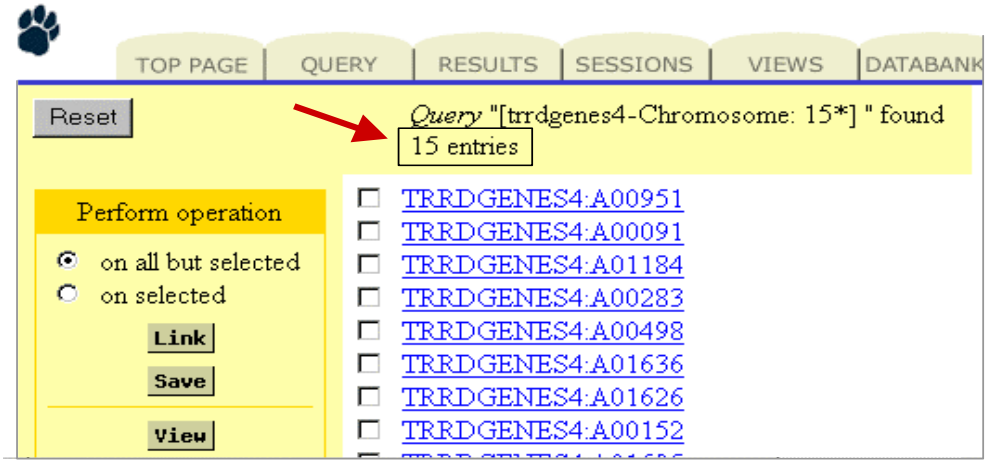
Example 7. Searching for the genes with particular chromosomal localisation

To display the list of genes localised at particular chromosome, select the data field 'Chromosome' and enter the chromosome number, for example '14*' or '15*'. Note that the asterisk should supplement the number, because the database may contain the record '14' as well as the record '14q32.32'.

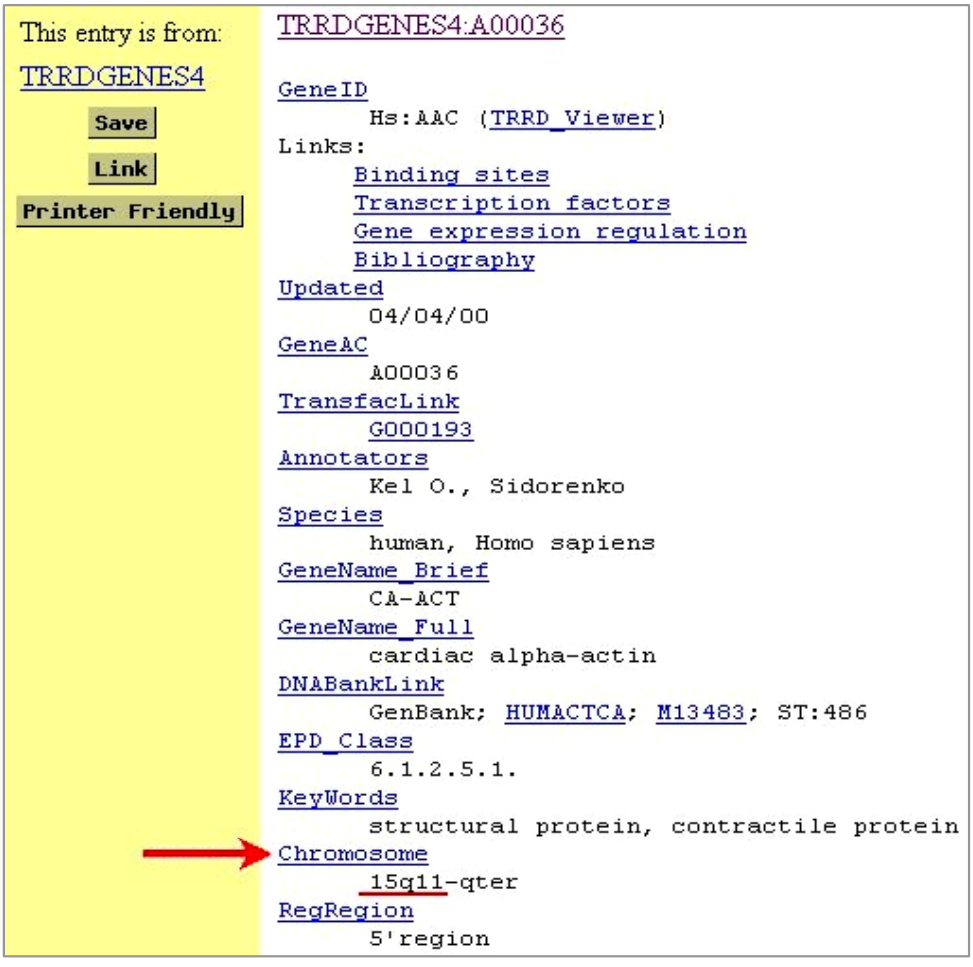
Example: querying the genes localised at chromosome 15:



The result of querying contains 8 entries:



Click the hyperlink pointed to by the arrow 1 in the figure above to receive the complete record data of the first found gene (depicted below):

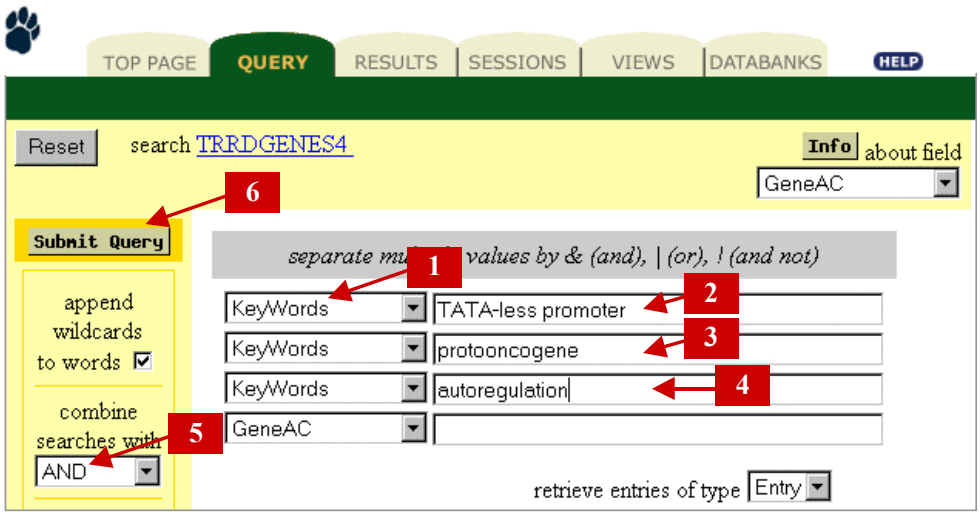


Example 8. Searching for the genes by KeyWords

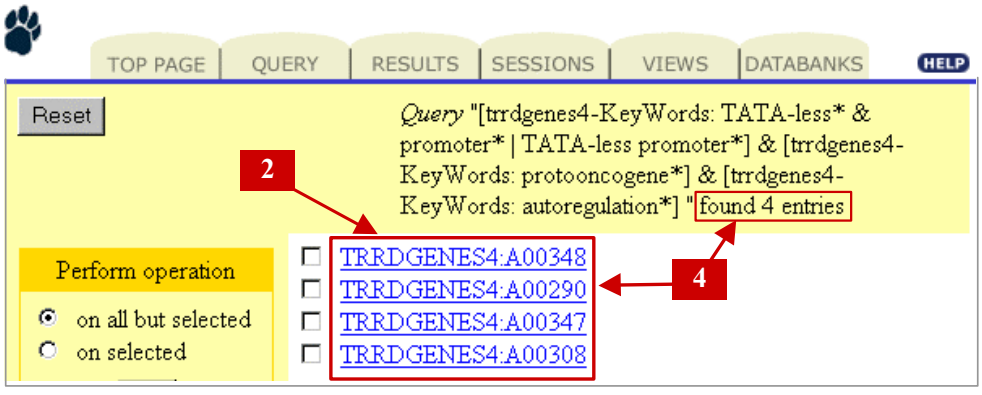
Enter some known to you characteristic properties of the gene (or its product) to be found into the field 'KeyWords'. By querying the keywords, it is possible to extract the following information:

- the list of genes with characteristic structural features, for example, by querying TRRD for keywords 'TATA-less promoter', 'multiple transcription initiation sites', 'GC-rich promoter', 'alternative exon', 'alternative promoters', 'Alu repeat';
- the list of genes with typical patterns of functioning, by ordering the query by the keywords 'alternative splicing', 'allergic triggering', 'autoregulation', 'DNA damage-induced';
- the list of genes by their products, by means of searching for keywords 'glycoprotein', 'hormone precursor', 'transcription factor', 'growth factor', 'enzyme';
- the list of genes according to peculiarities of their products' functioning by querying the keywords 'DNA binding protein', 'G-protein coupled receptor', 'mitogen', 'monooxygenase';
- the list of genes participating in functioning of this or that organism system, or the organism as a whole, by querying the keywords 'glyoxylate cycle', 'lignin synthesis', 'iron homeostasis', 'intracellular protein', 'housekeeping gene', 'homeotic gene', 'heme biosynthesis pathway', 'growth regulation', 'embryo-specific gene';
- the list of genes acting in formation of this or that organism structure, by ordering the query by keywords 'exocuticle', 'eye lens';
- the list of genes according to evolutionary homology, for example, by querying the keywords 'growth hormone family', 'heat shock-related').

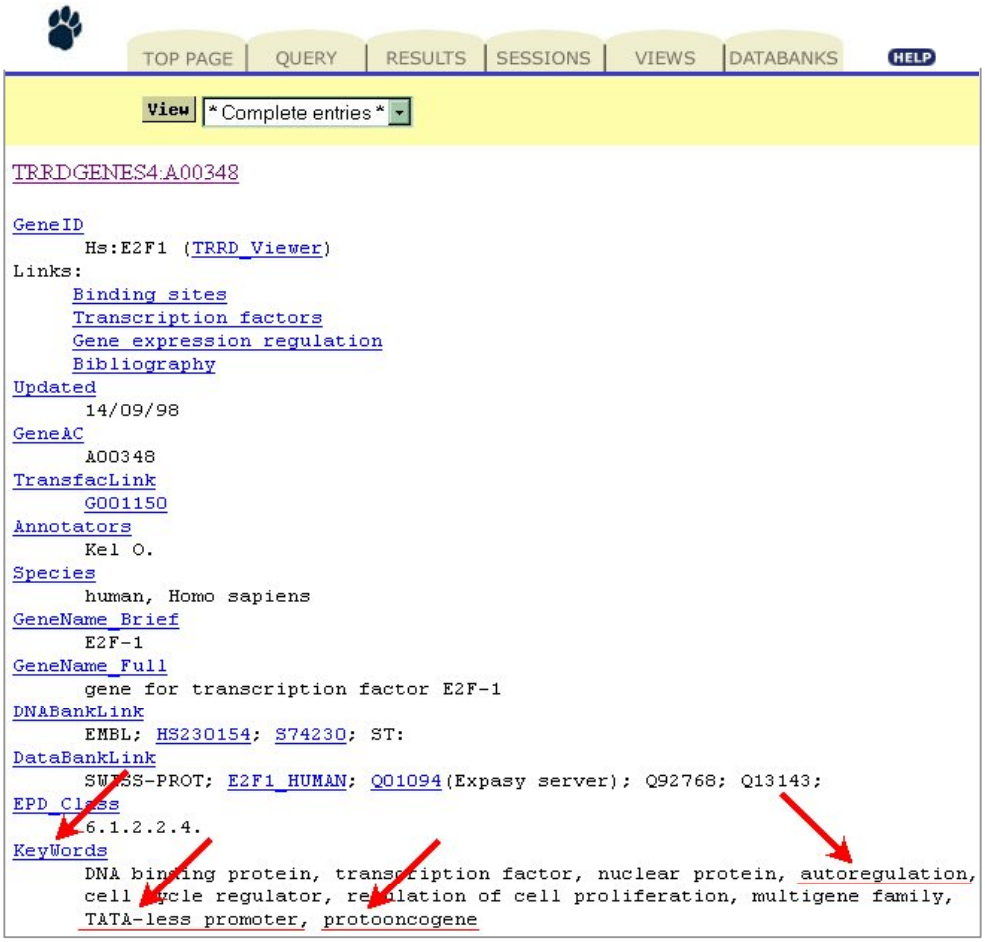
Example: to search for the genes by combination of keywords 'TATA-less promoter', 'protooncogene', 'autoregulation'. For this purpose, it is necessary to select the data field 'KeyWords', to input these keywords, to select combination 'AND' and to submit the query:



This will bring up the resulting list of **TRRD** entries. In our example, 4 entries were found (see arrows 1):

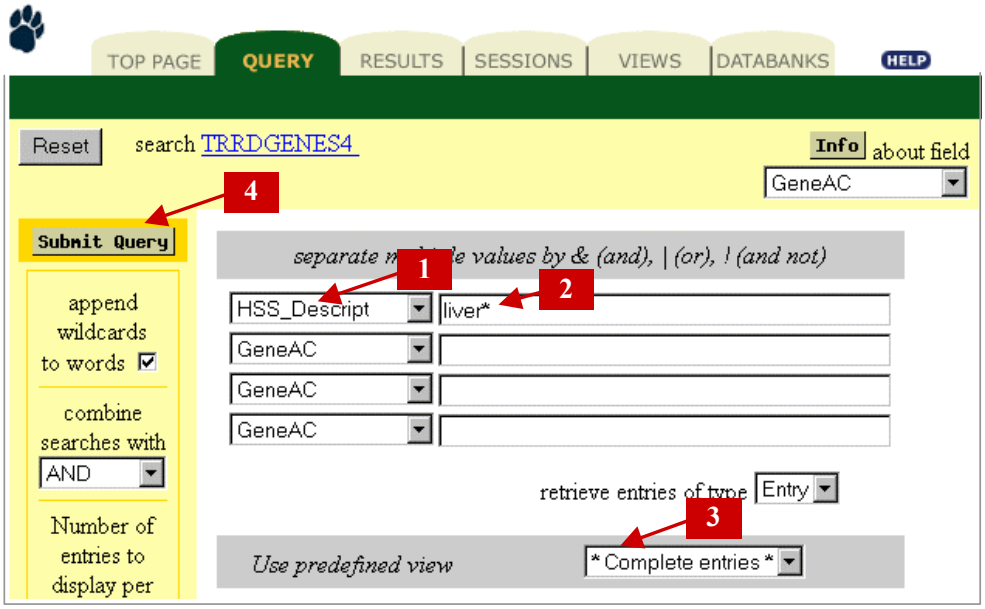


By clicking the above marked hyperlink A00348 (arrow 2), you may receive the complete record data related to this gene (shown in the figure below). The keywords due to which the entry of the gene A00348 was extracted are underlined by red and pointed to by arrows:



Example 9. Search for the liver-specific hypersensitive to DNase I sites

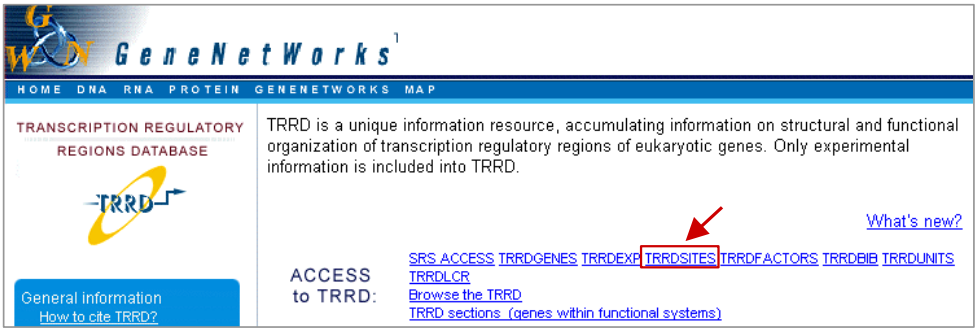
Use the field 'HSS_Descript'. Input into this field the name of an organ with an asterisk, 'liver*' in our case. You may choose the *Complete entries* option in the 'Use predefined view' drop-down list box for convenience. Submit the query:



A fragment of one of the **TRRD** entries found is shown in the figure:

Example 10. Search for the transcription factors binding sites by names

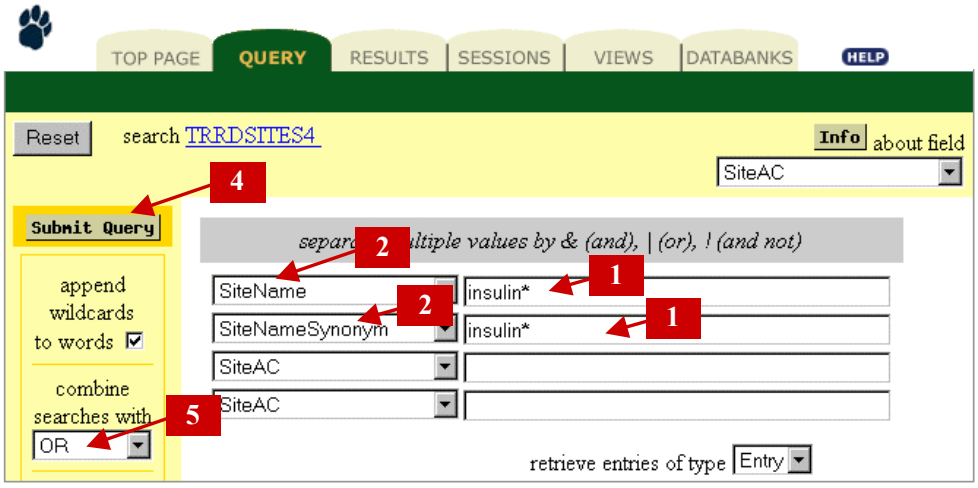
To display the list of transcription factors binding sites, use the standard Query Form of **TRRDSITES** table. Starting from **TRRD home page**, follow the hyperlink TRRDSITES:



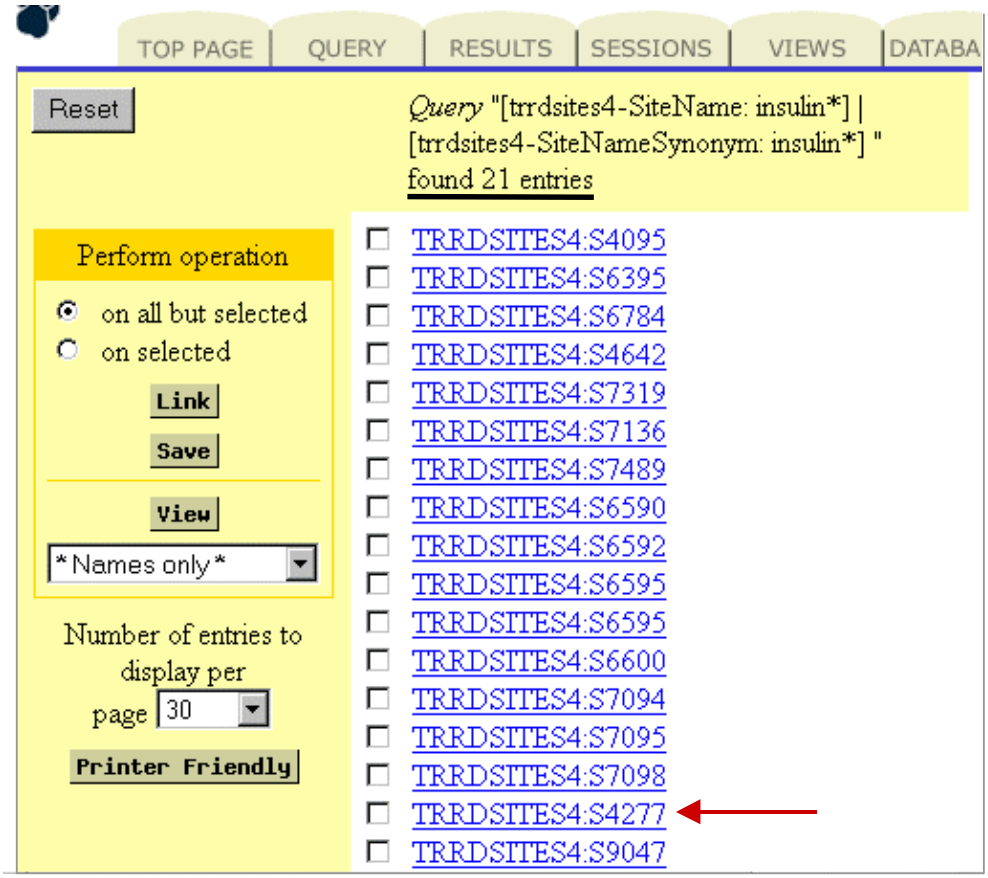
On **TRRDSITES** 'Library Information' page then displayed, click the 'Search' button:

This will bring up the standard Query Form (see the figure below).

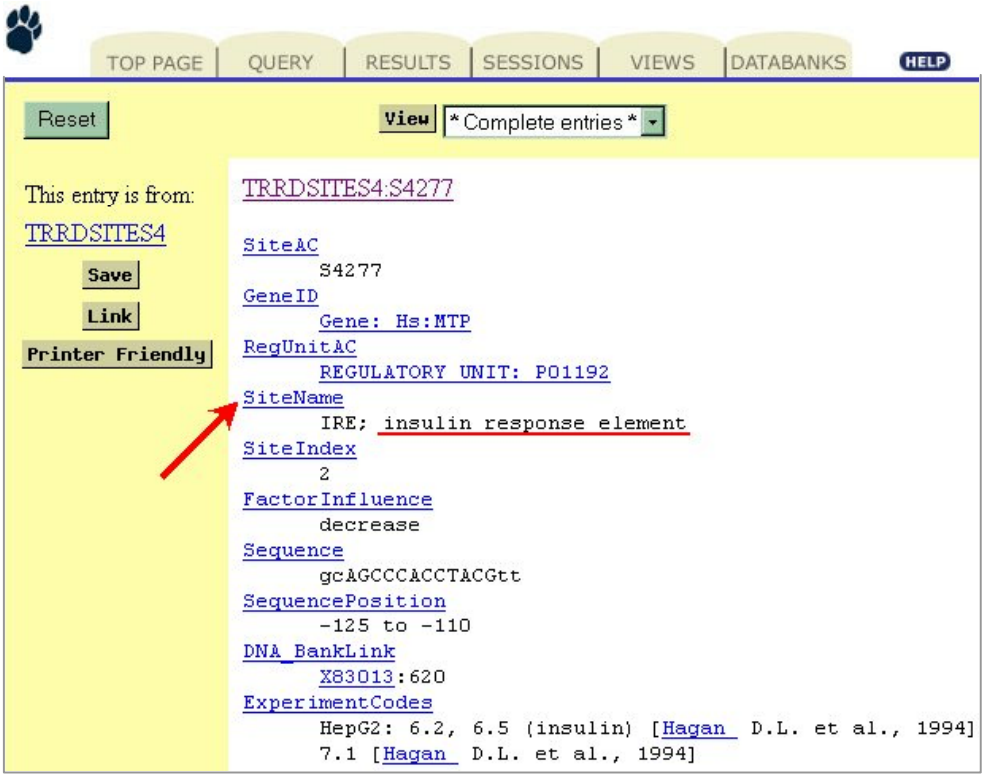
Example: Let us search for insulin responsive element and insulin responsive sequence by transcription factor binding site names and their synonyms. Since abbreviations of these regulatory sequences, IRE and IRS, coincide with abbreviations of interferone-regulated sites, it is recommended to set the query by complete name. Enter the term 'insulin*' (arrows 1) into the data fields 'SiteName' and 'SiteNameSynonym' (arrows 2) and combine the searches with 'OR' (arrow 3):



After clicking the screen button 'Submit Query' (arrow 4) you may look through the following resulting list of transcription factor binding sites:



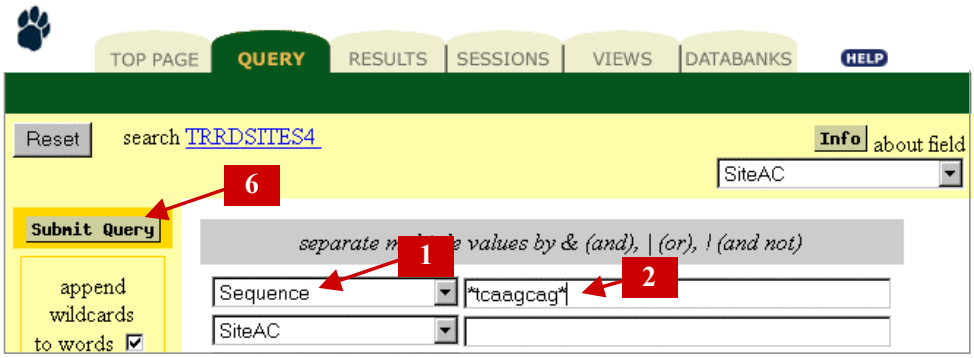
Click the first, for example, item in the list (pointed to by red arrow in the figure above) and by means of hyperlink you will receive the complete entry data of appropriate site:



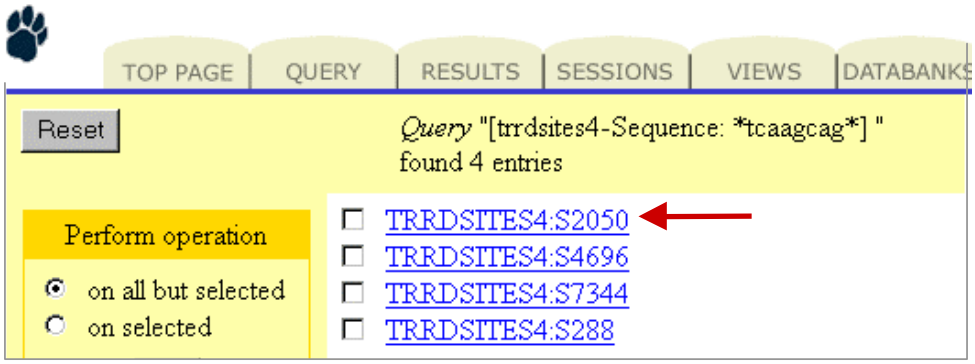
Example 11. How to search for a site by its sequence?

Use again the standard Query Form for the **TRRDSITES** table. Select the data field 'Sequence' and enter or copy from clipboard buffer the sequence of transcription factor binding site to be found. Note that the sequence to be entered should not contain any symbols except nucleotides (the only exception is for asterisks that may border the sequence, see the figure). In case of searching for long sequences, the sequences with deletions, insertions, duplications, etc., you should better use the option 'Blast search TRRD database'.

Let us search for transcription factor binding sites containing the sequence '`tcaagcag`'. Enter the sequence to be found `*tcaagcag*` limited by asterisks into the data field 'Sequence':



Submit the query and view the resulting list of transcription factor binding sites:

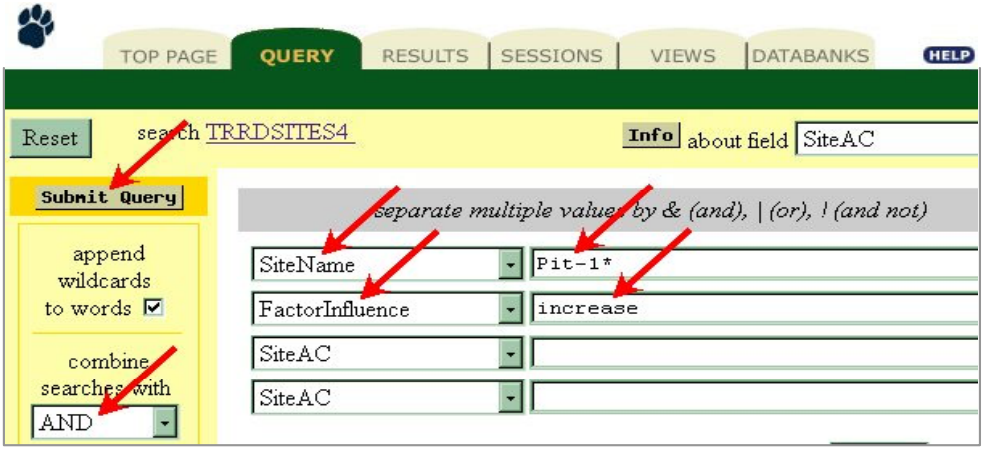


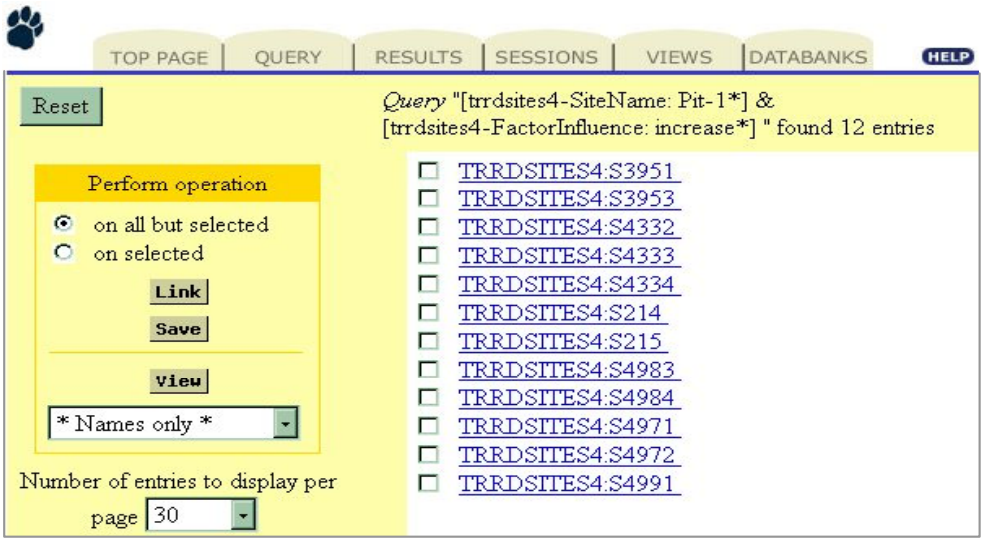
Click any site name (which is the hyperlink at the same time) to receive the complete entry content:

Example 12. How to search for sites by the factor name and the factor influence?

Start from the standard Query Form for the **TRRDSITES** table, as in previous cases. In order to search for the sites binding to particular transcription factors, which influence gene transcription in a particular manner, it is necessary to make combined query by two fields, 'SiteName' and 'FactorInfluence'. For this purpose, it is sufficed to enter into the data field 'FactorInfluence' one of the terms 'increase' or 'decrease'.

Let us search for transcription factor binding sites Pit-1, interaction of which with the factor of the same name activates transcription. Into the data field 'SiteName', enter the site name or its part (in this example, 'Pit-1*'). In the field 'FactorInfluence' enter 'increase'. Combine the search by 'AND', submit query and view the resulting list.

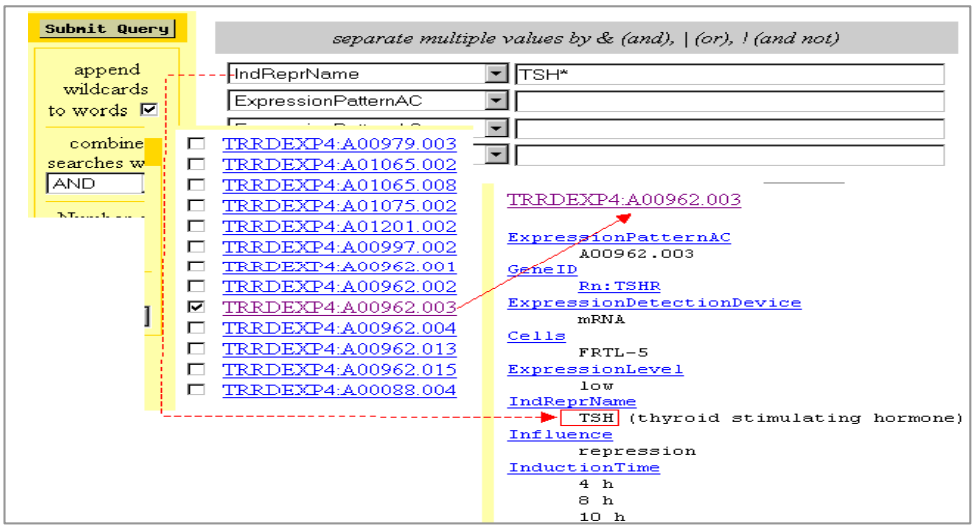




Example 13. How to search genes by inducer or repressor name?

Use the standard Query Form for the **TRRDEXP** table. In order to extract the list of genes, about which there is information that their expression is modified by a particular factor, it is necessary to query the data field 'IndReprName'. Under the term 'factor' we understand any impact, for example, hormone, transcription factor, diet, pathology, surgery, or emotional action.

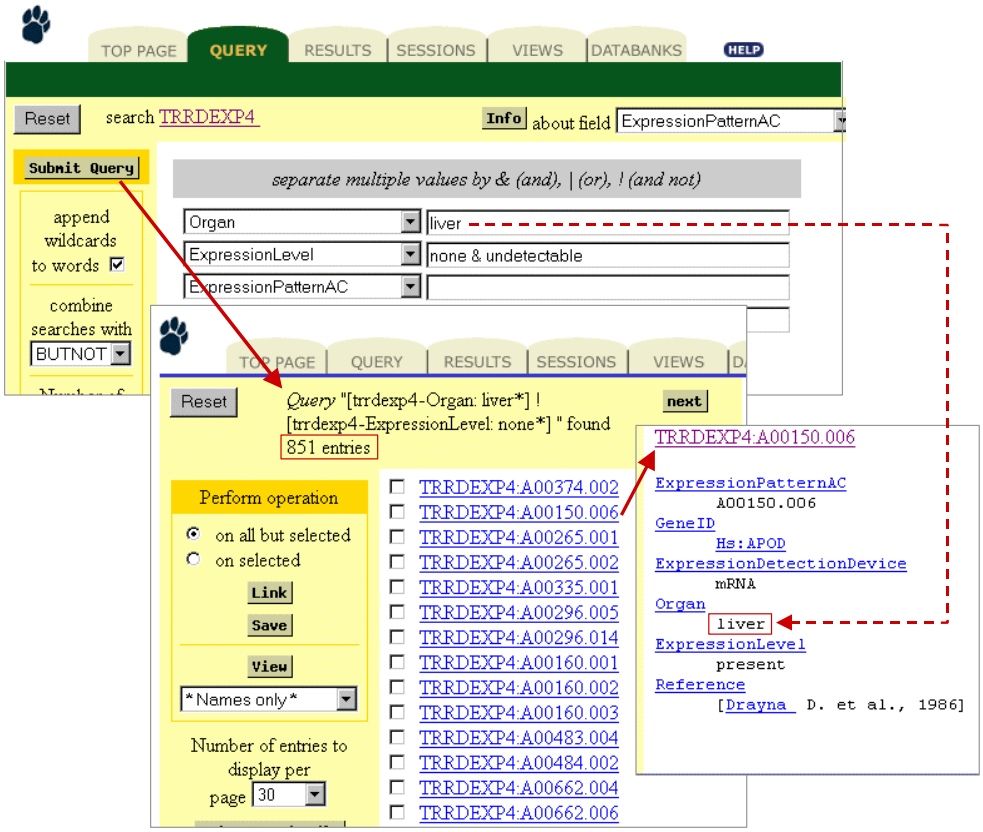
Example: Search for the genes that contain information about the influence of TSH (thyroid stimulating hormone) on their expression. Into the field 'IndReprName', enter 'TSH*':



Example 14. How to search expression patterns from genes that are expressed in particular organ?

Use again the standard Query Form for the **TRRDEXP** table. To search for expression patterns related to the genes expressed in a particular organ, it is necessary to make query by two fields, 'Organ' and 'ExpressionLevel'. It is necessary to account that among the terms in the field 'ExpressionLevel' may be the following: 'exclusive expression', 'high', 'low', 'maximal level', 'med', 'minimal level', 'none', 'present', 'undetectable', 'very high', 'very low'. To search for the gene patterns that are expressed in particular organ, it is necessary to except from the query the patterns of expression that contain the records 'none' and 'undetectable' in the field 'ExpressionLevel'.

Example: Search for expression patterns related to genes expressed in liver. Choose for the querying data field 'Organ', enter the term 'liver'. Choose the querying field 'ExpressionLevel', enter the terms 'none' and 'undetectable', combine them by the sign '&'. Combine the search by 'BUTNOT'. Submit the query:


TOP PAGE
QUERY
RESULTS
SESSIONS
VIEWS
DATABANKS
HELP
Reset
search TRRDEXP4
Info about field ExpressionPatternAC
Submit Query
separate multiple values by & (and), | (or), ! (and not)
append wildcards to words
combine searches with
BUTNOT
Organ
liver
ExpressionLevel
none & undetectable
ExpressionPatternAC
Query "[trrdexp4-Organ: liver*] ! [trrdexp4-ExpressionLevel: none*] " found
851 entries
next
Perform operation
on all but selected
on selected
Link
Save
View
* Names only *
Number of entries to display per page 30
TRRDEXP4:A00374.002
TRRDEXP4:A00150.006
TRRDEXP4:A00265.001
TRRDEXP4:A00265.002
TRRDEXP4:A00335.001
TRRDEXP4:A00296.005
TRRDEXP4:A00296.014
TRRDEXP4:A00160.001
TRRDEXP4:A00160.002
TRRDEXP4:A00160.003
TRRDEXP4:A00483.004
TRRDEXP4:A00484.002
TRRDEXP4:A00662.004
TRRDEXP4:A00662.006
TRRDEXP4:A00150.006
ExpressionPatternAC
A00150.006
GeneID
Hs:APOD
ExpressionDetectionDevice
mRNA
Organ
liver
ExpressionLevel
present
Reference
[Drayna D. et al., 1986]

CHAPTER 2. SITE RECOGNITION MODULE

1. B-DNA Site Video

1.1. B-DNA Site Video Databases

Release 2003

Databases description:

These databases are designed to study the sets of various transcription factor binding sites, providing evidence that transcription factor binding sites are characterized by specific sets of significant conformational and physico-chemical DNA properties.

Access to B-DNA Site Video Databases:
http://www.domain.com/mgs/gnw/bdna/

Databases content:

SRS table	Description	Number of entries
FEATURES	Knowledge base on significant B-DNA properties of sites	51
PROPERTY	Database on sequence-dependent conformational and physico-chemical B-DNA properties	38
SAMPLES	Database on functional site sequences	77
PROFILE_LIST	This SRS table accumulates sets of significant physico-chemical properties profiles of the sites.	6

List of biological tasks that could be solved by using the B-DNA Site Video Databases:

- to extract specific sets of significant conformational and physico-chemical DNA properties of various transcription factor binding sites;
- to browse information about all conformational and physico-chemical DNA properties that differ significantly for the sequences of transcription factor binding sites and random sequences (the sequences of transcription factor binding sites occurring in nature are stored in the SAMPLES database);
- for a site given, by using the B-DNA features selected for recognition of this particular site, to generate the C-program recognising this site on the basis of the site properties stored in the database B-DNA-VIDEO;

SRS tables format:

FEATURES

Line code	Field name	Field description
MI	Entry ID	This field indicates an identifier.
MN	Site Name	This field contains the name of transcription factor binding site.
HN	SCIENTIST	This field contains the link to SCIENTIST database and indicates a contributor of the entry.
DR	SRS-links	This field contains the links to the supplementary databases installed under SRS (SAMPLES, etc.).
WW	Web-link to Recognition Tools	This field contains the link to the Web-based tools implementing each C program documented within the entry and aimed to recognition of transcription factor binding site by its conformational and physico-chemical features within an arbitrary DNA sequence.
DP	Link to PROPERTY database	This field contains the link to the PROPERTY database installed under SRS.
PV	Property Name	This field contains the short name of the property investigated.
HL	Feature Indicator	This field indicates high or low deviation of the conformational or physico-chemical feature of the site studied from the corresponding value for random sequences.
AB	Analyzed Region	This field indicates the DNA region, for which the conformational or physico-chemical feature differing significantly the sites studied from the random sequences was revealed. Positions are given in bp relative to the site center.
UT	Utility	This field indicates the utility of the conformational or physico-chemical feature for discrimination of the sites studied from the random sequences.
ST	Means, Standard Deviation, False Negatives for Control Sequences	This field provides the mean value of the property at [a; b] region (first estimate), the standard deviation of this mean value (second estimate), and the I type error rate (false negatives), which have been determined for the independent control sequences of the site by the program given below in the field C-CODE.
NT	Means, Standard Deviation, False Positives for Random Sequences	This field provides the mean value of the property at [a; b] region (first estimate), the standard deviation of this mean value (second estimate), and the II type error rate (false positives), which have been determined for random sequences by the program given below in the field C-CODE.
FG	Graphical Representation of Test Results	This field represents the links to the graphical representation of the results obtained by the program, given below in the field C-CODE, over the independent control sequences of the site.
C-CODE		This field contains the recognition program by the sequence-dependent conformational or physico-chemical feature or Mean recognition in the 'C' language of the ANSI standard. Each C program documented within the B-DNA-Video entry has a check-box in the MENU window of the Recognition Tools (see field WW).

PROPERTY

See **Chapter 5.1. ACTIVITY Databases**.

SAMPLES

Line code	Field name	Field description
FI	SampleID	brief name of the sample
NM	SampleName	name of the sample and explanation of its biological meaning
OR	Organization	title of the organization
AU	Author	the first and last name of the author
DA	Date	date of creation
LU	LustUpdate	the date and the author of the last update/modification (Last update); in case of no modification, the date and the authors name are from the DA and AU fields
FV	FormatVersion	Number of the Format Version
ST	SiteTypeDescription	ST {X,Y} [Left_border,Right_border] Point; Description; Factor_name. - (Structure) description of the site (not a particular site, but type of the sites); several ST fields are allowed; X=0….9 - the number of site batch (sites are joined into batches if they in the aggregate they represent an integral structure-function unit); Y=0….9 – the number of the site in a site batch; [Left_border,Right_border] – The position of the site in case this particular site type possesses a fixed location relative to a specific point, for example, relative to transcription start. If no, put nothing; Point – the point relative to which the site location is fixed; If the site has no fixed position, put nothing Description - conventional name of the site (abbreviated name if available) Factor_name - a name of site-binding factor
WA		reference to 'Aligned' database
WF		reference to 'Features' database
WW		hyperlink to x-ray structure of a binding complex
ID	CardID	identifier of a card
AC	CardAC	accession number of a card
OS	Specia	organism specia
OC	Taxon	organism taxon
DR	SourceDatabase	reference to the source database
CC	Comments	comments
FT	FeatureTable	the field contains positions of site described in ST field. The format is as follows: FT {X,Y} [left;right]; Method X = 0…9 → number of the site batch Y = 0…9 → number of the site in the site batch [left;right] → positions from the start of the sequence Method = EXP (experimental), GBS (Gibbs Sampler alignment), RCG (Recognition Group alignment;
SQ	Sequence	nucleotide sequence

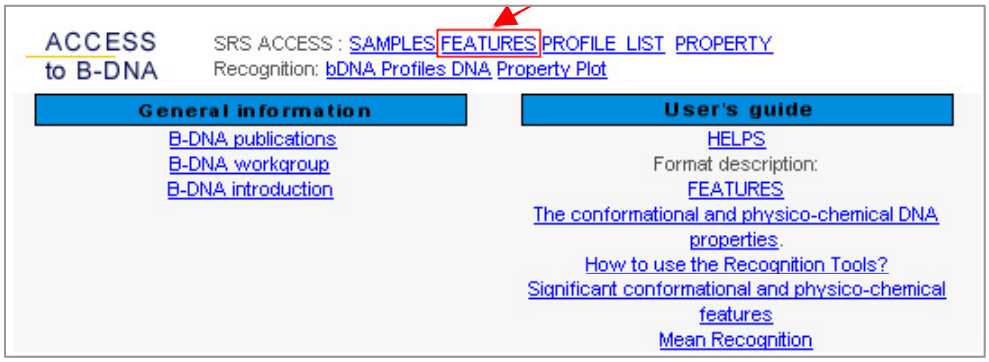
PROFILE_LIST

See **Chapter 6.1. PROFILES Databases**.

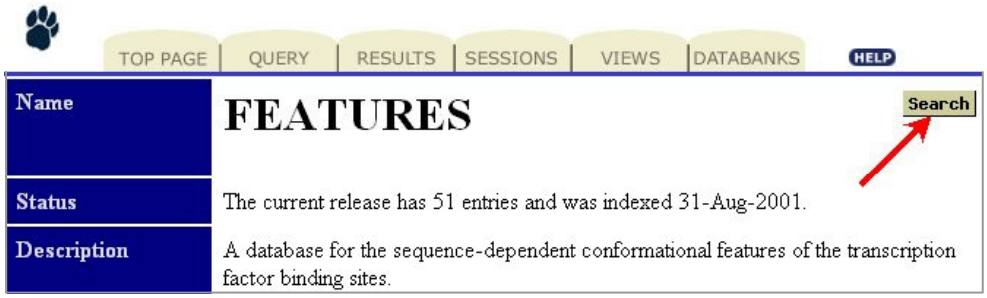
Example of SRS queries to the B-DNA Site Video Databases:

What are conformational and physico-chemical DNA properties differing significantly the transcription factor binding site AP-1 sequences from the random sequences?

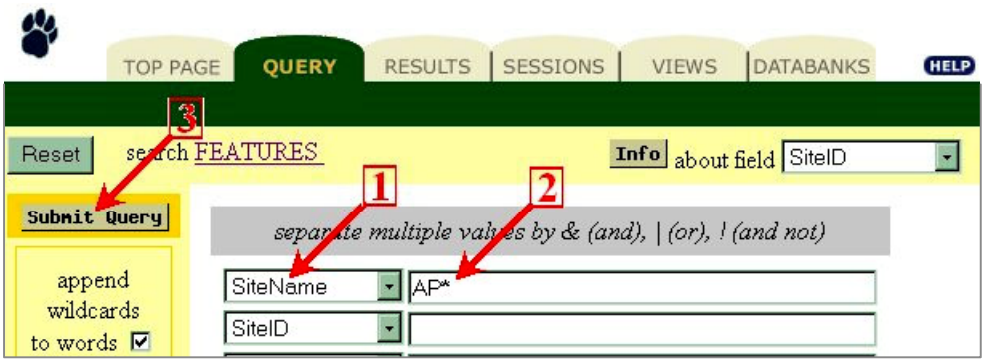
To make such a query, you should perform the following:
1. Choose 'FEATURES' SRS table on the page 'access to B-DNA':



2. This will bring up the home page of the chosen 'FEATURES' SRS table. Click the button 'Search':

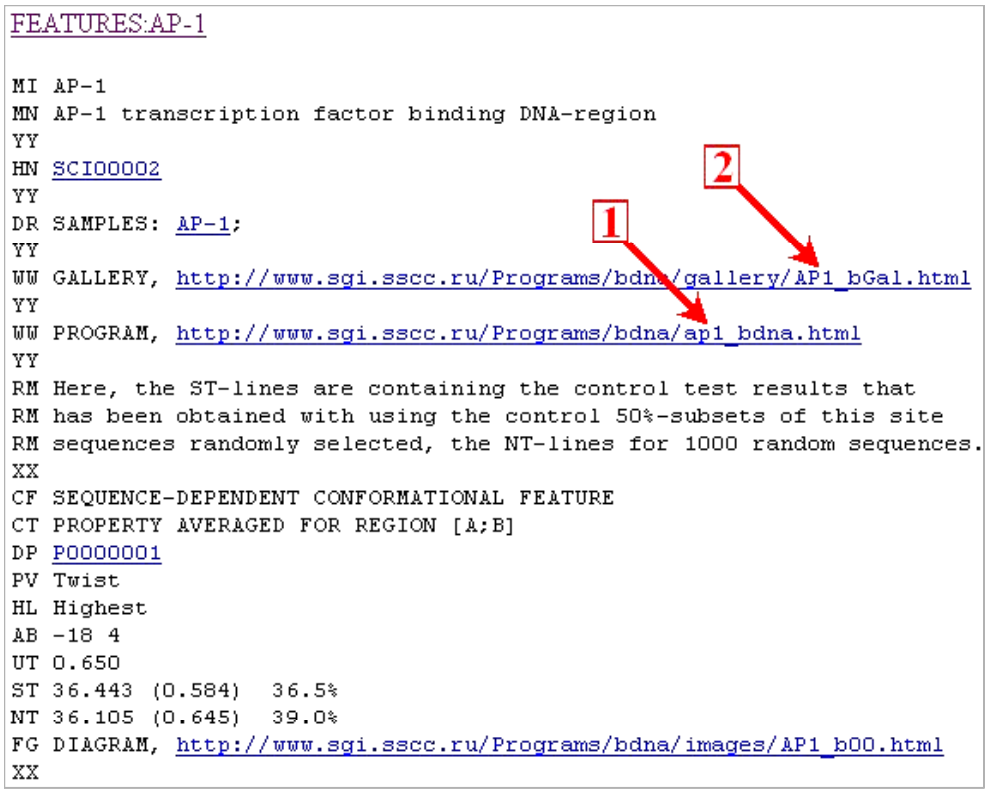


3. Select the fields to be searched for from the list. You will need the field 'SiteName' (1), Type terms to be searched in the text window: 'AP*' (2) (An asterisk marks "any symbol"). Then click the button 'Submit Query' (3):



4. This will bring up the resulting window with the list of matching entries (FEATURES AP-1). Click the link:

5. The query result will be displayed as the complete list of entries containing significant B-DNA features of the AP-1 site. Red arrows denote links to the recognition tools for this site (1) and graphical representation of the results for chosen property (2):



Comments and questions are welcome to Mikhail P. Ponomarenko (pon@bionet.nsc.ru).

1.2. Software

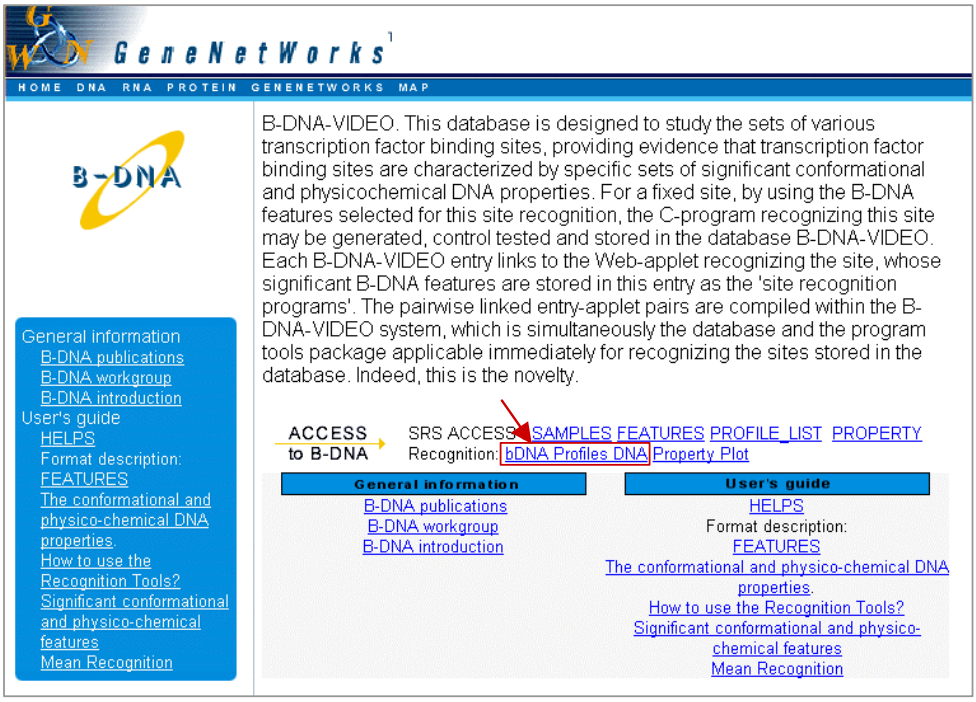
1.2.1. bDNA Profiles DNA. Feature-Recognition Tools

Release 2003

Program description:

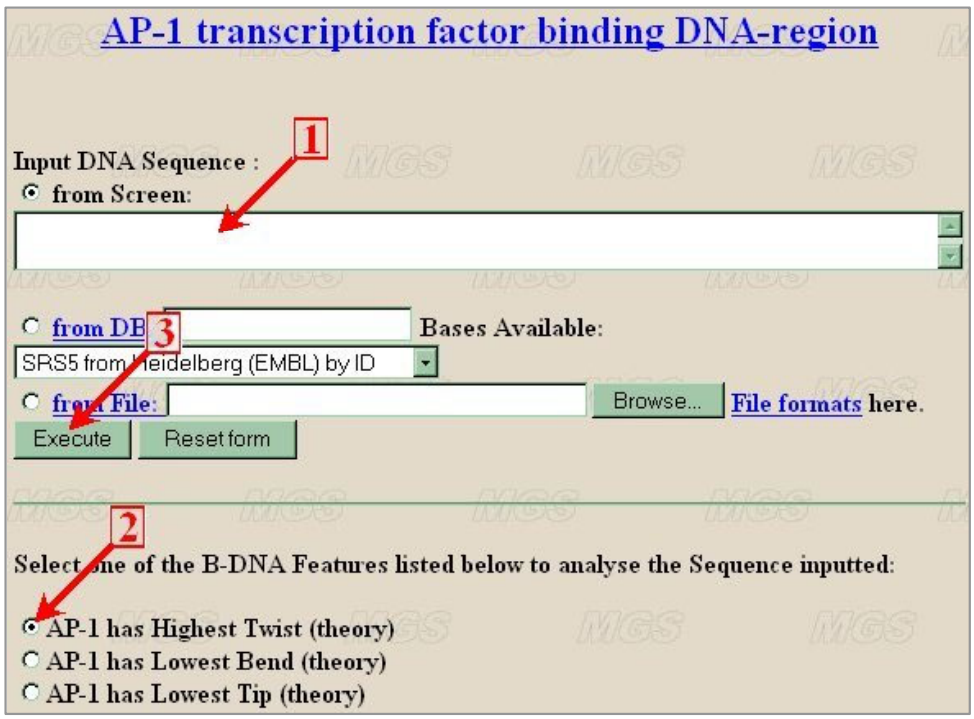
The transcription factor binding sites are characterized by specific sets of significant conformational and physico-chemical DNA properties. The site recognition programs are generated by the sets of significant conformational and physico-chemical DNA properties (DNA features).

Access to Recognition Tools:
http://www.domain.com/mgs/gnw/bdna/; then choose the link bDNA Profiles DNA:



The biological task that could be solved by using the program: DNA functional site recognition by most significant DNA conformational and physico-chemical properties.
Data input:

Input the DNA sequence of interest into the text window (1). Sequence should be in a plain text format (a, t, g, c in upper- or lower-cases, spaces or tabulation are accepted). The sequence to be input for analysis should be of maximal length of 32 kbp and minimal length of 100 bp.



Program options:

Select the necessary recognition program by clicking one of the radio buttons (2). Each one refers to a significant conformational or physico-chemical DNA property. Generalized (mean) recognition profile is calculated using all known significant conformational and physico-chemical features of the site, contained in **B-DNA features database**.

Program execution

Start the tools processing by clicking the button 'Execute' (arrow 3 in the figure above).

Data output

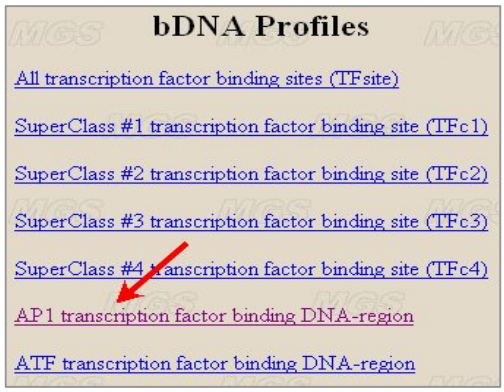
The tools output represents the profile of the score value. The positive peaks of this profile pinpoint to potential site recognized.

Example

Search for potential binding sites of the AP-1 transcription factor in promoter region of the human c-ets-1 protooncogene.

1) On the **B-DNA home page**, click the link bDNA Profiles DNA (see above, "**2) Access to Recognition Tools**" and illustrating figure).

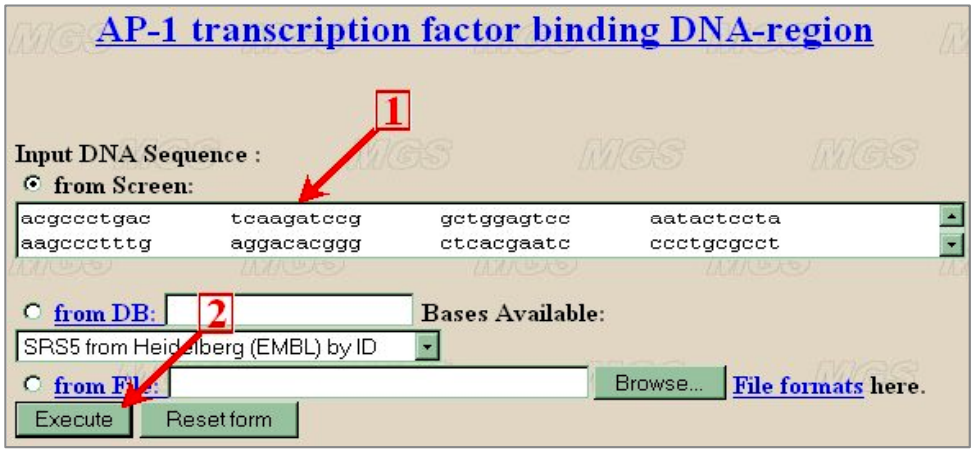
2) On the page 'bDNA Profiles', click the link AP1 transcription factor binding DNA-region:



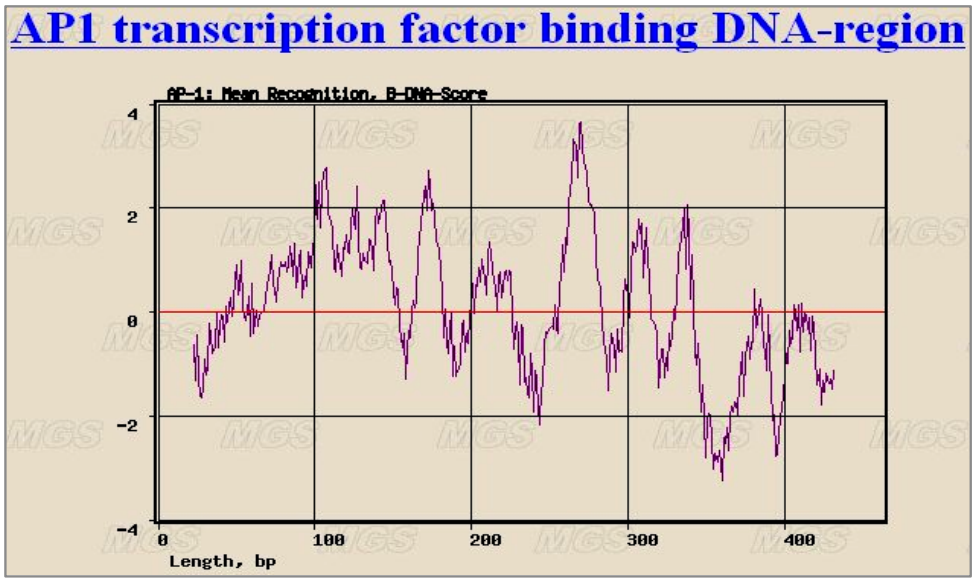
3) Input into the text-box (arrow 1) the following sequence:

```
acgccctgac  tcaagatccg  gctggagtcc  aatactccta
aagccctttg  aggacacggg  ctcacgaatc  ccctgcgcct
gcctgcacgc  tcgcttcatc  cacatgcctc  acgtcctgtg
tgtcagtctt  tgtggaatga  atgatgtaca  cgcacttgga
aaactatgct  gctactggga  gggggcgaga  gcgggtgacc
aagccctcaa  gaatgcgtgg  agaatcagac  ggactttccc
gaaacggtgg  aggcggcctg  tgcacccagc  ctgcacaccc
gctcccggcc  cttcccgccc  ctgcctgggc  tccgaggccc
ggggctccac  gcactgctcc  tccgcggtcg  gcccggcccc
gcgctcgcca  gcccttcctt  tcgctttggg  ccgcggcgga
gattggccgc  gtgctcggcc  cgcccccggc  tccccgccct
gccccgacgc  cccgcccctc  gct
```

(human c-ets-1 protooncogene sequence from -461 to +1 bp relatively transcription start site, GenBank; X65469) Use program options set by default. Click the button 'Execute' (arrow 2):



4) Data output window:



Comments and questions are welcome to Mikhail P. Ponomarenko (pon@bionet.nsc.ru).

1.2.2. DNA Property Plot

Release 2003

Program description:

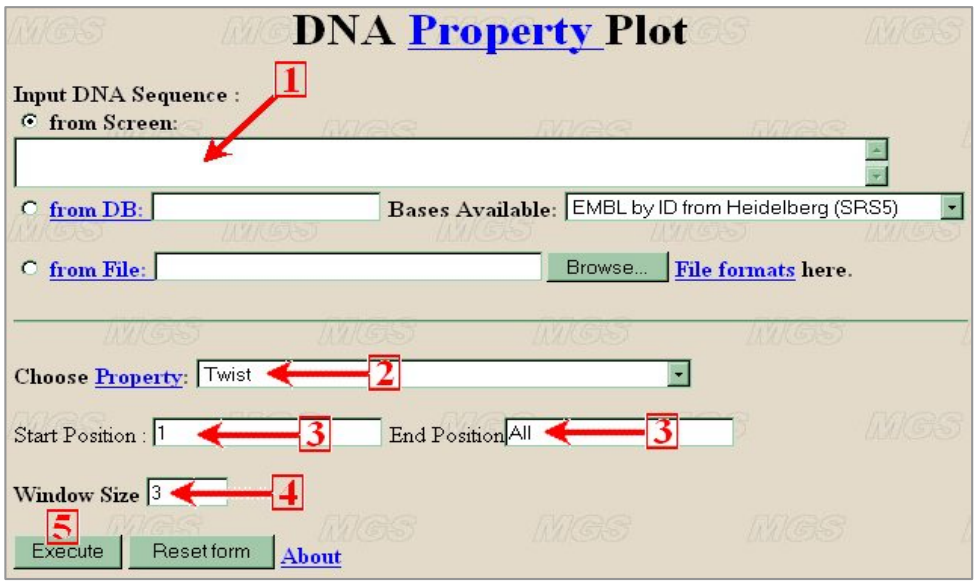
This program is designed to construct the profile of DNA conformational or physico-chemical properties.

Access to Property Plot:
http://www.domain.com/mgs/gnw/bdna/; then choose the link Property Plot.

The biological task that could be solved by using the program: investigation of DNA conformational or physico-chemical properties.

Data input

Input the DNA sequence of interest into the text-box 'Input DNA Sequence' (arrow 1 in the figure below). Sequence should be in plain textual format. Only the symbols a, g, c, t or A, T, G, C are permissible, line feeds and blanks are ignored. Sequence length should be not less than 10 bp and not longer then 1000 bp in order to view the results in details.



Program options:

a) Select the necessary DNA conformational physico-chemical property from the list by clicking the field 'Property' to choose the bDNA physico-chemical or conformational property of interest from the Property database (arrow 2).

b) Set the values in the fields 'Start Position' and 'End Position' to set the scope of the sequence region of interest (arrow 3).

c) Select the averaging window size value in the field 'Window Size' (arrow 4).

Program execution

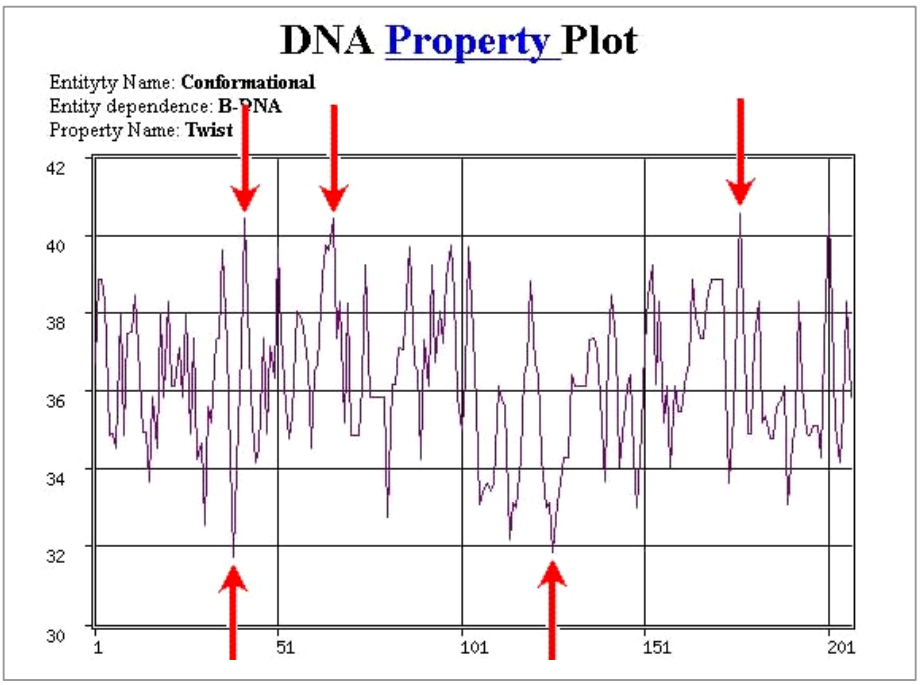
Start the program execution by clicking the button 'Execute' (5).

Data output

The output represents the profile of the selected physico-chemical property. The Y value on the plot corresponds to selected property magnitude at positions numbered on the X-axis.

Example

Data output for random sequence calculated for values of 'Twist' conformational property. Sequence length = 207, Window size = 3. Arrows mark regions displaying maximal and minimal twist values:



Comments and questions are welcome to Mikhail P. Ponomarenko (pon@bionet.nsc.ru).

CHAPTER 3. SELEX SYSTEM

1. SELEX Knowledge Base

Release 2003

Description:

SELEX Knowledge Base is designed for accumulation of experimental data on selected affinity-enriched sequences from different combinatorial libraries.

Access to SELEX Knowledge Base:
http://www.domain.com/mgs/gnw/selex/

SELEX Knowledge Base content:

SRS table	Description	Number of entries
SELEX_DB	a database on selected randomized DNA/RNA sequences	116
SELEX_BIB	a database on annotated in SELEX_DB papers	70
SELEX_TOOLS	a database on computer programs for recognition of sites accumulated in SELEX_DB	32

List of biological tasks that could be solved by using the SELEX Knowledge Base:

- to extract the sequences binding to a particular protein, which DNA or RNA binding site is annotated in **SELEX**;
- to extract sequence positions that are most important for binding of a particular DNA/RNA site;
- to recognise the site documented in **SELEX** within an arbitrary sequence in on-line mode;
- for a site accumulated in **SELEX**, to generate the C-program recognising this site according to appropriate weight matrix calculated by data stored in the database **SELEX**;
- planning of novel experiments applying **SELEX** technology.

SELEX Knowledge Base SRS tables formats:

SELEX_DB

Line code	Field name	Field description
ID	CardID	Identifier
AC	Accession	Accession number
CR	Referee	Reference on SELEX_DB database
OS	Specia	Organism
OC	Taxon	Taxon
NF	LigandName	Name of a ligand
KW	Key Words	Keywords
DA		Date of creation
DT		Date of the last update
FV		Release number
MN		Name of an entry
RF		Reference to the literature source
TE		Templates for amplification
EX		Type of an experiment
EC		Experimental conditions
NS		Number of sequences
AA		Aligned sequences as they are represented in the original paper
WA		Weight impact of the A nt at functionally important positions
WT		Weight impact of the T nt at functionally important positions
WG		Weight impact of the G nt at functionally important positions
WC		Weight impact of the C nt at functionally important positions
CN		Consensus sequence
DR		Links to the other databases
WW		Link to recognition tools
NM		Number of sequences in the set
SQ		Sequences
CC		Comments of an annotator concerning the functional role of a factor or peculiarities of consensus evaluation

SELEX_BIB

Line code	Field name	Field description
Id	CardID	Reference identifier
AU	Authors	Reference authors
TI	PaperTitle	Reference heading
SO	Source	Reference journal
VL	VolumeNumber	Reference journal volume
IS	Issue	Reference journal issue
YR	Year	Reference year

SELEX_TOOLS

Line code	Field name	Field description
Id	SiteID	Site identifier
MN	SiteName	Name of RNA/DNA site binding the ligand
SC	SC	Homology Score
HN		Author of the entry
DR SELEX_DB		Link to the SELEX_DB SRS Table
WW		Link to Recognition Tools
RM		Comments
AB		Sequence region
ST		Means, standard deviation, false negatives
NT		Means, standard deviation, false positives
C-CODE		C-code procedure for calculation of the data for an arbitrary DNA

Comments and questions are welcome to Mikhail Ponomarenko (pon@bionet.nsc.ru)

2. SELEX Profiles Program

Release 2003

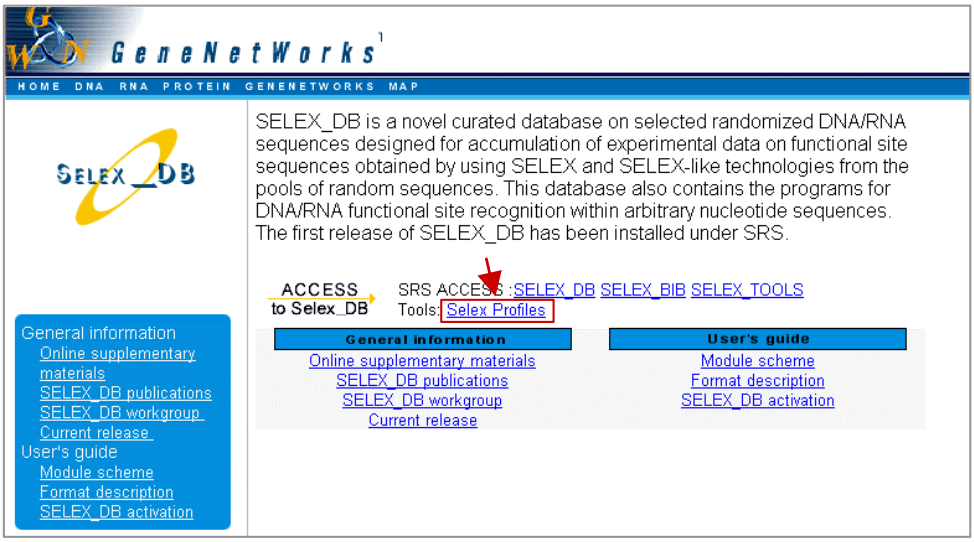
Program description:

The aim of the analysis is to determine positions of DNA in an arbitrary sequence binding to particular proteins, information about which is stored in the **SELEX_DB** database. The recognition procedure is based on 15 different variants of calculation, including homology score, matrix similarity, weighting consensus match scores, etc.

The biological task that could be solved by using the SELEX Profiles program:

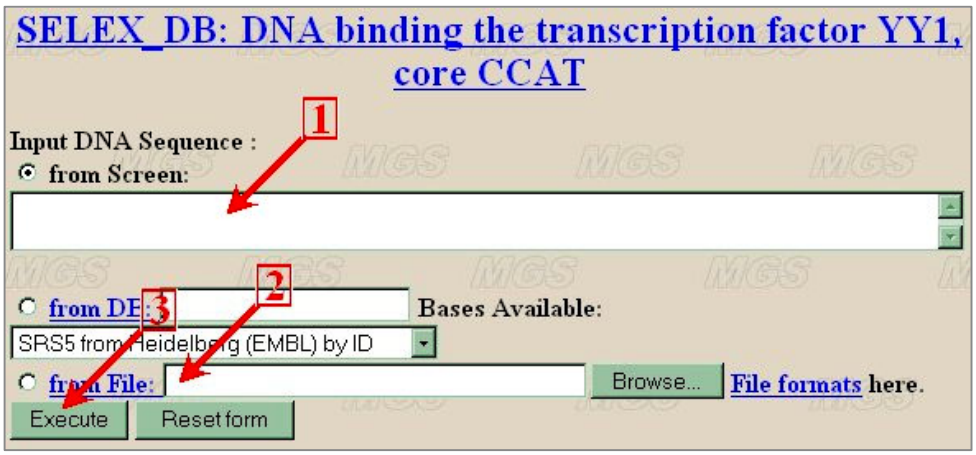
Searching for DNA sites binding to some proteins or transcription factors (in total 23 DNA binding sites) in an arbitrary sequence entered by a user.

Access to the SELEX Profiles program:
http://www.domain.com/mgs/gnw/selex/; link Selex Profiles.



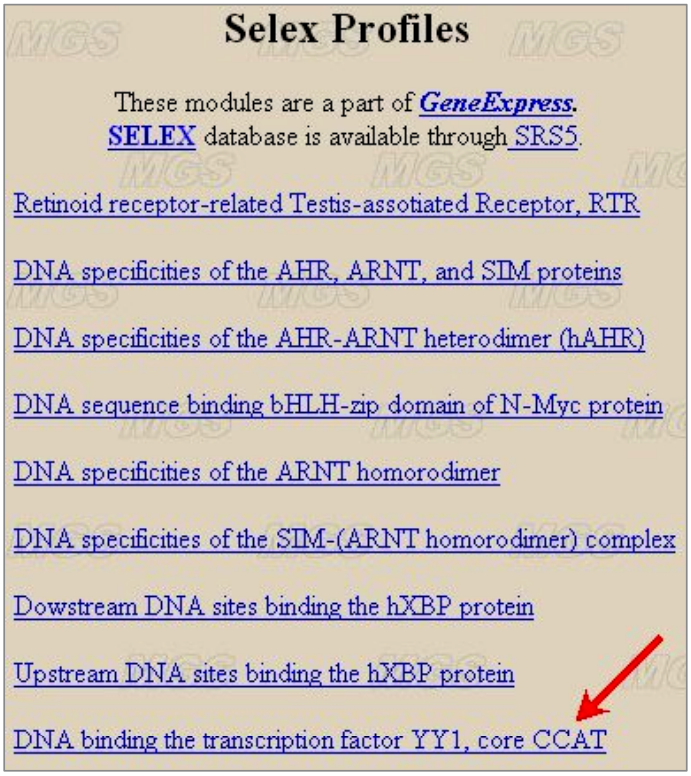
Data input

Input the data into the input textbox (1) shown in the figure below. The sequences should be entered in upper- or lower-case letters; line feeds and blanks are ignored.



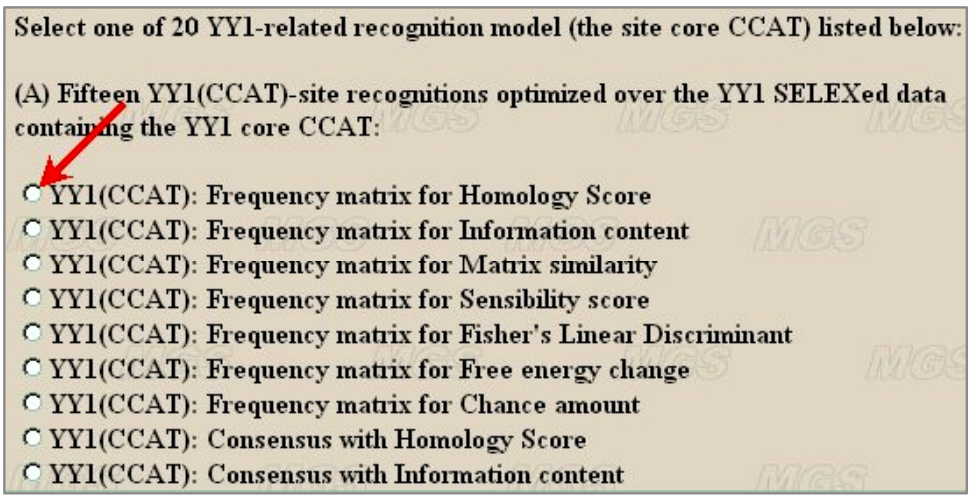
Program options:

By clicking the hyperlink, choose the protein or transcription factor from the list, to which you would compare your sequence with:



The sequence to be analysed could be entered in different ways: from the screen in the textbox (arrow 1 in the previous figure), by typing in from the keyboard or by cut & paste operation; or from file in 'FASTA' format: in this case, click the 'BROWSE' button and select the source file (arrow 2).

Choose one recognition model from the list by clicking the appropriate radio-button:



Program execution:

Click the button 'Execute' to execute the program (see arrow3 in the figure).

Data output:

The program output represents the recognition score profile.
Example

Let us analyse the fragment of the Moloney murine leukemia virus complete genome (EMBL:J02255) containing YY1 transcription factor.

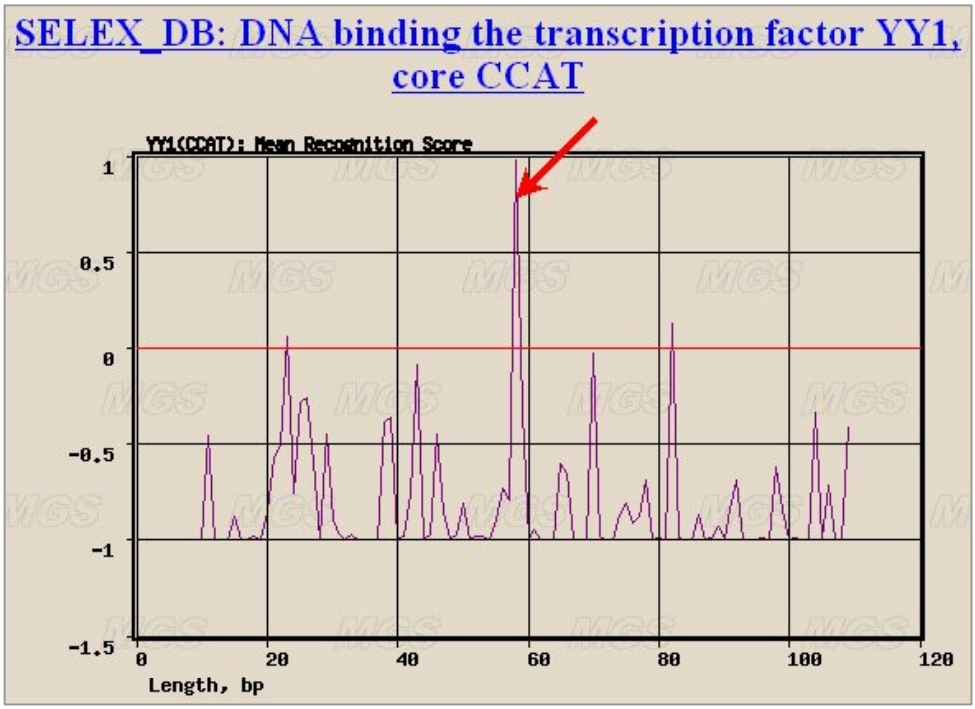
Select recognition model 'SELEX_DB: DNA binding the transcription factor YY1, core CCAT'.

For analysis, insert the fragment of this sequence in-between positions 7805-7924 into the text-box by choosing the option 'from Screen'.

Choose 'Mean Recognition Score' procedure.

Click the button 'Execute'.

The output window displays the YY-1 recognition score profile with the pick marked by arrow, which corresponds to the experimentally identified YY-1 transcription factor binding site (positions 7860-7868).



Comments and questions are welcome to Mikhail Ponomarenko (pon@bionet.nsc.ru)

CHAPTER 4. REGSCAN MODULE FOR DNA FUNCTIONAL SITE RECOGNITION

1. RegScan Databases

Release 2003

Databases description:

MATRIX Database accumulates oligonucleotide frequency matrices of transcription factor binding sites, the natural sequences of which are contained in the **SAMPLES** Database (see **Chapter 2.1.1. B-DNA Site Video Databases**).

Access to RegScan Databases:

http://www.domain.com/mgs/gnw/regscan/

link MATRIX for **MATRIX** database
link ALIGNED for **ALIGNED** database
link Property for **PROPERTY** database

Databases content:

SRS table	Description	Number of entries
MATRIX	Database of oligonucleotide frequency matrices of transcription factor binding sites	42
ALIGNED	A database of regulatory active genomic regions aligned by Gibbs Sampler. Investigation of transcriptional factors binding sites of eukaryotes.	43
PROPERTY	A distributed and intelligent database for the activities of the functional sites in DNA and RNA.	38

The biological task that could be solved by using RegScan Databases:

Recognition of functional sites by their specific oligonucleotide content.

SRS tables format:

MATRIX

Line code	Field name	Field description
MI	SiteID	This field indicates an identifier.
MN	SiteName	This field contains name of the transcription factor binding site.
HN	SCIENTIST	This field contains link to SCIENTIST database and indicates a contributor of the entry.
DR	SRS-link	This field contains the links to the supplementary databases installed under SRS (SAMPLES, ALIGNED, etc.).
WW	Web-link to Recognition Tools	This field contains the link to the Web-based tools implementing each C program documented within the entry to recognize the transcription factor binding site by its oligonucleotide frequency matrices within an arbitrary DNA sequence.
DP	Oligonucleotide Alphabet Length	This field indicates the length of oligonucleotide alphabet used by the recognition program given below in the field 'C-CODE'.
PV	Oligonucleotide Alphabet	This field indicates the oligonucleotide alphabet used by the recognition program given below in the field 'C-CODE'.
AB	Analyzed Region	This field indicates the DNA region, for which the weight matrix was constructed. Positions are given in bp relative to the first position of the site core multiply aligned by the Gibbs potential function (Lawrence C., 1994, Comput. Chem., 18, 255-258) (see ALIGNED database).
ST	Means, Standard Deviation, False Negatives for Control Sequences	This field provides the mean value of the Score(first estimate), the standard deviation of this mean value (second estimate), and the I type error rate (false negatives), which have been determined on the independent control sequences of the site by the program given below in the field 'C-CODE'.
NT	Means, Standard Deviation, False Positives for Random Sequences	This field provides the mean value of the Score(first estimate), the standard deviation of this mean value (second estimate), and the II type error rate (false positives), which have been determined on the random sequences by the program given below in the field 'C-CODE'.
FG	Graphical Representation of Test Results	This field represents the links to the graphical representation of the results obtained by the program, given below in the field 'C-CODE', over the independent control sequences of the site.
C-CODE	C-CODE	Recognition program by the Partial recognition Score or Mean recognition Score in the C language of the ANSI standard. Each C program documented within the MATRIX entry has a check-box in the MENU window of the Recognition Tools (field Web-link to Recognition Tools)

ALIGNED

Line code	Field name	Field description
FI	SampleID	brief name of the sample
NM	SampleName	name of the sample and explanation of its biological meaning
OR	Organization	title of the organization
AU	Author	the first and last name of the author
DA	Date	date of creation
LU	LustUpdate	the date and the author of the last update/modification (Last update); in case of no modification, the date and the authors name are from the DA and AU fields
FV	FormatVersion	Number of the Format Version
ST	SiteTypeDescription	ST {X,Y} [Left_border,Right_border] Point; Description; Factor_name. - (Structure) description of the site (not a particular site, but type of the sites); several ST fields are allowed; X=0….9 - the number of site batch (sites are joined into batches if they in the aggregate they represent an integral structure-function unit); Y=0….9 – the number of the site in a site batch; [Left_border,Right_border] – The position of the site in case this particular site type possesses a fixed location relative to a specific point, for example, relative to transcription start. If no, put nothing; Point – the point relative to which the site location is fixed; If the site has no fixed position, put nothing Description - conventional name of the site (abbreviated name if available) Factor_name - a name of the site-binding factor
WS		reference to 'SAMPLES' database
WC		reference to 'CONSENSUS' database
WM		reference to 'MATRIX' database
ID	CardID	identifier of an entry
AC	CardAC	accession number of an entry
OS	Specia	organism specia
OC	Taxon	organism taxon
DR	SourceDatabase	reference to the source database
FT	FeatureTable	This field contains positions of site described in ST field. The format is as follows: FT {X,Y} [left;right]; Method X = 0…9 → number of the site batch Y = 0…9 → number of the site in the site batch [left;right] → positions from the start of the sequence Method = EXP (experimental), GBS (Gibbs Sampler alignment), RCG (Recognition Group alignment;
SQ	Sequence	nucleotide sequence

PROPERTY

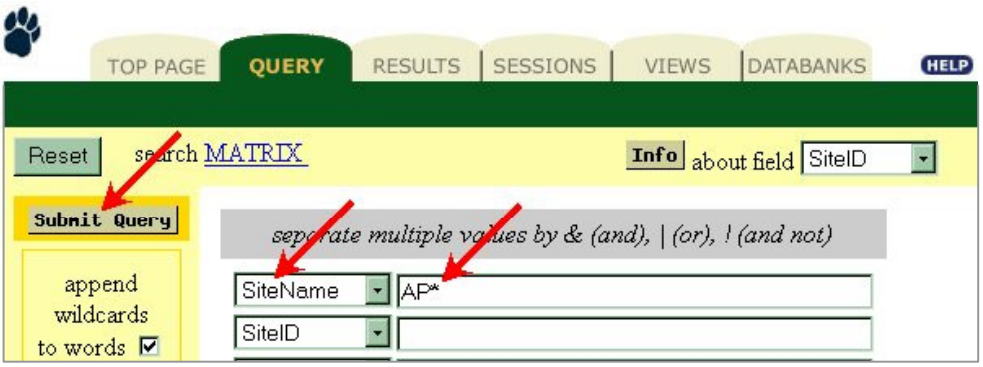
See **Chapter 5. ACTIVITY Databases.**

Example of query:

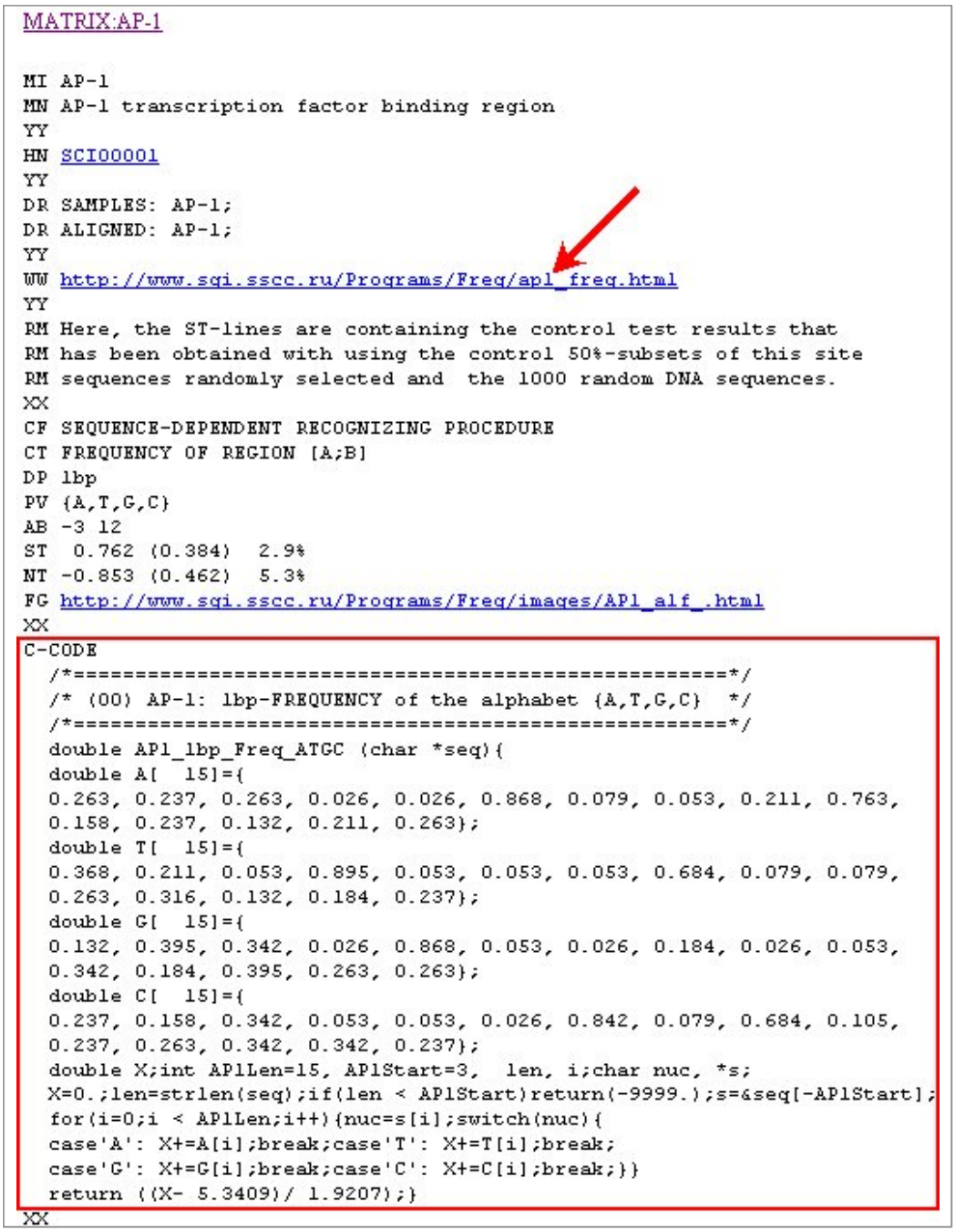
To extract C-code for the AP-1 transcription factor recognition from the **MATRIX Database**, or apply WWW-accessible **AP-1 Recognition Tool** to an arbitrary sequence of interest.

1. Click the button 'Search':

2. Select the fields to be searched for from the list. You will need the field 'SiteName'. Type term to be searched in the text window: 'AP*'; and click the button 'Submit Query':



3. This will bring up the resulting window with the list of matching entries (MATRIX: AP-1):



```
MATRIX:AP-1

MI AP-1
MN AP-1 transcription factor binding region
YY
HN SCI00001
YY
DR SAMPLES: AP-1;
DR ALIGNED: AP-1;
YY
WW http://www.sgi.sscc.ru/Programs/Freq/ap1_freq.html
YY
RM Here, the ST-lines are containing the control test results that
RM has been obtained with using the control 50%-subsets of this site
RM sequences randomly selected and  the 1000 random DNA sequences.
XX
CF SEQUENCE-DEPENDENT RECOGNIZING PROCEDURE
CT FREQUENCY OF REGION [A;B]
DP lbp
PV {A,T,G,C}
AB -3 12
ST  0.762 (0.384)  2.9%
NT -0.853 (0.462)  5.3%
FG http://www.sgi.sscc.ru/Programs/Freq/images/AP1_alf_.html
XX
C-CODE
  /*=======================================================*/
  /* (00) AP-1: lbp-FREQUENCY of the alphabet {A,T,G,C}  */
  /*=======================================================*/
  double AP1_lbp_Freq_ATGC (char *seq){
  double A[  15]={
  0.263, 0.237, 0.263, 0.026, 0.026, 0.868, 0.079, 0.053, 0.211, 0.763,
  0.158, 0.237, 0.132, 0.211, 0.263};
  double T[  15]={
  0.368, 0.211, 0.053, 0.895, 0.053, 0.053, 0.053, 0.684, 0.079, 0.079,
  0.263, 0.316, 0.132, 0.184, 0.237};
  double G[  15]={
  0.132, 0.395, 0.342, 0.026, 0.868, 0.053, 0.026, 0.184, 0.026, 0.053,
  0.342, 0.184, 0.395, 0.263, 0.263};
  double C[  15]={
  0.237, 0.158, 0.342, 0.053, 0.053, 0.026, 0.842, 0.079, 0.684, 0.105,
  0.237, 0.263, 0.342, 0.342, 0.237};
  double X;int AP1Len=15, AP1Start=3,  len, i;char nuc, *s;
  X=0.;len=strlen(seq);if(len < AP1Start)return(-9999.);s=&seq[-AP1Start];
  for(i=0;i < AP1Len;i++){nuc=s[i];switch(nuc){
  case'A': X+=A[i];break;case'T': X+=T[i];break;
  case'G': X+=G[i];break;case'C': X+=C[i];break;}}
  return ((X- 5.3409)/ 1.9207);}
XX
```

This entry displays the frequency matrices for AP-1 site, including C-code for recognition program (marked with red rectangle) and recognition tools for this site. The link to the entry is marked by arrow:

Comments and questions are welcome to Mikhail P. Ponomarenko (pon@bionet.nsc.ru).

2. Software For Transcription Factor Binding Site Recognition

2.1. Program BinomSite

Release 2003

Program description:

This software was developed to search for potential transcription factor binding sites by using binomial criterion for estimation of similarity score between the regions of the sequence analysed and the sequences of transcription factor binding sites.

Access to BinomSite:
http://www.domain.com/mgs/gnw/regscan/ link BinomSite

Biological task that could be solved by using BinomSite:

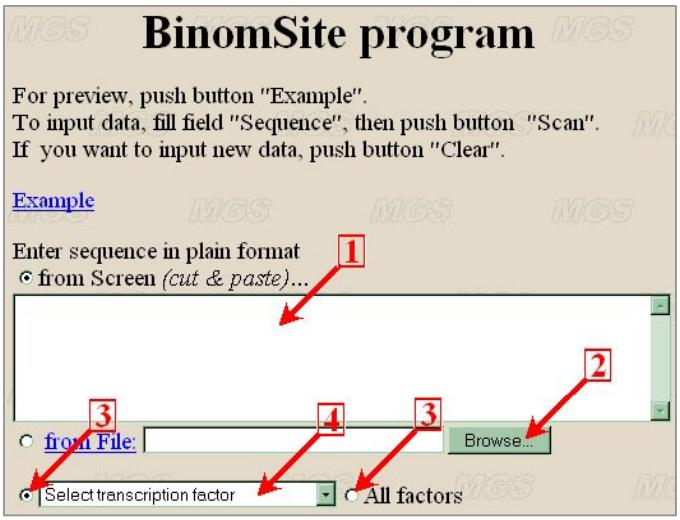
This program searches for potential transcription factor binding sites in the arbitrary sequence entered by a user.

Data input

To search for transcription factor binding sites, do the following:

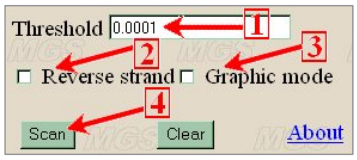
1) Enter the nucleotide sequence into the text-box (red arrow 1 in the figure below). It is possible to enter the sequence from clipboard or from file by using the standard dialog window, which pops up after clicking the screen button 'Browse' (arrow 2). Note that the sequence should contain only the symbols a, c, t, g, A, C, T, G and the length should range from 30 to 100 000 bp. If the sequence is entered from file, it should be of FASTA format.

2) By using the radio buttons (arrows 3) and drop-down list (arrow 4) located below the text-box 'Sequence', select the mode of searching for sites. It is possible to search for a single transcription factor binding site or for all of them contained in the list. If a single transcription factor is searched for, select its name in the list:



Program options:

Enter the significance level value into the text box 'Threshold' (red arrow 1 in the figure. The threshold value 10^{-4} = 0.0001 is set there). It is recommended to set the threshold ranging from 10^{-12} to 10^{-4}. The more is the threshold value, the more regions of the sequence analysed will be annotated by the program as true transcription factor binding sites. In case the threshold value is set near by 1 (or more), each position will be annotated as a site. In case the threshold value is equalling or less than zero, then the program will not annotate any site at all.



You may if necessary select for processing the complementary DNA sequence strand by clicking the check-box 'Reverse strand' (arrow 2). If this box is not checked, the search will be done in the sequence entered, "as is".

Choose the mode of the data output. For displaying the results graphically, click the check-box 'Graphic mode' (arrow 3). If this option is off, then the result will be output in a textual form.

Program execution

Start the program by clicking the button 'Scan' (see arrow number 4 in figure given above).

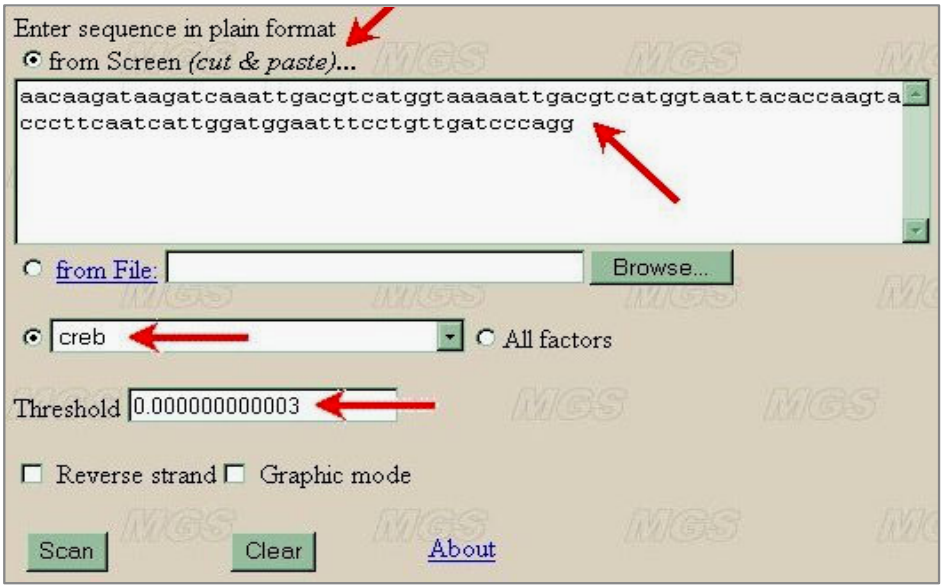
Data output

See the example below.

Example

As an example, which is invoked by clicking the appropriate hyperlink in the BinomSite starting page, the sequence of steroid-responsive region annotated in TRRD with the number 'P00082' and located in-between positions (-160; -60) relatively transcription start site was analysed. The data input is illustrated for searching for a single transcription factor binding site in direct strand. The example sequence has the CREB transcription factor binding site (`'caaatTGACGTCAtgg'`; TRRD number S173) at position -147 relatively transcription start site. This position corresponds to 14-th position in the regulatory region sequence, which was entered for searching in it potential CREB binding sites.

In the figure, the data input is illustrated for searching for a single transcription factor binding site, with the threshold $3 \cdot 10^{-12}$, in direct strand, with the textual data output:

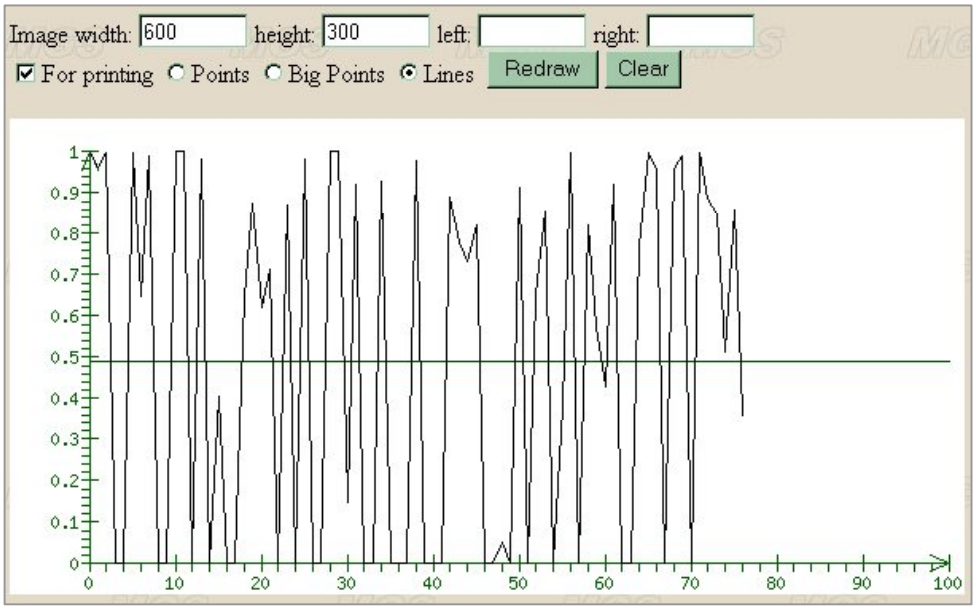


The textual data output is illustrated in the next figure. The program BinomSite displays a list of positions with potential transcription factor binding sites for a particular transcription factor, the name of this transcription factor, DNA strand ('+' or '-', that is, direct or complementary DNA chain), as well as the fragments of the sequence ordered that correspond to potential transcription factor binding sites in positions indicated:

Name	Position	Strand	Site
creb	12	+	ataagatcaaattgacgtcatggtaaaaa
creb	30	+	catggtaaaaattgacgtcatggtaattac

As seen, the program BinomSite has detected a potential site with the sequence '`ACGTCA`' at position 12, with the length of 29 nt and the sequence '`ataagatcaaattgACGTCAtggtaaaaa`'. This sequence includes the real CREB binding site annotated in TRRD. In the sequence analysed, the BinomSite program has detected one more potential CREB binding site at position 30.

The graphic data output is illustrated in the next figure. The program BinomSite displays the plot of similarity score at each position of the sequence analysed with the CREB transcription factor binding sites. The green horizontal line marks the average similarity score estimated for all positions of a sequence:



Comments and questions are welcome to Mikhail Pozdniakov (mike@bionet.nsc.ru)

2.2. Program MMSite

Release 2003

Program description:

This program realizes a novel approach for searching for potential transcription factor binding sites in the arbitrary sequence entered by a user.

Access to MMSite:
http://www.domain.com/mgs/gnw/regscan/ link MMSite

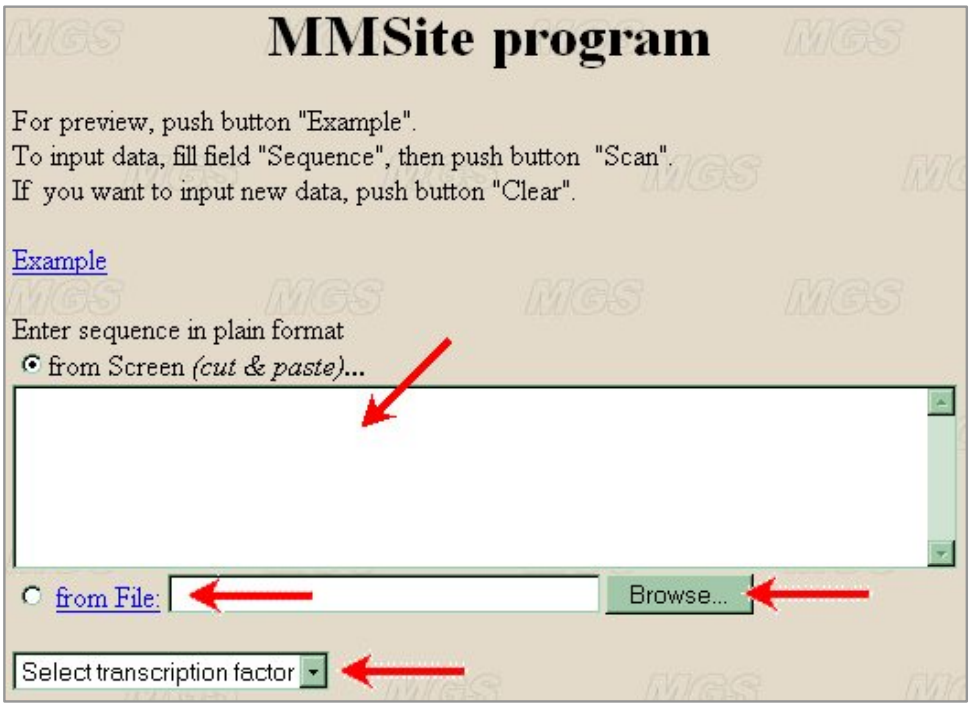
Biological task that could be solved by using MMSite:

This software was developed for searching for potential transcription factor binding sites in the arbitrary sequence entered by a user. User may enter one of possible four methods of transcription factor binding site recognition. This enables to choose appropriate recognition method for each type of transcription factor binding sites.

Data input:

To search for transcription factor binding sites, do the following:

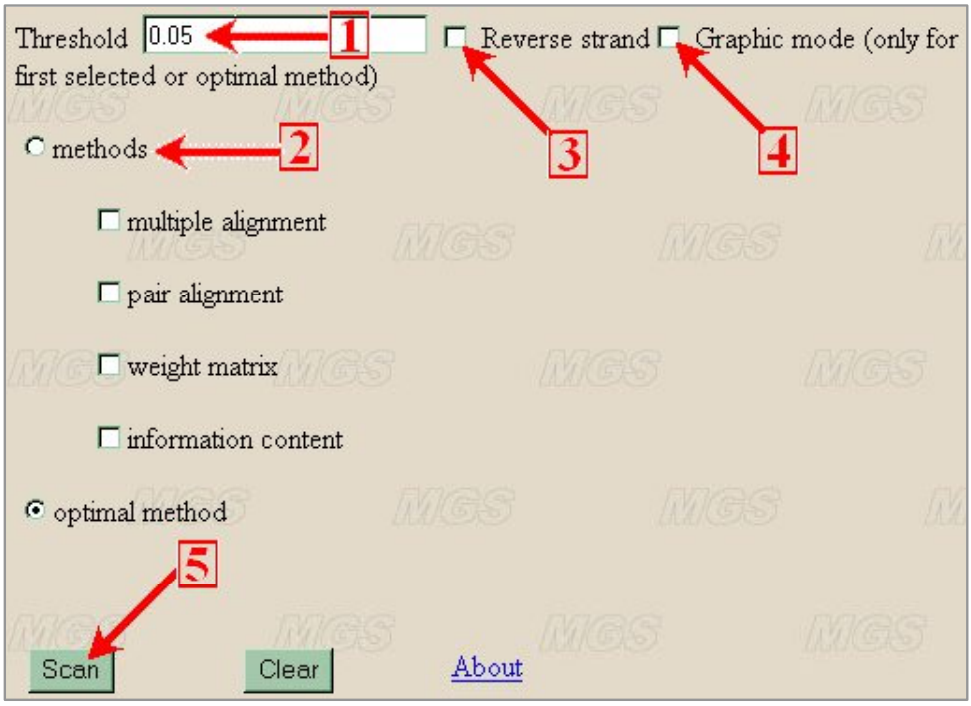
- Enter the nucleotide sequence into the text-box. It is possible to enter the sequence from clipboard or from file by using the dialog window 'Browse'. Note that the sequence should contain only the symbols a, c, t, g, A, C, T, G and the length should range from 30 to 3000 bp. If the sequence is entered from file, it should be of 'FASTA' format. The text-box for data input is marked by red arrow 1 in the figure.
- Choose the transcription factor that you want to recognise from the drop-down list 'Select transcription factor'. This list for choosing the name of transcription factor is also marked by red arrow in the figure.



Program options:

Select significance level value in the drop-down list 'Threshold' (arrow 1). It is recommended to set the threshold ranging from 10^{-5} to 10^{-2}. The more is the threshold value, the more regions of the sequence analysed will be annotated by the program as true transcription factor binding sites. In case the threshold value is set near by unit (or more), each position will be annotated as a site. In case the threshold value is equalling to zero or less than zero, then the program will not annotate any site at all. Click the check-box 'Methods' (arrow 2) for choosing the method or a combination of methods that you want to apply or the option 'Optimal method' that enables automated choosing of the best method for particular site of your interest. If you click the check-box 'Optimal method',

the program automatically selects one of four possible methods that will at best discriminate particular transcription factor binding sites indicated out of random sequences.



Program execution:

Start the program by clicking the button 'Scan' (see arrow number 5 in figure given above).

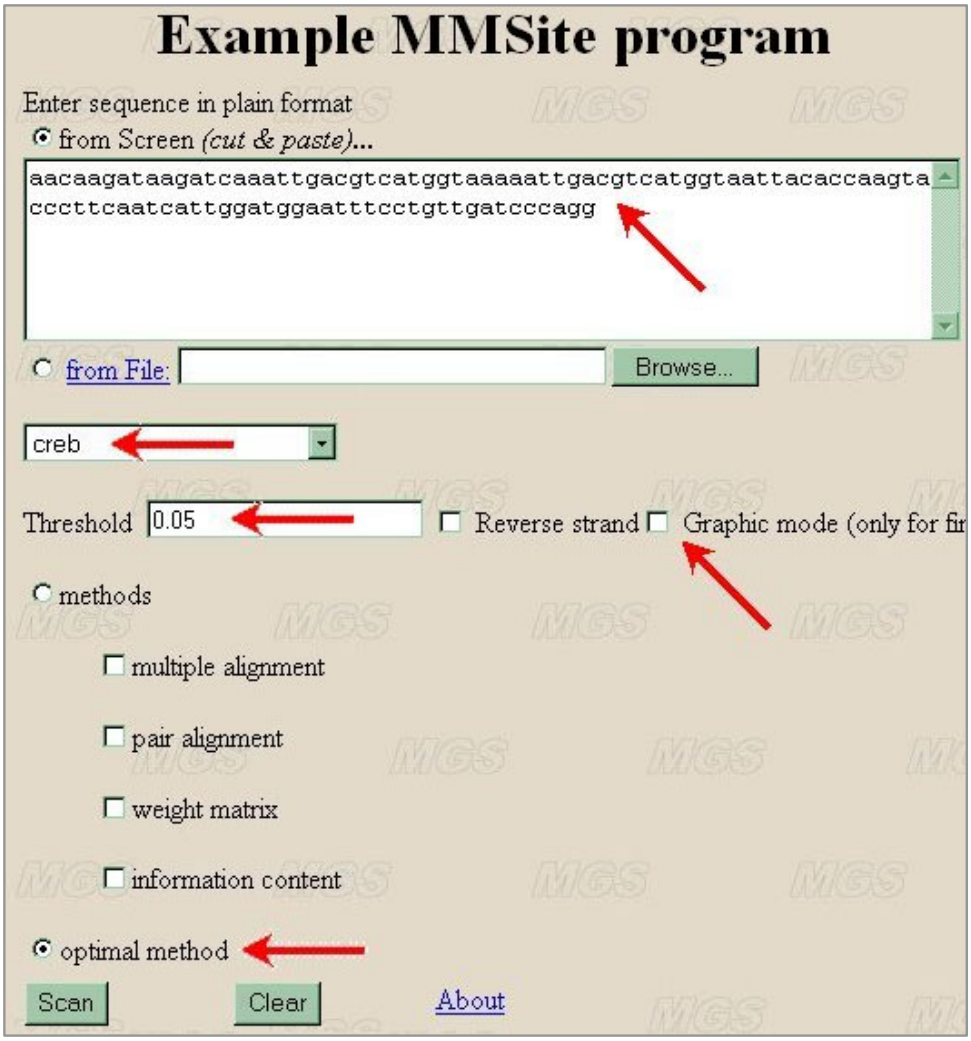
Data output:

1) Select the strand of DNA sequence to be analysed by clicking the check-box 'Reverse strand' (arrow 3). If this box is not checked, the search will be done in the sequence entered, otherwise, in complementary, reverse DNA strand.
2) Choose the mode of the data output. For displaying the results graphically, click the button 'Graphic mode' (arrow 4).

Example:

An example of data input, selecting the site type and recognition method of the program **MMSite**. As the example, the sequence of steroid-responsive region annotated in TRRD with the number 'P00082' and located in-between positions (-160; -60) relatively transcription start site.

In the figure, the data input is illustrated for searching for a single transcription factor binding site, with the threshold 0.05, in direct strand, with the textual data output:



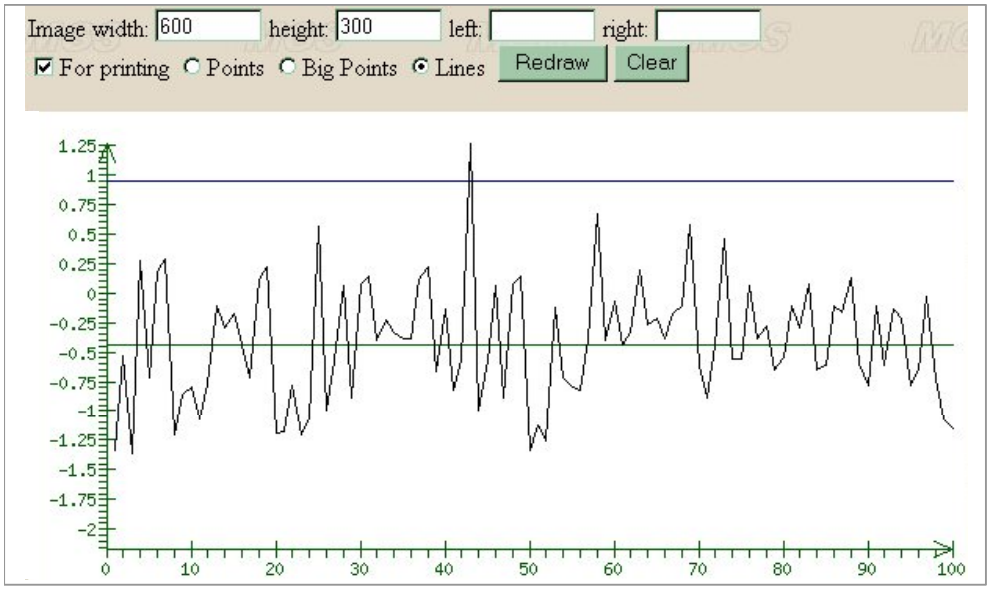
Below, the textual data output of the program 'MMSite' is illustrated. It displays the list of positions where potential transcription factor binding sites are found, the name of potential transcription factor binding sites, DNA strand ('+' or '-', that is, direct or inverse DNA), as well as the regions in particular sequence corresponding to potential binding sites in positions indicated. These regions are aligned with the sample of the sites of particular type, so, the symbol of deletion (-) may occur in the sequences of potential transcription factor binding sites '-'.

The data output by the program 'MMSite' searching for potential transcription factor binding sites in the sequence shown in figure below. The program 'MMSite' has found two potential 'CREB' transcription factor binding sites. The first of these sites is annotated in TRRD database.

As an example, we have performed an analysis of steroid-responsive region annotated in TRRD with the number 'P00082' and with experimentally detected 'CREB' transcription factor binding site ('S173', '`caaatTGACGTCAtgg`') at position -147 relatively transcription start site. Note that this position corresponds to position 14 of the sequence entered in above figure. The program 'MMSite' has detected potential 'CREB' binding site at position 21, which is a part of the sequence of real 'CREB' binding site annotated in TRRD. Also, the program has recognised three more potential 'CREB' binding site: at position 39, 53 and 64, which are not annotated in TRRD.

```
P-Level: 0.050000
Name                Position   Strand      Site                Method
creb                22            +      -cgtca      multiple alignment
creb                39            +      acgtca      multiple alignment
creb                53            +      acacca      multiple alignment
creb                64            +      ccttca      multiple alignment
```

Graphic mode of resulting data output. In the Figure, the plot illustrates the results of analysis by the program 'MMSite' of the example sequence shown in Figure above. The program displays similarity Score at each position of the sequence of interest with the 'CREB' transcription factor binding sites. The blue horizontal line marks the threshold set by user. The green horizontal line marks the average similarity Score estimated for all positions of a sequence. If the region displays similarity Score peak exceeding the blue line, this region is considered as potential binding site of a certain transcription factor.



Comments and questions are welcome to Mikhail Pozdniakov (mike@bionet.nsc.ru)

2.3. Programs for DNA Functional Sites Recognition

Release 2003

Programs description:

RegScan module contains programs for transcription factor binding sites recognition.

Access to the programs for DNA regulatory regions and functional sites recognition:

http://www.domain.com/mgs/gnw/regscan/;

- Click the link FreqMatrix for searching for the sites by the method of frequency matrices.
- Click the link Activity for searching for the sites by the method of binding activity of the site.
- Click the link bDNA for searching for the sites by the method of physico-chemical and conformational properties.

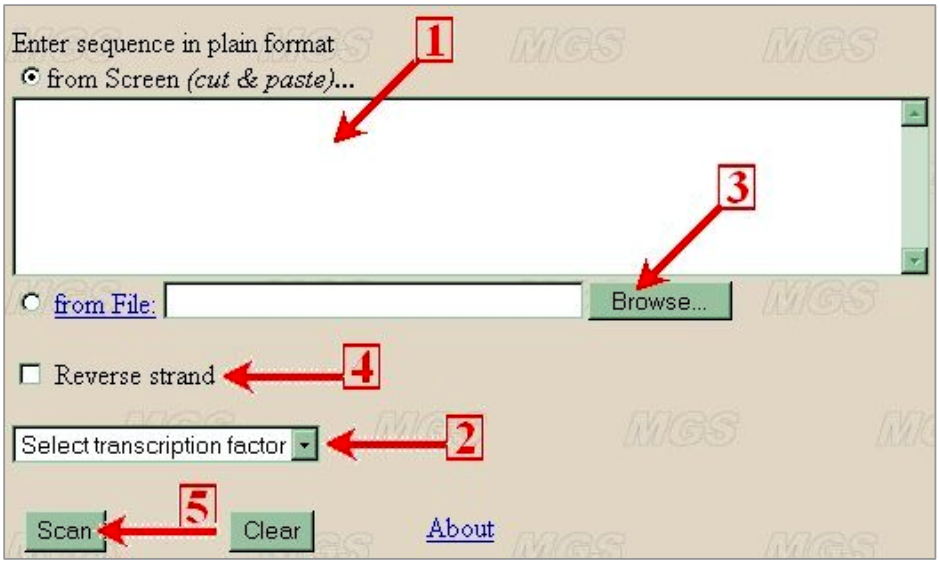
Biological task that could be solved by using the programs:

Recognition of transcription factor binding sites in DNA sequences by using 3 different methods, based on data on the frequency matrices, physicochemical and conformational properties or binding activity of a site.

Data input:

Input window of **RegScan Recognition Module** contains input textbox for sequence (1) of interest and the list of available regulatory regions or functional sites to recognise (2). It is also possible to input sequence from a text file by clicking the 'BROWSE' button (3).

Essential notes: sequences should be entered in FASTA format using upper- or lower-case letters; line feeds and blanks are ignored; Sequence length should not exceed 32000 bp and should not be less than 50 bp.



Programs options:

The search for transcription factor binding sites could be provided by the reverse strand, in this case, click the check box 'Reverse strand' (4).

Programs execution:

Click the button 'Scan' (5).

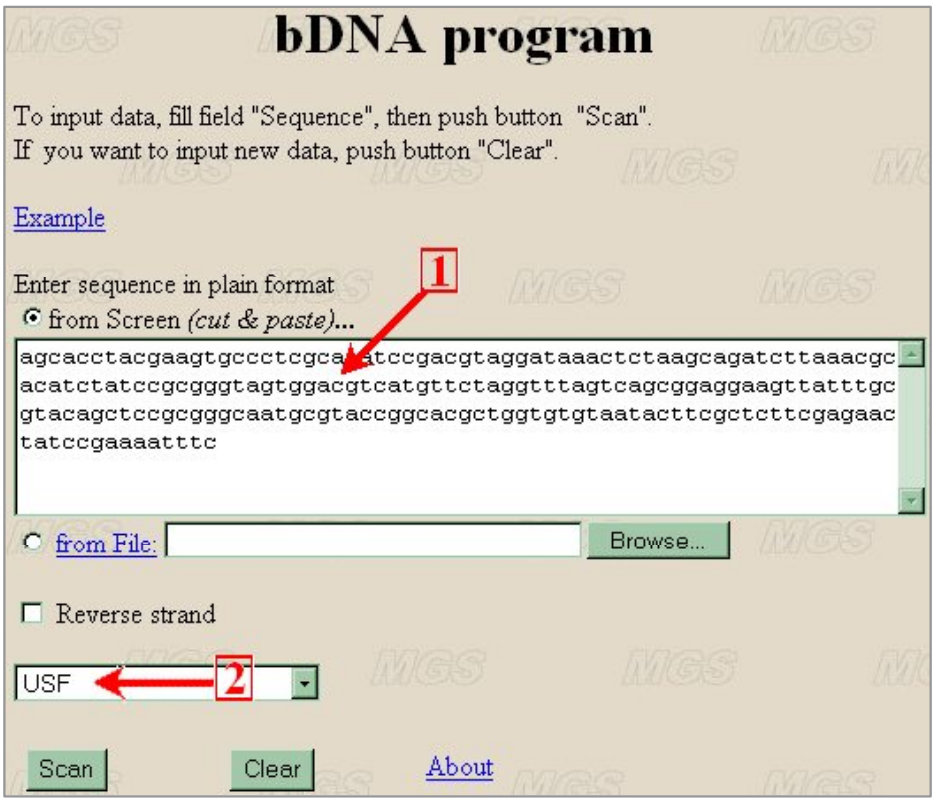
Data output:

The tools output represents the profile of the Score value. The positive peaks of this profile pinpoint to the potential site recognized.

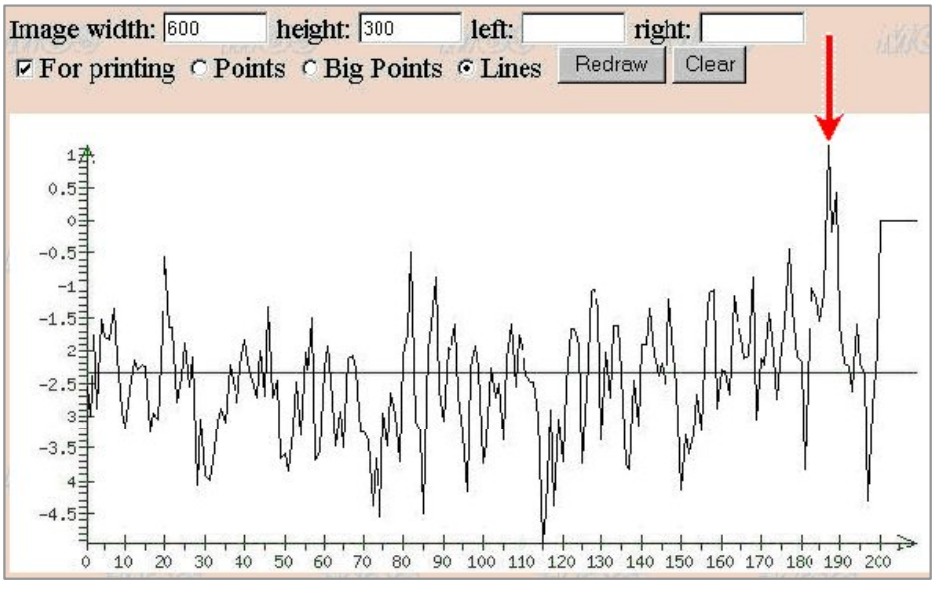
Example: Recognising of USF binding site in the sequence of interest

An example of recognising USF binding site in the sequence of interest by using one of the recognition program, **bDNA**.

Into the text-box, input the random sequence with the length 200 bp (1).
Select from the list (2) transcription factor USF:



After the program execution (by clicking the button 'Scan') view the graphic mode of resulting data output:



In the figure, the plot illustrates the results of analysis by the program **bDNA** of the example sequence shown in figure above. The program displays similarity Score at each position of the sequence of interest with the UTF binding sites. The horizontal line marks the average similarity Score estimated for all positions of this sequence. If the region displays similarity Score peak, this region is considered as potential binding site of a certain transcription factor (marked by arrow).

Comments and questions are welcome to Mikhail P. Ponomarenko, (pon@bionet.nsc.ru)

2.4. RGSiteScan Program

Release 2003

Program description:

This program is designed to predict the potential binding sites of transcription factors in the target DNA sequence.

Access to RGSiteScan program:

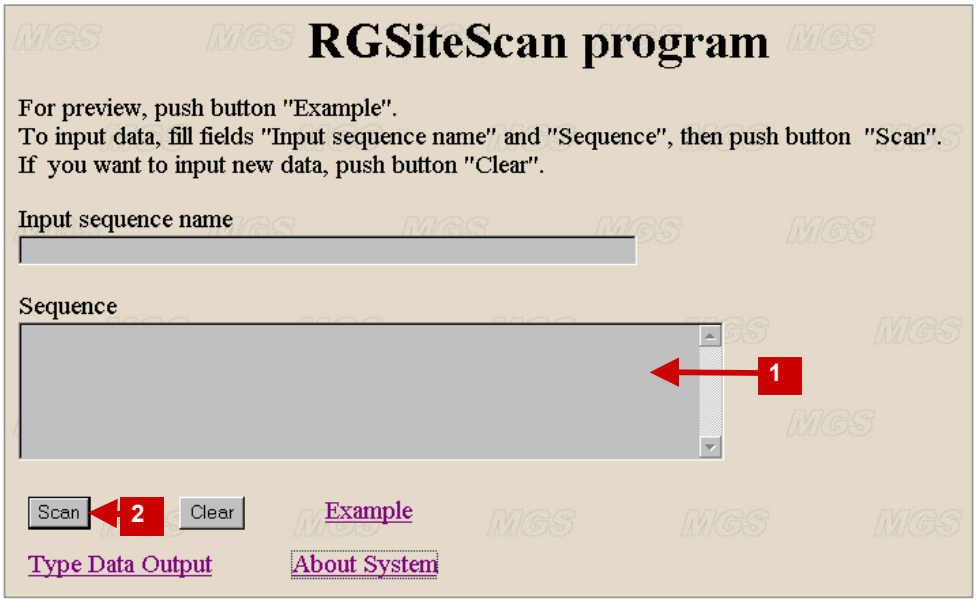
http://www.domain.com/mgs/gnw/regscan/ link RGSiteScan

List of biological tasks that could be solved by using the RGSiteScan program:

- Recognition of the potential binding sites of transcription factors in the target DNA sequence.

Data input

The sequence to be analysed should have the plain textual format (a, t, g, c, printed in lower case, without gaps) and the sequence length should be less than 32000 bp.
Enter the sequence to be analysed into the text-box 'Sequence' (1).



Program options:

It is not necessary to install parameters.

Program execution

The program is executed by clicking the button 'Scan' (2).

Data output

The results of the program execution are presented as the oligonucleotides that correspond to potential transcription factor binding sites. The names of these factors, as well as positions of respective oligonucleotides are also displayed.

Example

The task is to predict the potential binding sites of transcription factors in GGMYHE standard; DNA; VRT; (promoter region:1-2211 BP)

```
      ****************************************
      * Prediction of  potential  binding sites *
      * of  transcription   factors   in  given *
      * DNA-sequence on the base of recognition *
      *                groups.                 *
      ****************************************

The name of sequence: GGMYHE     standard; DNA; VRT; (promoter region:1-2211
The number of all predicted binding sites =     49

     Name       Position    Site
      AR          17        ttttct
      AR          92        ttttct
     COUP         98        tgacct
      AR         166        ttttct
    ICSBP        185        agtttc
     APF         196        gttaag
      AR         200        agaaaa
      AR         207        tgctct
      AR         217        agcaca
      AR         231        aaaaca
      AR         297        agaaaa
      AR         300        aaaaca
      AR         380        tgtcct
      AR         416        ttttct
      AR         427        aaaaca
     APF         450        attacc
    Pit-1        465        tatacat
     AP-1        473        tgactga
      AR         479        agaaaa
```

As a result, positions of 49 potential binding sites of transcription factors were predicted.

Comments and questions are welcome to
Yury Kondrakhin (kondrat@bionet.nsc.ru).

Example

The task is to predict the potential binding sites of transcription factors in GGMYHE standard; DNA; VRT; (promoter region:1-2211 BP)

```
      ****************************************
      * Prediction of  potential  binding sites *
      * of  transcription   factors   in  given *
      * DNA-sequence on the base of recognition *
      *                groups.                 *
      ****************************************

The name of sequence: GGMYHE     standard; DNA; VRT; (promoter region:1-2211
The number of all predicted binding sites =     49

     Name       Position    Site
      AR          17        ttttct
      AR          92        ttttct
     COUP         98        tgacct
      AR         166        ttttct
    ICSBP        185        agtttc
     APF         196        gttaag
      AR         200        agaaaa
      AR         207        tgctct
      AR         217        agcaca
      AR         231        aaaaca
      AR         297        agaaaa
      AR         300        aaaaca
      AR         380        tgtcct
      AR         416        ttttct
      AR         427        aaaaca
     APF         450        attacc
    Pit-1        465        tatacat
     AP-1        473        tgactga
      AR         479        agaaaa
```

As a result, positions of 49 potential binding sites of transcription factors were predicted.

Comments and questions are welcome to
Yury Kondrakhin (kondrat@bionet.nsc.ru).

2.5. RecGroup program

Release 2003

Program description:

The program is designed for searching for recognition groups of oligonucleotides of fixed length that are specific for transcription factor binding sites.

Access to RecGroup program:

http://www.domain.com/mgs/gnw/regscan/ link RecGroup

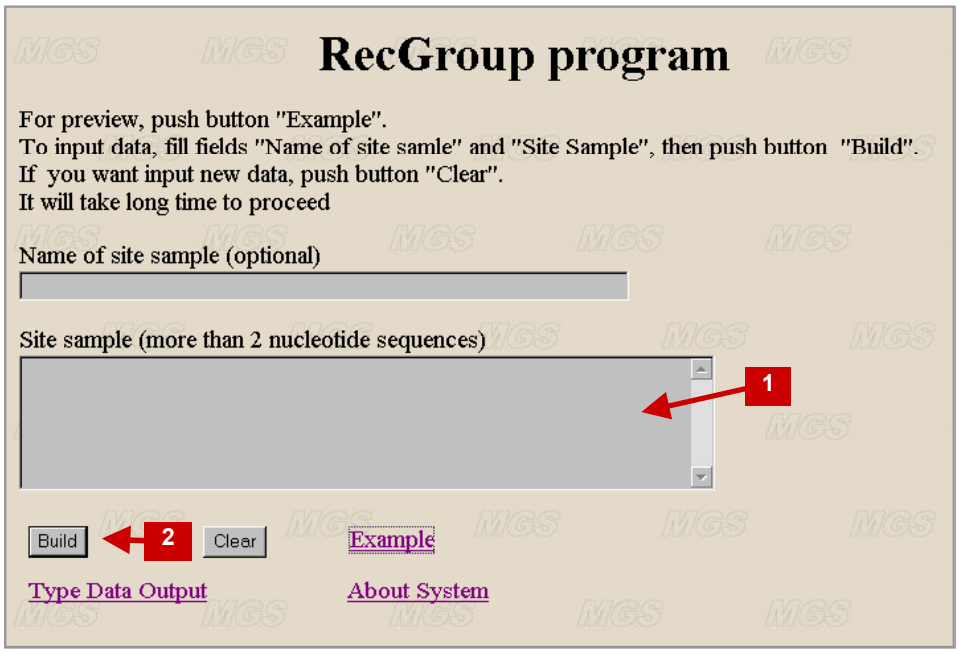
List of biological tasks that could be solved by using the RecGroup program:

- Recognition Group retains most of the information and is, thus, well suited to describe the transcription factor binding sites.

Data input

The sequences to be analysed should have the plain textual format (a, t, g, c in lower case, without spaces). Note that the number of sequences should be more than 2.

Enter the sequence to be analysed into the text-box 'Site sample' (1).



Program options

For the program execution, it is not necessary to set the parameters.

Program execution

The program is executed by clicking the button 'Build' (2).

Data output

The program outputs the results as a column of the weighted oligonucleotides.

Example

The goal is to build recognition group for a set of sequences containing NF-Y binding site. As a result, we get the list of weighted oligonucleotides, which could be used for recognition of transcription factor NF-Y binding site.

```
It will take long time to proceed. Please wait ...
 .........................................

      ************************************
      * Calculation of optimal parameters *
      * l (site   length)  and  ANM  (the *
      * admissible number of  mismatches) *
      * and construction  of  recognition *
      *                 group.            *
      ************************************

 The name of transcription factor:  Sequences Containing NF-Y
 The binding site length =   7
 Admissible number of mismatches =   2

 Recognition group:
agccaat     7
aaccaat     3
agccact     1
gaccaat     4
agctcat     1
agcctct     1
aggcaag     1
atgcaat     1
agccatg     1
ggcaaat     1
```

Comments and questions are welcome to
Yury Kondrakhin (kondrat@bionet.nsc.ru).

2.6. ARGO system

Release 2003

Program description:

The program is designed for searching for degenerated oligonucleotide motifs that are specific for gene regulatory regions.

Access to ARGO system:
http://www.domain.com/mgs/gnw/regscan/ link 'Search for degenerate oligonucleotide motifs'

List of biological tasks that could be solved by using the ARGO system:

- Oligonucleotide analysis of promoter regions,
- Oligonucleotide analysis of transcription factor binding sites,
- Oligonucleotide analysis of enhancers and silencers, as well as other regulatory elements.
- Oligonucleotide analysis of nucleosome binding sites.

Data input

The sequence to be analysed should have the plain text format (a, t, g, c in lower case, without spaces). Note that the number of sequences should be less than 100. Length of sequences should be less than 100 bp.

Enter the sequence to be analysed into the text-box 'Sequence'.

Program options

'Length of oligs' (1) – Length of oligonucleotides to be considered.

The length could be varied in dependence of the length of the signals to be searched for. The length value could vary from 6 to 12. The recommended length is 8.

'Hamming's distance value' (2).

In our case, the Hamming distance between two oligonucleotides is a number of not coinciding positions in these oligonucleotides. For example, for the pair of oligonucleotides AATGCGT and TAGGCTT, Hamming distance is equal to 4 (*A*G**T).

If Hamming distance R between oligonucleotides from different sequences is lower than the threshold value r_o, they are united into one class. We recommend to use r_o >(l/2) of the oligonucleotides length.

'Minimal share of sequences with olig' (3)

This parameter denotes the minimal percent of sequences in the set analysed, in which every found motif should present. This parameter ranges from 1 to100. Recommended value equals to 80.

'Minimal binomial probability' (4)

This parameter means the decimal logarithm of the significance level.

The oligonucleotide motif is considered significant, if it meets the following conditions: (1) the fraction f of the RGS containing the motif is higher than a certain level fo and (2) the binomial probability P(n,N) to observe this motif by accidence in n and more number of RGS from N RGS considered, is lower than a significance level a.

The value of this parameter may be varied from –1 to –999.

The more this value is close to –1, the more probable to find false motifs in the set of motifs recognised. If this value is close to –999, then the probability to loose from consideration the true signals that are present in a set.

We recommend to use lg(a)<-8.

'Alphabet size' (5)

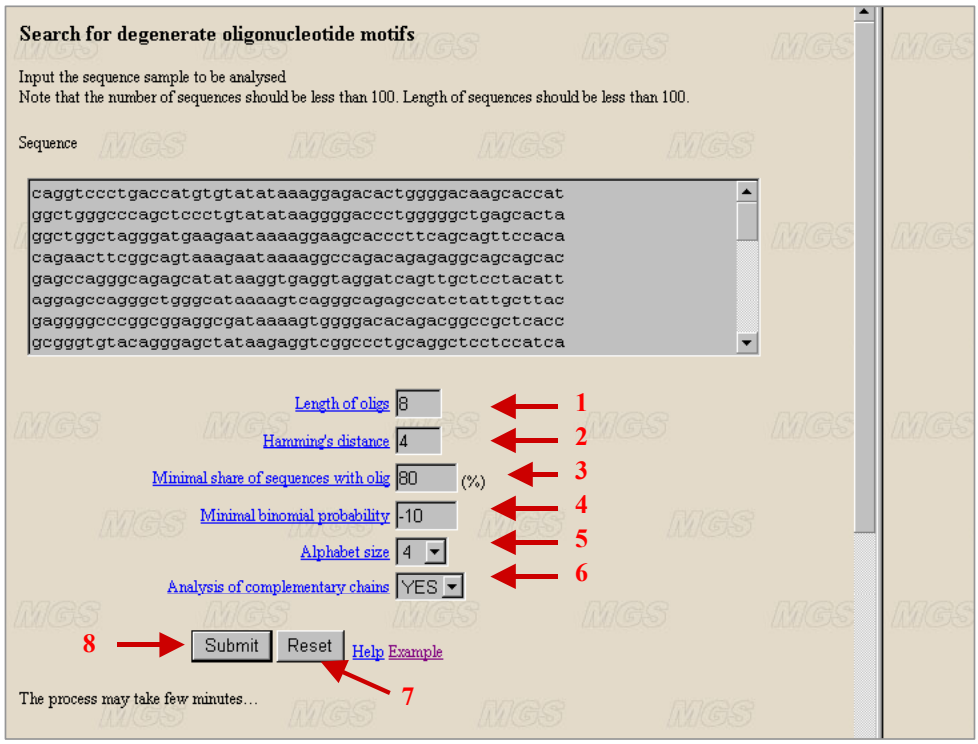
This parameter has two values:
4 – to be used for analysing the motifs with canonical letters A, T, G,C.
14 – to be used for analysis of motifs written by extended alphabet.

'Analysis of complementary chains' (6)

This parameter enables a user to analyse complementary chains of the sequences to be analysed. That is, the motifs could be searched for both in '+' and in '-' directions.

'Reset' (7)

This motif is needed to clear up the text-box 'Sequence'.

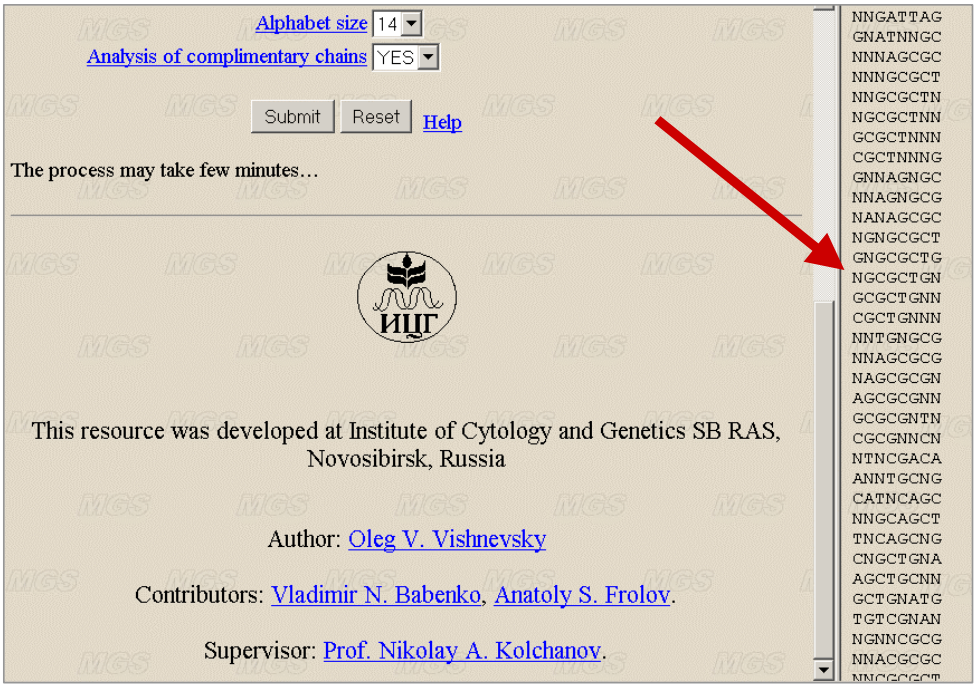


Program execution

The program is executed by clicking the button 'Submit' (8).

Data output

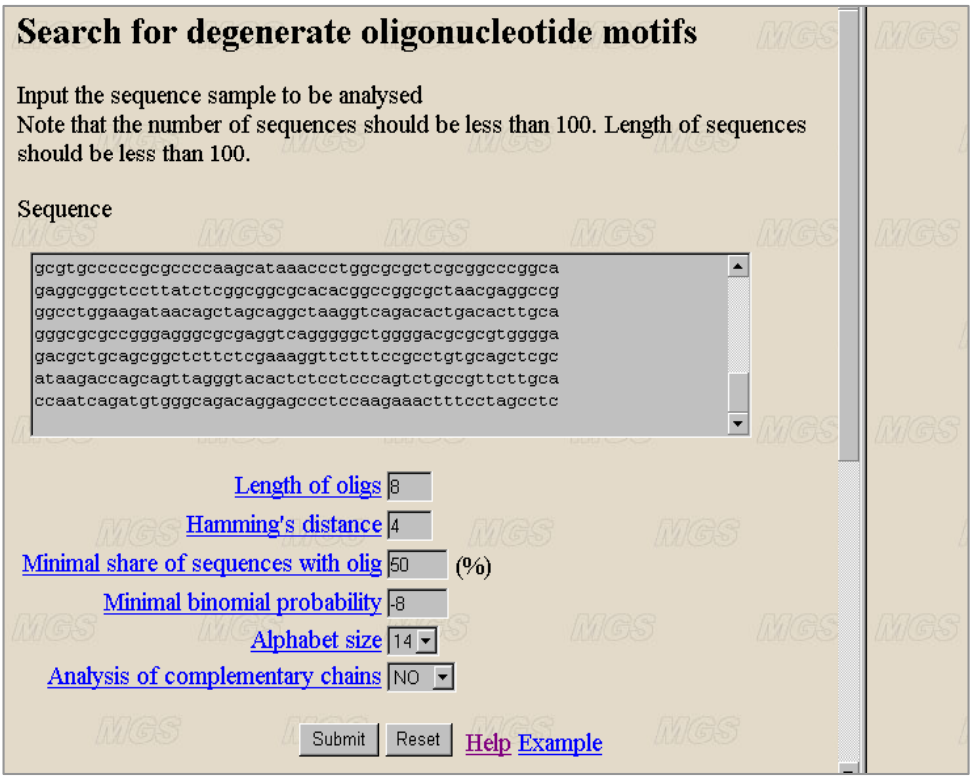
The program outputs the results as a column of the motifs found.



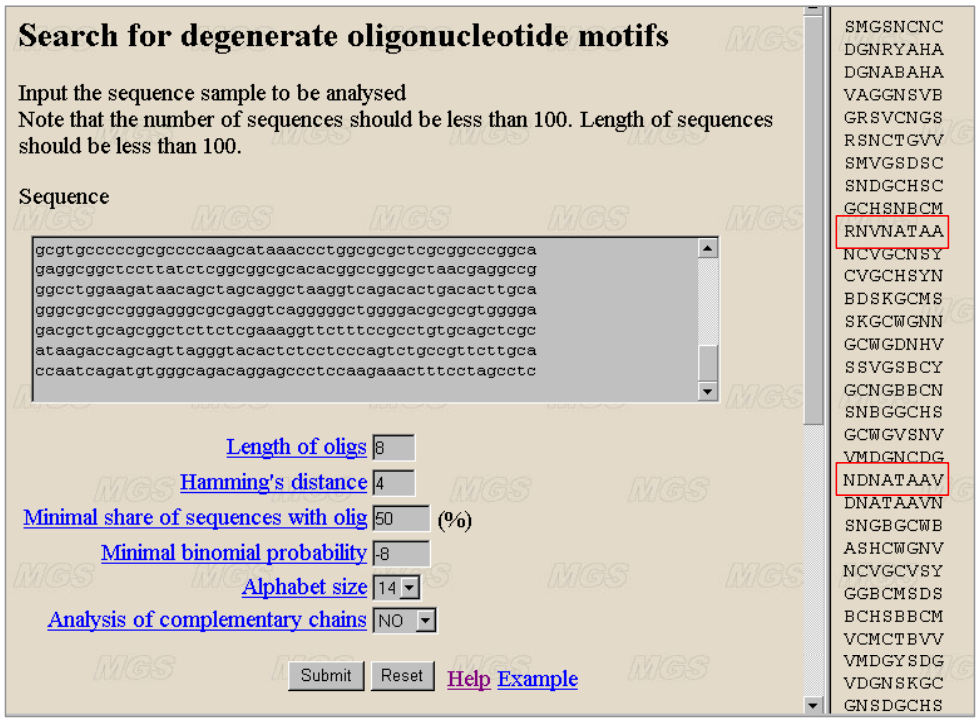
Example

The goal is to make oligonucleotide analysis of a set of sequences containing the regions of erythroid-specific promoters in-between -50 - +1 bp relatively transcription start site in order to find in these sequences the motifs that possibly correspond to the known transcription factor binding sites.

The parameters are set as follows:
Length of oligs - 8
Hamming's distance - 4
Minimal share of sequences with olig - 50%
Minimal binomial probability - -8
Alphabet size -14
Analysis of complementary chains – NO



As a result, we get the list of motifs, some of them are similar by nucleotide content, for example, to consensus transcription factor TBP binding sites (they are marked by red rectangles at the Figure).



Comments and questions are welcome to
Oleg V. Vishnevsky (oleg@bionet.nsc.ru).

3. Programs for Promoter Recognition

3.1. Program to recognize eukaryotic promoters

Release 2003

Program description

The program is designed for recognition of promoters in an arbitrary nucleotide sequence.

Access to the program
http://www.domain.com/mgs/gnw/regscan/ link 'Recog'

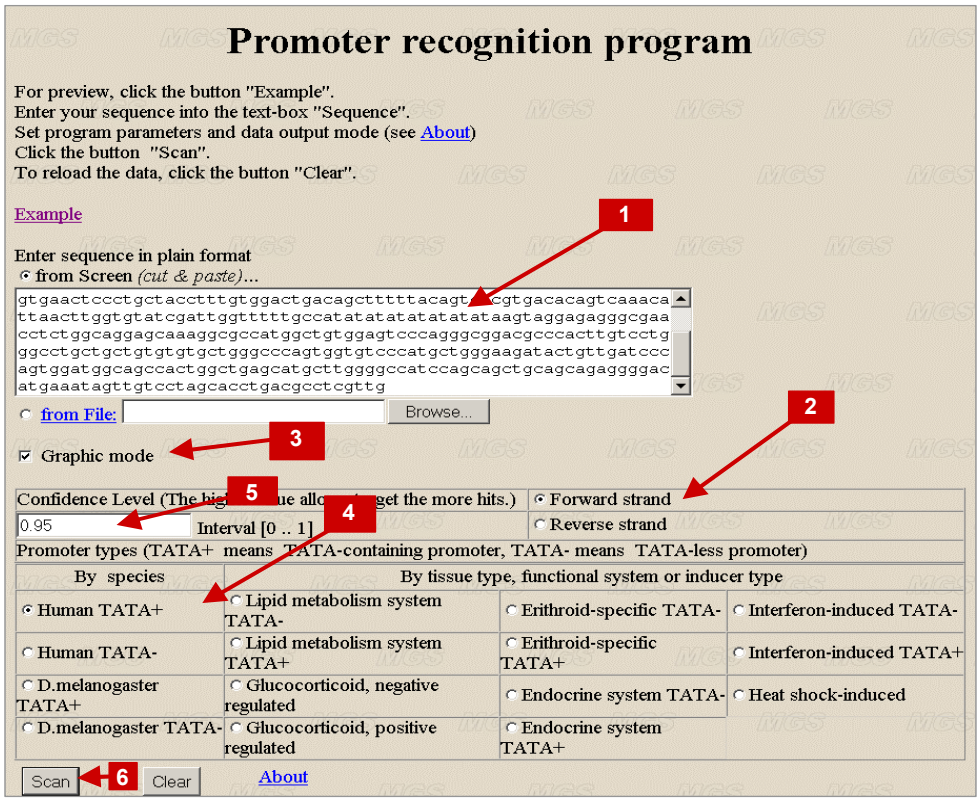
List of biological tasks that could be solved by using the program
- Recognition of promoters in DNA sequences.

Data input

Input DNA sequence to be analysed into the field 'Input DNA Sequence' (1). Sequence length should be at least 400 bp and at most 32 kbp. (a, t, g, c in upper or lower case, line feeds and blanks are ignored).

Program options

The analysis of complementary DNA strand is executed by clicking the check-box 'Reverse strand' (2). To select the data output mode, choose the option by clicking the check-box 'Graphic mode' (3). Choose the promoter type: by species (Human or Drosophila) or by cell type (erythroid), type of an inducer (interferon, heat shock), functional system (lipid or endocrine metabolism). In all these cases, only one of TATA-containing or TATA-less promoter types should be selected, i.e., in our example, positive or negative regulation by glucocorticoid inducer (4). Choose the confidence level in the check-box 'Confidence Level' (default value 0.95). By varying the Confidence Level and 'CF' value, it is possible to find a compromise between sensitivity and specificity. By default, the 'CF' value is set at 0.95; whereas the reasonable values are within the range from 0.5 to 0.99. The 'CF' value denotes the portion of correctly recognised promoters in the sequences of the training set. By increasing the 'CF' value, it is possible to get more hits (5).



Program execution

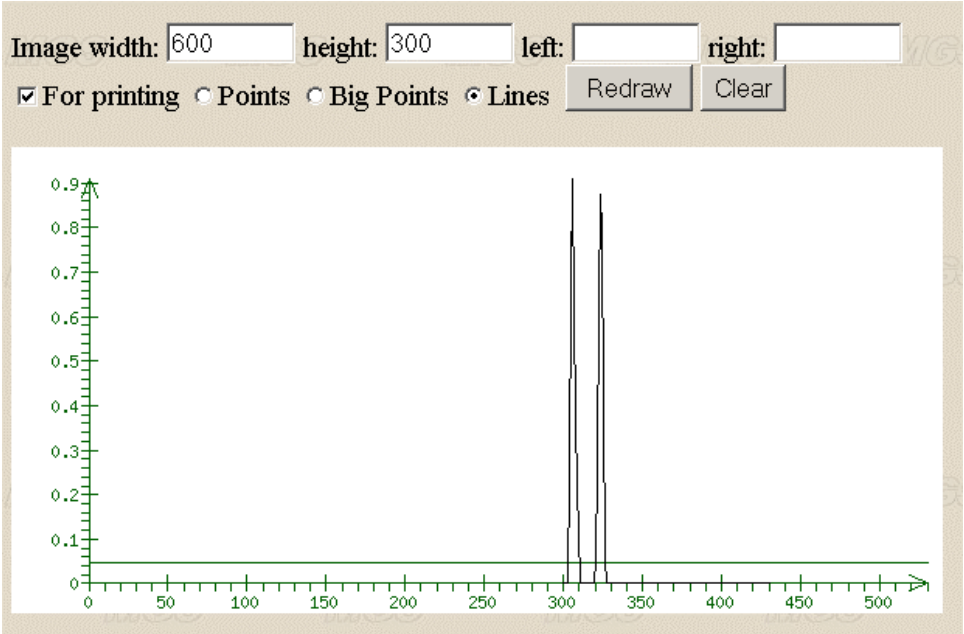
Click the button 'Scan' and wait for the program execution (6).

Output data

Output data could be presented in two modes: graphical representation and numerical delivery.

Example

The output gives the scores within the range in-between 0 and 1. The best hit equals to 1. For example, enter the sequence of the human bilirubin UDP-glucronosyltrasferase (1-1, UGT1A1) gene, AC AF352795, with the TATA-box sequence (279…295) and with transcription start at 314 nt. This gene is considered as TATA-containing human gene. Output data could be displayed as the graphical representation or as the numerical delivery. The graphical representation is shown below.



The numerical delivery is displayed as follows:

pos	value	method
	recon2	
305	0.245268	recon2
306	0.670882	recon2
307	0.908197	recon2
308	0.605870	recon2
309	0.375586	recon2
310	0.210228	recon2
311	0.112294	recon2
321	0.051683	recon2
322	0.262601	recon2
323	0.626194	recon2
324	0.876315	recon2
325	0.792281	recon2
326	0.400228	recon2
327	0.053702	recon2

Comments and questions are welcome to
Victor Levitsky (levitsky@bionet.nsc.ru)

3.2. ARGO-Viewer

Release 2003

1. Program description:

A software program ARGO-Viewer was designed for recognition of regulatory gene regions in arbitrary extended nucleotide sequences. Promoters are recognised in extended genome regions by estimating the set R of region-specific motifs found by the program ARGO.

2. Access to ARGO-Viewer system
http://www.domain.com/mgs/gnw/regscan/ link ARGO-Viewer

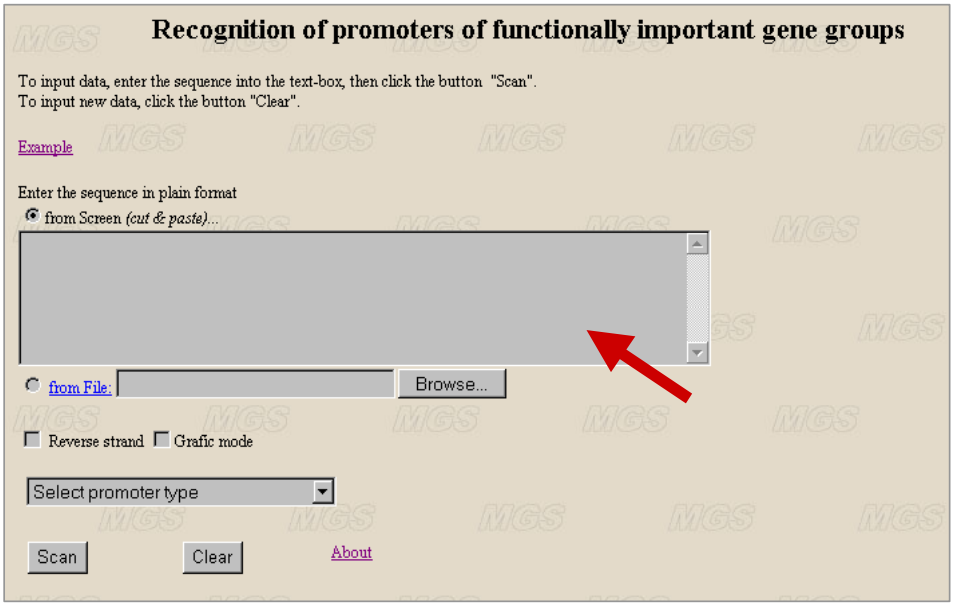
3. List of biological tasks that could be solved by using the ARGO system.

Recognition of promoter regions in tissue-specific genes in eukaryotes.

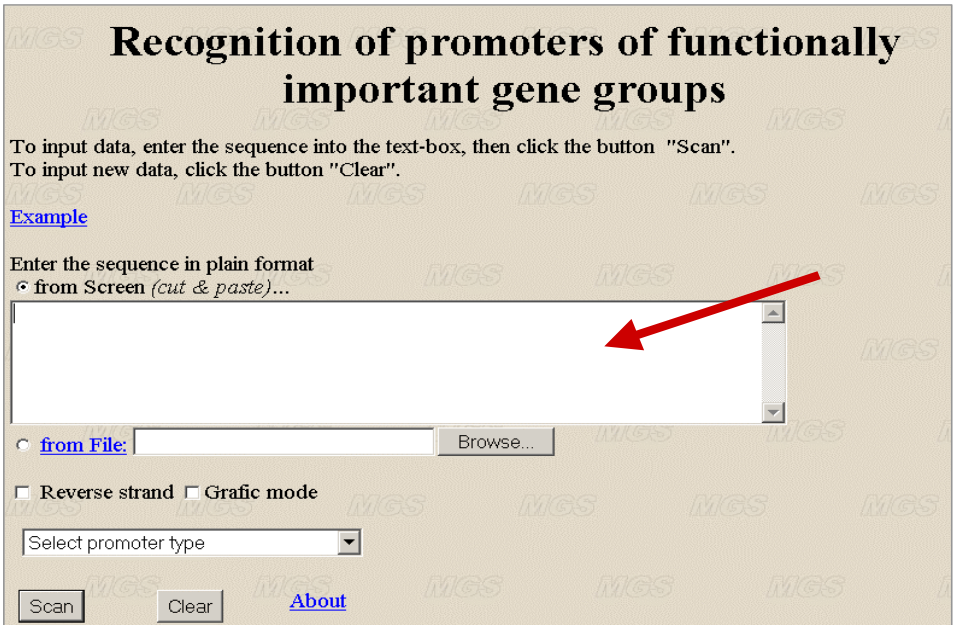
4. Data input

The sequence to be analysed should have the plain text format (a, t, g, c, printed in lower case, without gaps) and the length should range from 400 to 80000 bp.

Enter the sequence to be analysed into the text-box.



If you input the sequence from your file, use the dialog menu by clicking the button 'Browse'.

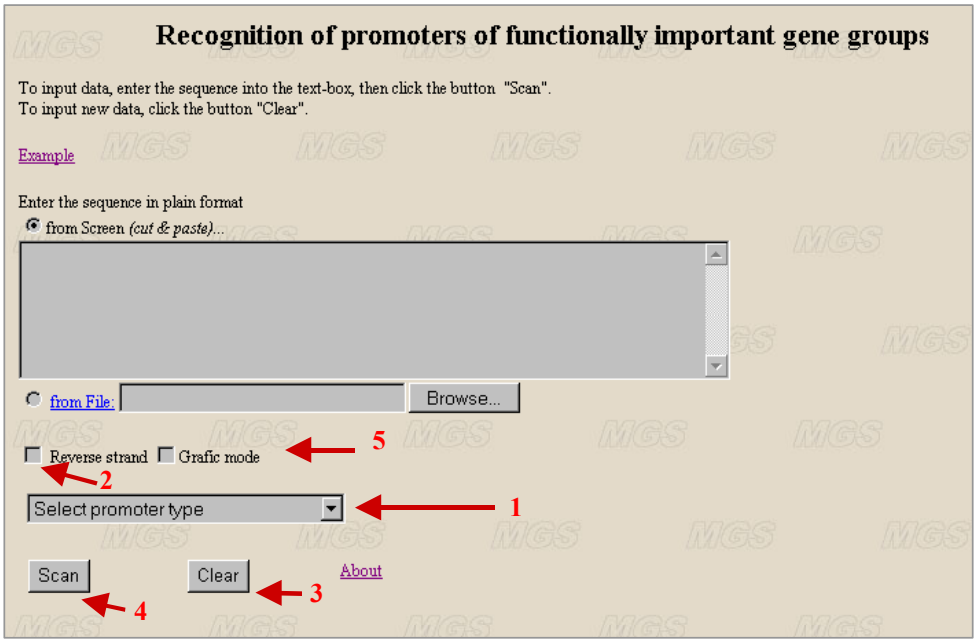


5. Program options

'Select promoter type' (1) is an option for choosing the type of tissue-specific promoters for recognition.
'Reverse strand' (2) is a parameter enabling perform recognition not only in direct strand, but in complementary strand too.
By clicking the button 'Clear' (3), it is possible to clear up the text box for data input.

6. Program execution

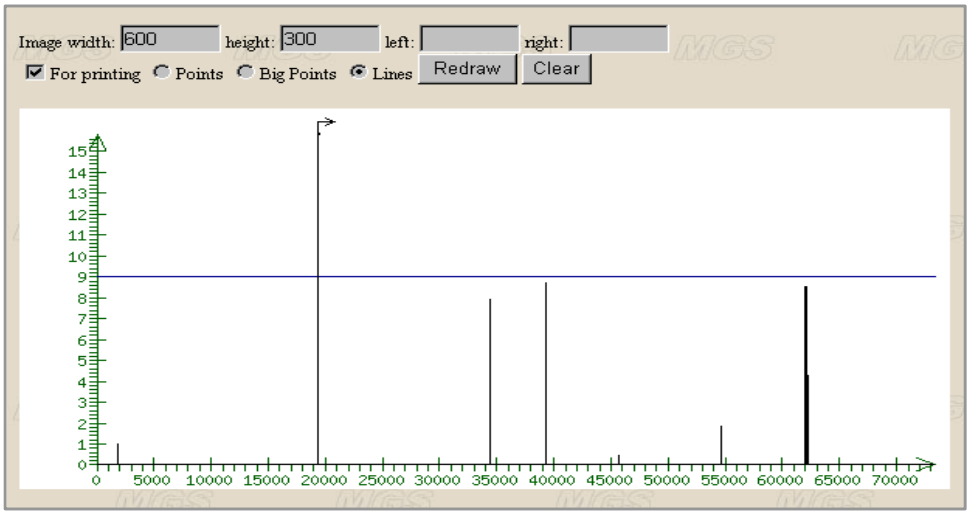
The program is executed by clicking the button 'Scan' (4).



7. Data output

The option 'Grafic mode' (5) is designed for graphical visualisation of recognition function profile.

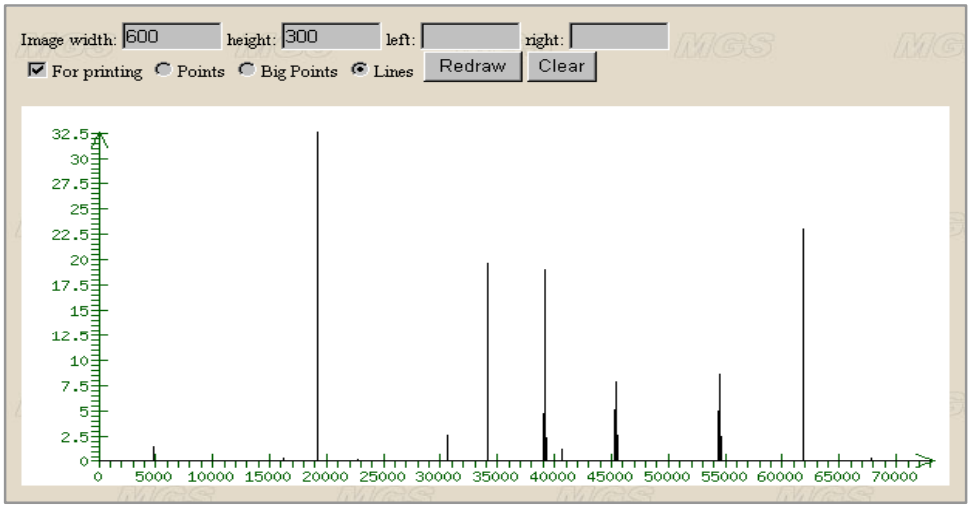
The results are output as a graphical profile of the sequence analysed. The peaks at the profile correspond to potential promoter regions.



Example

The task is to search for potential erythroid-specific promoter regions in Human beta globin region (ID HSHBB), of 73308 bp in length, mapped on the chromosome 11.
Choose the option 'Grafic mode' and select 'Erythroid-Specific Regulated genes' in the pull-down menu for choosing the type of promoters (these settings are marked by red rectangles in the Figure).

The resulting profile of the recognition function clearly illustrates 6 potential promoter regions, which correspond to experimentally detected transcription start sites at positions 19289, 34478, 39414, 54740, 62137 of five real promoters and to one promoter region of a pseudogene in vicinity of position 45557.



Comments and questions are welcome to
Oleg V. Vishnevsky (oleg@bionet.nsc.ru).

3.3. POLIIISCAN

Release 2003

1. Program description:

The program is based on a new method for recognition of type 2 promoters in the eukaryotic tRNA genes, which takes into consideration not only nucleotide distribution at individual box positions, but also the multiple interactions between the different parts of the boxes. The recognition system is determined by the module organization of the A and B boxes within intragenic promoters of the tRNA genes.

2. Access to POLIIISCAN

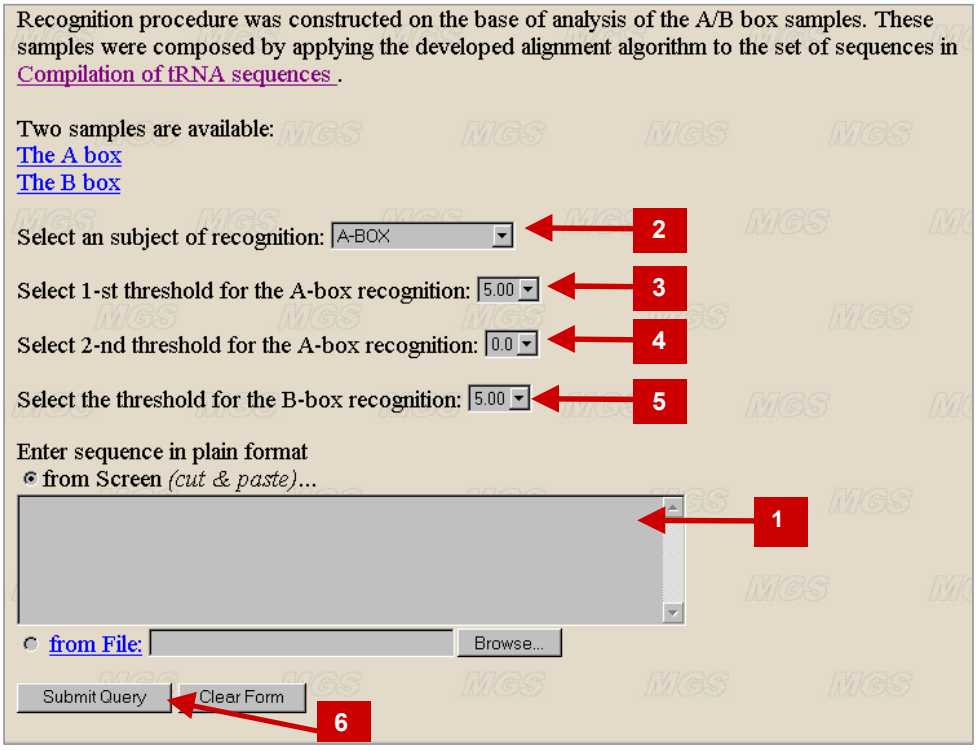
http://www.domain.com/mgs/gnw/regscan/ link 'PolIIIScan'

3. List of biological tasks that could be solved by using the POLIIISCAN system.

- the program is designed for the recognition of type 2 promoters in the eukaryotic tRNA genes

4. Data input

The sequence to be analysed should have the plain text format (a, t, g, c in lower case, without spaces). Enter the sequence to be analysed into the text-box (1).



5. Program options

Option 'Subject of recognition' (2) determines the type of regulatory region for recognition.
Option '1-st threshold for the A-box recognition' (3) orders the first limiting value for recognition of the A-box. It is recommended to set the values within the range from 5.00 to 7.70.
Option '2-nd threshold for the A-box recognition' (4) sets the second limiting value for recognition of the A-box. This value varies in-between 0.00 - 2.00.
Option 'threshold for the B-box recognition' (5) sets the first threshold value for recognition of the A-box. This value varies in-between 5.00 - 10.00.

6. Program execution

The program is executed by clicking the button 'Submit Query' (6).

7. Data output

The program outputs the results as a column of the regulatory regions found.

```
Prediction of the promoters
 1) A-Box: [   30,   40], tcacccatttgg F=5.627000, f=1.112957; B-Box: [   66,   76], gtctagtatcc F=5.506000,
 2) A-Box: [   39,   49], tggtatattaaa F=2.684000, f=0.089701; B-Box: [    1,   11], gaattctaatc F=6.413000,
 3) A-Box: [   39,   49], tggtatattaaa F=5.092000, f=0.455149; B-Box: [  110,  120], gtctgcaagcc F=5.173000,
 4) A-Box: [   44,   54], tattaaagatgt F=1.206000, f=0.584718; B-Box: [    1,   11], gaattctaatc F=6.413000,
 5) A-Box: [   44,   54], tattaaagatgt F=5.301000, f=0.156145; B-Box: [  110,  120], gtctgcaagcc F=5.173000,
 6) A-Box: [   49,   59], aagatgtgttgt F=1.972000, f=0.245848; B-Box: [    1,   11], gaattctaatc F=6.413000,
 7) A-Box: [   49,   59], aagatgtgttgt F=5.129000, f=0.720929; B-Box: [  110,  120], gtctgcaagcc F=5.173000,
 8) A-Box: [   57,   67], ttgtctactgtc F=3.135000, f=0.345514; B-Box: [    1,   11], gaattctaatc F=6.413000,
 9) A-Box: [   57,   67], ttgtctactgtc F=5.365000, f=0.850499; B-Box: [  110,  120], gtctgcaagcc F=5.173000,
10) A-Box: [   73,   83], atccctcaagta F=5.334000, f=1.272424; B-Box: [    1,   11], gaattctaatc F=6.413000,
11) A-Box: [  106,  116], aatagtctgcaa F=5.151000, f=0.345514; B-Box: [   65,   75], tgtctagtatc F=5.050000,
12) A-Box: [  118,  128], gccaggagtggt F=3.141000, f=0.451828; B-Box: [   65,   75], tgtctagtatc F=5.050000,
13) A-Box: [  118,  128], gccaggagtggt F=5.273000, f=1.106312; B-Box: [  169,  179], gggaggactgc F=5.102000,
14) A-Box: [  118,  128], gccaggagtggt F=5.273000, f=1.106312; B-Box: [  177,  187], tgcttgagctc F=5.020000,
15) A-Box: [  121,  131], aggagtggtggc F=3.273000, f=0.372094; B-Box: [   65,   75], tgtctagtatc F=5.050000,
16) A-Box: [  121,  131], aggagtggtggc F=5.426000, f=0.664452; B-Box: [  169,  179], gggaggactgc F=5.102000,
17) A-Box: [  121,  131], aggagtggtggc F=5.426000, f=0.664452; B-Box: [  177,  187], tgcttgagctc F=5.020000,
18) A-Box: [  121,  131], aggagtggtggc F=5.426000, f=0.664452; B-Box: [  191,  201], agtttgatatt F=5.372000,
19) A-Box: [  129,  139], tggctcatgtct F=2.157000, f=0.146180; B-Box: [   65,   75], tgtctagtatc F=5.050000,
20) A-Box: [  129,  139], tggctcatgtct F=2.157000, f=0.146180; B-Box: [   91,  101], ggaattagtca F=5.216000,
21) A-Box: [  129,  139], tggctcatgtct F=5.807000, f=1.617940; B-Box: [  169,  179], gggaggactgc F=5.102000,
22) A-Box: [  129,  139], tggctcatgtct F=5.807000, f=1.617940; B-Box: [  177,  187], tgcttgagctc F=5.020000,
23) A-Box: [  129,  139], tggctcatgtct F=5.807000, f=1.617940; B-Box: [  191,  201], agtttgatatt F=5.372000,
24) A-Box: [  129,  139], tggctcatgtct F=5.807000, f=1.617940; B-Box: [  195,  205], tgatattatcc F=5.026000,
25) A-Box: [  146,  156], tccagcactgga F=4.226000, f=0.634552; B-Box: [   91,  101], ggaattagtca F=5.216000,
26) A-Box: [  146,  156], tccagcactgga F=4.226000, f=0.634552; B-Box: [  109,  119], agtctgcaagc F=5.144000,
27) A-Box: [  146,  156], tccagcactgga F=5.386000, f=0.724253; B-Box: [  191,  201], agtttgatatt F=5.372000,
28) A-Box: [  146,  156], tccagcactgga F=5.386000, f=0.724253; B-Box: [  195,  205], tgatattatcc F=5.026000,
29) A-Box: [  151,  161], cactggagaggt F=2.086000, f=0.827244; B-Box: [   91,  101], ggaattagtca F=5.216000,
30) A-Box: [  151,  161], cactggagaggt F=2.086000, f=0.827244; B-Box: [  109,  119], agtctgcaagc F=5.144000,
31) A-Box: [  151,  161], cactggagaggt F=5.268000, f=1.305647; B-Box: [  191,  201], agtttgatatt F=5.372000,
32) A-Box: [  151,  161], cactggagaggt F=5.268000, f=1.305647; B-Box: [  195,  205], tgatattatcc F=5.026000,
33) A-Box: [  154,  164], tggagaggtaga F=2.795000, f=0.149502; B-Box: [   91,  101], ggaattagtca F=5.216000,
34) A-Box: [  154,  164], tggagaggtaga F=2.795000, f=0.149502; B-Box: [  109,  119], agtctgcaagc F=5.144000,
35) A-Box: [  154,  164], tggagaggtaga F=2.795000, f=0.149502; B-Box: [  122,  132], ggagtggtggc F=5.353000,
36) A-Box: [  154,  164], tggagaggtaga F=5.461000, f=0.345515; B-Box: [  191,  201], agtttgatatt F=5.372000,
```

Comments and questions are welcome to Yury Kondrakhin (kondrat@bionet.nsc.ru).

4. Other Programs

4.1. NASCA program

Release 2003

1. Program description:

The aim of the analysis is to determine the positions or regions of aligned sequences where chosen physico–chemical or conformational feature variations depend on each other. Such dependence could indicate to possible importance of these regions for functioning of the sites and could point to possible molecular mechanisms of these regulatory sequences functioning.

2. List of biological tasks that could be solved by using the Prediction programs:

- Searching for conformationally or physico-chemically interdependent regions in DNA sequence alignment of functional sites or regulatory regions.
- Searching for conformational or physicochemical properties of bDNA double helix which indicate to possible importance of regions for functioning of regulatory sequences and point to possible molecular mechanisms of functioning of these sequences.

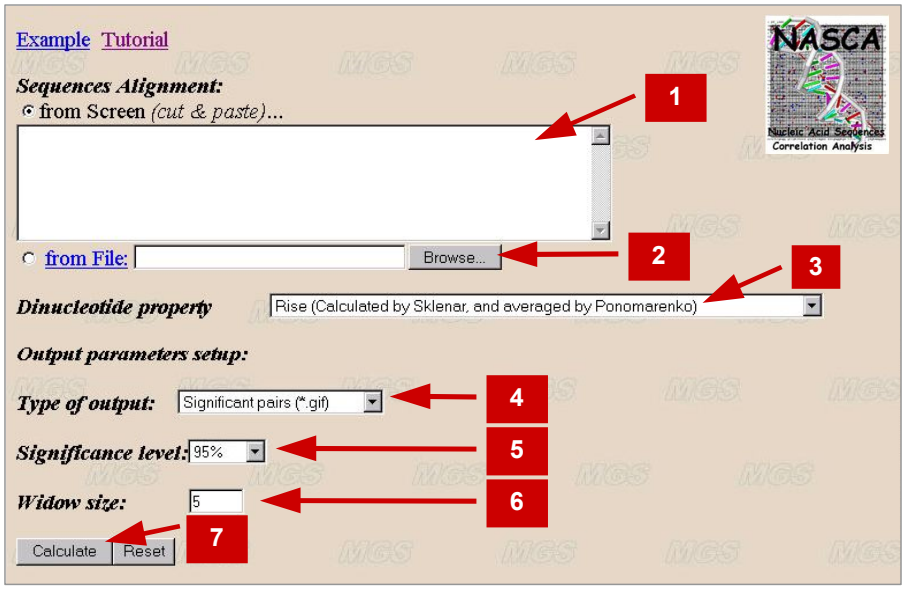
3. Access to the NASCA program
http://www.domain.com/mgs/gnw/regscan/ link 'NASCA'

4. Data input

Input the data into the input textbox (1) shown in the figure below for sequence alignment of interest.

Essential notes:

sequences should be entered in uper- or lower-case letters; line feeds and blanks are ignored; The sequence number should be at least 10. Note that accuracy of the method increases with sequence number. The sequences should be of equal length that should not exceed 130 bp and should not be less than 10. Gaps are forbidden, so positions of alignment that contain gaps will be eliminated from analysis.



5. Program options:

Input sequence alignment by different ways:

- from screen in the textbox (1) (by typing in from the keyboard or by cut & paste operation)
- or from file in 'FASTA' format: in this case, click the 'BROWSE' button (2) and select the source file.

Choose the property of interest in the list of available conformational and physico-chemical properties for calculation (3).

Choose the form of the output data display (4) in the 'Output parameters setup' menus. Output data may be represented by matrix colour picture with marked significantly correlated pairs (recommended), colour matrix diagram, HTML table of correlation coefficients matrix, and ASCII table of correlation coefficients matrix.

Choose significance level (5) in the 'Output parameters setup' menu to calculate the critical value for the correlation coefficients, which is calculated in accordance with the Student criterion at a chosen significance level.

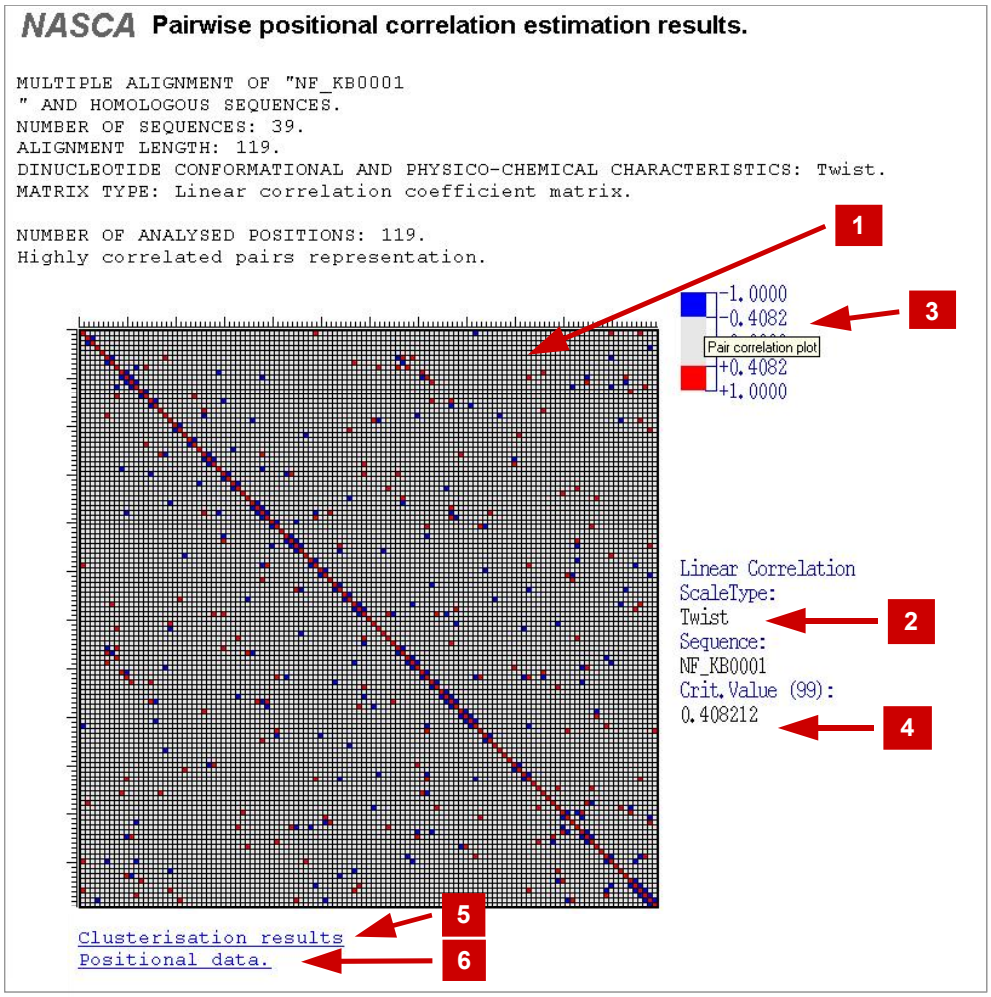
Type the value in the range 1-20 in the text-box 'Window size' (6) to choose the size of the window for clusterisation process. Values between 3 and 10 are recommended.

6. Program execution

Click the button 'Calculate' to execute the program (7).

7. Data output

The program output represents the matrix of correlation coefficients (1) of the chosen property (2) values for a given alignment. Insignificant correlations are marked grey, significant are coloured. The value scale is shown in the top right corner (3). The critical value (4) is calculated in accordance with the Student criterion at a chosen significance level.



Link to 'Clusterisation results' (5) will lead to representation of the matrix of positioning of significant blocks in correlation coefficients matrix calculated for chosen property values for a given alignment. By X- and Y-axes, positions of clusters are marked. The critical value is calculated in accordance with the Student criterion at a chosen significance level. Clusterisation was made for the window with the chosen size and chosen significance level calculated according to binomial distribution. The blocks are marked by colour. Red colour corresponds to the centre of significant block, where the number of positively and significantly correlated pairs exceeds the number of negatively and significantly correlated pairs. In other case, the centre position of significant block is marked by blue.

Link to 'Positional data' (6) leads to the table where results for significantly correlated pairs are represented in numerical values.

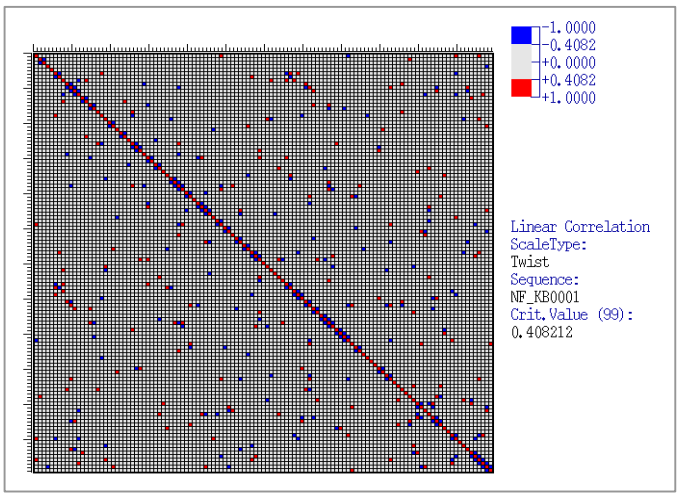
Example

Example of the program execution for alignment of 39 sequences of NF-kappaB transcription factor binding sites (available in the SAMPLES database), with the sample length of 120 nucleotides and chosen property 'Twist(Averaged for X-rays)' .

Results

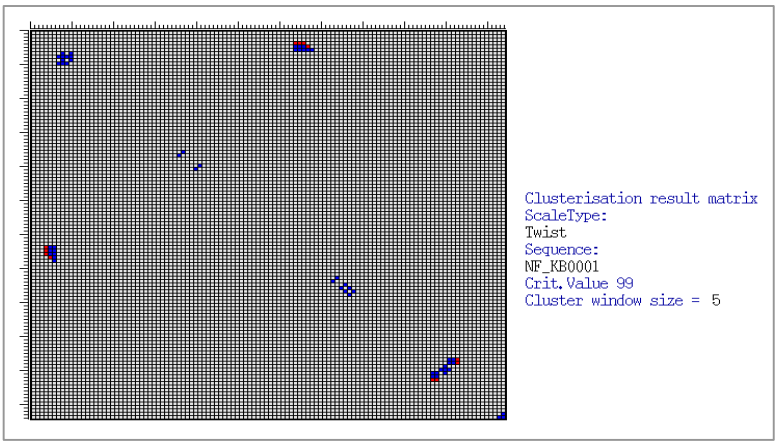
This figure exemplifies the matrix of correlation coefficients of the Twist (Averaged for X-rays) values for a sample of 39 sequences of NF-kappaB sites. Insignificant correlations are marked grey,

significant are coloured. The value scale is shown in the top right corner. The critical value is calculated in accordance with the Student criterion at a 99% significance level:



Matrix has both isolated elements corresponding to the pairs of significantly correlating positions and the clusters of such elements. To reveal such clusters (blocks), the clusterisation method could be applied. By clicking the hyperlink 'Clusterisation results', the results of the program execution are displayed.

This Figure illustrates positioning of significant blocks in correlation coefficients matrix calculated for Twist (Averaged for X-rays) property values for NF-kB sites. By X- and Y-axes, positions of clusters are marked. The value of correlation coefficient significance is estimated by the formula (2) under the confidence level of 99%. Clusterisation was made for the window with the size 5x5 and confidence level of 99%, calculated by the formula (3). The blocks are marked by colour. Red colour corresponds to the centre of significant block, where the number of positively and significantly correlated pairs exceeds the number of negatively and significantly correlated pairs. In other case, the centre position of significant block is marked by blue.



Thus, it is possible to detect the sequence regions that could be important for functioning of regulatory regions as well as relationships of these regions to each other. For example, as seen from the Figure, the cluster located in the upper central part indicates that the left flank of the site (positions 4-6) is related to the site centre (positions 74-79) according to the Twist property value. Thus, for the proper site functioning, there are functional restrictions for nucleotide content of the sequence in this regions.

Comments and questions are welcome to Dmitry Y Oshchepkov (diman@bionet.nsc.ru)

4.2. Program for estimation of stochastic complexity of genetic texts

Release 2003

Program description: The program is designed for estimation of symbolic data compression.

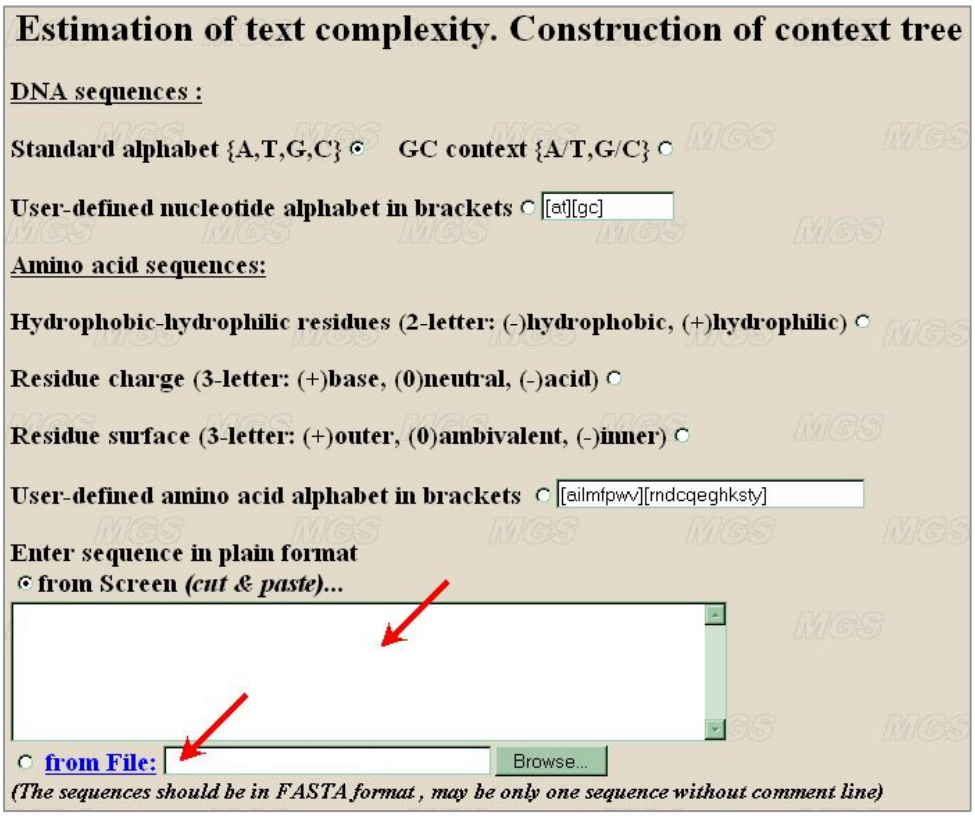
Access to the program:
http://www.domain.com/mgs/gnw/regscan/ link Complexity

List of biological tasks that could be solved by using the program for estimation of stochastic complexity:

- Estimation of general regularities in the context structure in a set of nucleotide sequences with low homology.
- Estimation of general regularities in the context structure in a set of amino acid sequences with low homology.
- Evaluation of homogeneity in a group of sequences.
- Detection of statistically significant short contexts in a sequence or in a group of sequences.

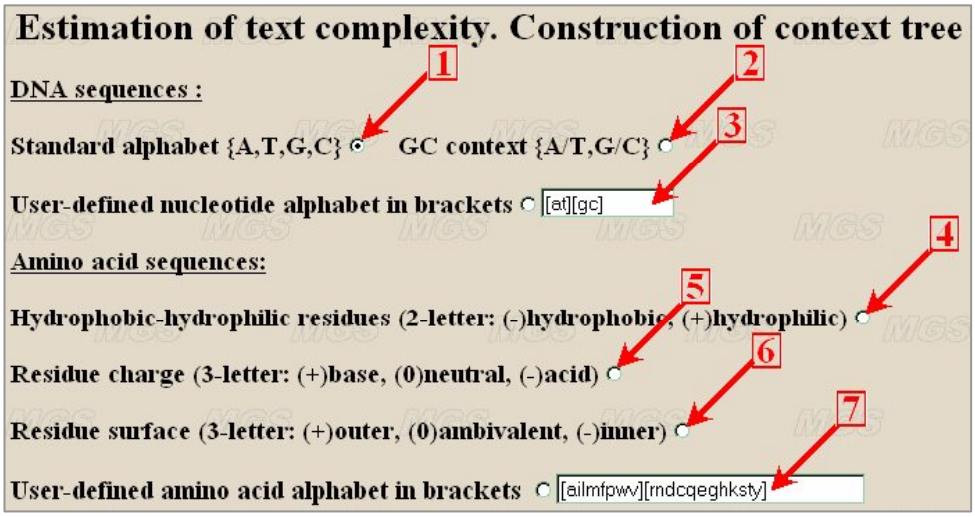
Data input.

The sequence to be analysed should be entered into the text-box in the FASTA-format. The divisor between separate sequences is the line with the first symbol '>'. The sequence can be entered from file of the user's computer by clicking the option 'From file'.



If there is only a single sequence to be analysed, you may input this sequence in a plain text format without the comment line. The program has no formal limitations for a sequence length. It is recommended to analyse sequence not less than 500 bp.

Program options.

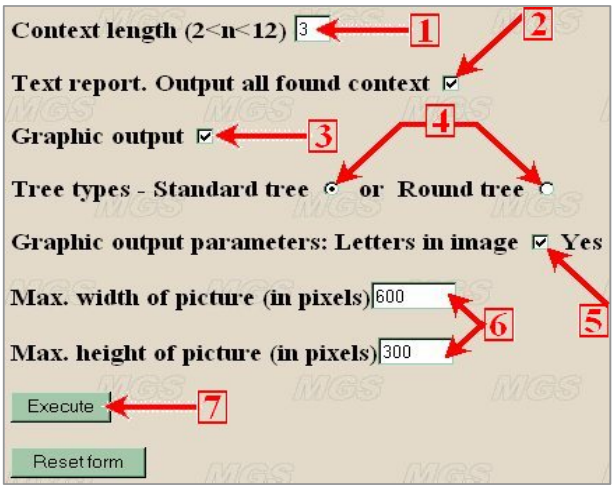


Let us determine the alphabet. If it is necessary to analyse DNA sequence, choose the alphabet for analysis of the text. It is possible to use the nucleotide symbols {A, T, G, C} (the standard alphabet, by default) (number 1 in the figure), or their combinations (for example, A/T, G/C} for studying GC-content (number 2). A user may use another combination of letters by choosing the option 'User defined alphabet'. Then a user may set in a text-box the alphabet (number 3), for example, in a form [AT][GC] or [TC][AG], or A[CGT]. The symbols in square brackets are interpreted as a single symbol. The symbols that are not indicated in square brackets are ignored.

To analyse amino acid sequences, choose the alphabet for analysis of the protein sequence. In this case, the parameters of nucleotide alphabet are ignored. The program searches for the contexts in the alphabet with the length up to 5 symbols. That is why 20 symbols denoting amino acid residues should be united in the groups. We suggest the following variants of grouping of 20 residues: by hydrophobicity-hydrophilicity (number 4 in the figure), by charge (number 5), by location (inner positioning or the surface one; number 6). A user may order his own variant of partitioning by ordering in appropriate window the line indicating how to group the symbols (number 7). By ordering by a user of his own alphabet, the residues that are not indicated will be ignored. In case the alphabet of more than 5 symbols is ordered, the estimation of complexity will be incorrect.

Only a single type of the alphabet should be chosen (see numbers 1-7).

The program parameters are ordered as indicated by arrows in the figure given below.



(1) The context length could vary from 2 to 12. The default value is 3. This is the length of the maximal context significant for generation of the symbol (dimension of the Markov model). It is recommended to input the maximal value of a parameter equalling to 12.

(2) By clicking the check box 'Text report' the program displays the complete statistical data in a text form. By default, the program will display these data. Since these data may have a large volume (it exceeds 4^d lines, where 'd' is the context length), a user may discard this option.

(3) To display the graphical data output, select the appropriate check box by default.
Select the form of graphical representation of the prefix context tree.

(4) By default, the tree with suspended vertexes (leaves) is ordered ('Standard tree'). The contexts are read from top to the bottom or from leaves to the root. This corresponds to disposition of contexts in the sequence from the left to the right (for DNA, from 5' to 3' end; for proteins, from N- to C-end). A user may choose the tree representation in a form of the radial tree. This representation is more compact. In this case, the contexts are read from outer surface to the centre (root of a tree).

(5) Additional parameters for the tree representation set the data output without denotation of vertexes (without letters) for simplicity. By default, the graphical output is displayed with the letters from the alphabet ordered by a user. If the alphabet is not standard, the symbols are replaced by numbers.

(6) A user may also set the options for the scale of graphical representation or to make the picture less for more convenient copying from screen and further usage. The absolute size of the picture by height and width in pixels are set in appropriate text-boxes. The maximal size is set by default. It is recommended to set the size of at least 200 pixels.

Program execution.

The program is executed by clicking the button 'Execute' (see number 7 in figure given above).

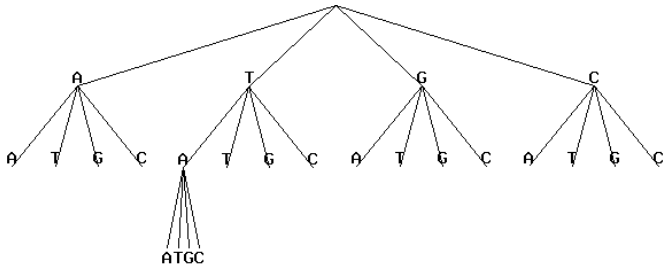
Data output.

The program outputs the results in a textual or graphical format as was indicated by a user, by means of clicking the check boxes 2 or 3 shown in the figure above.
The program displays in the text format the value of complexity for a sequence entered and a set of selected significant contexts.

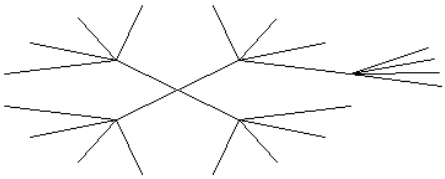
Example of the data output:

```
Set context len: 3
Show all statistics
-------------- Data information --------------
Alphabet size: 4
Alphabet contains: ATGC
Context length: 3
Number of contexts: 64
Complexity: 16410.867772
This is tree
in pseudographics (0-absent, 1-present)
/treebegin 4 ATGC
1
1111
1111000011111111
00000000000000000000000000000000000000000000000011110000000000000
Context - Variability - Complexity
AAC=103 32- 33-  7- 31- 197.7044
TAC= 48 21- 14-  4-  9-  93.0464
GAC=119 26- 53- 25- 15- 228.9963
CAC=126 39- 44-  8- 35- 238.3943
```

In a graphical representation, the visualisation of the context tree in a standard or radial form is displayed as a GIF-file. An example of representation of the context tree in a standard form is given below:



The same context tree could be represented in a radial form:

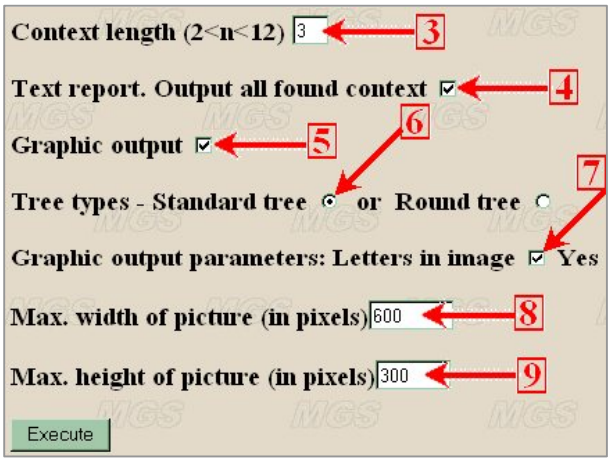


Example.

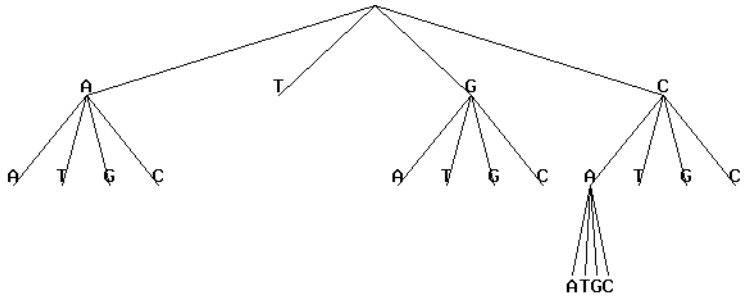
Let us consider the task of searching for context dependencies in the set of AP-1 binding sites (32 kB, Eukaryotic Transcription Factor Binding Sites Compilation http://wwwmgs.bionet.nsc.ru/mgs/dbases/nsamples/).

Enter this set in a FASTA-format into the input text box (step 1).
Set the standard alphabet {A, T, G, C} (step 2).

Set all the parameters by default.
Set the context length equal to 3 (step 3 in the figure below).
Set the textual output of contexts (step 4).
Set the graphic mode of the output data (step 5).
Set tree type as a standard tree (step 6).
Set parameter 'letters in image' equal to 'Yes' (step 7).
Set maximal width of the picture equal to 600 pixels (step 8).
Set maximal height of the picture equal to 300 pixels (step 9).



The results of the program execution will be displayed as the figure shown:



As seen from the figure, for the sequence considered, the statistically significant are the dinucleotides of the form NA, NG, NC, where N=A/T/G/C. This fact should be taken into account for constructing recognition programs. As illustrated additionally, the contexts of the form NAC are significant too.

Comments and questions are welcome to Orlov Yu.L. (orlov@bionet.nsc.ru)

CHAPTER 5. ACTIVITY SYSTEM

1. ACTIVITY Database

Release 2003

1. Database description:

We suggest a new approach to predict the activity of DNA functional sites that is focused on the perspicuity of the prediction in terms of "a probable molecular mechanism of the site functioning". The biological novelty of the method is in the involvement of physico-chemical and conformational DNA properties to provide clear interpretation of the obtained activity predictions in terms of a probable molecular mechanism of the site functioning.

2. Access to ACTIVITY database:
http://www.domain.com/mgs/gnw/activity/

3. Database content

SRS table	Description	Number of entries
ACTIVITY	A database for accumulation of activity values of DNA/RNA functional sites. Each entry in the ACTIVITY database describes a set of "sequence-activity" data measured in a fixed experimental system. The entry presents the "sequence-activity" data in a computable format.	554
KNOWLEDGE	It contains significant features of DNA/RNA site sequences determined experimentally (i.e. sequence-activity relationships under fixed experimental conditions), and the programs for predicting activity of these sites.	22
REFERENCE	Database on references to original publications describing the experimental data on DNA/RNA site sequences with known activity magnitudes.	265
PROPERTY	Database on sequence-dependent conformational and physico-chemical B-DNA parameters	38
WEIGHT	Database on weight	33
SCIENTIST	Database on scientists annotating the literature data	29

4. List of biological tasks that could be solved by using ACTIVITY database

- To browse all available experimental data concerning exploration of DNA functional sites activity.
- To explore data about the significant features of the DNA/RNA site sequences determined experimentally (i.e. sequence-activity relationships under fixed experimental conditions). Information is stored within KNOWLEDGE supplementary database.
- To predict possible type of DNA/RNA functional sites activity in the sequence of interest. For a fixed type of activity, the predicting C-program is stored in the KNOWLEDGE database.
- ACTIVITY is useful for molecular biology, pharmacogenetics, metabolic engineering, drug-design, and biotechnology.

5. SRS table format

ACTIVITY

Line code	Field name	Field description
OC	Taxon Specificity	This field describes the taxon specificity of the gene containing the site analyzed, if the sequences are not synthetic.
FF	Site Name	This field describes the name of the site analyzed and the site localization in the gene, if the sequences are not synthetic.
AN	Type of Activity's Measurement	This field describes the magnitude measured in an experiment.
AU	Measurement Units	This field describes the units of a magnitude.
PN	Sequence Phasing Point	This field describes the site sequence phasing point.
SC	Site's Variant	This field contains the name of the site's variant analyzed.
SQ	Site Sequence	This field contains the sequence analyzed ("+"-chain, –5'-3' direction).
SA	Activity Magnitude	This field contains the value of a magnitude measured in an experiment for the site's variant with the sequence given in the field SQ.
SD	Standard Deviation	This field contains the standard deviation value of a magnitude measured in an experiment for the site's variant with the sequence given in the field SQ.
PA	Position of the Phasing Point in the Sequence	This field indicates position of the phasing point (see field PN) relative to the start of the sequence given in the field SQ.

KNOWLEDGE

Line code	Field name	Field description
MI	Entry ID	This field indicates an identifier.
MN	Entry Name	This field contains name of the entry.
HN	Link to SCIENTIST database	This field contains link to the SCIENTIST database and indicates a contributor of the entry.
DA	Link to Activity database	This field contains link to the ACTIVITY database.
DR	SRS-link	This field contains the links to other databases installed under SRS (ACTIVITY, SCIENTIST, SYSTEM, PROPERTY, etc.).
WW	Web-link to Activity Tools	This field contains the link to the Web-based tools implementing each C program documented within the entry to recognize the site by its significant feature or predicting activity value within an arbitrary DNA sequence.
GF	Mathematical Model	This field describes the mathematical model
CT	Computational Method	This field describes the computational method.
DW	Link to WEIGHT database	This field contains the link to the WEIGHT database installed under SRS.
DP	Link to PROPERTY database	This field contains the link to the PROPERTY database installed under SRS.
PV	Property Name	This field contains short name of the property investigated
HL	Indicates Feature Deviation	This field indicates high or low deviation of the conformational or physico-chemical feature of the site studied from the random sequences.

AB	Analyzed Region	This field indicates the DNA region, for which the conformational or physico-chemical feature differing significantly the sites studied from the random sequences was revealed. Positions are given in bp relative to the first position of the site sequences.
UT	Utility	This field indicates the utility of the context-dependent feature for discrimination of the sites studied from the random sequences
LC	Linear Correlation Coefficient	This field describes the linear correlation coefficient.
AL	Significance Level	This field describes the significance level.
FG	Graphical Representation of Test Results	This field represents the links to the graphical representation of the results obtained by the program, given below in the field C-CODE, over the independent control sequences of the site.
C-CODE		This field contains the program calculating the profile of the sequence-dependent feature or activity predicted values along arbitrary sequence in the 'C' language of the ANSI standard. Each C program documented within the KNOWLEDGE entry has a check-box in the MENU window of the Activity Tools (see field WW).

REFERENCE

Line code	Field name	Field description
RN	Entry ID	This field contains an identifier.
RA	Authors	This field indicates the authors of the article
RT	Title	This field contains an article title.
RJ	Journal	This field contains the journal name
RV	Volume	This field contains the number of the journal volume.
RP	Pages	This field indicates the numbers of pages.
RY	Year	This field indicates the year of publication.
RR	Abstract	This field contains the abstract of an article.
RS	Data Source	This field indicates the number of a figure or a table, which contain the data used in the ACTIVITY database.
DR	SRS-link	This field contains the links to the supplementary databases (ACTIVITY, SYSTEM, etc.).

PROPERTY

Line code	Field name	Field description
MI	Entry ID	This field indicates an identifier.
MN	EntityName	Each PROPERTY entry describes the only B-DNA property, either "conformational" or "physico-chemical" one.
MD	EntityDependence	Entity Dependence
ML	Step	Step
RN	Reference	The field RN links to the database REFERENCE citing the literature sources, out of which the property values were taken
PN	PropertyName	This property is entitled by the field PN
PM	Method_for_PropertyDefinition	Method for Property Definition
PV	PropertyID	
PU	Unit_for_Property Expression	contains the property unit
DINUCLEOTIDE		quantitative value assigned to each dsDNA dinucleotide step

SCIENTIST

Line code	Field name	Field description
ID	Identifier	Identifier
HH	Authors	Authors name
HP	Tel	Telephone
HF	FAX	FAX
HE	Email	E-mail
HL	Laboratory	Laboratory name
HI	Institute	Institute name
HD	Department	Department name
HM	Minister	Ministry name
HC	Country	Country name
HA	Address	Home address

WEIGHT

Line code	Field name	Field description
ID	EntityID	This field indicates an identifier.
MN	EntityName	This field contains name of the entry.
MD	EntityDependence	Entity Dependence
WP	WeightName	Weight Name
WL	WeightPicking	Weight Picking
WW	WeightWidth	Weight Width
WI	WeightID	Weight ID

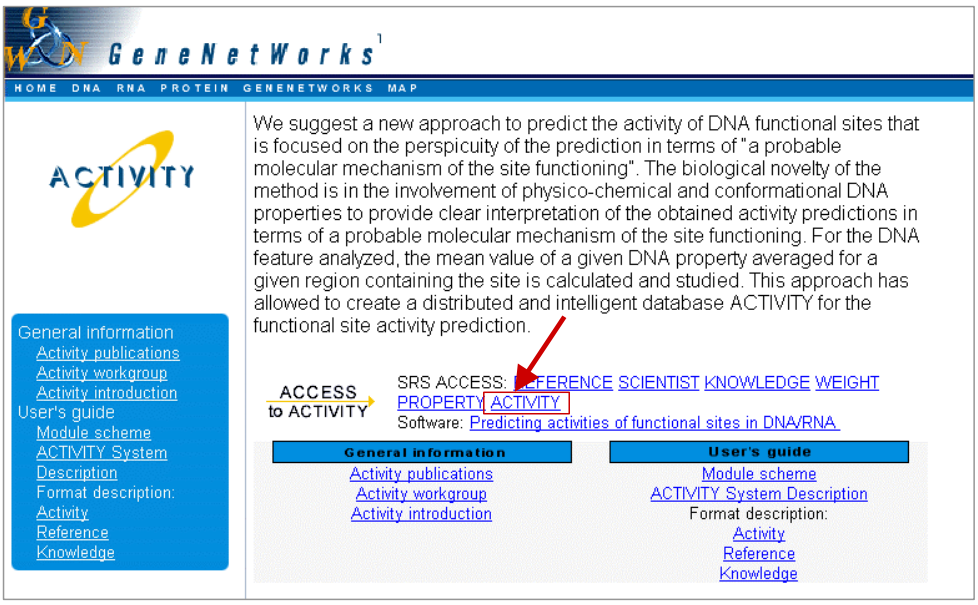
Examples of SRS queries to the ACTIVITY database

The query is:

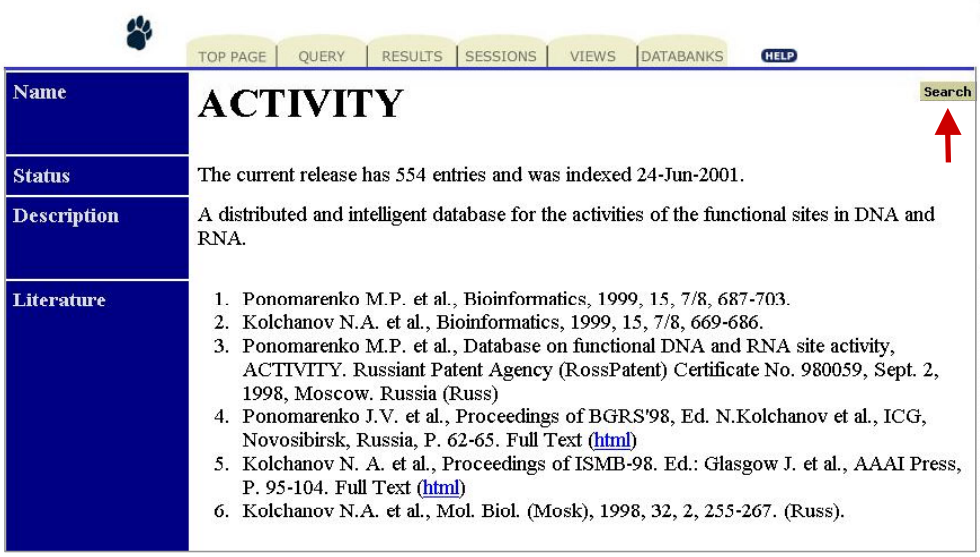
What experimental data is available concerning relative binding activity of different variants AP-1 site sequences?

To make such a query, you should perform the following:

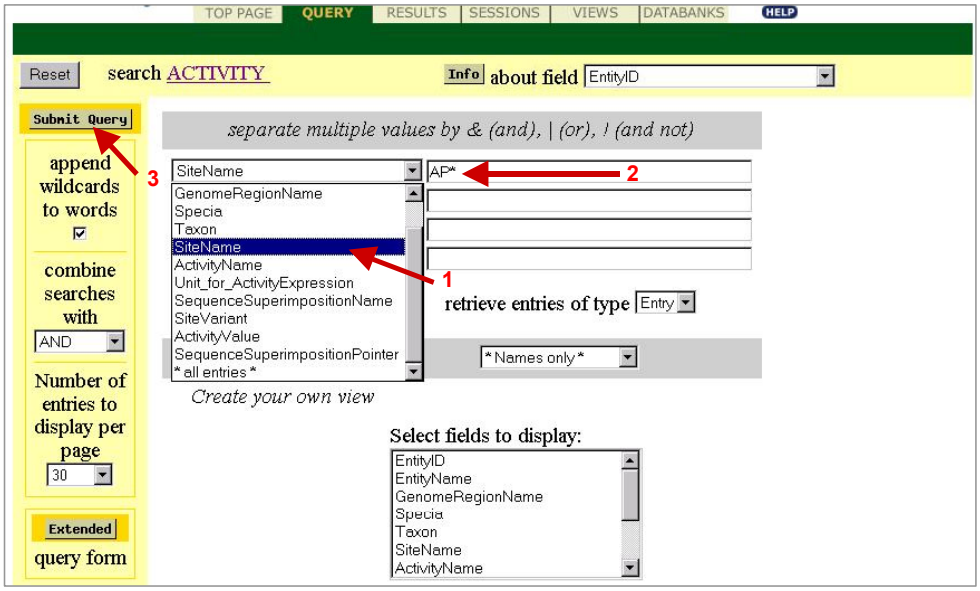
1. Choose ACTIVITY SRS table on the page 'access to ACTIVITY'.



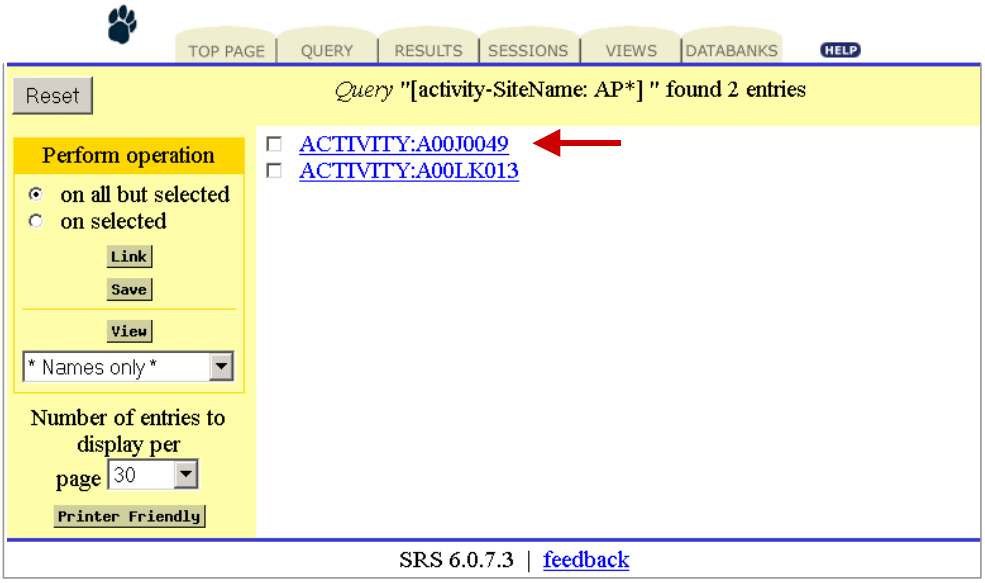
2. This will bring up the home page of the chosen ACTIVITY SRS table. Click the button 'Search'.



3. Select the fields to be searched for from the list. You will need the field 'SiteName' (1), Type terms to be searched in the text window: 'AP*' (2)
(An asterisk means "any symbol") and click the button 'Submit Query' (3).



4. This will bring up the resulting window with the list of matching entries/
To extract all information contained in the entry, click any link



5. The query result is the whole set of information about experiment and relevant binding activity of the AP-1 site sequence variants based on the features of the AP-1 site.

ACTIVITY:A00J0049

```
MI A00J0049
MN binding activity of Jun-Fos dimer to AP-1 and CRE sites in vitro
YY
HN Expert Database: SCI00002
YY
OG synthetic oligonucleotides
OS in vitro
OC EUKARYOTA
YY
FF AP-1 or CRE site
AN relative binding activity
AU percent
PN start of sequences
YY
TD Site 1125 (TRRD)
TD Site 2281 (TRRD)
TD Site 2203 (TRRD)
TD Site 2284 (TRRD)
TD Site 2286 (TRRD)
TD Site 1383 (TRRD)
TD Site 1952 (TRRD)
TD Site 2374 (TRRD)
TD Site 1503 (TRRD)
TD Site 1892 (TRRD)
TD Site 1893 (TRRD)
TD Site 981 (TRRD)
TD Site 982 (TRRD)
TD Site 1903 (TRRD)
TD Site 1904 (TRRD)
TD Site 2074 (TRRD)
TD Site 2075 (TRRD)
TD Site 2240 (TRRD)
```

Comments and questions are welcome to Mikhail P. Ponomarenko, (pon@bionet.nsc.ru).

2. Predicting Activities of Functional Sites in DNA/RNA

1. Program description

The identification of the sequence-dependent DNA features correlating with affinity magnitudes of DNA sites interacting with a protein could pinpoint to molecular event limiting this protein/DNA recognition machinery. This approach is realized in computer system ACTIVITY containing the databases on site activity and on conformational and physical-chemical DNA/RNA parameters. By using the system ACTIVITY, an analysis of some sites was provided and the methods for predicting the site activity were constructed.

2. Access to Predicting activities of functional sites in DNA/RNA

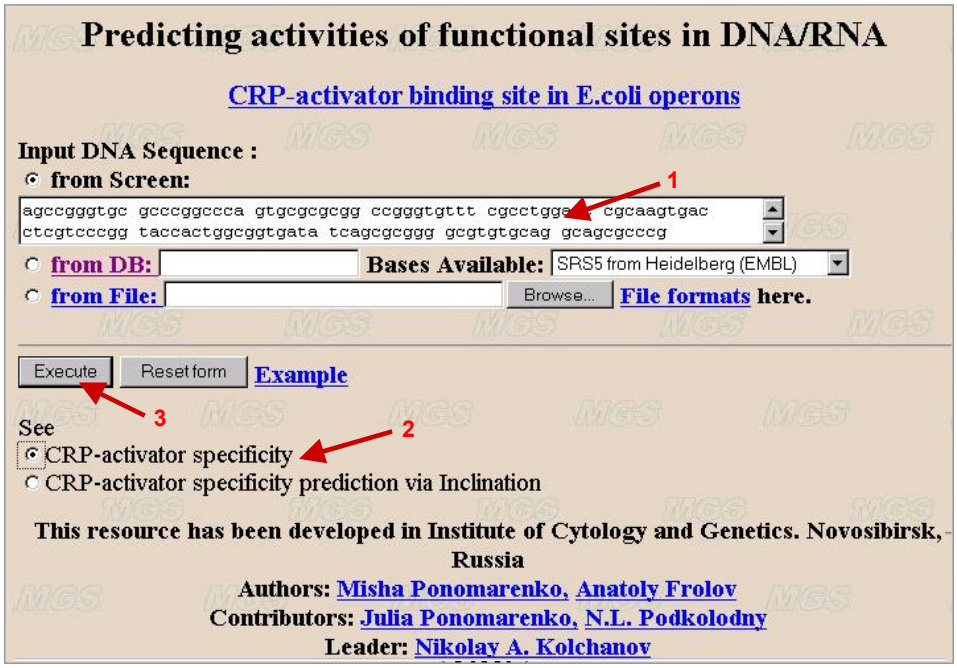
http://www.domain.com/mgs/gnw/activity/ link 'Predicting activities of functional sites in DNA/RNA'

3. List of biological tasks that could be solved by using the program

- Prediction of site activity based on its primary nucleotide sequence.

4. Data input

Input the DNA sequence of interest into the field 'Input DNA Sequence' (1). Sequence should be in a plain text format (a, t, g, c in up- or low case, tabulation or spaces are accepted). The sequence to be entered should be of at most length of 32 kbp and of least length equaling to the site sequence length.



5. Program options

Specify feature of prediction program by clicking one of check-boxes (2).

6. Program execution

Start the tools processing by clicking the button 'Execute' (3).

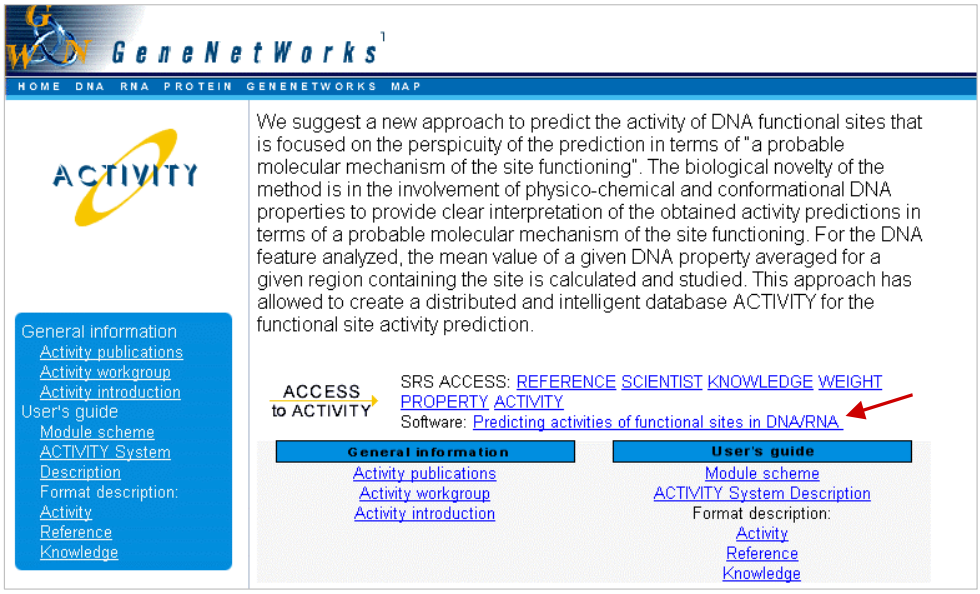
7. Data output

The tools output represents the profile of the significant feature or predicting activity value.

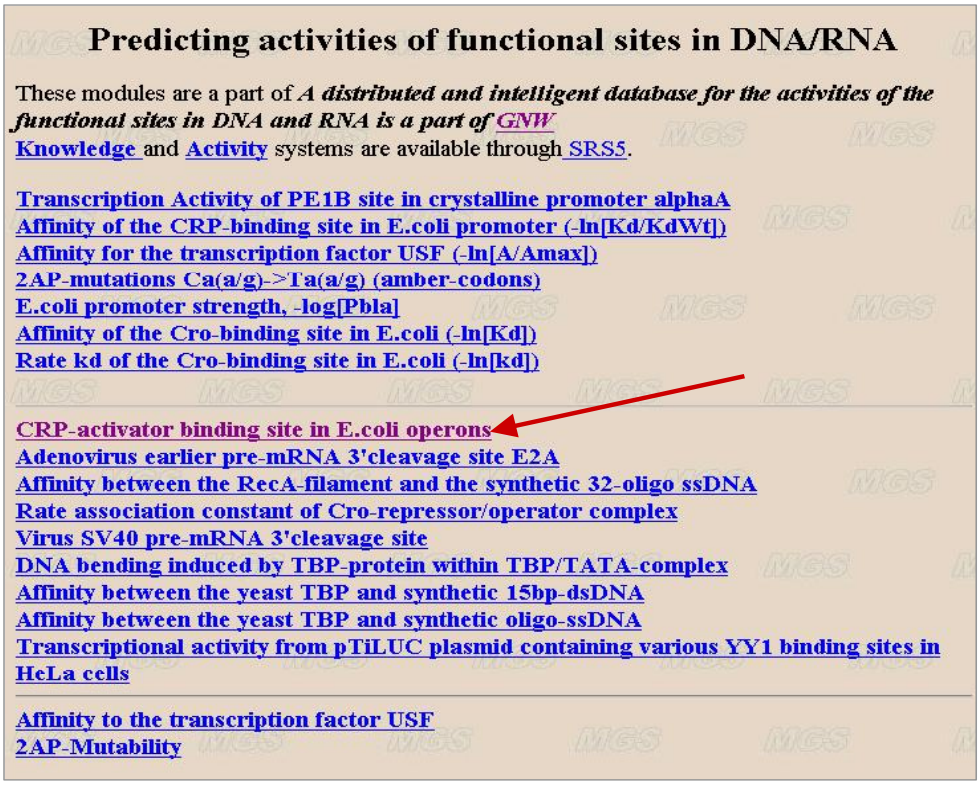
Example

To search for activity of the CRP-activator binding site in E.coli operons

1. Choose the link ‘predicting activities of functional sites in DNA/RNA’.



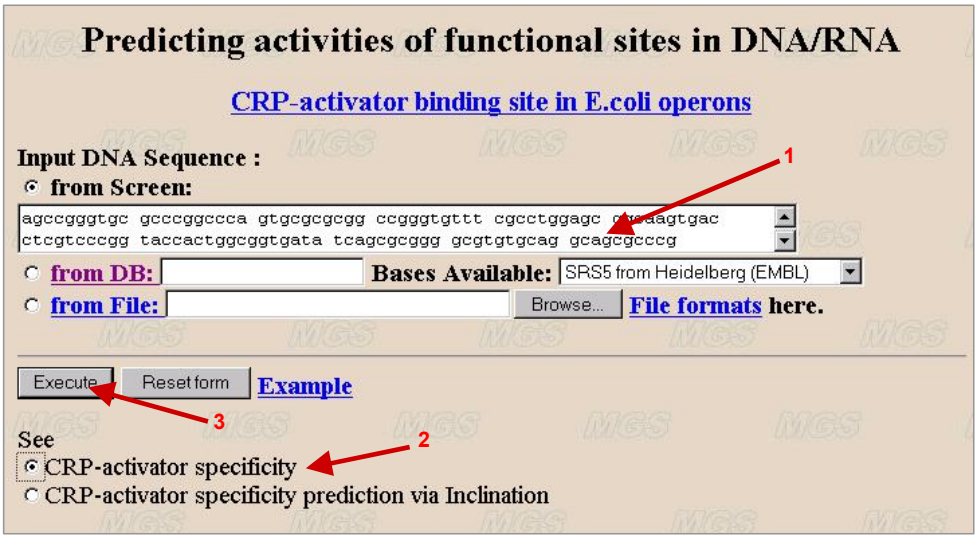
2. Choose what kind of activity should be predicted. Each C program documented within the KNOWLEDGE entry has a check-box in the MENU windows:



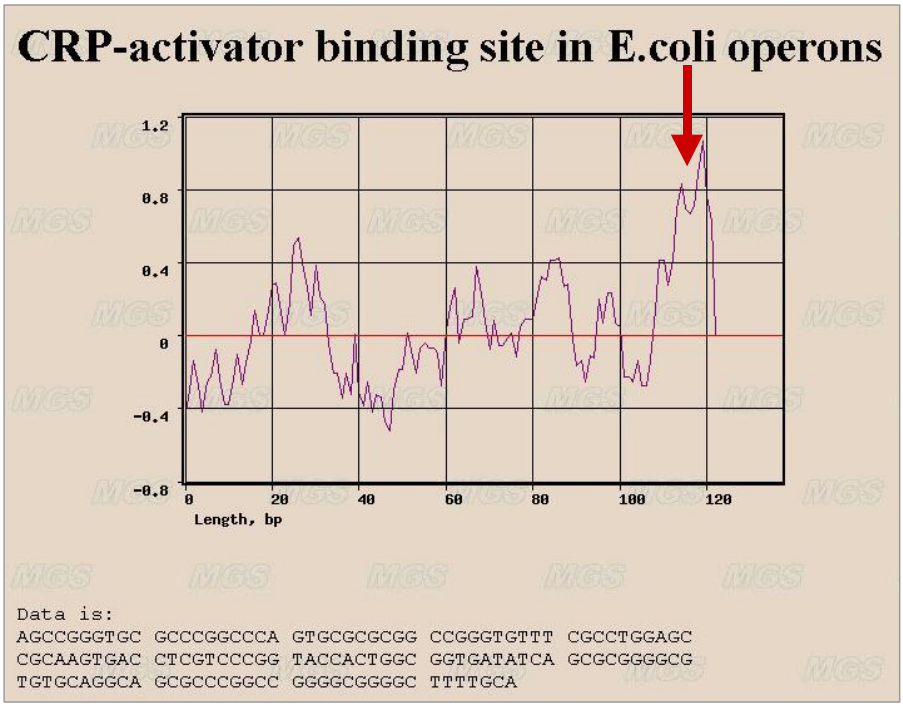
3. Input the sequence of the CRP-activator binding site into the text-box window (1)

```
ggaactgggc   ggagttaggg   gcgggatggg   cggagttagg   ggcgggacta   tggttgctga
ctaattgaga   tgcatgcttt   gcatacttct   gcctgctggg   gagcctgggg   actttccaca
ggaactgggc   ggagttaggg   gcgggatggg   cggagttagg   ggcgggacta   tggttgctga
ctaattgaga tgcatgcttt gcatacttct gcctgctggg gagcctgggg actttccaca
```

Choose the option ‘CRP-activator specificity’ (2) and click the button ‘Execute’ (3).



4. The output represents the profile of the Score value. The positive peaks of this profile point to the potential site recognized.



Comments and questions are welcome to Mikhail P. Ponomarenko, (pon@bionet.nsc.ru).

CHAPTER 6. DNA NUCLEOSOMAL ORGANISATION

1. PROFILES Databases

Release 2003

1. Database description

PROFILES and PROFILE_LIST databases accumulate an information about the profiles of conformational and physico-chemical DNA parameters, their significant extremum points and linear trends.

2. Access to PROFILES Databases

http://www.domain.com/mgs/gnw/nucleosom/ links 'PROFILE' 'PROFILE_LIST'

3. Database content

SRS table	Description	Number of entries
PROFILE_LIST	This SRS table accumulates sets of significant physico-chemical properties profiles of the nucleosome formation sites.	6
PROFILE	database on profiles of conformational and physico-chemical B-DNA properties	228

4. List of biological tasks that could be solved by using the PROFILES and PROFILE_LIST databases

- Searching for and characteristics of extremal values and linear trends for the profiles of conformational and physico-chemical properties of nucleosomal DNA

5. SRS table format

PROFILE_LIST

Line code	Field name	Field description
ID	ProfileListID	Identifier
AC	AccessNumber	Accession number
SD	SiteDescription	Site description

PROFILE

This SRS table accumulates significant physico-chemical properties profiles of the nucleosome formation sites.

SRS table format

Line code	Field name	Field description
ID	ProfileID	idendifier of an entry
SD	SiteDescription	site description
PW	ProfileWindow	profile window size
PA	ProfileAverage	profile average value
PD	StandardDeviation	standard deviation
ET	ExtremumType	type of extremum: minimum or maximum
EV	ExtremumValue	profile value in the position of extremum
EL	ExtremumLevel	significance level of extremum
GC	SignCriteria	criteria for gradient significance evaluation

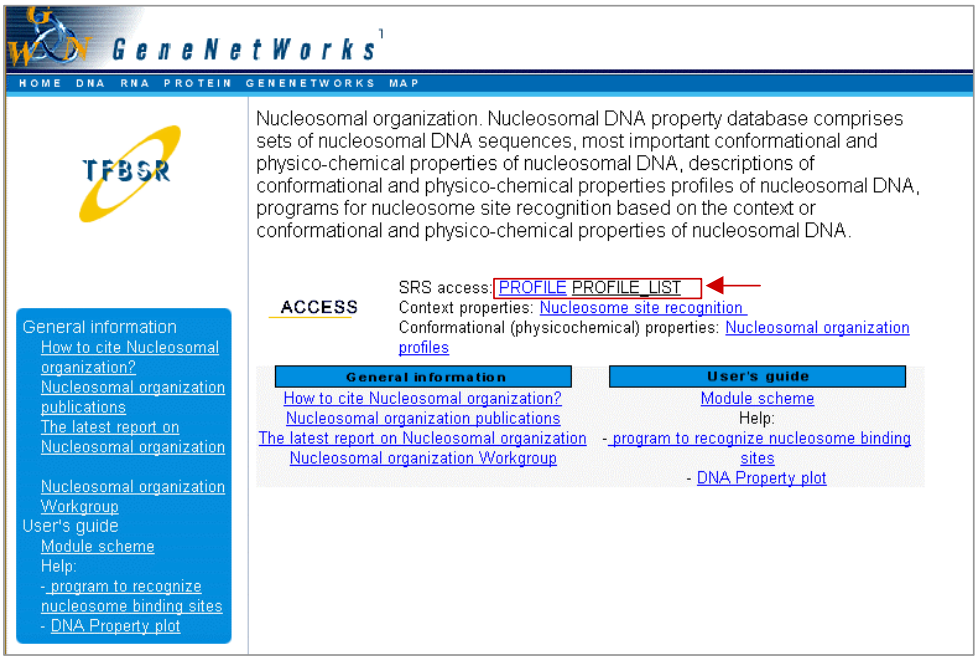
The database PROFILE_LIST contains entry headers, which present site samples and the lists of DNA conformational and physico-chemical properties. Each property in the list refers to individual property entry, which is contained in the database PROFILE. The property entry presents the profile of DNA conformational physico-chemical property and description of its extremum points and linear trends. To construct a profile, various sliding window sizes were applied. The evaluation of the best suited window size was made by comparison of the sample profile with the profile constructed for random sequences with the same dinucleotide content as that of the site. The best selected profile is included into the database entry. Then the profiles were checked for linear trends and the resulted data are also compiled in the database.

Example of SRS queries to the PROFILE database

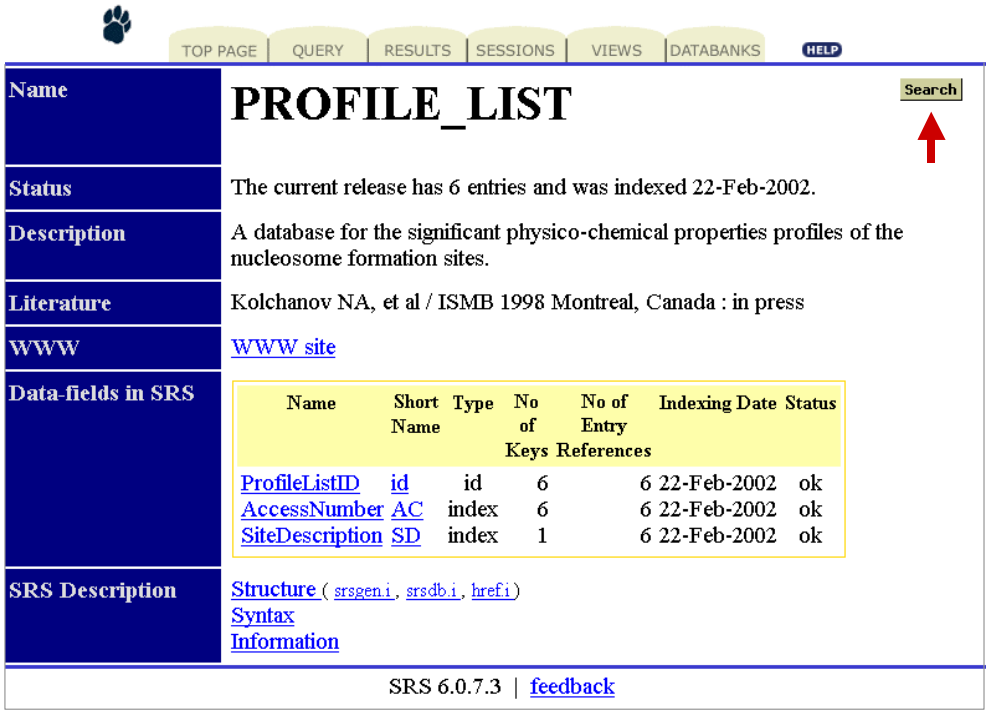
The example of querying the database is as follows:

Which conformational and physico-chemical DNA properties profiles are significant for nucleosome binding DNA-region in 5' & 3' gene regions?

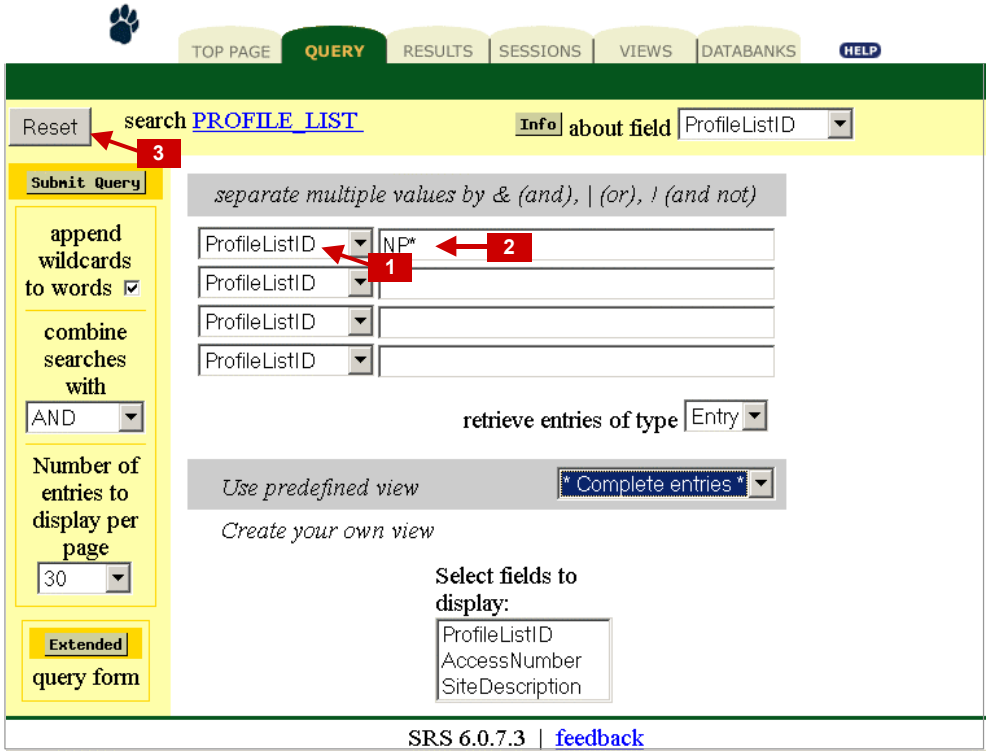
To make such a query, you should perform the following:
1. Choose ‘PROFILE_LIST’ SRS table on the page 'access to DNA NUCLEOSOMAL ORGANIZATION'.



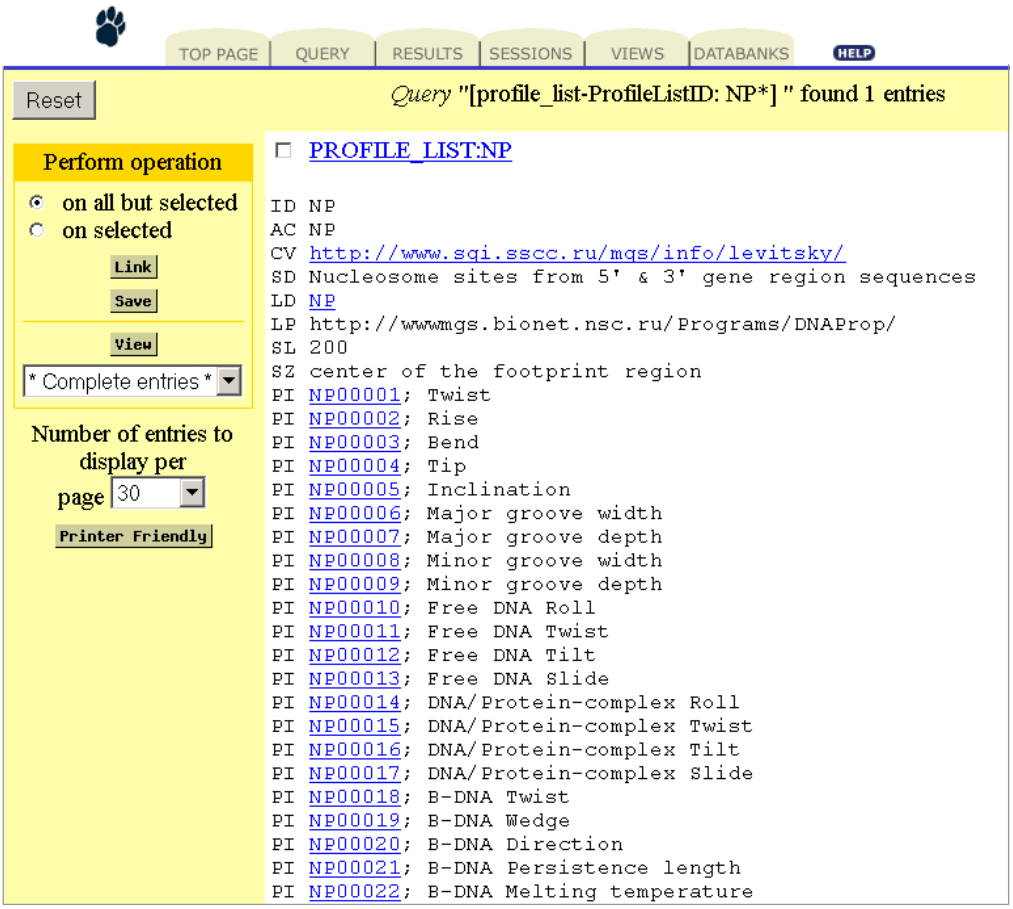
2. This will bring up the home page of the chosen PROFILE_LIST SRS table.
Click the button 'Search'.



3. Select the fields to be searched for from the list. You will need the field 'ProfileListID' (1), Type terms to be searched in the text window: 'NP*' (2)
(An asterisk means "any symbol") and click the button 'Submit Query' (3).



The query result will be displayed as the complete list of entries containing significant B-DNA property profiles of nucleosome binding DNA in 5' & 3' gene regions.
Click the link ' NP00001; Twist'.



The result will be displayed as the complete entry containing the profile description of the nucleosome binding DNA in 5' & 3' gene regions of the conformational B-DNA property 'Twist'

```
PROFILES:NP00001

PI NP00001
SD Nucleosome sites from 5' & 3' gene region sequences
PR P0000001; Twist
IP http://www.sqi.sscc.ru/Programs/acts2/images/TWIST.html
YY
CR Student criterion
FG http://www.sqi.sscc.ru/Programs/NucSitRec/images/ptwistl_1.gif
PW 9
FG http://www.sqi.sscc.ru/Programs/NucSitRec/images/ptwistl_2.gif
PA 36.538
PD 0.123
PN 2    2
ET Minimum
EN 1
EV 36.230
EL 0.02
EP -20.5
ET Minimum
EN 2
EV 36.230
EL 0.02
EP 20.5
ET Maximum
EN 1
EV 36.830
EL 0.02
EP 41.5
ET Maximum
EN 2
EV 36.830
EL 0.02
EP -41.5
YY
GC Fisher-Snedecor criterion
FG http://www.sqi.sscc.ru/Programs/NucSitRec/images/ptwistl.gif
GW 29
GX [-44.5, -16.5]
GM -30.5
GA 36.544688
GL 0.050000
GK -0.026441
GI 0.021562
GF Y(X) = 36.544688 + (-0.026441 * (X + 30.5))
YY
GW 29
GX [16.5, 44.5]
```

Comments and questions are welcome to Victor Levitsky (levitsky@bionet.nsc.ru)

2. Software

2.1. Nucleosome binding site recognition

Release 2003

1. Program description

Program generates a profile of the so-called nucleosome positioning potential.

2. Access to the program:
http://www.domain.com/mgs/gnw/nucleosom/ link 'Nucleosome site recognition'.

3. Biological task that could be solved by using the program:

Recognition of nucleosome formation sites by oligonucleotide content.

4. Input Data

Input the DNA sequence into the field 'Input DNA Sequence' (1). The sequence Length should be at least 160 bp and at most 32 kbp. The sequence should be in a plain textual format (a, t, g, c in upper- or lower case, line feeds and blanks are ignored). The sequence to be analysed should have the maximal length of 32 kb and the minimal length of 160 bp). The dinucleotide relative abundance distance was chosen as an additional restriction for input data to exclude the sequences with poor dinucleotide content. Positions of the sliding window that are ignored by the program are marked by colour in the graphical representation of the data output and by symbol *, in numerical delivery of the program.

5. Program options

The analysis of complementary DNA strand is executed by clicking the check-box 'Reverse strand' (2). To select the data output mode, choose the option by clicking the check-box 'Graphic mode' (3). To select the output profile transformation mode, choose the option by clicking the check-box 'Standardization by dispersion' (5).

6. Program execution

Click the button 'SCAN' and wait for the program execution (4).

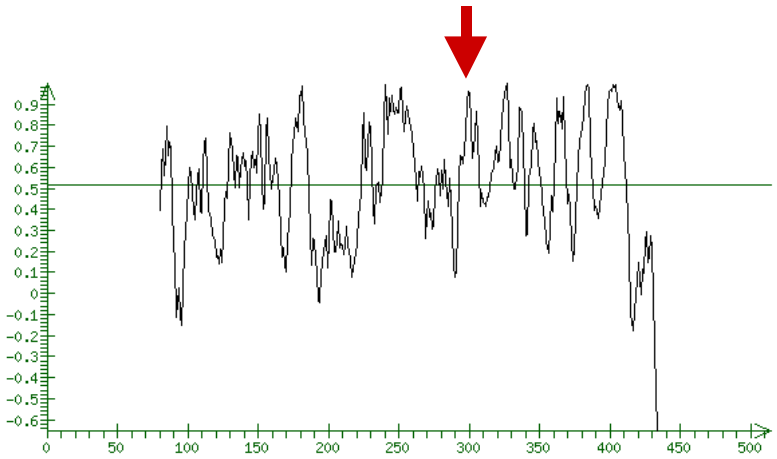
7. Output Data

Output data could be displayed in two modes: graphical representation and numerical delivery.

Example

Nucleosome potential is constructed in a way so that its mean value over the training set of nucleosome site sequences equals +1; whereas over the set of non-site (random) sequences, –1. This means that the values close to +1 correspond to a higher probability of nucleosome positioning. The option 'Standardization by dispersion' transforms nucleosome potential values to the interval ≤ 1, here value +1 corresponds to the best prediction, and interval [0; 1] corresponds to reliable nucleosome formation sites. For example, we consider here the analysis of *Xenopus laevis* TFIIIA gene, with the center of nucleosome site placed approximately at 290 position relatively the sequence start (marked in the Figure above). Click the button 'Example' to display the example of the program execution.

The graphical representation is illustrated in the Figure below.



This Figure illustrates numerical delivery.

81	0.396568
82	0.589130
83	0.686781
84	0.562456
85	0.657379
86	0.793726
87	0.653407
88	0.716841
89	0.688931
90	0.414675
91	0.231619
92	0.151196
93	-0.115555
94	-0.051597
95	0.026091
96	-0.114321
97	-0.150909
98	0.020630
99	0.209199
100	0.380815
101	0.510850

Comments and questions are welcome to Victor Levitsky (levitsky@bionet.nsc.ru)

2.2. Recognition tools

1. Program description

The nucleosome binding sites are characterized by specific sets of significant conformational and physico-chemical DNA properties. The site recognition programs are generated by the sets of significant conformational and physico-chemical DNA properties (DNA features).

2. Access to RECOGNITION TOOLS

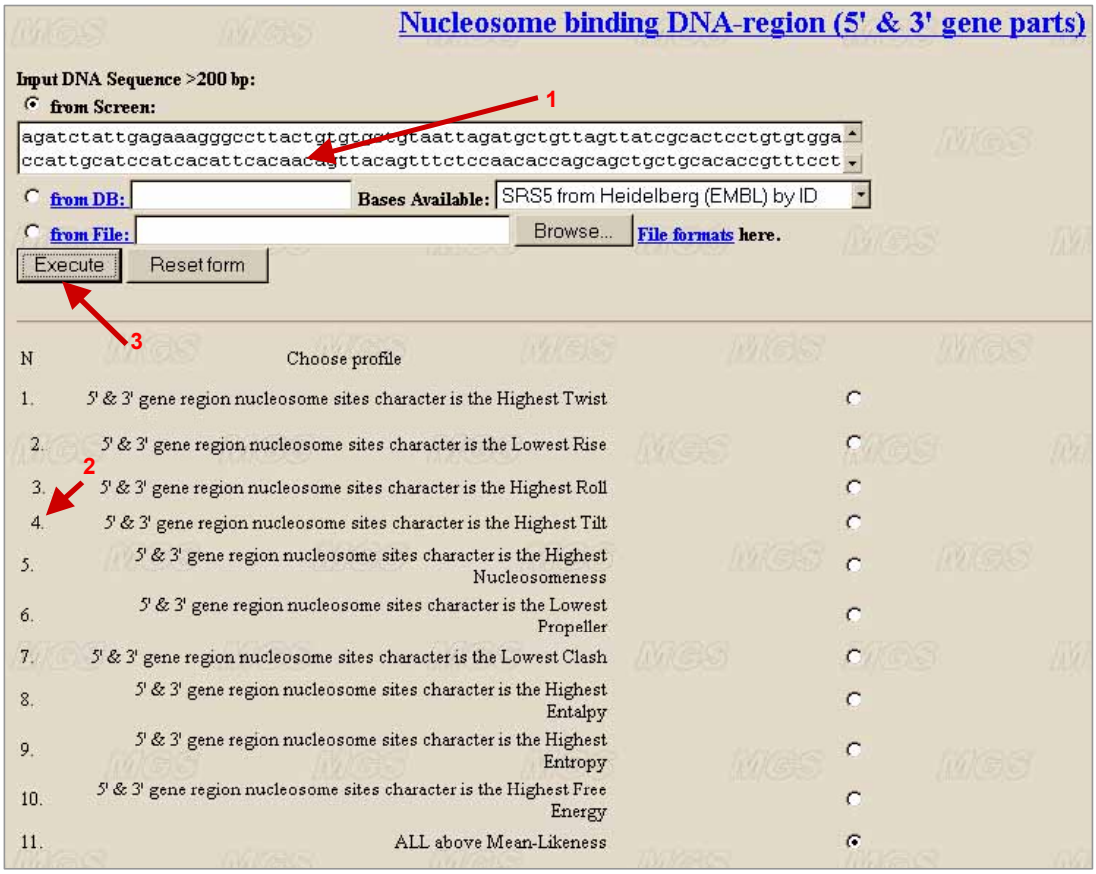
http://www.domain.com/mgs/gnw/nucleosom/ link 'Nucleosomal organization profiles'.

3. List of biological tasks that could be solved by using the program

- nucleosome site recognition by the most significant DNA conformational and physico-chemical properties.

4. Data input

Input the DNA sequence of interest into the text window (1). Sequence should be in a plain textual format (a, t, g, c in up- or low cases, spaces or tabulation are accepted). The sequence to be input for analysis should be of maximal length of 32 kbp and minimal length of 160 bp.



5. Program options

Select the necessary recognition program by clicking one of the check-boxes (2). Each check-box refer to a significant conformational or physico-chemical DNA property. Generalized (mean) recognition profile is calculated using all known significant conformational and physico-chemical features, contained in B-DNA features database.

6. Program execution

Start the tools processing by clicking the button ‘Execute’ (3).

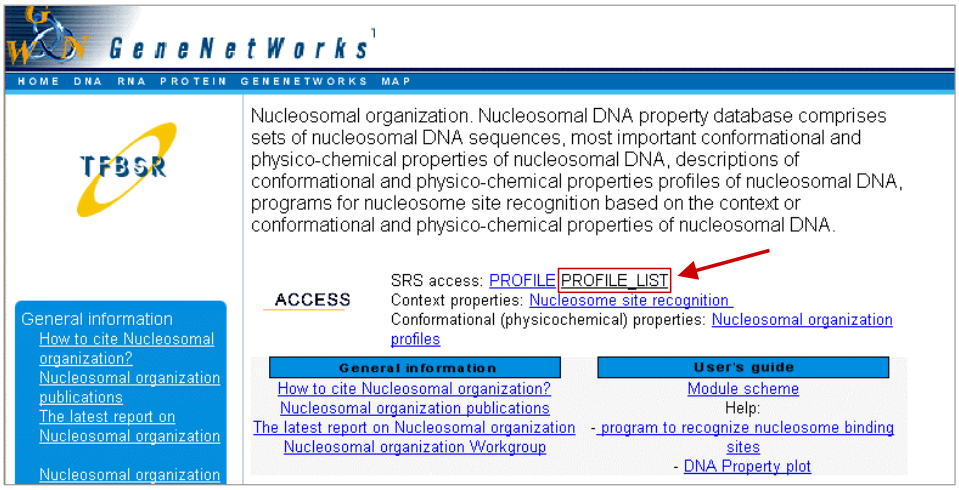
7. Data output.

The tools output represents the profile of the Score value. The positive peaks of this profile pinpoint to potential nucleosome site recognized.

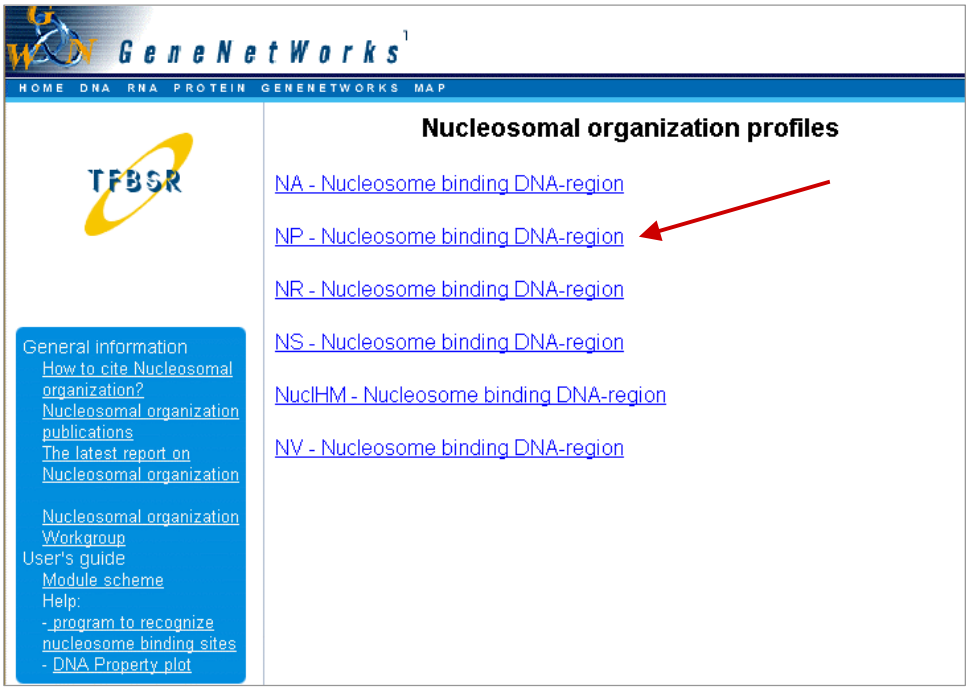
Example

Search for potential nucleosome sites of the *Xenopus laevis* TFIIIA gene 5' region and partial coding region, EMBL ID XLTF3A5.

1. On the DNA NUCLEOSOMAL ORGANIZATION home page, click the link ‘Nucleosomal organization profiles’.



2. On the page ‘Nucleosomal organization profiles’ click the link 'NP nucleosome binding DNA-region'



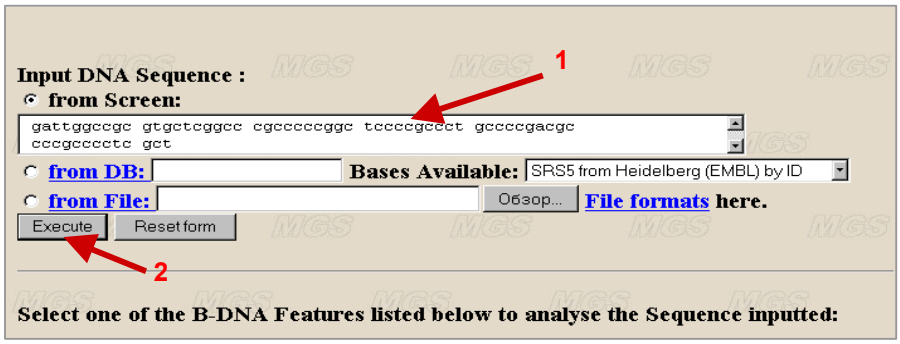
3. Input into the text-box (1) the following sequence:

```
agatctattgagaaagggccttactgtgtgctgtaattagatgctgttagttatcgcactcctgtg
tggaccattgcatccatcacattcacaacagttacagtttctccaacaccagcagctgctgcacac
cgtttcctcggcttcatgtattatcacgtgctccactaggactcaaccactaagaacgaggggagt
gtccagaaacacccaacttgtgaaataacacatcgctgacataaaacacaggaatttaacatcctt
ttttttaagtttgcgagcgcaattactgaaggctaaagctgcaagggacacaggaaagggctgatt
gccaatcctttcagacatcgcaaaacttcccgatgcatgtgcgataatggtttgtcctagagctat
ataaacaggcacacatggcggctacagtggcttctacaagttcagaggaagccgagggcagcttag
ttactgaaggagagatgggagagaaggcgctgccggtggtgtataagcggtac
```

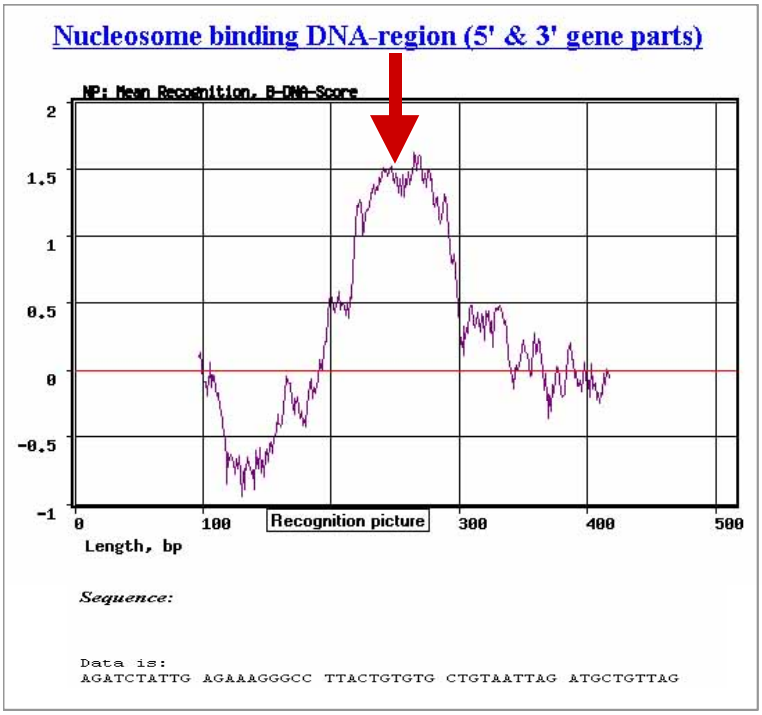
(*Xenopus laevis* TFIIIA gene 5' region and partial coding region, EMBL ID XLTF3A5).

Use program options set by default.

Click the button 'Execute' (2)



4. Data output window:



Comments and questions are welcome to Victor Levitsky (levitsky@bionet.nsc.ru)

PART II. RNA INTEGRATION LEVEL.

CHAPTER 1. LEADER RNA.

1. LEADER_RNA Knowledge Base.

LEADER_RNA Knowledge base description.

The LEADER_RNA was designed to evaluate translational activity of mRNAs of mammals, dicot plants and monocot plants.

Access to LEADER_RNA prediction program:
http://www.domain.com/mgs/gnw/leader/

LEADER_RNA content

Knowledge base LEADER_RNA is linked to the prediction program allowing to evaluate similarity of user-defined mRNA to high and low expression ones. It consists of five data files installed at the SRS platform. Note: this knowledge base may not be used efficiently by itself but as a data source for both the prediction program and mRNA features description.

SRS table	Description	Number of entries
LEADER_SQ	sequences of 5'UTRs of high- and low-expression eukaryotic mRNAs	879
LEADER_REF	brief summaries of several papers concerning mRNA translational activity.	22
LEADER_WHY	contains brief description of mRNA contextual features influencing translatability	27
LEADER_KN	contextual features of mRNA 5'UTRs significantly different between high and low expression genes and C-codes of the prediction program	34
LEADER_SCI	reference authors	2

4. List of biological tasks that could be solved by using the LEADER database

- evaluation of mRNA translational efficiency in higher plant or mammalian cells. It answers to the following question: to what extent the user-defined mRNA is similar to mRNAs of high or low expression genes? Consequently, similarity of this mRNA 5'UTR on this region of high expression genes can indicate high expression rate of its gene.
- gene engineering experiments (e.g., for prediction of translation properties of modified mRNAs in model experiments, as well as for evaluation of translation efficiency of transgene mRNA in new host organism).

LEADER_RNA knowledge base: SRS table format

Note, this knowledge base may be used efficiently only in combination with the prediction program to get info on some analysis details: it was not designed for independent usage.

Leader_SQ

Line code	Field name	Field description
MI	EntityID	Identifier of taxon-specific subdatabase
MN	EntityName	General information on the database
KN	CrossRef	Reference on LEADER_KN database
OG	GenomeRegionName	Description of stored genes
OS	Specia	Taxon info
OC	Taxon	Taxon info
FF	SiteName	Gene region
AN	ActivityName	Expression characteristics
AU	Unit of ActivityExpression	Estimation type, 1 or –1 for high and low expression mRNAs, respectively
PN	SequenceSuprempositionName	Functional site (numbered 1)
SC	SiteVariant	Identifier of corresponding EMBL entry
SQ	Sequence	5'UTR nucleotide sequence
SA	ActivityValue	sequence translational activity (1 or –1, see AU)
PA	SequenceSuprempositionPointer	Sequence length up to functional site (PA)

Leader_WHY

Line code	Field name	Field description
MI	EntityID	mRNA 5'UTR feature name
MN	EntityName	mRNA 5'UTR feature type
MD	EntityDependence	Gene region
ML	Step	Sequence feature
RN	CrossRef	Reference on Leader_REF database
PN	PropertyName	Type of analysis (models and hypothesis)
PM	Method for property definition	Type of analysis (statistical, modelling, etc)
PV	Property ID	Details of analysis scheme (training test, GA, modelling, etc)
PU	Unit for Property Expression	significance criteria
REASONS	Description	Basis of this feature usage as a discriminative characteristic in the prediction process

Leader_REF

Line code	Field name	Field description
RN	Identifier	Reference identifier
RA	Authors	Reference authors
RT	Title	Reference heading
RJ	Journal	Reference journal
RV	Vol	Reference journal volume
RP	PP	Reference journal pages
RY	Year	Reference year
RR	Abstract	Reference abstract

Leader_KN

Line code	Field name	Field description
CF	KnowledgeName	Type of mRNA feature
CT	KnowledgeNature	influence on mRNA translational activity
DP	FeatureName	Feature name
PV	SequenceFeature	Types of a sequence
AB	Values	Mean values of the features for high and low expression mRNAs
LC	LinearCorrelation	Significance of statistical difference
ST	ContrHigh	Results of prediction of a high expression control set: <mean values> <standard errors> <the percentages of incorrect prediction>
NN	ContrLow	Results of prediction of a low expression control set: <mean values> <standard errors> <the percentages of incorrect prediction>
C-CODE	CODE	C-code of the prediction program

Comments and questions are welcome to Alex Kochetov (ak@bionet.nsc.ru)

2. Software.

2.1. Programs for mRNA Translatability Prediction

Release 2003

Program description:

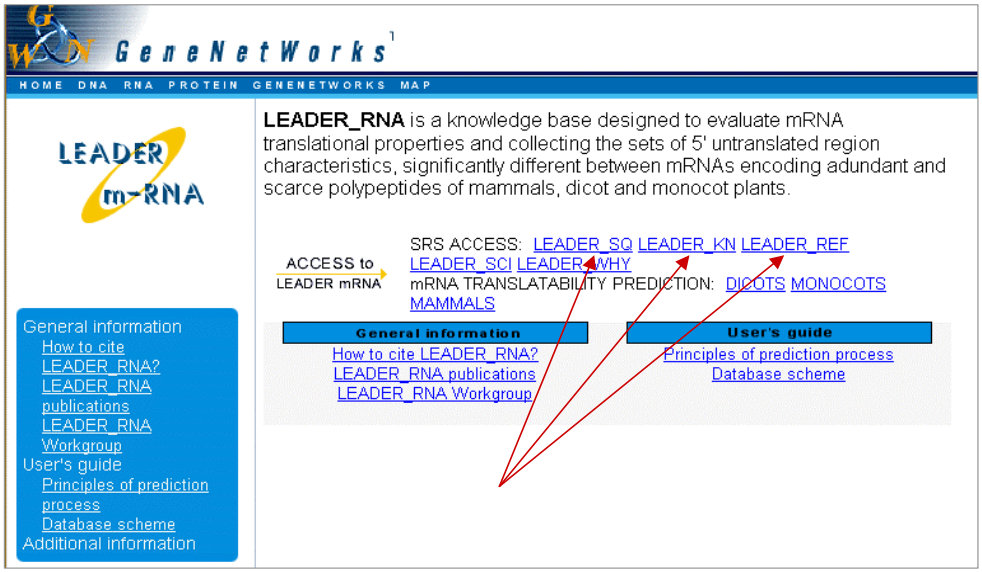
LEADER_RNA contains distinct programs for prediction of mRNA translation efficiency in organisms of each of these taxa: dicot plants, monocot plants and mammals.

List of biological tasks that could be solved by using Prediction programs:

- evaluation of mRNA translational efficiency in higher plants or mammals. The program estimates to what extent the mRNA defined by a user is similar to mRNAs of high or low expression genes? Consequently, similarity of a particular mRNA 5'UTR to that of high expression genes possibly may indicate the fact that this gene is also expressed at high expression rate.
- gene engineering experiment design (e.g., for prediction of translation properties of modified mRNAs in model experiments, as well as for evaluation of translation efficiency of transgene mRNA in a novel host organism).

Access to Prediction programs:

http://www.domain.com/mgs/gnw/leader/ link 'DICOTS' (dicot plants)
http://www.domain.com/mgs/gnw/leader/ link 'MONOCOTS' (monocot plants)
http://www.domain.com/mgs/gnw/leader/ link 'MAMMALS' (mammals)



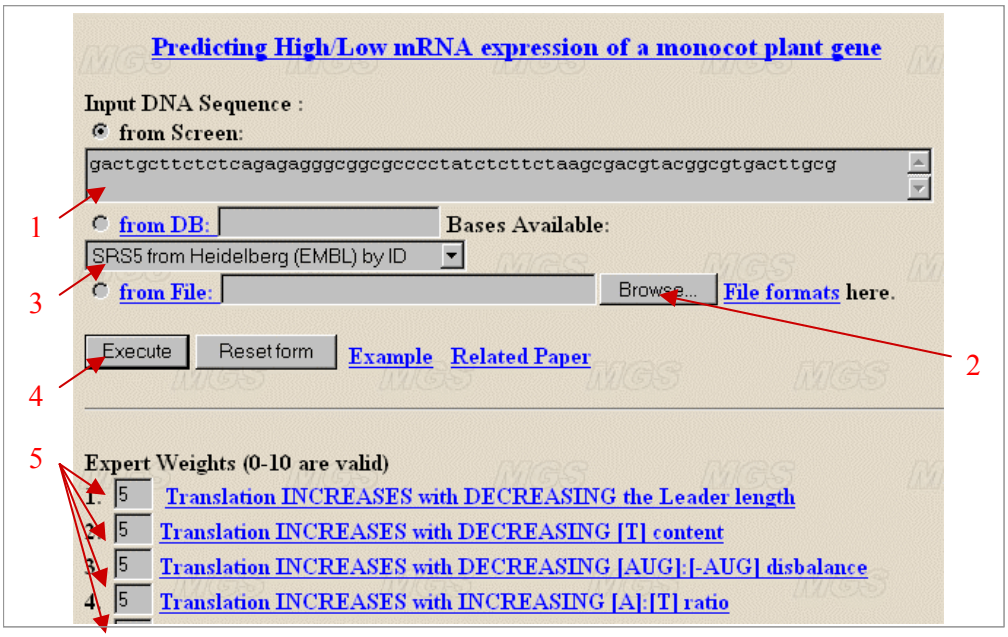
Data input

Input window of LEADER_RNA prediction program is presented below in the Figure. A user may select one of three available taxa (dicot plants, monocot plants, and mammals).

Next window contains input text-box for 5'UTR sequence and the list of discriminative features (determined by choosing of appropriate taxon at the previous page). It is possible to input 5'UTR sequence by several ways:

- from screen (by typing the sequence from the keyboard) (1).

- from file at your PC (by browsing it from the user's PC hard disk). For this operation, you should use the file with a single 5'UTR sequence in a FASTA format. Click the button 'BROWSE' (2) and choose the source file.
- from databases (currently, only EMBL and GeneBank are included in the database list; in these cases the entries should contain 5'UTR sequence only; in future, we plan to add the links to specialised mRNA 5'UTR databases) (3).



Essential notes:
(1) 5'UTR sequences should be entered in upper- or lower-case letters; both T and U letters are allowed (anyway, T is considered as U); line feeds and blanks are ignored;
(2) DO NOT include start codon of the ORF in 5'UTR sequence. INCLUDE 5'UTR up to -1 nucleotide (located just upstream AUG codon of the main ORF):
(3) Try to use COMPLETE 5'UTR sequence (if possible) to increase prediction accuracy.
(4) You may choose weights (see 5 in the Figure) of discriminative features (ranging from 0 to 10; 5 by defaults). This feature is likely to be helpful for users experienced in translation machinery organization in eukaryotes.

Program options

The list of discriminative features in the figure demonstrating the input window of the program is marked by number 5. It contains various contextual characteristics used in prediction process. By default, all these features are used with the same expert weights. Prediction program is based on the mean recognition approach, so the contribution of a single feature into the final evaluation is likely not so critical. The simplest way is to use the default options (each parameter equals to 5).

However, this program provides a user with opportunity to change the expert weights of discriminative features in order to optimise prediction process. Each 5'UTR feature listed is supplied with a brief description allowing to evaluate its significance and to assign the weight. In some cases, it may be useful to apply features, which significance was supported experimentally (e.g., weight of start codon context, the presence of leader AUG triplets and weights of their contexts).

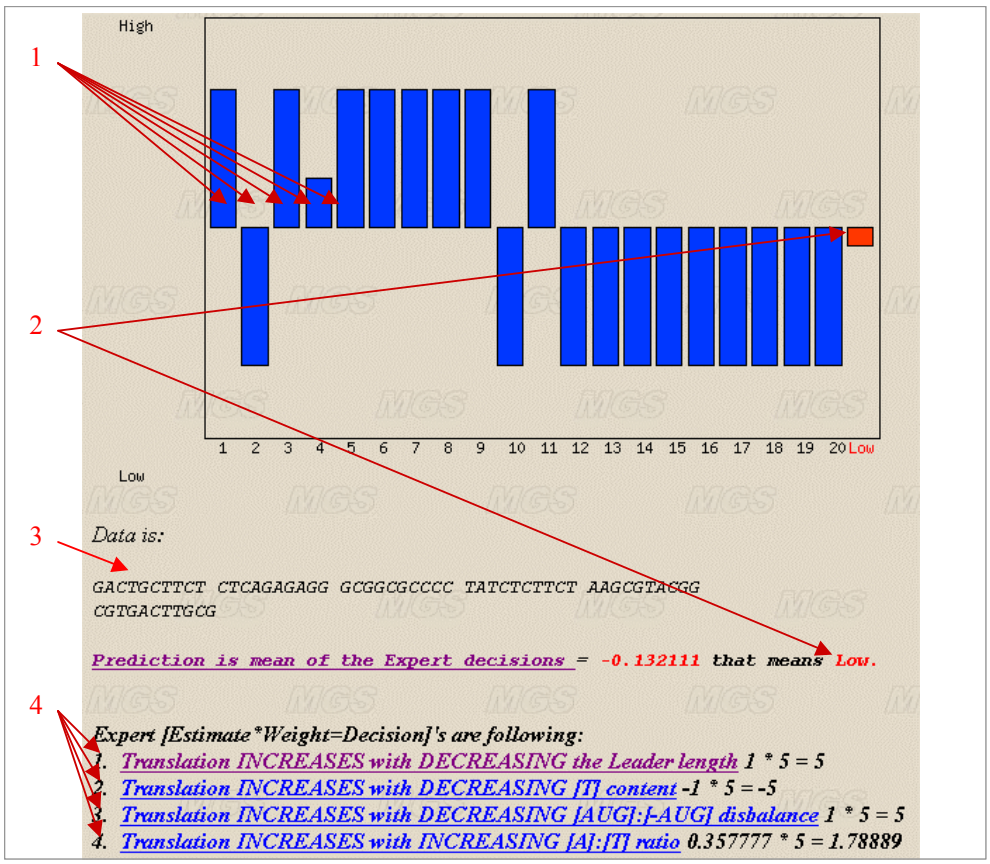
Program execution

Click the button ‘Execute’ (4).

Data output

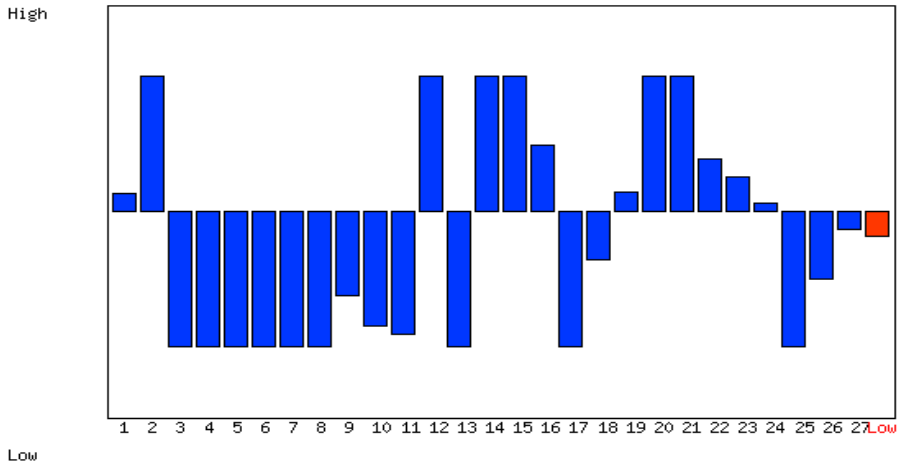
This window presents the results of prediction (demonstrated both in the figure and table). Values of each of discriminative feature are shown (4). It is possible to evaluate similarity of 5’UTR analysed with those of high and low expression mRNAs of corresponding taxon (1).

Each field in the table of discriminative features (4) contains links to LEADER_RNA knowledge base allowing extract additional information. The input sequence is also shown (3). Integral prediction result is marked by 2.



Examples

(1) Random sequence with equal shares of nucleotides (default parameters were used for prediction).



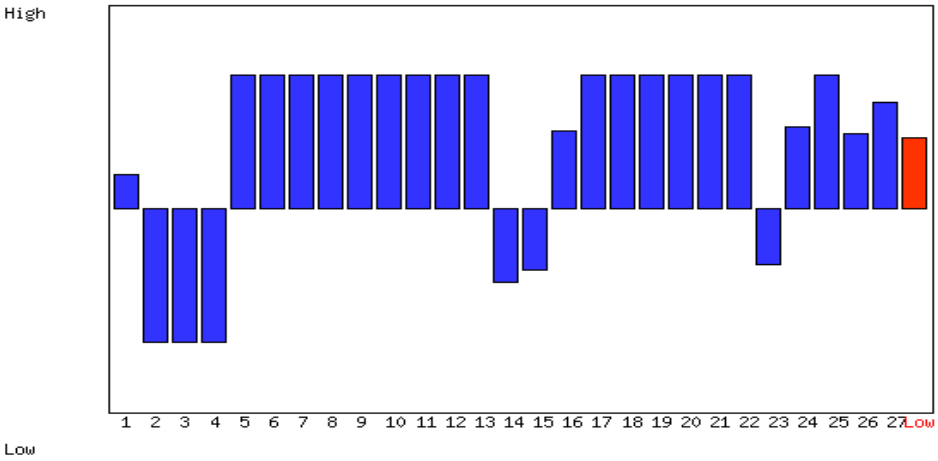
Data is:

TGGCCGGTGG CTTATGGACT GTCACAGCTC GTGTCACTGA TGAAGATAAG
GGTATCCTCC CGCTGTCGGA CCATAAAGAA CCACACGATC ATTTTAATCG

***Prediction is mean of the Expert decisions* = -0.181138 that means Low.**

Expert [Estimate*Weight=Decision]'s are following:
1. Translation INCREASES with DECREASING the Leader length 0.137257 * 5 = 0.686285
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 -0.623815 * 5 = -3.11908
10. Translation INCREASES depends on the "-3 position" rule -0.840345 * 5 = -4.20173
11. Translation INCREASES with DECREASING [AUG] "-3"-ruled -0.903936 * 5 = -4.51968
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 1 * 5 = 5
15. Translation INCREASES with DECREASING Low-consensus matches 1 * 5 = 5
16. Translation INCREASES with INCREASING High-ShortFreqMatr 0.483426 * 5 = 2.41713
17. Translation INCREASES with INCREASING High/Low 1bp-FreqRatio -1 * 5 = -5
18. Translation INCREASES with INCREASING High/Low 1bp(KM)-FreqRatio -0.360723 * 5 = -1.80362
19. Translation INCREASES with INCREASING High/Low 2bp(KM)-FreqRatio 0.146544 * 5 = 0.73272
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 0.3873 * 5 = 1.9365
23. Translation INCREASES with INCREASING High/Low 3bp(ATGCx)FreqRatio 0.250668 * 5 = 1.25334
24. Translation INCREASES with INCREASING High/Low 5bp(ATGCx)FreqRatio 0.057873 * 5 = 0.289365
25. Translation INCREASES with INCREASING High/Low 3bp(KxM)-FreqRatio -1 * 5 = -5
26. Translation INCREASES with INCREASING High/Low 5bp(KxM)-FreqRatio -0.497674 * 5 = -2.48837
27. Translation INCREASES with INCREASING High/Low 7bp(KxM)-FreqRatio -0.127296 * 5 = -0.63648

(2) 5'UTR sequence of dicot plant gene with the high level of expression (EMBL accession number: AT30SRS13) (default parameters were used for this prediction).

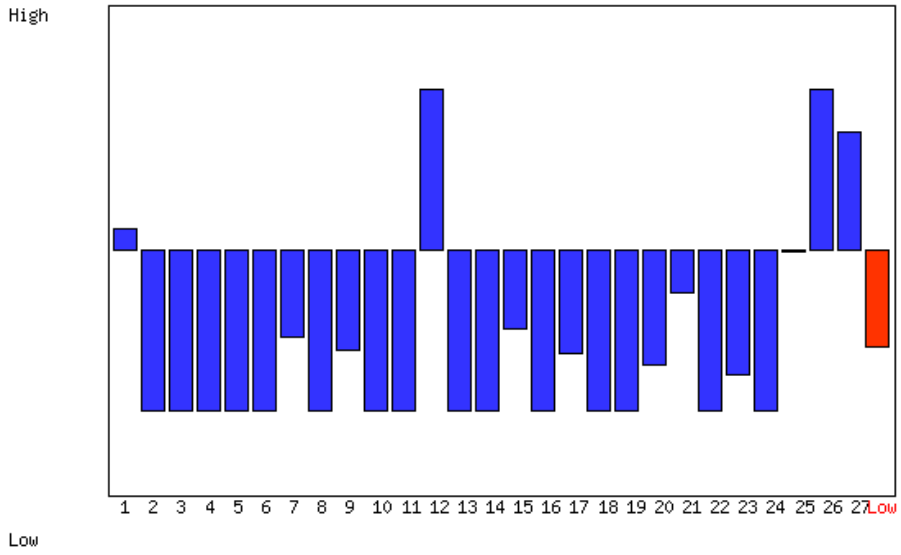


Data is:
CGTTTGCCTT ATCCGTTCAG CTCATCTTCT TCTTCTTCTT CGTCACCTCT
GAATTAGTTT CCCAGAATCC GAAATTCCTA GGAAGAGAAC ATAACA

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**

(3) 5'UTR sequence of dicot plant gene with low level of expression (EMBL accession number: AT31834) (default parameters were used for the prediction).



Data is:

TTCAACAACT GTGTTATCTC TATGAACAGT TCCTTTGATG AAGAACACAA
AGTGAAAGTT GCTGTCTTTA TAACCAGGAT TTGGTAATTC CCATTGTTTC

Prediction is mean of the Expert decisions* = -0.598938 *that means Low.

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

1. *Translation INCREASES with DECREASING the Leader length 0.137257 * 5 = 0.686285*
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 -0.537042 * 5 = -2.68521*
8. *Translation INCREASES with DECREASING [AUG] framed -1 * 5 = -5*
9. *Translation INCREASES with DECREASING [AUG] optimized -0.623815 * 5 = -3.11908*
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 -1 * 5 = -5*
15. *Translation INCREASES with DECREASING Low-consensus matches -0.487738 * 5 = -2.43869*
16. *Translation INCREASES with INCREASING High-ShortFreqMatr -1 * 5 = -5*
17. *Translation INCREASES with INCREASING High/Low 1bp-FreqRatio -0.639225 * 5 = -3.19612*
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 -0.708711 * 5 = -3.54355*
21. *Translation INCREASES with INCREASING High/Low 5bp(KM)-FreqRatio -0.264822 * 5 = -1.32411*
22. *Translation INCREASES with INCREASING High/Low 6bp(KM)-FreqRatio -1 * 5 = -5*
23. *Translation INCREASES with INCREASING High/Low 3bp(ATGCx)FreqRatio -0.776691 * 5 = -3.88346*
24. *Translation INCREASES with INCREASING High/Low 5bp(ATGCx)FreqRatio -1 * 5 = -5*
25. *Translation INCREASES with INCREASING High/Low 3bp(KxM)-FreqRatio -0.00717 * 5 = -0.03585*
26. *Translation INCREASES with INCREASING High/Low 5bp(KxM)-FreqRatio 1 * 5 = 5*
27. *Translation INCREASES with INCREASING High/Low 7bp(KxM)-FreqRatio 0.736632 * 5 = 3.68316*

Comments and questions are welcome to Alex Kochetov (ak@bionet.nsc.ru).

2.2. MatrixSS: Building of E-score plot for RNA sequence

Release 2003

Program description: 'MatrixSS' enables searching for the regions with the stable secondary RNA structure in extended (from 250 to 100000 nt) genome sequences.

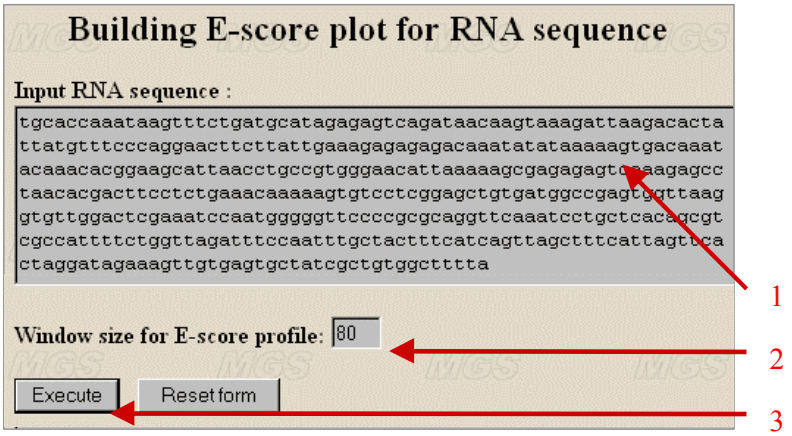
Access to 'MatrixSS':
http://www.domain.com/mgs/gnw/garna/ link 'Primary structure analysis (MatrixSS program)'

Biological tasks that could be solved by using the 'MatrixSS':

- The program MatrixSS is designed for searching in genome sequences for the genes encoding RNA with a potentially significant secondary structure (SS).

Data input

Into the text-box (1), enter or insert from the clipboard the RNA sequence to be analysed. The sequence should be within the range of 250 – 100000 nt in length. Use the alphabets ATGC(atgc) or AUGC(augc) (any other symbol will give an error). Blanks or line folding are ignored.



Program options

Set the window length for calculation of the E-score value (2). The window length should correspond to the length of the RNA molecule searched for (by default, the window length equals to 80 nt, which is the length of the tRNA molecule).

Program execution

Click the button 'Execute' (3).

Data output

As the data output, two graphical plots are presented. The first one shows the distribution of the E-score value along the sequence analysed. The dotted line marks the mean E-score value typical for the structural RNAs (tRNA, rRNA, etc.). The solid lines to the up and down of the dotted line mark the limits of E-score variation in the class of structural RNA. The regions of the resulting line, which lie between these limits, are marked by red. These are the regions that potentially encode structural RNA genes.
The second graphical plot is a coloured E-score matrix. This matrix describes cross-complementarity of the potential or real RNA regions to each other. The colour of the matrix

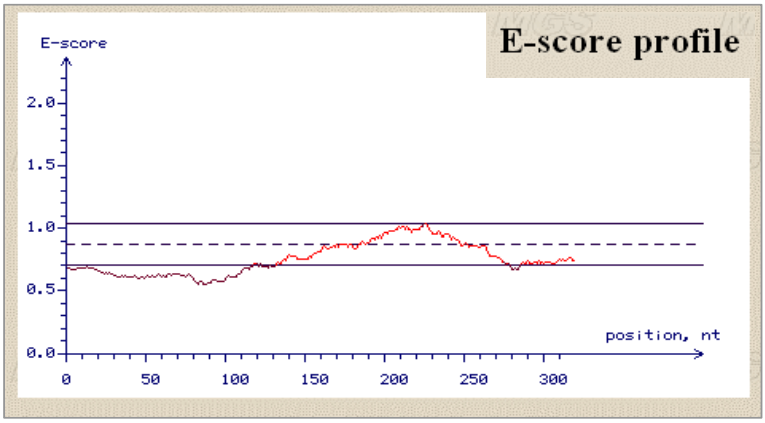
elements varies with respect to the expected stability of RNA SS, which is formed by two these RNA regions. Red colour corresponds to the higher values of expected stability.

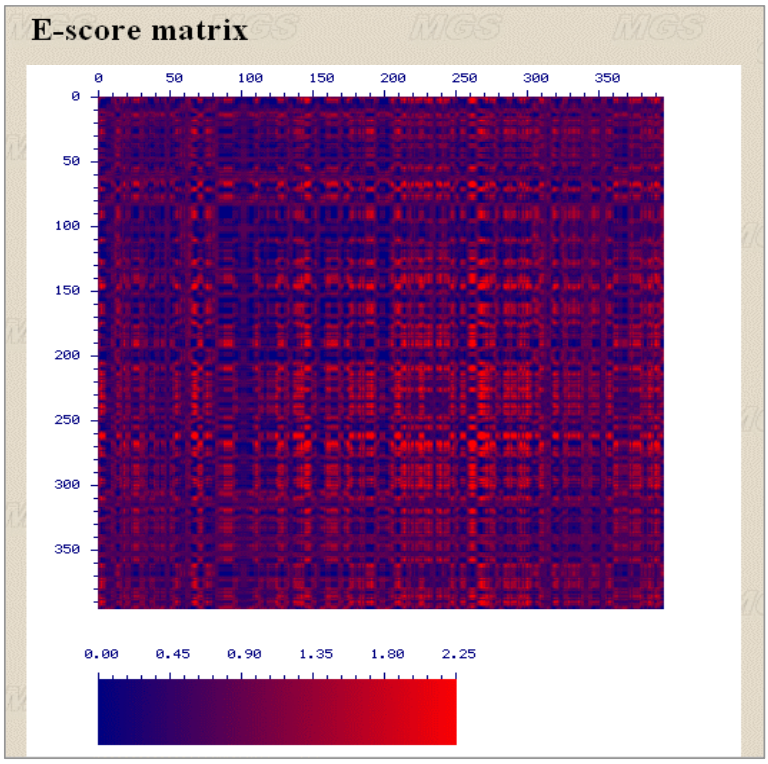
Example

As an example, we search for the tRNA Ser sequence (of 82 nt in length), which is present in the sequence extracted from the EMBL databank (ID AF184043). The total length of this sequence is 400 nt. The tRNA Ser gene is located in-between positions 218 - 299 nt.

In this search we use 'window size' equaling to 80.

The output is represented by the following plots:





As seen in the Figure, by the example of tRNA Ser, the region of predicted potential SS contains the real tRNA gene, that is, the program made correct prediction.

Comments and questions are welcome to Igor Titov (titov@bionet.nsc.ru)

2.3. GArna Program

Release 2003

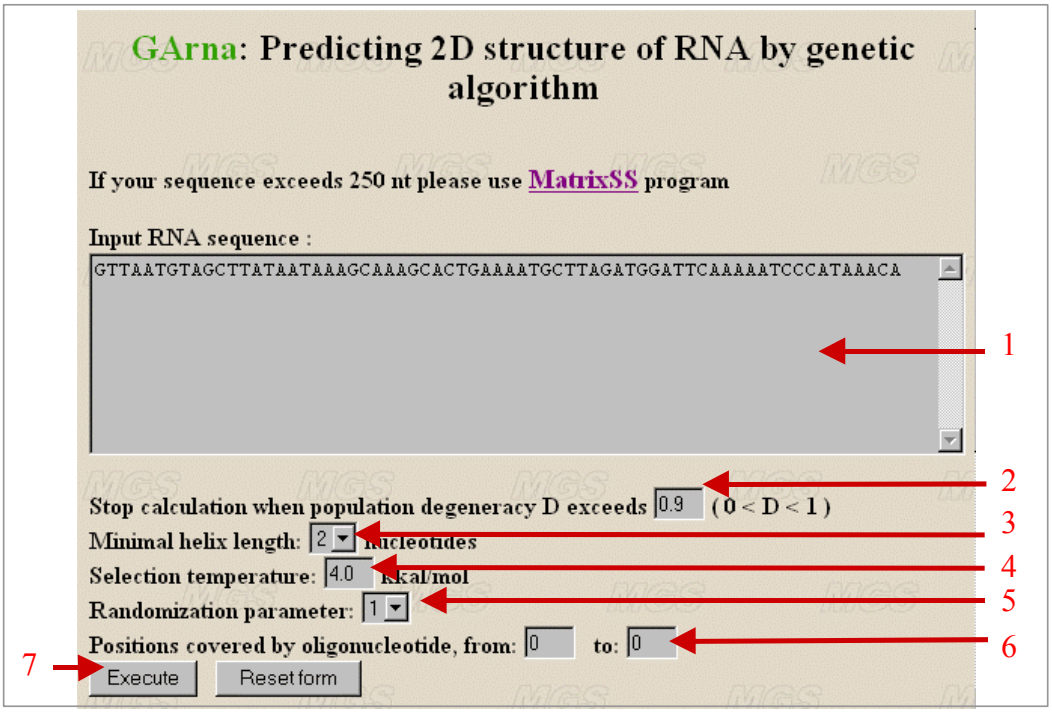
1. Program description:
GArna is a program for secondary structure prediction in short (up to 250 nt) RNA sequences and its visualization.

2. Access to 'GArna':
http://www.domain.com/mgs/gnw/garna/ than follow link ' Secondary structure prediction (GArna program)'

3. Biological tasks that could be solved by using the GArna program:
- to calculate and visualise optimal secondary structure in short (at most 250 nt) RNA sequences
- to account available experimental information about single strand RNA regions
- to evaluate stability of the secondary RNA structure in comparison to the random sequences of the same length and nucleotide content (by calculating Z-score value of the secondary structure energy)
- to model rearrangements of the secondary RNA structure under the action of the anti-sense oligonucleotide.

4. Data input
Into the text-box, enter or insert from the clipboard the RNA sequence to be analysed. The sequence should be within the range of 5 – 250 nt in length. Use the alphabets ATGC or AUGC (any other symbol will give an error). Blanks or line feeds are ignored.

5. Program options
The program GArna has the following parameters:
- Stop criterion D (option 2 on the figure). This value varies within the range $0 < D < 1$. The less is D, the more quick is calculation and less the prediction accuracy (default value giving the most accurate prediction is 0.9). The low stop criterion values (about 0.5) may be applied for searching for alternative structures and/or kinetic intermediates for particular RNA.
- Minimal helix length (option 3 in the figure). This parameter determines the least stem length necessary to form SS. Default value is 2. Minimal value equals to 2, maximal value equals to 6. The more is 'Minimal helix length' value, the more quick is calculation and less the accuracy of the secondary structure prediction.
- Selection temperature (option 4 on the figure). It is recommended to use default value 4 kcal/mol. Any other value gives less precise prediction.
- Randomization parameter (r.p.; option 5 in the figure). Any integer value is allowed. If the particular r.p. is ordered (under constancy of the rest parameters), the program always outputs one and the same result. By varying r.p., different results could be obtained. This parameter may be used, for example, to determine the robustness of prediction.
- Positions covered by oligonucleotide (from X to Y) (option 6 in the figure). This parameter indicates the range of nucleotide positions, which are prohibited to form the duplexes with the other nucleotides in the RNA molecule studied. This parameter is used (a) for calculation if the anti-sense oligonucleotide is present, (b) for calculations using *a priori* information about single-stranded regions of SS. X and Y values could vary within the range from 1 to sequence_length, X is less or equals to Y. By default, the values X=0, Y=0 are set. These values mean that all nucleotides in the sequence analysed may form duplexes (since numeration in the input sequence begins with 1). For calculations with non-zero X and Y values, z-score value is not specified.



6. Program execution

Click the button 'Execute' (7).

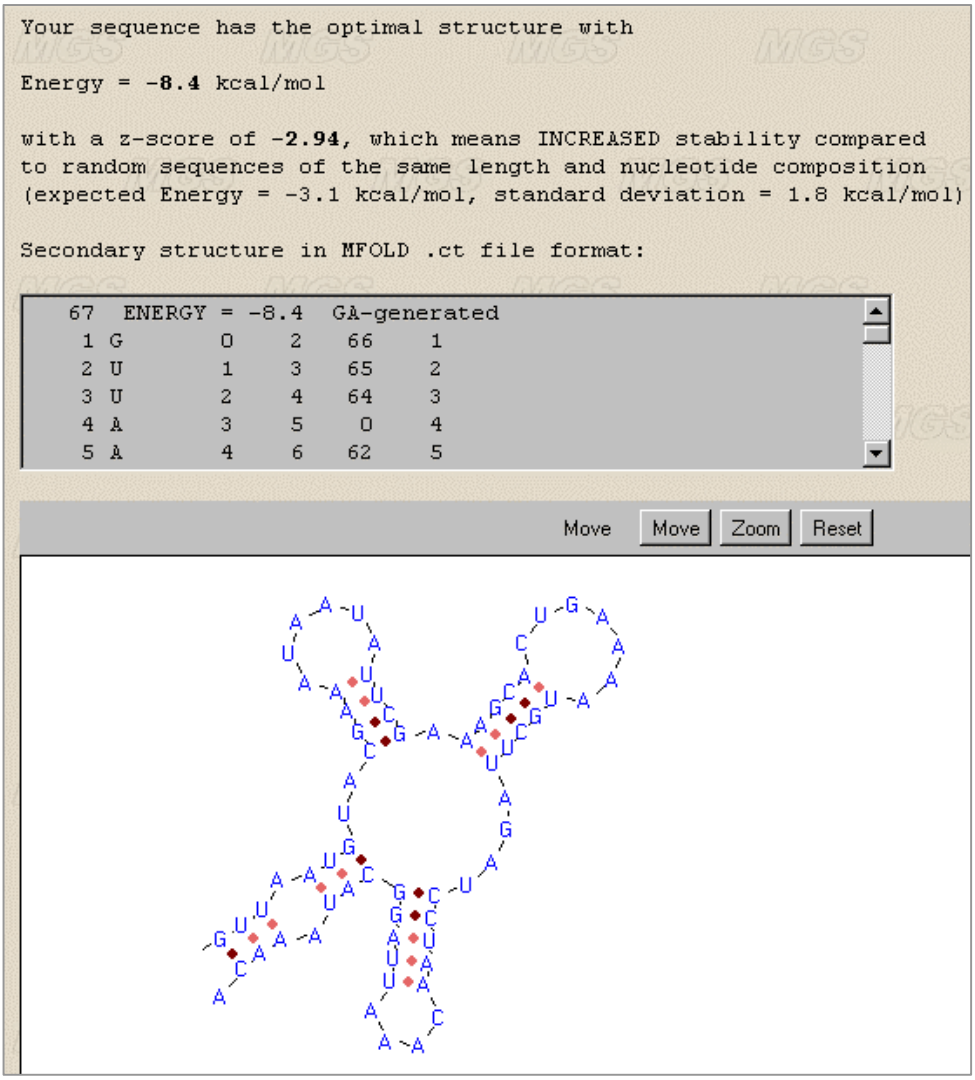
7. Data output

Program execution will bring up the resulting window with the data output. They include the following:

- the energy of the secondary structure calculated;
- deviation of stability of the secondary structure calculated from expected value for the random nucleotide sequences of the same length and content (z-score);
- graphical representation of the secondary structure calculated;
- representation of the calculated secondary structure in the format CT, which is convenient to use while exploiting the other programs (RNAstructure, GCG, etc.) for analysis of secondary RNA structure. [CT File Format description. A CT (Connectivity Table) file contains secondary structure information for a sequence. It may contain multiple structures for a single sequence. When entering a structure to calculate the free energy, the following format shown below should be used. The number at the beginning of the first line is the number of bases in the sequence. Next is the title of the structure. Each of the following lines provides information about a base in the sequence, each base being described in its own line. First is the base number, n, followed by the base denotation (A,C,G,U, or T). The third column is n-1, whereas the fourth column is n+1. The fifth column is the number of the base to which n is paired. The lack of pairing is indicated by 0 (zero). The last column is called the natural numbering.]

Example

The calculation of rat Phe tRNA secondary structure (the entry DF5280 in the database accumulating tRNA sequences; http://www.uni- bayreuth.de/departments/biochemie/sprinzl/trna/). All options for this example were set to default values.



Comments and questions are welcome to Igor Titov (titov@bionet.nsc.ru)

PART III. PROTEIN INTEGRATION LEVEL.

CHAPTER 1. DATABASES.

1. EnPDB

Release 2003

1. Database description:
ENPDB is search system for query implementation to the PDB (Protein Data Bank).

2. Access to ENPDB database:
http://www.domain.com/mgs/gnw/enpdb/

3. Database content

SRS table	Description	Number of entries
ENPDB	Search system for query implementation to the PDB	16777

4. List of biological tasks that could be solved by using the ENPDB database

- To provide extended querying of the PDB database accumulating spatial protein structures and their functional characteristics.

5. SRS table format

ENPDB

Line code	Field name	Field description
ID	ID	PDB ID code. This identifier is unique for PDB
HEADER	Header	PDB classification of the entry.
DATE	Date	Deposition date is the date when the coordinates were received by PDB. The date field contains mainly the date when the entry was created, always stored in the index as an eight-digit number of the format "yyyymmdd" (y = year, m = month, and d = day), e.g., "19940117". It is also possible to type the date to be searched in a different, more intuitive, format: "dd-mmm-yy" or "dd-mmm-yyyy", e.g., "1-jan-97" or "01-jan-1997".
TITLE	Title	This field contains the title for experiment or analysis described in the entry. The field content corresponds to that in PDB.

Line code	Field name	Field description
COMPOUND	Compound	It describes the macromolecules contained in the entry. Each macromolecule of the entry is defined with a set of token: value pairs, and is referred to as a component of the COMPOUND field. The field content corresponds to that in PDB. For each macromolecular component, the molecule name, synonyms, number assigned by the Enzyme Commission (EC), and other relevant details are specified.
COMPOUND	Molecule	This specialised field is not present in the entry as a separate line. It contains names of macromolecules from the COMPND of PDB and is designed to search for entries by the names of macromolecules.
COMPOUND	Synonym	This specialised field is not present in the entry as a separate line. It contains synonymic names of macromolecules from the COMPND of PDB and is designed to search for entries by the synonymic names of macromolecules.
COMPOUND	EC	It contains the Enzyme Commission number associated with the molecule. If there is more than one EC number, they are presented as a comma-separated list.
COMPOUND	BioUnit	If a MOLECULE functions as a part of a larger biological unit, the entire functional unit may be described. The field content is selected from the COMPND of PDB using the token BIOLOGICAL_UNIT.
SOURCE	Gene	It identifies the gene through the gene names taken from the SOURCE field of PDB.
MOL_SOURCE	MolSource	It specifies biological and/or chemical sources of all the biological molecules in the entry. There are three values: BIOLOGICAL, SYNTHETIC, and MIXED. SYNTHETIC means that all the molecules with the entry were chemically synthesised; BIOLOGICAL, all the molecules were not synthesized; and MIXED, some molecules were synthesized, whereas the rest are natural.
SOURCE	Source	This field specifies the biological source of each biological molecule in the entry. Sources are described by both the common and scientific names, e.g., genus and species. Strain and/or cell line for immortalized cells are given when they help to uniquely identify the biological entity studied. Note that the content of this field is not a replica of the SOURCE of PDB. The original PDB filed is divided into two parts: all the information concerning the biological source is retained in this field, while all the data related to synthesis is comprised in the new field SYNTHESIS. We believe that this division allows user to specify the region to be searched for more precisely.
SOURCE	Synthesis	This field specifies the data on expression systems, e.g. strain, variant, cell line, etc. The content originates from the SOURCE of PDB. See also the description of Field Source.
KEYWORD	Keyword	It contains keywords describing the macromolecule. The content corresponds to that of KEYWDS of PDB.
TECHNIQUE	Technique	It identifies the experimental technique used. This may refer to the type of radiation and sample, or include the spectroscopic or simulation technique. The content originates from the EXPDTA of PDB.
AUTHOR	Author	It indicates the names of the experts responsible for the contents of the entry and corresponds to the AUTHOR field in PDB.
JRNL	Jrnl	It indicates the reference to original publication that describes the experiment and defines the coordinate set. Its content originates from the JRNL field of PDB.

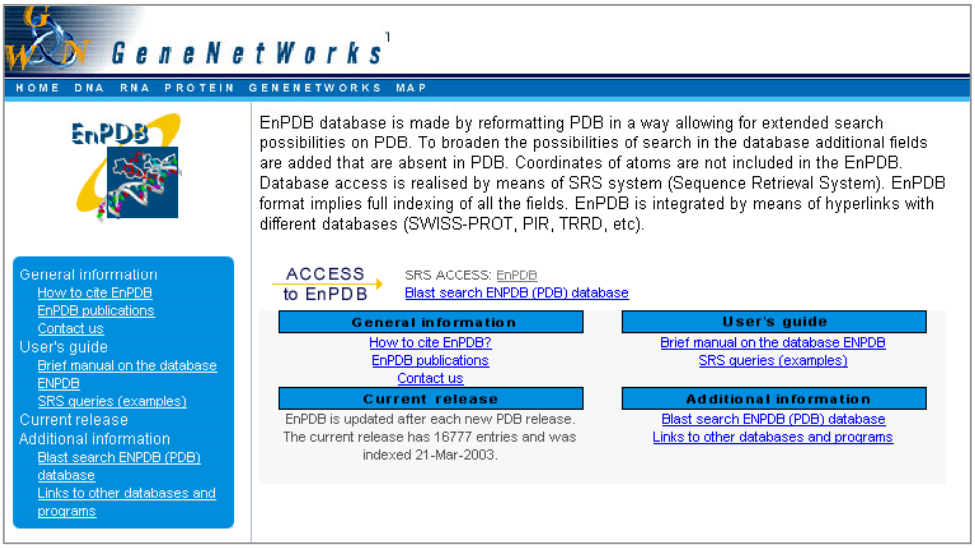
Line code	Field name	Field description
JRNL AUTH	JrnlAuthor	It contains the list of authors of the paper cited or contribution to a larger work. Its content originates from the JRNL field of PDB.
JRNL TITL	JrnlTitle	It specifies the title of the reference and is used for the title of a journal article, chapter, or part of a book. Its content originates from the JRNL field of PDB.
JRNL REF	JrnlRef	It contains name of the publication. Its content originates from the JRNL field of PDB.
JRNL VOLUME	JrnlVolume	It contains the volume of the publication. Its content originates from the JRNL field of PDB.
JRNL YEAR	JrnlYear	It indicates the year of the publication. Its content originates from the JRNL field of PDB.
REMARK_1	Remark_1	It lists important publications related to the structure described in the entry. These citations are chosen by the depositor. The content originates from the REMARK 1 of PDB.
RESOLUTION	Resolution	It is derived from REMARK 2 in the PDB file. No resolution is given for NMR structures and models. The field indicates the highest resolution in Angstroms used in building the model.
CHAIN_AMOUNT	ChainAmount	It indicates the number of chains in the entry, calculated from the SEQRES field of PDB.
CHAIN_SIZES	ChainSizes	It specifies the lengths of the chains in the entry, calculated from the data contained in the SEQRES of PDB.
HELIX_AMOUNT	HelixAmount	It indicates the number of helices in the entry and is derived from the MASTER field of PDB.
SHEET_AMOUNT	SheetAmount	It indicates the number of beta-sheet structures in the entry and is derived from the MASTER field of PDB.
DNA_RNA_AMOUNT	DnaRnaAmount	It specifies the number of DNA/RNA strands in the entry, calculated from the SEQRES field of PDB.
PROTEIN_AMOUNT	ProteinAmount	It indicates the number of protein chains, calculated from the SEQRES field of PDB.
HET_AMOUNT	HetAmount	It indicates the number of unusual residues, such as prosthetic groups, inhibitors, solvent molecules, and ions, supplemented with their coordinates. The data are calculated from the HET field of PDB.
HETEROGEN	Heterogen	It gives the chemical name and the synonyms of unusual residues, such as prosthetic groups, inhibitors, solvent molecules, and ions, supplemented with their coordinates. The data are calculated from the HETNAM and HETSYN fields of PDB.
LINK_EMBL	LinkEmbl	It links to EMBL Data Bank through SWISS-PROT. For example, we find a SWISS-PROT entry with references to both PDB and EMBL entries. In this case, we consider that the PDB and EMBL entries are linked.
LINK_PIR	LinkPir	It links to PIR Data Bank through SWISS-PROT.
LINK_SWISS-PROT	LinkSwissProt	It links to SWISS-PROT Data Bank as its entries contain references to PDB.
LINK_TRANSFAC	LinkTransfac	It links to Transfac Data Bank through SWISS-PROT. For example, we find a SWISS-PROT entry with references to both PDB and TRANSFAC entries. In this case, we consider that the PDB and TRANSFAC entries are linked.
LINK_TRRD4	LinkTrrd4	It links to TRRD Data Bank.

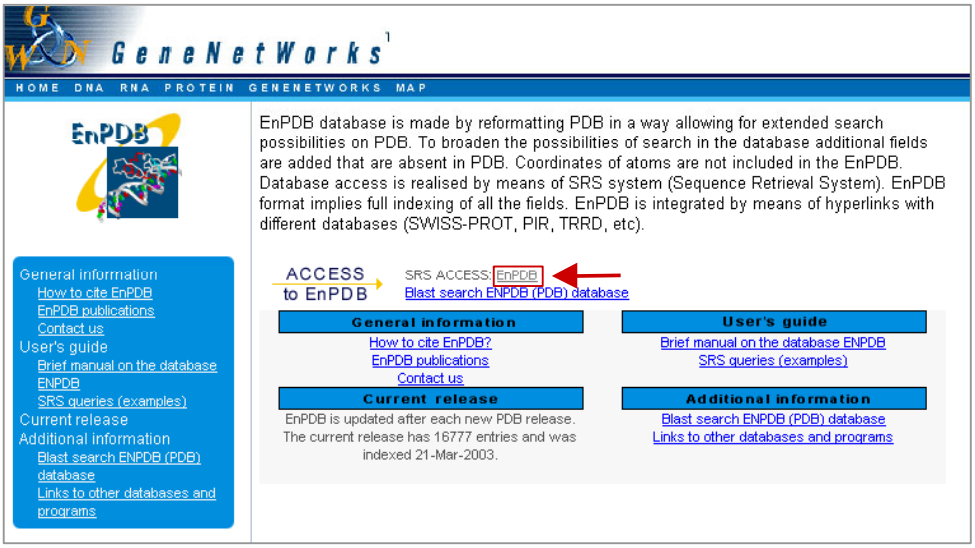
Examples of SRS queries to the ENPDB database

Example 1

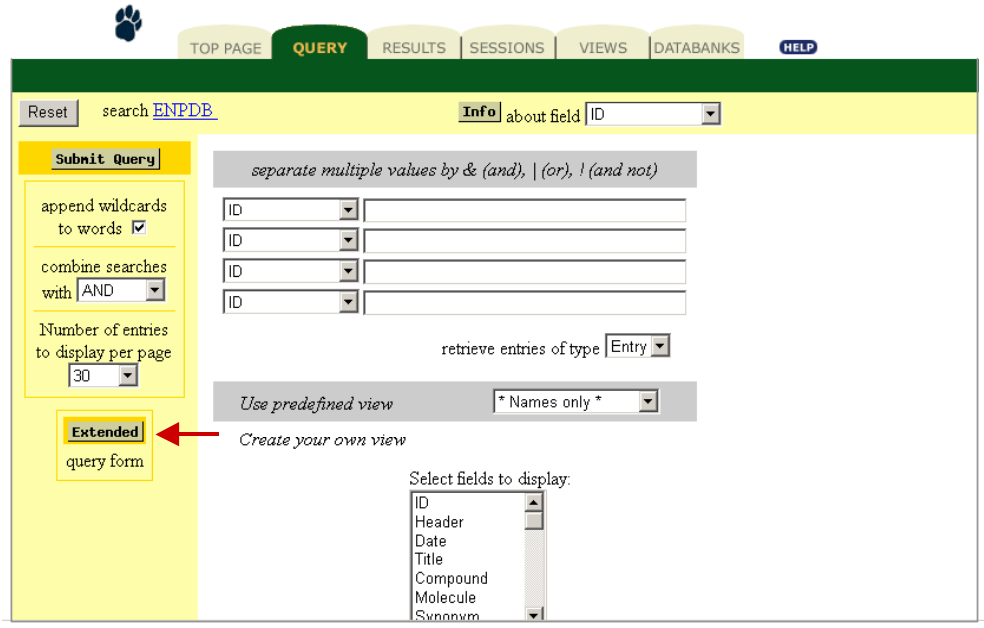
The query is 'To search for receptors of TNF, with the known tertiary structures represented in PDB'.

To make such a query, you should perform the following:
1. Choose the option SRS ACCESS: ENPDB as shown in the Figure by red arrow.

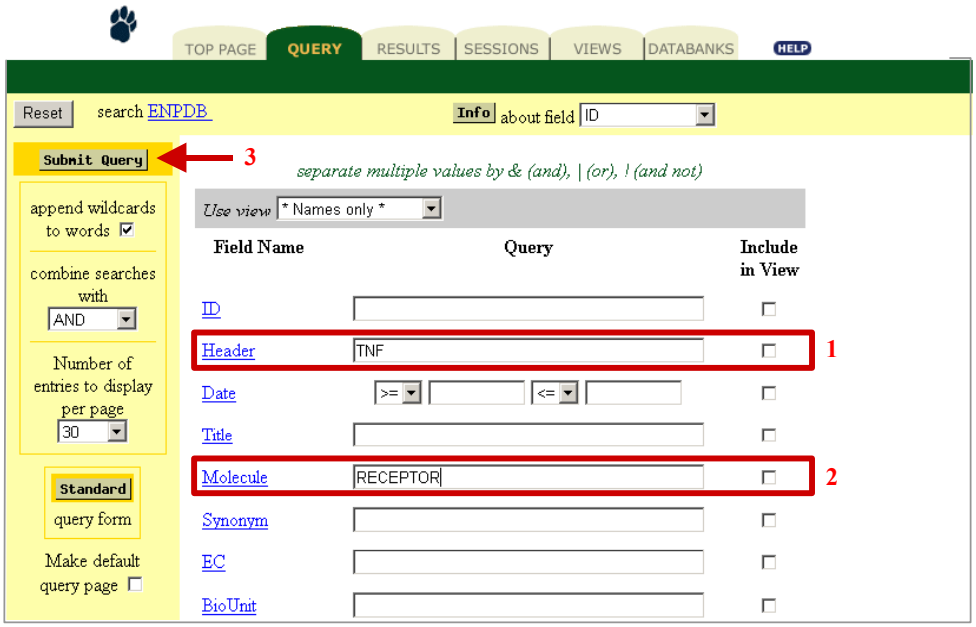




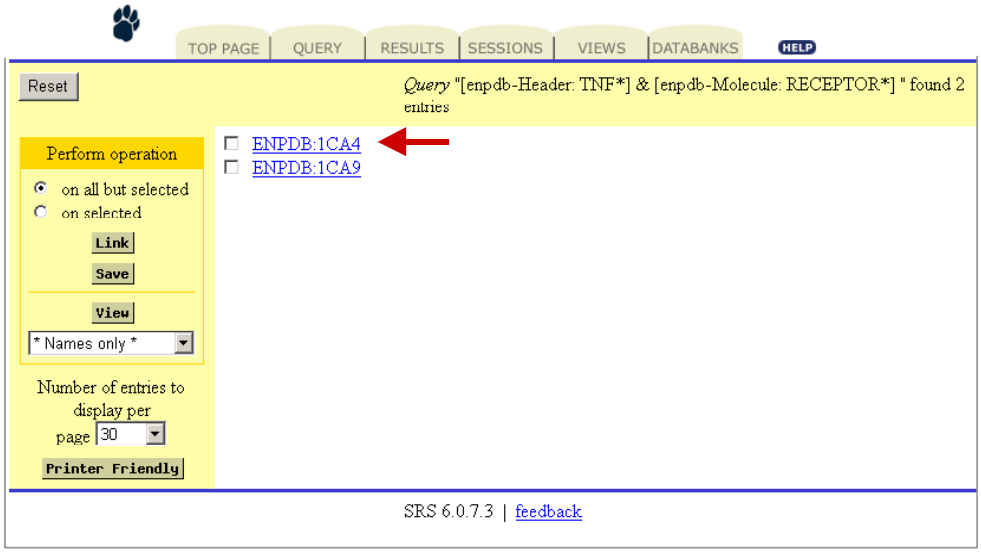
2. To load the extended search form, click the button ‘Extended’.



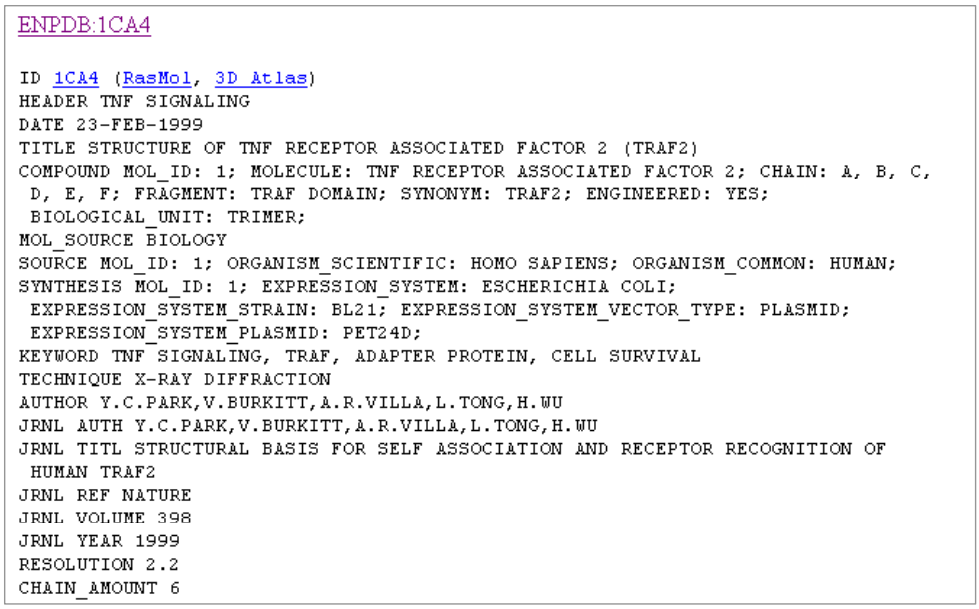
3. In the field ‘Header’ choose ‘TNF’ (1). In the field ‘Molecule’ input ‘RECEPTOR’(2). Click the button 'Submit Query’(3).



4. This will bring up the 'Query Results' page with the entries found in ENPDB. This page is illustrated in the Figure. To load the complete text of the entry, click the hyperlink of the entry.



5. Click the name of the entry. This will bring up the page illustrated in the figure.



ENPDB:1CA4

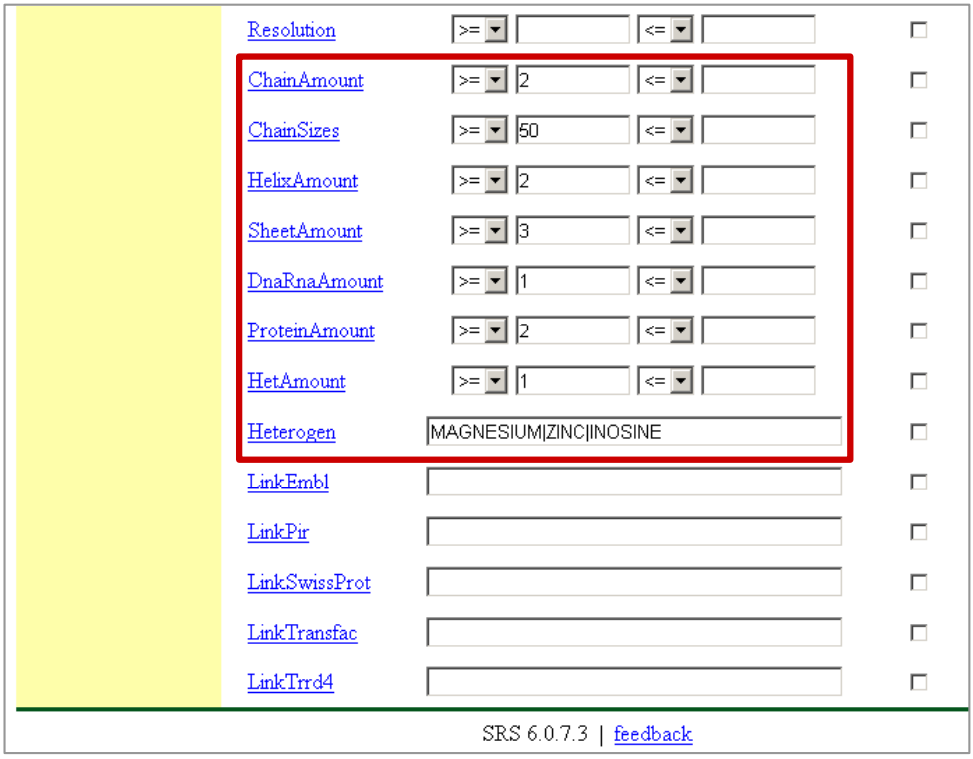
```
ID 1CA4 (RasMol, 3D Atlas)
HEADER TNF SIGNALING
DATE 23-FEB-1999
TITLE STRUCTURE OF TNF RECEPTOR ASSOCIATED FACTOR 2 (TRAF2)
COMPOUND MOL_ID: 1; MOLECULE: TNF RECEPTOR ASSOCIATED FACTOR 2; CHAIN: A, B, C,
 D, E, F; FRAGMENT: TRAF DOMAIN; SYNONYM: TRAF2; ENGINEERED: YES;
 BIOLOGICAL_UNIT: TRIMER;
MOL_SOURCE BIOLOGY
SOURCE MOL_ID: 1; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN;
SYNTHESIS MOL_ID: 1; EXPRESSION_SYSTEM: ESCHERICHIA COLI;
 EXPRESSION_SYSTEM_STRAIN: BL21; EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
 EXPRESSION_SYSTEM_PLASMID: PET24D;
KEYWORD TNF SIGNALING, TRAF, ADAPTER PROTEIN, CELL SURVIVAL
TECHNIQUE X-RAY DIFFRACTION
AUTHOR Y.C.PARK,V.BURKITT,A.R.VILLA,L.TONG,H.WU
JRNL AUTH Y.C.PARK,V.BURKITT,A.R.VILLA,L.TONG,H.WU
JRNL TITL STRUCTURAL BASIS FOR SELF ASSOCIATION AND RECEPTOR RECOGNITION OF
 HUMAN TRAF2
JRNL REF NATURE
JRNL VOLUME 398
JRNL YEAR 1999
RESOLUTION 2.2
CHAIN_AMOUNT 6
```

Example 2.

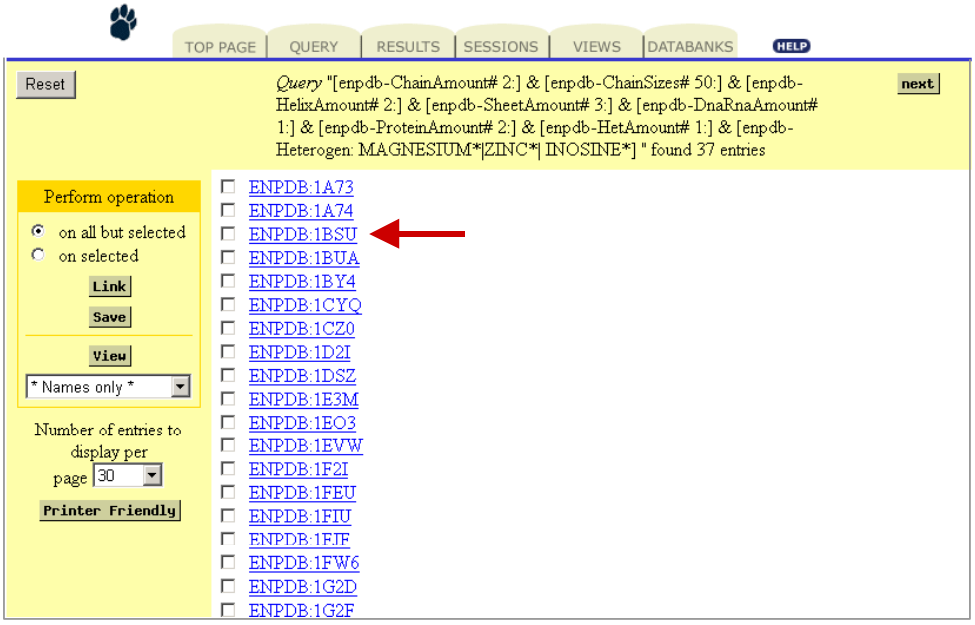
The query is 'To find all protein complexes with the ordered quantitative parameters: the number of DNA/RNA molecules in the complex is at least 1; the number of proteins in the complex is at least 2; number of protein chains is at least 2; the protein chain length is at least 50; the number of alpha-helices in the protein is at least 2; the number of beta-sheets is at least 3; the number of ligands in the complex is at least 1; the ligand name is one of the following MAGNESIUM, ZINC, or INOSINE'

To make such a query, you should perform the following:

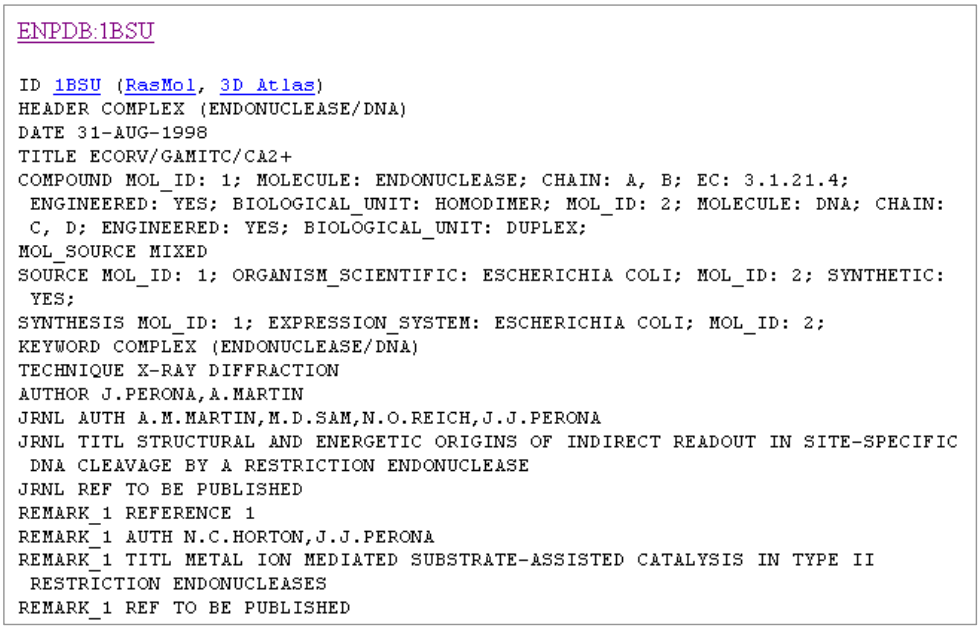
1 To input these parameters, in the field 'ChainAmount' select the sign ">=" and enter the value 2. In the field 'ChainSizes' choose the sign ">=" and input the value 50. In the field 'HelixAmount' choose the sign ">=" and insert the value 2. In the field SheetAmount choose the sign ">=" and enter the value 3. In the field 'DnaRnaAmount' choose the sign ">=" and input the value 1. In the filed 'ProteinAmount' choose the sign ">=" and input the value 2. In the field 'HetAmount' choose the sign ">=" and input the value 1. In the field Heterogen input MAGNESIUM|ZINC|INOSINE. Click the button "Submit Query". This query is illustrated in the Figure.



2. After submitting the query, you will see the 'Query Results' page with the list of entries found in ENPDB database. To load the complete text of the entry, click the hyperlink with the entry name. This will bring up the page illustrated in the next Figure.



3. Click the name of the entry. This will bring up the entry page illustrated in the Figure.



ENPDB:1BSU

```
ID 1BSU (RasMol, 3D Atlas)
HEADER COMPLEX (ENDONUCLEASE/DNA)
DATE 31-AUG-1998
TITLE ECORV/GAMITC/CA2+
COMPOUND MOL_ID: 1; MOLECULE: ENDONUCLEASE; CHAIN: A, B; EC: 3.1.21.4;
 ENGINEERED: YES; BIOLOGICAL_UNIT: HOMODIMER; MOL_ID: 2; MOLECULE: DNA; CHAIN:
 C, D; ENGINEERED: YES; BIOLOGICAL_UNIT: DUPLEX;
MOL_SOURCE MIXED
SOURCE MOL_ID: 1; ORGANISM_SCIENTIFIC: ESCHERICHIA COLI; MOL_ID: 2; SYNTHETIC:
 YES;
SYNTHESIS MOL_ID: 1; EXPRESSION_SYSTEM: ESCHERICHIA COLI; MOL_ID: 2;
KEYWORD COMPLEX (ENDONUCLEASE/DNA)
TECHNIQUE X-RAY DIFFRACTION
AUTHOR J.PERONA,A.MARTIN
JRNL AUTH A.M.MARTIN,M.D.SAM,N.O.REICH,J.J.PERONA
JRNL TITL STRUCTURAL AND ENERGETIC ORIGINS OF INDIRECT READOUT IN SITE-SPECIFIC
 DNA CLEAVAGE BY A RESTRICTION ENDONUCLEASE
JRNL REF TO BE PUBLISHED
REMARK_1 REFERENCE 1
REMARK_1 AUTH N.C.HORTON,J.J.PERONA
REMARK_1 TITL METAL ION MEDIATED SUBSTRATE-ASSISTED CATALYSIS IN TYPE II
 RESTRICTION ENDONUCLEASES
REMARK_1 REF TO BE PUBLISHED
```

4 Click the 3D Atlas link to visualize the entry by 3D Browser.

```
ENPDB:1BSU

ID 1BSU (RasMol, 3D Atlas)
HEADER COMPLEX (ENDONUCLEASE/DNA)
DATE 31-AUG-1998
TITLE ECORV/GAMITC/CA2+
COMPOUND MOL_ID: 1; MOLECULE: ENDONUCLEASE; CHAIN: A, B; EC: 3.1.21.4;
 ENGINEERED: YES; BIOLOGICAL_UNIT: HOMODIMER; MOL_ID: 2; MOLECULE: DNA; CHAIN:
 C, D; ENGINEERED: YES; BIOLOGICAL_UNIT: DUPLEX;
MOL_SOURCE MIXED
SOURCE MOL_ID: 1; ORGANISM_SCIENTIFIC: ESCHERICHIA COLI; MOL_ID: 2; SYNTHETIC:
 YES;
SYNTHESIS MOL_ID: 1; EXPRESSION_SYSTEM: ESCHERICHIA COLI; MOL_ID: 2;
KEYWORD COMPLEX (ENDONUCLEASE/DNA)
TECHNIQUE X-RAY DIFFRACTION
AUTHOR J.PERONA,A.MARTIN
JRNL AUTH A.M.MARTIN,M.D.SAM,N.O.REICH,J.J.PERONA
JRNL TITL STRUCTURAL AND ENERGETIC ORIGINS OF INDIRECT READOUT IN SITE-SPECIFIC
 DNA CLEAVAGE BY A RESTRICTION ENDONUCLEASE
JRNL REF TO BE PUBLISHED
REMARK_1 REFERENCE 1
REMARK_1 AUTH N.C.HORTON,J.J.PERONA
REMARK_1 TITL METAL ION MEDIATED SUBSTRATE-ASSISTED CATALYSIS IN TYPE II
 RESTRICTION ENDONUCLEASES
REMARK_1 REF TO BE PUBLISHED
```

4.

```
ENPDB:1CA4

ID 1CA4 (RasMol, 3D Atlas)
HEADER TNF SIGNALING
DATE 23-FEB-1999
TITLE STRUCTURE OF TNF RECEPTOR ASSOCIATED FACTOR 2 (TRAF2)
COMPOUND MOL_ID: 1; MOLECULE: TNF RECEPTOR ASSOCIATED FACTOR 2; CHAIN: A, B, C,
 D, E, F; FRAGMENT: TRAF DOMAIN; SYNONYM: TRAF2; ENGINEERED: YES;
 BIOLOGICAL_UNIT: TRIMER;
MOL_SOURCE BIOLOGY
SOURCE MOL_ID: 1; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN;
SYNTHESIS MOL_ID: 1; EXPRESSION_SYSTEM: ESCHERICHIA COLI;
 EXPRESSION_SYSTEM_STRAIN: BL21; EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
 EXPRESSION_SYSTEM_PLASMID: PET24D;
KEYWORD TNF SIGNALING, TRAF, ADAPTER PROTEIN, CELL SURVIVAL
TECHNIQUE X-RAY DIFFRACTION
AUTHOR Y.C.PARK,V.BURKITT,A.R.VILLA,L.TONG,H.WU
JRNL AUTH Y.C.PARK,V.BURKITT,A.R.VILLA,L.TONG,H.WU
JRNL TITL STRUCTURAL BASIS FOR SELF ASSOCIATION AND RECEPTOR RECOGNITION OF
 HUMAN TRAF2
JRNL REF NATURE
JRNL VOLUME 398
JRNL YEAR 1999
RESOLUTION 2.2
CHAIN_AMOUNT 6
```

Comments and questions are welcome to Vladimir Ivanisenko (salix@bionet.nsc.ru)

2. PDBSite Database

Release 2003

1. Database description: PDBSite is a database on protein active sites and their spatial environment.

2. Access to PDBSite database:
http://www.domain.com/mgs/gnw/pdbsite/

3. Database content

SRS table	Description	Number of entries
PDBSITE	database on protein active sites and their spatial environment	2842

4. List of biological tasks that could be solved by using the PDBSite database:

- search for proteins according the data on protein function and structural characteristics of protein active sites;
- search for proteins based on the data about sites subjected to biochemical modifications;
- search for functional sites on the basis of data on biological classification of proteins.

5. SRS table format

PDBSITE

PDBSite database contains the fields of two types: those that could be queried and searched or not. The fields that could not be queried contain additional information on the site structure and their surroundings.

Below is given the description of fields that could be queried.

Line code	Field name	Field description
ID	ID	Entry identifier. This identifier is unique within PDBSite.
PDBID	PDBID	PDB ID code. This identifier is unique within PDB.
HEADER	Header	It contains PDB classification for the entry. The field content corresponds to that in PDB.
TITLE	Title	It contains the title for experiment or analysis described in the entry. The field content corresponds to that in PDB.
KEYWORD	Keyword	It contains keywords describing the macromolecule. The content corresponds to that of KEYWDS of PDB.
MOLECULE	Molecule	It contains names of macromolecules from the COMPND of PDB and is designed to search for entries by the names of macromolecules.
NUM_SITE_CHAINS	NumSiteChains	It contains number of different chains, which the residues of the site belong to.
SITE_DESCR	SiteDescr	It contains description of the site. The content corresponds to that of SITE_DESCRIPTION sub-field of REMARK 800 field of PDB.
RESIDUE_NOTAA	ResidueNotAA	It contains the names of residues that are not amino acids but contained in the site.
SITE_CHAINS	LenSite	It contains number of residues in the site.
SURR_CHAINS	LenSurround	It contains number of residues in the site environment.

EXPOSURE	ExposureSite	It contains average exposure of residues of the site.
EXPOSURE	ExposureSurround	It contains average exposure of residues of the site environment.
DISCONTINUITY	Discontinuity	It characterises discontinuity of the site by its primary structure.

Below is the list of fields that could not be queried, but contain additional information on the site structure and its surroundings.

MOLECULE - contains names of macromolecules from the COMPND of PDB and is designed to search for entries by the names of macromolecules.

MOLCHAINS – contains identifier of chains of a macromolecule.

CHAIN_ID - contains chain identifier for the site and its environment.

POS – indicates the positions of site residues and its surroundings in a protein sequence.

RESNAME - contains names of site residues and its environment.

EXPOSE - contains exposure of each of the residues of the site and its environment.

SITE_SEQ_PROFILE – Contains relative frequency of an amino acid type at each position of the site in homologous proteins.

Physico-chemical parameters of the site and its surroundings are listed in a special table. Types of physico-chemical characteristics are given in the columns of this table. The order of physico-chemical characteristics in the columns of the table is indicated in the line ORDER. In the lines, the type of a site and its surroundings is indicated, as well as the way of calculation of the physico-chemical parameter for the site and its surroundings. Three types of physico-chemical parameters are listed in the table: average, sum and module of spatial moment.

In the lines SITE, there are physico-chemical parameters calculated for the site. The lines FULL_SURROUND contain physico-chemical parameters calculated for all residues of site surroundings. The lines EXPOSED_SURROUND contain physico-chemical parameters calculated for exposed residues of the site surroundings. The lines BURIED_SURROUND contain physico-chemical parameters calculated for buried residues of site surroundings.

The table is organized in the following way.

The first is the line ORDER.

Then follows the line AVERAGE that indicates that in lines SITE, FULL_SURROUND, EXPOSED_SURROUND and BURIED_SURROUND the average values of physico-chemical parameters will be indicated. These lines are placed below the line AVERAGE.

Next is the line SUM indicating that physico-chemical parameters listed in the lines below were calculated as the sum values. Then follows the line SPATIAL MOMENT indicating that physico-chemical parameters listed in subsequent lines SITE, FULL_SURROUND, EXPOSED_SURROUND and BURIED_SURROUND were calculated as module of spatial moment. This is the end of the table.

PAIRWISE - contains pairwise distances between residues of the site.

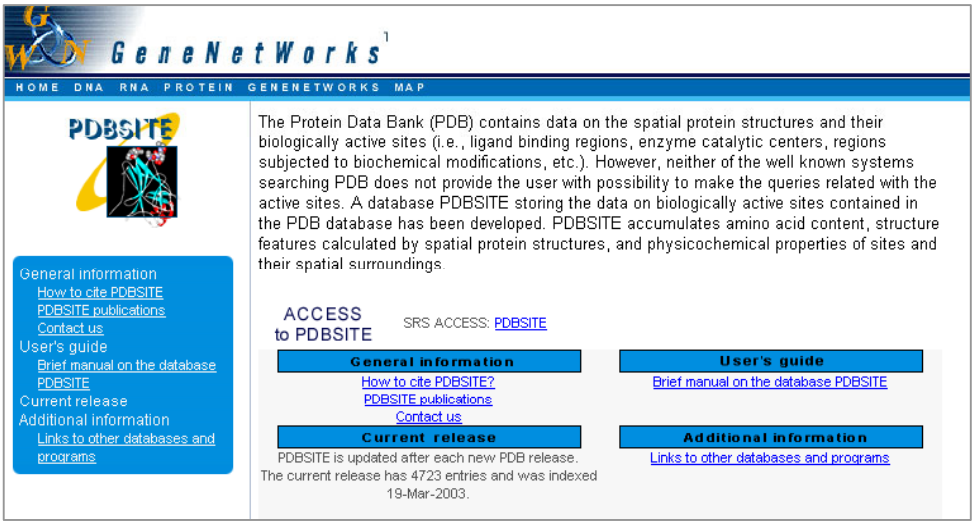
COORDINATES CA_ATOMS - contains C-alpha atom co-ordinates of site residues.

COORDINATES CENTRE_MASS - contains centre mass co-ordinates of site residues.

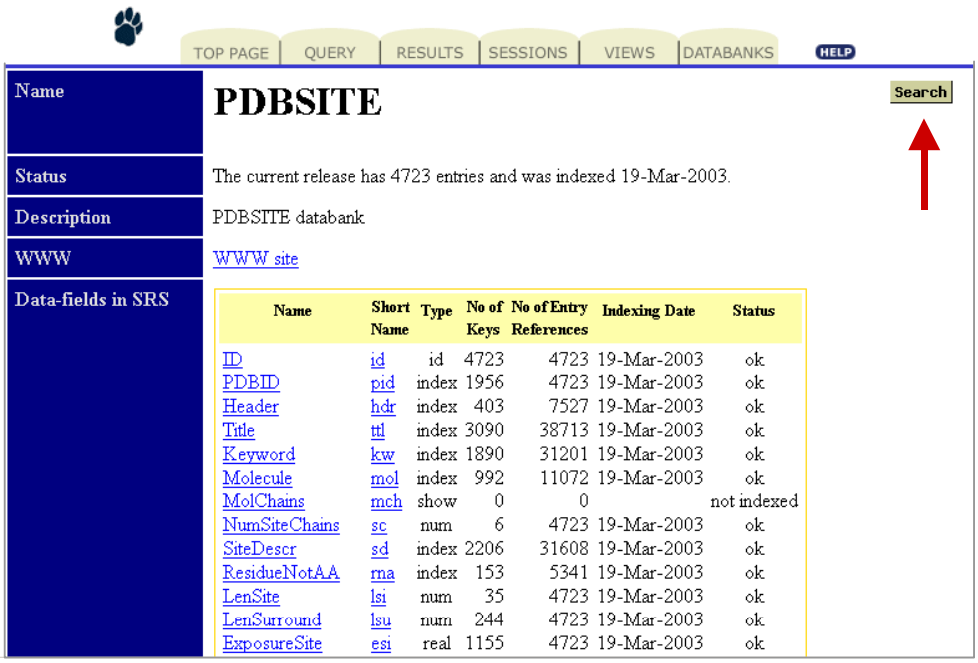
Examples of SRS queries to the PDBSite database

Example 1
The query is
'Find all proteins containing Zinc-binding site**'.**

To make such a query, you should execute the following:
1. Choose PDBSite SRS table on the page 'SRS access'.

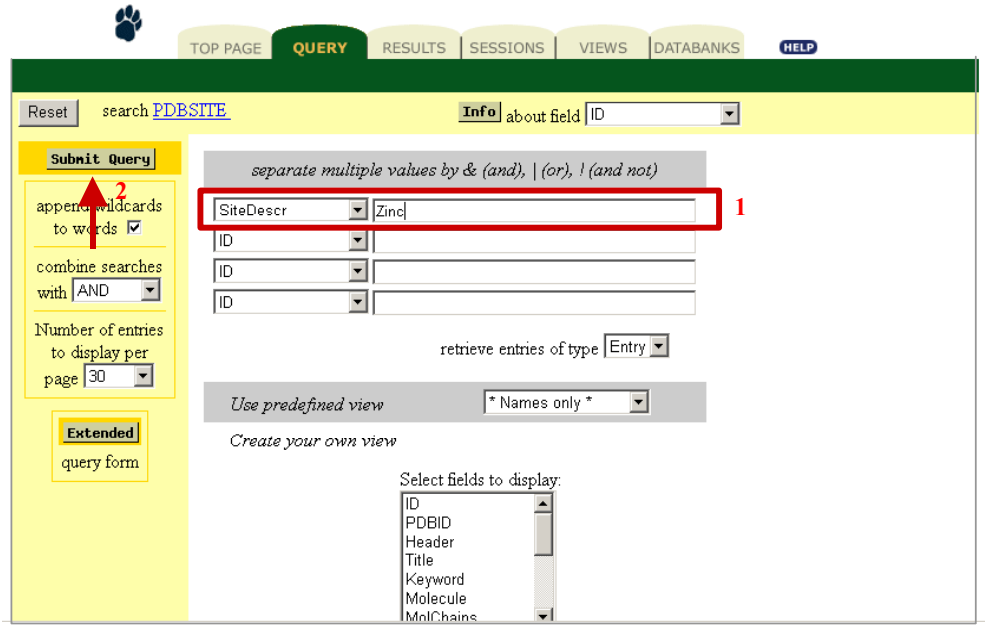


2. Click the button ‘Search’.

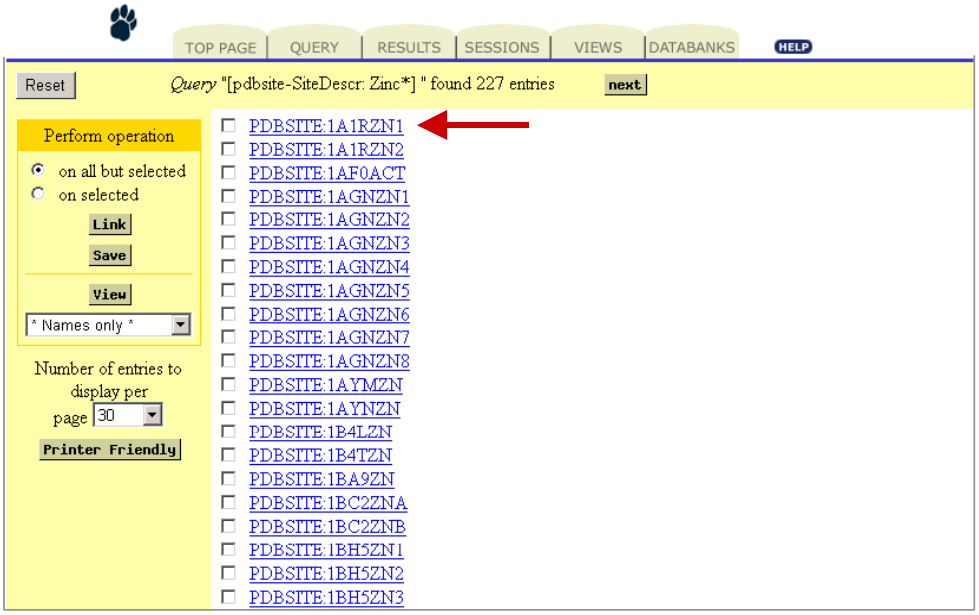


Name	Short Name	Type	No of Keys	No of Entry References	Indexing Date	Status
ID	id	id	4723	4723	19-Mar-2003	ok
PDBID	pid	index	1956	4723	19-Mar-2003	ok
Header	hdr	index	403	7527	19-Mar-2003	ok
Title	ttl	index	3090	38713	19-Mar-2003	ok
Keyword	kw	index	1890	31201	19-Mar-2003	ok
Molecule	mol	index	992	11072	19-Mar-2003	ok
MolChains	mch	show	0	0		not indexed
NumSiteChains	sc	num	6	4723	19-Mar-2003	ok
SiteDescr	sd	index	2206	31608	19-Mar-2003	ok
ResidueNotAA	rna	index	153	5341	19-Mar-2003	ok
LenSite	lsi	num	35	4723	19-Mar-2003	ok
LenSurround	lsu	num	244	4723	19-Mar-2003	ok
ExposureSite	esi	real	1155	4723	19-Mar-2003	ok

3. Click the field 'SiteDescr' from the drop-down menu. For querying Zinc-binding sites insert 'Zinc' in the text-box located to the right (1). Click the button 'Submit Query'(2).



4. The results of the querying operation will be a list of PDBSITE entries. To display the complete text of an entry, click the hyperlink with the entry name.



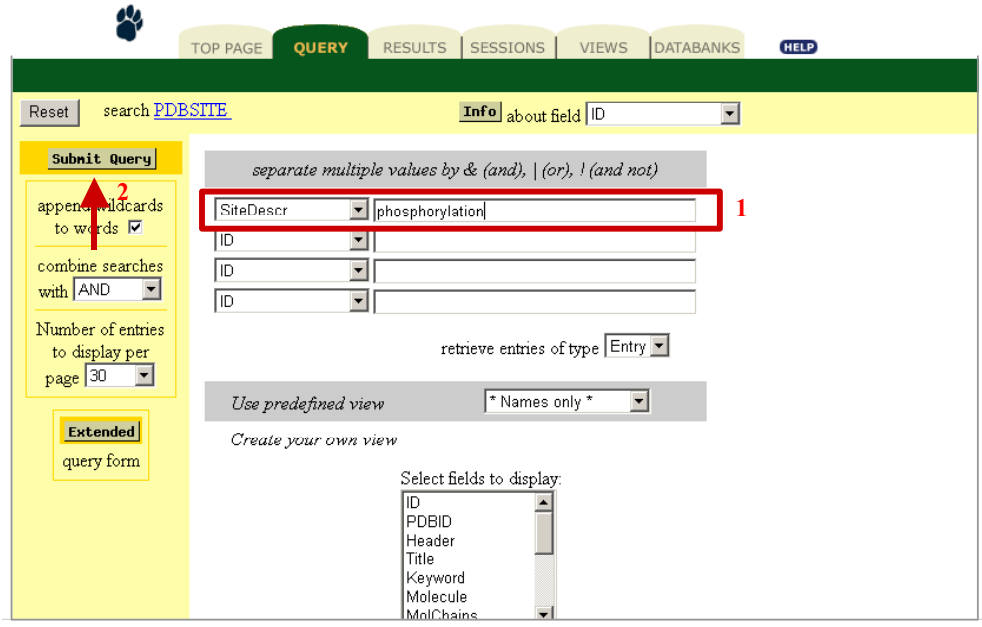
5. Complete text of the entry found

PDBSITE:1A1RZN1

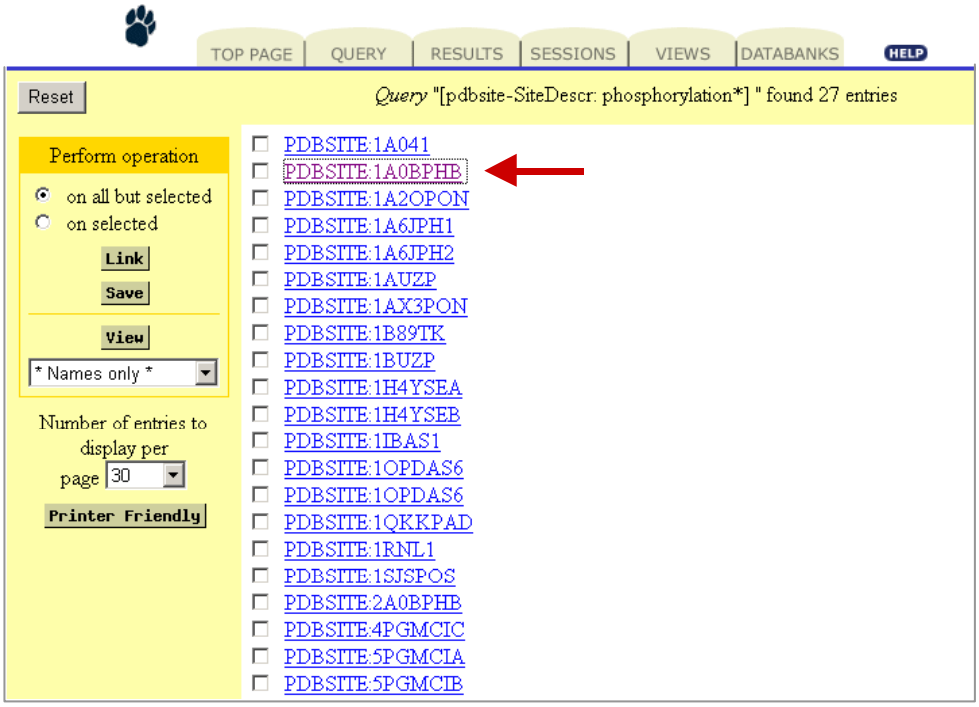
```
ID 1A1RZN1
PDBID 1A1R
HEADER VIRAL PROTEIN
TITLE HCV NS3 PROTEASE DOMAIN:NS4A PEPTIDE COMPLEX
KEYWORD VIRAL PROTEIN, SERINE PROTEASE, NONSTRUCTURAL PROTEINS, COFACTOR PEPTIDE, HEL
MOLECULE NS3 PROTEIN
MOL_CHAINS AB
NUM_SITE_CHAINS 1
SITE_DESCR ZINC BINDING SITE, MOLECULE A.
RESIDUE_NOTAA ZN
NUMBER_OF_AA 3 22
EXPOSURE 13.667 14.864
DISCONTINUITY 15.333
SITE_CHAINS AAA
POS 123 125 171
RESNAME CCC
EXPOSE 9 32 0
SITE_SEQ_PROFILE
  PDBNo   V   L   I   M   F   W   Y   G   A   P   S   T   C   H   R   K   Q   E   N
  123 A   0   0   0   0   0   0   0   0   0   0   0   0  99   0   0   0   0   0   0
  125 A   0   0   0   0   0   0   0   0   0   0   0   0  99   0   0   0   0   0   0
  171 A   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0
```

Example 2
The query is
'Find all proteins containing the sites subjected to phosphorilation**'.**
To make such a query, you should perform the following:

1. Select from the drop-down menu the field SiteDescr. For searching for phosphorylation sites enter in the text-box to the right ‘phosphorylation’(1). Click ‘Submit Query’ button (2).



2. This brings up a list of PDBSite entries displayed in the 'Query Results' page. To display the complete text of an entry, click the hyperlink with the entry name.



3. The complete text of the 'phosphorilation site' entry is shown in Figure.

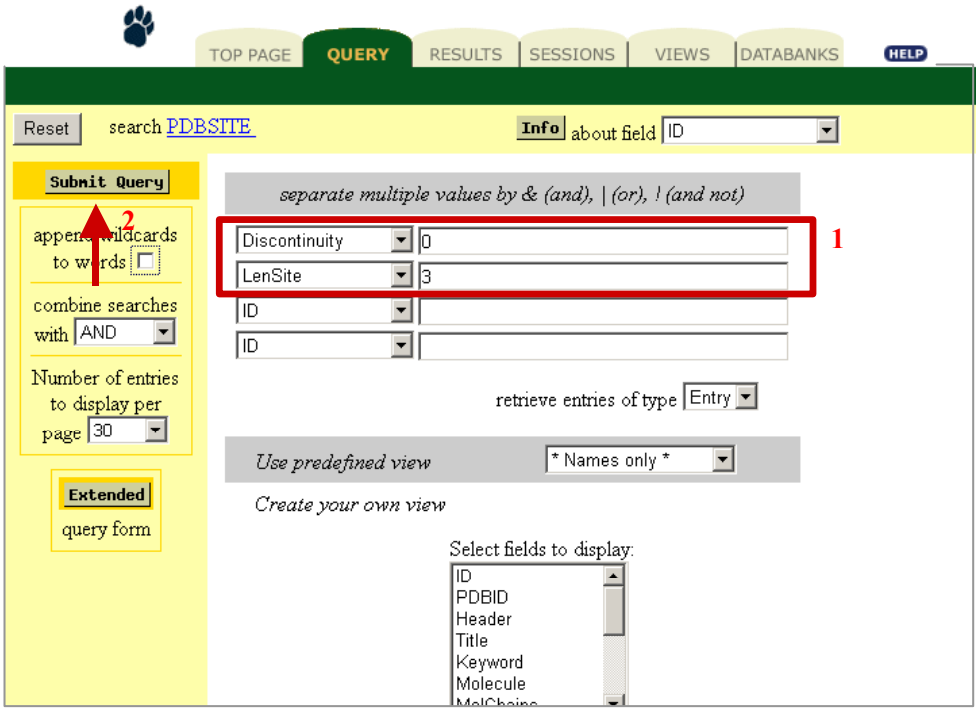
PDBSITE:1A0BPHB

```
ID 1A0BPHB
PDBID 1A0B
HEADER HISTIDINE KINASE
TITLE HISTIDINE-CONTAINING PHOSPHOTRANSFER DOMAIN OF ARCB FROM ESCHERICHIA COLI
KEYWORD HISTIDINE KINASE, PHOSPHOTRANSFER, TWO-COMPONENT SYSTEM, FOUR-HELIX BUNDLE
MOLECULE AEROBIC RESPIRATION CONTROL SENSOR PROTEIN ARCB
MOL_CHAINS
NUM_SITE_CHAINS 1
SITE_DESCR PHOSPHORYLATION SITE.
NUMBER_OF_AA 1 8
EXPOSURE 35.000 17.250
DISCONTINUITY 0.000
SITE_CHAINS
POS 715
RESNAME H
EXPOSE 35
SITE_SEQ_PROFILE
  PDBNo   V   L   I   M   F   W   Y   G   A   P   S   T   C   H   R   K   Q   E   N
  715     0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0
SURR_CHAINS
POS 711 713 714 716 717 718 719 737
RESNAME VEGKIKGQ
```

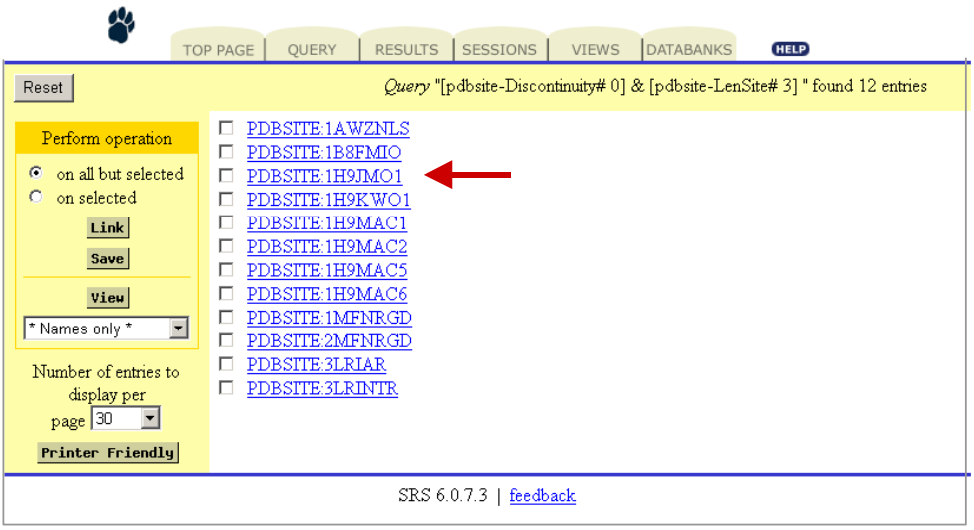
Example 3
The query is
'Find all proteins containing discontinuous sites, with the length of 3 amino acid residues**'.**

To make such a query, you should perform the following:

1. Select from the drop-down menu the field 'Discontinuity'. In the text-box located to the right enter '0'. In menu below select the field 'LenSite'. In the field to the right enter the value '3'. Click the button 'Submit Query'.



2. The results of the querying operation will be a list of PDBSITE entries. To display the complete text of an entry, click the hyperlink with the entry name.



3. Complete text of the entry found is seen in Figure.

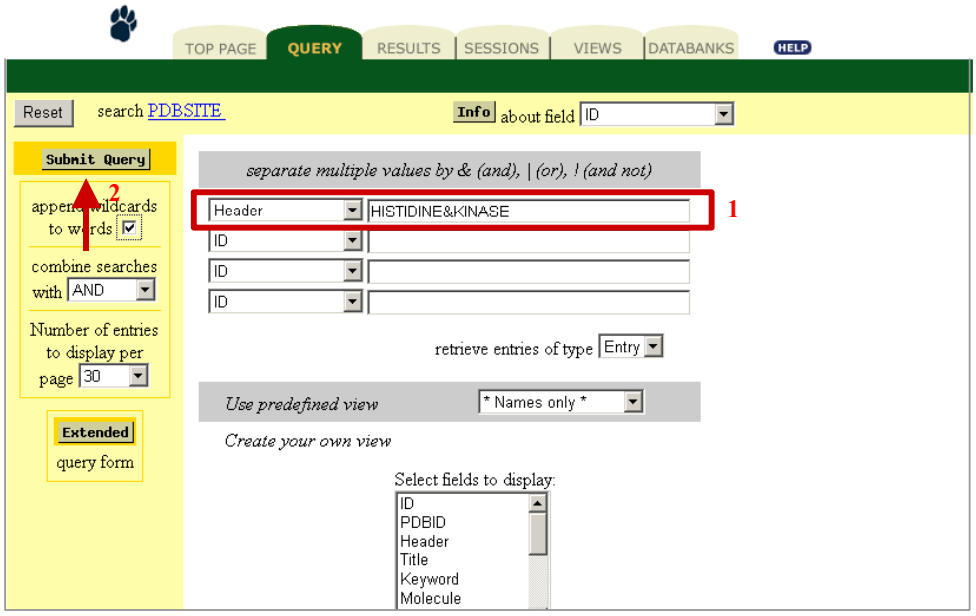
PDBSITE:1MFNRGD

```
ID 1MFNRGD
PDBID 1MFN
HEADER CELL ADHESION PROTEIN
TITLE SOLUTION NMR STRUCTURE OF LINKED CELL ATTACHMENT MODULES OF MOUSE FIBRONECTIN C
KEYWORD CELL ADHESION PROTEIN, RGD, EXTRACELLULAR MATRIX, HEPARIN-BINDING, GLYCOPROTE
MOLECULE FIBRONECTIN
MOL_CHAINS
NUM_SITE_CHAINS 1
SITE_DESCR CELL ADHESION SITE.
NUMBER_OF_AA 3 10
EXPOSURE 46.000 40.900
DISCONTINUITY 0.000
SITE_CHAINS
POS 168 169 170
RESNAME RGD
EXPOSE 28 45 65
SITE_SEQ_PROFILE
  PDBNo   V   L   I   M   F   W   Y   G   A   P   S   T   C   H   R   K   Q   E   N
  168     3   0   0   0   0   0   0   3   8   5  11   3   0   3  50   8   5   0   3
  169     0   0   0   0   0   0   0  53   3   0   0   0   0   5   5   0  34   0   0
  170     3   3   0   0   0   0   0   0   5   5   0   0   0   0   5  32  11   5   0
SURR_CHAINS
```

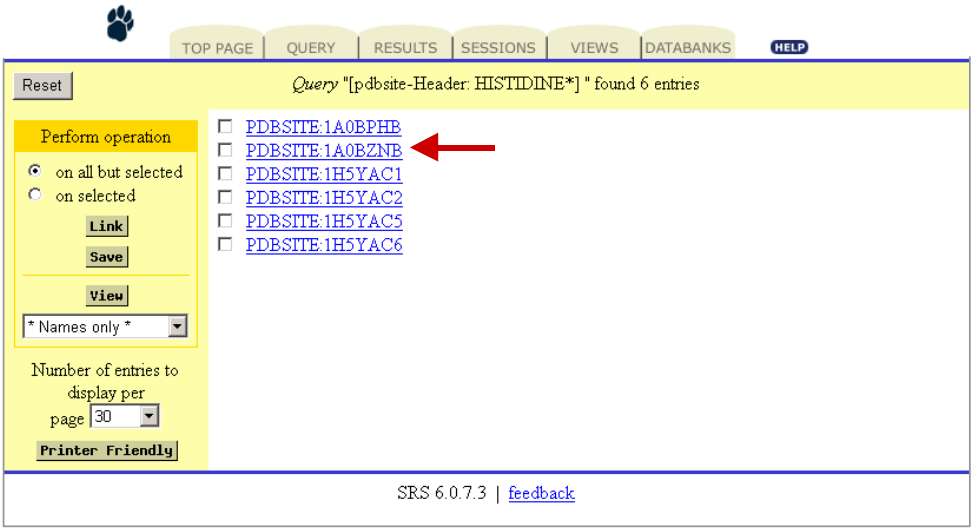
Example 4
The query is
'Find all sites for the protein family of histidine kinases**'.**

To make such a query, you should perform the following:

1. Select the option from the drop-down menu ‘Header’. In the text-box enter ‘HISTIDINE&KINASE’(1). Click the button ‘Submit Query’(2).



2. Submitting the query will bring up the list of PDBSITES entries found. To display the complete text of an entry, click the hyperlink with the entry name.



3. Complete text of the entry found is shown in Figure.

PDBSITE:1A0BZNB

```
ID 1A0BZNB
PDBID 1A0B
HEADER HISTIDINE KINASE
TITLE HISTIDINE-CONTAINING PHOSPHOTRANSFER DOMAIN OF ARCB FROM ESCHERICHIA COLI
KEYWORD HISTIDINE KINASE, PHOSPHOTRANSFER, TWO-COMPONENT SYSTEM, FOUR-HELIX BUNDLE
MOLECULE AEROBIC RESPIRATION CONTROL SENSOR PROTEIN ARCB
MOL_CHAINS
NUM_SITE_CHAINS 1
SITE_DESCR ZN BINDING SITE.
NUMBER_OF_AA 4 39
EXPOSURE 37.500 19.897
DISCONTINUITY 6.750
SITE_CHAINS
POS 728 746 754 758
RESNAME HDEE
EXPOSE 11 75 33 31
SITE_SEQ_PROFILE
  PDBNo   V   L   I   M   F   W   Y   G   A   P   S   T   C   H   R   K   Q   E   N
  728     0   0   0   0   0   0   0   0   0   0   0   0   0  75  25   0   0   0   0
  746     0   0   0   0  13   0   0   0   0   0   0   0   0   0   0   0   0  25   0
  754     0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  13  75   0
  758     0  13   0   0   0   0  13   0   0   0   0   0   0   0   0   0   0  75   0
```

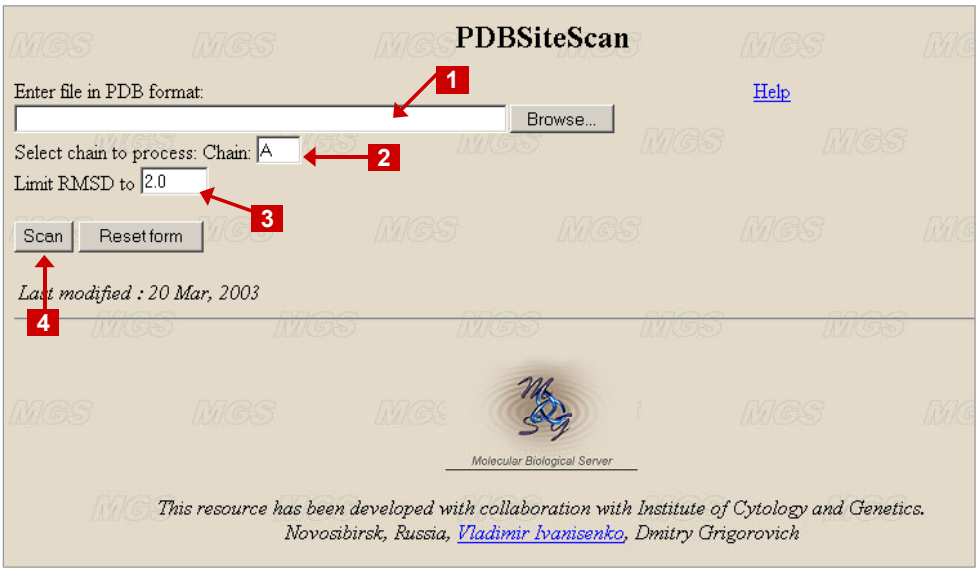
Comments and questions are welcome to Vladimir Ivanisenko (salix@bionet.nsc.ru)

3. PDBSiteScan

Release 2003

Program description:

The program PDBSiteScan automatically performs the best superposition of sites from PDBSite with the 3D structure of a protein under study.
The algorithm developed is based on exhaustion of all the possible combinations of protein positions to be compared with the site. The program realizes the following steps. At the first step, the amino acids of a fragment and the site are compared. If they are identical, their 3D structures are compared at the second step. If the RSMD obtained is lower that the user-specified value, the fragment examined, this fragment is added to the list of results.



Access to PDBSiteScan:
http://www.domain.com/mgs/gnw/pdbsitescan/ link ' Search for Functional Sites in Protein Tertiary Structures '

List of biological tasks that could be solved by using the PDBSiteScan

- detection of potential active sites in the 3D structure;
- structural alignment with known functional sites stored in PDBSite.

Data input

Input a filename of tertiary structure of a protein under study into the text window (1). The tertiary structure should be in PDB format.

Program options

Input a chain ID of the protein to be analyzed (2).
The program selects all sites for which the root mean square deviation (RMSD) for atoms *N*, *Cα*, *C* does not exceed a level specified by the user in the text window (3).

Program execution

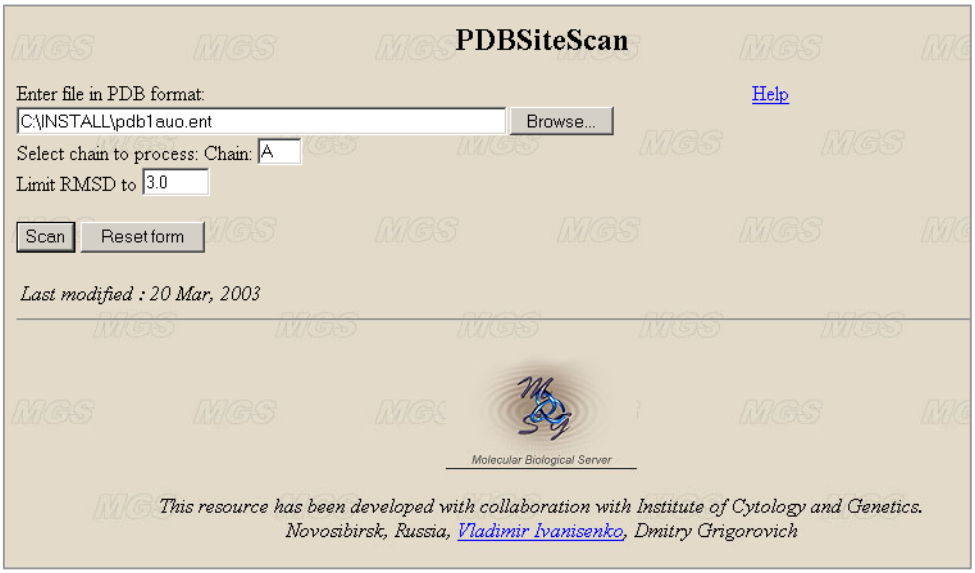
Start the tools processing by clicking the button 'Scan' (4).

Program output

The program output includes links to the PDBSite database, brief description of the site, RMSD value, positions of the protein fragment, and the list of residues.

Example

Let us consider the application of the program by the example of recognition of catalytic site in the protein 1ELV (the family HYDROLASE).
This figure demonstrates the input data for PDBSiteScan.



This figure demonstrates the result of PDBSiteScan operation.

```
PDBSiteID 1AUOACA
PDBID 1AUO
SITE_DESCR CATALYTIC TRIAD.
RMSD   0.000
Chain for each residue: AAA
Positions: 114 168 199
Residues: S D H
END

PDBSiteID 1AUOACB
PDBID 1AUO
SITE_DESCR CATALYTIC TRIAD.
RMSD   0.265
Chain for each residue: AAA
Positions: 114 168 199
Residues: S D H
END

PDBSiteID 1AURACB
PDBID 1AUR
SITE_DESCR CATALYTIC TRIAD.
RMSD   0.465
Chain for each residue: AAA
Positions: 114 168 199
Residues: S D H
END
```

Comments and questions are welcome to Vladimir Ivanisenko (salix@bionet.nsc.ru)

4. Artificial Selected Peptides/Proteins Database (ASPD)

Release 2003

Database description:

ASPD is a curated database on selected from randomized pools proteins and peptides designed for accumulation of experimental data on protein functionality obtained by in vitro directed evolution methods (phage display, ribosome display, SIP etc.)

Access to ASPD database:

http://www.domain.com/mgs/gnw/aspd/

Database content

SRS table	Description	Number of entries
ASPD_ALIGN	Experiments	195
ASPD_REF	Literature	112

Biological tasks that could be solved by using the ASPD database

- Find sequences of proteins and peptides selected by phage display by querying the database with biological keywords
- Annotate proteins by finding its homologs in ASPD

SRS table format

ASPD_ALIGN

Field name	Field description
Identifier	Unique identifier of an entry. It has the format PH1XXNNN, where X is a letter of latin alphabet, N is a number (0-9)
Lit_reference	Link to the literary reference in the ASPD_REF database (the identifier of the corresponding entry in ASPD_REF having the form PH4XXNNN)
Target	the substrate for binding to which the peptides or proteins were selected
Template	the native protein or peptide performing the function for which peptides from random pool were selected
Keywords	A list of keywords (separated with commas) describing the biological relevance of the entry.
Link	Links to external databases. Has the form Link DbName EntryId, where the DbName is one of the following: SwissProt, Prosite, PDB, Enzyme, and EntryId is the identifier of the corresponding entry in these databases.
Comment	The referent's comments concerning conditions of the experiment (free text)
Consensus	The general form of peptides retrieved by in vitro evolution (in PROSITE signature), usally provided by the authors of the original paper.
Number of sequences	Number of different amino acid sequences in the entry
Alignment	Author's alignment of amino acid sequences (those retrieved from random pools are denoted with numbers, others (native or constructed) with letters.

ASPD_REF

Field name	Field description
Identifier	Unique identifier of an entry. It has the format PH4XXNNN, where X is a letter of latin alphabet, N is a number (0-9)
Authors	Authors of the paper
Title	Title of the paper
Journal	Journal name
Volume	Volume
Year	Year
Pages	Page number(s)
Medline	PubMed ID
Corresponding_Author	Full name of the corresponding author with e-mail address

Comments and questions are welcome to Vadim Valuev (valuev@bionet.nsc.ru).

5. DCS - Database on residue coordination spheres

Release 2003

1. Database Description

DCS is database on structural and physico-chemical information of amino acid residue coordination spheres in protein 3D structures. Current release contain data from DNA-binding proteins structures (46 non-homologous protein chains). We concentrate on this specific subset of proteins, because they are involved in gene regulation function.

2. Access to DCS database:

http://www.domain.com/mgs/gnw/dcs/

3. Database content:

DCS database contains of three tables, related to environment class data (DCS_CLUST), protein data (DCS_PROT) and residues data (DCS_RES). Brief description of tables is presented in Table 1.

Table 1. Description of data tables in DCS database.

SRS table	Description	Number of entries
DCS_CLUST	List of coordination sphere clusters with similar physico-chemical properties	9
DCS_PROT	List of proteins used for database derivation	46
DCS_RES	List of coordination spheres and their characteristics for residues in protein dataset	4240

4. Biological tasks that could be solved using DCS database

Database aimed to find residues that have similar physico-chemical environment in protein structures and annotate protein residues in terms of their environment classes.

5. How DCS database was constructed?

The construction of DCS database involved the following stages: (1) determining local spatial environment of residues; (2) calculating physicochemical characteristics of the local environment of individual residues; (3) clustering the local environment with respect to their physicochemical properties; (4) analyzing the resulting classes of physicochemical characteristics of the environment and their interrelations with the structural properties of residues; and (5) analyzing the similarity of spatial environments for various amino acid types.

Data. Spatial structures of the transcription factors displaying a degree of sequence similarity not exceeding 40% were used for the analysis. These structures included (chain names are indicated in parenthesis): 1AIS(B), 1BH9(A,B), 1BM8, 1BOR, 1BVO(A), 1C7U(A),1CF7(A,B), 1CI6(A), 1CQT(I), 1D8J(A), 1DH3(A), 1DL6(A), 1DP7(P), 1ENW(A), 1EO0(A), 1EQF(A), 1EXE(A), 1F3U(A,B), 1F4S(P), 1F62(A), 1G2Y(A), 1GD2(E), 1HKS, 1HLO(A), 1I27(A), 1I4W(A), 1JFI(A,B), 1K99(A), 1MNM(A,C), 1NCS, 1PUE(E), 1QQH(A), 1SKN(P), 1SP1, 1TBA(A,B), 1TF3(A), 1TFI, 1YTF(B,C), and 3HSF. The total number *N* of the residues analyzed amounted to 4240.

Structural properties of residues. We used two parameters, calculated by the program DSSP (Kabsch and Sander, 1983), as structural properties, namely, secondary structure of residues and the surface area accessible to water. These characteristics were calculated for each protein chain

separately from others. Additionally, we determined the residues being in contact with DNA that have at least one atom at a distance less than 4 Å to any atoms of DNA.

Determination of local spatial environment of residues. It was assumed that the local spatial environment of the *i*th (central) residue is formed by the residues whose Cα atoms were located at a distance not exceeding 7 Å from the Cα atom of the *i*th residue. No exclusions were made for the immediate neighbor residues. We considered that the neighbors immediate in the primary structure also contribute to the local environment of a residue in question.

Physicochemical characteristics of the local environment of individual residues. Values of five amino acid properties, taken from (Bogardt *et al*., 1980)—volume, polarity, isoelectric point, hydrophobicity, and the surface area accessible to water—were used while describing the physicochemical characteristics. The spatial environment of the *i*th residue was characterized with the vector $f_i = \{f_k, k = 1,\ldots 5\}$; each component of the vector corresponded to the value of *k*th amino acid property averaged over the resides forming the local spatial environment.

Classification of physicochemical properties of local environment. To classify the local spatial environment of individual residues according to their physicochemical characteristics, we used the hierarchical cluster analysis (Sneath and Sokal, 1973). The squared Euclidian distance $d_{ij} = [\sum_{k=1}^{5} (f_{ik} - f_{jk})^2]$ was used as a measure of the distance between characteristics of the environment for a pair of residues *i, j*. Unweighted paired grouping method with arithmetic mean (UPGMA) was used for constructing the similarity tree.

6. SRS table format

DCS

Field name	Field description
DCS_CLUST	
ID	Environment class identifier
SZ	Size of class (number of environments of this class)
AM	Average solvent accessibility for cluster members
FM	Average physico-chemical properties for environment residues of cluster members
SC	Secondary structure class counts for cluster residues
SF	Secondary structure class fraction for cluster
CM	Residues - members of this class
DCS_PROT	
ID	Protein identifier
DB	Reference to protein structure in PDB (format is ID:Chain)
SZ	Size of protein (number of residues of this protein)
PM	Residues - members of this protein
DCS_RES	
ID	Residue identifier
PR	Reference to protein
CR	Reference to environment class
NM	Residue name
AA	Amino Acid type of residue
PI	Residue index in protein structure
SI	Residue index in protein sequence
SS	Secondary structure according to DSSP program
AC	Solvent Accessibility according to DSSP
DC	Residue is in contact with DNA if DC=1, not in contact if DC=0, if DC=-1 information not available N/A.
CN	Coordination number (number of contact residues)
AF	Frequencies of different amino acid types in coordination sphere (list in order "ARNDCQEGHILKMFPSTWYV")
FX	Average physico-chemical properties for environment residues

Comments and questions to Dmitry Afonnikov (ada@bionet.nsc.ru).

7. References

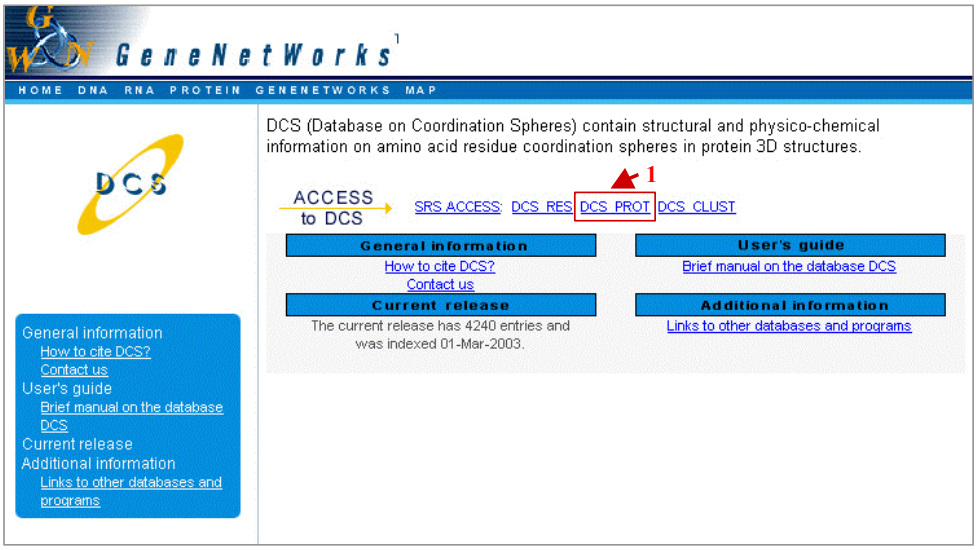
1. Bogardt, R.A., Jr, Jones, B.N., Dwulet, F.E., Garner, W.H., Lehman, L.D., and Gurd, F.R. (1980). Evolution of the amino acid substitution in the mammalian myoglobin gene. *J. Mol. Evol.,* **15**:197-218.
2. Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers,* **22**:2577-2637.
3. Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* San Francisco: W.H. Freeman and Co.

Examples of SRS queries to the DCS database

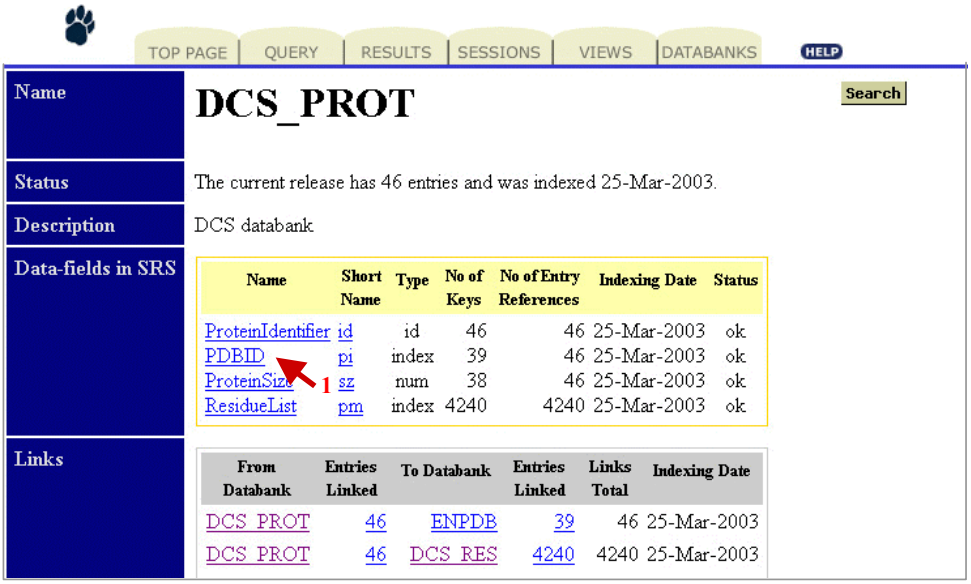
Example 1

The query is 'To search for TBP protein (PDB identifier 1AIS), represented in DSC'.

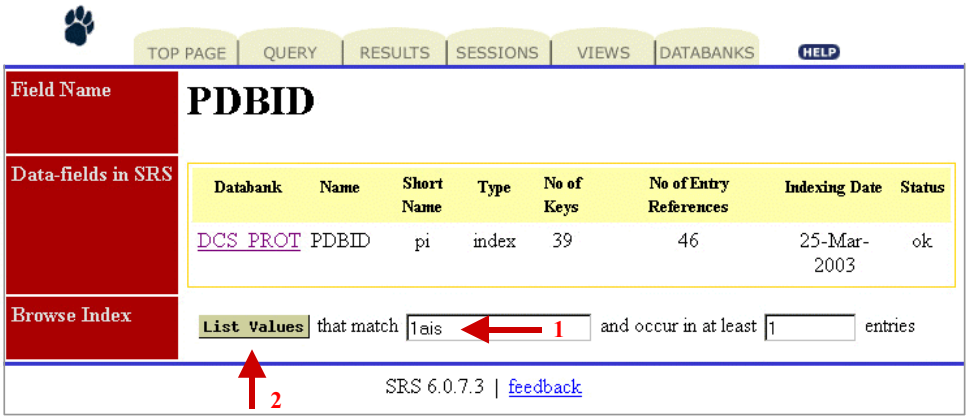
To make such a query, you should perform the following:
1. From DCS start page follow the link 'DCS_PROT' database (1).



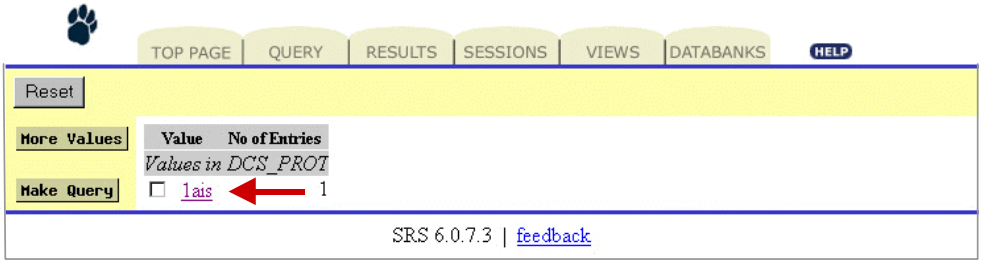
2. This will bring up the 'DCS_PROT' page. This page is illustrated in the Figure. To perform database query, click 'PDBID' hyperlink (1).



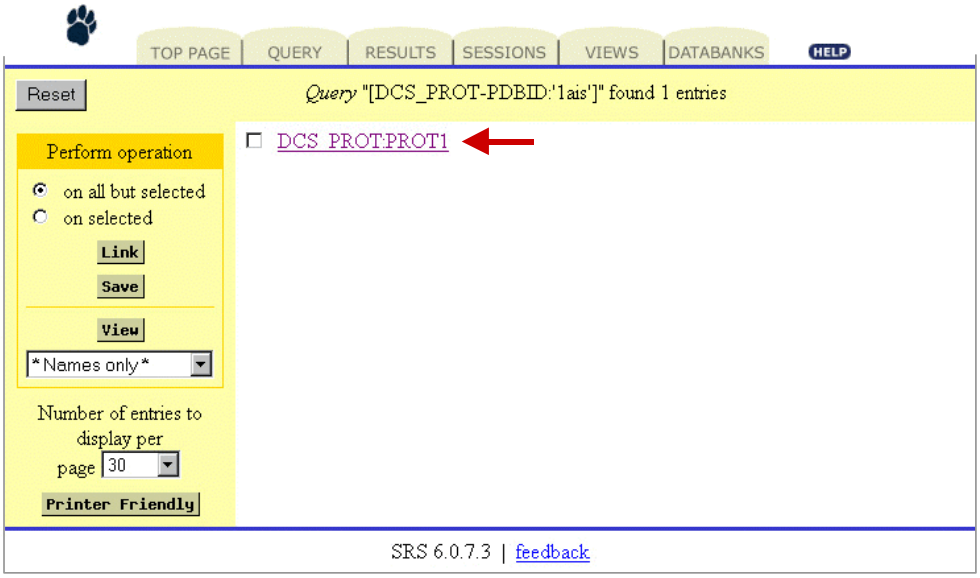
3.This will bring up the query page for 'PDBID' field of the DCS_PROT table. Input '1ais' string into query field (1). Click the 'List Values' button (2).



4. This will bring up the 'Query Results' page with the entries found in DCS_PROT. This page is illustrated in the Figure. Click the hyperlink of the entry.



5. This will bring up the page illustrated in the figure. Click to the hyperlink to PROT1 entry.



6. This will bring up the resulting page of the PROT1 entry.

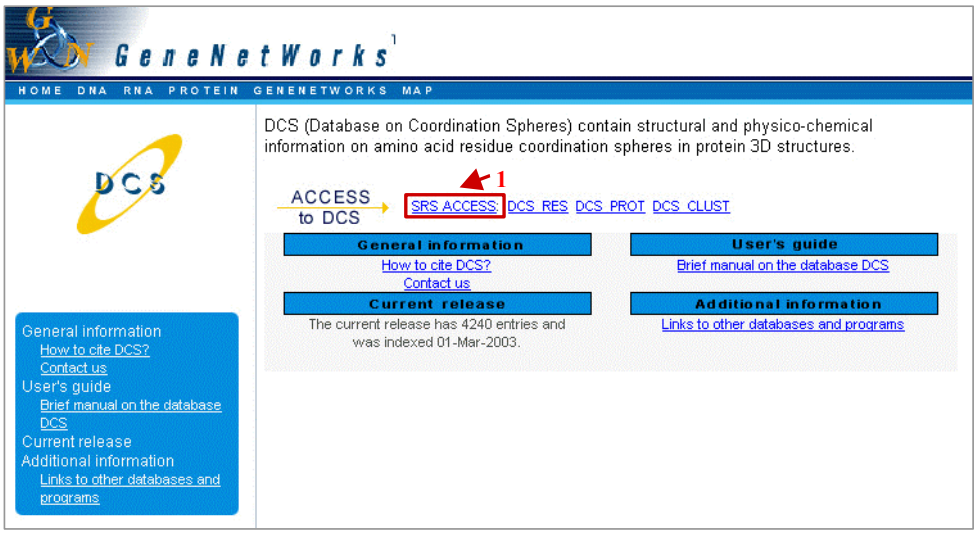
DCS_PROT:PROT1

```
ID PROT1
DB 1AIS:B
SZ 193
PM
 RES1 RES2 RES3 RES4 RES5 RES6 RES7 RES8 RES9 RES10
 RES11 RES12 RES13 RES14 RES15 RES16 RES17 RES18 RES19 RES20
 RES21 RES22 RES23 RES24 RES25 RES26 RES27 RES28 RES29 RES30
 RES31 RES32 RES33 RES34 RES35 RES36 RES37 RES38 RES39 RES40
 RES41 RES42 RES43 RES44 RES45 RES46 RES47 RES48 RES49 RES50
 RES51 RES52 RES53 RES54 RES55 RES56 RES57 RES58 RES59 RES60
 RES61 RES62 RES63 RES64 RES65 RES66 RES67 RES68 RES69 RES70
 RES71 RES72 RES73 RES74 RES75 RES76 RES77 RES78 RES79 RES80
 RES81 RES82 RES83 RES84 RES85 RES86 RES87 RES88 RES89 RES90
 RES91 RES92 RES93 RES94 RES95 RES96 RES97 RES98 RES99 RES100
 RES101 RES102 RES103 RES104 RES105 RES106 RES107 RES108 RES109 RES110
 RES111 RES112 RES113 RES114 RES115 RES116 RES117 RES118 RES119 RES120
 RES121 RES122 RES123 RES124 RES125 RES126 RES127 RES128 RES129 RES130
 RES131 RES132 RES133 RES134 RES135 RES136 RES137 RES138 RES139 RES140
 RES141 RES142 RES143 RES144 RES145 RES146 RES147 RES148 RES149 RES150
 RES151 RES152 RES153 RES154 RES155 RES156 RES157 RES158 RES159 RES160
 RES161 RES162 RES163 RES164 RES165 RES166 RES167 RES168 RES169 RES170
 RES171 RES172 RES173 RES174 RES175 RES176 RES177 RES178 RES179 RES180
 RES181 RES182 RES183 RES184 RES185 RES186 RES187 RES188 RES189 RES190
 RES191 RES192 RES193
//
```

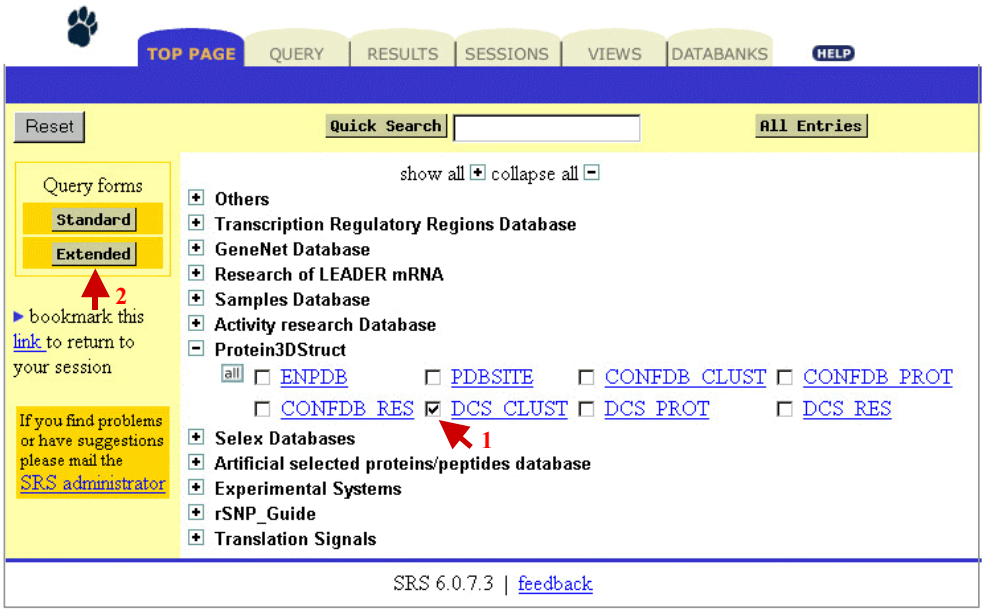
Example 2

The query is 'To search for environment class with average solvent accessibility of central residue between 70 and 90'.

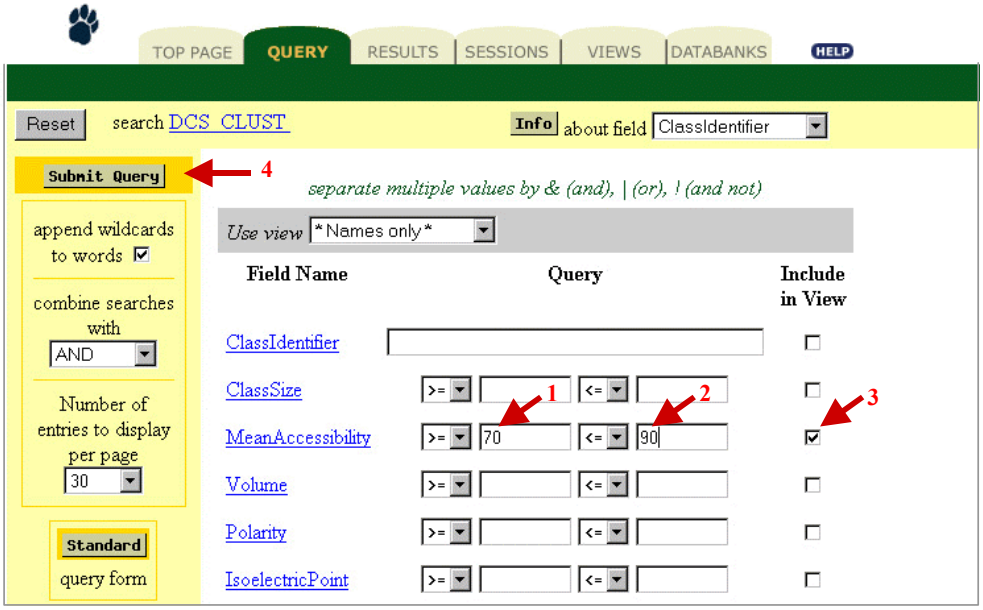
To make such a query, you should perform the following:
1. From DCS start page follow the hyperlink 'SRS_ACCESS' (1).



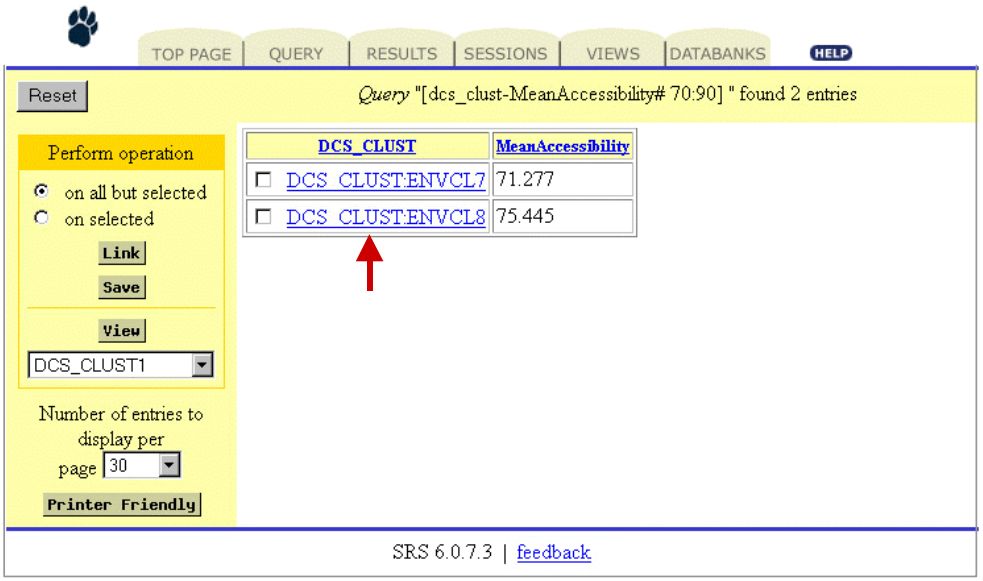
2. This will bring up the 'SRS top page'. This page is illustrated in the Figure. To perform database query, select DCS_CLUST database (1) and click extended query button (2).



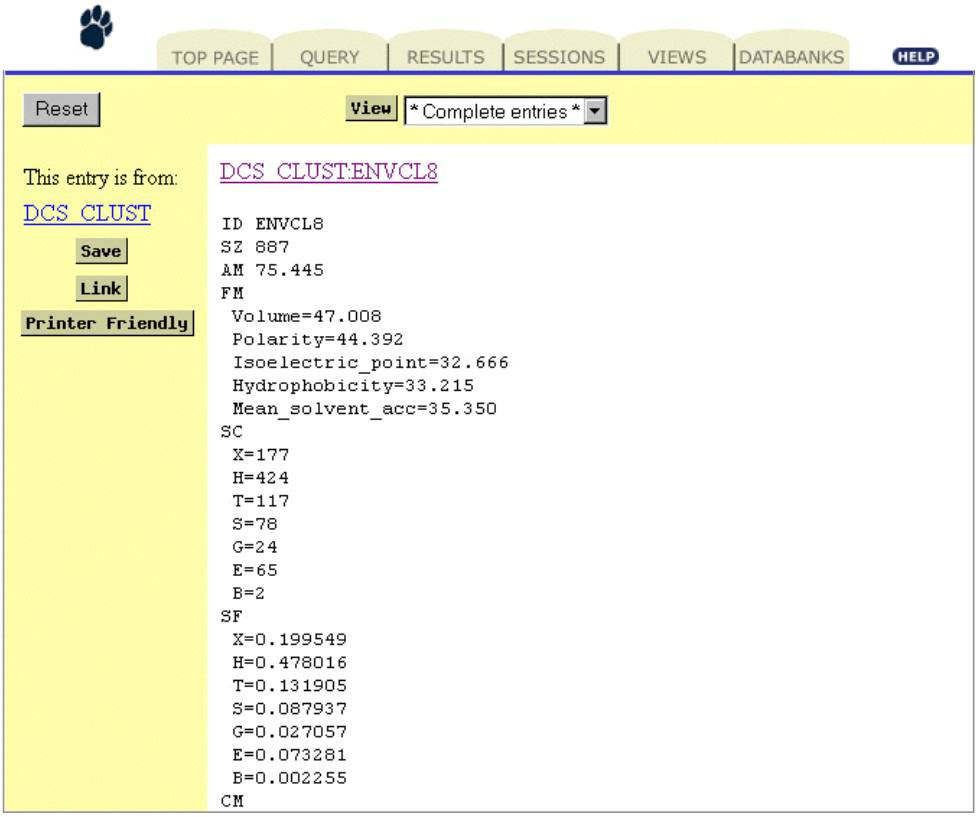
3. This will bring up the query page. This page is illustrated in the Figure. To perform database query, input lower limit 70 for 'Mean Accessibility' field (1), upper limit 90 for the same field (2), click 'Include in View' button for this field (3). Click 'Submit Query button' (4).



4. This will bring up the result page with the list of environment classes satisfying the above conditions. This page is illustrated in the Figure. To view environment class entry click to the corresponding hyperlink (1).



5. This will bring up the resulting page of the ENVCL8 entry. This page is illustrated in the Figure.



6. Database on local conformations of protein chains -"conformons" (ConfDB)

Release 2003

1. Database description:
ConfDB contain structural information on conformational properties of short fragments (9 aa in length) in 3D structures of *C.elegans* proteins.

2. Access to ConfDB database:

http://www.domain.com/mgs/gnw/confdb/

3. Database content:

SRS table	Description	Number of entries
CONFDB_RES	List of peptides	790
CONFDB_PROT	List of proteins used for database derivation	5
CONFDB_CLUST	List of clusters of peptides with similar conformations	320

4. Biological tasks that could be solved using ConfDB database:
Find peptides that have similar local conformations in protein structures.
Annotate protein residues in terms of local conformations of polypeptide chains.

5. How ConfDB was constructed?

Database constructed from short polypeptide segments. They are collected from set of protein structures by sliding window of 9 residues. These segments are overlapped. All of peptide segments were classified in conformational space in conformational clusters. Nearest neighbour method was used with r.m.s.d. cutoff 1.2 A for CA-atoms.

6. SRS table format:

CONFDB_RES

Field name	Field description
ID	Peptide identifier
PR	Reference to protein
CR	Reference to CLUSTER table
SP	Index of start position in protein structure
SS	Index of start position in protein sequence
SQ	peptide sequence
PC	peptide coordinates for CA atoms (columns: SP;SS;Amino Acid 1-letter code; X coordinate; Y coordinate; Z coordinate.

CONFDB_PROT

Field name	Field description
ID	Protein identifier
DB	Reference to protein structure in PDB (format is ID:Chain)
SZ	Size of protein (number of peptides in this protein)
PM	Peptides - members of this protein

CONFDB_CLUST

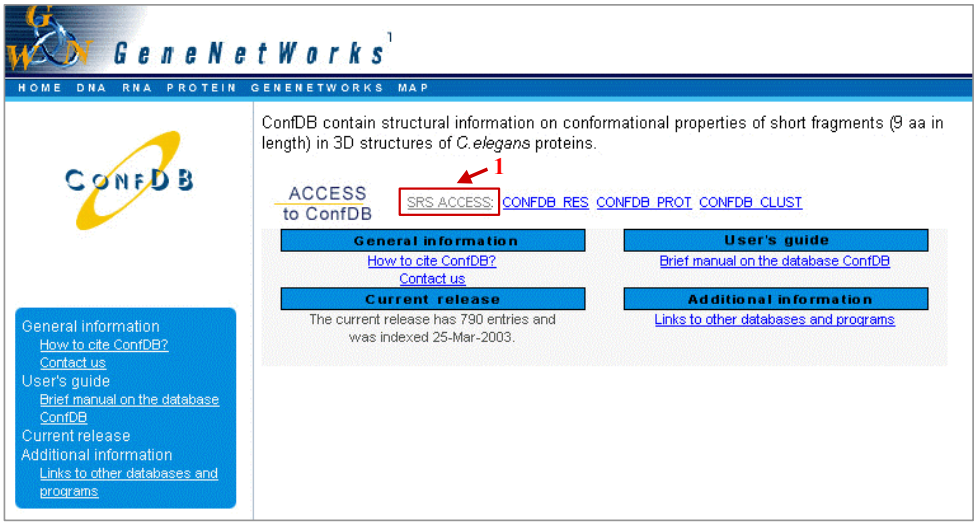
Field name	Field description
ID	Conformon cluster identifier
SZ	Size of cluster (number of peptides)
CM	Peptides - members of this class

Comments and questions to Dmitry Afonnikov (ada@bionet.nsc.ru).

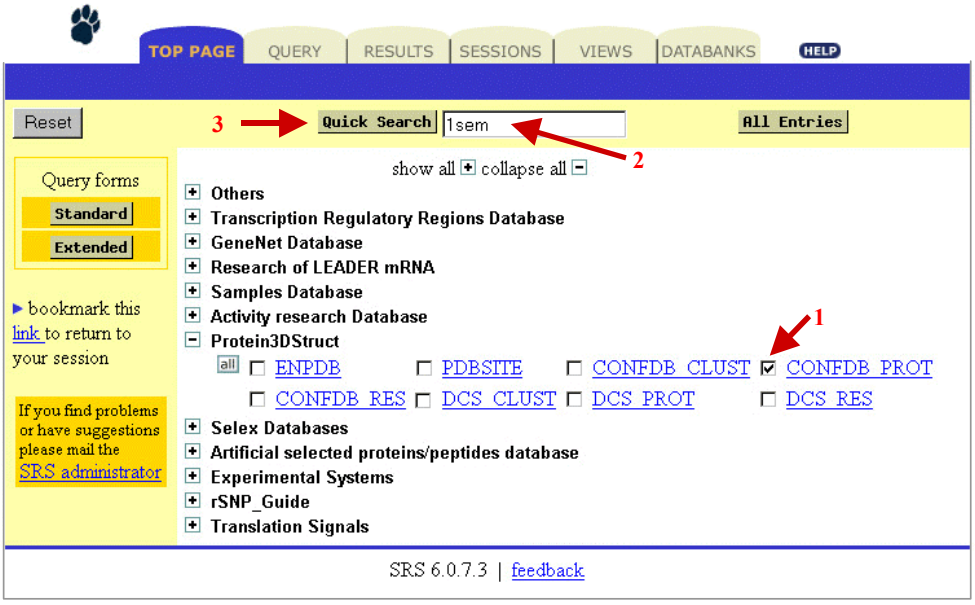
Example of SRS query to the ConfDB database

The query is 'To search for signal transduction protein (PDB identifier 1SEM), represented in ConfDB'.

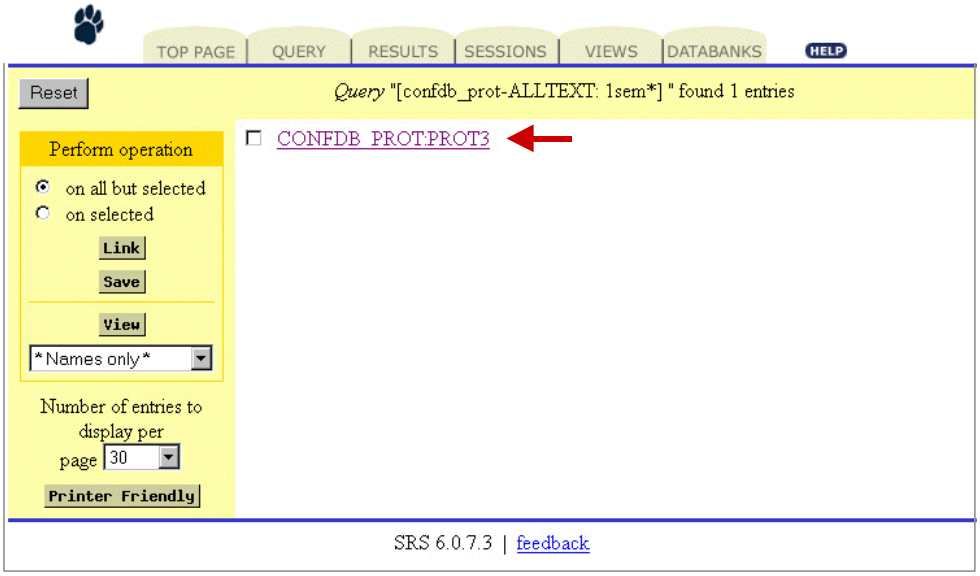
To make such a query, you should perform the following:
1. From ConfDB start page follow the link 'SRS_ACCESS' (1).



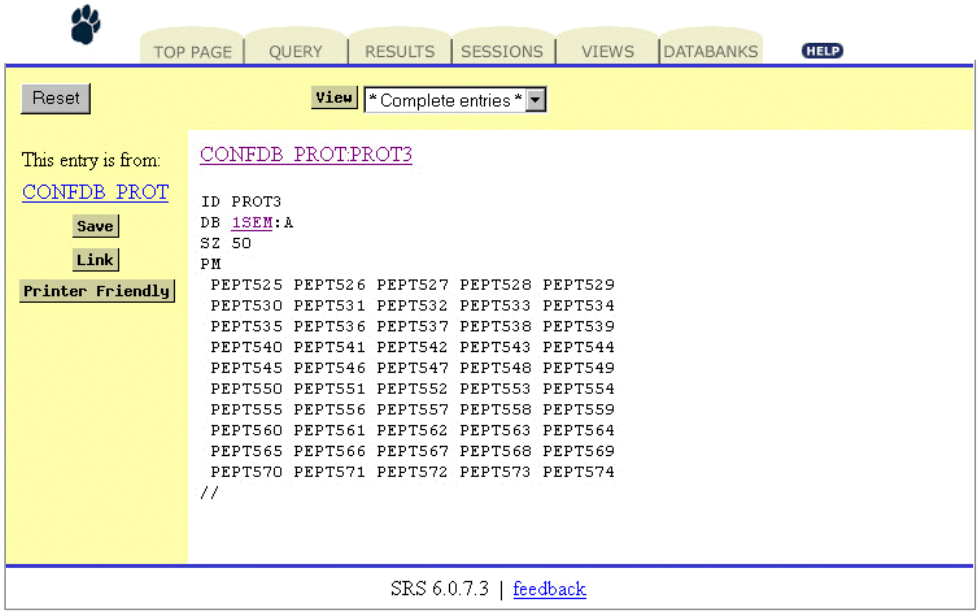
2. This will bring up the 'SRS top page'. This page is illustrated in the Figure. To perform database query, select CONFDB_PROT database (1), input query string '1sem' into query field (2) and click 'Quick Search' button (3).



3. This will bring up the page illustrated in the figure. Click to the hyperlink to PROT3 entry.



4. This will bring up the resulting page of the PROT3 entry.



CHAPTER 2. SOFTWARE

1. CRASP

1.1. Analysis of pairwise positional correlations

The program package CRASP, module P_CRASP.

Release 2003

Program description:

The program package CRASP has been developed for analysis of co-ordinated substitutions of amino acid residues in aligned sequences of protein families. The module P_CRASP is designed for evaluation of pairwise dependency of physico-chemical properties of amino acid residues.

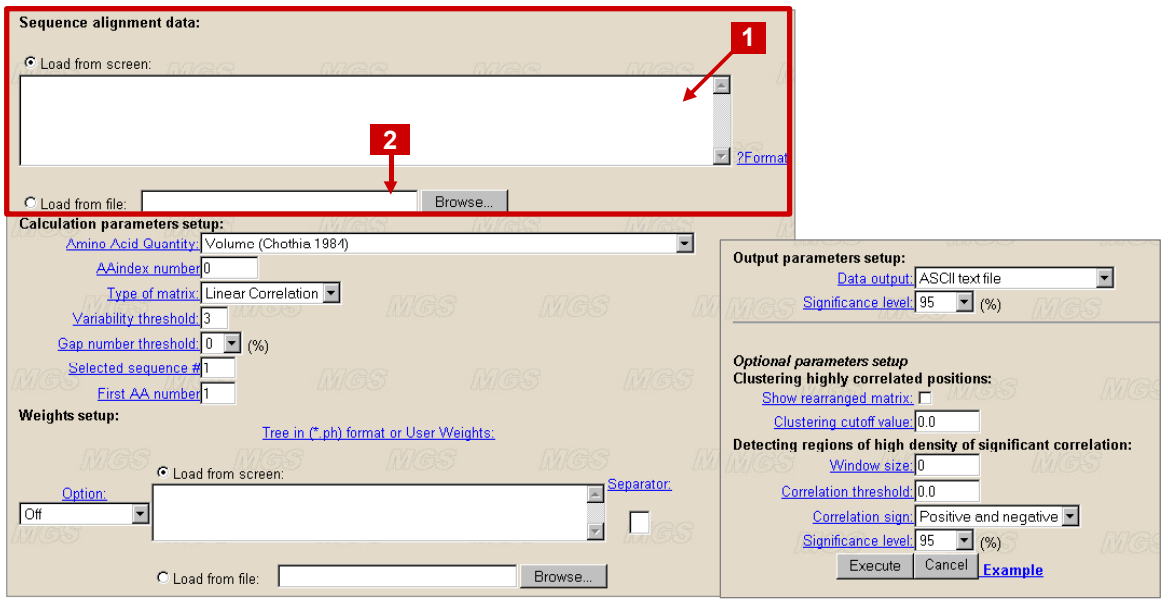
Access to CRASP package:
http://www.domain.com/mgs/gnw/crasp/ link 'Analysis of pairwise positional correlations'

List of biological tasks that could be solved by using the CRASP package

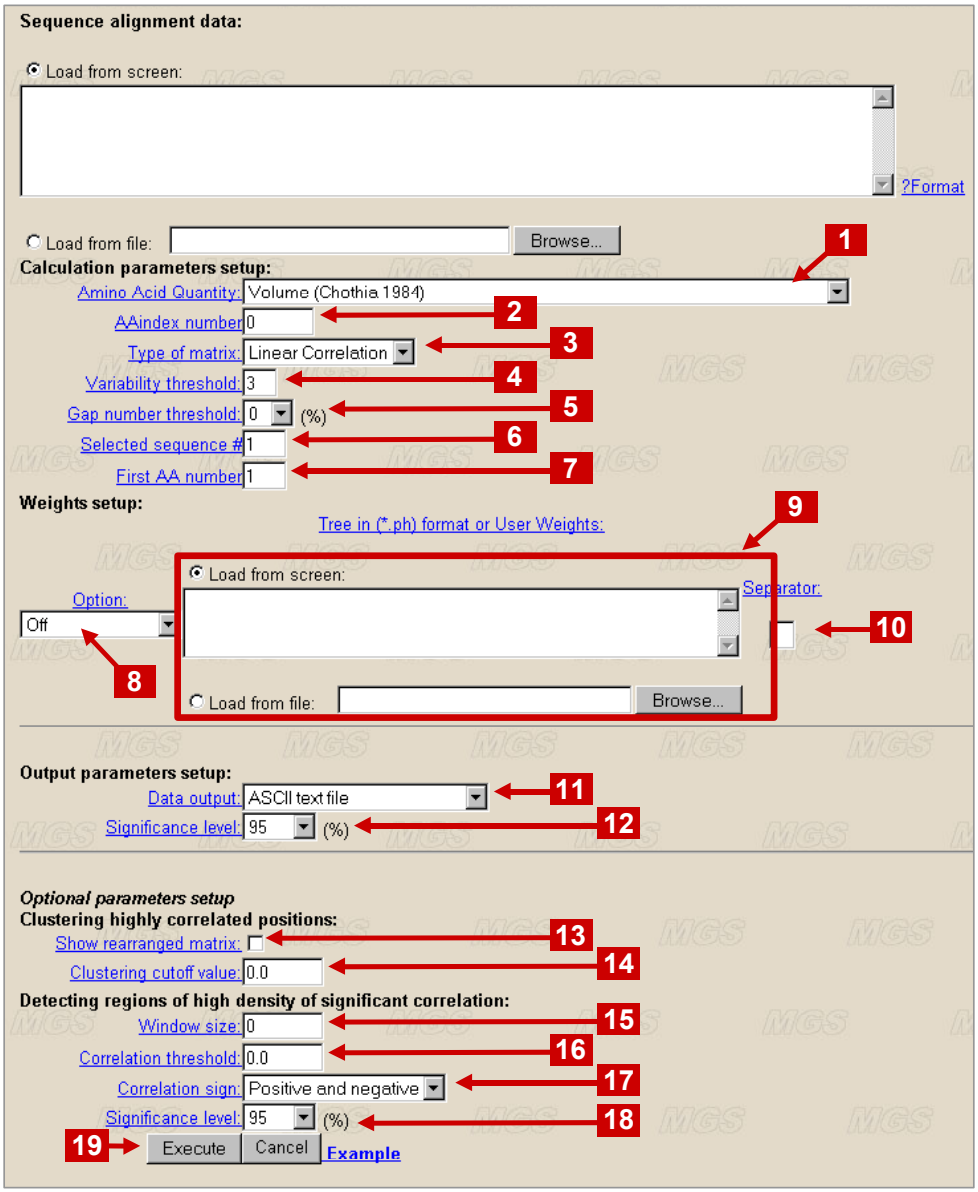
- evaluation of the mutual interdependency of physico-chemical property values in the pairs of positions of multiple alignment;
- searching for conserved (variable) physico-chemical properties of a protein

Data input

The data could be entered either from the browser screen in the appropriate text-box for data input (1) (option “Load from screen”) or from the file (2) from the user’s computer (option “Load from file”). The sequences should be aligned and entered in the FASTA format. There are no restrictions for the sequence length in alignment and the number of sequences.



Program options.



(1) Choose one of 36 physico-chemical amino acid properties. See also (2).
(2) Choose amino acid properties by the number in the database AAindex. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. The limits of this parameter range within 0-434. In case the parameter value is chosen to be equal to "0", then the choice of the property is determined by the menu (1). By default, the parameter equals to "0".
(3) Choose the type of the matrix to be calculated:

- 'Covariation' - covariation matrix calculation
- 'Linear correlation' - linear correlation coefficients calculation
- 'Partial correlation' - partial correlation coefficients calculation

By default, the option 'Linear correlation' is installed.

(4) Installation of the threshold of variability of positions. Here the minimal admissible number of different types of residues in a column of alignment is ordered. If the number of different types of residues is less than the ordered one, then such alignment column is excluded from calculation of a matrix. This parameter varies within the range 1-20. By default, the value "3" is installed.
(5) This option sets the threshold for the number of gaps in alignment column (expressed in percentage to the number of sequences). If the number of gaps exceeds the ordered value, then this column will be excluded from calculations of a matrix. This parameter is chosen from the drop-down menu within the range "0"-"15" %.

(6) This option sets the number of a sequence that will be the reference sequence in alignment. The reference sequence will be used for denoting positions of alignment. The admissible values vary from "1" to the number of sequences in alignment.

(7) This option orders the number of the first amino acid in alignment reference sequence. This value will be used for numeration when denoting alignment positions. Any whole positive numbers are accessible.

(8) This option is designed for choosing the regime of data weighting. See also description of the parameter (9).
- "Off" - all the sequence weights are equal to 1;
- "Vingron & Argos" - the method suggested by Vingron and Argos (1985);
- "User defined" - the weight coefficients are introduced by the user;
- "Felsenstein" - the method is suggested by Felsenstein (1985) and its calculation is based on phylogenetic tree data. If you these data are available, this weighting scheme is recommended.

By default, the option "Off" is installed.

(9) This option is designed for ordering weighting parameters if the option (8) is installed as "User defined" or for phylogenetic tree, if the option (8) is set as "Felsenstein". The weights determined by user are entered in a text format with separator. The symbol of separator is ordered by the parameter (10). Phylogenetic tree is entered in a text form in the format PHYLIP (*.ph). It is possible to enter data either from browser screen or from file.

(10) This option is for denoting the separator symbol. This parameter is installed in case the option (8) is set as "User defined". It is possible to use any symbol, for example, ";", ",", ":". By default, separator is set as the symbol "semicolon".
(11) this option sets the output format of the matrix calculated:
- "ASCII text file" - ASCII text file (with HTML header);
- "HTML table" - HTML-table;
- "Matrix color diagram" - color diagram for matrix elements in GIF-format;
- "Significant pairs" - color diagram for statistically significant correlation coefficients in GIF-format (not defined for covariation matrix)

By default, the option "ASCII text file" is on.
(12) This option denotes the significance level of correlation coefficient that is chosen from menu.
(13) This is an additional option – graphical output of the matrix so that positions forming the cluster are closely located in the matrix. By default, this option is off. See also (14).

(14) This parameter is absolute value of correlation coefficient that determines the level of clusterisation. This parameter ranges within the limits "0" – "1". If this parameter equals to "0", then as the threshold is taken the critical value of correlation coefficient under given level of significance (parameter (12)).

(15) Parameters (15)-(18) determine visualisation of the matrix regions with predominance of correlation coefficients that exceed the threshold r_t. The matrix region is ordered as the rectangle window. The parameter (15) determines its size. In case this parameter is set as "0", then this option is not fulfilled. The window size could be from "1" to the size of the matrix itself. By default, the value "0" is installed.

(16) This parameter sets the threshold r_t of correlation coefficient ranging within the limits "0" – "1". If this parameter is set as "0", then the threshold value is the critical value of correlation coefficient under significance level given (parameter (12)).

(17) This parameter sets the sign of correlation coefficients.
- "Only positive" – positive relationship only;
- "Only negative" – negative ones;
- "Positive and negative" – both sign;

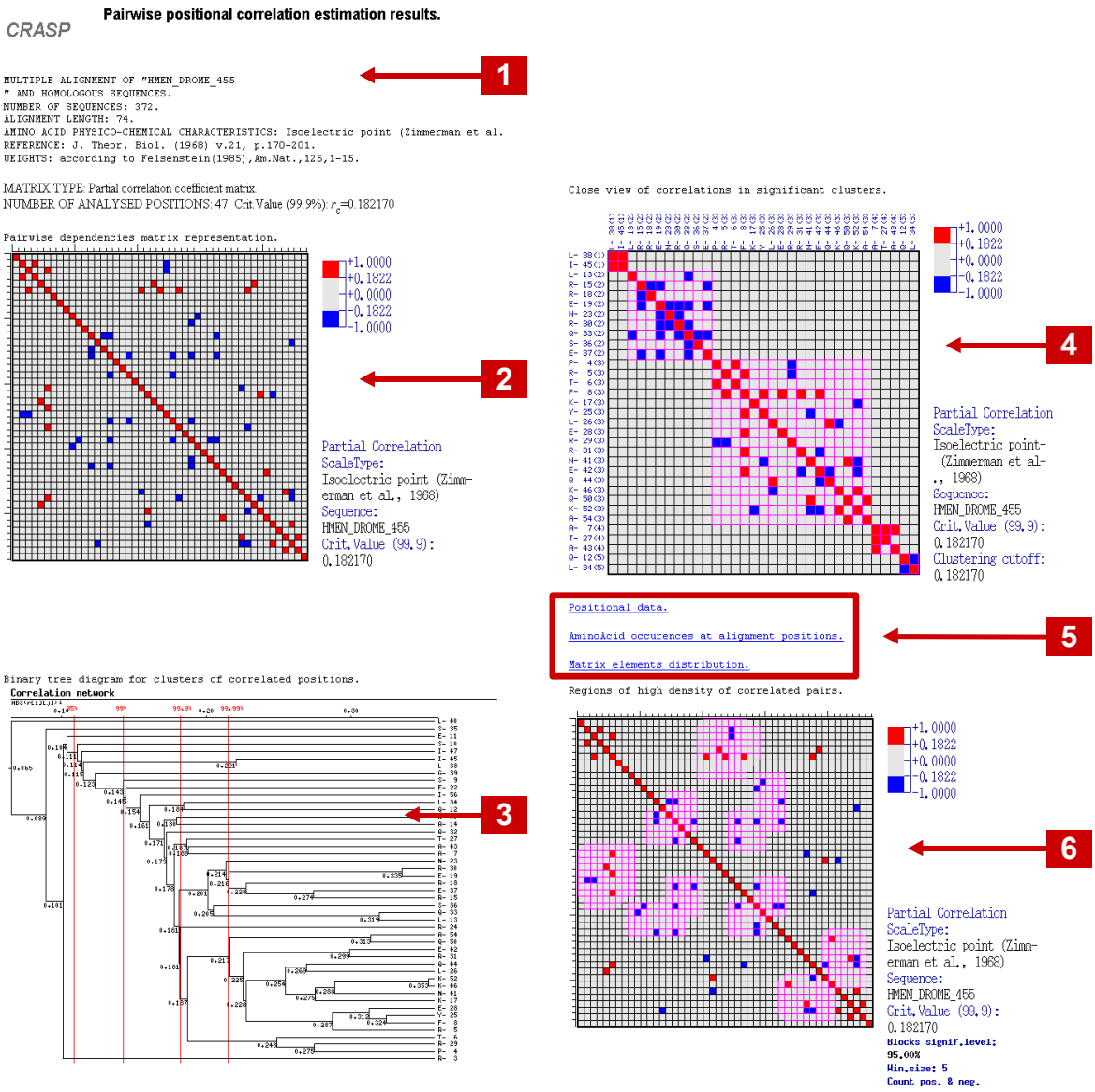
default option is "Positive and negative".

(18) Significance level. This value should be chosen from the menu.

Program execution.

(19) For the program execution, click the button "Execute".

Program output.



(1) This option is for dispalying calculation parameters.

(2) This option is for dispalying correlation matrix.

(3) This option is for dispalying the diagram of hierarchical clusterisation of positions.

(4) This option sets displaying of the ordered correlation matrix such that cluster positions in this matrix are closely located.

(5) This option enables to display additional information: parameters of distribution of physico-chemical property in columns (Positional data), amino acid ocurrences at alignment columns (AminoAcid occurences at alignment positions), matrix elements distribution.

(6) This option enables to display the regions of a matrix with predominance of significant correlation coefficients that are marked by violet colour.

Example:

The examples of data input for each program module could be displayed by the hyperlink (1):

Comments and questions are welcome to Dmitry Afonnikov (ada@bionet.nsc.ru).

1.2. Analysis of protein integral physico-chemical characteristics

The program package CRASP, module F_CRASP.

Release 2003

Program description:

The program package CRASP has been developed for analysis of co-ordinated substitutions of amino acid residues in aligned sequences of protein families. The module F_CRASP evaluates the impact of co-ordinated substitutions into the constancy (or variability) of chemical and physical protein characteristics.

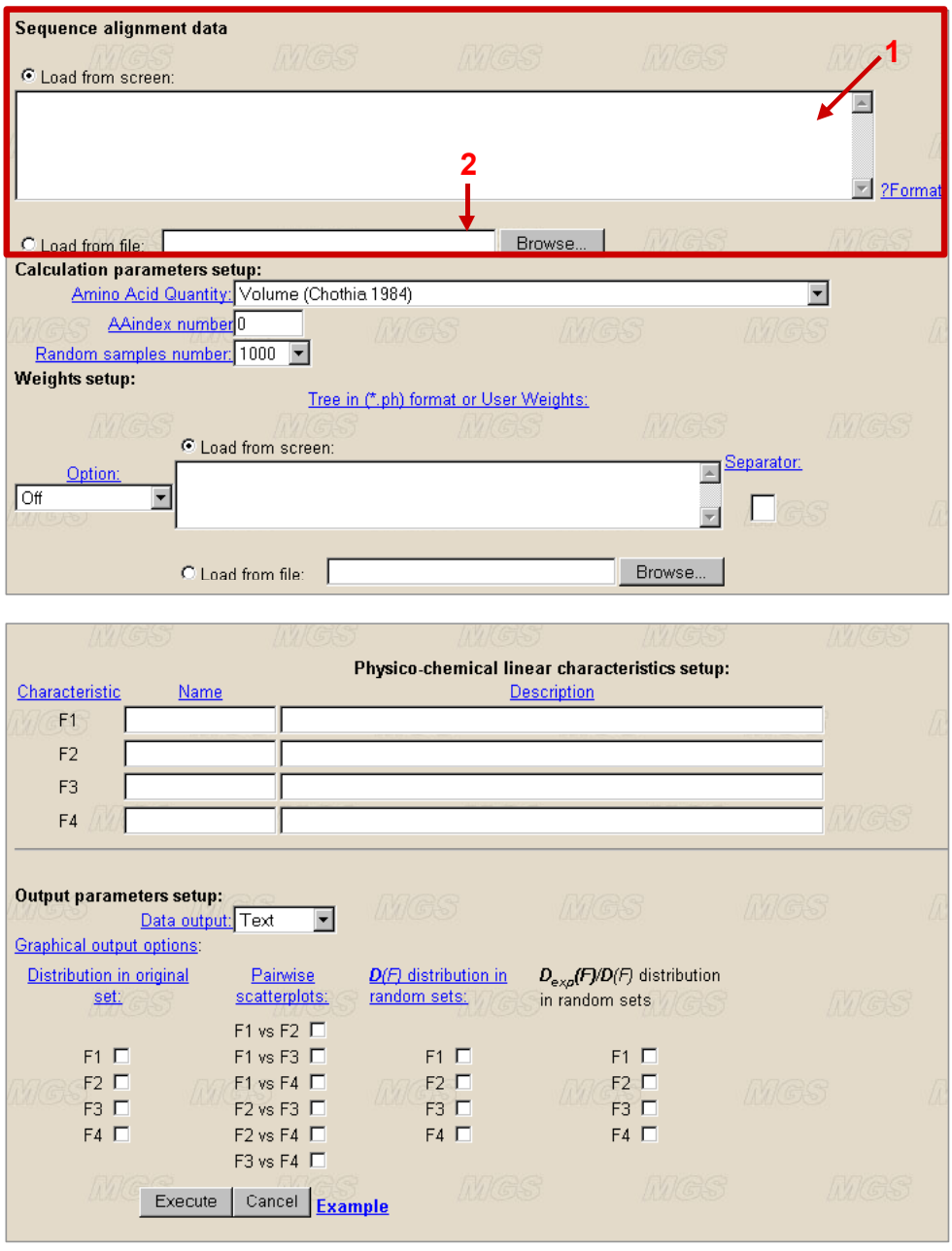
Access to the CRASP package:
http://www.domain.com/mgs/gnw/crasp/ link 'Analysis of protein integral physico-chemical characteristics'

List of biological tasks that could be solved by using the CRASP package:
evaluation of the impact of co-ordinated substitutions into the constancy (or variability) of chemical and physical protein characteristics.

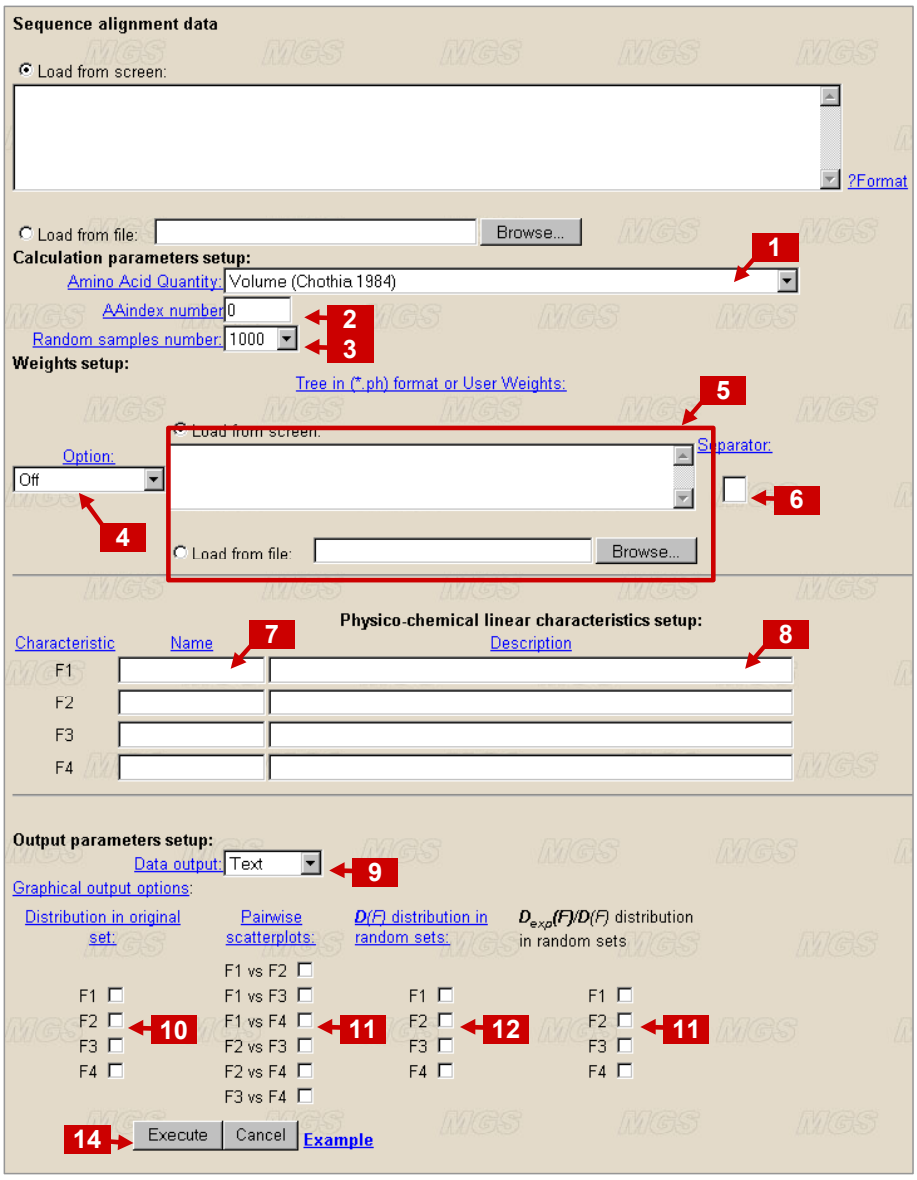
The program package CRASP consists of two modules. The first module (P_CRASP) is designed for evaluation of the pairwise correlation of physico-chemical properties of the residues.

Data input

The data could be entered either from the browser screen in the appropriate text-box for data input (1) (option “Load from screen”) or from the file (2) from the user’s computer (option “Load from file”). The sequences should be aligned and entered in the FASTA format. There are no restrictions for the sequence length in alignment and the number of sequences.



Program options.



(1) Choose one of 36 physico-chemical amino acid properties. See also (2).

(2) Choose amino acid properties by the number in the database AAindex (Tomii, K. and Kanehisa, M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 9, 27-36.). The limits of this parameter range within 0-434. In case the parameter value is chosen to be equal to "0", then the choise of the property is determined by the menu (1). By default, the parameter equals to "0".

(3) This option determines the number of random sets generated for testing statistical significance of the impact of co-ordinated substitutions into the constancy (variability) of the conserved parameter. The value of this parameter is chosen from the pull-down menu.

(4) This option determines the regime of data weighting. See also description of the parameter (9).

- "Off" - all the sequence weights are equal to 1;
- "Vingron & Argos" - the method suggested by Vingron and Argos (1985);
- "User defined" - the weight coefficients are introduced by the user;
- "Felsenstein" - the method is suggested by Felsenstein (1985) and its calculation is based on phylogenetic tree data. If you these data are available, this weighting scheme is recommended.

By default, the option “Off” is installed.

(5) The parameters of data weighting should be entered if the option (8) is set as “User defined” or, for phylogenetic tree, if the option (8) is ordered as “Felsenstein”. The weights determined by a user are entered in a textual format with separators. The symbol of a separator is ordered by the parameter (10). Phylogenetic tree is entered in a textual form, in the PHYLIP (*.ph) format. It is possible to enter the data either from screen, or from file.

(6) This option determines the symbol-separator, which is ordered if the option (8) is set as “User defined”. It is possible to use any symbol, i.e., “;”, “,”, “:”. By default, the symbol-separator is the “ semicolon”.

(7) This option determines the name of integral characteristics. The symbol line should be at most of 50 symbols.

(9) This option is designed for setting parameters of integral characteristics. To setup integral physico-chemical characteristics user should use the following format: “x1(npos1); x2(npos2); ...; xn(nposn);” xi - arbitrary numbers in a floating point format. nposi - corresponding positions of alignment (not a protein positions). They can be enumerated in an arbitrary form (using ',' and '-' symbols), i.e., “(1-3,4,5,30-44)” denotes positions from 1 to 5 and from 30 to 44. For example, net value of certain amino acid characteristic at the protein positions corresponding to alignment positions 6-8 is described as “1.(6-8);” Projection of alpha helical momentum (for helix positions 1 to 5) could be expressed as “1.(1); -0.17(2); -0.94(3); 0.5(4); 0.77(5);” where cos(0°)=1; cos(100°)=-0.17; cos(200°)=-.94; .... In total, up to 4 different integral protein characteristics related to one and the same physico-chemical property of residues (parameters (1) or (2)) could be ordered.

(9) This option orders the data output format:
- “Text” displays the textual data output;
- “Graphics” displays the resulted data as a set of images.

By default, the option “Text” is entered.

(10) This option enables to output information about distribution of integral characteristics value in the set analysed.

(11) This option enables to display information about dependency between two integral characteristics.

(12) This option enables to display information about distribution of dispersions of integral characteristics in random sets.

(13) This option enables to display information about the ratio between dispersion of characteristics in random sets and dispersion, which is expected in case the substitutions are independent (non-correlating).

Program execution.

(14) For the program execution, click the button “Execute”.

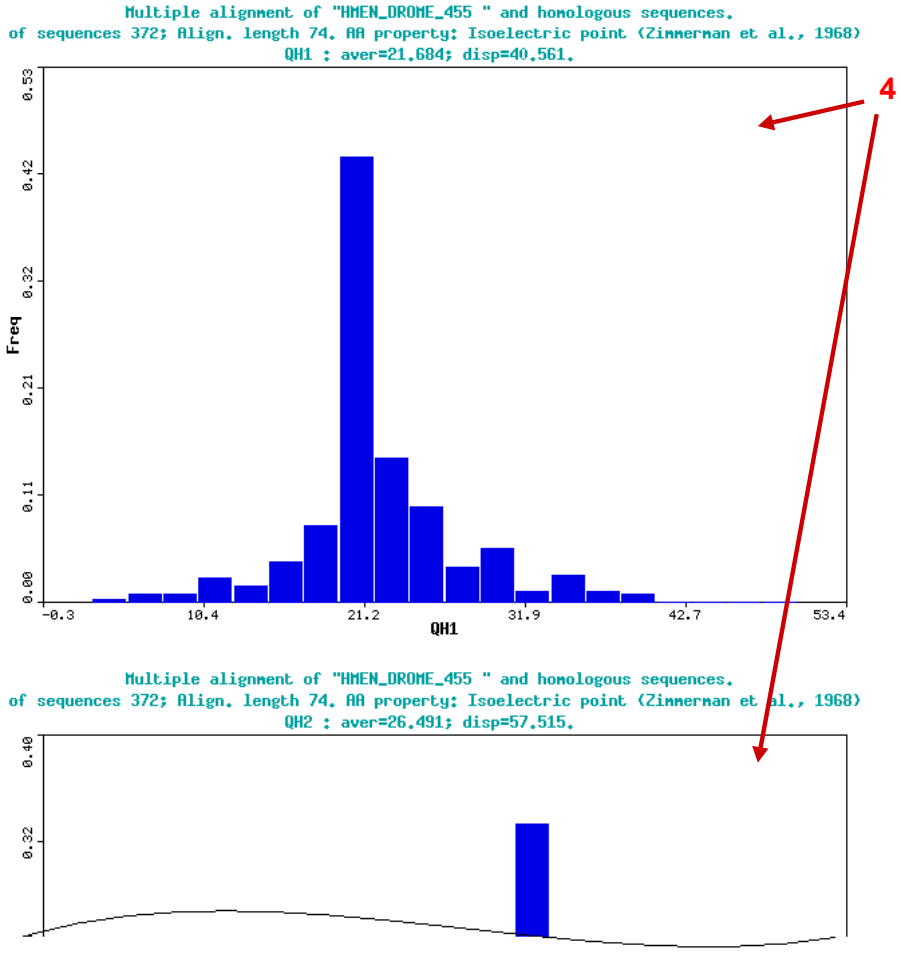
Program output.

CRASP **Integral characteristic calculation results**

```
MULTIPLE ALIGNMENT OF "HMEN_DROME_455 " AND HOMOLOGOUS SEQUENCES.
NUMBER OF SEQUENCES: 372.
ALIGNMENT LENGTH: 74.
AMINO ACID PHYSICO-CHEMICAL CHARACTERISTICS: Isoelectric point (Zimmerman et al., 1968).
REFERENCE: J. Theor. Biol. (1968) v.21, p.170-201.
WEIGHTS: according to Felsenstein(1985),Am.Nat.,125,1-15.
 Mean(M) and variance(V):
1: QH1         21.68394   40.56061
2: QH2         26.49128   57.51518
3: QH1+QH2     48.17522   80.70230
4: QH1-QH2     -4.80733   115.44930

! WARNING:
! All graphs corresponding to original sample data represent
! contrasts  calculated according to Felsenstein(1985),Am.Nat.,125,1-15.


============== F(i) values distribution in original sample  ================
```



(1) This option displays the output of main parameters of calculations.

(2) This option displays the mean values and dispersions of characteristics ordered.

(3) This warning becomes visible if the method of data weighting by Felsenstein(1985) was used.

(4) Below the histograms and the scattering plots are displayed that were ordered by the parameters (10)-(13). In the heading of each plot, the alignment parameters are indicated (number of sequences and the length of alignment); type of the physico-chemical characteristics of amino acid residues;

mean and dispersion of characteristics considered; regression and correlation coefficients and the intervals of their significance.

Example:

The examples of data input for each program module could be displayed by the hyperlink (1) shown in the picture fragment below:

Comments and questions are welcome to Dmitry Afonnikov (ada@bionet.nsc.ru).

PART IV. GENE NETWORKS INTEGRATION LEVEL

CHAPTER 1. GENENET DATABASE

Release 2003

1. Database description:

The GeneNet database is designed to accumulate the information about structure and functional organisation of gene networks

2. Access to GeneNet database:
http://www.domain.com/mgs/gnw/genenet/

3. Database content

SRS table	Description	Number of entries
GN_GENE	Genes	1006
GN_RNA	RNAs	342
GN_PROTEIN	proteins and their complexes	1766
GN_SUBSTANCE	nonproteinaceous substances	241
GN_RELATION	relationships between entities	3634
GN_SCHEME	description of gene networks diagrams	37
GN_SCHEME_ENTITY	entities (elementary structures)	3530
GN_SCHEME_RELATION	relationships in gene networks	3634
GN_COMPARTMENT	cell compartments	126
GN_ORGANISM	species	93
GN_PROCESS	input and output processes	128
GN_CELL	cells, cell lines, tissues and organs	393
GN_BIBLIOGRAPHY	references to the papers	1980
GN_EXPERT	GeneNet annotators	32

4. List of biological tasks that could be solved by using the GeneNet database

- to extract the list of entities that are involved in functioning of a particular gene network and select the items by species, compartment, type of an entity;
- to extract the list of all reactions and regulatory relations for a particular gene network;
- to browse information about all relationships that involve the protein of interest;
- to extract the list of genes, transcription of which is induced by a particular transcription factor;
- to view reactions that involve a protein, as well as its role in them.

5. SRS table format

GN_GENE

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	<Species abbreviation>:<gene abbreviation>
DT	EntryInfo	<date>; <annotator name>; <created/updated>
OS	Species	<Latin name> (<English name>)

SN	ShortName	Short gene name
NM	FullName	Full gene name
SY	Synonym	Synonym(s) of a gene name
SO	CellTableLink	Link to the GN_CELL table <Cell IC>
CH	Chromosome	Chromosomal localisation
RE	Inducer_Repressor	Inducer/repressor name
PN	Protein	Link to the GN_PROTEIN table
DR	DatabaseLink	<database name>; <database first AC>; <database ID>
RF	LiteratureReference	Link to the GN_BIBLIOGRAPHY table
CC	Comments	Comments

GN_RNA

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	<Species abbreviation>:<RNA abbreviation>
DT	EntryInfo	<date>; <annotator name>; <created/updated>
OS	Species	<Latin name> (<English name>)
SN	ShortName	Short RNA name
NM	FullName	Full RNA name
SY	Synonym	Synonym(s) of a RNA name
RF	LiteratureReference	Link to the GN_BIBLIOGRAPHY table
CC	Comments	Comments

GN_PROTEIN

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	<Species abbreviation>:<protein abbreviation>
DT	EntryInfo	<date>; <annotator name>; <created/updated>
OS	Species	<Latin name> (<English name>)
SN	ShortName	Short protein name
NM	FullName	Full protein name
SO	CellTableLink	Link to the GeneNet CELL table <Cell IC>
FN	FunctionalState	Functional state <active/inactive/no data>
MM	Multimerization Level	Multimerization state <monomer/homodimer/heterodimer/multimer/no data>
MD	Modifications	Protein modifications <phosphorylated/non-phosphorylated/no data>
GN	Gene	Link to the GeneNet GENE table <Gene IC>
RF	Reference	Link to the GN_BIBLIOGRAPHY table
CC	Comments	Comments

GN_SUBSTANCE

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	Substance abbreviation
DT	EntryInfo	<date>; <annotator name>; <created/updated>
SN	ShortName	Short substance name
NM	FullName	Full substance name
CC	Comments	Comments

GN_RELATION

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
DT	EntryInfo	Entry info: <date>; <annotator name>; <created/updated>
RE	Relation	Relation code <<type of the component entering into reaction>><component identifier>^<component localisation> -> <<type of the component that is product of the reaction>><component identifier>^<component localisation>

IN	Input	Input entity
TY	RelationType	Relation class <control/reaction>
OU	Output	Output entity
CO	RegTarget	Code of the controlled relation <<type of the component entering into reaction>><component identifier>^<component localisation> -> <<type of the component that is product of the reaction>><component identifier>^<component localisation>
EF	ReactionType	Reaction type <direct/indirect>
AT	InfluenceType	Type of regulatory event <increase/decrease/switch on/switch off>
RF	Reference	Link to the GN_BIBLIOGRAPHY table
CC	Comments	Comments

GN_SCHEME

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
NM	SchemeName	Gene network name
DT	EntryInfo	Entry info: <date>; <annotator name>; <created/updated>
MD	GeneNetDynamic Model	Link to the GeneNet dynamic model
TP	RepresentationLevel	Level of the gene network representation <cell/organism>
EN	EntityList	Link to the list of entities included in the gene network
RA	ReactionList	Link to the list of reactions in the gene network
RE	RegulatoryEventList	Link to the list of regulatory events in the gene network
DE	Description	Gene network description
OS	Species	<Latin name> (<English name>)
RR	Reference	Link to the GN_BIBLIOGRAPHY table
DR	DatabaseLink	Links to other databases

GN_SCHEME_ENTITY

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
SC	SchemeId	Link to the GN_SCHEME table
NM	SchemeName	Gene network name
ET	EntityType	Entity type <gene/process/protein/RNA/Substance>
OS	Species	<Latin name> (<English name>)
EN	EntityName	Entity name
SU	Compartment	Compartment localisation

GN_SCHEME_RELATION

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
SC	SchemeId	Link to the GN_SCHEME table
NM	SchemeName	Gene network name
RT	RelationType	Relation type <reaction/regulatory event>
IN	Input	Input entity
OU	Output	Output entity
CO	RegTarget	Code of the controlled relation <<type of the component entering into reaction>><component identifier>^<component localisation> -> <<type of the component that is product of the reaction>><component identifier>^<component localisation>

GN_COMPARTMENT

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	Compartment code
AC	IdCode2	Compartment name
DT	EntryInfo	Entry info: <date>; <annotator name>; <created/updated>
CC	Comments	Comments

GN_ORGANISM

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	Species_Latin	Species Latin name
AC	Species_English	Species English name
SN	Abbreviation	Species abbreviation
OC	Classification	Classification

GN_PROCESS

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	Process code
DT	EntryInfo	Entry info: <date>; <annotator name>; <created/updated>
SN	ShortName	Process name
NM	FullName	Complete process name
CC	Comments	Comments

GN_CELL

Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	<Species abbreviation>:<item abbreviation>
DT	EntryInfo	Entry info: <date>; <annotator name>; <created/updated>
OS	Species	<Latin name> (<English name>)
SN	ShortName	Abbreviated name
NM	FullName	Complete name
SY	Synonym	Synonymous name
RF	Reference	Link to the GN_BIBLIOGRAPHY table
CC	Comments	Comments

GN_BIBLIOGRAPHY

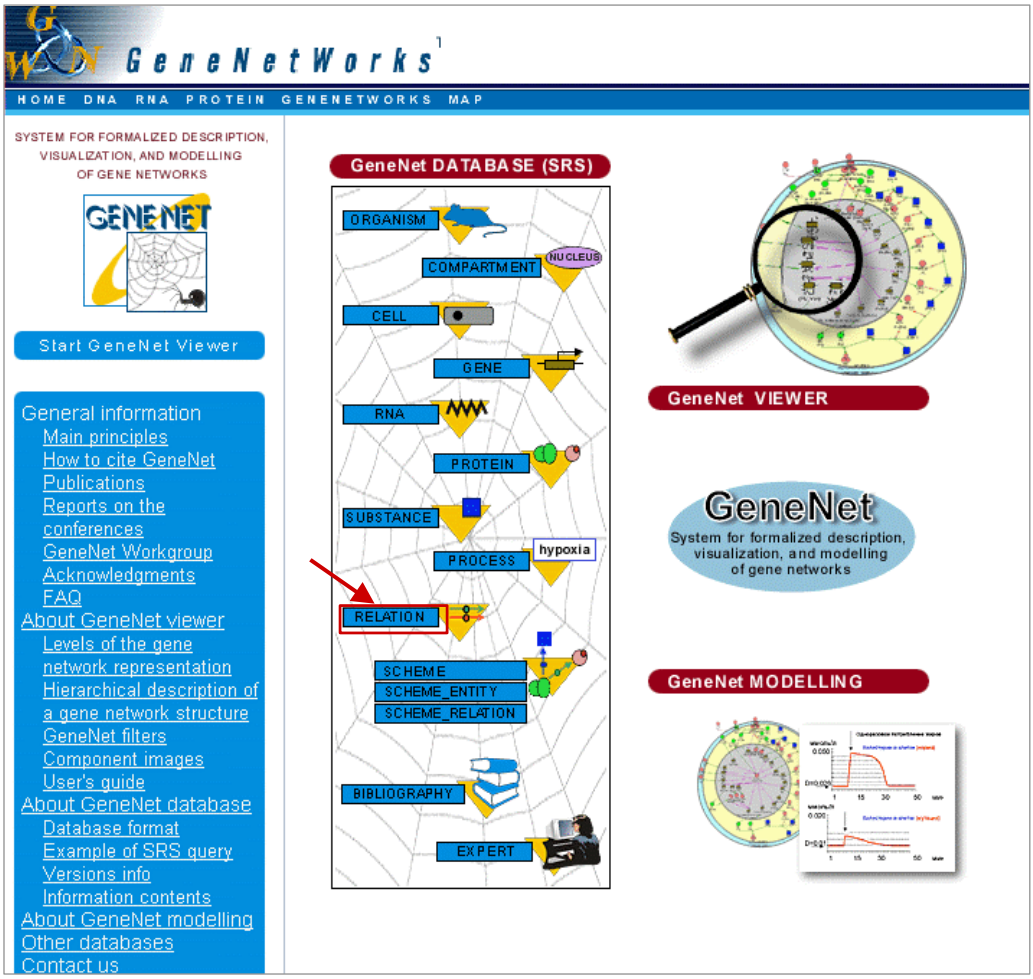
Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	<Species abbreviation>:<item abbreviation>
AU	Author	Authors of the paper
TI	Title	Title of the paper
SO	Journal	Journal name
VL	Volume	Volume
IS	Issue	Issue
YR	Year	Year
PG	Pages	Pages
ML	Medline_UI	Link to the MEDLINE

GN_EXPERT

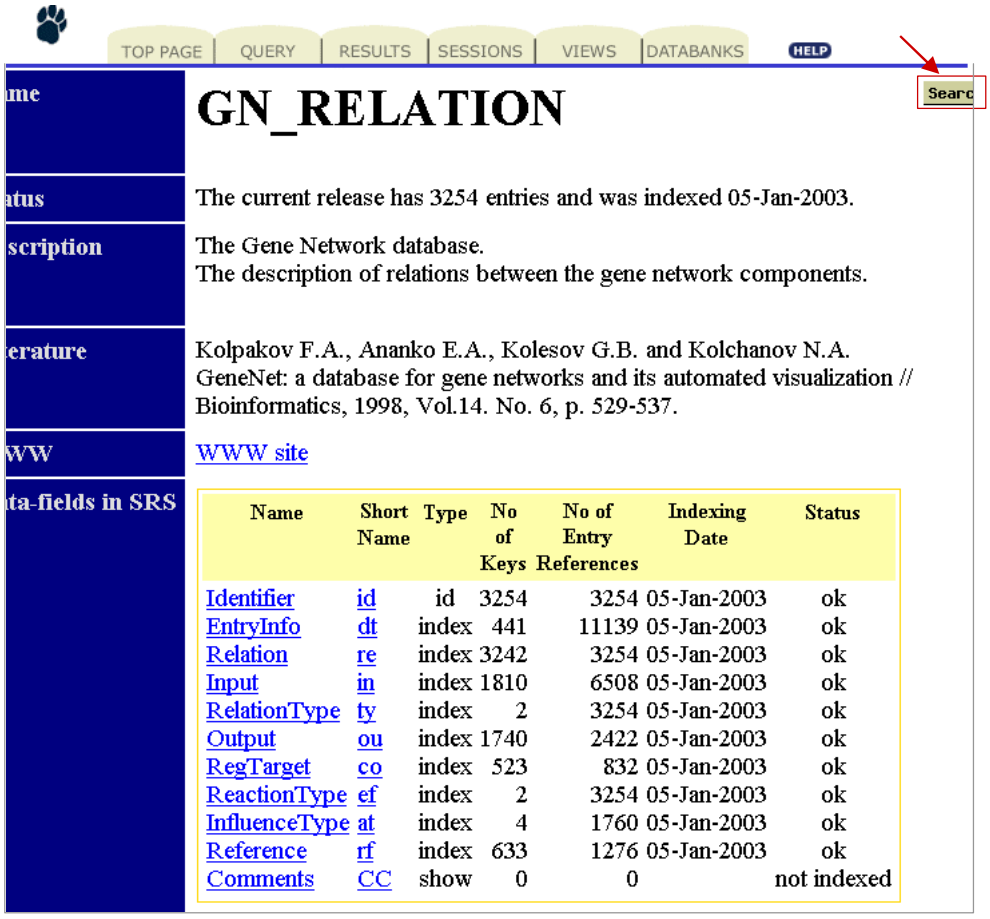
Line code	Field name	Field description
ID	Identifier	GeneNet identifier
IC	IdCode	Abbreviation of the annotator
NM	Name	Name of the annotator
LB	Laboratory	Laboratory
OR	Organization	Organization
CO	Country	Country
EM	EMail	E-mail address

Example 1
To search for all relationships described in the GeneNet database, in which IRF gene or protein partivcipate.

To make such a query, you should perform the following:
1 Choose GeneNet SRS table (RELATION) on the page 'SRS access'.



2. This will bring up the home page of the chosen GeneNet SRS table (RELATION). Click the button 'Search'.



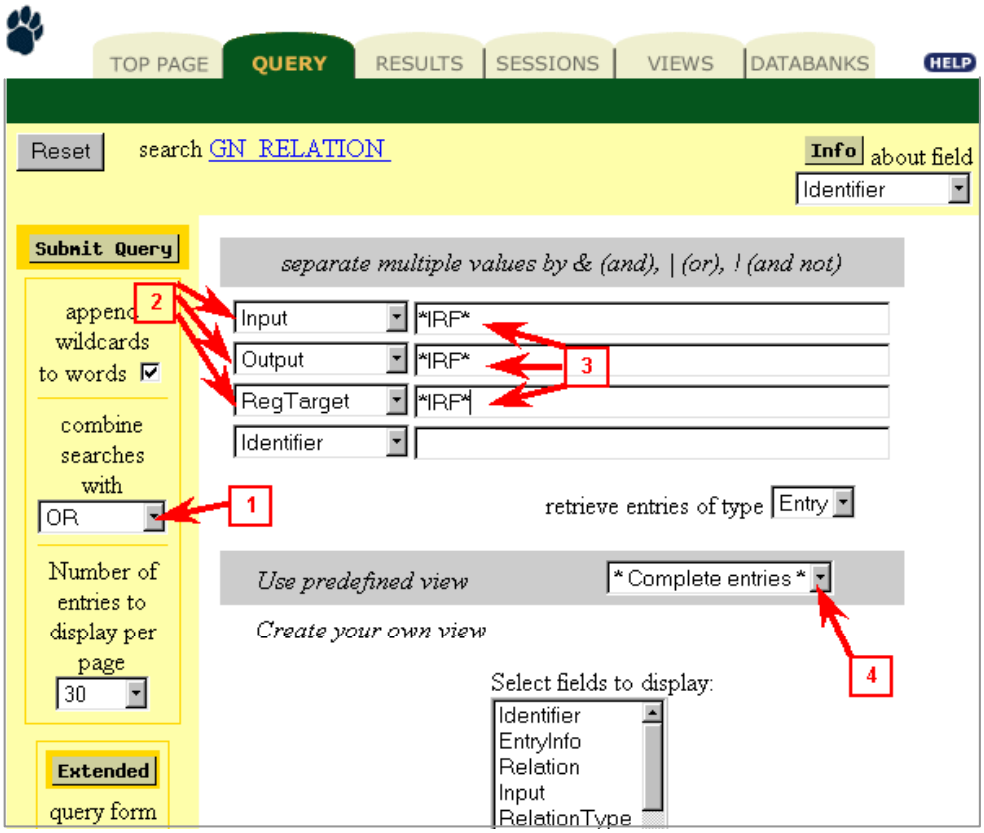
Name	Short Name	Type	No of Keys	No of Entry References	Indexing Date	Status
Identifier	id	id	3254	3254	05-Jan-2003	ok
EntryInfo	dt	index	441	11139	05-Jan-2003	ok
Relation	re	index	3242	3254	05-Jan-2003	ok
Input	in	index	1810	6508	05-Jan-2003	ok
RelationType	ty	index	2	3254	05-Jan-2003	ok
Output	ou	index	1740	2422	05-Jan-2003	ok
RegTarget	co	index	523	832	05-Jan-2003	ok
ReactionType	ef	index	2	3254	05-Jan-2003	ok
InfluenceType	at	index	4	1760	05-Jan-2003	ok
Reference	rf	index	633	1276	05-Jan-2003	ok
Comments	CC	show	0	0		not indexed

3. Select the mode 'OR' from the list of the combine searches (1).

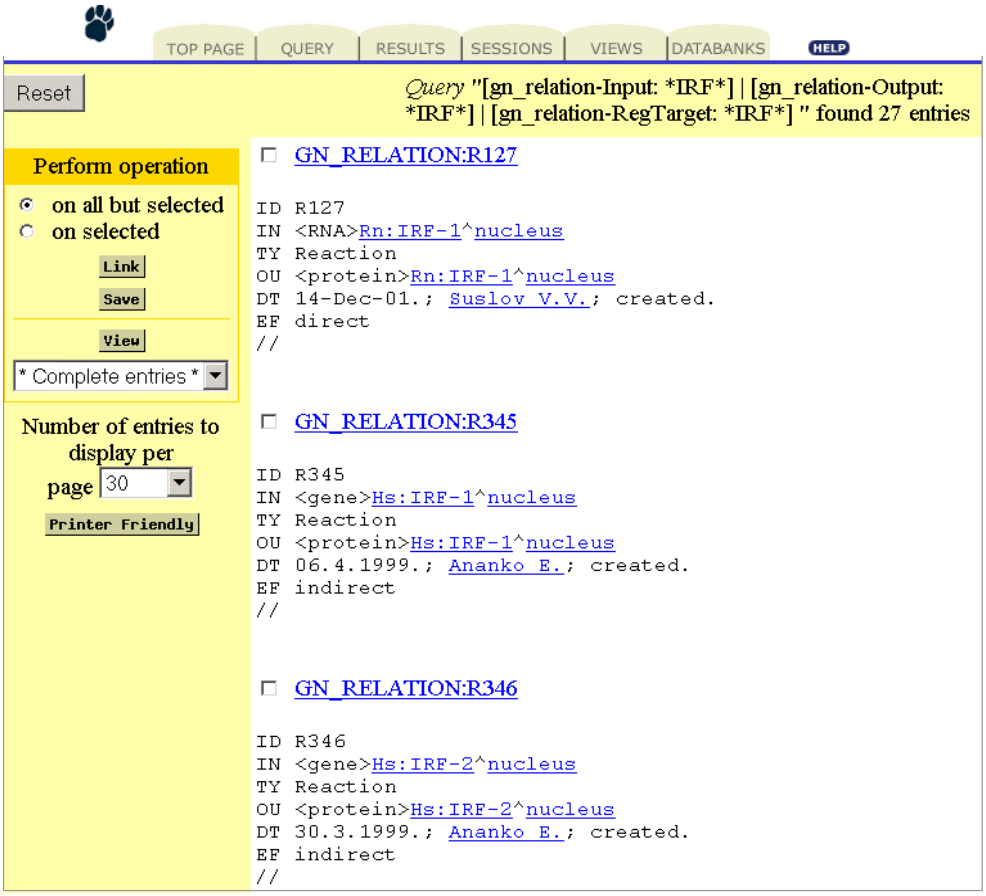
Select the fields to be searched for from the lists (2). You will need the following fields: 'Input', 'Output', 'RegTarget'.

Type term *IRF* to be searched in the text windows (3). (An asterisk marks "any symbol")

Then choose the parameter 'Complete entries' for displaying the query results (4) and click the button 'Submit Query' (5).



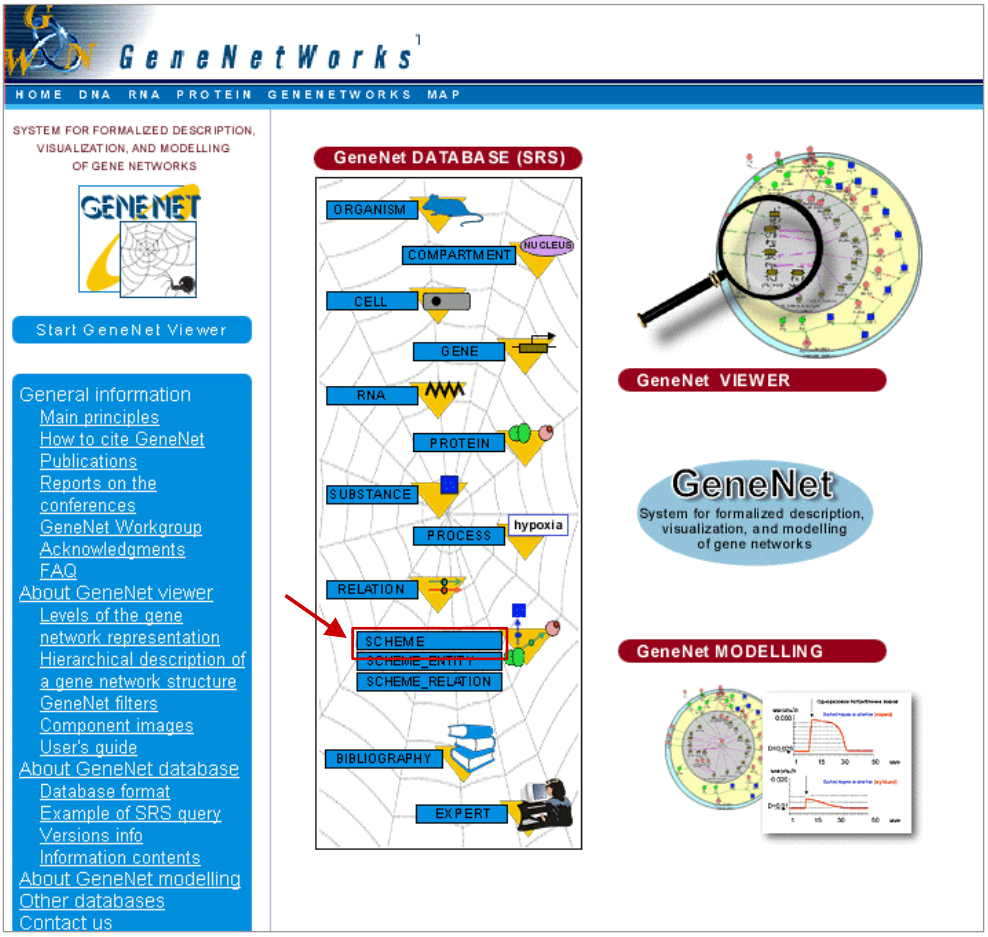
4. The query result is shown below.



Example 2

To visualise description of the diagram 'Macrophage activation' and the list of all entities and relationships entering this diagram.

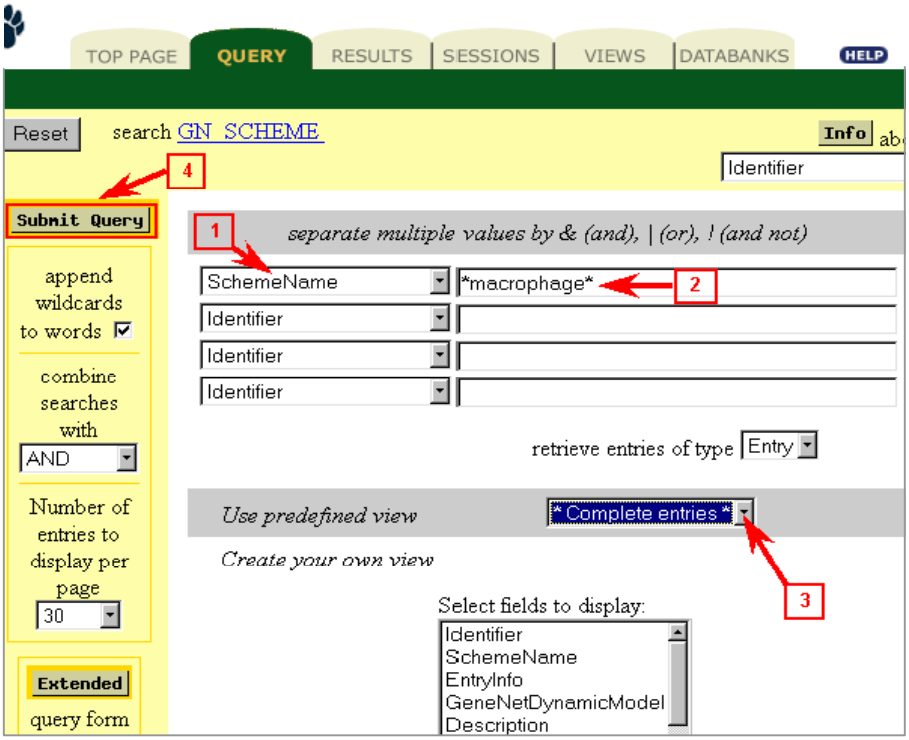
To make such a query, you should perform the following:
1 Choose GeneNet SRS table (SCHEME) on the page 'SRS access'.



2. In the Query Form for GN_SCHEME select the field 'SchemeName' to be searched for from the list (1).

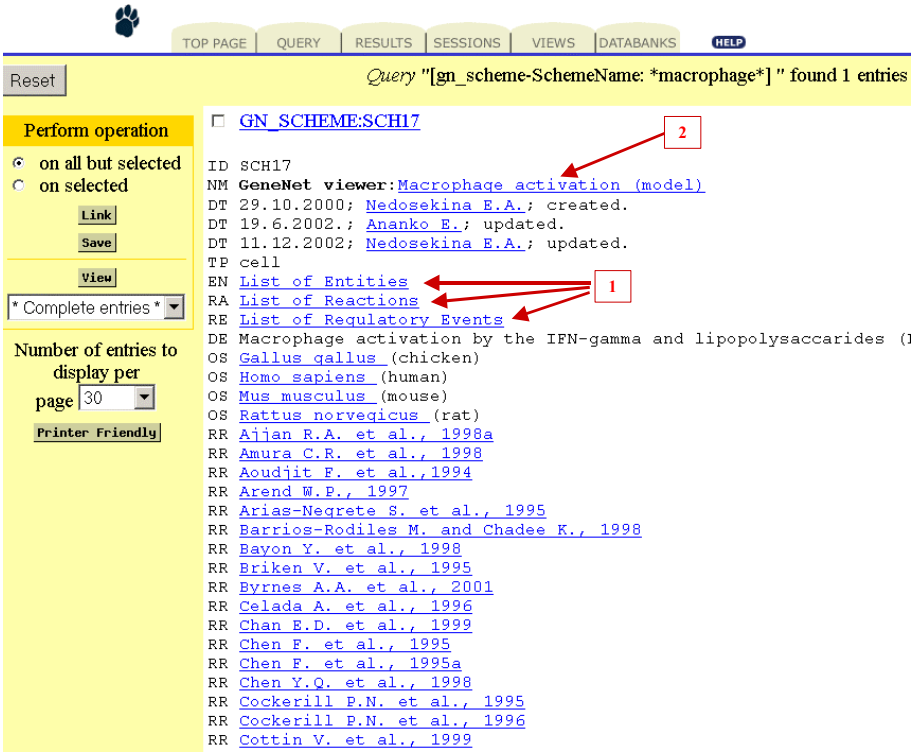
Type term *macrophage* to be searched in the text windows (2). (An asterisk marks "any symbol").

Then choose the parameter 'Complete entries' for displaying the query results (3) and click the button 'Submit Query' (4).



3. This will bring up the description of the gene network of macrophage activation in the GeneNet database.

By clicking the links 'List of Entities', 'List of Reactions', 'List of Regulatory Events' (1), you may extract the complete lists of Entities, Reactions, and Regulatory Events for this gene network. Also, here are the links to GeneNet viewer (2).



Comments and questions are welcome to Elena Ananko (eananko@bionet.nsc.ru).

CHAPTER 2. GENENET SOFTWARE

1. GeneNet Viewer.

Release 2003

1. Program description:
Software for graphical representation of the information stored in the GeneNet database

2. Access to GeneNet Viewer
http://www.domain.com/mgs/gnw/systems/genenet/applet_genenet_viewer.shtml

3. List of biological tasks that could be solved by using the GeneNet Viewer:
- To display the general visualisation of the structure and functional organisation of the gene network described in the GeneNet database;
- To view, which data were obtained for particular species or organisms.

4. Data input

Data input is available only via the special program Data Input GUI (http://www.domain.com/mgs/gnw/systems/genenet/). To use this program via the Internet, it is necessary to register by the address http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/registration.html.

5. Program options

System of filters

The data obtained from different species are summarised in the scheme of the gene network. As a result, the diagram may contain several equivalent objects (for example, homologous genes or proteins from different species). The equivalent components of the gene network are displayed as a single image.

By default, the gene network diagram is drawn from all the information given in the corresponding entry from the scheme table.

A system of filters enable a user to select for visualisation only the entities and relations that have been experimentally identified for a specified organism or specified cell types as well as those specific to the cell response to specified external stimuli. For this purpose, the GeneNet is provided with filters of three types according to:
1. Species
2. Cell type
3. Inducer

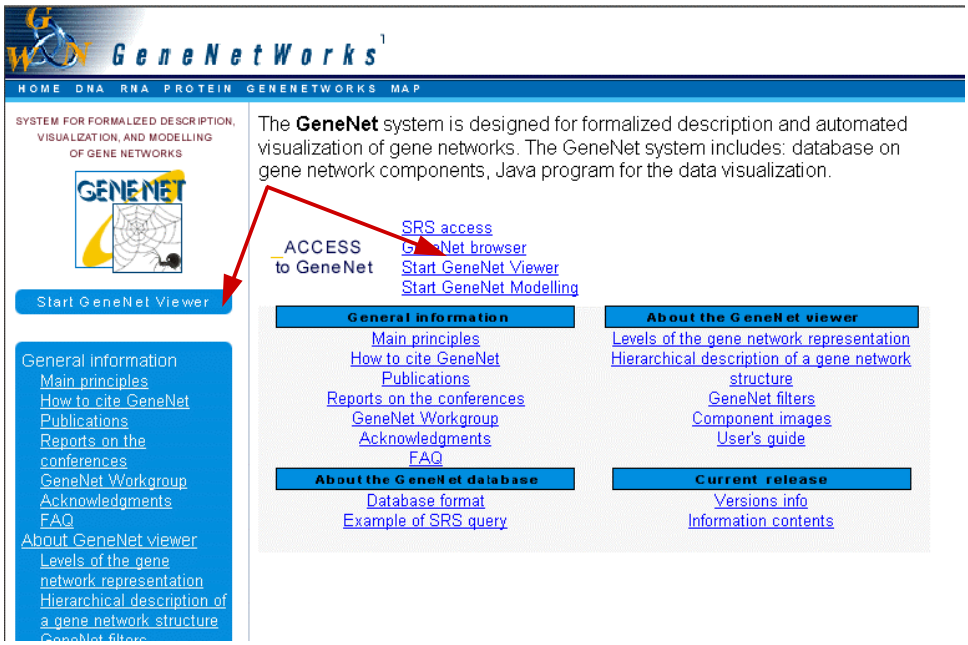
All the three filters can be used at the same time. As a result, only the objects that meet the requirements of all the filters, will be displayed in the diagram.
The system of filters is available via the menu 'Filters' (see Example).

Zooming

Zooming of a diagram is ordered through the menu 'Diagram', 'Zoom" (see Example).

6. Program execution

The program is executed by clicking the respective links at the GeneNet Home Page (they are marked by red arrows in the figure), as well as from SRS entries in GN_SCHEME table via the special link in the field 'NM GeneNet viewer:'



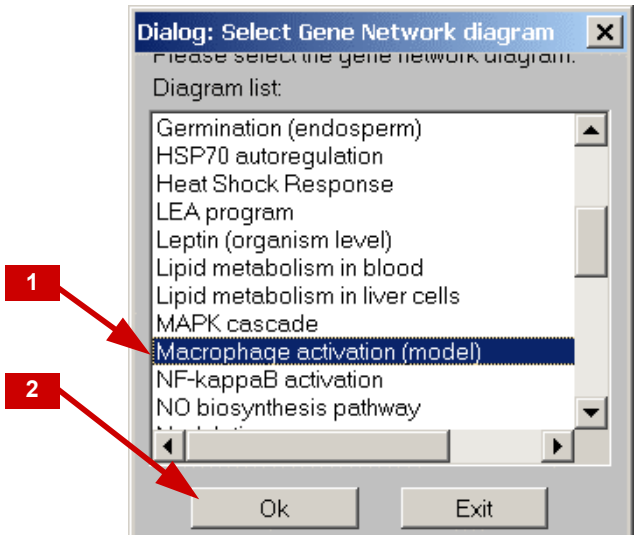
7. Data output

Data are displayed as a graphical diagram of structure and functional organisation of a gene network.

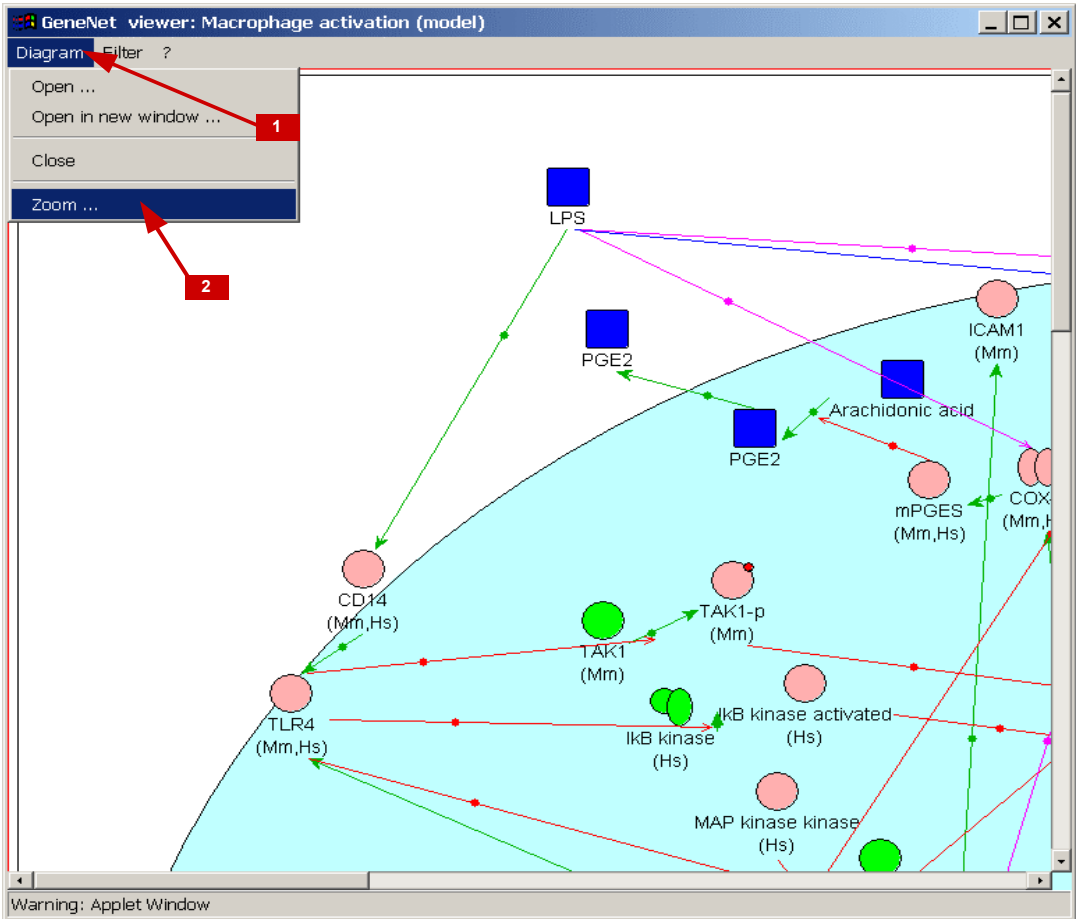
Example

Visualisation of the diagram 'Macrophage activation (model)' by means of GeneNet Viewer

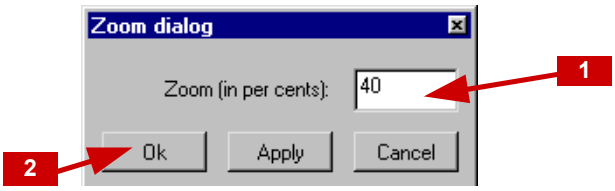
1. Start GeneNet Viewer. After data loading into the dialog window, select the option 'Macrophage activation (model)' (1) and click the button 'OK' (2)



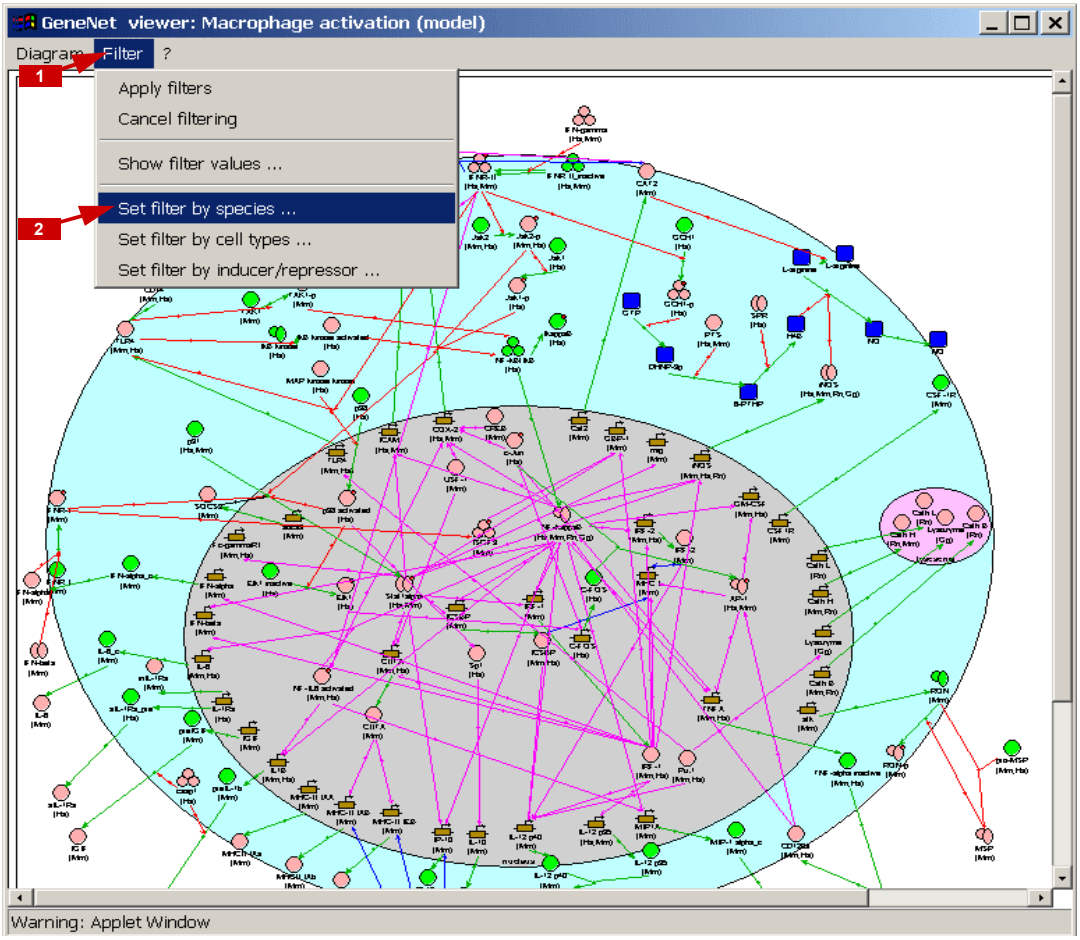
2. This will bring up the graphical diagram of the gene network Macrophage activation (model). In order to modify the zoom of the diagram, select in the menu 'Diagram' (1) the option 'Zoom' (2).



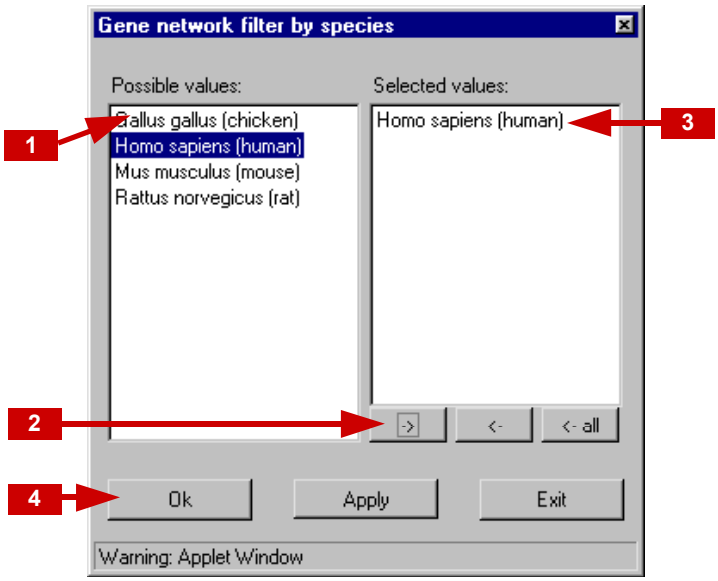
3. Enter the percentage value ‘40’ into the textual window (1) and click the button 'OK' (2).



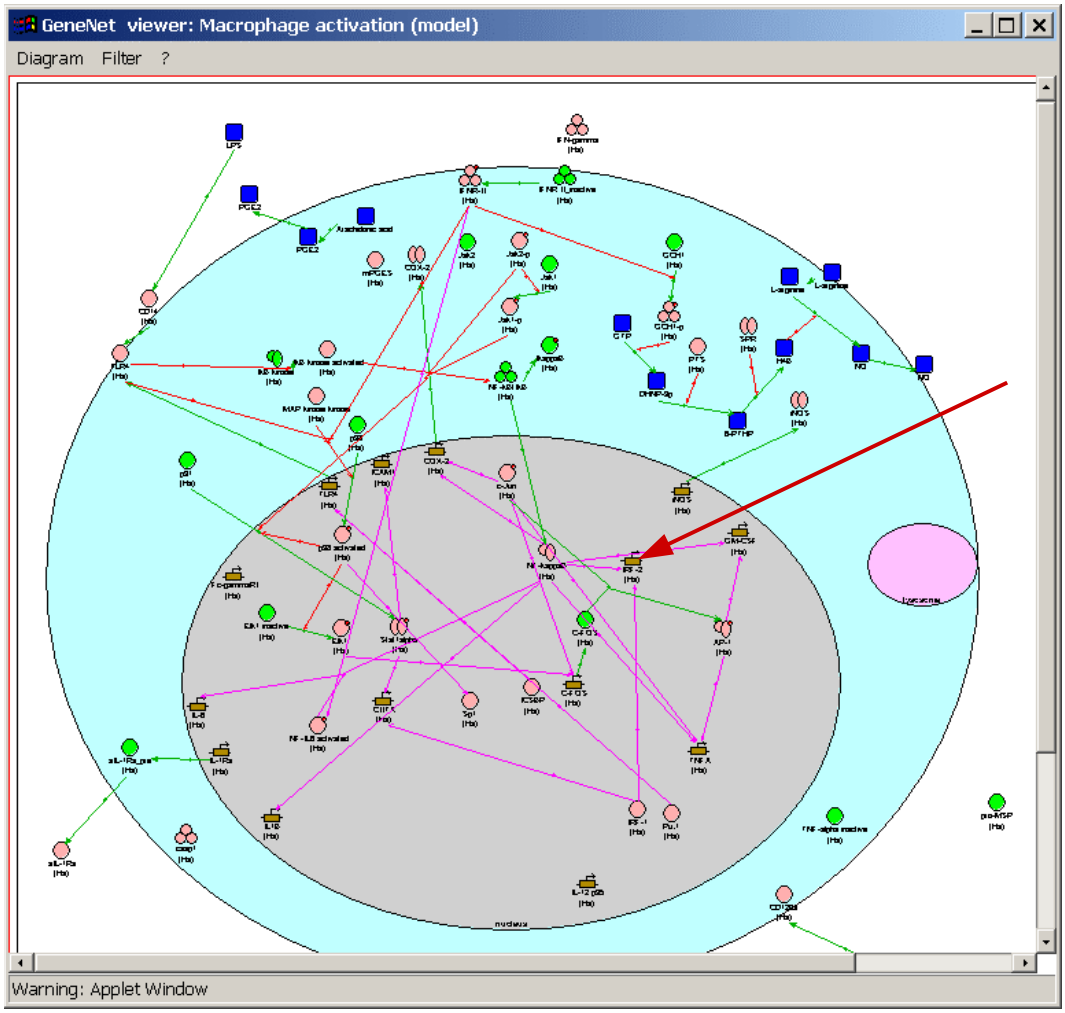
4. This will bring up the window with the altered zooming of the diagram.
In order to display the visualisation of data obtained only for human cells, select in the menu 'Filter' (1) the value 'Set filter by species' (2).



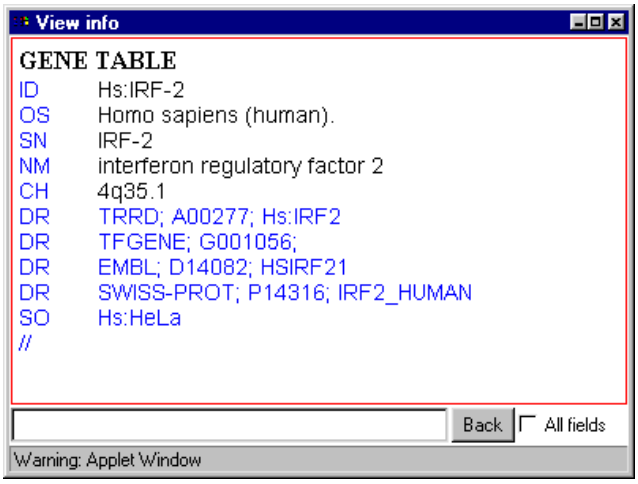
5. In the dialog window for setting the filters, choose the value 'Homo sapiens' (1), click the button '->' (2). This will bring up the option 'Homo sapiens' in the window from the right (3). Click the button 'OK' (4).



6. This will bring up the diagram displaying the data obtained by experiments only with human cells. The data obtained for the cells of the other organisms will disappear from the diagram. To display textual description of an object in the GeneNet database, click this object by mouse twice.



7. Textual description of the chosen object will be displayed in the special text window, in which the hyperlinks to the other databases are also available.



Comments and questions are welcome to Elena Ananko (eananko@bionet.nsc.ru).

2. GeneNet Modelling

Release 2003

1. Program description:

GeneNet Modeling is a software for operating with mathematical models of gene networks dynamics. The current version of the GeneNet Modelling operates with three models:

- Model of cholesterol biosynthesis regulation
- Model of erythrocyte maturation regulation
- Model of macrophage activation under the action of interferon gamma and/or lipopolysaccarides (LPS).

2. Access to GeneNet Modeling
http://www.domain.com/mgs/gnw/gn_model/

3. List of biological tasks that could be solved by using the GeneNet Modeling:

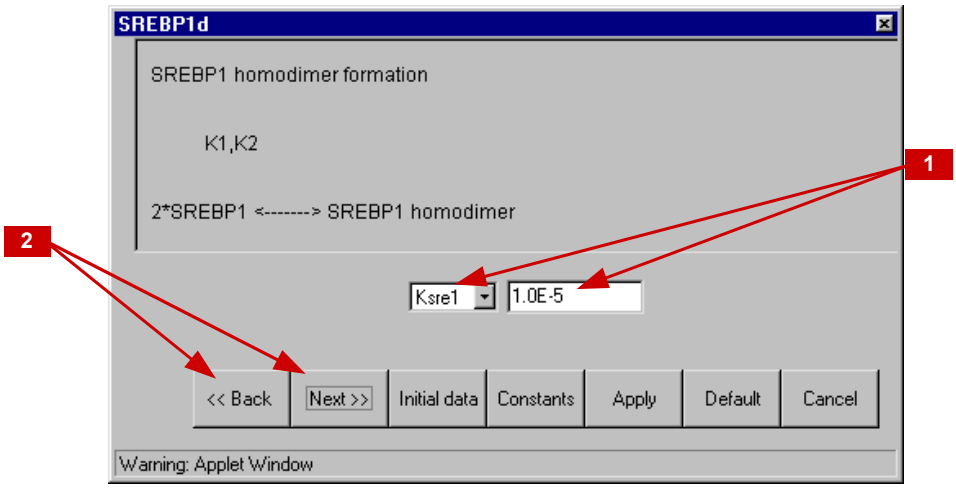
- viewing the reaction of a gene network in response to alteration of initial concentration values of the system's components;
- viewing the reaction of a gene network in response to alteration of the values of concentrations of the system's components during fixed periods of time;
- viewing the gene network behavior in response to alteration of parameters determining the rates of elementary processes;
- computer simulation of gene mutation (disruption of a gene or enzyme function, alteration of rates of translation, transcription, transportation, utilisation processes, etc.) by varying parameters of a gene network mathematical model.

4. Data input

Novel structure components and interactions might be added into existing model only be means of the special program Data Input GUI (http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/), which could be used via Internet after registration by the address http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/registration.html.

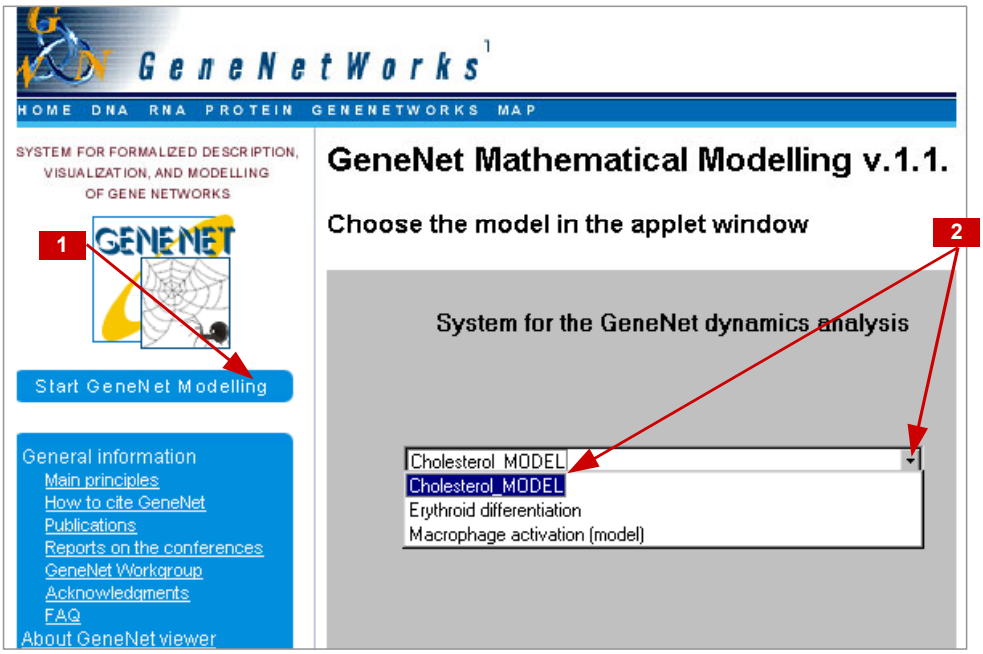
5. Program options

In the regime of the graphical interface it is possible to modify reaction rates and initial concentrations of some components. By clicking by mouse the component chosen, user dispalys the dialog window, in which the parameters available for a given gene network component could be modified (1). It is possible to view all the parameters of this component available for alteration by clicking the buttons 'Next' and 'Back' (2).
Note: if the dialog window is not open by clicking the object, this means that alteration of parameters of this object is prohibited for user.



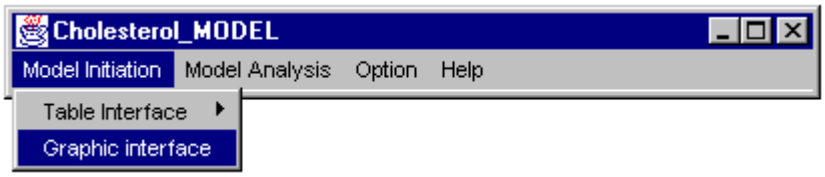
6. Program execution

Click 'Start GeneNet Modelling' (1)
This will be your working window of the GeneNet Modelling package.
Choose the model to operate with (2).



If you click the option ‘Cholesterol_MODEL’ in the drop-down menu ‘Select a model’, this will bring up the starting page of cholesterol biosynthesis gene network modeling.

Choose on the menu bar the options ‘Model Initiation’ and ‘Graphic interface’ to initiate the graphical interface of the model of cholesterol biosynthesis.



7. Data output

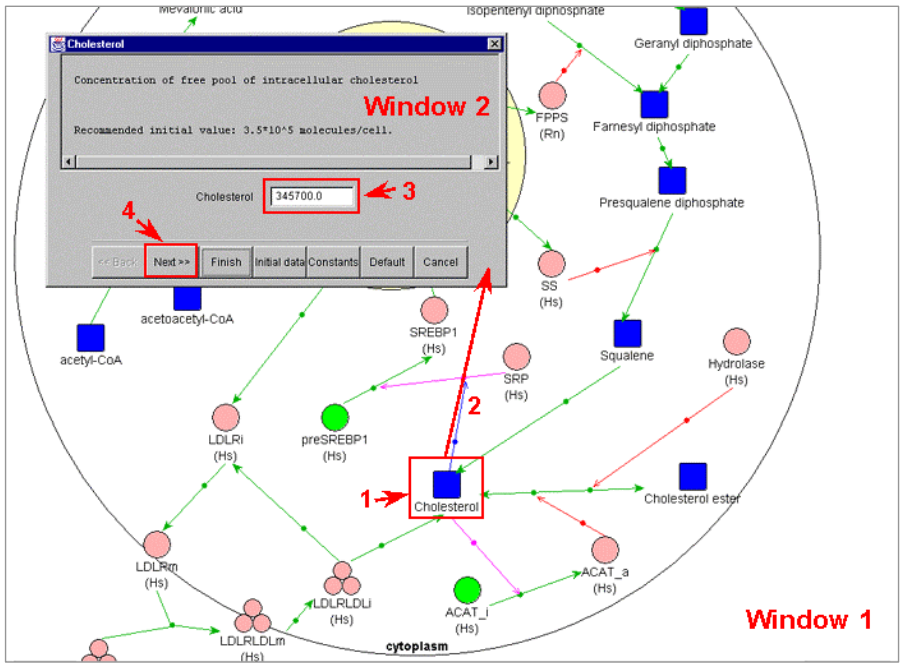
The data could be output both in graphical representation and in a form of a table (see Example).

Example 1

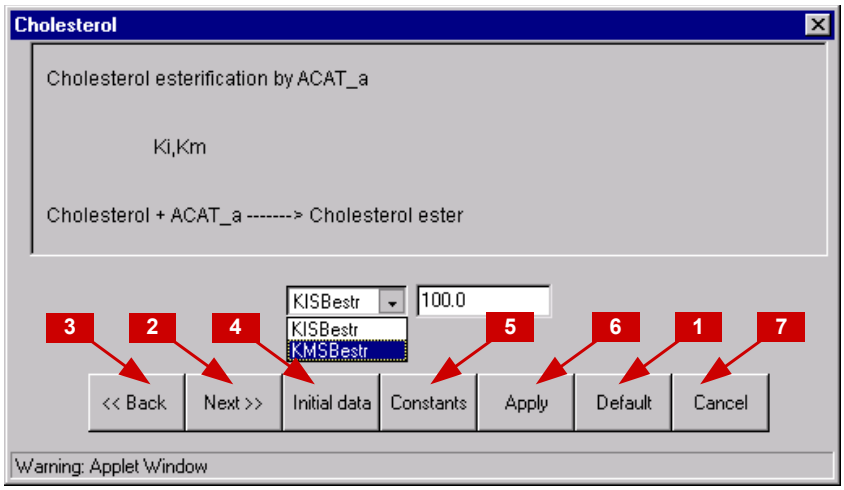
Operating with dynamic model of inner cellular cholesterol concentration.

Use it in parallel with the tutorial window. Tutorial pages will guide you how to work with mathematical models of gene networks.

1. Modification of parameters of a component.
Select the component 'Cholesterol' at the diagram (1).
By clicking the component you will get the dialog window (2). You may change the value of initial concentration (3) of the component selected.
If you want to restore initial values that are input by default, click the button ‘Default’.
For mogifying the next parameter, click the button 'Next' (5).



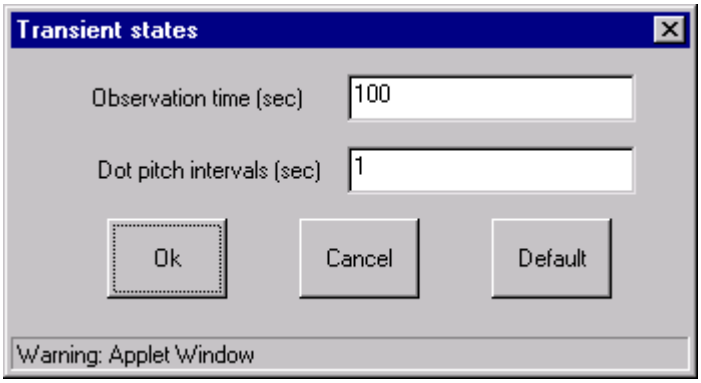
2. This brings up the dialog window with description of reaction(s) in which the selected component is involved. You may change the reaction constant(s) value(s).
If you would like to return to initial values of the constants in this window installed by default, click the button 'Default' (1).
After finishing the editing of reaction, click the button ‘Next’ (2) if you want to edit the following reaction in which the selected component is involved. For returning to preceding window, click the button 'Back' (3).
If you want to return the window for editing initial values, click the button 'Initial data' (4).
If you want to edit constant(s), click the button 'Constants' (5).
After finishing the editing of initial value(s) and constants of reactions in which the selected component is involved, click the button ‘Apply’ (6) to safe the changes.
Click the button ‘Cancel’ (7) to return back to initial data.



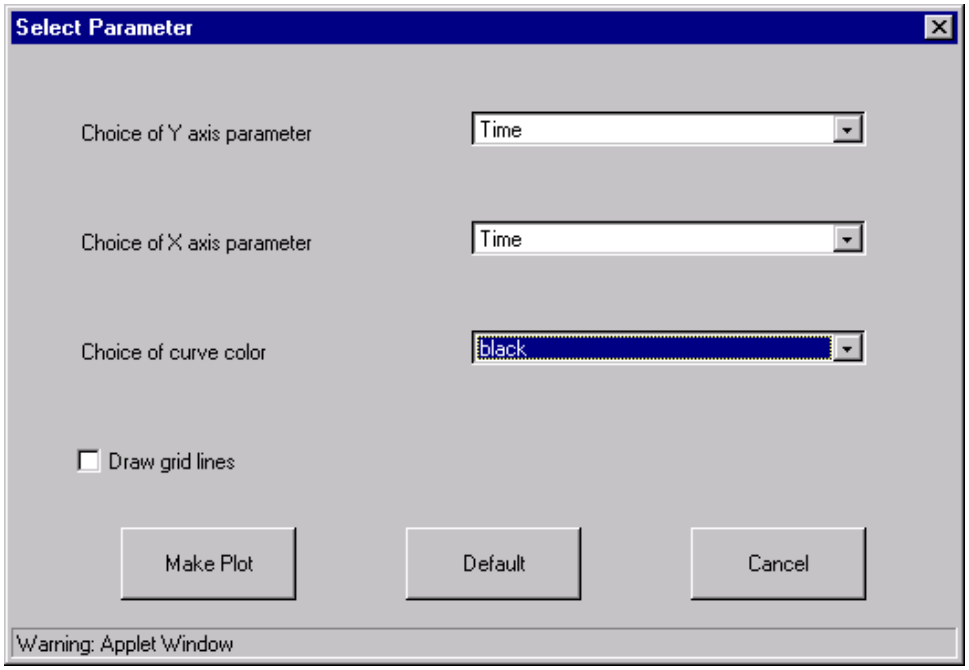
Analogously, user may change the parameters (initial values and reaction constants) of the other components involved in modeling of the gene network.

3. Go to the model window and click from the drop-down menu 'Model Analysis' the view 'Transient states'.

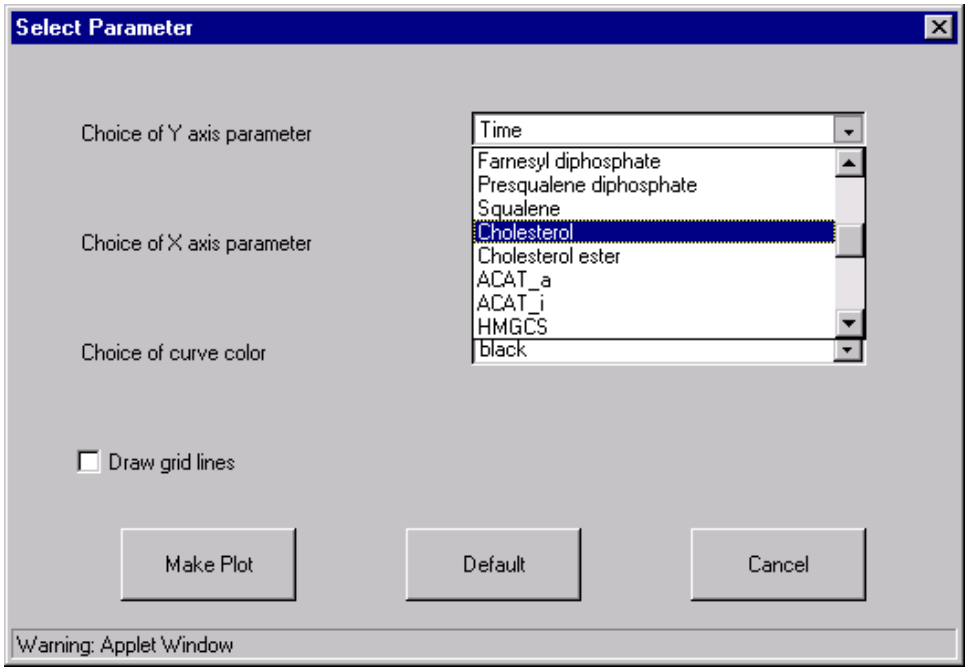
This will bring up the dialog window for editing the observation time and dot pitch of the output data.



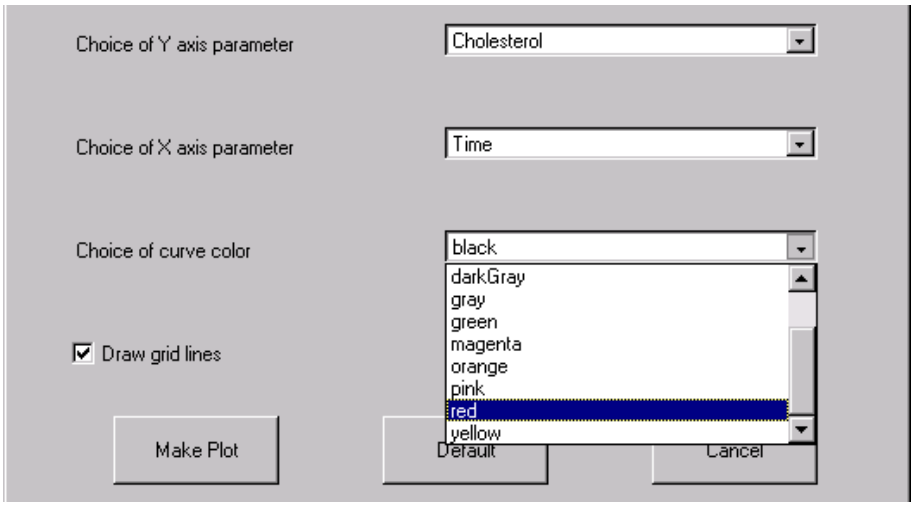
For default settings, click the button 'Default'.
To go back to editing of reaction constants and (or) initial values of components of mathematical model of a gene network, click the button 'Cancel'.
Select the observation parameters and click ‘OK’.
Wait for calculation of the model.
Next, you will get the dialog window for ordering the view of graphical representation data output.



4. Choose the parameters for X and/or Y axis from the drop down menu ‘Choice of Y axis parameter’ and ‘Choice of X axis parameter’.



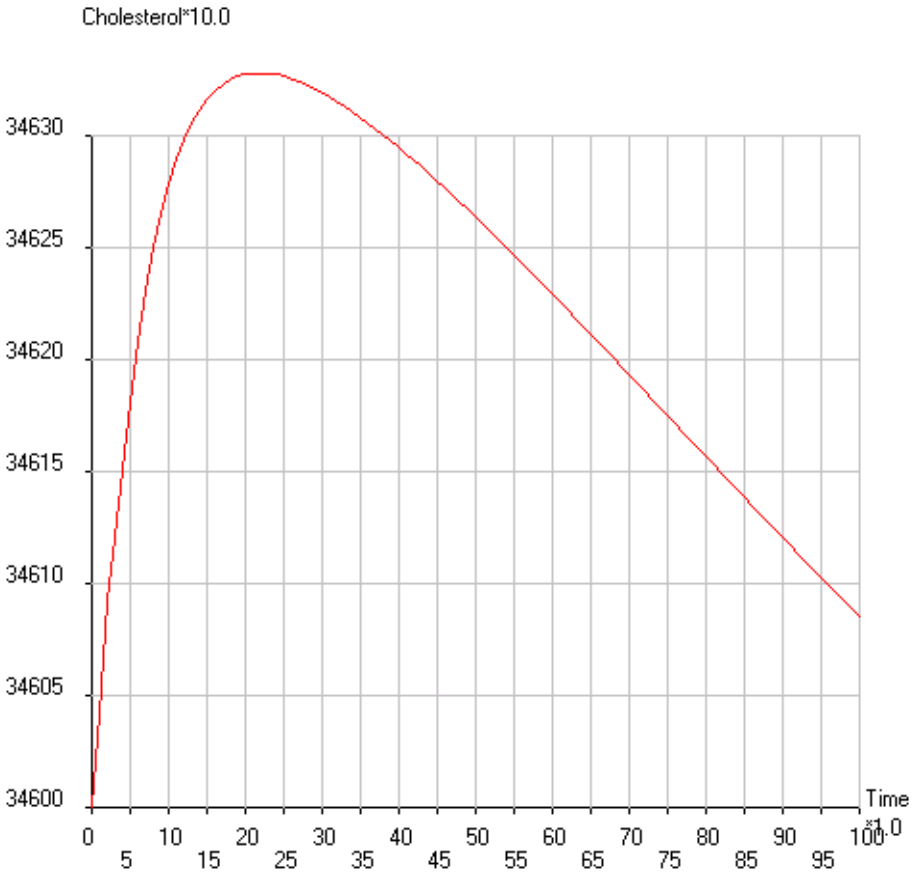
5. Select the colour of the curve at the plot from the drop down menu ‘Curve Color’.



For visualization of grid lines at the plot, tick off the check box ‘Draw grid lines’.

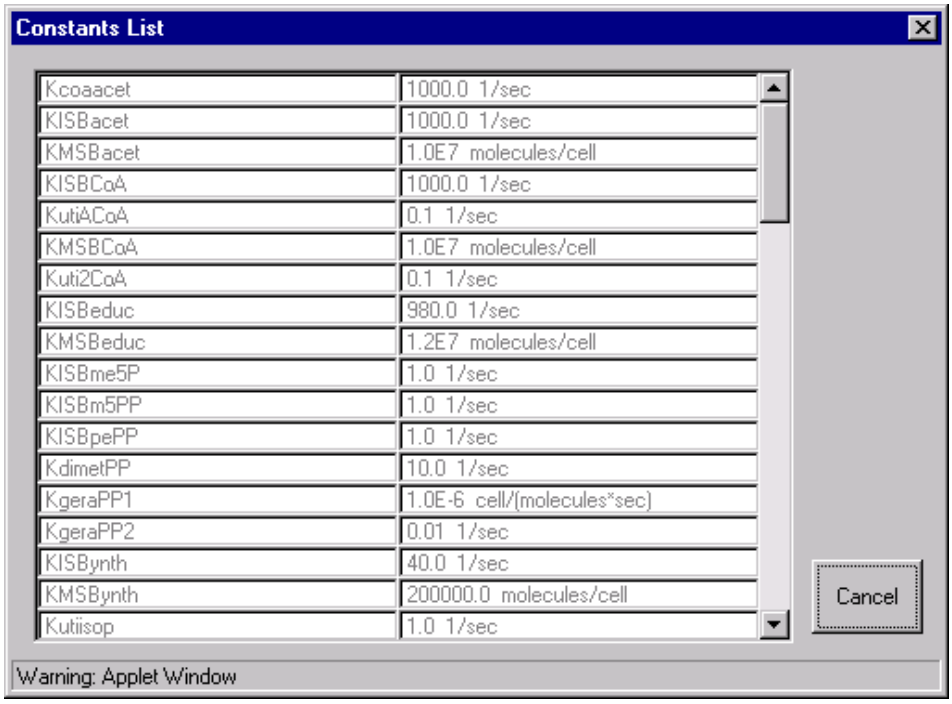
When the button 'Default' is clicked, the first component from the drop down menu ‘Choice of Y axis parameter’ and the option ‘Time’ from the drop down menu ‘Choice of X axis parameter’ will be selected.

6. Click the button 'Make Plot'.
This brings up the plot resulted.



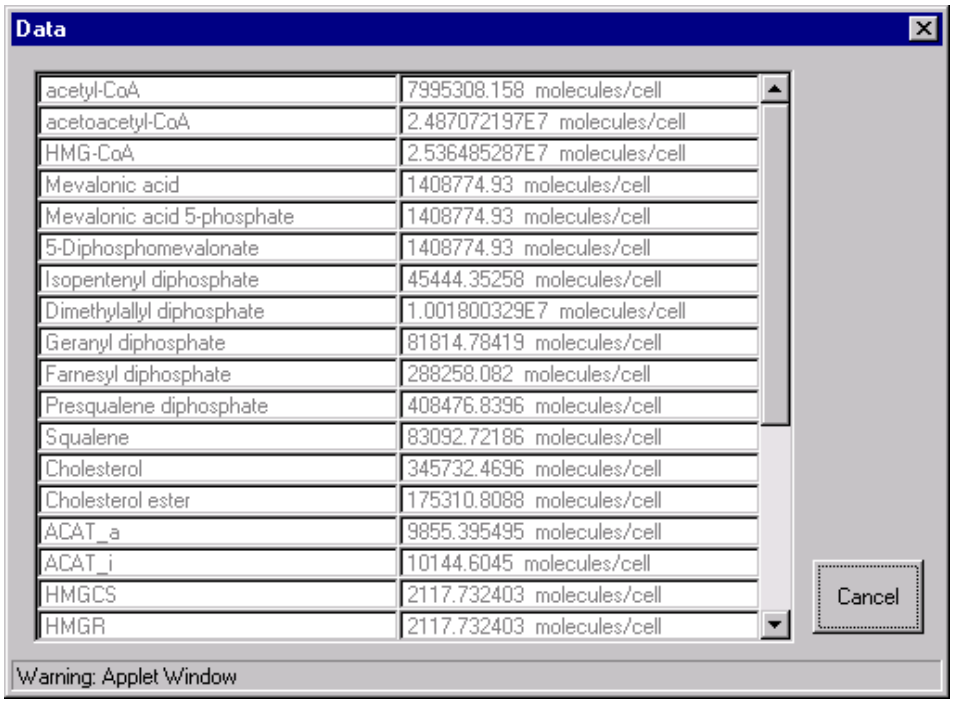
If necessary, you may view the lists of reaction constants and final component concentrations. For this purpose, go to the model window, click the drop-down menu below the ‘Model Analysis’ button, click the view 'Stationary states' and find the view you want, 'Data' or 'Constant'.

If you click the view 'Constant', this will display the following table:



This table lists the names of constants that are involved in mathematical model, their values and units.

If you selected the view 'Data', this will bring up the following table:



This table lists the names of dynamic variables of the gene network, values of concentrations of dynamic variables at the end of calculations and measurement units.

Example 2.

Modelling of a mutation by the example of mutation decreasing cell concentration of LDL receptor mRNA.

For modelling of a mutation, it is necessary to do the following:
1. Select the model of interest. In our case, click the option 'Cholesterol_MODEL' in the drop-down menu 'Select a model', this will bring up the starting page of cholesterol biosynthesis gene network modeling (see Example 1);
1. Load the graphical interface of the model selected;
2. Click the component named LDLRi that denotes the concentration of LDL receptors, which are located inside the cell and have not reached the cell membrane. A part of this LDL pool consists of the native LDL synthesized from mRNA;
3. By clicking the button 'Next', select the following reaction: *LDLR gene* + *SREBP1 homodimer* $\xrightarrow{Ki,\ Km}$ *LDLR* ;
4. For decreasing the LDL synthesis rate, for example, for 30%, input the value '0.7' in the text-box 'KISBldlr';
5. Click the button 'Enter';
6. Calculate the model.

In order to compare the results obtained to the results of the model with initial parameters or some other set of parameters, it is necessary to close the windows 'Select Parameter' and 'Transient states' and input the necessary set of parameters (reaction constant values and initial concentrations of components of a gene network). Then calculate the model again. Make necessary plots and compare them with the plots for the previous set of parameters.
By analogy, mutations described in the Section 'Application of the models: Computer simulation of mutations of the Low Density Lipoprotein (LDL) receptor gene' could be modeled, as well as many other mutations influencing the system of cell cholesterol biosynthesis.

Example 3.

Modeling of the EKLF gene mutation that modifies the functioning of erythrocyte maturation gene network.

This mutation completely arrests the synthesis of the EKLF transcription factor. For simulation of this mutation, do the following:

1. Select the model ‘Erythroid_differentiation’.
2. Upload graphic interface of the model chosen.
3. Click the component named GATA1-p, which determines the inner-cellular concentration of the key transcription factor GATA-1;
4. Choose the reaction *EKLF gene* $\xrightarrow{Ki,\ Km,\ GATA1-p}$ *EKLF mRNA* by clicking the button 'Next';
5. Enter the value of the constant KISBgmre equaling to 0;
6. Click the button 'Enter';
7. Calculate the model.

In order to compare the results obtained to the results of the model with initial parameters or some other set of parameters, it is necessary to close the windows ‘Select Parameter’ and ‘Transient states’ and input the necessary set of parameters (reaction constant values and initial concentrations of components of a gene network). Then calculate the model again. Make necessary plots and compare them with the plots for the previous set of parameters.

3. Mathematical models of gene networks in SBML format

3.1 About mathematical models of gene networks in SBML format

This web page contains 38 mathematical models in SBML format (Level 1 and Level 2) of gene networks represented in *GeneNet* database.
What is the SBML?
SBML - *Systems Biology Markup Language* is a common representation language for storing biochemical models. SBML is based on *XML*, and contains structures for representing compartments, species and reactions, as well as optional unit definitions, parameters and rules (constraints). All information about SBML you can find on *this web page*.
All compartments, species and reactions in the models of gene networks are the same as in *GeneNet* database.
There are some software packages supporting SBML Level 1:
The Systems Biology Workbench (*SBW*) and *SBW-enabled modules* (*Jarnac*, a biochemical simulation package for Windows; *JDesigner*, a visual biochemical network layout tool and *others*).
Gepasi is a software package for modeling biochemical systems. It simulates the kinetics of systems of biochemical reactions and provides a number of tools to fit models to data, optimize any function of the model, perform metabolic control analysis and linear stability analysis.

It is necessary to download the software packages from web sites given above for working with *the mathematical models of gene networks*.

3.2 List of mathematical models in SBML format of gene networks represented in GeneNet database

1.	Adipocyte	SBML Level 1	SBML Level 2
2.	Antiviral response	SBML Level 1	SBML Level 2
3.	Apoptosis (general scheme)	SBML Level 1	SBML Level 2
4.	Cell Cycle (G0/G1-S transition)	SBML Level 1	SBML Level 2
5.	Cholesterol metabolism (intracellular) 1	SBML Level 1	SBML Level 2
6.	Cholesterol metabolism (intracellular) 2	SBML Level 1	SBML Level 2
7.	Cholesterol metabolism (intracellular) 3	SBML Level 1	SBML Level 2
8.	Erythroid differentiation	SBML Level 1	SBML Level 2
9.	Germination (endosperm)	SBML Level 1	SBML Level 2
10.	Heat Shock Response	SBML Level 1	SBML Level 2
11.	HSP70 autoregulation	SBML Level 1	SBML Level 2
12.	LEA program	SBML Level 1	SBML Level 2
13.	Leptin (organism level)	SBML Level 1	SBML Level 2
14.	Lipid metabolism in blood	SBML Level 1	SBML Level 2
15.	Lipid metabolism in liver cells	SBML Level 1	SBML Level 2
16.	Lipid metabolism (integral diagram)	SBML Level 1	SBML Level 2
17.	Macrophage activation (model)	SBML Level 1	SBML Level 2
18.	MAPK cascade	SBML Level 1	SBML Level 2
19.	NF-kappaB activation	SBML Level 1	SBML Level 2

20.	NO biosynthesis pathway	SBML Level 1	SBML Level 2
21.	Nodulation	SBML Level 1	SBML Level 2
22.	Oxidative stress response (glutathione)	SBML Level 1	SBML Level 2
23.	Photomorphogenesis	SBML Level 1	SBML Level 2
24.	Plant-pathogen	SBML Level 1	SBML Level 2
25.	Principal cell of CCD	SBML Level 1	SBML Level 2
26.	REDOX-REGULATION	SBML Level 1	SBML Level 2
27.	Seed reserve mobilisation (1): carbohydrates	SBML Level 1	SBML Level 2
28.	Seed reserve mobilisation (2): lipids and phosphates	SBML Level 1	SBML Level 2
29.	Seed reserve mobilisation (3): proteins	SBML Level 1	SBML Level 2
30.	Seed reserve mobilisation (4): regulatory relationships	SBML Level 1	SBML Level 2
31.	Seed reserve mobilisation (5): the general diagram	SBML Level 1	SBML Level 2
32.	Seed reserve mobilisation (organism level)	SBML Level 1	SBML Level 2
33.	Steroidogenesis (adrenal cortex)	SBML Level 1	SBML Level 2
34.	Steroidogenesis (sex steroids)	SBML Level 1	SBML Level 2
35.	Storage protein biosynthesis (dicot)	SBML Level 1	SBML Level 2
36.	Storage protein biosynthesis (monocot)	SBML Level 1	SBML Level 2
37.	Thermotolerance	SBML Level 1	SBML Level 2
38.	Thyroid system	SBML Level 1	SBML Level 2

Comments and questions are welcome to Alexander V. Ratushny (ratushny@bionet.nsc.ru).