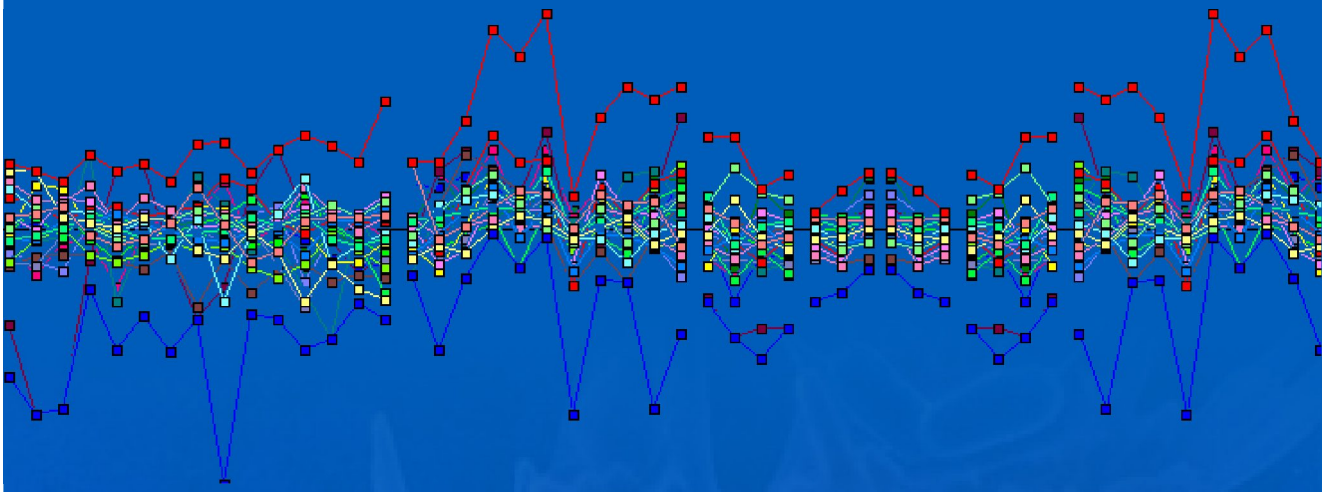
SoftBerry

www.softberry.com

SOFTWARE PRODUCTS FOR GENOMIC AND PROTEOMIC RESEARCH



EUKARYOTIC GENOME ANNOTATION PIPELINE

BACTERIAL GENOME ANNOTATION AND OPERON PREDICTION

GENOME MAPPING AND COMPARISON

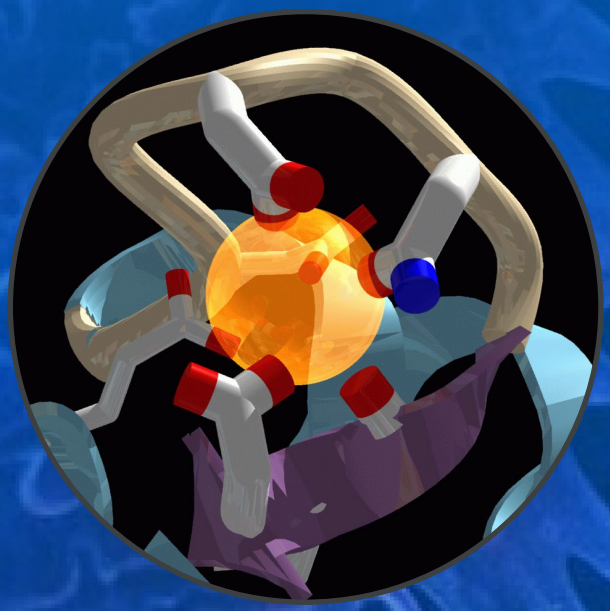
PROTEIN SUB-CELLULAR LOCALIZATION

PROMOTER PREDICTION

EST CLUSTERING

GENOME AND PROTEIN 3D VIEWERS

EXPRESSION DATA ANALYSIS



2003

Content

FGENESH++C: FULLY AUTOMATIC EUKARYOTIC GENOME ANNOTATION PIPELINE *Page 1*

FGENESB: BACTERIAL GENOME ANNOTATION PIPELINE *Page 2*

FGENESV: VIRAL GENOME ANNOTATION PIPELINE *Page 3*

EUKARYOTIC GENE, PROMOTER AND FUNCTIONAL SITE PREDICTORS *Page 4*

PROTCOMP: THE PROGRAM FOR PREDICTING PROTEIN SUBCELLULAR LOCALIZATION *Page 6*

GENOME COMPARISON AND MAPPING PROGRAMS *Page 8*

GENOME EXPLORER: POWERFUL TOOL FOR INTEGRATING GENOMIC INFORMATION WITH EXPRESSION DATA *Page 10*

3-D VISUAL WORKS: PROTEIN/DNA 3D VIEWER *Page 10*

SELTAG: TOOL FOR ANALYSIS OF EXPRESSION DATA *Page 11*

PROTEIN STRUCTURE ANALYSIS PROGRAMS *Page 11*

CORPORATE PROFILE *Page 13*

FGENESH++C: FULLY AUTOMATIC EUKARYOTIC GENOME ANNOTATION PIPELINE

Based on fastest and most accurate *ab initio* gene prediction program, FGENESH (see page 4), Softberry fully automatic genome annotation pipeline, FGENESH++C, is the best available.

It involves the following steps:

1. RefSeq mRNA mapping by EST_MAP program - mapped genes are excluded from further gene prediction process.
2. *Ab initio* FGENESH gene prediction .
3. Search of all products of predicted genes through NR database for protein homologs.
4. Fgenesh+ gene prediction on sequences with found protein homology.
5. Second run of *ab initio* gene prediction in regions free from predictions made on stages 1 and 4.
6. Run of FGENESH gene predictions in large introns of known and predicted genes.

Special variants of FGENESH++C can take into account synteny - for example, human-mouse, and direct protein-to-DNA mapping for improved gene finding.

Examples of using FGENESH++C and its individual elements:

Yu *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296:79-92. As part of rice genome sequencing project, the team led by Beijing Genomics Institute has compared several well-known *ab initio* gene prediction programs and shown that FGENESH is by far the most accurate (see Fig.1 on page 4). As a result, their rice genome annotation was based almost exclusively on FGENESH results.

Goff *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296:92-100 and supplement. Second rice genome sequencing and annotation project also used FGENESH as primary source of gene predictions.

Galagan *et al.* (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422:859-868. Neurospora genome annotation based on FGENESH and FGENESH+.

http://genome.ucsc.edu: FGENESH++C gene predictions at Santa Cruz human, mouse and rat genome assembly, and FGENESV prediction of SARS genome. When in browser window, unhide "Fgenesh++ genes" or "FGENESV+" genes to see predictions.

FGENESH and, by extension, FGENESH++C, use taxon-specific gene finding parameters for improved *ab initio* gene prediction: we can currently supply them for human, mouse, *Drosophila*, *C.elegans, S.pombe*, *Neurospora*, *Anopheles, Plasmodium, Arabidopsis*, tobacco and monocot plants. Gene prediction accuracy on new genomes can be improved by creating custom-trained parameters. In the past eighteen months, we trained eight new sets by order of several customers: see http://www-genome.wi.mit.edu/ annotation/fungi/magnaporthe/ gene_finding.html for an example from MIT/Whitehead Institute.

FGENESB: BACTERIAL GENOME ANNOTATION PIPELINE

Softberry FGENESB bacterial **gene and operon** prediction pipeline includes the following features:

- Automatic training of gene finding parameters for new bacterial genomes using only genomic DNA as an input.
- Highly accurate HMM-based gene prediction.
- Operon prediction that combines several approaches including promoters and terminator identification.
- Automatic annotation of predicted genes by homology with COG and NR databases.

FGENESB gene prediction engine is one of the most accurate prokaryotic gene finders available: see Table 1 for its comparison with two other popular gene prediction programs.

Table 1 Comparison of three popular bacterial gene finders. Accuracy estimate was done on a set of difficult short genes that was previously used for evaluating other bacterial gene finders (http://opal.biology.gatech.edu/GeneMark/genemarks.cgi). First set (51set) has 51 genes with at least 10 strong similarities to known proteins. Then 72set has 72 genes with at least two strong similarities, and 123set has 123 genes with at least one protein homolog.

Here are the prediction results on these three sets for GeneMarkS and Glimmer (calculated by Besemer *et al.* (2001) *Nucl. Acids Res.* 29:2607-2618) and FGENESB (calculated by Softberry, three iterations of FGENESB-Train script).

	Sn (exact predictions)	Sn (exact+overlapping predictions)
123set:		
Glimmer	57.0%	91.1
GeneMarkS	82.9	91.9
FgenesB	89.3	98.4
72set:		
Glimmer	57.0%	91.7
GeneMarkS	88.9	94.4
FgenesB	91.5	98.6
51set:		
Glimmer	51.0%	88.2
GeneMarkS	90.2	94.1
FgenesB	92.0	98.0

Example of FGENESB output - the beginning of complete annotation of *E.coli* genome:

```
Prediction of potential genes in microbial  genomes
Time:   Wed Sep 18 15:15:22 2002
Seq name: gi|16127994|ref|NC_000913.1| Escherichia coli K12, complete genome
Length of sequence - 4639221 bp
Number of predicted genes - 4479, with homology - 4247
Number of transcription units - 2440, operons - 927
N   Tu/Op   Conserved  S        Start    End    Score
            pairs(N/Pv)
                       +  Prom     54 -  113    4.1
1  1 Op  1     .       +  CDS     190 -  255     99
                       +  Term    280 -  315    5.9
2  1 Op  2   2/0.147   +  CDS     337 - 2799   2448  ## COG0527 Aspartokinases
3  1 Op  3  12/0.000   +  CDS    2801 - 3733    775  ## COG0083 Homoserine kinase
4  1 Op  4     .       +  CDS    3734 - 5020   1480  ## COG0498 Threonine synthase
                       +  Term   5044 - 5084    4.4
                       +  Prom   5029 - 5088    2.9
5  2 Tu  1     .       +  CDS    5234 - 5530    170  ## orf, hypothetical protein
[Escherichia coli K12]^Agi|6686173|
                       -  Term  5520 - 5564    2.2
6  3 Op  1   2/0.147   -  CDS   5683 - 6459    885  ## COG3022 Uncharacterized BCR
7  3 Op  2     .       -  CDS   6529 - 7959   1032  ## COG1115 Na+/alanine symporter
                       -  Prom   8016 - 8075    4.3
                       +  Prom   8132 - 8191    1.9
8  4 Tu  1   1/0.517   +  CDS    8238 - 9191   1318  ## COG0176 Transaldolase
                      +  Term   9199 - 9240    9.5
```

FGENESV: VIRAL GENOME ANNOTATION PIPELINE

FGENESV annotation pipeline is based on fastest and most accurate viral gene prediction program. Generic-parameters version, FGENESV0, is suited for small (<10 kb) genomes, while custom-parameters version trains itself by using larger viral genome sequence as an input. FGENESV supports alternative genetic codes, which are not uncommon among viruses. Annotation part of the pipeline, **FGENESV-Annotator** script, annotates found genes by homology with COG and NR databases. This script can also identify low scoring genes if they have known homologous protein.

Example of FGENESV output annotation of SARS coronavirus genome is given below.

```
Softberry Fgenesv-annotator Prediction of potential genes in viral  genomes
 Time:   Mon Apr 14 23:26:46 2003
 Seq name: gi|29826276|gb|AY274119.1| SARS coronavirus TOR2, complete genome
 Length of sequence - 29736 bp
 Number of predicted genes - 11, with homology - 5
     N   S             Start         End    Score

     1   +    CDS        250 -     13398    10746  ## polyprotein [Bovine coronavirus]
     2   +    CDS      13584 -     21470     6365  ## polyprotein [Murine hepatitis virus]
     3   +    CDS      21477 -     25244     3084  ##
     4   +    CDS      25253 -     26077      571
     5   +    CDS      26102 -     26332      206
     6   +    CDS      26383 -     27048      365  ## glycoprotein (Matrix glycoprotein)
(Membrane glycoprotein)
     7   +    CDS      27059 -     27250      158
     8   +    CDS      27258 -     27626      277
     9   +    CDS      27764 -     27883      104
    10   +    CDS      27849 -     28103      215
    11   +    CDS      28105 -     29373      850  ## protein [Murine hepatitis virus]
```

EUKARYOTIC GENE, PROMOTER AND FUNCTIONAL SITE PREDICTORS

FGENESH. One of the most popular and accurate *ab initio* gene finders available (see Figure1 and Table 2), FGENESH is an engine behind our FGENESH++C annotation pipeline. It is also the fastest genegene prediction program - 50 to 100 times faster than Genscan. FGENESH can be supplied with data sets specifically trained for several taxonomic groups - see page 1 for their list. Custom parameter sets for specific organism or group can also be created. Recently released ver. 3 of FGENESH supports non-standard GC-donor splice sites and has various output options.

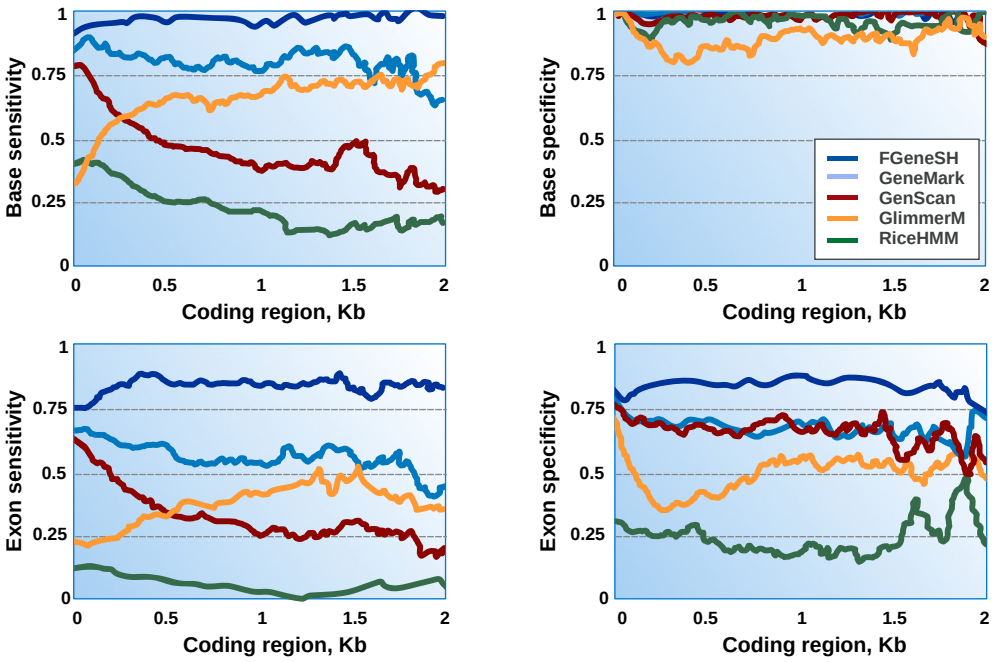


Figure 1. Performance of different gene prediction programs on rice genes as a function of gene position. Reproduced from Yu *et al.* (2002) *Science* 296:79-92.

Table 2. Comparison of three popular gene prediction programs on 42 semiartificial genomic sequences containing 178 known human gene sequences (900 exons). DIGIT is a program that combines outputs of these three gene finders. Sensitivity is percentage of known exons that were predicted correctly. Selectivity is percentage of predicted exons that are correct. Reproduced with changes from Yada *et al.*, 2002 Cold Spring Harbor Genome Sequencing and Biology Meeting, May 7-11, 2002.

Program	Sensitivity	Specificity	Missed Exons, %	Wrong Exons, %
FGENESH	**77.1**	**65.7**	**9.6**	**23.2**
GenScan	**66.5**	**44.9**	**12.0**	**40.9**
HMMGene	**69.6**	**36.6**	**15.5**	**55.5**
DIGIT	**78.6**	**77.7**	**14.0**	**15.1**

FGENESH+. This derivative of FGENESH uses information on homologous proteins for dramatically improved gene prediction, if such homologies can be found (see Table 3).

Table 3. Comparison of FGENESH and FGENESH+ on specially selected set of "difficult" genes. For each of these genes, *ab initio* gene prediction by FGENESH produced at least one missing or wrongly predicted exon. The set included 61 human genes (370 exons), all with known protein homologs from other organisms.

Program	Correctly predicted genes	Sensitivity, exons	Specificity, exons	Sensitivity, nucleotides	Specificity, nucleotides
FGENESH	0 (0%)	0.63	0.68	0.86	0.83
FGENESH+	46 (75%)	0.82	0.85	0.96	0.98

FGENESH_C uses information on homologous mRNA/EST to improve accuracy of gene prediction.

FGENESH-2. Variant of FGENESH that uses homology between two genomic DNA sequences, such as human and mouse, as an extra factor for more accurate gene prediction.

FGENES. This pioneering gene recognition program uses discriminant analysis algorithm, which sets it apart from most other popular gene finders, usually utilizing Hidden Markov Model based algorithms. That makes FGENES well suitable as a supplemental gene-predictor for exhaustive annotation tasks. The ***FGENES-M*** variant can predict multiple splice forms of same gene.

Promoter Prediction Programs TSSW, TSSG and PromH. Two of the most accurate eukaryotic PolII promoter prediction programs, ***TSSW*** and ***TSSG***, are based on discriminant analysis combining characteristics of functional elements of regulatory sequence with the database of regulatory motifs. TSSG is the most accurate stand-alone promoter prediction program available: it correctly predicts 50-60% of promoters, and 80-85% of promoters predicted by TSSG are true promoters. Accuracy of TSSW is only slightly lower (see Table 4). ***PromH*** is an enhancement of TSSW that adds information on syntenic regions of mouse and human genomes into prediction algorithm. That results in additional 20% accuracy improvement, especially pronounced on TATA+ promoters. TSSW and PromH programs contain elements of older version of Transfac database (www.biobase.de), and their use may require Transfac license.

TSSP plant PolII promoter recognition program. TSSP is plant variant of TSSG promoter prediction program, trained on PlantProm database developed by Softberry in collaboration with Royal Holloway, University of London, and utilizing our RegSite database for motif search.

NSITE can be used for search for consensus patterns of functional motifs with statistical estimation of their significance. The program contains elements of Transfac database and a set of additional functional motifs, and may require Transfac license for its use.

NSITE-PL is plant variant of NSITE that utilizes our RegSite database instead of Transfac.

Short descriptions some other gene and functional site finders are given below:

Program	Short Description
FEX	Prediction of internal exons on genomic sequence
BESTORF	Prediction of potential coding fragments on mRNA/EST sequence
PolyaH	Prediction of polyA sites
SPL and SPLM	Two splice site finders that use different statistical approaches
NSITEM	Search for regulatory motif conserved in a set of sequences
BPROM	Prediction of bacterial σ^{70} promoters

Table 4. Results of promoter search on genes with known mRNAs by different promoter-finding programs. Reproduced with changes from Liu and States (2002) *Genome Research* 12:462-469. TSSG is close second in percentage of promoters predicted and has by far the fewest false positive predictions.

Program	Set 1 (133 promoters)		Set 2 (120 promoters)	
	True Predictions	False Predictions	True Predictions	False Predictions
PROSCAN1.7	32 (24%)	18 (36%)	30 (25%)	22 (42%)
NNPP2.0	56 (42%)	41 (42%)	26 (22%)	50 (66%)
PromFD1.0	88 (66%)	43 (33%)	69 (58%)	57 (45%)
Promoter2.0	8 (6%)	100 (93%)	14 (12%)	92 (88%)
TSSG	75 (56%)	10 (12%)	62 (52%)	18 (23%)
TSSW	57 (43%)	29 (34%)	58 (48%)	20 (26%)

PROTCOMP: THE PROGRAM FOR PREDICTING PROTEIN SUBCELLULAR LOCALIZATION

ProtComp combines several methods of protein localization prediction: neural networks-based prediction, direct comparison with homologous proteins of known localization; prediction of certain functional peptide sequences, such as signal peptide, signal-anchor, transit peptides of mitochondria and chloroplasts and transmembrane segments; and search for certain localization-specific motifs. The program includes separately trained recognizers for animal/fungal and plant proteins, dramatically improving recognition accuracy, which for major compartments, such as plasma membrane, nuclear and extracellular, achieves 80-90% level. Current, fifth version of ProtComp was retrained on much larger sample of proteins with known localization, resulting in dramatical improvement of accuracy (see Table 5.)The logical scheme of ProtComp algorithm is given on Figure 2.

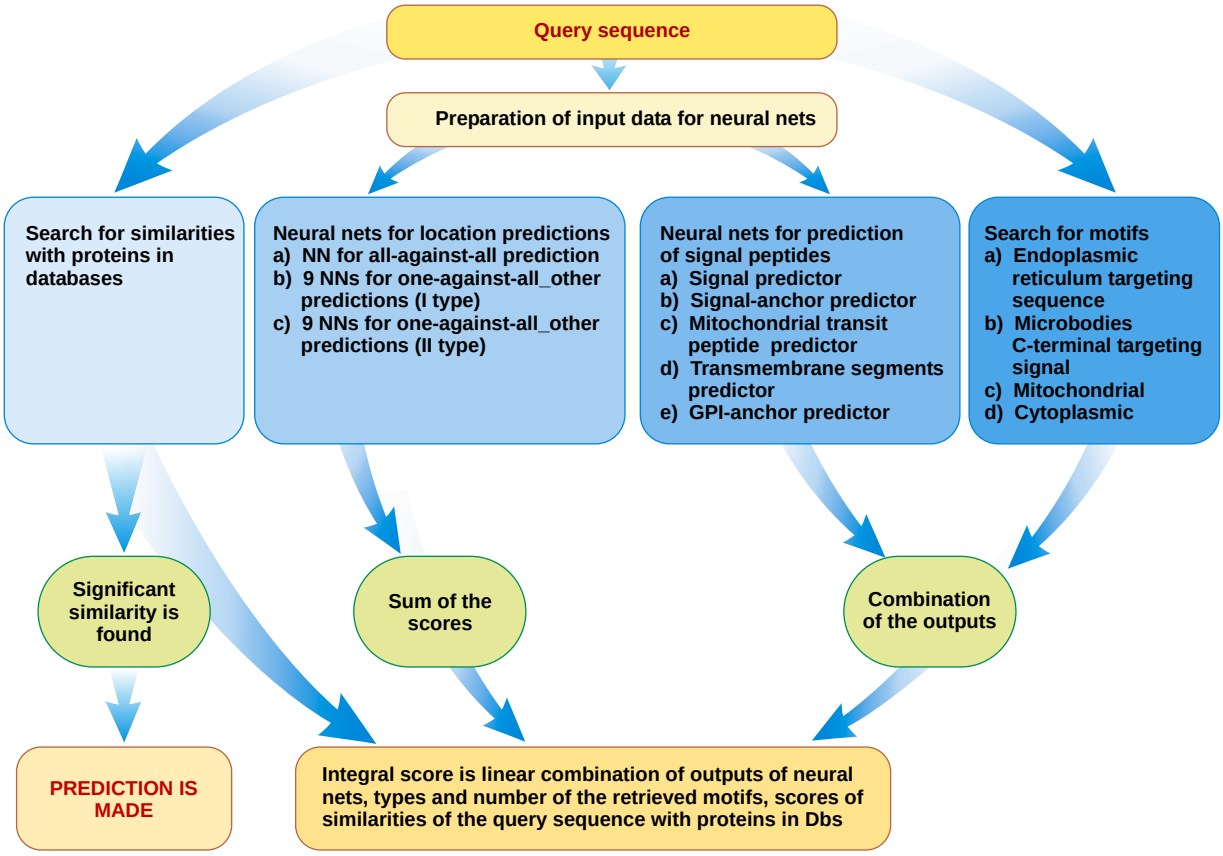


Figure 2. Logical scheme of protein subcellular localization prediction by ProtComp.

Table 5. Results of testing ProtComp on a set of 977 plant, fungal and animal proteins with known localization not included in training set.

Compartment	Sample Size	Percent predicted correctly ver. 4	ver. 5
Nucleus	200	80	88
Plasma Membrane	200	80	87
Extracellular	162	69	83
Cytoplasm	200	46	63
Mitochondria	85	76	82
Endoplasmic Reticulum	46	67	83
Peroxisome	35	95	97
Lysosome	23	69	91
Golgi	26	57	77

GENOME COMPARISON AND MAPPING PROGRAMS

EST_MAP: a program for fast mapping of a set of mRNAs/ESTs to a chromosome sequence. Using EST_MAP, we mapped a set of 11,000 sequences of full mRNAs from NCBI reference to 52-MB unmasked Y chromosome fragment in about 20 minutes on one 500-MHz Alpha processor. EST_MAP is a part of FGENESH++C genome annotation pipeline, where it maps RefSeq sequences to a query genome at very early stages of annotation. EST_MAP takes into account statistical features of splice sites for more accurate mapping. Its variant **EST_MAP_P** is used for mapping a set of protein sequences to genome with accounting for splice sites.

FMAP: a program for instant sequence-to-genome mapping. It takes FMAP only one to two seconds to map a sequence to human genome, which allows users to work on genome mapping projects interactively. It can be integrated with Softberry Genome Explorer for grafical presentation of results. Used in conjunction with DBScan, FMAP provides very accurate sequence alignment, and with EST_MAP accounts for splice sites for more accurate mRNA/EST mapping.

DBScan is a program for sequence database search and alignment. It now includes functions of Scan2, which was discontinued as a stand-alone program. DBScan is much faster than BLAST, especially on multimegabyte sequences. Its DBScanP varaint is used for high-speed protein database search, for instance, in high-throughput installations of FGENESH++C pipeline.
DBScan is also a very sensitive sequence alignment program, which makes it useful for conserved motif search to discover new regulatory elements: see Fig. 3 as an example.

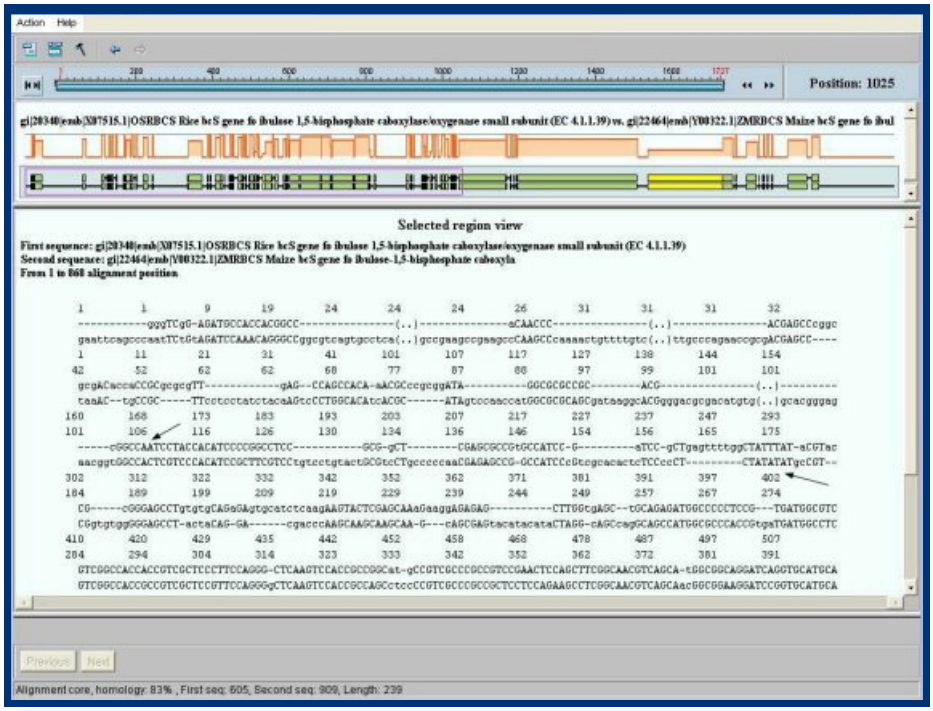


Figure 3. DBScan alignment of 5' regions of rice and maize ribulose bisphosphate carboxylase (*rbcS*) genes. Very small known conserved motifs, such as CAAT and TATA boxes (pointed to by black arrows), are properly aligned. We can see other even more conserved sequences in promoter region which may represent regulatory motifs.

Human-Mouse-Rat Genome Synteny is, so far, the most comprehensive collection of information about homology between human, mouse and rat genomic regions. Compared to NCBI homology maps, Softberry map contains significantly more genes and is directly linked to genomic sequences. The data were generated by Softberry programs for gene prediction, EST/RNA mapping and gemomic sequence comparison. Sequences from human (November 2002), February 2002 mouse (February 2002) and rat (November 2002) genome drafts were used to create sets of about 4,000 aligned syntenic regions for each pair, containing ~19,000 human genes mapped to mouse genome, ~17,000 human genes mapped to rat genome, and ~19,000 mouse genes mapped to rat genome, and *vice versa*.

FCLUST: FAST EST Clustering and Alignment Program. FCLUST options include accounting for known mRNAs, repeat masking and many others. Example of consensus EST sequences and their alignments in text and graphical form, as produced by FCLUST from millions of GenBANK mouse ESTs, are available at http://www.softberry.com/berry .phtml?topic=clust. Example of cluster analysis results in graphic form is shown at Fig 4.

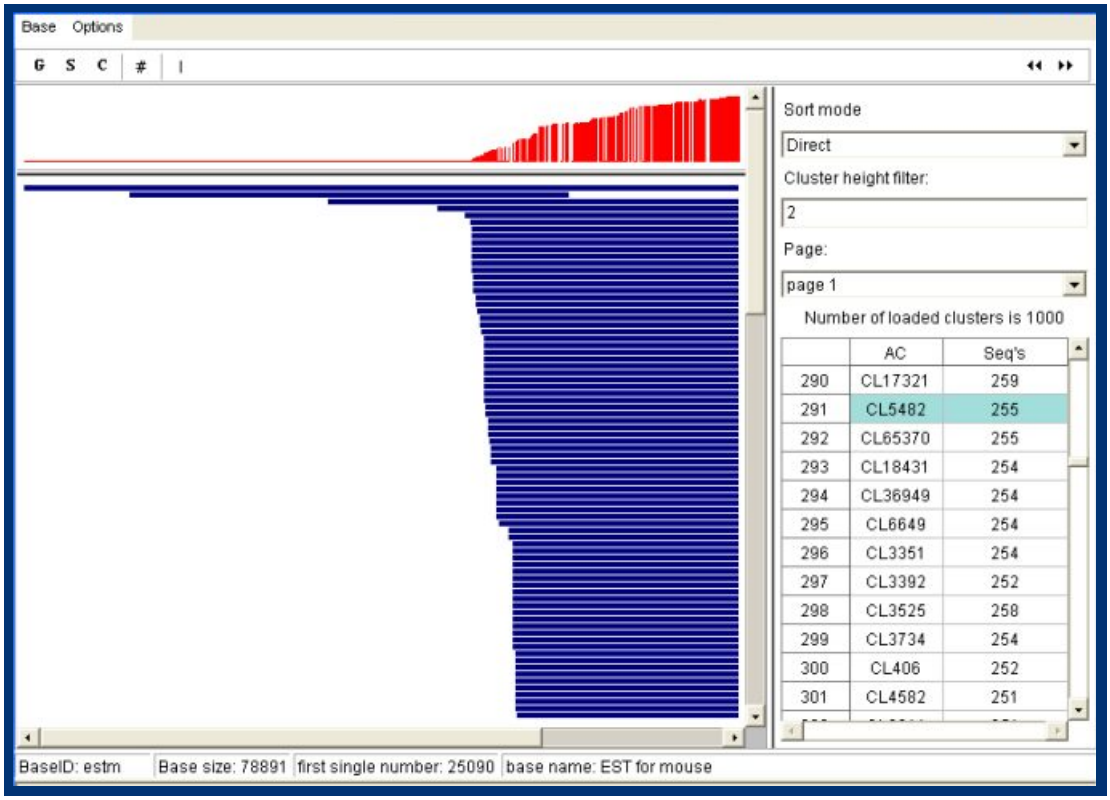


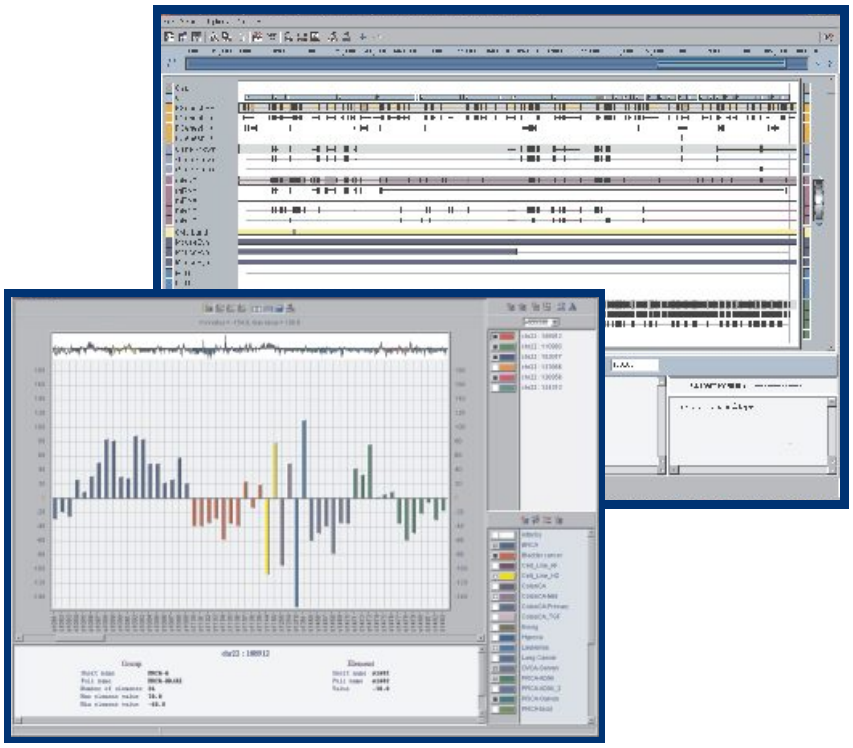
Figure 4. Example of cluster analysis by FCLUST.

GENOME EXPLORER: POWERFUL TOOL FOR INTEGRATING GENOMIC INFORMATION WITH EXPRESSION DATA

Softberry Genome Explorer is a powerful viewer of genomic sequences with capability for sequence homology search (one second on human genome using ***FMAP*** algorithm), simple pattern, feature ID or word search, retrieval of nucleotide and amino acid sequences of features, expression data on individual genes and many others.Genome Explorer is easily customizable to include wide variuety of data from outside sources. Annotation of public version of human genome draft can be viewed in Genome Explorer at www.softberry.com/berry.phtml?topic=chrvis. It includes known and predicted (FGENESH++C) genes, mapped ESTs and mRNAs and many other features (see Fig. 5).

Figure 5.

Softberry Genome Explorer: Graphical presentation of fragment of human chromosome 1 (back window) and of expression levels of one of genes located on that fragment in different tissues (front window).



3-D VISUAL WORKS: PROTEIN/DNA 3D VIEWER

3-D Visual Works is a viewer for protein and DNA 3D structures (See Figure 6). It allows easy mapping between structure and sequence information, so that particular aminoacid residues can easily be located on a 3D structure. The viewer can generate about ten different model types such as ribbon/solid ribbon, ball and sticks, CPK etc., and have endless ways of manipulating with atoms/residue/chains colors and methods of visualization. 3-D Visual Works also allows structure-to-structure alignment and comparison of two molecules. It can be run as a Java applet through web site or as standalone Java-based application. C++ version is currently under development.

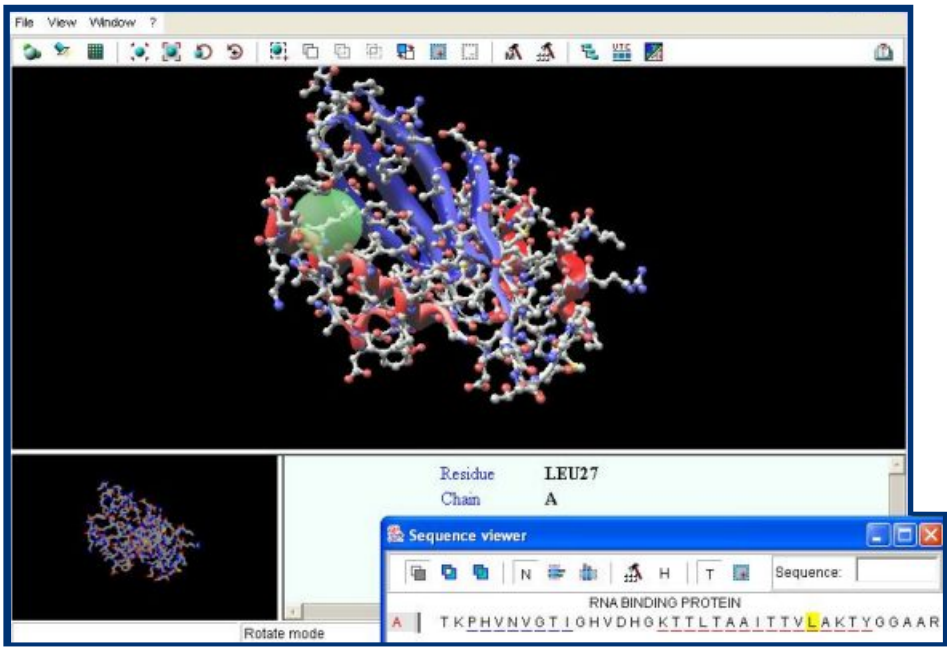


Figure 6. First 120 residues of translation elongation factor from E.coli as rendered by Softberry 3-D viewer.

SELTAG: TOOL FOR ANALYSIS OF EXPRESSION DATA

SelTag is one of the most elegant tools for analysis of expression data (see Fig. 7). It can analyze all or marked groups of genes or tissues, select tissue-specific genes based on complex criteria, provide visual representation of expression data, identify genes correlatively expressed in a given set of tissues, select disease-specific genes with particular characteristics, such as receptors or secreted proteins. SelTag is available in Windows and Java-based variants.

PROTEIN STRUCTURE ANALYSIS PROGRAMS

SSENVID: Protein secondary structure and environment assignment from atomic coordinates. SSENVID recognizes secondary structural elements in proteins from their atomic coordinates. It performs same tasks as DSSP by Kabsch and Sander (1983) or STRIDE by Frishman & Argos (1995), analyzing both hydrogen bonds and mainchain dihedral angles, as well as some probabilistic measures. SSENVID also computes accessible surface area, polarity and environment classes as defined by Bowie, Luthy, Eisenberg (1991). The program's new feature is prediction of probability (quality) of secondary structure assignment for each amino acid. SSENVID computes 3D protein characteristics which can be used in structure prediction by measuring the compatibility between protein sequences and known protein structures.

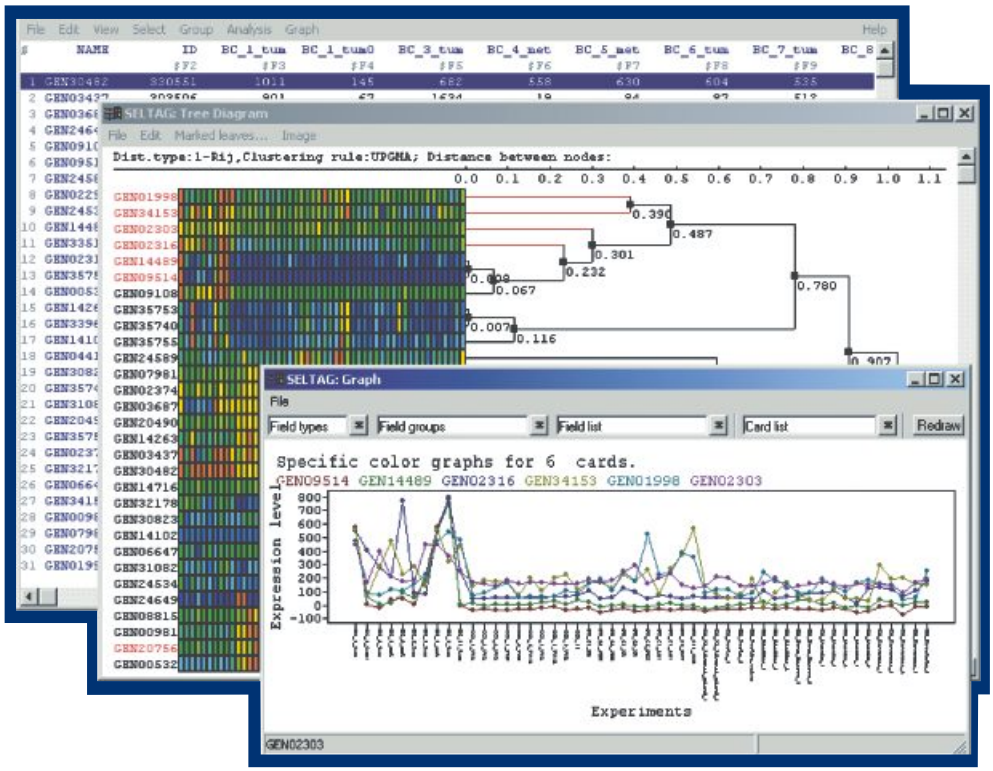


Figure 7. Tissue-specific expression of selected genes as presented In SelTag window.

GETATOMS is a program for modeling atomic coordinates of a protein with unknown 3D structure. It uses main chain coordinates from 3D structure of similar protein, which sequence is aligned with a query protein. Loops in alignment are restored using homology with known protein structures and subsequent optimization. The program computes 3D protein coordinates of a query protein and estimates quality of produced 3D structure using Steric_Score similar to described in JMB (1997), 267, 1268-1282, VDW_Score similar to JMB(1981) v.153,p.1087-1109 and Bump Score - a number of atomic pairs that have sterically forbidden overlap. Resulting 3D structure can be visualized using 3D viewers such as RasMol or Softberry 3D Visual Works viewer.

SSPAL is a program for prediction of protein secondary structure by using local alignments. Overall 3-state prediction gives about 75% correctly predicted residues.

PSITE searches for PROSITE patterns with statistical estimation. The program method is based on statistical estimation of expected number of prosite pattern in a given sequence. It uses PROSITE database of functional motifs (author: Amos Bairoch,1995).

CORPORATE PROFILE

Softberry, Inc. is a leading developer of software tools for genomic and proteomic research, with primary areas of expertise and interest in the following areas:

Genome annotation
Functional site identification in DNA and proteins
Sequence database managing
Genome comparison
Expression data analysis
Protein structure prediction
Protein compartment (destination) prediction.

List of licensees of Softberry programs includes over forty leading pharmaceutical, biotech and agricultural companies, as well as scientific institutions, from around the world. Some of them are listed here.

Amgen
AstraZeneca
Baylor College of Medicine
Bristol-Myers Squibb
Celera Genomics
Cold Spring Harbor Laboratory
Genentech
Hitachi
Human Genome Sciences
Incyte Pharmaceuticals
Lexicon Genetics
Millennium Pharmaceuticals
MIT (Whitehead Institute)
Monsanto
Pioneer Hi-Bred International
Protein Design Labs
RIKEN Genomic Sciences Center
Roche Bioscience
Schering AG
SmithKline Beecham
Syngenta
The Institute for Genomic Research (TIGR)
Washington University Genomic Center
Yamanouchi Pharmaceutical Co.

In addition to licensing our own software, Softberry provides services on custom genome annotation and training gene finding parameters, custom software development and compiling software packages for bioinformatics departments.

Softberry is involved in collaborative relationships with Lion Bioscience, Biomax Informatics and CTC Laboratory Systems, which distribute our programs and data as parts of their own products or as stand-alone modules.

Some of our products are briefly described in this leaflet. Information on many more can be found on our web site, where many of these products can also be tried.

SoftBerry

Softberry, Inc. - www.softberry.com
116 Radio Circle, Suite 400, Mount Kisco, NY 10549, USA
Tel (914) 242-3592, Fax (914) 242-3593
softberry@softberry.com