 Nucleic Acids Research, 2002, Vol. 30, No. 1 299–300

SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein

Antje Krause*, Stefan A. Haas, Eivind Coward and Martin Vingron

Max-Planck-Institute for Molecular Genetics, Computational Molecular Biology, Ihnestrasse 73, 14195 Berlin, Germany

Received September 21, 2001; Accepted October 1, 2001

ABSTRACT

We have integrated the protein families from SYSTERS and the expressed sequence tag (EST) clusters from our database GeneNest with SpliceNest, a new database mapping EST contigs into genomic DNA. The SYSTERS protein sequence cluster set provides an automatically generated classification of all sequences of the SWISS-PROT, TrEMBL and PIR databases into disjoint protein family and superfamily clusters. GeneNest is a database and software package for producing and visualizing gene indices from ESTs and mRNAs. Currently, the database comprises gene indices of human, mouse, *Arabidopsis thaliana* and zebrafish. SpliceNest is a web-based graphical tool to explore gene structure, including alternative splicing, based on a mapping of the EST consensus sequences from GeneNest to the complete human genome. The integration of SYSTERS, GeneNest and SpliceNest into one framework now permits an overall exploration of the whole sequence space covering protein, mRNA and EST sequences, as well as genomic DNA. The databases are available for querying and browsing at http://cmb.molgen.mpg.de.

INTEGRATED DATABASES

SYSTERS

The SYSTERS protein sequence cluster set (1) consists of the hierarchical classification of all known sequences from the SWISS-PROT (2), TrEMBL and PIR (3) sequence databases into disjoint protein family clusters and superfamilies. The classification is based on an all-against-all database search using gapped BLAST (4) with a subsequent hierarchical clustering. The sequences in every cluster have been multiply aligned using CLUSTALW (5) and for each cluster an unrooted phylogenetic tree is available. All multiple alignments are annotated with known domains from the Pfam database of protein domain families (6) and clusters can be selected directly from a list of Pfam domains. A new protein sequence can be searched against the database of multiple alignments using the similarity searching tool SSMAL (7). For each cluster, an MView (8) output is generated and from the resulting partial multiple alignment a majority consensus sequence is calculated. All consensus sequences together build a database searchable with BLAST. Precomputed BLAST searches of the GeneNest consensus sequences against the SYSTERS protein consensus sequences were evaluated to generate links from SYSTERS to GeneNest and vice versa.

GeneNest

GeneNest (9) is a database and software package for the generation and visualization of gene indices based on EST and mRNA sequences. Currently, the database comprises gene indices of man (based on UniGene), mouse, *Arabidopsis thaliana* and zebrafish. All cDNA/mRNA sequences related to an organism are extracted either directly from the EMBL (10) database or from an already clustered UniGene (11) database. A preprocessing step includes vector clipping, repeat annotation and marking of regions of low sequence quality in order to restrict processing to data of high quality. In further steps, these sequences are clustered and all members of each cluster are assembled into one or more contigs. Roughly speaking, each cluster represents a single gene, whereas contigs of a cluster reflect different transcripts of that gene. A schematic view of the assembled clusters is presented on the GeneNest web site. Detailed information about sequences and their preprocessing results, as well as information about open reading frames, similarities between clusters or protein homologies, can be accessed interactively. GeneNest can be queried using BLAST against the consensus sequences or by keyword search. GeneNest is tightly linked to SYSTERS and SpliceNest as well as to external resources like EMBL.

SpliceNest

SpliceNest (12) is a web-based graphical tool to explore gene structure based on a mapping of the expressed sequence tag (EST) consensus sequences (contigs) from GeneNest to the complete human genome. Assuming that a cluster normally represents a single gene, every contig of a cluster is aligned separately to the same genomic region, using the spliced alignment program sim4 (13). Differences between the contigs may correspond to alternative splicing, but they can also be due to low sequence quality, genomic contamination or other artifacts. The alignments are visualized in a diagram showing the exon/intron structure of all contigs of a single cluster (i.e. gene) simultaneously, mapped on the common genomic sequence. Exons are represented as colored bars and introns as arrows. The visualization facilitates the identification of genuine splice variants. Furthermore, candidate loci of alternative splicing are automatically identified and highlighted. If a cluster has several matches in the genome, a ranked list of all matches is provided. Each contig is linked to the corresponding GeneNest assembly, giving easy access to information about individual EST and mRNA sequences. Other links point to detailed

*To whom correspondence should be addressed. Tel: +49 30 8413 1404; Fax: +49 30 8413 1152; Email: krause_a@molgen.mpg.de

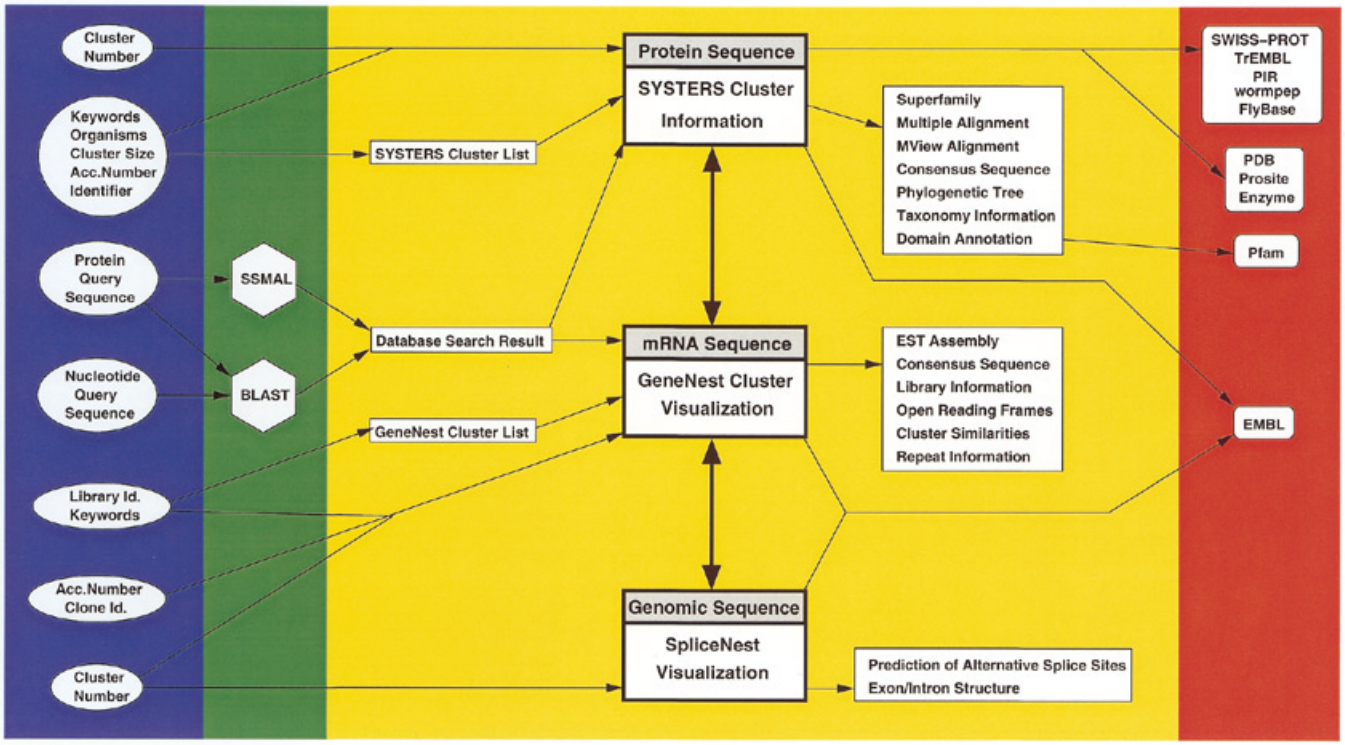

Figure 1. The integration of SYSTERS, GeneNest and SpliceNest into one framework. Possible queries to the databases are given on the left (blue), followed by the underlying query tools BLAST and SSMAL (green). The features and interactions of the SYSTERS, GeneNest and SpliceNest databases are shown in the middle (yellow) and links to external resources on the right (red).

alignments, related entries in the EMBL database or raw sequences. A toolbar allows zooming into the alignment. The current version of SpliceNest uses the GeneNest assembly based on human UniGene and the Golden Path genomic sequence (14).

SUMMARY

The three otherwise independent databases GeneNest, SpliceNest and SYSTERS are now fully linked with each other and to other major databases (Fig. 1). This allows navigating, e.g. from a protein to its UniGene cluster assembly and on to its genomic position and structure. Alternatively, one might enter via a sorted list of UniGene clusters on a chromosome and link from a particular cluster to its gene product in the context of a protein family. Thus, the linking of these databases facilitates navigation of sequence space between genomic DNA and protein sequences and families.

ACKNOWLEDGEMENTS

We acknowledge financial support from Bundesministerium für Bildung und Forschung (BMBF) and Deutsches Human Genom Projekt (DHGP).

REFERENCES

1. Krause,A. and Vingron,M. (1998) A set-theoretic approach to database searching and clustering. *Bioinformatics*, **14**, 430–438.
2. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
3. Barker,W.C., Garavelli,J.S., Hou,Z., Huang,H., Ledley,R.S., McGarvey,P.B., Mewes,H.W., Orcutt,B.C., Pfeiffer,F., Tsugita,A. *et al.* (2001) Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**, 29–32. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 35–37.
4. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
5. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
6. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
7. Nicodème,P. (1998) SSMAL: similarity searching with alignment graphs. *Bioinformatics*, **14**, 508–515.
8. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
9. Haas,S.A., Beissbarth,T., Rivals,E., Krause,A. and Vingron,M. (2000) GeneNest: automated generation and visualization of gene indices. *Trends Genet.*, **16**, 521–523.
10. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Lombard,V., Lopez,R., Parkinson,H. *et al.* (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **29**, 17–21. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 21–26.
11. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
12. Coward,E., Haas,S.A. and Vingron,M. (2002) SpliceNest: visualization of gene structure and alternative splicing based on EST clusters. *Trends Genet.*, **18**, in press.
13. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
14. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.