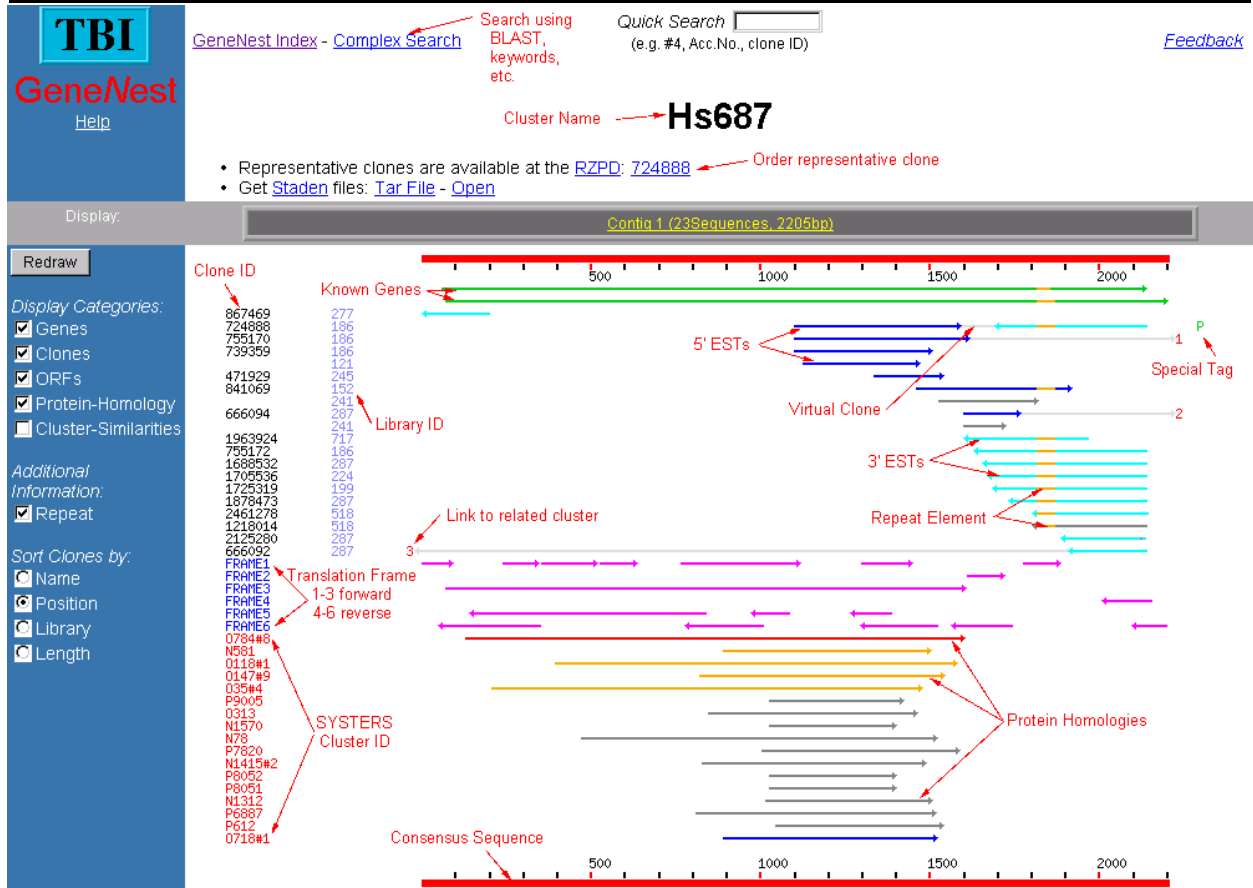


Figure 2. Visualization of a GeneNest cluster providing an interactive interface
(<http://www.dkfz.de/tbi/services/GeneNest/index>)



GeneNest: automated generation and visualization of gene indices (Draft version)

Expressed sequence tags (ESTs), introduced by Adams *et al.* in 1991¹, are a rapidly growing resource for analyzing genes. Although ESTs may be of low sequence quality they are extremely useful for detecting new genes, determining the genomic structure of a gene (exon-intron boundaries, alternative splicing)² or for expression studies³.

Since EST sequence information is highly redundant a single gene may be covered by thousands of ESTs each representing different parts of that gene. In order to simplify the analysis of specific genes several efforts have been made to cluster sequences belonging to the same gene⁴⁻⁶ resulting in so-called gene indices. Some commonly used gene indices are Unigene⁴ at NCBI (National Center for Biotechnology Information), the TIGR (The Institute for Genomic Research) gene indices⁵ and STACK⁶ at SANBI (South African National Bioinformatics Institute). In particular Unigene and the TIGR gene indices differ mainly in the clustering strategy used and in the presentation of cluster related information⁷. Clusters of the

TIGR gene indices are summarized by a database of consensus sequences each reflecting a single transcript. Additionally, the relative order of sequences within a cluster is roughly sketched. In contrast, sequences in Unigene are clustered less stringently such that alternative splice variants fall into the same cluster. Sequences derived from the same clone also may be clustered only based on their annotation. Unigene focuses on an

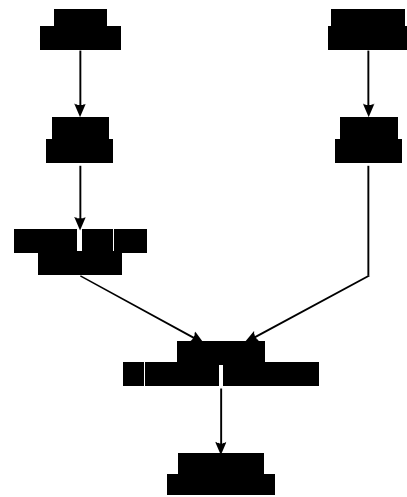
Figure 1. Schematic overview of the processing steps used by GeneNest

extensive linkage of clusters to related information e.g. mapping data or protein homologies.

Generation of gene indices

We developed GeneNest, a software and database for automated generation and visualization of gene indices.

The generation of the GeneNest gene indices starts either with a database of sequences extracted from the EMBL database, or from a Unigene database of ESTs already clustered



(Fig. 1). All sequences are subjected to clipping based on an extensive quality check. As a result of this step repeats, vector sequences as well as low quality regions are

masked. Similarities between these cleaned-up sequences are then determined using BLAST⁸ and sequences with sufficient local similarity are clustered. Sequences clustering together are assembled in order to determine their relative positions and to obtain a representative consensus sequence. A cluster may be split into several contigs each reflecting a group of sequences sharing global similarity. Such contigs are often caused by alternative splicing, ESTs derived from hnRNA or other artefacts like chimeric sequences. In a final step, a web site presenting all these data is generated automatically.

In some projects (in particular for *A. thaliana*) we treated genomic sequences containing coding sequence annotations (CDS) as potential genes. This strategy increases the number of sequences contributing to a gene index drastically compared to Unigene or the TIGR gene indices, thus also leading to an improved clustering.

Querying gene indices

The usefulness of a gene index strongly depends on its accessibility by the user. Frequently, private sequences have to be compared against the gene index database. Therefore, GeneNest offers a BLAST search against a database of representative consensus sequences. Queries on either an accession number, clone or library identifier or any keyword can be performed. Comprehensive information about a cluster can be downloaded as a Staden package⁹ project containing the alignment of all sequences related to that cluster.

Visualization

The GeneNest visualization serves as an entry point to the gene index database as well as to external databases. A contig of sequences sharing sequence similarities is the basic unit visualized by GeneNest (Fig. 2). All sequences are sketched by an arrow indicating the direction of this sequence.

References

- 1 Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656
- 2 Mironov, A.A. *et al.* (1999) Frequent alternative splicing of human genes. *Genome Res.* 9, 1288–1293
- 3 Schmitt, A.O. *et al.* (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumor tissue. *Nucleic Acids Res.* 27, 4251–4260

Additionally, the type (mRNA/gene or EST) or annotated direction of a sequence is reflected by different colours. If two ESTs are derived from the same clone they are connected by a grey line indicating the putative clone sequence. Specific features like repeats or poly A signals are also colour coded. As far as possible, clone identifiers are directly linked to institutions where these clones can be ordered. Clones labeled by a 'P' are part of a non-redundant clone set available at the Resource Center of the German Human Genome Project (RZPD). Each contig is represented by a single consensus sequence summarizing the sequence content of this contig. Since each consensus sequence reflects a single mRNA one could expect to find at least a partial open reading frame (ORF). Putative ORFs longer than 30 bases are symbolized by arrows providing the predicted amino acid sequence. In order to integrate a contig into the context of sequence homologies GeneNest displays homologies to other contigs/clusters as well as precomputed protein homologies to the SYSTERS protein consensus sequences¹⁰. These homologies are again marked by an arrow where the respective colour indicates the degree of sequence similarity.

All items visualized can be accessed interactively by clicking on the appropriate symbol thus linking to more detailed information, databases or related institutes. In the case of contigs composed of a large number of sequences the visualization of features can be turned on or off, optionally allowing the user to focus only on the data of interest.

Conclusions

Currently, the GeneNest database (<http://www.dkfz.de/tbi/services/GeneNest/index>) comprises gene indices of Man (based on Unigene), Mouse, *Arabidopsis thaliana* and Zebrafish. GeneNest combines properties of both Unigene and the TIGR

gene indices. Similar to Unigene, clusters are representing sequences related to one gene. However, GeneNest clusters are solely based on sequence homologies, but links between clusters containing sequences of the same clone are visualized. Because of the high rate of misannotation in public databases this strategy avoids clustering of unrelated sequences, still presenting all sequence relationships to the user. Contigs which are only generated by GeneNest often reflect single transcripts in a similar way as clusters of the TIGR gene indices. The comparison of contigs indirectly provides insight into the genomic structure of the gene represented. The comprehensive database of consensus sequences summarizing every putative transcript can be used as an efficient tool for searching homologies to private sequences. This way multiple searching of sequence databases, often containing only fragments of transcripts, can be avoided.

The interactive interface of GeneNest together with its compact presentation of cluster/gene related data minimizes the manual interaction for the user. GeneNest is especially useful for scientists analyzing so far unknown genes.

Future directions

The usefulness of gene indices on one hand depends on the source of sequences used. On the other hand gene indices have to be updated in a regular fashion to guarantee an optimal and complete summary of information related to a single gene. A regular update of the GeneNest is planned. Furthermore, the set of gene indices will be extended to several other organisms of scientific interest like rat and rice. With respect to the increasing amount of genomic sequence data available we also plan to integrate more detailed information about the genomic structure of a gene as well as alternative splice variants into the GeneNest visualization.

- 4 Adams, M.D. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377, 3–174
- 5 Schuler, G.D. *et al.* (1997) Pieces of the puzzle: expressed sequence tags and the catalogue of human genes. *J. Mol. Med.* 75, 694–698
- 6 Burke, J. *et al.* (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* 8, 276–280
- 7 Bouck, J. *et al.* (1999) Comparison of gene indexing databases. *Trends Genet.* 15, 159–162
- 8 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
- 9 Staden, R. (1996) The Staden sequence analysis package. *Mol. Biotechnol.* 5, 233–241
- 10 Krause, A. *et al.* (2000) The SYSTERS protein sequence cluster set. *Nucleic Acids Res.* 28, 270–272

Authors: Haas, S.A., Beissbarth, T., Rivals, E.[§], Krause, A., Vingron, M.

Department of Theoretical Bioinformatics

Im Neuenheimer Feld 280

D–69120 Heidelberg

Germany

Correspondence to: s.haas@dkfz.de

[§] Current address: LIRMM, UMR CNRS 5508, 161, rue Ada, 34392 Montpellier Cedex 5