positive feedback from these external users, we recommend our RUMMAGE annotation service to everyone who wants to get a quick and comprehensive overview of genomic sequence data.

The RUMMAGE Sequence Annotation Service is available at http://gen100.imb-jena.de/~baumgart/rummage/register.html. The URL leads to a registration form that has to be submitted before the first use. This is necessary to provide a user-specific password, which ensures confidential treatment of the sequence data and the corresponding annotation results. As soon as the password is assigned, each user may run as many jobs as desired.

References
1 Hattori, M. (2000) The DNA sequence of human chromosome 21. *Nature* 405, 311–319
2 Huang, X. (1994) GC rich region search tool. *Compu. Appl. Biosci.* 10, 219–225
3 Larsen, F. (1992) CpG islands as gene markers in the human genome. *Genomics* 13, 1095–1107
4 Burge, C. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94
5 Uberbacher, E.C. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. U. S. A.* 88, 11261–11265
6 Zhang, M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. U. S. A.* 94, 565–568
7 Solovyev, V.V. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* 22, 5156–5163
8 Thomas, A. (1994) A probabilistic model for detecting coding regions in DNA sequences. *IMA J. Math. Appl. Med. Biol.* 11, 149–160
9 Lowe, T.M. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964
10 Altschul, S.F. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
11 Huang, X. (1996) Fast comparison of a DNA sequence with a protein sequence database. *Microb. Comp. Genomics* 1, 281–291
12 Hofmann, K. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27, 215–219
13 Prestridge, D.S. (1995) Predicting Pol II promoter sequences using transciption factor binding sites. *J. Mol. Biol.* 249, 923–932

GeneNest: automated generation and visualization of gene indices

Expressed sequence tags (ESTs), introduced by Adams *et al.* in 1991 (Ref. 1), are a rapidly growing resource for analysing genes. Although ESTs might be of low sequence quality they are useful for detecting new genes, determining the genomic structure of a gene (such as exon–intron boundaries and alternative splicing sites)[2] and for gene-expression studies[3].

Because EST sequence information is highly redundant, a single gene might be covered by thousands of ESTs, each representing different parts of that gene. There have been several attempts to simplify the analysis of specific genes by clustering sequences belonging to the same gene[4–6], resulting in, so-called, gene indices. Some commonly used gene indices are Unigene (Ref. 4) at the National Center for Biotechnology Information (NCBI), the Institute for Genomic Research (TIGR) gene indices[5] and STACK (Ref. 6) at the South African National Bioinformatics Institute (SANBI). The Unigene and TIGR gene indices differ mainly in the clustering strategy used and presentation of cluster-related information[7]. Clusters of 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 sketched roughly. 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, based on their annotation. The Web presentation of Unigene at NCBI has extensive links between clusters and related information, such as mapping data or protein homologies.

Generation of gene indices

We have developed GeneNest (http://www.dkfz.de/tbi/services/GeneNest/index), a software and database for automated generation and visualization of gene indices.

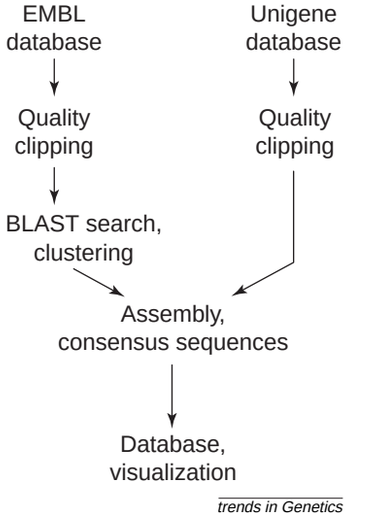
Generation of the GeneNest gene indices starts either with a database of sequences extracted from the EMBL database or from an already clustered database of ESTs from Unigene (Fig. 1). All sequences are subject to clipping, based on an extensive quality check. As a result of this step, repeats and vector sequences as well as low quality regions are masked. Similarities between these 'cleaned-up' sequences are then determined using BLAST (Ref. 8) and sequences are clustered if a near-perfect match extends over at least half of the shorter sequence. Sequences in a cluster are assembled in order to determine their relative positions and to obtain a representative consensus sequence. A cluster might be split into several contigs, each reflecting a group of sequences with global similarity. Such contigs are often caused by alternative splicing, ESTs derived from hnRNA or other artefacts such as chimeric sequences. In a final step, a Website presenting all these data is generated automatically.

In some projects, in particular for *Arabidopsis thaliana*, genomic sequences containing coding sequence annotations have been treated as potential genes. This strategy increases the number of sequences contributing to a gene index drastically, compared to Unigene or the TIGR gene indices, and, thus, also leads to an improved clustering.

Querying gene indices

The usefulness of a gene index depends strongly on its accessibility to the user. Most frequently, privately generated sequences are compared against the gene-index database. To this end,

FIGURE 1. Generation of GeneNest indices

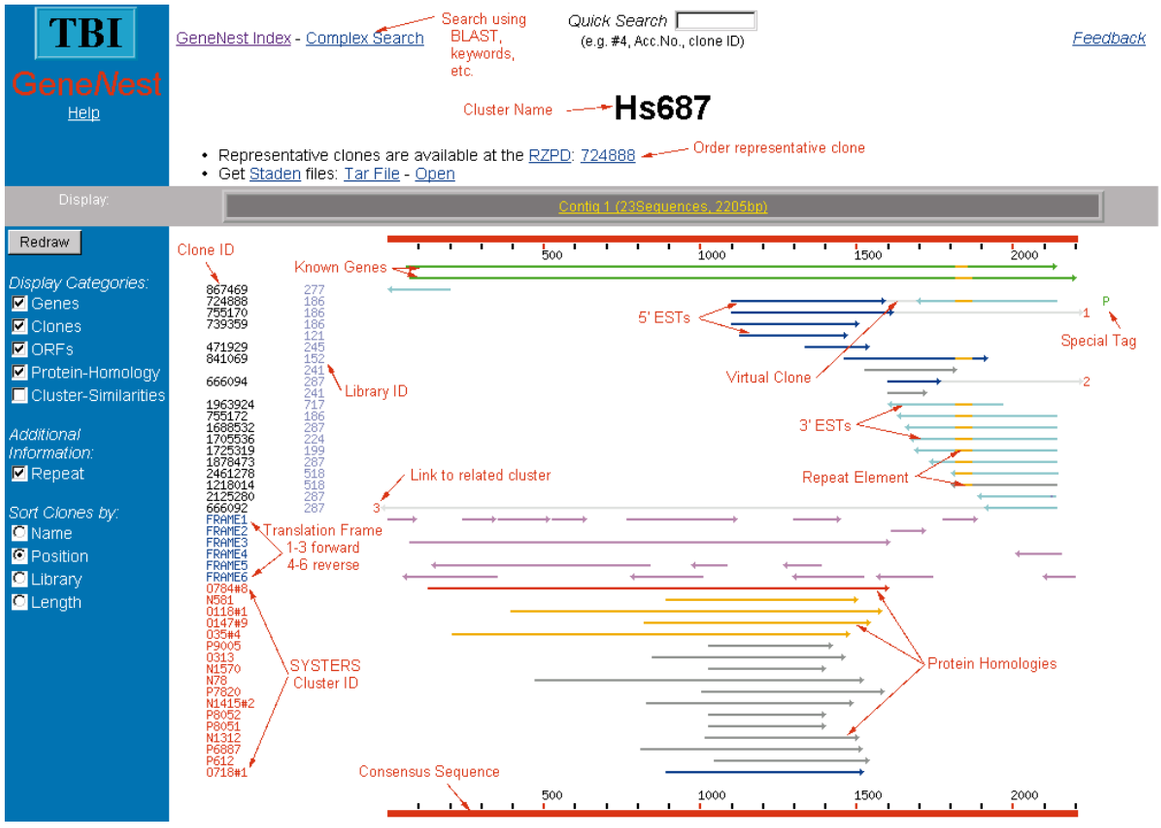


GeneNest is a database and software package for producing and visualizing gene indices from expressed sequence tags. The processing steps involved in the automated generation of gene indices are shown.

Stefan A. Haas
s.haas@dkfz.de
Tim Beissbarth
t.beissbarth@dkfz.de
Eric Rivals
rivals@lirmm.fr
Antje Krause
a.krause@dkfz.de
Martin Vingron
m.vingron@dkfz.de

Deutsches Krebsforschungs-zentrum, Department of Theoretical Bioinformatics, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.

FIGURE 2. GeneNest visualization



The basic GeneNest results are a contig of sequences that share sequence similarity. They appear as an interactive interface, allowing the user to access other homologous DNA, RNA and protein sequences from various other databases. GeneNest can be accessed from http://www.dkfz.de/tbi/services/GeneNest/index.

GeneNest offers a BLAST search against a database of representative consensus sequences. Queries can include an accession number, clone or library identifier, or any keyword. Comprehensive information about a cluster can be downloaded as a Staden package project[9] 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 depicted by an arrow that indicates the direction of the sequence. Additionally, the derivation of the sequence (mRNA, gene or EST) or annotated direction of a sequence is reflected by different colours. A grey line connects ESTs derived from the same clone indicating the putative sequence of the clone. Specific features, such as repeats or polyadenylation signals, are also colour-coded. As far as possible, clone-identifiers are linked to institutions where they can be ordered. Clones labelled '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 that summarizes the sequence content of the contig. Because each consensus sequence reflects a single mRNA one should expect to find at least one partial open reading frame (ORF) within each contig. The predicted amino acid sequences of putative ORFs longer than 30 nucleotides are also marked with an arrow. In order to integrate a contig into the context of sequence homologies, GeneNest indicates homologies between other contigs and clusters, as well as precomputed protein homologies to sequences within the SYSTERS protein cluster sets[10]. Again, an arrow indicates these homologies with the colour indicating the degree of sequence similarity.

All items can be accessed interactively by clicking on the appropriate symbol, thus linking to more detailed information, databases or related institutes. When contigs are composed of a large number of sequences the number of features visualized can be altered, allowing the user to focus on part of the data.

Conclusions

Currently, the GeneNest database comprises gene indices of humans (based on Unigene), mouse, *A. thaliana* and Zebrafish. GeneNest combines properties of both Unigene and the TIGR gene indices. As with Unigene, clusters represent sequences related to one gene. However, GeneNest clusters are based solely on sequence homologies although 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, but still presents all sequence relationships to the user. Contigs generated by GeneNest often reflect single transcripts in a similar way to gene indices generated by TIGR and the comparison of contigs provides insight into the genomic structure of a gene. The comprehensive database of consensus sequences that summarize every putative transcript is an efficient tool for searching homologies to private sequences and avoids searching of multiple sequence databases that often contain only fragments of transcripts. The interactive interface of GeneNest, together with its compact presentation

of cluster- and gene-related data, minimizes the manual input required from the user.

Future directions

On the one hand, the usefulness of gene indices depends on the source of sequences used. On the other hand, gene indices must be updated regularly to guarantee an optimal and complete summary of information about a single gene. We plan to update GeneNest three times per year. Furthermore, the set of gene indices will be extended to several other organisms including rats and rice. Because of 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.

Acknowledgements

This work is supported by a grant from the German Human Genome Project (DHGP) and the Zentrum zur Identifikation von Genfunktionen durch Insertionsmutagenese bei *Arabidopsis thaliana* (ZIGA).

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
4 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
5 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
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

Population genetics: past and future

Evolutionary Genetics: From Molecules to Morphology
edited by Rama S. Singh and Costas B. Krimbas
Cambridge University Press, 2000, £60.00 hbk (702 pages) ISBN 0 521 57123 5

A tribute to the work in population genetics of Richard C. Lewontin on the occasion of his 65th birthday, this excellent collection reviews many areas of population genetics – both their historical development and their current research focus. The 32 chapters are divided into eight sections, beginning with the history of population genetics. The majority of the book is a discussion of molecular and phenotypic variation within and between populations, and our ability to detect selection acting on this variation. Mechanisms of speciation and the interface of population genetics and behaviour and ecology are dealt with in the final two sections. Many of the authors have worked with Lewontin, and their names are a roll call of most of the leading practitioners. Although Lewontin is most associated with the use of starch gel electrophoresis of enzymes as a means of assessing variation in populations, initially using *Drosophila pseudoobscura*, this volume reminds us of the breadth (and depth) of his work in evolutionary biology. It is clear that he would have been regarded as a major contributor to the field without the allozyme experiments. Moreover, apart from the insight that we gain into Lewontin's career, the consistently high quality of the chapters (there are very few weak spots) means that this is admirable introduction to the great issues of 20th-century population genetics. I would recommend it wholeheartedly, for example, to a new graduate student moving into evolutionary genetics from a related field such as molecular biology. Considering the cross-fertilization between subdisciplines that characterizes evolutionary biology, the amount of repetition between chapters is surprisingly low, and is mainly limited to the history of population genetics and Lewontin's role in it.

The editors' choice of title, 'From Molecules to Morphology', represents the scope of the book in the prosaic sense that the subjects of the chapters range from the molecular to the morphological. The title also hints at what Lewontin rightly thinks is still a major problem confronting evolutionary genetics, that is, the difficulties of integrating population variation at the genetic and phenotypic levels. The cover illustration, reproduced from Lewontin's *The Genetic Basis of Evolutionary Change*[1] and included in his chapter here, also reflects this. A rigorous evolutionary biology based on genes and gene sequences is surely impossible unless the impact of the genes on phenotypes are known. Without this knowledge, even if we understand the causal relationship between observable aspects of individual phenotype and the environment-dependent expected phenotype that is fitness, we will not be able to predict the selection acting on the genes themselves.

In his more general writings, Lewontin is known for his view that the knowledge of genes alone is inadequate for predicting, for example, the minutiae of behavioural differences between people or peoples. This idea is congruent with his insistence on the need for knowledge of the epigenetic relationship between genes and phenotypes in population genetics. In a sense, this is a question about developmental biology, but the understanding of development that is required is not the main focus of current studies of developmental evolution. The issue here is not how genes cause the differences between a mollusc and an echinoderm, for example, but how genes and environment interact to create variation within a species. At this level, the impact of genetic differences on individual traits might be less than that of the environment, and genetic–environment interactions might be so strong and prevalent that any attempt to partition the phenotypic variance into its genetic and environmental components will fail.

One disconcerting aspect of the history of population genetics revealed in these accounts is that once the elaborate and beautiful theory of mathematical population genetics had been created, mainly before the development of the technical means to produce the data sets for analysis, future progress in the field was dominated by the sequential

John Brookfield
john.brookfield@nottingham.ac.uk

Institute of Genetics, University of Nottingham, Nottingham, UK NG7 2UH.