Annotating eukaryote genomes

Suzanna Lewis*, Michael Ashburner† and Martin G Reese‡

The Genome Annotation Assessment Project tested current methods of gene identification, including a critical assessment of the accuracy of different methods. Two new databases have provided new resources for gene annotation: these are the InterPro database of protein domains and motifs, and the Gene Ontology database for terms that describe the molecular functions and biological roles of gene products. Efforts in genome annotation are most often based upon advances in computer systems that are specifically designed to deal with the tremendous amounts of data being generated by current sequencing projects. These efforts in analysis are being linked to new ways of visualizing computationally annotated genomes.

Addresses
*Berkeley Drosophila Genome Project, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3200, USA;
e-mail: suzi@fruitfly.berkeley.edu
†Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK and EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK;
e-mail: m.ashburner@gen.cam.ac.uk
‡Berkeley Drosophila Genome Project, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3200, USA and Neomorphic Inc., 2612 8th Street, Berkeley, CA 94710, USA;
e-mail: mgreese@lbl.gov

Current Opinion in Structural Biology 2000, **10**:349–354

0959-440X/00/$ – see front matter

Abbreviations

EST	expressed sequence tag
GASP	Genome Annotation Assessment Project

Introduction: the problem of annotation

How big is the problem?

Recent advances in sequencing technology are making the generation of whole genome sequences commonplace. Capillary sequencers speed the production of raw data. Changing tactics from the traditional mapping and sequencing of clones in series to an integrated simultaneous mapping and sequencing approach (a.k.a. whole genome shotgun) has significantly reduced the amount of time it takes to sequence a genome. These improvements in genomic sequencing are possible because of software advances that fully exploit mapped clone constraint data and directly attack the problems that repetitive sequences cause during sequence assembly.

At present, several very large-scale genomic sequencing projects are complete or are expected to be complete within a few months. These initial genome sequences represent key model organisms in genetics and include five eukaryotes, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana*, as well as a draft human sequence. In a few years, sequencing new genomes and individuals will become routine practice. Yet, as raw sequences these offer little and thus place major demands on biocomputational professionals and biologists to interpret the genome.

What do we mean by ‘annotation’?

The process of interpreting raw sequence data into useful biological information is the process of annotation. Annotations describe the genome and transform raw genomic sequences into biological information by integrating computational analyses, auxiliary biological data and biological expertise. Traditionally, small-scale studies of isolated genes carried out in an individual researcher’s laboratory use a combination of computational and experimental methods that permit very detailed descriptions of their features. They offer a narrow, but deep view. In contrast, the best current results from the annotation of large eukaryotic genomes provide a complete perspective and overview of the entire genome, but are rather superficial and incompletely describe individual genes. They offer a broad, but shallow view. At present, the annotation of large-scale sequences is a compromise but, ideally, the aim is to have both breadth and depth in our description of the genome.

What tools are available for annotation?

Substrate

The very first strategic decision is choosing what sequence to annotate. This is especially important for shotgun sequencing projects, but is also an issue for clone-based projects. In general, annotation should begin as early in a project as is possible, because the analysis of the sequence will often identify problems in the raw sequence or in its assembly. A consequence is that annotation is being added to a moving target and tracking annotations forward to newer versions of the sequence presents a computational problem that must be solved. In some respects, this is a local data management issue and tools to solve this problem will be specific to individual sites and databases. If annotations are to be shared and transferred between sites and sequences, however, it does become a more general problem. At present, there are no tools for mapping annotations onto arbitrary pieces of sequence, although we point out that local sequence alignments may offer a rather general basis for a solution to this problem.

Identification

Having decided on the substrate for annotation, we are faced with the problem of identification. The types of features that can be detected and described in the sequence include the location of the protein-coding genes; the structures of those genes (including untranslated regions and control elements, in addition to the exon–intron structure for all possible transcripts); the probable translations of every transcript into a protein product; the location of repetitive sequences and their nature; and the location of

Table 1

Useful URLs.

Title	References	URL
Gene-finding software		
FGENEH	[43]	http://genomic.sanger.ac.uk/gf/gfs.shtml
GENEID	[44,45]	http://www1.imim.es/geneid.html
GENIE	[14,15,46]	http://www.fruitfly.org/seq_tools/genie.html
GENSCAN	[47]	http://CCR-081.mit.edu/GENSCAN.html
HMMGene	[48]	http://www.cbs.dtu.dk/services/HMMgene/
GeneMarkHMM	[49]	http://genemark.biology.gatech.edu/GeneMark/
GRAIL	[50]	http://compbio.ornl.gov/
GlimmerM	[7]	http://www.tigr.org/softlab/glimmer/glimmer.html
GeneBuilder	[12]	http://www.itba.mi.cnr.it/webgene
Wise2/Genewise	[51]	http://www.sanger.ac.uk/Software/Wise2/
BLOCKS	[21•,22]	http://blocks.fhcrc.org
Formats and tools		
GFF		http://www.sanger.ac.uk/Software/formats/GFF/
GFFTOOLS		http://www1.imim.es/~jabril/GFFTOOLS/
Bioxml		http://www.bioxml.org
Model organism projects and databases		
WormBase (*C. elegans*)		http://www.wormbase.org/
SGD (*S. cerevisiae*)		http://genome-www.stanford.edu/Saccharomyces/
TAIR (*A. thaliana*)		http://www.arabidopsis.org/
MGD (mouse)		http://www.informatics.jax.org/
FlyBase (*Drosophila*)		http://flybase.bio.indiana.edu/
BDGP (*D. melanogaster*)		http://www.fruitfly.org
EDGP (*D. melanogaster*)		http://edgp.ebi.ac.uk
Annotation-related sites		
Genome Annotation Assessment Project (GASP1)		http://www.fruitfly.org/GASP1
Gene Ontology (GO project)		http://www.geneontology.org
InterPro		http://www.ebi.ac.uk/interpro/

genes encoding noncoding RNAs. This is only a partial list and can easily be expanded. What is important to remember is that the identification of these essential elements of the genomic sequence is a necessary, but insufficient basis for annotation.

There are two major classes of technique for the prediction of genes — *ab initio* methods and homology-based methods. In prokaryotes and in some simple eukaryotes (such as *S. cerevisiae*), genes normally have single continuous open reading frames and adjacent genes are separated by short intergenic regions. By contrast, genes in most eukaryotes can be very complex, with many exons, introns that may be tens of kilobases in length, noncoding 5′ and 3′ exons, and alternatively spliced products. In addition, complex relationships among genes may be quite frequent, for example, genes contained within the introns of other genes and adjacent series of highly related genes. The consequence is that any *ab initio* method must combine the prediction of gene components (exons, introns, splice sites and so on) with the prediction of a model for how these components may be assembled into a gene.

Computational gene finding has evolved steadily over the past 20 years and excellent reviews in this area have been written by Fickett and Tung [1], Claverie [2], Guigó [3] and Burge and Karlin [4]. In 1998, Haussler [5] categorized gene-finding methods as either ‘signal sensors’ or ‘content sensors’. In broad terms, signal sensor methods exploit descriptions of pertinent sites, such as splice junctions, start and stop codons, branch points, promoters, termination of transcription and others, to identify genes. Content sensor methods employ models that are based upon extended lengths of sequence, such as exons and introns. Most recent eukaryotic gene identification methods combine both signal sensor and content sensor approaches.

Gene finding in bacterial genomes is close to being a solved problem. In 1999, Ramakrishna and Srinivasan [6] reported an improved and adapted GeneScan algorithm for bacterial and organellar genomes that has a sensitivity of 100% for *Plasmodium falciparum* and 98% for the *Mycoplasma genitalium* and *Haemophilus influenzae Rd* genomes, and a specificity at the nucleotide level of 0.25% when compared with biological sequence annotation [6]. Salzberg *et al.* [7] tackled a more difficult problem. This group built a gene-finding system for a small complex eukaryotic organism, the malaria parasite *P. falciparum*. They based their new system (GlimmerM) on an existing Markov chain program that functioned well for bacteria. The new program models the more complex splice sites of eukaryotes and pioneered the use of interpolated Markov chains for gene finding.

Eukaryotic genomes pose a still more difficult problem. At the 7th International Conference on Intelligent Systems in Molecular Biology, computational biologists shared their efforts in attacking this problem in an annotation assessment experiment called GASP (Genome Annotation Assessment Project) [8••]. The experiment was a blind test and was performed on a well-studied 2.9 Mb sequence region of the *D. melanogaster* genome [9]. At the base level, the best programs reached a sensitivity of 95%, whereas for most of the seven participating groups, the specificity was approximately 90%. At the exon level, the programs achieved up to 78% sensitivity, but only a little over 50% specificity. At the gene level, the average program's sensitivity was over 60% and its specificity below 40%. (For a review of the metrics used in evaluating gene-finding performance, see Burset and Guigó [10].) The clearest outcome from this experiment is that clean datasets from well-studied regions of genomes are absolutely essential to effectively evaluate, compare and refine existing methods. It is likewise clear that they are equally essential to improve the training sets used in gene prediction. Therefore, it is not too surprising that most recent advances in gene finding have come from improved methods for creating training sets.

Rogozin *et al.* [11] have developed a new method that is able to derive and train a model for coding regions of completely new gene families. The method computationally recognizes evolutionarily conserved coding regions, rather than relying upon the annotations in public databases. This enables the generation of the large high-quality datasets describing complex gene models that are required for a typical gene-finding program. The program is called SYNCOD and it is now integrated into a more complete gene-finding system called GeneBuilder [12]. GeneBuilder integrates information from different signal sensors, such as promoters, splice sites, start and stop codons, and the 3′ untranslated regions, with statistical properties from content sensors for coding sequences. In addition, information about homologous proteins, EST (expressed sequence tag) sequences and repetitive elements is integrated into a gene structure model using a dynamic programming method. This approach is similar to the pioneering work by Stormo and Haussler [13], Kulp *et al.* [14], Reese *et al.* [15] and Burge and Karlin [4].

A second source of improvement in gene prediction comes from gradual adaptations and extensions to existing programs. Popular programs such as GenScan, Genie, Fgenes+, HMMGene, Genemark and GRAIL have all been improved to automatically train the methods and models for new organisms (see [8••] for an assessment).

Progress in the application of signal sensors that model binding sites and other features in genomic DNA has been presented in the areas of promoter recognition [16•,17•], start codons in bacteria [18] and genomic repeats [19•,20].

Characterization

Identification leads to the third major problem — the characterization of these annotated features (elements). This characterization must be done in several ways: in terms of the relationships between the sequences of the elements and other sequences (both within the genome being annotated and within other genomes); in terms of the structure of the elements (e.g. the protein domains of predicted proteins); and in terms of the predicted function of the elements (e.g. what inferences can be drawn concerning the biological function of a predicted protein).

Characterizations on the basis of homology have, traditionally, meant using methods such as BLAST or FASTA for detecting regions of similarity between the sequences being analyzed and the universe of sequences available from the major public sequence databases at either the nucleic acid or the (predicted) protein level. These techniques remain absolutely invaluable and new variations are extending the capabilities of homology-based methods [21•,22]. In addition, a new class of method that exploits EST data is now supplementing these. Although the stated purpose of most large-scale EST sequencing programs was gene discovery, it has turned out that these sequence resources are invaluable both for gene prediction and for confirming models of gene structure. Indeed, they are the only reliable method for detecting 5′ and 3′ untranslated regions. The alignment of EST (or full-length cDNA) sequences with genomic DNA is a specialized alignment problem that must take into account splice site models and the expectation of a single open reading frame in determining a match between two sequences. SIM4 [23] and ACEMBLY [24] are two examples of this type of software.

Determining the protein domains of the predicted proteins is a crucial part of turning raw sequence into biologically relevant data. There are several independent methods now available to construct databases of patterns in protein sequences. These include PROSITE [25], PRINTS-S [26], PFAM [27,28], PRODOM [29] and BLOCKS [21•,22]. A new database, InterPro [30••], has begun the task of integrating these into a single resource (the beta release of October 1999 includes PROSITE, PRINTS and PFAM). The advantage is that not only do these databases now share a common nomenclature and documentation for protein domains and patterns, but also that tools can be built to incorporate the variety of methods for scanning sequences for predicted domains. In a manner somewhat analogous to using alignment data to augment gene predictions, Wise2 [31] exploits protein domain knowledge from PFAM to enhance the quality of gene and exon predictions from primary sequence data.

A tentative conclusion regarding the function of a newly predicted gene can be drawn from its similarities and motifs. The Gene Ontology collaboration (the 'GO' project at URL http://www.geneontology.org) is an effort to use carefully defined terms to describe function, process and cellular location. One major advantage is that when different 'single organism' databases adopt the same vocabulary, then the community will have a powerful method for exploring functional aspects of the genomes of several different organisms. The GO project is integrated closely with InterPro, so that the association of protein motifs with functional descriptions will be easily maintained. The GO project is also creating a sequence set containing only those genes to which human curators have actively assigned a function term in order to enable tentative functional assignment on the basis of a combination of sequence similarity and InterPro motifs. At this time, the GO project is limited to its originators from FlyBase, SGD (*Saccharomyces* Genome Database) and MGD (Mouse Genome Database) (for URLs, see Table 1). By the spring of 2000, gene associations and expansion of the vocabularies will be extended to include other organisms.

Quality assessment

Logically independent of, but pragmatically in parallel to solving the third major problem is the fourth — quality control. We must assure ourselves of the accuracy and completeness of the data. For example, by comparing the transcript set to an EST set, one can evaluate (approximately) whether all of the protein-coding genes have been identified. Other tests must be made to assess the correctness of intron–exon structures and the protein products, to detect fused genes and split genes, and to correlate transcripts with known sequenced genes. These tests require simple tools but, generally, the existing tools are *ad hoc* and inconsistently applied. This is a crucial area that calls for improvement.

Large-scale genomic features

Identification and characterization deal with individual genes, but the genome is larger than the sum of its genes. In other words, there is the problem of describing the genome as a whole. For example, regularities in the arrangement of genes along the chromosomes may reveal some insight of biological interest or, as another example, evolutionary history may be deduced from the overall structure of the annotated genome. Tools that summarize the number of gene families in a genome; that extract high-quality genes for improving training sets; that evaluate the presence of genes in heterochromatin; that characterize gene distribution over the genome; and that profile averages of such features as intron length, GC content, number of exons and intergenic spacing will all illuminate our understanding of the genome.

Jareborg *et al.* [32] focused on the alignment of noncoding regions of 77 orthologous mouse and human gene pairs using a new method to identify conserved genomic regions. Other recent examples of genome-wide comparisons published in 1999 come from Elofsson and Sonnhammer [33], two articles by Marcotte, Pellegrini *et al.* [34•,35], Enright *et al.* [36•] and Andrade *et al.* [37].

Scaling up

Genomes are large and the only way to cope with the volume of data is to automate as much as possible. For data that cannot be automated, tools must be created to maximize the efficient handling of the data by human curators. There are currently two such automated pipelines in the public area: the Oakridge Genome Annotation Channel (http://compbio.ornl.gov/channel/) [38] and ensEMBL (http://ensembl.ebi.ac.uk/) [39••]. The nascent form of ensEMBL was used in annotating human chromosome 22 [40], the largest contiguous sequence yet described.

Delivering the annotated genome

Unless the methods and results are widely distributed to the community (or to customers), annotation is a self-indulgent exercise. Mechanisms to publish both the tools and the data are essential to complete the task. Because of the volume of the data and its complexity, this task is not straightforward. At opposite ends of the spectrum are bulk transfers of data that require standard data exchange formats to be accepted and supported by an array of utilities and graphical browsers, and sophisticated query tools, so that answers can be quickly located in the vast sea of data.

There are a number of complementary efforts underway to develop syntaxes that are rich enough to semantically capture the data. The Gene-finding format (GFF) is the most well developed. At http://www.bioxml.org, there are contributions such as GAME (Genome Annotation Markup Elements) for describing annotation data in XML. Efforts at the OMG (Object Management Group) are aimed at describing biological objects for exchange of information using CORBA (http://www.omg.org/homepages/lsr/).

Visualization and browsing are essential if annotation is to be used by biologists. A number of different Java-based browsers are available (Genome Channel [38], GeneScene [41] and Jalview [42]), but there are none that, as yet, provide all of the functionality required. In addition, programs for converting GFF to Postscript, which can then be viewed as a static image or print-out, are now available (Abril, GFF2PS: http://www1.imim.es/~jabril/GFFTOOLS/).

Conclusions

This review has discussed the initial steps required to annotate eukaryotic genomes. Considerable progress has been achieved in '*ab initio*' methods for gene prediction (see the several publications in *Genome Research*, volume 10, 2000 that are related to GASP [8••]). Nevertheless, there remains the need for further work, especially in the areas of gene characterization and classification. The importance of the visualization of biological data has eventually found its place and progress here has been tremendous. Many eukaryote genomes, large and small, will be sequenced in their entirety in the next few years. Their annotation and analysis will continue to present challenging problems to both computer scientists and biologists.

Note added in proof

Genome Research (volume 10, issue 4, 2000) has recently published a comprehensive collection of articles related to genome annotation, all focusing on GASP. In addition, details concerning the first large-scale usage of the GO and InterPro databases in annotation were published in papers describing the *Drosophila* genome sequence [52,53].

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fickett JW, Tung CS: **Assessment of protein coding measures.** *Nucleic Acids Res* 1992, **20**:6441-6450.

2. Claverie JM: **Computational methods for the identification of genes in vertebrate genomic sequences.** *Hum Mol Genet* 1997, **6**:1735-1744.

3. Guigó R: **Computational gene identification.** *J Mol Med* 1997, **75**:389-393.

4. Burge CB, Karlin S: **Finding the genes in genomic DNA.** *Curr Opin Struct Biol* 1998, **8**:346-354.

5. Haussler D: **Computational genefinding.** *Trends Biochem Sci* 1998, **suppl**:12-15.

6. Ramakrishna R, Srinivasan R: **Gene identification in bacterial and organellar genomes using GeneScan.** *Comput Chem* 1999, **23**:165-174.

7. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H: **Interpolated Markov models for eukaryotic gene-finding.** *Genomics* 1999, **59**:24-31.

8. •• Reese MG, Hartzell G, Harris NL, Ohler U, Lewis SE: **Genome annotation assessment in *Drosophila melanogaster*.** *Genome Res* 2000, **10**:483-501.

The April 2000 issue of *Genome Research* (see Note added in proof) includes both descriptions and evaluations of the Genome Annotation Assessment Project.

9. Ashburner M, Misra S, Roote J, Lewis SE, Blazej R, Davis T, Doyle C, Galle R, George R, Harris N *et al.*: **An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*. The Adh region.** *Genetics* 1999, **153**:179-219.

10. Burset M, Guigó R: **Evaluation of gene structure prediction programs.** *Genomics* 1996, **34**:353-367.

11. Rogozin IB, D'Angelo D, Milanesi L: **Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences.** *Gene* 1999, **226**:129-137.

12. Milanesi L, D'Angelo D, Rogozin IB: **GeneBuilder: interactive *in silico* prediction of gene structure.** *Bioinformatics* 1999, **15**:612-621.

13. Stormo GD, Haussler D: **Optimally parsing a sequence into different classes based on multiple types of evidence.** *Ismb* 1994, **2**:369-375.

14. Kulp D, Haussler D, Reese MG, Eeckman FH: **A generalized hidden Markov model for the recognition of human genes in DNA.** *Ismb* 1996, **4**:134-142.

15. Reese MG, Eeckman FH, Kulp D, Haussler D: **Improved splice site detection in Genie.** *J Comput Biol* 1997, **4**:311-323.

16. • Knudsen S: **Promoter2.0: for the recognition of PolII promoter sequences.** *Bioinformatics* 1999, **15**:356-361.

This paper, together with [17•], describes statistically clean approaches to the hard problem of identifying promoters in eukaryotic genome sequences.

17. • Ohler U, Harbeck S, Niemann H, Noth E, Reese MG: **Interpolated Markov chains for eukaryotic promoter recognition.** *Bioinformatics* 1999, **15**:362-369.

See annotation to [16•].

18. Frishman D, Mironov A, Gelfand M: **Starts of bacterial genes: estimating the reliability of computer predictions.** *Gene* 1999, **234**:257-265.

19. • Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-580.

This paper presents a novel method for detecting repeat elements in DNA in a rigorous statistical framework.

20. Kurtz S, Schleiermacher C: **REPuter: fast computation of maximal repeats in complete genomes.** *Bioinformatics* 1999, **15**:426-427.

21. • Henikoff JG, Henikoff S, Pietrokovski S: **New features of the Blocks Database servers.** *Nucleic Acids Res* 1999, **27**:226-228.

The BLOCKS database is a tremendous resource for identifying protein motifs. It was used extensively in the annotation of the *Drosophila* genome (see Note added in proof).

22. Henikoff S, Henikoff JG, Pietrokovski S: **Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations.** *Bioinformatics* 1999, **15**:471-479.

23. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8**:967-974.

24. ACEMBLY on World Wide Web URL: http://alpha.crbm.cnrs-mop.fr/acembly/

25. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucleic Acids Res* 1999, **27**:215-219.

26. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W: **PRINTS-S: the database formerly known as PRINTS.** *Nucleic Acids Res* 2000, **28**:225-227.

27. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2000, **28**:263-266.

28. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL: **Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.** *Nucleic Acids Res* 1999, **27**:260-262.

29. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons.** *Nucleic Acids Res* 2000, **28**:267-269.

30. •• Fleischmann W, Moller S, Gateau A, Apweiler R: **A novel method for automatic functional annotation of proteins.** *Bioinformatics* 1999, **15**:228-233.

Semiautomated protein function assignment using InterPro is the best resource for comprehensively studying newly discovered genes.

31. Wise2 on World Wide Web URL: http://www.sanger.ac.uk/Software/Wise2/

32. Jareborg N, Birney E, Durbin R: **Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs.** *Genome Res* 1999, **9**:815-824.

33. Elofsson A, Sonnhammer EL: **A comparison of sequence and structure protein domain families as a basis for structural genomics.** *Bioinformatics* 1999, **15**:480-500.

34. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D:
• **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**:83-86.
Studying protein domains in complete genomes gives many clues about the evolution and function of genes.

35. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285**:751-753.

36. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein**
• **interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86-90.
Studying protein domains in complete genomes gives many clues about the evolution and function of genes.

37. Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C *et al.*: **Automated genome sequence analysis and annotation.** *Bioinformatics* 1999, **15**:391-412.

38. Mural RJ, Parang M, Shah M, Snoddy J, Uberbacher EC: **The Genome Channel: a browser to a uniform first-pass annotation of genomic DNA.** *Trends Genet* 1999, **15**:38-39.

39. ensEMBL on World Wide Web URL:
•• http://www.ensemble.org/
A landmark in the automatic annotation of eukaryotic sequences on a very large scale.

40. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ *et al.*: **The DNA sequence of human chromosome 22.** *Nature* 1999, **402**:489-495.

41. GeneScene on World Wide Web URL: http://www.fruitfly.org/annot/genescene-launch-static.html

42. Jalview on World Wide Web URL: http://www.ebi.ac.uk/~michele/jalview/contents.html

43. Solovyev VV, Salamov AA, Lawrence CB: **Identification of human gene structure using linear discriminant functions and dynamic programming.** *Ismb* 1995, **3**:367-375.

44. Guigo R: **Computational gene identification: an open problem.** *Comput Chem* 1997, **21**:215-222.

45. Guigo R: **Assembling genes from predicted exons in linear time with dynamic programming.** *J Comput Biol* 1998, **5**:681-702.

46. Kulp D, Haussler D, Reese MG, Eeckman FH: **Integrating database homology in a probabilistic gene structure model.** *Pac Symp Biocomput* 1997:232-244.

47. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78-94.

48. Krogh A: **Two methods for improving performance of an HMM and their application for gene-finding.** *Ismb* 1997 **5**:179-186.

49. Besemer J, Borodovsky M: **Heuristic approach to deriving models for gene-finding.** *Nucleic Acids Res* 1999, **27**:3911-3920.

50. Uberbacher EC, Mural RJ: **Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach.** *Proc Natl Acad Sci USA* 1991, **88**:11261-11265.

51. Birney E, Durbin R: **Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison.** *Ismb* 1997, **5**:56-64.

52. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: **The genome sequence of *Drosophila melanogaster.*** *Science* 2000, **287**:2185-2195.

53. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W *et al.*: **Comparative genomics of the eukaryotes.** *Science* 2000, **287**:2204-2215.