# Genome annotation techniques: new approaches and challenges

Alistair G. Rust, Emmanuel Mongin and Ewan Birney

**As more of the human genome draft sequence is finished, and genomes from other organisms begin to be sequenced, the demand for accurate and reliable genome annotation will increase significantly. To facilitate this industrial-scale genome annotation, automated bioinformatics solutions are increasingly required. As a result, automatic genome annotation systems have become more important in gene discovery within recent years. The design of such large-scale bioinformatics systems is an evolving and dynamic field, based on central cores of bioinformatics software tools and relational databases. Not only must these systems efficiently manage and integrate large volumes of genomic data, but they must also deliver accurate gene predictions and effectively distribute annotation data to the biosciences community.**

**Alistair G. Rust**
**Emmanuel Mongin and**
***Ewan Birney**
European Bioinformatics
Institute (EMBL-EBI),
Wellcome Trust Genome
Campus
Hinxton
Cambridge
UK CB10 1SD
tel: +44 1223 494420
fax: +44 1223 494468
*e-mail: birney@ebi.ac.uk

▼ With the increase in sequencing capacity around the world over the past few years, the amount of human genomic sequence being submitted to database repositories has been increasing at an exponential rate. Although considerable effort is still being expended on turning the draft sequence into the finished sequence, attention is now turning to the processing of genomes from other species [1,2]. Current figures put the number of eukaryotic genomes being sequenced in part at over 170 [3], with at least 14 in the process of being fully sequenced. This exponential increase in raw sequence needs to be matched by increases in the accurate annotation of this huge variety of genomes.

Accurate annotation of the human genome and other species is an essential element in supporting current drug discovery efforts. Validating potential drug targets against genomic sequence requires accurate annotation in the first instance to make this procedure worthwhile [4]. Accurate genome annotation has traditionally been achieved via manual curation, using the experience of expert individuals to annotate sequence by hand. Although manual curation can attain high degrees of accuracy, it cannot keep pace with the throughput of multi-species sequence now being demanded. As a result, bioinformatics solutions are increasingly required to develop automatic annotation techniques to support and complement the manual curation process [5,6].

## Automatic genome annotation pipelines

To handle the large, industrial-scale throughput of sequence data, one of the key developments in recent years has been the implementation of automatic genome annotation pipelines. The primary goal of the pipeline process is to deliver highly accurate and reliable genome annotations, using the widest possible range of evidence from available databases. In essence, pipelines are the integration of suites of bioinformatics software tools with multiple databases, to manage automatically the analysis and storage of genomic sequence.

As pipelines have evolved, the trend has been to move away from single algorithm methods and towards consensus-based approaches, whereby the combined results of gene predictors and similarity search methods are used, for example, to generate more reliable predictions. The design of current pipeline annotations reflects this methodology, such that genomic sequences pass through several successive levels of algorithms. Each layer of processing provides further refinement of annotation detail.

Although no two pipelines are exactly alike, the basic structure of an annotation pipeline is depicted in Fig. 1. (Literature regarding pipelines is scarce, thus, Fig. 1 is an attempt to summarize

1359-6446/02/$ – see front matter ©2002 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(02)02289-4

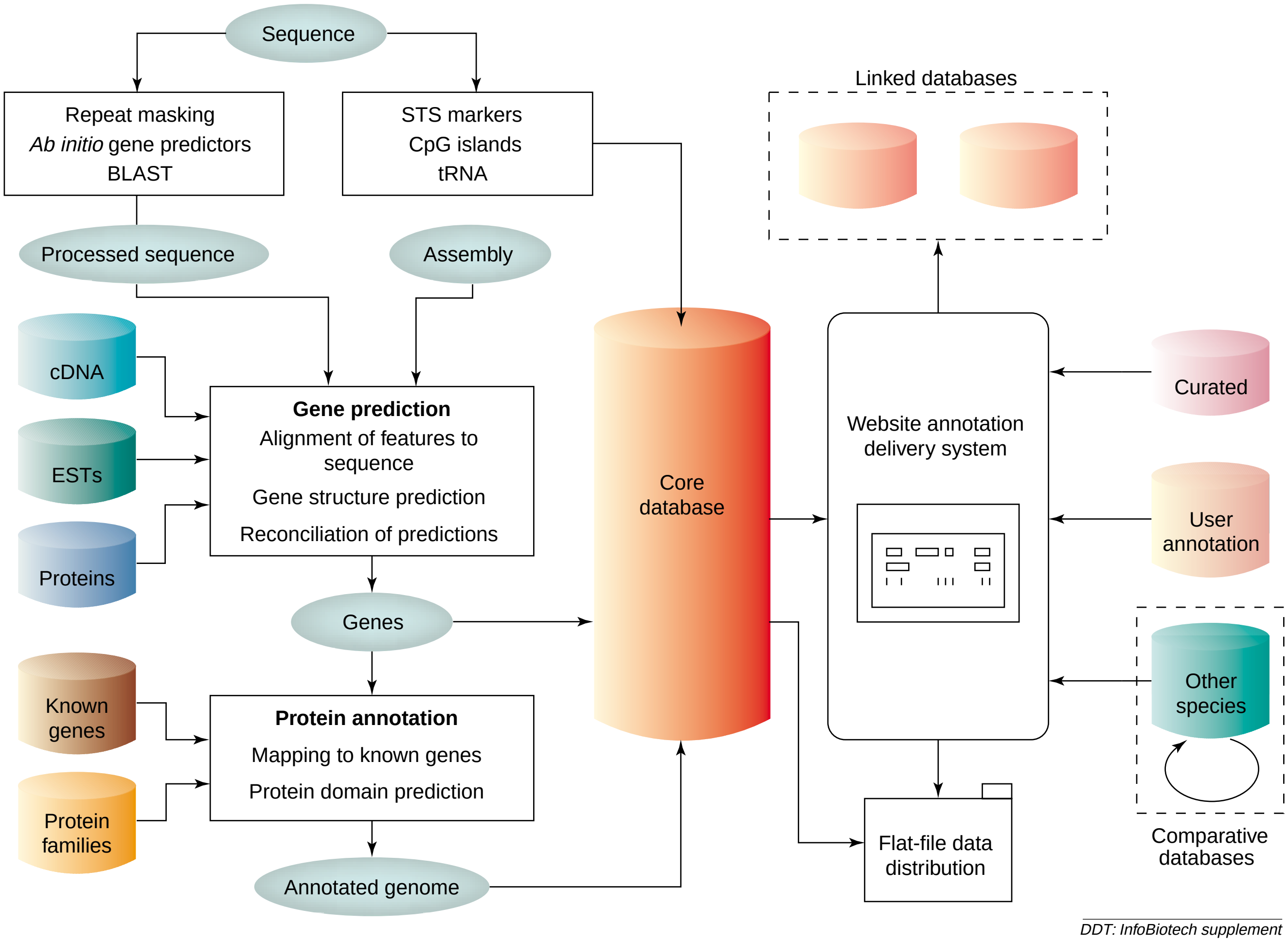


Figure 1. The generic structure of an automatic genome annotation pipeline and delivery system. The diagram can be simply partitioned into the analysis pipeline stage on the left and delivery system on the right, where data is distributed between the two stages via the central, core database. Raw, sequence data enters the pipeline and flows down through the gene prediction analyses before being stored in the core database. The website annotation delivery system is a key feature as it serves to integrate the gene predictions and features from the core database alongside multiple, external databases.

a generic approach.) The definition of a pipeline used here includes the analysis of raw sequence data into gene predictions, and the storage of these predictions and other features within a relational database. The annotation process is then completed with the integration of external databases and the distribution of annotation data via websites and downloadable data files.

Several genomic pipelines exist worldwide. Publicly funded projects include Ensembl at the European Bioinformatics Institute (EBI)/Sanger Institute [7], the NCBI Analysis Pipeline, and the Oak Ridge National Laboratories (ORNL) Genome Channel. There are also several commercial pipelines that underlie the Celera Discovery System [2,8], Incyte Genomics LifeSeq system and the Paracel GeneMatcher2 solution (see Box 1 for related URLs). (Several browsers, for example the one provided by UCSC, include *ab initio* gene predictions, but they do not operate complete pipeline systems to perform these analyses.)

## From raw sequence to gene predictions

### Raw sequence pre-processing

The initial stages of pipelines are typically very similar, in that they simply process raw sequence data in preparation for subsequent analysis stages. This includes masking known repeats and low complexity sequences using RepeatMasker [http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl] and identifying homology matches using BLAST (Basic Local Alignment Sequence Tool) [9]. Pre-processing analyses can be targeted at species other than human, provided that the relevant species libraries and databases are available. Scans for other features, such as sequence tagged site (STS) markers and CpG islands, can also be performed at this stage.

## Box 1. Useful human genome annotation and browser URLs

**Automated annotation pipelines**

- EBI/Sanger Institute Ensembl Project: http://www.ensembl.org/Homo_sapiens/
- NCBI Human Genome Browser: http://www.ncbi.nlm.nih.gov/genome/guide/human/
- The Oak Ridge National Laboratories Genome Channel: http://compbio.ornl.gov/channel/
- Celera Discovery System: http://cds.celera.com/
- Incyte Genomics – Genomics Knowledge Platform: http://www.incyte.com/incyte_science/technology/gkp/
- Paracel GeneMatcher2 System: http://www.paracel.com/products/gm2.html

**Human genome browsers**

- UCSC Human Genome Browser: http://genome.cse.ucsc.edu/cgi-bin/hgGateway/
- Softberry Genome Explorer: http://www.softberry.com/berry.phtml?topic=genomexp
- Viaken Enterprise Ensembl Solution: http://www.viaken.com/ns/solutions/ensembl.html
- LabBook Inc. Genomic Explorer Suite: http://www.labbook.com/products/ExplorerSuite.asp
- University of Tokyo Gene Resource Locator Browser: http://grl.gi.k.u-tokyo.ac.jp/

**Other useful sites**

- The Institute for Genomic Research (TIGR): http://www.tigr.org/
- Human Genome Central: http://www.ensembl.org/genome/central/ and http://www.ncbi.nih.gov/genome/central/

## Gene prediction

The actual stages of gene prediction or gene building, are where the specifications of the pipelines diverge, as a wide variety of algorithms and source databases are used in the prediction process. This is perhaps the most dynamic stage of the pipeline, because as new algorithms are developed and more databases become available, pipeline components are frequently modified to enhance the annotation process. It is not feasible to review all the possible combinations of tools here, so the following description simply aims to summarize the key software packages used, and how they are integrated with various data sources to provide annotations.

Gene prediction can be characterized by taking individual homology search matches and *ab initio* computational predictions, aligning them to the genomic sequence, and then making predictions of gene structure. Figure 2 is a simplified, general overview of the gene prediction process. The data sources of homology matches can be either protein or DNA, and each class of matches is typically analysed separately. In most cases, final gene predictions are achieved by reconciling prediction alignments from the different techniques to produce plausible transcripts. The aim is to reduce the redundancy arising from overlapping gene features, such as predicted exons.

*Predictions based on protein matches.* Analysed protein features can be derived from two different subsets of databases. For human genome analysis, subsets of databases containing only human sequences, such as RefSeq [10] (incorporated in the NCBI, Ensembl and Celera pipelines) and SWISS-PROT [11] (Ensembl), can be used to target the gene building specifically to known, characterized genes. Alternatively, matches from non-human databases (such as EMBL-vertrna [12]), can similarly be used to produce predictions.

*Predictions based on DNA sequence.* One key set of resources for gene prediction algorithms are full-length cDNAs (which are incorporated into all pipelines from sources such as GenBank [13] and EMBL [12]), because they not only contain exon structures, but also 3′ and 5′ untranslated regions of mRNA (UTRs). Recently, expressed sequence tag (EST) sequences have also been integrated into the pipelines of NCBI, Ensembl and Celera for exon and transcript predictions. However, EST sequences are by nature highly variable in terms of their quality, and thus the reconciliation of EST predictions is prone to heavy redundancy of predicted gene structures.

Ab initio *gene prediction programs. Ab initio* gene predictors rely on the statistical qualities of exons rather than on homologies. Examples of such prediction programs include Genscan [14] (used by Celera, ORNL and Ensembl), fgenesh [15] (used in the Softberry and UCSC browsers as well as Celera's pipeline) and GrailEXP [16] (ORNL). The direct application of *ab initio* gene prediction programs is not, however, completely sufficient. The prediction algorithms are prone to many false-positives, are often unable to predict correctly spliced variants, and cannot predict 3′ and 5′ UTRs. However, using prediction programs in conjunction with other data sources, such as EST data, can provide valuable sources of annotation data.

In bioinformatics terms, the choice of alignment and predictions tools used depends on the source of the original feature or match. For example, gene predictions from protein alignments are often performed by the package Genewise [17] (used, for example, by Ensembl and Celera). For DNA-based features, gene structure can be determined using est2genome [18] (Ensembl and Celera) or SIM4 [19].

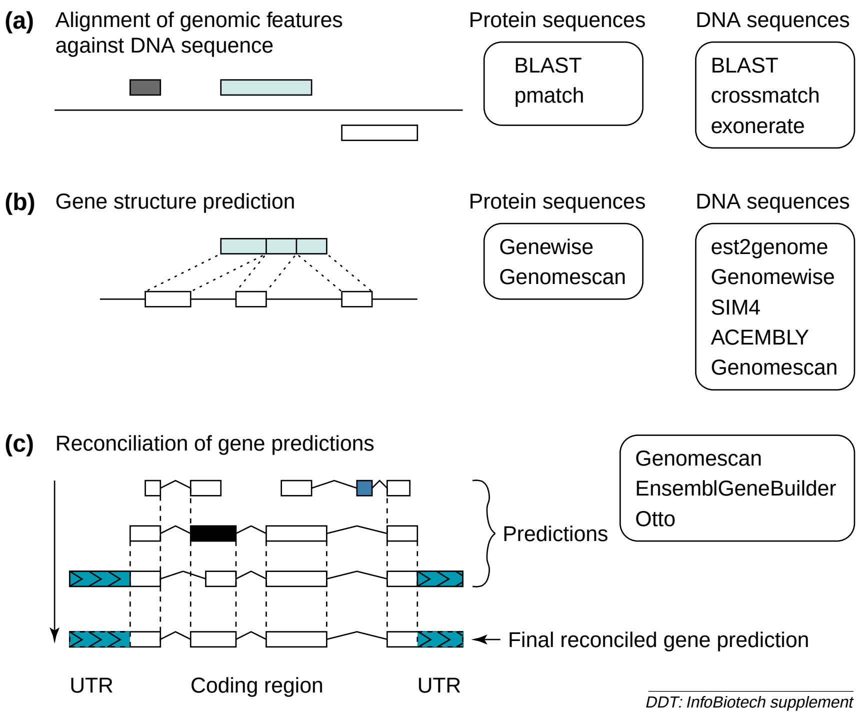
## Gene function characterization

### Mapping to known genes

The combination of manually curated sequence databases (such as RefSeq and SWISS-PROT) with powerful similarity search algorithms enable predicted gene sequences to be mapped to known genes and proteins. If a match exists between a predicted gene and a corresponding record in a curated database, gene characterization data can then be retrieved. Once a prediction is assigned a mapping, links to other databases, such as HUGO (as linked by NCBI, UCSC and Ensembl) [20], can then be included in the annotation process.

### Protein domain annotation

Protein domains are also essential in determining the function of predicted genes. Different databases can be used for annotation, including Pfam [21], PRINTS [22], PROSITE [23], ProDom [24], BLOCKS [25] and SMART [26]. Each of these databases has, however, been designed for specific problems and therefore have their own inherent strengths and weaknesses. To address this issue, different domain signatures are being integrated into the Interpro project (http://www.ebi.ac.uk/interpro/), creating a unique characterization for a given protein family, domain or functional site [27]. Domains of the protein sequences can then be identified using this signature method. The use of Interpro provides the least-redundant and extensive annotation currently available.



*DDT: InfoBiotech supplement*

Figure 2. A simplified schematic of algorithmic gene prediction. The prediction process can be summarized in three stages: **(a)** alignment, **(b)** prediction and **(c)** reconciliation of predictions. A variety of bioinformatics software packages are used during prediction and a cross-section of typical tools are listed within the oval boxes, under the type of genomic sequence for which they can be used. The reconciliation of gene predictions in (c) highlights how predictions via several sources are reconciled to give a more accurate prediction. In this hypothetical example, the final prediction is determined through the addition of UTRs, the removal of fragmented exon predictions (shaded block in the top prediction), and the adjustment of internal exon boundaries (the black block is the preferred prediction).

### Gene ontology

The function of a predicted protein can be obtained from its similarity to already annotated proteins, or from its domains. A common vocabulary is an important requirement to enable definition of this function. The Gene Ontology (GO) project (http://www.geneontology.org/) aims at defining such common terms to specify molecular function, biological process and cellular location [28]. This project is integrated with Interpro, and sets of annotated sequences are related to GO terms. Thus, predictions with assigned mappings to proteins or genes, using either similarity matches with curated genes or using Interpro annotations, can be given functional GO definitions. GO terms are now integrated into the browsers provided by Ensembl, NCBI and Celera.

## Sharing genome annotations

An important consideration of the annotation process is the presentation of information. The end-users of annotation information can range from lab biologists, who simply wish to browse particular annotated regions visually, to bioinformatics specialists, who wish to perform specific statistical, data-mining tasks. Ideally, the database and software systems should handle these contrasting demands.

### Website display and ftp sites

Website distribution is by far the most common means of distributing genome annotation. An example of the Ensembl project's web interface is shown in Fig. 3. The features contained in this sample webpage, such as gene predictions, database alignments, repeat features and markers, are those commonly displayed across the range of website interfaces. A list of URLs for human genome browsers is given in Box 1.

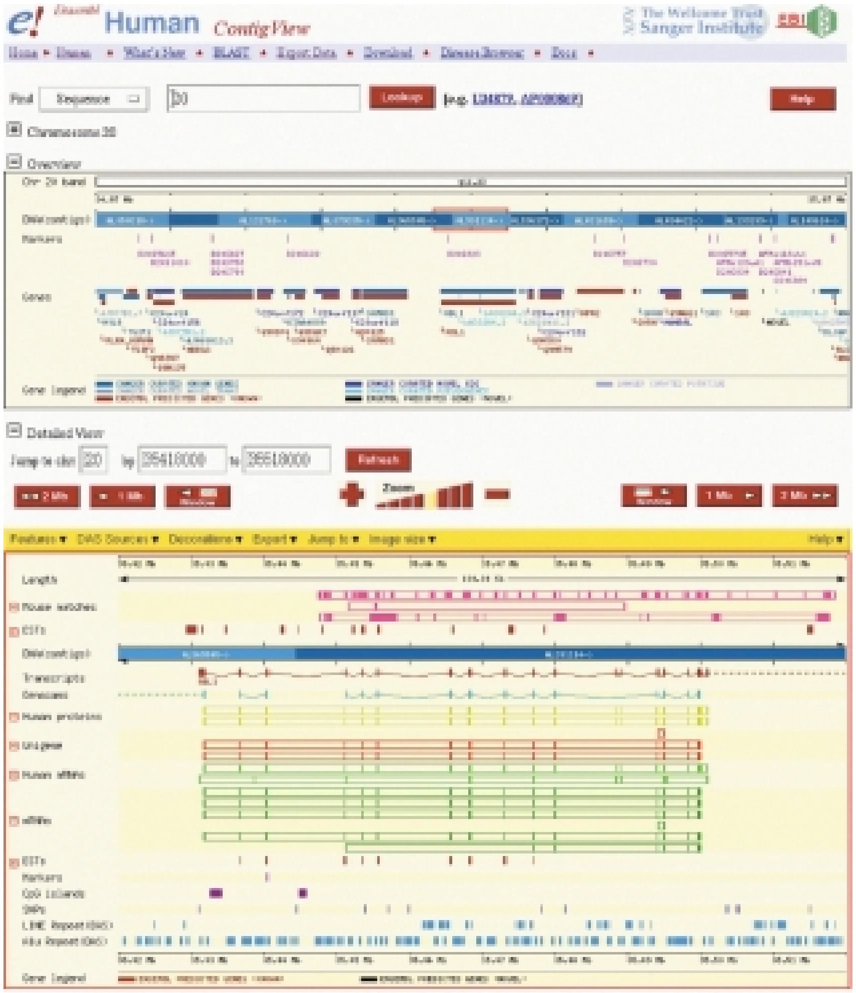


Figure 3. A typical human genome annotation website interface. The screenshot is taken from the Ensembl Contigview page, displaying predictions for the human gene RBL1 (Ensembl gene ID ENSG00000080839 in release 3.26.1). The display is typical of most annotation websites in that it presents gene predictions with all the biological evidence used in making the prediction. In this example, corroborating evidence is provided from alignments to protein databases, expressed sequence tag (EST) data and algorithmic (Genscan) predictions. Other important features such as sequence tagged site (STS) markers and SNPs are displayed in their own separate tracks on the display. This screenshot also contains two new sources of annotation information that have recently become available, namely alignment matches against mouse genomic sequence, and data provided from Distributed Annotation System (DAS) sources (see text for further details). From this page it is possible to access more detailed information, including sequence alignments, splice site predictions and records from external databases. Genome annotation websites are also becoming increasingly reconfigurable, enabling the display of tracks to be defined interactively by the user.

One of the key advantages of website browsers is that their use does not require expert bioinformatics skills and they are thus more accessible to a wide range of researchers wishing to gain access to genomic annotation. The majority of genome website browsers, including those at UCSC, NCBI and Ensembl, enable simple queries to be performed, such as searching for genes by their accession identifiers or visualizing a specific region of a chromosome. There are, however, limitations to this presentation of data because it makes it difficult to perform large-scale data mining. The distribution of databases via ftp sites is one solution, enabling more experienced users to retrieve the data they require and to run analyses locally. For example, the websites at UCSC, NCBI and Ensembl offer the facility to download not only raw genomic sequence data, but also annotated datasets that were generated using their locally run analyses.

## Open annotation

Obviously, most genomes are too complex to be analysed and processed by a single group, and some believe that open-source annotation is the solution to this problem [29]. Open annotation enables researchers to have access to annotations available in the community and to share their own contributions with the community. The main problem of open annotation is the need for a common protocol between systems that enables genome data to be freely exchanged. Two such approaches that aim to define standards have been proposed by the AGAVE (Architecture for Genomic Annotation, Visualization and Exchange) (http://www.agave.org) and the Distributed Annotation System (DAS) [30] projects.

AGAVE is an XML-based format, developed by DoubleTwist, to encourage the exchange of genomic information and the development of related tools. With the closure of DoubleTwist, however, the future of AGAVE depends on whether the open-source community created around this project takes over the responsibility of continuing with its development. DAS is similarly an XML-based format. It relies on a common 'reference' or 'baseline' sequence upon which the annotation is based, and DAS sources are added as further tracks to supplement the annotation. DAS data are controlled solely by its provider, and any group adhering to the DAS specification can exchange and compare their data, without requiring the central coordination of data. Examples of DAS sources serving human data can be found within Ensembl and at TIGR (http://www.tigr.org/tdb/DAS/DAS.shtml).

Both projects are still in their relative infancies, but as the uptake in usage increases, they can offer effective mechanisms for sharing annotation by providing standards to which the biosciences community can adhere. Both AGAVE and DAS were not, however, designed to provide any controls over the quality or accuracy of submitted data. Sources of shared annotation must therefore be judged by the end-user and will be based on the reputation of the laboratory or group supplying the annotation data in addition to any supporting literature references.

## Challenges facing automatic annotation systems

### Data warehousing: a solution for large-scale data mining

Owing to its complexity and volume, genomic annotation data does not lend itself to being easily searched or queried. Technically, there are two problems that inhibit complex queries on genomic data. First, the desired query statement might be too complex to implement, even for an enthused bench biologist, requiring detailed knowledge of the relational database language. Second, the computing power needed might be too expensive in most cases for queries performed on large, monolithic databases. Solutions to these issues have already been implemented by the business sector using data warehousing, which segregates information into denormalized databases, enabling fast querying and data retrieval. This has prompted equivalent proposals for molecular biology [31], and an initial implementation has been set up by the Ensembl project [7]. The ability to extract datasets of interest efficiently can result in subsequent stages of statistical analyses or data mining. There are currently a large variety of data-mining tools available, examples of which can be found in the suite of tools provided by the NCBI (http://www.ncbi.nlm.nih.gov/Tools/).

### The requirement to remain flexible

The development of automated annotation pipelines is an evolving process. Some of the key requisites of a pipeline are that it must be both flexible and adaptive, representing a significant computational challenge for both the hardware and software systems. One of the driving forces behind these requirements is that as the quality of sequences and assemblies continue to improve, redundant sequences are replaced with new, superior sequences. This demands a flexible system in which new, individual sequences can be added and analysed without disrupting the whole system.

Similarly, the architecture of a pipeline must be easily extensible. As new, improved algorithms and methodologies are developed, they should be able to be incorporated into the analysis process without redesign of the system. Furthermore, as a greater variety of species are sequenced, pipelines must be able to perform optimized, species-specific analyses. Pipelines are therefore becoming increasingly reconfigurable to incorporate specific sets of databases and arbitrary software configuration parameters.

## Future opportunities

### Comparative genomics

As more genomes are sequenced and become publicly available in the next few years, comparative genomics will become one of the greatest areas of development. For example, the publicly funded mouse genome project has recently released 7X coverage of genomic sequence in addition to a preliminary assembly [32], enabling cross-comparisons between this model organism and the human genome. Such comparisons offer several significant opportunities for drug discovery.

In terms of human sequence annotation, cross-species analyses will improve the accuracy of gene predictions and refine the definition of gene function. One attractive use of comparative genomics is to use the presence of conserved sequences to deduce all functional regions in a genome. Protein coding genes are likely to be highly conserved between closely related species (e.g. mouse and human), and other regions, such as RNA genes and regulatory regions, could also be elucidated. Although the development of algorithmical frameworks is just starting in these areas (e.g. Q-RNA [33]), it holds great promise.

There is, however, much work to be done in the development of bioinformatics tools to analyse the conservation of sequences between genomes. Although analysis and display tools such as Vista [34], Synplot [35] and FamilyJewels (http://family.caltech.edu) are being developed, the integration of such tools with the current automated approaches is still in its infancy. The design of genome browsers and websites that can intelligently display and annotate comparative results is also still a developing field of bioinformatics [for an example see the Artemis Comparison Tool (http://www.sanger.ac.uk/Software/Artemis/ACT/)].

### Integrating and delivering new data

Data integration will be one of the main avenues of opportunity for genomic projects over the next decade. Integration can be defined in two ways [7], horizontal and vertical integration.

*Horizontal integration*. As more genomes become available, genomic systems should be able to cross-match species that can be sensibly compared. For example, this will be important given the variety of available vertebrate genomes. The inclusion of the single comparative database in Fig. 1 represents just one of the possible multiple-species databases.

*Vertical integration*. New flows of data coming from proteomics and microarray sources will soon have to be incorporated. These pose new challenges, such as the highly complex and varied nature of microarray data. Therefore, initiatives to define standards for such data are crucial [36] because they will make the integration of new data sources into existing annotation systems a more seamless process.

## Concluding remarks

Automatic genome annotation systems have progressed a long way within a short period of time and are becoming increasingly important in gene discovery. Although the field of automated genome annotation systems is constantly evolving, the systems themselves are grounded upon central cores of bioinformatics software tools and associated relational databases. The evolution process is set to continue at a rapid pace: as the number of sequenced genomes increases, the demands on integration of new genomes into the current systems will increase. This, in turn, will drive the demand for an openess towards the distribution of annotation data, and to the delivery of genomic data in forms suitable for large-scale data mining.

## Acknowledgements

We would like to thank the the Ensembl team for many useful discussions and suggestions during the preparation of this review. Thanks also to Mark Robinson of the Biocomputation Group at the University of Hertfordshire, UK, for help in the formatting of Fig. 3.

## References

1 International Human Genome Sequencing Consortium (2001) A physical map of the human genome. *Nature* 409, 934–941

2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

3 Bernal, A. *et al.* (2001) Genomes online database (GOLD): a monitor of genome projects worldwide. *Nucleic Acids Res.* 29, 126–127

4 Reiss, T. (2001) Drug discovery of the future: the implications of the human genome project. *Trends Biotechnol.* 19, 496–499

5 Lewis, S. *et al.* (2000) Annotating eukaryote genomes. *Curr. Opin. Struct. Biol.* 10, 349–354

6 Searls, D.B. (2000) Using bioinformatics in gene and drug discovery. *Drug Discov. Today* 5, 135–143

7 Hubbard, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41

8 Kerlavage, A. *et al.* (2002) The Celera Discovery System. *Nucleic Acids Res.* 30, 129–136

9 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

10 Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* 29, 137–140

11 Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48

12 Stoesser, G. *et al.* (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res.* 30, 21–26

13 Benson, D.A. *et al.* (2002) GenBank. *Nucleic Acids Res.* 30, 17–20

14 Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94

15 Solovyev, V.V. *et al.* (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3, 367–375

16 Xu, Y. and Uberbacher, E.C. (1997) Automated gene identification in large-scale genomic sequences. *J. Comput. Biol.* 4, 325–338

17 Birney, E. and Durbin, R. (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5, 56–64

18 Mott, R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *CABIOS* 13, 477–478

19 Florea, L. *et al.* (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8, 967–974

20 Wain, H. *et al.* (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.* 30, 169–171

21 Bateman, A. (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280

22 Attwood, T.K. *et al.* (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res.* 30, 239–241

23 Falquet, L. *et al.* (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* 30, 235–238

24 Corpet, F. *et al.* (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* 28, 267–269

25 Henikoff, S. *et al.* (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15, 471–479

26 Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234

27 Apweiler, R. *et al.* (2000) InterPro – An integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 16, 1145–1150

28 The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29

29 Hubbard, T. and Birney, E. (2000) Open annotation offers a democratic solution to genome sequencing. *Nature* 403, p. 825

30 Dowell, R.D. *et al.* (2001) The distributed annotation system. *BMC Bioinformatics* 2, p. 7

31 Schönbach, C. *et al.* (2000) Data warehousing in molecular biology. *Briefings Bioinform.* 1, 190–198

32 Lindblad-Toh, K. *et al.* (2001) Progress in sequencing the mouse genome. *Genesis* 31, 137–141

33 Rivas, E. and Eddy, S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2, p. 8

34 Mayor, C. *et al.* (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16, 1046–1047

35 Göttgens, B. *et al.* (2001) Long-range comparisons of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res.* 11, 87–97

36 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365–371