38–42 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg083

Ensembl 2002: accommodating comparative genomics

M. Clamp, D. Andrews, D. Barker, P. Bevan, G. Cameron[1], Y. Chen[1], L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond[1], T. Hubbard*, A. Kasprzyk[1], D. Keefe[1], H. Lehvaslaiho[1], V. Iyer, C. Melsopp[1], E. Mongin[1], R. Pettett, S. Potter, A. Rust[1], E. Schmidt[1], S. Searle, G. Slater[1], J. Smith, W. Spooner, A. Stabenau[1], J. Stalker, E. Stupka[1], A. Ureta-Vidal[1], I. Vastrik[1] and E. Birney[1]

The Wellcome Trust Sanger Institute, and [1]European Bioinformatics Institute (EMBL–EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

Received September 30, 2002; Accepted October 2, 2002

ABSTRACT

The Ensembl (http://www.ensembl.org/) database project provides a bioinformatics framework to organise biology around the sequences of large genomes. It is a comprehensive source of stable automatic annotation of human, mouse and other genome sequences, available as either an interactive web site or as flat files. Ensembl also integrates manually annotated gene structures from external sources where available. As well as being one of the leading sources of genome annotation, Ensembl is an open source software engineering project to develop a portable system able to handle very large genomes and associated requirements. These range from sequence analysis to data storage and visualisation and installations exist around the world in both companies and at academic sites. With both human and mouse genome sequences available and more vertebrate sequences to follow, many of the recent developments in Ensembl have focusing on developing automatic comparative genome analysis and visualisation.

INTRODUCTION

A genome sequence provides a natural framework about which to organise biological data. In the short time in which genome sequences have been available, genome databases have proved invaluable resources to researchers. Ensembl has been involved in the continued analysis of human data, analysis of the mouse genome (Mouse Genome Sequencing Consortium, Nature, submitted), analysis of the *Anopheles gambiae* genome (Holt *et al.*, *Science*, in press) and the *C. briggsae* genome. In addition, the Fugu project, utilising the open nature of the Ensembl system, used a clone of the Ensembl system for their analysis (1). The availability of the genome sequence of these provides for a great insight into much of their biology, however it also enables the process of linking these species at the DNA level to define regions of synteny and construct a catalogue of orthologous genes. The pairwise linkage of genomes is the first stage in the genome wide organisation of a number of vertebrates, as both rat and zebrafish genomes will soon reach 'draft' status. The new challenge is to find ways to use automatic comparative genomics analysis to organise and integrate this multi-species data. The knowledge about genes and proteins in individual genomes that is brought together in this way must be made to add to the completeness of our view of vertebrates, without simply complicating it.

The Ensembl project (2–4) is attempting to address this challenge with a number of specific projects focused on comparative genomics, whilst maintaining and improving its role of per-genome analysis. Ensembl remains an entirely open project with all data freely available and code openly licensed. Ensembl has developed a strong developer network of users in both academia and industry and is being installed both to mirror Ensembl generated data and used as a software foundation for user projects.

NEW FEATURES

Comparative genome analysis

As new genomes are annotated by Ensembl, extra information can be obtained using comparative genomics methods. There are three types of comparative information available in Ensembl: fine grained DNA–DNA alignments; orthologous protein information and large scale synteny data. Currently, these data are available between human and mouse but in the future Ensembl will provide comparative analysis between drosophila and mosquito and a three way comparison of fugu,

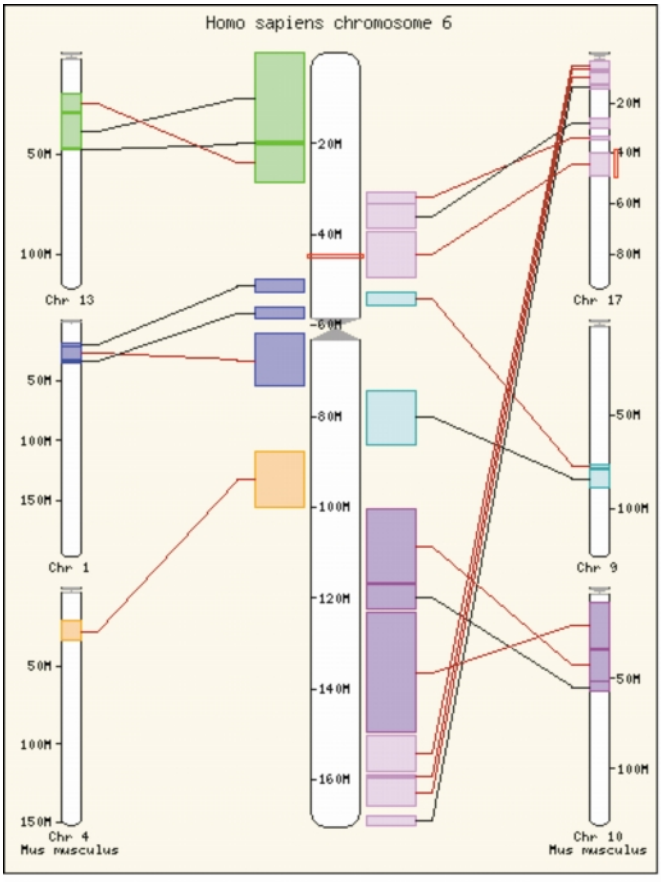
*To whom correspondence should be addressed. Tel: +44 1223494983; Fax: +44 1223494919; Email: th@sanger.ac.uk

human and mouse. As new genomes arrive (e.g. zebrafish and rat) these will be incorporated into the Ensembl compara database.

The fine-grained DNA–DNA analysis is done using an initial fast alignment step using exonerate (2) to isolate similar regions in the two genomes being compared. A more sensitive alignment is then done using BLASTN. The individual BLASTN alignments are displayed as a track on contigview. As these are clickable, the user can easily move between regions sharing similarity in different genomes.

These DNA alignments are also used to find syntenic blocks between two genomes and these are displayed via a new web interface, syntenyview, and example of which is shown in Figure 1. To eliminate paralogous matches between the two genomes, the DNA–DNA alignments are clustered chromosome by chromosome based on their direction and distance from the previous match. To reduce the effect of misassemblies, only regions that are longer than 100 kb are displayed on the web site. As in contigview (2) the user can click on a region of one genome's chromosome and go directly to the other genome.

The final set of comparative data produced by Ensembl is done at the protein level rather than the DNA level and attempts to identify orthologous proteins in two genomes. The process starts by using BLASTP to compare all one genome's proteins to the other genome's proteins. A core set of protein



Human Chromosome 6

Jump to chromosome 6 Lookup

Jump to mapview for chromosome statistics.

Homology Matches

Homo_sapiens Genes		*Mus_musculus* Homologues
SUPT3H (44.92 Mb)	->	ENSMUSG00000038954 (chr 17 : 44.01 Mb)
RUNX2 (45.39 Mb)	->	RUN2_MOUSE (chr 17 : 43.61 Mb)
CLIC5 (45.87 Mb)	->	Q9CYD1 (chr 17 : 43.19 Mb)
ENSG00000174868 (46.05 Mb)	->	ENSMUSG00000039317 (chr 17 : 43.14 Mb)
ENPP4 (46.11 Mb)	->	ENSMUSG00000023961 (chr 17 : 43.10 Mb)
ENPP5 (46.13 Mb)	->	Q9EQG7 (chr 17 : 43.08 Mb)
DSCR1L1 (46.19 Mb)	->	NM_030598 (chr 17 : 42.84 Mb)
CYP39A1 (46.52 Mb)	->	NM_018887 (chr 17 : 42.67 Mb)
UCP4_HUMAN (46.62 Mb)	->	Q9D6D0 (chr 17 : 42.64 Mb)
PLA2G7 (46.67 Mb)	->	NM_013737 (chr 17 : 42.57 Mb)
MEP1A (46.76 Mb)	->	NM_008585 (chr 17 : 42.48 Mb)
O94858 (46.82 Mb)	->	ENSMUSG00000040779 (chr 17 : 42.45 Mb) ENSMUSG00000040993 (chr 17 : 42.41 Mb)
Q8TDT4 (46.98 Mb)	->	ENSMUSG00000041293 (chr 17 : 42.27 Mb)
TNFRSF21 (47.20 Mb)	->	Q9EPU5 (chr 17 : 42.01 Mb)
CD2AP (47.45 Mb)	->	O88903 (chr 17 : 41.79 Mb)

Navigate Homology Matches

Upstream (<44.92 Mb) Downstream (>47.60 Mb)

Figure 1. Screenshot of Ensembl syntenyview. An overview of the synteny between human chromosome 6 and mouse is shown. The central chromosome is the human chromosome and the surrounding ones are the corresponding mouse chromosome. Ensembl clusters DNA–DNA matches into syntenic regions eliminating any paralogous DNA–DNA hits. Each segment is clickable and displays a menu showing the chromosomal coordinates of the match on both organisms and two extra options which take you to the contigview of one or other of the genomes. On the right hand side is a list of the potential orthologous proteins between the two genomes in the region identified by the red rectangle in left hand panel.

pairs is extracted where one protein is the top BLAST hit in the other genome and vice versa. These are commonly known as reciprocal BLAST hits or seeds (5). An extra set is then found by looking in between the seeds on both genomes and finding extra pairs that have high similarity and also conserve gene order. These protein pairs are displayed on the right hand side of the syntenyview and in the geneview pages.

As more genomes become available, comparison data will be available across all the species. If the species are similar enough this will be done both at the protein and the DNA level. For more distant species this will be provided through clustering of the annotated proteins in the family database.

Apollo annotation viewer/editor

Apollo is a java based genome browser and editor and has been developed as a collaboration between Ensembl and the Berkeley Drosophila genome project (http://www.bdgp.org/). As well as providing a view of Ensembl data similar to contigview, the fine-grained DNA–DNA alignments and the protein level comparative data can also now be displayed. Figure 2 shows an example of this between human chromosome 20 and mouse chromosome 2. The protein level information is displayed with orthologous protein pairs linked in pink. Distributed Annotation System (DAS) support has also been added and allows the user to combine their local data with that available through DAS servers. As for the main Ensembl project, Apollo is an open source and is available from gmod.sourceforge.org.

EnsemblMart: data mining for genomes

With the advent of genome sequences for vertebrates many biological questions can be inverted from ‘how can I find the protein kinase involved in this process’ to ‘give me a list of all protein kinases that satisfy some criteria’. To facilitate such a query one needs a flexible, easy to use data-mining framework. To address this a new data-mining interface, EnsemblMart, has been developed that allows easy data mining via the martview web interface.

For example, one can dump 2000 bp upstream of each protein kinase containing domain on chromosome 1. The

Figure 2. Screenshot of Apollo. This is a detailed synteny view in Apollo with one genome displayed on the top and the other genome on the bottom. In the middle are links between orthologous genes as identified by the Ensembl synteny generating software. The genes from both genomes are read from Ensembl databases and the orthologous gene pairs can either be read from an Ensembl compara database or from a flat file. The central panel can be used to scroll back and forth along the syntenic region or to centre the display on a particular region by clicking on one of the coloured matches. The top and bottom Apollo panels behave just as a normal, single Apollo panel and can be zoomed, scrolled, collapsed and also link out to a web page (as configured by the user).

interface is aware of comparative genomic analysis calculated elsewhere in Ensembl, so one can for example also restrict this query by relationships to other organisms, such as the presence of a mouse ortholog. Similarly, one can also require the output to contain the corresponding mouse ortholog.

The underlying MySQL database for EnsemblMart is a query-optimised schema designed for quick queries and easy data discovery and is re-generated directly from the core Ensembl databases each time an organism database is updated. This database, along side all other Ensembl databases, can also be accessed directly from our open MySQL server at kaka.sanger.ac.uk, username anonymous. The same machine can alternatively be connected to via an ODBC interface, from programs such as Excel. These experimental, completely raw interfaces are mainly of use to bioinformaticians.

ENHANCEMENTS

Other than these major new features, there have been continuous enhancements (and also bug fixes) to Ensembl over the year. Users are recommended to read the what's new pages accompanying every release as frequently user interface improvements are subtle, but can save researchers considerable time. Some of the more significant improvements listed here.

Ensembl genome annotation

The overall principle of the three step Ensembl gene building system (2) remains the same, however the details of its implementation have been refined with each human genome release. The accuracy and coverage of the genes built by this automatic system continue to improve, as assessed by comparison with the gene structures of the current 'gold standard' of manually annotated and experimental verified finished human chromosomes 20 (6) and 22 (7). Sanger finished clones are initially annotated and updated by Sanger Institute vertebrate annotation group (http://www.sanger.ac.uk/HGP/havana/) and experimentally investigated by the experimental gene annotation group (http://www.sanger.ac.uk/Teams/Team69/). Ensembl works closely with these groups to integrate this annotation into Ensembl web displays and provide feedback.

An area where the existing Ensembl gene building system has been weak is in predicting alternative transcript forms, since the core gene building machinery does not rely on ESTs directly, as this results in too many false positives. To partly address this, a separate set of gene predictions are now being made, built entirely from human ESTs using the Ensembl EST GeneBuilder.

Ensembl maps ESTs to the genome using a combination of Exonerate, BLAST and EST2Genome. These are then processed by merging the redundant ESTs and setting splice-sites to the most common ends. This method finds the correct internal splice-sites, clusters 5′ and 3′ ESTs into UTRs and joins the fragments into longer transcripts structures. The resulting transcripts are processed by Genomewise, which finds the longest ORF across each one.

Alternative transcripts are predicted where there is at least one alternatively spliced EST and each EST gene has a supporting evidence page showing which ESTs have been used to construct it. At present, these EST genes are not classed as 'Ensembl genes' and therefore do not have Ensembl stable identifiers, however we are working to combine the EST and core gene builders in a way to increase alternative transcript coverage without decreasing gene prediction accuracy.

Ensembl web site

All interfaces have continued to be refined during the year, with probably the most development carried out to the 'workhorse' interface to genome sequence contigview. Refinements include toggle controls to switch between single-line and multi-line track displays; screen width configuration; contig orientation indicators and a gap type track. New tracks include an Eponine (8) track showing transcription start site predictions. Speed has been significantly improved by the use of the Ensembl-lite denormalised database that now provides much of the data for these pages and has been optimised for web queries.

New interfaces are the martview data mining interface (see above), goview and haploview. In goview we have integrated the standard GO browser from the GO consortium (9). GO is an ontology of gene function, process and location terms (e.g. 'protein phosphorylation' or 'cell cycle processing'). The GO data for human is inherited directly from SWISS-PROT GOAH work. The haploview interface provides access to haplotype data, currently available for human chromosome 22 (10). To access this data, turn on the haplotypes track in contigview and where haplotypes are shown, click on them to jump to haploview.

What is unlikely to be apparent to the user, are the underlying changes to the webcode to make it more multi-species orientated, which allow it to support all the species presented in Ensembl from a single codebase.

Finally, the integration of DAS (11,12) servers with the website has greatly increased. New DAS tracks on contigview include NCBI Transcript models, NCBI GenomeScan predictions, Acembly Transcript models and Ensembl mapped RefSeqs. Improvements have also been made to the interfaces to allow you to add DAS tracks from your own servers and to upload your own data directly for display. It is clear that the usage of DAS to integrate user data with our baseline annotation has increased greatly over the year.

Ensembl software system

Maintaining the circa 500 000 lines of code that supports and runs the Ensembl project is a major task in itself. Over 2002, Ensembl has transitioned to a revised schema and code base that principally has involved a more complete compliance to code and schema standards. For example, mixed case columns have been removed from the schema definitions and foreign key relationships are consistently named. In the code, the previous loose convention of separating the 'biological' objects from the database aware 'adaptor' objects is now consistent across the database with a consistent style of function name.

In addition to the Perl code base, there is a parallel Java code base with a common design between the two language bindings. As with the Perl code, the biological objects versus database aware adaptor objects is rigorously followed. The

Java layer is currently used for stable ID transfer and as a backend data adaptor for Apollo.

Ensembl data analysis pipeline

The data analysis pipeline has had a number of improvements, in particular the processing of ESTs and cDNAs as part of the EST Gene analysis. In addition, the work with the *A. gambiae* (Holt *et al.*, *Science*, in press) and *C. briggsae* project has introduced more configuration options to allow the pipeline to adapt to these invertebrate genomes. For example, the heuristics about maximum intron size have to be adjusted between vertebrates and invertebrates.

At the more technical level, the Ensembl pipeline system has improved its handling of complex data conditions which previously took manual work to fix; for example, areas of the genome which are almost complete masked by repeats, and so often triggered software errors in programs such as GenScan (13) when presented with high N content are now recognised and special processing rules applied. Different scheduler systems, such as PBS and GridEngine as well as LSF can now be used.

One innovation has been the compact storage of gapped alignments by storing the maximum extent of the matches and then a text string which encodes the placement of gaps inside the alignment. This text string format was first introduced in exonerate and represents the state path of the alignment process. Colloquially inside Ensembl this is called a 'cigar line' and its adoption has shrunk the number of rows in the feature table around 4-fold.

FUTURE DIRECTIONS

Ensembl remains focused on providing a genome information infrastructure of use to many researchers, principally via the web. As well as providing the baseline annotation for a number of genomes, Ensembl is continuously trying to improve all aspects of its work, from software engineering through to data analysis. Year 2003 promises a number of new genomes (in particular rat and zebrafish) but also technology improvements such as more complete comparative information spanning multiple vertebrates. Other projects include professional user level and technical documentation manuals and integration with rich ontology based systems such as GenomeKnowledgeBase (www.genomeknowledge.org).

CONTACTING ENSEMBL

Ensembl is a joint project of the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI), both of which are located on the Wellcome Trust Genome Campus, Cambridge, UK. To receive announcements about updates, subscribe to the 'announce' mailing list: majordomo@ebi.ac.uk 'subscribe ensembl-announce'. To follow the day-to-day development of Ensembl subscribe to the 'development' mailing list: majordomo@ebi.ac.uk 'subscribe ensembl-dev'. Requests for information and support can be sent to helpdesk@ensembl.org, which is a fully supported helpdesk. Extensive additional documentation can be found on the Ensembl web site, including installation guides and tutorials, both about using the software system and the web interface.

ACKNOWLEDGEMENTS

The Ensembl project is principally funded by the Wellcome Trust with additional funding from EMBL. We are grateful to users of our web site and the developers on our mailing lists for much useful feedback and discussion.

REFERENCES

1. Aparicio,S., Chapman,J., Stupka,E., Putnam,N., Chia,J.M., Dehal,P., Christoffels,A., Rash,S., Hoon,S., Smit,A. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, **297**, 1301–1310.
2. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
3. Birney,E., Clamp,M.E. and Hubbard,T.J. (2002) Databases and tools for browsing genomes. *Annu. Rev. Genome. Hum. Genet.*, **3**, 293–310.
4. Rust,A.G., Mongin,E. and Birney,E. (2002) Genome Annotation Techniques: New approaches and challenges. *Drug Discov. Today*, **7**, S70–S76.
5. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
6. Deloukas,P., Matthews,L.H., Ashurst,J., Burton,J., Gilbert,J.G., Jones,M., Stavrides,G., Almeida,J.P., Babbage,A.K., Bagguley,C.L. *et al.* (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature*, **414**, 865–871.
7. Dunham,I., Shimizu,N., Roe,B.A., Chissoe,S., Hunt,A.R., Collins,J.E., Bruskiewich,R., Beare,D.M., Clamp,M., Smink,L.J. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
8. Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
9. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
10. Dawson,E., Abecasis,G.R., Bumpstead,S., Chen,Y., Hunt,S., Beare,D.M., Pabial,J., Dibling,T., Tinsley,E., Kirby,S. *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, **418**, 544–548.
11. Hubbard,T. and Birney,E. (2000) Open annotation offers a democratic solution to genome sequencing. *Nature*, **403**, 825.
12. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.
13. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
14. Holt,R.A., Subramanian,G.M., Halpern,A., Sutton,G.G., Charlab,R., Nusskern,D.R., Wincker,P., Clark,A.G., Ribeiro,J.M., Wides,R. *et al.* (2002) The genome sequence of the malaria mosquito Anopheles gambiae. *Science*, **298**, 129–149.