D468–D470 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh038

Ensembl 2004

E. Birney[1], D. Andrews, P. Bevan, M. Caccamo, G. Cameron[1], Y. Chen[1], L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X. M. Fernandez-Suarez[1], P. Gane, B. Gibbins, J. Gilbert, M. Hammond[1], H. Hotz, V. Iyer, A. Kahari[1], K. Jekosch, A. Kasprzyk[1], D. Keefe[1], S. Keenan, H. Lehvaslaiho[1], G. McVicker[1], C. Melsopp[1], P. Meidl, E. Mongin[1], R. Pettett, S. Potter, G. Proctor, M. Rae[1], S. Searle, G. Slater[1], D. Smedley[1], J. Smith, W. Spooner, A. Stabenau[1], J. Stalker, R. Storey, A. Ureta-Vidal[1], C. Woodwark[1], M. Clamp and T. Hubbard*

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

Received September 16, 2003; Accepted September 18, 2003

ABSTRACT

The Ensembl (http://www.ensembl.org/) database project provides a bioinformatics framework to organize biology around the sequences of large genomes. It is a comprehensive and integrated source of annotation of large genome sequences, available via interactive website, web services or flat files. 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. The facilities of the system range from sequence analysis to data storage and visualization and installations exist around the world both in companies and at academic sites. With a total of nine genome sequences available from Ensembl and more genomes to follow, recent developments have focused mainly on closer integration between genomes and external data.

INTRODUCTION

Genome sequences provide a natural framework about which to organize biological data. In the short time in which they have been available, genome databases have proved invaluable resources to researchers. Ensembl provides one of the most popular sources of automatic analysis and integration of large genome sequence data and is a joint project between the EBI and the Sanger Institute. It now contains nine genomes: five vertebrates: human, mouse, rat, fugu, zebrafish; two worms: *Caenorhabditis briggsae* and *Caenorhabditis elegans* and two insects: *Drosophila melanogaster* and *Anopheles gambiae*. Ensembl has been involved in the continued analysis of human data, analysis of the mouse genome (1), analysis of the *A.gambiae* genome (2) and the *C.briggsae* genome. Ensembl gene predictions have also formed the core set of annotations for the forthcoming rat 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 to be used as a software foundation for user projects. Several papers describing specific aspects of Ensembl have recently been submitted (3–6). This paper briefly outlines some of the developments of the project since the report last year (7).

NEW DEVELOPMENTS

Regular update cycle

To streamline the handling of this ever changing and increasing amount of data, from February 2003, Ensembl adopted a monthly release cycle, allowing improvements to the web interface and database schema to be released monthly, with new data being incorporated as it became available. Database dumps and flat files are released in sync with updates to the website.

Pre-ensembl website

A full Ensembl annotation of a genome takes some weeks to complete. To provide users with immediate access to newly released genome assemblies Ensembl now offers a pre-ensembl website (http://pre.ensembl.org/) with limited functionality. This can be made available only a few days after the release of the genome and provides BLAST and SSAHA searching, placement of all known proteins, repeat masking and *ab initio* gene predictions.

Otter: an extended Ensembl schema for gene curation

During the year, Ensembl developed a new software component called Otter. Otter is an Ensembl database, but with an extended schema and an associated client/server system to support manual gene annotation. The Sanger Institute vertebrate annotation system is being migrated to use Otter, which will then put both automatic (Ensembl) and manual annotation under a single software framework and

*To whom correspondence should be addressed. Tel: +44 1223 494983; Fax: +44 1223 494919; Email: th@sanger.ac.uk

help greatly with subsequent data integration. The Otter server communicates with annotation clients via an XML format, which allows easy exchange and verification of annotation generated with different systems.

The Apollo genome browser (4), a GMOD component (http://www.gmod.org/) under joint development by Ensembl and the Berkeley *Drosophila* genome project (http://www.bdgp.org/), can be used as an annotation client for Otter. Apollo has also been extended to display data from DAS (distributed annotation system) servers. As an editor, Apollo has the advantage of being able to view and edit annotation in a comparative genomic context: by connecting to two Otter servers (e.g. human and mouse) and an Ensembl compara database containing pre-calculated synteny information between the two genomes, it is possible to view annotation for both genomes and edit each in the context of the synteny with the other.

ENHANCEMENTS

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

Ensembl genome annotation and comparative analysis

The quality of the annotation produced by the core automatic gene building system has continued to improve, with builds delivered on seven genome assemblies during the year. The most recent is the first version of the finished human genome sequence (NCBI33) announced in April, which also has pseudogenes automatically predicted. In parallel with gene building, comparative analysis is now routinely carried out for each new assembly. DNA synteny is generated between human, mouse and rat and putative gene orthologues between all five vertebrates and between each of the two worms and insects are automatically generated.

Ensembl website

Last year's move to the new schema enabled the development of significant enhancements to the Ensembl webviews. These include the addition of a fourth basepair level panel to Contigview, showing nucleotide, six frame amino acid translation and restriction enzyme site features. Additional pre-processing of SNP data during the building of the Ensembl-lite database (a denormalized database to speed web access), with respect to other annotation, has allowed Contigview, Transview and Protview to be extended to show SNPs against transcripts and their protein products, including labelling of synonymous and non-synonymous coding SNPs. Other enhancements to Contigview include labelled syntenic blocks shown on the overview panel and access to a new interface, Dotterview, from DNA conservation tracks on the detailed view panel. Dotterview is a web interface to the program Dotter, showing a dotplot of DNA similarity by default over a 10 kb window in two genomes, with Ensembl annotation. The interface for adding DAS (8) sources to Contigview has continued to be developed, giving the user much greater control over display of each source.

EnsemblMart: data mining for genomes

Ensembl has continued to import new externally generated data sets and resources into its system. These are frequently available in contigview via the DAS source menu; however, many are also being incorporated into EnsemblMart as additional data mining indicies. Examples include the STACK expression database eVOC nomenclature (collaboration with SANBI); rat QTLs and microarray identifiers from Affymetrix and others. All of these data types are queryable via the Mart data mining interface, which has increased substantially in functionality over the year and now has its own 'What's new' web pages and includes such functionality as integration with the ArrayExpress microarray repository at EBI.

Ensembl software system

The flexibility of components of the Ensembl software system are increasingly leading to their reuse elsewhere. Within the Sanger Institute alone, the Ensembl pipeline is being used to support gene curation by both the Wormbase and Havana (vertebrate annotation) groups. Havana is also in the process of making use of the Otter database for storing its gene annotation. The Ensembl website code has been reused to power the Vega website (http://vega.sanger.ac.uk/), which shows curated annotation of vertebrate genomes collected from a number of annotation groups into a single database. The fact that Ensembl data are also being served via DAS servers (8) is encouraging data to be combined in novel ways to provide specialist data displays. The website code has already been reused to build Contigview-like webviews of a virtual database composed entirely of different DAS sources.

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. 2004 promises a number of new genomes (e.g. chicken, chimp and honey bee) but also continued technology and presentation improvements, such as new views of cross-species data, organized around the putative gene orthologues predicted by the comparative analysis pipeline.

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 website, including installation guides and tutorials, about using both the software system and the web interface.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
2. 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.
3. Birney,E., Clamp,M.E. and Hubbard,T.J. (2002) Databases and tools for browsing genomes. *Annu. Rev. Genom. Hum. Genet.*, **3**, 293–310.
4. Lewis,S.E., Searle,S.M., Harris,N., Gibson,M., Lyer,V., Richter,J., Wiel,C., Bayraktaroglir,L., Birney,E., Crosby,M.A. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.
5. Hoon,S., Ratnapu,K.K., Chia,J.M., Kumarasamy,B., Juguang,X., Clamp,M., Stabenau,A., Potter,S., Clarke,L. and Stupka,E. (2003) Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res.*, **13**, 1904–1915.
6. Clamp,M. (2003) The Jalview Java Alignment Editor. *Bioinformatics*, in press.
7. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
8. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.