The EnsEMBL Project – Open Source in Bioinformatics

Genomes for All

In the search for new medicines, the genetic information about humans and animals is a source of information we are just tapping. An Open Source project called EnsEMBL helps in useful work, researching the causes and helping us understand the symptoms of medical conditions.

BY STEFFEN MÖLLER AND LUCA TOLDO

Pharmacology calls upon many ways to produce the drugs of the future. One typical procedure sounds quite simple: in endless test series, researchers splice chemical substances into test tube grown cell cultures, observing the results. If there are any results, some kind of interaction between the cell proteins and the candidate must have taken place. Bioinformatics ensure that this kind of test is not purely random, by storing information on known proteins, predicting the existence of new proteins and even endeavors to describe their characteristics.

Proteins are products of our genes. They are encoded in cells as a sequence of four distinct nucleic acids, which are typically abbreviated to A, C G and T. The process of ascertaining this sequence is referred to as genome sequencing, a topic that has been in the headlines quite frequently.

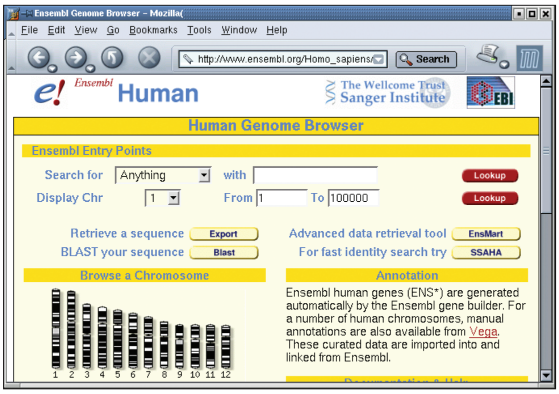
Human genome sequencing was officially declared "finished" on April 14th this year. The human genome was sequenced twice, in fact: once by a US company called Celera, which now markets this data, and by an international network of academic sequencing centers. The Sanger Centre in Cambridge [4, 10], sponsored by the Wellcome Trust, contributed the lion's share of the European sequencing effort.

The software used to manage, collate and present the tremendous amounts of data involved is as Open Source as the data itself: EnsEMBL [1], a project by the Sanger Centre in co-operation with the European Bioinformatics Institute (EMBL-EBI) [2, 3]. Apache, MySQL, Perl and increasingly, Java have been used to implement the project.

The source is distributed via CVS, and database dumps via FTP. In addition to the human genome, data is also available for zebra fish, mouse, rat, worm, fruit fly, mosquito and fugu, an Asiatic blowfish. These are the genomes of more complex organisms closest to completion. They were selected on account of their significance in the lab, with regard to understanding a single disease, malaria, from a parasite carried by the mosquito, plasmodium falciparum, which has also been sequenced, and from the blowfish because of its extraordinary cytological characteristics.

Using EnsEMBL

The user first has to select an organism to investigate. To navigate the program, one then searches for a gene or protein with specific characteristics, or selects a chromosome. From this point one can display equivalent chromosome areas in different organisms.



Ensembl Genome Browser – Mozilla(
File Edit View Go Bookmarks Tools Window Help
Search
e! Ensembl
The Wellcome Trust Sanger Institute
EBI
1
2
3
4
5
6
7
8
9
10
11
12

http://www.ensembl.org/Homo_sapiens/

Human

Human Genome Browser

Ensembl Entry Points

Search for Anything with Lookup

Display Chr 1 From 1 To 100000 Lookup

Retrieve a sequence Export

BLAST your sequence Blast

Advanced data retrieval tool EnsMart

For fast identity search try SSAHA

Browse a Chromosome

Annotation

Ensembl human genes (ENS*) are generated automatically by the Ensembl gene builder. For a number of human chromosomes, manual annotations are also available from Vega. These curated data are imported into and linked from Ensembl.

Figure 1: The EnsEMBL entry point for the human genome. Clicking on a chromosome allows you to browse the genome

Comparing the sequences reveals that both genomes are extremely similar. At some stage there must have been a common ancestor, a mammal from which both humans and mice evolved. This is why results from experiments with mice may be applied to humans.

The "ContigView" (see Figure 2) is the most important view type. A "contig" is a contiguously known area of a chromosome. The areas outside the contigs are either unsequenced, or the sequence has not been reproduced often enough to confirm it.

In the ContigView researchers can find references to information in other databases, describing the molecular basis of a disease, or providing functional and structural classifications of proteins. Data can be exported from the system either in image formats or spreadsheets.

Sequencing the genome reveals DNA sequences without any further annotation. Only a small proportion of the DNA sequence actually codes for protein. The function of the vast majority of genomic DNA, while far from being understood, is assigned to the regulation of gene expression. Investigating the location of coding areas is particularly interesting and one of the main focuses of the EnsEMBL founders' research activities. This involves a combination of multiple algorithms [7], some of which concentrate on the DNA sequence, and others use alter-native data sources, such as the protein database, SWISS-PROT [5].

An Adaptable System

No matter how competent and quick the EnsEMBL developer team may be, there is always something that some research group needs implemented immediately. Fortunately, this is no problem, thanks to Open Source and the use of standard technologies.

EnsEMBL is also an extremely useful knowledge base for commercial enterprises. A local EnsEMBL installation in the LAN of a research company can help to protect intellectual property and allows for interoperability with internal databases. An interface to the Distributed Annotations System (DAS) [6] is provided for the latter, although this often requires a source code extension.

EnsEMBL's MySQL databases are quite large occupying a total of 65 GBytes in the author's local installation. One can also expect the amount of information available to continue growing, even should no new organisms be added, as genome descriptions continue to improve.

At present there is no mechanism available to provide partial updates of the knowledge base. The task is difficult as it involves handling the growth of the knowledge base with respect to the amount of information it contains and changes to the database schema. However, the EnsEMBL has made a MySQL server available for public use via the Internet for any researchers who want to avoid the need for local updates.

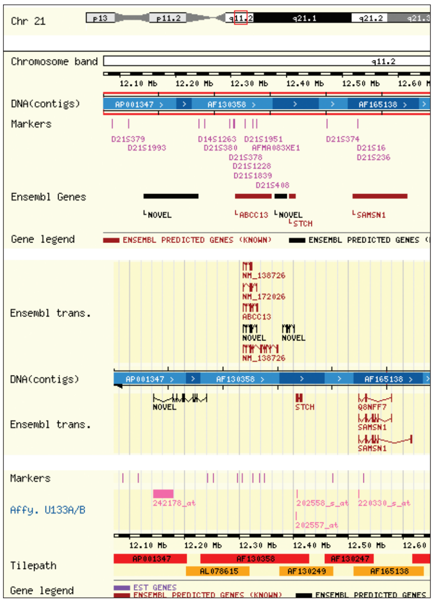


Figure 2: Section from the ContigView showing the position of genes on the chromosome

In projects of this kind, where computer scientists need to work hand in hand with scientists from other disciplines, documentation is particularly important. EnsEMBL provides a useful online help system, and researchers needing to extend their local system can always contact the extremely helpful development team by email or via the mailing list to request support.

Conclusion

The availability of the source code makes EnsEMBL a valuable tool for any Bioinformatics service department and many research groups.

For over two years the system has served multiple gigabytes of data, and MySQL has proved its value as a robust tool. A run-of-the-mill computer with 256 MBytes of main memory, a 50 GByte hard disk and Linux is all you need to run EnsEMBL locally. The MySQL database for the human genome weighs in at some 16 GBytes, and the mouse runs to about eight.

This opens up a world of opportunities for biology teaching, amateur researchers, or even the next generation of research scientists, allowing them to familiarize themselves with modern genetic research methods. ■

Gen-Sections and Disease

EnsEMBL can help to understand genetically caused disease, primary role is to assist in the generation of hypotheses. Lab tests and clinical research are always required to provide proof. Some diseases are accompanied by translocation of part of a chromosome, and may be visible using a microscope.

For example, it is known that MDS (myelodisplastic syndrome) [8], a precursor to a particular form of leukaemia (blood cancer), is accompanied by the translocation of part of chromosome 11 to chromosome 21. However, it is non-trivial to discover the genes that are affected, and why it causes cancer. It is also unknown whether this collates with the results of other research into MDS.

This is where EnsEMBL steps in, as it allows the user to select genes by reference to their position on the chromosome. Selections based on other features, such as a known link to a specific disease, are also possible [9].

Searching for "Myelodisplastic syndrome" returns genes, but not within the affected chromosomes, 11 or 21.

This means closer searching of the appropriate regions in the ContigView. And doing so reveals the gene HSPA8 in region 11q24 which is known to control growth genes. Uncontrolled growth is typical for cancer. Our first hypothesis would be that translocation changes the extraction rate of this gene, thus making appropriate control of cell growth difficult.

Chromosome 21 also provides a few hints. Genes NRIP1 and SAMSN1 in region 21q11.2 also catch the eye, as their involvement in other cancer types has already been demonstrated. EnsEMBL additionally shows that this region of chromosome 21 contains, or may contain, genes that no-one has found before ("novel" genes). A specific investigation of the region could result in the discovery of a new therapeutic approach.

INFO

[1] EnsEMBL: *http://www.ensembl.org*

[2] European Molecular Biology Laboratory: *http://www.embl-heidelberg.de*

[3] EMBL-EBI, European Bioinformatics Institute: *http://www.ebi.ac.uk*

[4] Sanger Centre: *http://www.sanger.ac.uk*

[5] Protein database, SWISS-PROT: *http://www.expasy.org/sprot/*

[6] DAS interface: *http://www.biodas.org*

[7] Genewise Algorithms: *http://www.cgen.com/products/genewise.htm*

[8] MDS Causes: *http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=3DR= etrieve&db=3DPubMed&list_uids=3D10451699&dopt=3DAbstract*

[9] Online Mendelian Inheritance of Man: *http://www.ncbi.nlm.nih.gov*

[10] Sanger Centre Press Release: *http://www.sanger.ac.uk/Info/Press/2003/030414.shtml*

THE AUTHORS

Steffen Möller is a research scientist at the University of Rostock's Proteom Center. Luca Toldo is involved in R & D for Merck's Scientific Information Systems / Biological Services department, located at Darmstadt, Germany.