

Shopping in the genome market with EnsMart

Abstract

Life scientists who work with the super-market of genome data will find the EnsMart database and software package (www.ensembl.org/EnsMart/) offers a valuable door to a wealth of genes and genome features. This popular multi-organism genome database can be installed and used on your own Unix computer with relative ease. It offers a flexible, fast and practical data-mining framework for computer-savvy biologists and bioinformaticians.

Keywords

Genome databases, human genome, comparative genomics, data mining, open source software

Introduction

Bioinformatics has a strong focus on the rich and growing datasets of life sciences. With genome information in particular, a transition is underway from common use of data files to an essential need for integrated database software¹. Integration provides the only practical way to extract knowledge, and answer biological questions of these diverse data. Today biologists seek primary genome information from among many genome web interfaces. Genomics web portals often lack methods for effectively mining large subsets of genomes, or are limited in the questions one can pose to the underlying complex data. The Ensembl project^{2,3} provides a new breed of integrated software and data to bridge the gap between bulk files and web portals.

Ensembl is a recognized leader in providing genome annotation of human and other eukaryote genomes. The project provides its work as portable, open-source software and data that can be copied and installed on other computers. EnsMart is a self-contained addition to Ensembl software and data, providing access to commonly used parts of this genome data. Species genomes collected in EnsMart (version 14.1) include human, mouse, rat, zebrafish, fugu, fruitfly, mosquito, *C. elegans*, and *C. briggsae*, and comprise some 30 gigabytes of data (the largest portion in DNA sequences). EnsMart provides fast genome information search and retrieval for human and several other popular eukaryote species. Its query building operations allow one to select data by species, cross-species homology, by gene expression, by location on genomes and regions around genes, as summarized with examples in Table 1. One can retrieve data as tables for spreadsheets and databases, as Fasta sequence format, and other forms. The public web site at www.ensembl.org/EnsMart/ provides a super-market for easy selection and retrieval of complex genome data.

A new ability in this version allows one to copy and use EnsMart locally, including its web interface and databases. Many scientists will find the public web site provides for their needs, without requiring computing expertise to use. But if you want more than the public

web site offers, a few hours of work by someone familiar with common informatics techniques will provide your own copy of EnsMart. Benefits are apparent if you make frequent use of Ensembl data, integrate it with your own data, want to make queries not offered at the web site, mine data in large quantities, and want to avoid slowness or out-of-service at www.ensembl.org, such as encountered during review. This project offers a good paradigm for genome informatics: integrated software and databases that can be copied, intelligently queried and combined with one's own data to extend beyond web services provided at the source.

Table 1: EnsMart features and example uses⁴.

<i>Data integration includes</i>	<i>Results provided include</i>
<ul style="list-style-type: none"> ○ Gene and protein annotation ○ Disease information ○ Expression data ○ Sequence variation ○ Cross-species analyses 	<ul style="list-style-type: none"> ○ Homology ○ SNPs affecting proteins ○ Retrieval by external identifiers ○ Retrieval by expression vocabulary ○ Customised sequence datasets ○ Microarray annotation tools
<i>Example uses of EnsMart</i>	
<ul style="list-style-type: none"> ○ Mouse homologues for human disease genes. ○ Coding SNPs for all novel kinases. ○ Genes on chromosome 1 expressed in liver. ○ Ensembl genes mapped to RefSeq identifiers. ○ Upstream sequence for all Ensembl genes mapped to U95A chip. ○ Disease related genes between markers (eg D10S255 and D10S259). ○ Transmembrane proteins with an Ig-MHC domain (IPR003006) on chromosome 2. ○ Genes with associated coding SNPs on chromosomal band 5q35.3. ○ Novel GPCRs with nonsynonymous SNPs. ○ Rat homologues of human disease genes expressed in brain. ○ Complete genomic annotation of mouse MG_U74 Affymetrix chip. ○ Genomic location and description of all mouse, rat and fugu homologues of all human genes, which have transmembrane domains, are expressed in cardiovascular system and have non-synonymous snps 	

The EnsMart query process operates in three steps: Select a genome focus, filter these by criteria, and output formatted results. First you select the primary result of interest: species genome and its gene, EST gene or SNP contents. Next, filter the available set of these to satisfy your questions, by choosing criteria among genome region, known genes or user-specified lists of gene ID, where these are expressed in anatomy and development stage, homology to other species, protein domains and SNP attributes. Finally decide on results output of features, structures, SNPs, and sequences, including IDs from Ensembl and many other databases, protein and microarray attributes, disease associations, species homologies, as well as file formats suited to your spreadsheet, database or other uses.

Details of use

Documentation provided with EnsMart includes on-line instructions and tutorials, along with installation instructions, outlining the program, and its operations. Although installing your own copy requires computing skills, these are not extensive. An introductory book such as Gibas and Jambeck⁶ covers the skills needed. If you find installation instructions for EnsMart are obscure, a bioinformatics or computing student or staff will be able to help.

Basic open-source software components and requirements include MySQL database, Perl language and several library packages, Apache web server, and CVS source code management tool. You will need a modern Unix computer (including Linux and MacOSX) with 30 to 60 gigabytes of free disk, depending on datasets you install. If this has a current operating system, most of the prerequisite software will be included; remaining parts can be added from Internet sources per instructions.

The EnsMart software package employs simple web forms for searching combined with a full set of help, data and software downloads. Its standard operation is as web server where you fill out forms in a Query Builder with your web browser, and then download results, choosing among multiple result formats including MS Excel, HTML or tabular text. One can access the database directly with network programs that talk the database language SQL. With additional software, one can use the in-development MartExplorer, or new Web object access programs that use the SOAP standard. This can be useful for organizations developing integrated web-based bioinformatics services.

EnsMart software and database (version 14.1 reviewed June 2003) was tested on a current Mac OS X computer, where it performed properly. The package is supported for Linux and other Unix systems also. Data is available by Internet file transfer from source and world mirror sites⁵. Loading this data into MySQL is straightforward, taking just a few minutes of personal time (hours for the computer), and can be automated for updates. Installation of the instructed EnsMart data filled 20 GB, rather than 60 GB as described, though this amount would be needed for all Ensembl genome data.

Software is installed using the concurrent versions system (CVS) popular among developers. During review, the instructed installation of EnsMart failed. One problem with CVS for basic software installation is that it lacks browsing and other methods to determine how to get around an error. As fallback, I used the instructions for installing the full Ensembl software set, including EnsMart, which succeeded. After reporting the error, the Ensembl staff corrected it and the EnsMart software CVS installation worked properly.

Instructions given for configuring your local installation are brief, and proved not quite complete. After editing a half-dozen configuration lines per instructions, the first trial of EnsMart web server failed. Adding and removing entries in the Apache server configuration solved this problem. This database software system, while complex, is well constructed by the project members. It is smart in determining which databases and tables are installed: the initial invocation causes an extensive check and reporting of available and missing data sets. If you later add new sets, you may need to re-invoke this configuration check. After tests, I was able to remove uninteresting organisms from the data set without causing problems.

Following installation and invocation, the local copy offered web forms and choices essentially identical to the main EnsMart server (though labeled as version 13). The queries done locally returned the same results as for the version 14 at the Ensembl.org site. This web interface is well suited to providing an understandable, widely usable set of operations, but it won't make all data miners happy due to common limitations of web interfaces.

A tool in development called MartExplorer is one that data-miners will appreciate. It allows programs to search and retrieve from these databases directly, bypassing the web server and browser. It will also allow the addition of new database tables, to combine one's project data with EnsMart. MartExplorer includes example application programs, command-line tools and a basic graphic interface, so those with basic software skills can interface their tools and databases. It provides for automated uses and scripting, offering more flexibility than a web interface. It can be programmed in Java and Python languages, and runs on MS-Windows as well as Unix and other Java enabled systems. In less than a half hour, I was able to copy the development version of MartExplorer from Ensembl.org, compile it with a standard Java build tool, launch 'martgui', and retrieve data from a local database. Though it has not yet been officially released, and work remains to flesh out MartExplorer, it is a good start for mining genomes available in EnsMart.

Discussion

EnsMart provides bioinformatics and bioscience researchers with usable access to a large, changing volume of popular genome data. Drawing on Ensembl databases and software, it provides simple but sophisticated ways to answer questions that combine aspects of genome regions, gene homologies, expression, protein actions, and nucleotide polymorphisms needed in genome research. This collection is part of a wider range of information essential to genome research, which as yet requires one to traverse many web portals, each with distinctive methods and values, for answers to complex biological questions^{1,7}. The approaches taken with EnsMart promise a route to effective integration of genome data with one's project needs. Others are taking advantage of the robust Ensembl software for genome data management such as the Gramene project⁸ for rice and grass genomes.

EnsMart is most useful when used in conjunction with other genome analysis tools. Data drawn from it are ready for spreadsheet, database and BLAST analyses. Visual mapping tools such as VISTA and PipMaker⁹ can produce maps with data from EnsMart, to aid design and analysis of genomics experiments. EnsMart includes less specialized visualization tools, but can be integrated with the Apollo¹⁰ visual genome browser and annotation editor. One can install and use both on a workstation, combining one's project results with Ensembl source data.

There are a few areas where additions and improvements to EnsMart could make it more useful. An optional package of the released software for FTP download would make it easier to install, where CVS access can prove problematic and awkward for those unaccustomed to it. Providing a Perl bundle would ease the installation of required Perl modules. Missing from documentation is an explanation of how to update EnsMart after you install it. Software distributed by CVS provides for near automatic updating. Data however need to be updated in MySQL using steps similar to a first installation. Some features of

EnsMart use the full Ensembl databases for fetching species data, using HTML links that return to Ensembl.org for details. If you find local use of EnsMart helpful, installing full Ensembl genome sets follows a similar procedure.

How valuable is having your own genome market with EnsMart? If you would like to use quantities of genome data in your work, it offers new levels of access and integration. EnsMart also is a learning and discovery tool for bioscientists, providing an easily learned query/retrieval interface to complex, integrated multi-organism genome data. Those who master it and find rewarding results may well want to move to the next step for answering questions more complex than web portals allow. MartExplorer software will help one take this step. Computer-using life scientists in growing numbers will use EnsMart to reap the benefits of knowledge hidden in these gigabytes of genome data.

Acknowledgments. Arek Kasprzyk of the Ensembl project provided helpful comments. This work was supported in part by NSF grant 0090782 and NIH grant 1R01HG002733-01 to D. Gilbert.

*Don Gilbert
Biology Department, and
Center for Genomics and Bioinformatics
Indiana University,
Bloomington, Indiana 47405 USA
Tel: +1 812 855 0587
E-mail: gilbertd@bio.indiana.edu*

References

1. Stein, L. 2003. Integrating Biological Databases. *Nature Reviews Genetics*. 4: 337-345. doi:10.1038/nrg1065
2. Clamp, M., *et al.* (2003). Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Research*, 31: 38–42. doi: 10.1093/nar/gkg083
3. Hubbard, T. *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Research*, 30: 38-41
4. EnsMart (URL: <http://www.ensembl.org/EnsMart/>)
5. EnsMart data (URL: ftp://ftp.ensembl.org/pub/current_mart/. US mirror: [ftp or rsync://bio-mirror.net/biomirror/ensembl/current_mart/](ftp://rsync://bio-mirror.net/biomirror/ensembl/current_mart/))
6. Gibas, C and Jambeck, P. (2001). Developing Bioinformatics Computer Skills. O'Reilly & Assoc., Sebastopol, CA, USA. 427 pp.
7. Valencia, A. (2002). Search and retrieve: Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? *EMBO Reports* 3(5): 396 - 400. doi: 10.1093/embo-reports/kvf104
8. Ware, D. H. *et al.* (2002). Gramene, a Tool for Grass Genomics. *Plant Physiology*, 130: 1606 - 1613. doi: 10.1104/pp.015248

9. Pennacchio, L.A. and Rubin, E.M. (2003). Comparative genomic tools and databases: providing insights into the human genome. *J. Clin. Invest.* 111:1099–1106. doi: 10.1172/JCI200317842.
10. Lewis, S. E. et al. (2002). Apollo: a sequence annotation editor. *Genome Biol.* 3, R0082.1–R0082.14.