PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

The desktop genome

ArticleInfo		
ArticleID	:	3881
ArticleDOI	:	10.1186/gb-2000-2-1-reports2001
ArticleCitationID	:	reports2001
ArticleSequenceNumber	:	16
ArticleCategory	:	Web report
ArticleFirstPage	:	1
ArticleLastPage	:	5
ArticleHistory	:	RegistrationDate : 2000–11–24 Received : 2000–11–24 OnlineDate : 2000–12–22
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	

ArticleContext	:	130592211

Colin Semple

Abstract

The Ensembl software automatically produces and archives annotation of the draft human genome sequence.

Content

The Ensembl software automatically produces and archives annotation of the draft human genome sequence. Launched in 1999, this site aims to produce 'baseline' annotation for genomic sequence data, including unfinished 'draft' sequence. In other words, it aims to tell you where genes and other important landmarks are in a genome, starting with the human genome. Given the fragmented, incomplete state of most of the human genome, Ensembl has undoubtedly taken on a heroic task. Known genetic markers and repetitive sequences are detected within each genomic sequence. Each sequence is also submitted to a variety of computational analyses to find any genes present. First, each sequence is submitted to the gene prediction program Genscan, which identifies putative coding exons on the basis of various statistical analyses. In essence, the program looks for regions that match the composition, lengths and other features of known exonic, intronic or intergenic sequence. The exons identified are translated into peptides, and used to search online protein and expressed sequence databases using BLAST. Those exons matching a database sequence are considered to be real or 'confirmed'. Confirmed exons are then combined into predictions of gene structures according to certain rules. Alternatively, if the exons show sufficiently high similarity to a known protein, the gene structure is predicted by another program called Genewise, which uses these similarities to predict more accurate gene structures.

And now the caveats. Genscan can detect most (perhaps between 70 and 90%), but not all, of the genes present in human genomic sequences, and by no means all novel genes show significant similarity to previously known sequences. This means that the set of confirmed genes in Ensembl is necessarily a conservative one. This problem may be further exacerbated by the quality of the sequence submitted. Ensembl uses the Human genome project working draft sequence assemblies (or 'golden path') produced at the University of California at Santa Cruz, which are known to contain many gaps and misassemblies. The golden path appears to be particularly unreliable in regions where it is composed of many small genomic sequence fragments from recently sequenced bacterial artificial chromosomes (BAC) clones. Many of these fragments are assembled in the wrong order and/or orientation. Such misassemblies will be a further source of errors and omissions in Ensembl. Although Genscan can detect the presence of most genes, it is substantially less successful in predicting their correct exonic structures (as with other *ab initio* gene predictions). This means that many, if not most, of the gene structures in Ensembl will be incorrect, or 'partial' predictions in Ensembl parlance.

In spite of these difficulties Ensembl remains a useful tool for the cautious biologist. It should detect the presence of most genes in a given fragment of genomic sequence and indicate their location in the genome on the basis of the best mapping data available. In addition it has a stab at predicting gene structures that should be accurate if the gene in question has a close homolog which is already known. Most aspects of the analysis Ensembl carries out are the subject of active research, so improvements in performance, as a result of the inclusion of new sequence data and algorithms, will be ongoing. Having secured major funding earlier this year, the database promises to become the most important source of annotation for the draft human genome.

Navigation

Substantial thought and effort has gone into the Ensembl site design. The result is certainly user friendly, and not just by the standards of computational biology. The web interface to the database achieves the laudable aim of providing seamless access to the draft genome. The user can sink down through cytogenetic ideograms of whole chromosomes, to large unfinished sequence contigs several megabases long and then smaller fragments of individual BAC clones only kilobases long. Along the way, a graphical display shows the relative positions of predicted exons, sequence similarities, single-nucleotide polymorphisms (SNPs) detected by the SNP Consortium and repetitive sequences. A recent addition to Ensembl is the ability to browse disease genes that are included in Online Mendelian Inheritance in Man (OMIM) and found in the draft genome sequence. Data retrieval is well catered for: text searches of all entries, BLAST searches of all sequences archived and bulk downloads of all Ensembl data are possible. All Ensembl gene identifiers should remain constant between data releases.

Reporter's comments

Timeliness

The present release '07' is based on the golden path assemblies from 15 June 2000.

Best feature

Omitting finished chromosomes, Ensembl is the only publicly available attempt to annotate the human genome sequence. In a sellers' market such as this, it is commendable that so much effort has been expended in making the interface to the database as friendly as possible. The user is led through series of graphical viewers that make this large and complex dataset readily interpretable to any biologist with a web browser. The Ensembl philosophy of making all annotation for the draft genome

publicly available is mirrored in the approach to software development; all source code from the project is freely available.

Worst feature

The Ensembl system is a work in progress and makes no claims to be a source of flawless data; in certain cases Ensembl annotation can be perplexing. Ensembl appears to be easily confused by genes that have similar pseudogenes. If one searches Ensembl for unmapped gene X, which is very similar (say 95% identical) to unmapped pseudogene Y, then assuming they are both present in Ensembl, it is difficult to distinguish the two on the basis of Ensembl data. Fundamentally this is because Ensembl does not provide the full BLAST alignments that the confirmed genes depend upon. This omission may have naive users floundering for some time. The best generic solution at the moment is to get hold of the two suspect Ensembl genes, align them with the sequence of gene X and assume that the Ensembl gene with the worst match to gene X is the pseudogene.

Wish list

Updates of the Ensembl database have been erratic. Since the latest Emsembl update in June, there have been three further golden path assemblies, with the current one released on 7 October, 2000; during this time a further 3,982 clones have been incorporated into the golden path, reducing the number of assembled contigs by 188,302. It would be nice if this time lag between the golden path and Ensembl could be narrowed. With mouse and other eukaryotic genomes on the horizon, the opportunity will arise for Ensembl to provide a friendly interface for comparative genomics. Ensembl will soon process mouse sequence and it will be interesting to see how capabilities for comparative analysis evolve.

Related websites

An analogous annotation system to Ensembl has been developed at Oak Ridge National Laboratory. It is based on different gene prediction software and is called The genome channel. It catalogs genomic annotation for a number of organisms; it has so far processed only a fraction of the human sequence data of Ensembl, however. The National Center for Biotechnological Information (NCBI) *Homo sapiens* genome view provides an interface to view human draft sequence contigs produced at the NCBI, indicating the relative positions of known markers and clones. It will also provide gene predictions in the future. Should you wish to develop your own annotation, the Human genome central website provides a good summary of the available data sources for the draft genome.

Table of links

Ensembl

Genscan

BLAST

Genewise

Human genome project working draft

SNP Consortium

Online Mendelian Inheritance in Man

The genome channel

Homo sapiens genome view

Human genome central

References

1. Ensembl.