Biological information: making it accessible and integrated (and trying to make sense of it)

Tim Hubbard

Sanger Institute, Cambridgeshire, UK

Received on April 8, 2002; accepted on June 15, 2002

ABSTRACT

The availability of the genome sequences of human and mouse, human sequence variation data and other large genetic data sets will lead to a revolution in understanding of the human machine and the treatment of its diseases. The success of the international genome sequencing consortia shows what can be achieved by well coordinated large scale public domain projects and the benefits of data access to all. It is already clear that the availability of this sequence is having a huge impact on research worldwide.

Complete genome sequences provide a framework to pull all biological data together such that each piece has the potential to say something about biology as a whole. Biology is too complex for any organisation to have a monopoly of ideas or data, so the collection, analysis and access to this data can be contributed to by research institutes around the world. However, although it is possible for all this data to be accessible to all through the internet, the more organisations provide data or analysis separately, the harder it becomes for anyone to collect and integrate the results. To address these problems of intergration of data, open standards for biological data exchange, such as the ‘Distributed Annotation System’ (DAS) are being developed and bioinformatics (Dowell et al., 2001) as a whole is now being strongly driven by the open source software (OSS) model for collaborative software development (Hubbard and Birney, 1999). The leading provider of human genome annotation, the Ensembl project (http://www.ensembl.org), is entirely an OSS project and has been widely adopted by academic and commercial organisations alike (Hubbard et al., 2002).

Accurate automatic annotation of features such as genes in vertebrate genomes currently relies on supporting evidence in the form of homologies to mRNAs, ESTs or protein. However, it appears that sufficient high quality experimentally curated annotation now exists to be used as a substrate for machine learning algorithms to create effective models of biological signal sequences (Down and Hubbard, 2002). Is there hope for ab initio prediction methods after all?

REFERENCES


