

Meeting report

Gene prediction: the end of the beginning

Colin Semple

Address: Department of Medical Sciences, Molecular Medicine Centre, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK.
E-mail: Colin.Semple@ed.ac.uk

Published: 28 July 2000

Genome **Biology** 2000, **1**(2):reports4012.1–4012.3

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2000/1/2/reports/4012>

© Genome**Biology**.com (Print ISSN 1465-6906; Online ISSN 1465-6914)

A report from the conference entitled Genome Based Gene Structure Determination, Hinxton, UK, 1-2 June, 2000, organised by the European Bioinformatics Institute (EBI).

The draft sequence of the human genome will become available later this year. For some time now it has been accepted that this will mark a beginning rather than an end. A vast amount of work will remain to be done, from detailing sequence polymorphisms to discovering the complexities of the transcriptome - the totality of sequences transcribed - and, ultimately, the proteome - all the proteins encoded by the genome. All of this work will, to a greater or lesser extent, depend on all the genes having been correctly identified. It will be necessary to document not only the coding exons of each gene but also non-coding exonic sequence and regulatory sequences. As this conference made clear, however, the production of genomic sequence has outstripped our ability to reliably predict such features computationally.

Traditionally, gene prediction programs that rely only on the statistical qualities of exons have been referred to as performing *ab initio* predictions (from the Latin: from the beginning). *Ab initio* prediction of coding sequences is an undeniable success by the standards of the machine-learning algorithm field, and most of the widely used gene prediction programs belong to this class of algorithms. It is impressive that the statistical analysis of raw genomic sequence can detect around 77-98% of the genes present, which was the range of sensitivity reported at the conference. This is, however, little consolation to the bench biologist, who wants the complete sequences of all genes present, with some certainty about the accuracy of the predictions involved. As Ewan Birney (European Bioinformatics Institute, UK) put it, what looks impressive to the computer scientist is often simply wrong to the biologist.

Reducing genomes to genes

All *ab initio* gene prediction programs have to balance sensitivity against accuracy. It is often only possible to detect all the real exons present in a sequence at the expense of detecting many false ones. Alternatively, one may accept only predictions scoring above a more stringent threshold but lose those real exons that have lower scores. The trick is to try and increase accuracy without any large loss of sensitivity; this can be done by comparing the prediction with additional, independent evidence. For example, one may increase confidence in a predicted coding exon by detecting the presence of a sequence within it which codes for a known protein domain. The patterns made in DNA sequences by such domains may be detected using probabilistic models known as hidden Markov models (HMMs). Predictions for exons that contain non-coding sequence or untranslated regions (UTRs) can be refined by comparison with ESTs (expressed sequence tags) - sequences representing fragments of mRNA sequence that include coding and/or UTR sequence. The latest generation of gene prediction programs take advantage of such similarity-based approaches to complement *ab initio* predictions. For example, Ed Uberbacher (Oak Ridge National Laboratory, USA) unveiled the latest incarnation of the Grail program - GrailEXP [<http://grail.lsd.ornl.gov/>] - which uses EST data to find both UTR boundaries and short exons, typically problematic areas for *ab initio* predictions. Anders Krogh (Centre for Biological Sequence Analysis, Denmark) also showed improvements in accuracy for his gene prediction program HMMgene [<http://www.cbs.dtu.dk/services/HMMgene/>] when similarity-based evidence was incorporated.

Regulatory regions in the human genome are estimated to occupy ten times the sequence length of coding sequences. Prediction of regulatory sequences remains troublesome as they are invariably short sequences matching a rather vague consensus pattern that arise frequently by chance in genomic sequence. Thomas Werner (GSF - National

Research Centre for Environment and Health, Germany) outlined a novel approach to circumventing the low sequence conservation of functionally equivalent promoters. He treats promoter regions as clusters of small, locally conserved sequence motifs or 'modules'. It would seem that there are specific restrictions on the spacing and ordering of modules within a promoter region, imposed by the requirements of the regulatory protein complexes that bind there. His program PromoterInspector [http://genomatix.gsf.de/free_services/] uses such restrictions to increase the accuracy of promoter prediction, achieving impressive results over large genomic sequences such as human chromosome 22, where it reached levels of specificity of more than 98% (that is, it generated less than 2% false positives when compared to the published annotation). As with gene prediction, however, there is a trade-off with sensitivity, and only around 33% of known promoters were found. Michael Zhang (Cold Spring Harbor Laboratory, USA) has begun the task of incorporating regulatory motif detection into his gene prediction software MZEF. As with many other programs, MZEF is most successful when predicting internal, coding, exons, so the idea is to tailor new programs to other classes of exon. For instance, models of initial exons could incorporate upstream promoter regions and final exon models could include poly(A) addition sites.

Comparative genomics may be the unexploded bomb in gene structure prediction, capable of sweeping away many of the ambiguities in human gene predictions. Regions of sequence conserved between species can reveal novel coding sequences and, more importantly, non-coding features that could not otherwise be detected. Mikhail Gelfand (Centre for Biotechnology, Russia) outlined strategies for finding regulatory regions by comparison of bacterial species. He showed how the discovery of regulatory elements for heat-shock protein genes allowed the detection of such elements at other loci and consequently the detection of novel co-regulated genes that may be involved in the heat-shock response. The mouse genome sequence will be available within a year or two and will doubtless provide a popular resource for comparisons with human sequence, particularly in detecting regulatory elements and refining exon boundaries. But more than one speaker concluded that comparisons between mouse and human sequences may not be as instructive as those between human and other species. For instance, Roderic Guigo (Institut Municipal de Investigacio Medica, Spain) found that chicken sequences may be more helpful in predicting coding exons, as they show good conservation in coding regions but diverge substantially elsewhere. Conservation between mammals seems to be more widespread, and so creates less clear distinctions between conserved and divergent sequences. Webb Miller (Pennsylvania State University, USA) presented a new program, PipMaker [<http://bio.cse.psu.edu/pipmaker/>] (Percentage Identity Plot Maker) for graphically viewing sequence conservation along genomic sequences.

Making ambiguity clear

Because *ab initio* prediction is far from perfect, and adding other evidence can improve gene prediction, there have been several efforts to develop graphical interfaces for the comparison of results. These interfaces allow simultaneous examination of the plethora of results generated by gene prediction programs along with sequence similarities. The idea is to show explicitly where evidence from different sources is contradictory or in agreement. Human intervention then takes the form of 'polishing' annotation: making decisions about the reliability of predicted features and designing experiments to support or refute them. A graphical interface called Artemis [<http://www.sanger.ac.uk/Software/Artemis/>], developed in the Sanger Centre (poster presented by Kim Rutherford), is designed to allow users to edit the features displayed. Artemis has been used extensively for annotation of Sanger Centre projects up to 4 Mb in size. These have been mainly pathogen genomes, but the program is now being used in the annotation of the genome of the fission yeast *Schizosaccharomyces pombe*. Various attempts are being made to automate the polishing process and reduce the amount of human intervention necessary. Richard Durbin (Sanger Centre, UK) described a new program called GAZE, which is an offshoot of his ACEDB database software. GAZE integrates evidence from multiple sources to come up with graphical representations of gene predictions. We were also introduced to Ensemble [<http://www.ensembl.org/>] by Ewan Birney. It takes exons predicted *ab initio* that are confirmed via similarity results and assembles the exons into predicted genes that are presented graphically. In this way it provides a 'base line' annotation for many of the fragmentary, unfinished human genomic sequences in the EMBL database. Around 2.9 Gb of the existing draft and finished sequence of the human genome have been processed by Ensemble and the results are freely available from the Ensemble website. Ensemble gene identifiers will remain stable during rearrangements and extension of draft sequences on the way to a definitive human genome sequence, which is at least three years away. Similar 'industrial scale' analyses are being run using the Genome Channel [<http://grail.lsd.ornl.gov/tools/channel/>], according to Ed Uberbacher; the genome channel is an analysis pipeline processing draft human genome sequence with a different combination of programs from Ensemble.

Automatic for the people

The Cold Spring Harbor Genome conference earlier this year saw the creation of Genesweep [<http://www.ensembl.org/genesweep.html>], a 'gene sweepstake' where participants can bet on the final number of human genes that will be found. The spread of bets reflects the current uncertainty among workers in human genomics, ranging from 27,462 up to 200,000. Interestingly, the mean is currently 62,598, much lower than the ballpark figure of 100,000 we have all become accustomed to. When the winner of Genesweep is announced

in 2003, what will be the reward for the rest of us? Once there is a complete set of known and predicted genes in which we have high confidence, it is to be made publicly available via the internet, and the way these data are presented will be influenced by the experiences and software taken from other genome projects. Various speakers at the meeting discussed archiving genomic annotation data for projects in the plant *Arabidopsis thaliana*, the fruit fly *Drosophila melanogaster* and completed human chromosomes. Michael Ashburner (EBI) stressed the importance of consistency in genome annotation across species and described the Gene Ontology (GO) project [<http://www.geneontology.org/>]. GO is an attempt to rigorously describe all the genes in a genome according to the molecular functions of their products and the biological processes and cellular components with which they are associated. The classification and standardized terminology of GO were used in the annotation of the *D. melanogaster* genome, and it is hoped that GO will become a community-curated entity, providing a democratic but central vocabulary for annotation.

Perhaps the real measure of the success of computational gene predictions is in their successful integration into the biologists' toolbox. Tim Hubbard (Sanger Centre) described the strategy for annotation of the first completed human chromosome sequence, of chromosome 22 [<http://www.sanger.ac.uk/HGP/Chr22/>]. Here the strength of computational predictions - their sensitivity - was exploited to generate a set of candidate exons. These candidates could then be used as the starting point for laboratory work to discover the actual mRNA sequences. For chromosome 22, 94% of genes were found and at least partially predicted by *ab initio* methods, but 16% of real exons were not predicted at all computationally and only 20% of predicted gene structures were correct. So, computational techniques give us invaluable clues to gene structure, but in the end it will be the addition of work by bench scientists that will provide the full picture. With this in mind, moves have begun to establish a Distributed Sequence Annotation System (DAS) [<http://stein.cshl.org/das/>] to democratize genome annotation. The idea is to designate a central 'reference server' which stores essential mapping and sequence data for the genome, and multiple 'annotation servers' maintained elsewhere by a range of different groups. Researchers interested in a given region of the genome would use a web browser-like application to download and integrate different features from servers of their choice. Thus, a reliable central annotation can be maintained in parallel with a diversity of less confidently predicted features that may or may not turn out to be useful. It seems inevitable that human beings will have to take on the final tasks of gene prediction and annotation once the machines have had a first pass. One way or another, human intervention is still essential in computational gene prediction and, it would seem, the more humans involved, the better.