DOI:
10.1038/nrg2253

MILESTONE 21

Fishing for genes

BANANASTOCK

Gene-prediction tools have co-evolved with advances in genome-sequencing capabilities, to make sense of the ever-increasing amount of data.

Early programs sifted through sequences to identify open reading frames. Later, richer representations of the features that distinguish coding from noncoding sequences were used, and methods treated gene prediction as a pattern-recognition problem. These programs, such as Genie, Genscan, GLIMMER and FGENESH, used linear-discriminant analysis, Markov models, neural networks or a combination of methods to detect the variation. Relatively simple models can be used for microbial gene identification: GLIMMER was applied to 10 completed microbial genomes in 1999.

The presence of exon splice sites complicates eukaryotic gene prediction. Each eukaryotic gene is 'marked' by start and stop codons and has splice sites (for example, ATG … GT-AG … GT-AG … Stop). Markov model-based applications then complete the identification — these algorithms are designed to compare sliding windows (several bases in length) to patterns that the program has 'learned' in a given genome. To learn, the program is trained on a part of the sequence for which gene information has already been determined experimentally.

Genie was one of the first programs developed to mine the human genome, and it identified up to 85% of the known protein-coding bases — a performance on par with that of other programs available in 1996. More recently, an abundance of expressed sequence data has improved prediction accuracy. FGENESH and a related suite of programs use an algorithm similar to that of Genie on a first pass; then, a BLAST comparison (see Milestone 15) to databases of exon products is used to mark confirmed exons.

The latest gene-annotation tools, such as Ensembl and Gnomon, integrate an even greater amount of information, including protein homology, cDNA and expressed sequence tag data. Ensembl was written to analyze the draft human genome, and since then has been used to annotate various vertebrate genomes, including zebrafish and human among others. Gnomon was developed to analyse other genomes, and was recently used for honey bee gene annotation.

Today, we take it as a given that the raw genomic data are presented in a meaningful way, as annotated sequences. Yet more is to come, as with increasing amounts of data to base their models on, computational biologists can train programs to make more accurate predictions and to mine the genome at an ever-increasing level of detail.

Irene Kaganman, Senior Copy Editor,
Nature Methods

ORIGINAL RESEARCH PAPERS Kulp, D. *et al.* A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4**, 134–142 (1996) | Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997) | Delcher, A. L. *et al.* Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999) | Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000) | Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14**, 942–950 (2004)

WEB SITES

Ensembl: http://www.ensembl.org
FGENESH: http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfs
Genie: http://www.fruitfly.org/seq_tools/genie.html
Genscan: http://genes.mit.edu/GENSCAN.html
GLIMMER: http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi
Gnomon: http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.html