

# Predicting full-length transcripts

Michael R. Brent

Accurate prediction of the complete structures of protein-coding genes, including 5'-untranslated regions, is crucial for full interpretation of genome sequence. A new report describes the statistical properties of the first exons of genes, and presents a novel statistical method for predicting their boundaries. First exons, which are partially or completely non-coding, have been relatively neglected by gene-finding algorithms. When integrated with other genome annotation tools, this new method will improve the technology of automated genome interpretation.

Published online: 17 May 2002

Instead of wondering at the exponentially increasing number of bases in GenBank, we now wonder at the exponentially increasing number of sequenced eukaryotic genomes. This change of units reflects the triumph of both high-throughput sequencing and shotgun assembly algorithms. But the technology of genome interpretation has not kept pace. The most common biological approach to elucidate the locations and structures of protein-coding genes is random sequencing of cDNA libraries. End sequences (ESTs) are relatively inexpensive to obtain, but often do not cover the full length of the mRNA. Furthermore, the 5' end of the cDNA is often not the 5' end of the mRNA owing to the non-processivity of reverse transcriptase. Constructing and sequencing full-length cDNAs can be difficult and expensive, although concerted efforts have recently yielded a large collection of full-length mouse cDNAs [1]. Nonetheless, random sequencing of cDNA libraries fails to identify a significant fraction of genes whose expression is low, tissue restricted, developmentally restricted, or restricted to specific growth conditions. Computational gene-structure prediction methods are currently the best hope for elucidating genes missed by random cDNA sequencing. However, predictions by themselves are not enough; they

must be verified by directed mRNA identification methods, such as PCR from cDNA libraries, RT-PCR from mRNA preparations, and hybridization to DNA microarrays. Even small improvements in the accuracy of gene prediction can have dramatic effects on the cost-effectiveness of verifying predictions on a genomic scale.

## Gene-finding algorithms

Algorithms for identifying coding regions in a genomic sequence have been a topic of investigation for nearly 20 years (e.g. Refs [2–8]; for reviews, see Refs [9–11]). Until recently, all such algorithms were characterized by one of two clearly distinguishable approaches. Alignment-based algorithms rely on aligning some other biological sequence to the genomic sequence, typically a partial or complete cDNA sequence from either the same gene or a related gene [12–14]. Because the introns are removed from the cDNA sequence, such alignments can help elucidate the locations of coding regions. Pure alignment-based methods can be quite powerful when a full-length cDNA from a closely related gene has been sequenced. However, they are limited because alignable cDNA sequences are not available for many genes, for the reasons described above. By contrast to alignment-based algorithms, *ab initio* algorithms use the genomic DNA to be annotated as their only biological sequence input. These algorithms rely on identifying patterns in genomic DNA that are characteristic of particular parts of genes, such as introns, coding regions and splice sites. This approach relies on statistical pattern recognition, with hidden Markov models being one of the most common frameworks. The performance of *ab initio* gene-structure prediction algorithms reached a plateau in the middle 1990s, with systems like GENSCAN [8] yielding the best results on vertebrate sequences.

Regardless of approach, research was focused until recently on genomic sequences known to contain a single

gene – typically short pieces of DNA that were specifically cloned and sequenced by investigators with a particular interest in their products. In the second half of the 1990s, high-throughput sequencing of large-insert clones began to yield large quantities of anonymous genomic sequence, with human sequence taking center stage. In human, an anonymous 100-kb sequence could contain between zero and ten or more protein-coding genes. Such sequences define a very different, and much harder gene-structure prediction problem. The best available algorithms were able to predict exact gene structures with accuracy in the range of 10–20%, although they could identify individual exons with 50–60% accuracy and coding nucleotides with 85–90% accuracy.

## Recent advances in gene finding

Recently, several hybrid algorithms have been developed for combining *ab initio* gene prediction with alignments between two closely related genomes, such as mouse and human. Examples include TWINSKAN [15], sgp [16] and SLAM. These are fundamentally *ab initio* algorithms but, in different ways, they favor the prediction of genes in one genome that are similar to regions of the other genome. Such algorithms are motivated by the idea that functionally important parts of a genome, including coding regions and perhaps also splice sites, regulatory regions and other features, are likely to be under selective pressure to maintain function and hence are likely to diverge more slowly than the regions under neutral selection. Such algorithms have begun to improve the accuracy of gene-structure prediction on anonymous genomic sequence for the first time since the publication of GENSCAN five years ago. However, it appears that only about half of conserved sequence between mouse and human is coding, so the improvements so far still leave the problem of gene-structure prediction far from solved.

### Finding first exons

Gene-structure prediction systems such as the ones described above are focused on identifying the coding regions of genes, although some make a desultory effort at annotating promoters and 3'- and 5'-untranslated regions (UTRs). Indeed, so little importance has been attached to UTRs that the term *exon* has come to mean contiguous translated sequence in the gene-finding literature, whereas the more general meaning of the term is contiguous transcribed sequence. The reason for this is not merely rampant protein centricism, but also a paucity of known, full-length transcript sequences. Because transcript sequencing is always achieved by reverse-transcription, and reverse-transcriptase tends to fall off before the 5' end of the RNA, sequences obtained for transcripts are often too short. Furthermore, it is difficult to know whether an mRNA sequence is full length.

Recently, the 'oligo-capped' cDNA libraries of Suzuki, Sugano and colleagues have greatly increased the number of known, full-length mRNA sequences [17]. Davuluri *et al.* [18] used these and other putative full-length human transcripts to construct a first-exon database by mapping 2139 transcript sequences to their genomic loci. Analysis of this database revealed that 39% of the first exons were completely non-coding, and that these were on average shorter than the partially coding first exons (151 bp versus 348 bp). They also studied the relation between so-called 'CpG islands' [19] and first exons by calculating the percentage of CpG dinucleotides in windows of 201 bp. The maximum percentage, over all windows starting within 500 bp upstream of the transcription start site, was computed. Davuluri *et al.* found a sharply bimodal distribution, with a clear separation between transcripts that have an upstream window with >6.5% CpG (dubbed CpG-related FEs) and those that do not. By this criterion, ~70% of the first exons in their database were CpG-related, and in 76% of these, the window with greatest CpG percentage overlapped the first exon.

Based on these and other statistical patterns, Davuluri *et al.* developed a program called FirstEF (first exon finder) for predicting the locations of first exons in genomic sequence. FirstEF

combines models of the region upstream of the transcription start site (termed the 'promoter region'), the first exon itself, and the first splice donor site. Computational experiments reported by Davuluri *et al.* suggest that FirstEF is more effective at finding promoter regions than are other programs designed specifically to identify these regions, such as PromoterInspector [20]. This improvement is likely, at least in part, to be because splice donor sites are easier to recognize statistically than promoters. FirstEF exploits this by favoring promoter regions just upstream of likely splice donors. In spite of these improvements, Davuluri *et al.* recommend a two-pass procedure in which a full gene-prediction program is first used to locate the 5' ends of genes approximately. FirstEF is then used to refine the prediction within a relatively small region around the 5' end of the predicted gene.

### Potential impact of FirstEF

Regardless of its potential as a tool for genome annotation, FirstEF represents important progress for those interested in understanding and modeling transcription. Furthermore, the ability to identify first exons is important for identifying *cis*-regulatory elements. The impact of FirstEF on the accuracy of gene-structure prediction and genome annotation in general will depend on how successfully it can be integrated with state-of-the-art systems for predicting the remainder of the gene. The two-stage procedure recommended by Davuluri *et al.* is only one possible method of combining FirstEF with full gene-structure prediction programs. In the long run, the most promising approach might be integration at the level of the probability model. That is, the insights gained from FirstEF, and perhaps the specific models used in it, must be combined with probabilistic models of the rest of the gene. In this way, features such as the presence and relative location of CpG islands can help locate features such as translation initiation sites, splice sites and polyadenylation sites, and vice versa. Because FirstEF uses several features that are already partially accounted for in current gene-prediction models, it remains to be seen how much leverage can be gained from such integration. As the first

system able to predict completely non-coding first exons, however, FirstEF serves as a wake-up call to a gene-prediction community that has tended to sleep through the 5' UTR and then awaken with a startle at the translation-initiation site. If our gene predictors took note of the first exon, they might be more alert to the start of translation.

### Summary

The accuracy of gene prediction is not likely to improve so much in the next five years that experimental verification becomes superfluous. However, integration of the insights and models provided by FirstEF into complete gene-prediction systems could provide another boost in accuracy at a time when genome comparison is also starting to pay off. Each boost of this sort brings down the cost of sequence-directed gene verification, by increasing the ratio of positive to negative results. Improvements in gene-prediction accuracy, combined with the continuing automation of routine laboratory tasks, is bringing the goal of identifying almost every human gene within reach.

### Acknowledgements

M.R.B. is supported, in part, by the National Science Foundation (DBI-0132436) and the National Institutes of Health (HG02278).

### References

- 1 Kawai, J. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature* 409, 685–690
- 2 Fickett, J.W. (1982) Recognition of protein-coding regions in DNA sequences. *Nucleic Acids Res.* 10, 5303–5308
- 3 Guigo, R. *et al.* (1992) Prediction of gene structure. *J. Mol. Biol.* 226, 141–157
- 4 Staden, R. and McLachlan, A.D. (1982) Codon preference and its use in identifying protein-coding regions in long DNA sequences. *Nucleic Acids Res.* 12, 505–519
- 5 Xu, Y. *et al.* (1994) An improved system for exon recognition and gene modeling in human DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 376–384
- 6 Snyder, E.E. and Stormo, G.D. (1993) Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res.* 21, 607–613
- 7 Snyder, E.E. and Stormo, G.D. (1995) Identification of protein-coding regions in genomic DNA. *J. Mol. Biol.* 248, 1–18
- 8 Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94

- 9 Burge, C.B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* 8, 346–354
- 10 Stormo, G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.* 10, 394–397
- 11 Reese, M.G. *et al.* (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10, 483–501
- 12 Gelfand, M.S. *et al.* (1996) Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. U. S. A.* 93, 9061–9066
- 13 Florea, L. *et al.* (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8, 967–974
- 14 Mott, R. (1997) EST\_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* 13, 477–478
- 15 Korf, I. *et al.* (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics* 17, S140–S148
- 16 Wiehe, T. *et al.* (2000) Genome sequence comparisons: hurdles in the fast lane to functional genomics. *Brief Bioinform.* 1, 381–388
- 17 Suzuki, Y. *et al.* (2000) Statistical analysis of the 5'-untranslated region of human mRNA using 'Oligo-Capped' cDNA libraries. *Genomics* 64, 286–297
- 18 Davuluri, R.V. *et al.* (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.* 29, 412–417
- 19 Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261–282
- 20 Scherf, M. *et al.* (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* 297, 599–606

**Michael R. Brent**

Campus Box 1045, Washington University,  
St Louis, MO 63130, USA.  
e-mail: [brent@cs.wustl.edu](mailto:brent@cs.wustl.edu)

# Predicting full-length transcripts

## Response from Michael Q. Zhang

Bioinformatics is driven by genomic data. All computational methods depend on the available data and all prediction accuracies depend on the quality and quantity of the training data. Traditionally gene prediction has focused on the protein coding sequence, because it directly relates to gene function, and therefore most high-quality sequence data is around coding regions. As large-scale genomic sequencing projects pick up speed, and as the need to understand mountains of gene expression data expands, attention is turning to regulatory sequences that are often found in the non-coding regions. Identification of boundaries of genes has become a pressing need.

FirstEF (First Exon Finder) is the first attempt to capitalize on newly available genomic data (largely produced by Sugano and colleagues at the University of Tokyo). However, there still many false-positives in gene prediction results. This is partly because current validation methods are limited. A current estimate suggests that the majority (>90%) of sequenced human transcripts are 5'-truncated, with at least 10% of human genes not represented anywhere in the public cDNA/EST databases. Almost all the 'full-length' cDNA sequencing projects stop short at the 5' end (so should really be called 'full-length up to ATG', the translation not transcription start site).

Here, I want to challenge technology-minded experimental biologists to produce massive real full-length mRNAs, and to hunt down every possible

### Box 1. Potential approaches

- RT-PCR + *in-silico* prediction have, to a certain extent, successfully identified a large number of novel transcripts [a]. But going after every possible tissue and developmental stage is still not possible with current technology. Can drugs (such as TSA, or TSA and AZT) be used to de-repress a large number of genes so that many novel transcripts, not otherwise expressed, could be harvested? Disrupting chromatin structure would also relieve the need for many tissue-specific activators/enhancers of transcription.
- ChIP + *in-silico* prediction [b] or 5'-oligo-capped' cDNA sequencing [c] have been useful in identifying many promoters and 5' exons. Massive parallel 5'-RACE should also be designed, and given first exon predictions [d], an 'array of transcriptional reporters' construct would be especially useful.

#### References

- a Das, M. *et al.* (2001) Assessment of the total number of human transcription units. *Genomics* 77, 71–78
- b Kel, A.E. *et al.* (2001) Computer-assisted identification of cell cycle-related genes – new targets for E2F transcription factors. *J. Mol. Biol.* 309, 99–120
- c Suzuki, Y. *et al.* (2000) Statistical analysis of the 5'-untranslated region of human mRNA using 'oligo-capped' cDNA libraries. *Genomics* 64, 286–297
- d Davuluri, R. *et al.* (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.* 29, 412–417

transcript. Only then will people start to appreciate how many 'false-positives' are actually real. All computational biologists are facing the same dilemma: because computational prediction has far surpassed the capacity of experimental validation, the lack of complete and accurate data has impeded further development of better bioinformatics. Recently, several funding agencies have realized the importance of codevelopment of high-throughput experimental validation technologies and genome informatics. It is time for computational biologists to reach out and work together with bench scientists to design the kind of large-scale experiments that are desperately needed (Box 1).

I am counting on biotechnologists to come up with more ingenious ideas and pragmatic solutions to the dilemma.

**Michael Q. Zhang**

Watson School of Biological Sciences,  
Cold Spring Harbor Laboratory,  
1 Bungtown Road, PO Box 100,  
Cold Spring Harbor, NY 11724, USA.  
e-mail: [mzhang@cshl.org](mailto:mzhang@cshl.org)

Published online: 17 May 2002

Personal subscribers to  
*Trends in Biotechnology* have  
FREE online access

Go to <http://tibtech.trends.com>

If you have any questions e-mail:  
[info@current-trends.com](mailto:info@current-trends.com)