Recent advances in gene structure prediction

Michael R Brent[1] and Roderic Guigó[2]

De novo gene predictors are programs that predict the exon-intron structures of genes using the sequences of one or more genomes as their only input. In the past two years, dual-genome *de novo* predictors, which exploit local rates and patterns of mutation inferred from alignments between two genomes, have led to significant improvements in accuracy. Systems that exploit more than two genomes simultaneously have only recently begun to appear and are not yet competitive on practical tasks, but offer the greatest hope for near-term improvements. Dual-genome *de novo* prediction for compact eukaryotic genomes such as those of *Arabidopsis thaliana* and *Caenorhabditis elegans* is already quite accurate. Although mammalian gene prediction lags behind in accuracy, it is yielding ever more useful results. Coupled with significant improvements in pseudogene detection methods, which have eliminated many false positives, we have reached the point where *de novo* gene predictions are being used as hypotheses to drive experimental annotation via systematic RT-PCR and sequencing.

Addresses
[1] Laboratory for Computational Genomics, Campus Box 1045,
Washington University, One Brookings Drive, St Louis,
Missouri 63130, USA
e-mail: brent@cse.wustl.edu
[2] Research Group in Biomedical Informatics, Institut Municipal
d'Investigació Mèdica/Universitat Pompeu Fabra/Centre de Regulació
Genòmica, Barcelona, Catalonia, Spain

Current Opinion in Structural Biology 2004, **14**:264–272

This review comes from a themed issue on
Sequences and topology
Edited by Peer Bork and Christine A Orengo

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2004.05.007

Abbreviations

EHMM	evolutionary HMM
EST	expressed sequence tag
HMM	hidden Markov model
indels	insertions and deletions
ORF	open reading frame
PPT	poly-pyrimidine tract
RT-PCR	reverse transcription-polymerase chain reaction
TSS	transcription start sites
UTR	untranslated region

Introduction

The past two years have seen the flowering of the genomic era, a period during which metazoan genome sequencing has been transformed from a major international event to a common undertaking that barely makes the covers of scientific journals, much less popular newspapers. The wealth of raw data generated by this technological triumph has greatly accelerated scientific progress even while it remains far from fully analyzed. It has also driven a series of advances in computational genome analysis, including methods for predicting the exon-intron structures of genes. Such methods can be divided into those that make use of expression data (including sequences from cDNAs and potentially data from hybridization experiments) and those that use only the sequences of one or more genomes (*de novo* or *ab initio* methods). The focus of this review is recent developments in *de novo* gene prediction for the genomes of higher eukaryotes.

De novo gene predictors can be categorized into those that use a single genome sequence, those that use two genome sequences to infer local rates and patterns of mutation along the genome, and those that use more than two genomes for the same purpose. Single-genome predictors reached a state of relative maturity with the development of systems based on hidden Markov models (HMMs) (e.g. GENSCAN [1], GENIE [2] and HMMGENE [3]) and related models (e.g. GENEID [4] and FGENESH [5]). Dual-genome *de novo* predictors (e.g. SGP-2 [6••], SLAM [7••] and TWINSCAN [8,9••]) have led to the greatest practical improvement in the accuracy of prediction over the past two years. Systems that exploit more than two genomes simultaneously (e.g. [10••,11]) have only recently begun to appear and are not yet competitive on practical tasks, but offer the greatest hope for near-term improvements in accuracy.

Since the first animal and plant genomes were sequenced, *de novo* gene finders have been part of the standard toolbox for genome annotation and analysis. With the advent of dual-genome predictors, the accuracy for compact genomes, such as that of *Arabidopsis thaliana*, has become so good that one-half to two-thirds of all known genes are predicted exactly right, from the start codon through every splice site to the stop codon, and most of the imperfect predictions are only slightly off ([12]; Chaochun Wei, personal communication). The accuracy for mammalian genomes has lagged behind owing to inherent challenges, such as the large number of pseudogenes and small fraction of coding sequence, that affect all mammalian annotation methods. Although dual-genome *de novo* systems now correctly predict about 75% of all known exons at both splice sites, only 15–20% of known gene structures are predicted correctly throughout the coding region [6••,9••]. Annotation pipelines such as ENSEMBL [13], which require homology to known expressed sequences, are somewhat more accurate at predicting exons of known genes [9••], but they tend to miss many predicted exons and genes that can be verified experimentally [14••,15,16]. Perhaps the most significant development of the past year in mammalian annotation has been the application of recently developed pseudogene detection methods [17,18], which have eliminated many false positives from both *de novo* and pipeline-style annotation. Indeed, the advent of dual-genome systems, together with the elimination of many pseudogenes, has improved the *de novo* prediction accuracy to the point where systematic reverse transcription-polymerase chain reaction (RT-PCR) and sequencing of *de novo* predictions is a cost-effective complement to sequencing of random cDNA clones, even in mammalian genomes [19].

Single-genome predictors

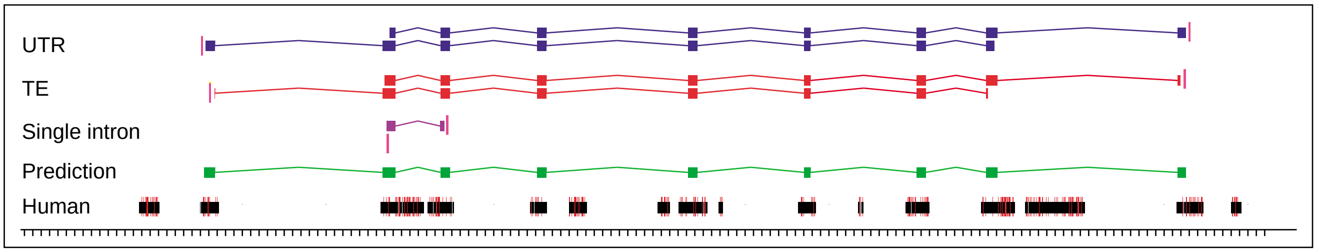
Although methods based on the comparison of two or more genomes have greater accuracy, there are good reasons to develop improved single-genome *de novo* predictors. First, they are easier to train and faster to run than multi-genome predictors, and are consequently the first systems used to annotate newly sequenced genomes. Second, although comparative methods exploit an underlying sequence alignment, the information from the alignment is usually integrated with information from intrinsic sequence signals (e.g. splice sites) and the compositional biases characteristic of protein-coding regions (e.g. codon usage). When explicitly investigated [9••,20•], the relative contribution of the sequence alignment to the final gene prediction has been found to be smaller than that of the intrinsic coding sequence patterns. Third, comparative gene prediction requires the sequenced genomes of two or more species at the appropriate phylogenetic distance (see below); when there is no second genome at an appropriate distance, one must resort to single-genome predictors.

Recent advances for single-genome predictors have focused on the problem of training and parameter estimation. Often, newly sequenced genomes lack large enough samples of known genes from which to estimate model parameters. Although in such cases it is common practice to use a genome predictor that has been trained on another species, recent analysis indicates that gene finders trained on a foreign genome tend to perform suboptimally (I Korf, personal communication). To address this limitation, Korf introduces the idea of 'bootstrap parameter estimation', in which a foreign gene finder is run on a novel genome and the resulting predictions are used to estimate the parameters for gene prediction for the novel genome. Kotlar and Lavner [21] pursue a different strategy that emphasizes detecting periodic correlations between nucleotide positions. These measures are universal and can be computed without the need for previous training.

Dual-genome predictors

Dual-genome gene predictors rely on the fact that functional regions of a genome sequence — protein-coding genes in particular — are more conserved during evolution than non-functional ones (Figure 1, bottom two tracks). Over the past four years, several programs have been developed that exploit sequence conservation between two genomes to predict genes. A wide variety of strategies have been explored. In the pair HMM approach (e.g. SLAM [7••]), a joint probability model for sequence alignment and gene structure is used to express the different types of alignments one expects to see in, for example, coding regions and introns. Alignment and gene prediction are performed simultaneously using a dynamic programming algorithm that combines the classic algorithms for alignment and HMM decoding. In the 'informant genome' approach (e.g. SGP-2 [6••] and TWINSCAN [8,9••]), alignments are performed first using standard tools such as TBLASTX or BLASTN, and these alignments are used to inform prediction

Figure 1



A TWINSCAN prediction (green, subsequently identified as rat aspartylglucosaminidase). TWINSCAN uses the blocks of alignment from the human genome (black), and the mismatches and gaps within the blocks (red) to predict the most likely gene structure. ENSEMBL predicted only small fragments of two exons in this gene, due to a fragmentary rat protein in public databases. RT-PCR and direct sequencing were performed using primer pairs designed around a single intron, in the predicted first and last exons (TE), and the predicted UTRs. Primers are shown as tall pink blocks at the same level as the sequence they yielded and aligned amplicon sequences from these experiments are shown in purple, red and blue, respectively. The left single intron primer did not yield high-quality sequence. Reproduced from [15].

algorithms that are extensions of successful single-genome predictors. SGP-2, SLAM and TWINSCAN were used in the comparative analysis of the human and mouse genomes [22]. These systems significantly outperformed single-genome predictors while identifying many exons and genes not found by expression-based annotation pipelines such as ENSEMBL [14••].

One of the surprises that emerged from comparison of the human and mouse genomes was that, whereas coding regions are mostly conserved between human and rodent genomes, most conserved regions are not protein coding. Depending on the tool, almost 40% of the human genome can be aligned with the mouse genome [22], but less than 2% encodes proteins. Most dual-genome predictors discriminate indirectly between coding and non-coding conservation by relying heavily on intrinsic coding sequence patterns.

In this regard, several researchers have been recently attracted to the problem of discriminating coding from non-coding conservation in cross-genome sequence alignments. Although to some extent implicit in the algorithms used in most dual-genome predictors, Nekrutenko *et al.* [23,24] explicitly use the ratio of non-synonymous over synonymous substitutions. There tend to be more synonymous than non-synonymous substitutions in open reading frames (ORFs) that encode proteins, as they are under selective pressure to maintain protein function. In non-coding ORFs, no such distinction is seen. By systematically computing this measure for segments conserved between human and mouse, Nekrutenko *et al.* [23] found evidence of more than 13 000 exons absent from all annotations of the human genome. Moore and Lake [25], on the other hand, convert sequence alignments to numerical series, so that alignments in coding regions result in series with well-defined frequency bands, whereas those in non-functional regions result in noisy series. They use the Wiener filter to eliminate the noise in the frequency space, in this way uncovering the alignments that are likely to occur in coding regions. Finally, Noguchi *et al.* [26] introduce an index relative to frame recovery in cross-species sequence alignments. If insertions or deletions cause a frame shift in coding regions, other nearby insertions or deletions will usually restore the reading frame. Frame recovery is relatively rare in non-coding alignments.

With increasing frequency, eukaryotic genome sequencing stops at a coverage level that makes full genome assembly unfeasible. In this regard, the informant genome method, which exploits short, interrupted alignments, has an advantage over the pair HMM method, which requires long continuous alignments from orthologous (not paralogous) regions. Flicek *et al.* [9••] and Parra *et al.* [27] investigated the effect of the level of mouse genome sequence coverage on the accuracy of the predictions for the human genome obtained by TWINSCAN and SGP-2, respectively. Both found that performance increases steadily up to threefold coverage but more slowly thereafter, leveling off at about fourfold [9••].

With the availability of an increasing number of genome sequences, it has become important to understand how phylogenetic distance affects the accuracy of dual-genome gene predictions. Investigating this question, Guigó and Wiehe [28], and Zhang *et al.* [20•] concluded that the optimal reference for comparative analysis of the human genome would be a species more distant than mouse. Wang *et al.* [29] bracket the optimal distance for annotating mouse between chicken (which is too far) and rat (which is too close). All three groups agree that, among the genome sequences available in late 2003, rodents were the best for annotation of the human genome. Ultimately, however, it may prove better to avoid this choice by simultaneously using all available genome comparisons to inform gene prediction.

Multi-genome predictors

Dual-genome *de novo* systems work by using alignments between two genomes to draw inferences about the rate of evolution at each nucleotide. If the two sequences match at a particular base, that base is conserved; if they do not, it is not conserved. Although this has proven effective in practice, it is clearly a crude measure of evolutionary rate. Using multiple alignments among several genomes can provide a more precise measure of evolutionary rate and, in principle, this should lead to greater precision in comparative gene prediction. Furthermore, dual-genome predictors for mammalian genomes have had the greatest success using relatively distant genomes, such as mouse and human. However, there are inherent uncertainties in reconstructing the lineages of genomic regions for two such distantly related organisms because so many rearrangements, segmental duplications, retrotranspositions and other events have occurred since their latest common ancestor. One possible solution is to infer evolutionary rate from many closely related species instead of two more distant species. For example, Boffelli *et al.* [30] have observed that the collective divergence of the higher primates, as a group, is comparable to the divergence of human and mouse, yet their genomes can be aligned much more accurately than those of human and mouse.

Recently, the continuous time Markov chains that are standard for describing the evolution of a particular residue have been combined with the discrete HMMs that are standard for describing the functions of nucleotides within the sequence of a gene. The combined models have been called evolutionary HMMs (EHMMs) [11,31] and phylo-HMMs [10••,32]. The input to these models is a multiple sequence alignment among several genomes. To paraphrase Siepel and Haussler, phylo-HMMs model molecular evolution as a Markov process in two dimensions: a substitution process over time at each site in the aligned genomes, which is guided by a phylogenetic tree; and a process by which the rate of evolution changes from one site to the next.

All phylo-HMM models assume that the rate of evolution of a nucleotide depends on its function, but the most elaborate phylo-HMM models [10••,32] also consider the possibility that the evolutionary rate may vary from one region of a genome to another. Furthermore, they allow the probabilities of mutation at each site to depend on the observed pattern of mutation in the previous few sites. Although traditional evolutionary models focus on substitutions rather than insertions and deletions (indels), the modeling of indels is critical for gene prediction. The reason is that the patterns of indels differ greatly between coding and non-coding regions; coding regions tend to have far fewer indels and, where they do occur, they are usually in multiples of three, preserving the reading frame. The best approach to indel modeling proposed so far seems to be to treat each pattern of gaps within a multiple alignment column separately. Each HMM state in the original model is copied once for each possible combination of gaps in an alignment column and separate probability models are estimated for each such state. One limitation of this state copying method is that it expands the number of HMM states by a factor of roughly 10–30, depending on the number of genomes modeled [10••].

Although this approach is very sophisticated from a mathematical modeling perspective, it has not yet yielded practical improvements in the accuracy of gene prediction. This is because the complexity of the evolutionary models leaves little room for the complex gene structure models needed to outperform state-of-the-art dual-genome systems. For example, none of the phylo-HMM gene finders has yet incorporated non-geometric models of exon length. Indeed, one of the lessons of dual-genome gene finders is that the best performance was obtained by modifying state-of-the-art single-genome systems rather than by building new systems that rely primarily on the signal from natural selection. Nonetheless, efforts to exploit alignments among multiple genomes to improve prediction accuracy are likely to bear fruit in the next year or two.

Combining the output of gene predictors

Human annotators and automated genome annotation 'pipelines' [13,33] generally operate by combining information that ultimately derives from cDNAs (expressed sequence tags [ESTs], full-length cDNA sequences and conceptual translations) with information from one or more *de novo* gene finders. Human annotators use their intuition and experience to synthesize the often contradictory evidence into a single gene structure, whereas pipelines generally use rules based on the intuition and experience of their designers. The rules are often simply priorities, for example, use an exon predicted by GENSCAN if it is supported by an EST alignment but does not overlap a GENEWISE [34] protein alignment. Recently, several systems have been developed to combine such evidence sources in a more mathematically principled way, using evidence weighting and dynamic programming algorithms [12,35,36]. When expression data are used, it is difficult to estimate performance on unknown genes from performance on known genes, as known genes are more likely to yield expression data than unknown genes (that is, in most cases, how they came to be known). Nonetheless, several papers suggest that combining different *de novo* gene predictors with one another [12,36] and with expression data [12,35] improves accuracy. Such systems may be improved further by current efforts to increase the accuracy of cDNA alignment [37], cross-species gene-structure mapping [38] and cross-species EST alignment [39].

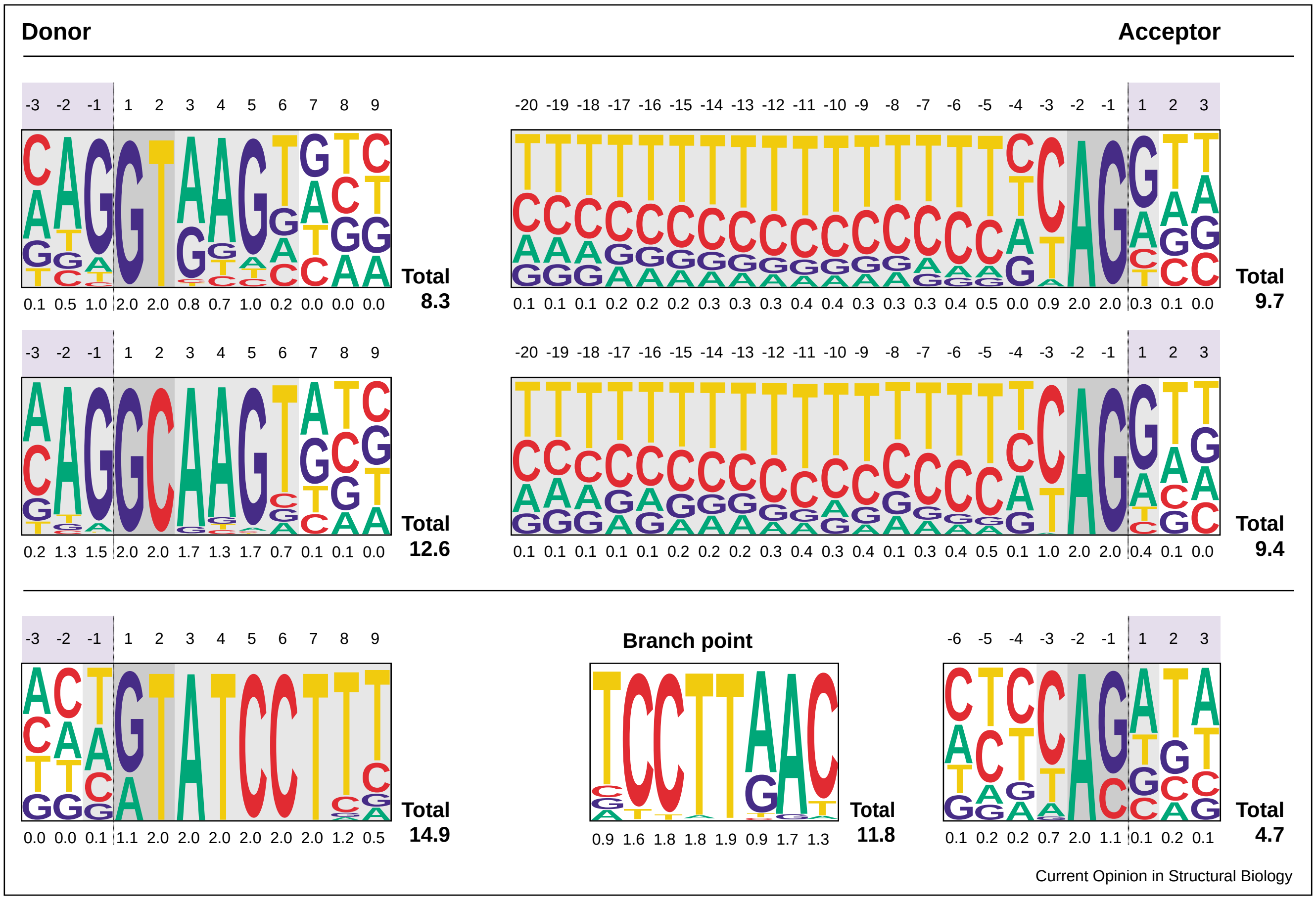
Improving models of DNA sequences and their roles in protein production

De novo gene prediction programs work primarily by recognizing patterns in genomic sequences that are characteristic of splice sites, translation initiation and termination sites, protein-coding regions, poly-adenylation sites and sites with other specific functions in gene expression. In most systems, pattern recognition is based on probability models for each of these functions. For example, given a DNA sequence, the splice donor model assigns a likelihood to the proposition that the sequence functions as a splice donor. The more accurate the model, the more it assigns high likelihoods to true splice donor sequences and low likelihoods to other sequences.

Splice sites are among the most powerful signals used by gene prediction programs, so accurate splice site modeling is crucial for achieving high accuracy. Most current gene prediction systems require introns to begin with GT and end with AG, as roughly 99% of introns in sequenced genomes do. However, GC–AG splice sites have a particularly strong splice donor consensus (Figure 2) and incorporating models of them has been found to improve accuracy (Ping Hu, unpublished).

Splice donor models typically consider about 3 nucleotides of exon and 6 nucleotides of intron (-3 to $+6$, where 0 is the exon-intron boundary). Splice acceptor models consider a 'core' acceptor site covering roughly 6 nucleotides of intron and 3 nucleotides of exon (-6 $+3$), as well as the upstream poly-pyrimidine tract (PPT), which extends another 15–30 nucleotides into the intron. Some gene finders also use branch point models, which also fall 15–30 base pairs upstream of the core acceptor site [40] (Figure 2). The most common types of models for splice sites are position-specific (inhomogeneous) 0^{th} or 1^{st} order Markov chains, often called WMMs and WAMs,

Figure 2



Current Opinion in Structural Biology

Sequence logos representing the splice site sequences of human introns. Although close to 99% of human introns start with GT at the donor site and end with AG at the acceptor site (top logos), there are some that use GC at the donor site (middle logos). These introns show a strong donor site consensus (12.6 bits of information in GC donors versus 8.3 in GT donors). In U12 introns (bottom logos), which are spliced out by a different mechanism, the donor site consensus is even stronger (14.9 bits). The acceptor site lacks the long PPT characteristic of most human introns, but in compensation it has a strong consensus at the branch point (11.8 bits), which is absent in most human introns. No *de novo* gene prediction program incorporates models for U12 introns and they are systematically mispredicted in the annotation of eukaryotic genomes. Logos were made with the PICTOGRAM program [65].

respectively. The 0th order Markov chains (WMMs) model each position in the splice site independently; a probability is computed for each base of the input sequence occurring in the corresponding position of the splice site and these probabilities are multiplied together to give a likelihood for the entire sequence. First-order inhomogeneous Markov chains (WAMs) condition these probabilities on the immediately preceding base, capturing dependencies between adjacent positions. GENSCAN improved performance by modeling non-adjacent dependencies in splice donor sites using a decision-tree-like model [1]. For example, this model captures the fact that splice donors without a guanine in position +5 are much more likely to have adenines in positions +3 and −2 than donors that have a guanine in +5.

Recently, several new probability models have been introduced for modeling non-adjacent dependencies between positions in splice sites, including increment of diversity quadratic discriminant analysis (IDQD) [41], support vector machines (SVMs) [42••,43], maximum entropy models (MEMs) [44,45] and Bayesian networks [46]. These methods appear to yield modest but real improvements over previous models of the core splice sites and PPT. Perhaps more important is the discovery of signals further into the intron that help differentiate true splice sites from false splice sites ([42••]; see also [47]). This discovery suggests that extending splice site models further into the intronic regions is likely to improve their accuracy. Similarly, the recent elucidation of splice enhancers within the exon suggests another direction in which splicing models can be usefully extended [48•].

New research also suggests that long introns display a much narrower range of splice site sequences than short introns [49•]. The interpretation of this observation is that the splicing signals for long introns must be stronger to compete with all the other potential splice sites in the intron (also see [50]). It should be possible to take advantage of this dependence between intron length and splice signal to avoid predicting long introns that lack strong signals. This may allow gene finders to predict long introns more accurately — currently, they tend to avoid predicting long introns altogether [51]. Current gene finders also assign too low a probability to short introns, favoring the intermediate range more than they should. This is because, for technical reasons, fully accurate models of intron length increase the time it takes to run gene prediction programs beyond acceptable limits. However, increases in computing power have made it possible to use models that, although not fully accurate, are much more accurate than those that were used in the past. This was accomplished by creating a highly accurate submodel for short- and medium-length introns, while using fast, moderately accurate models for very long introns [52•]. Separating intron models by length also makes it possible to model their splice sites differently. The combination of more accurate intron length models with length-specific splice site models can be expected to contribute to greater accuracy in the future.

Extending the functionality of gene predictors

Despite progress in modeling sequence signals that function in gene expression, overall the models underlying current computational methods are still quite simple, encoding a rather naive view of the eukaryotic gene. Almost without exception, computational gene finders predict only the coding fraction of a single spliced form of non-overlapping, canonical protein-coding genes. They deal poorly, if at all, with untranslated regions (UTRs), alternative spliced forms, overlapping or embedded genes, short intronless genes and rapidly evolving genes that are conserved poorly across genomes. They also tend to fail with highly atypical genes, such as those with unusual codon bias, those with non-canonical splice sites (Figure 2) and those that code for selenoproteins.

Although comparative analysis [14••] does not seem to support the existence of many more than 25 000 human genes, the discrepancy between this estimate and the results of genome-wide transcriptional surveys [53–55] remains puzzling. It cannot thus be ruled out that a significant fraction of the protein-coding content of the human genome — corresponding either to entirely novel genes or to fragments of already known genes — may remain undetected.

Efforts thus have also been made towards extending the functionality of computational gene finders, endowing them with a richer underlying model of the eukaryotic gene, which should ultimately lead to the production of more comprehensive automatic gene catalogs of eukaryotic genomes. We will discuss here recent progress in three different areas: prediction of UTRs, prediction of alternative splicing and prediction of selenoprotein genes.

UTRs are poorly predicted by computational methods: they do not show the characteristic sequence bias of coding regions, are less conserved across species and transcription start sites (TSS) exhibit a poor sequence consensus. On the other hand, although cDNA sequences and EST libraries often contain good representations of the 3′ end of genes, extending cDNA sequences to cover the whole 5′ region is often technically difficult [56]. Recently, Bajic and Seah [57] improved the prediction of TSS using information about CpG islands and signals in the downstream promoter region. Prediction of TSS is at the interface between gene and promoter prediction, a field on its own (see [58] for a recent review).

De novo prediction of the alternative splicing forms of genes is an open problem for which there is currently no adequate solution. ESTs are still the primary source of evidence and advances have been reported in methods for reconstructing a few of the true splice forms of a gene from many EST sequences (e.g. [59]). The estimate by Thanaraj *et al.* [60] that more than 60% of alternative splicing events are conserved between human and mouse offers hope that genome comparison, in conjunction with either *de novo* or expression-based methods, could contribute to the delineation of the alternative forms of eukaryotic genes. In addition, it has been suggested that the 'suboptimal' (e.g. second most probable) annotations found by single-genome [61] or dual-genome [62] gene prediction programs may be good candidates for alternatively spliced forms.

Selenoproteins highlight the limited ability of current systems to deal with exceptions to the canonical rules defining eukaryotic genes and illustrate the power of comparative genomics strategies to overcome them. In selenoproteins, the codon TGA is translated into a selenocysteine residue, whereas gene prediction programs assume without exception that the TGA triplet stops translation. Selenoprotein genes are absent from all automatic annotations of eukaryotic genomes. Although they are not abundant, selenoproteins participate in important physiological processes. Recently, methods have been reported that extend single-genome predictors to capture the specific sequence features characterizing selenoprotein genes and that, after cross-genome comparisons of the predictions, led to the identification of novel mammalian and vertebrate selenoprotein families [63,64].

Experimental verification and refinement of predicted gene structures

High-throughput sequencing of genomes and cDNA libraries is sometimes described as a 'data-driven' approach to biology, in contrast to the traditional hypothesis-driven approach. In keeping with this spirit, gene prediction systems are typically run on entire genomes, the results are published or distributed on web sites, and it is hoped that some of the predictions might influence the hypotheses pursued by experimental biologists. Recently, however, the limitations of the data-driven approach to elucidating gene structures by sequencing random cDNA clones have become apparent [19]. At the same time, the increasing accuracy of dual-genome prediction methods has led to a shift whereby experimental follow-up is being scaled up, rather than being left to individual investigators. For example, Guigo *et al.* [14••] performed RT-PCR and direct sequencing on short segments of hundreds of predicted mouse genes that were not in the annotation produced by the ENSEMBL pipeline. They found that verification rates were very high for mouse predictions that were similar to human predictions with at least one conserved intron location. Moreover, the genes that were verified in this way were expressed in significantly fewer tissues, on average, than previously known genes. More recently, Wu *et al.* [15] showed that complete predicted genes could be amplified and sequenced from primers in the UTRs with high success rates, even for genes not found by the pipeline approach (Figure 1). Thus, high-throughput biology has come full circle, incorporating a hypothesis-driven approach. Unlike traditional methods, however, these hypotheses are being generated in the thousands by increasingly accurate *de novo* gene prediction programs.

Concluding remarks

De novo gene prediction for compact eukaryotic genomes is already quite accurate. Although mammalian gene prediction lags behind in accuracy, it is yielding ever more useful results. In particular, the use of *de novo* gene predictions as hypotheses to drive experimental annotation based on systematic RT-PCR and sequencing will improve mammalian annotation greatly in the coming year. As the new approaches described above are integrated with state-of-the-art gene finders, *de novo* accuracy can be expected to improve further. Eventually, a better understanding of the molecular mechanisms involved in gene expression, and the incorporation of this knowledge into the theoretical models underlying *de novo* gene predictors, may lead to systems that are accurate enough to render both experimental verification and manual curation largely unnecessary.

Acknowledgements

We thank David Shteynberg for help with Figure 1 and Josep F Abril for help with Figure 2. MB is supported, in part, by grants HG02278, HG003150 and AI051209 from the National Institutes of Health, and DBI-0091270 from the National Science foundation. RG is supported by grants from the Plan Nacional de I+D (Spain), QLK3-CT-2002-02062 from the European Community and HG003150-01 from the National Institutes of Health.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78-94.

2. Kulp D, Haussler D, Reese MG, Eeckman FH: **A generalized hidden Markov model for the recognition of human genes in DNA**. *Proc Int Conf Intell Syst Mol Biol* 1996, **4**:134-142.

3. Krogh A: **Two methods for improving performance of an HMM and their application for gene finding**. *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:179-186.

4. Guigó R, Knudsen S, Drake N, Smith T: **Prediction of gene structure**. *J Mol Biol* 1992, **226**:141-157.

5. Salamov AA, Solovyev VV: ***Ab initio* gene finding in *Drosophila* genomic DNA**. *Genome Res* 2000, **10**:516-522.

6. •• Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R: **Comparative gene prediction in human and mouse**. *Genome Res* 2003, **13**:108-117.

This paper describes the program SGP-2, and its application to the comparative analysis of the human and mouse genomes. SGP-2 is an extension of the GENEID program that incorporates genome alignments obtained with TBLASTX. The paper describes the details of the implementation and a number of factors that influence performance.

7. •• Alexandersson M, Cawley S, Pachter L: **SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model**. *Genome Res* 2003, **13**:496-502.

This paper introduces the generalized pair HMM, a hybrid of generalized HMMs for gene prediction and pair HMMs for sequence alignment, and describes its implementation in the SLAM program. In SLAM, the gene prediction and the sequence alignment are obtained simultaneously. SLAM, together with SGP-2 and TWINSCAN, was used in the comparative analysis of the human and mouse genomes.

8. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction**. *Bioinformatics* 2001, **17(suppl 1)**:S140-S148.

9. •• Flicek P, Keibler E, Hu P, Korf I, Brent MR: **Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map**. *Genome Res* 2003, **13**:46-54.

This paper describes how the program TWINSCAN was used to annotate the entire human genome, using alignments from the entire mouse genome. This was one of the most specific *de novo* annotations of the human genome ever, predicting about 60% as many exons and genes as GENSCAN while achieving a slighter higher sensitivity to known exons and genes. The paper includes a detailed analysis of TWINSCAN's accuracy and the factors that influence it, such as the number of available shotgun reads from the informant genome.

10. •• Siepel AC, Haussler D: **Computational identification of evolutionarily conserved exons**. In *RECOMB 2004: Proceedings of the Eighth Annual International Conference on Computational Molecular Biology: 2004 March 27–31; San Diego*. New York: ACM Press: 2004:177-186.

This paper describes a gene prediction system that uses a multiple alignment among several vertebrate genomes to predict gene structures simultaneously in all the input sequences. It is based on a phylogenetic-HMM, a generalization of the pair HMM that models the relationship between the functions of genomic nucleotides and their patterns of evolution through a phylogenetic tree. The particular phylo-HMM model described in this paper is the most sophisticated and fully developed model proposed so far. Such methods are expected to lead to greater gene prediction accuracy in the near future.

11. Pedersen JS, Hein J: **Gene finding with a hidden Markov model of genome structure and evolution**. *Bioinformatics* 2003, **19**:219-227.

12. Allen JE, Pertea M, Salzberg SL: **Computational gene prediction using multiple sources of evidence**. *Genome Res* 2004, **14**:142-148.

13. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T *et al.*: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**:38-41.

14. Guigó R, Dermitzakis ET, Agarwal P, Ponting C, Parra G,
•• Reymond A, Abril JF, Keibler E, Lyle R, Ucla C *et al.*: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes**. *Proc Natl Acad Sci USA* 2003, **100**:1140-1145.

This paper describes a computational pipeline that exploits the human and mouse genome sequences to produce a set of gene predictions with a very high rate of experimental verification by RT-PCR. The first stage of the pipeline is the analysis of the genome sequence with dual-genome predictors. The second stage of the pipeline is based on the observation that almost all mouse genes have a human counterpart with highly conserved exonic structure. Therefore, computational predictions are retained only if the protein predicted in mouse aligns with a predicted human protein with at least one predicted intron in the same location. Experimental verification rates for these predictions reached 76%.

15. Wu JQ, Shteynberg D, Arumugam M, Gibbs RA, Brent MR: **Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing**. *Genome Res* 2004, **14**:665-671.

16. Dewey C, Wu JQ, Cawley S, Alexandersson M, Gibbs R, Pachter L: **Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat**. *Genome Res* 2004, **14**:661-664.

17. Zhang Z, Harrison PM, Liu Y, Gerstein M: **Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome**. *Genome Res* 2003, **13**:2541-2558.

18. Torrents D, Suyama M, Zdobnov E, Bork P: **A genome-wide survey of human pseudogenes**. *Genome Res* 2003, **13**:2559-2567.

19. The MGC Project Team: **The status, quality and expansion of the NIH full-length cDNA project (MGC)**. *Genome Res* 2004, **14**:in press.

20. Zhang L, Pavlovic V, Cantor CR, Kasif S: **Human-mouse gene**
• **identification by comparative evidence integration and evolutionary analysis**. *Genome Res* 2003, **13**:1190-1202.

This paper provides a detailed analysis of three important components of comparative gene predictors: the selection of the most appropriate reference genome; the selection of the most appropriate comparative features to be included in the gene prediction framework; and the selection of the architecture to integrate these comparative features.

21. Kotlar D, Lavner Y: **Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions**. *Genome Res* 2003, **13**:1930-1937.

22. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al.*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.

23. Nekrutenko A, Chung WY, Li WH: **An evolutionary approach reveals a high protein-coding capacity of the human genome**. *Trends Genet* 2003, **19**:306-310.

24. Nekrutenko A, Chung WY, Li WH: **ETOPE: evolutionary test of predicted exons**. *Nucleic Acids Res* 2003, **31**:3564-3567.

25. Moore JE, Lake JA: **Gene structure prediction in syntenic DNA segments**. *Nucleic Acids Res* 2003, **31**:7271-7279.

26. Noguchi H, Yada T, Sakaki Y: **A novel index which precisely derives protein coding regions from cross-species genome alignments**. *Genome Inform Ser Workshop Genome Inform* 2002, **13**:183-191.

27. Parra G, Blanco E, Guigo R: **GeneID in *Drosophila***. *Genome Res* 2000, **10**:511-515.

28. Guigó R, Wiehe T: **Gene prediction accuracy in large DNA sequences**. In *Frontiers in Computational Genomics*. Edited by Koonin EV, Galperin MY. Norfolk, UK: Caister Academic Press; 2003:1-33. [Saier MH Jr (Series Editor): Functional Genomics Series, vol 3.]

29. Wang M, Buhler J, Brent MR: **The effects of evolutionary distance on TWINSCAN, an algorithm for pairwise comparative gene prediction**. In *The Genome of Homo Sapiens*. Edited by Stillman B, Stewart D. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press; 2004:125-130.

30. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome**. *Science* 2003, **299**:1391-1394.

31. Holmes I, Bruno WJ: **Evolutionary HMMs: a Bayesian approach to multiple alignment**. *Bioinformatics* 2001, **17**:803-820.

32. Siepel AC, Haussler D: **Combining phylogenetic and hidden Markov models in biosequence analysis**. In *RECOMB 2003: Proceedings of the Seventh Annual International Conference on Computational Molecular Biology: 2003 April 10–14; Berlin*. Edited by Miller W, Vingron M, Istrail S, Pevzner P, Waterman MS. New York: ACM Press; 2003:277-287.

33. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E *et al.*: **Database resources of the National Center for Biotechnology Information: update**. *Nucleic Acids Res* 2004, **32**:D35-D40.

34. Birney E, Durbin R: **Using GeneWise in the *Drosophila* annotation experiment**. *Genome Res* 2000, **10**:547-548.

35. Howe KL, Chothia T, Durbin R: **GAZE: a generic framework for the integration of gene-prediction data by dynamic programming**. *Genome Res* 2002, **12**:1418-1427.

36. Pavlovic V, Garg A, Kasif S: **A Bayesian framework for combining gene predictions**. *Bioinformatics* 2002, **18**:19-27.

37. Volfovsky N, Haas BJ, Salzberg SL: **Computational discovery of internal micro-exons**. *Genome Res* 2003, **13**:1216-1221.

38. Meyer IM, Durbin R: **Gene structure conservation aids similarity based gene prediction**. *Nucleic Acids Res* 2004, **32**:776-783.

39. Brendel V, Xing L, Zhu W: **Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus**. *Bioinformatics* 2004, in press.

40. Tolstrup N, Rouze P, Brunak S: **A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites**. *Nucleic Acids Res* 1997, **25**:3159-3163.

41. Zhang L, Luo L: **Splice site prediction with quadratic discriminant analysis using diversity measure**. *Nucleic Acids Res* 2003, **31**:6214-6220.

42. Zhang XH, Heller KA, Hefter I, Leslie CS, Chasin LA: **Sequence**
•• **information for the splicing of human pre-mRNA identified by support vector machine classification**. *Genome Res* 2003, **13**:2637-2650.

This paper describes an analysis of sequence patterns found in intronic regions upstream of splice acceptors and downstream of splice donors. Current gene prediction programs predict splice site locations using only the regions relatively near to a potential splice site, but this paper suggests that gene prediction may be improved by extending splicing models further into the intron.

43. Saeys Y, Degroeve S, Aeyels D, Van De Peer Y, Rouze P: **Fast feature selection using a simple estimation of distribution algorithm: a case study on splice site prediction**. *Bioinformatics* 2003, **19(suppl 2)**:II179-II188.

44. Yeo G, Burge CB: **Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals**. In *RECOMB 2003: Proceedings of the Seventh Annual International Conference on Computational Molecular Biology: 2003 April 10–14; Berlin*. Edited by Miller W, Vingron M, Istrail S, Pevzner P, Waterman MS. New York: ACM Press; 2003:322-331.

45. Arita M, Tsuda K, Asai K: **Modeling splicing sites with pairwise correlations**. *Bioinformatics* 2002, **18(suppl 2)**:S27-S34.

46. Castelo R, Guigo R: **Splice site identification by idlBNs**. *Bioinformatics* 2004, in press.

47. Majewski J, Ott J: **Distribution and characterization of regulatory elements in the human genome**. *Genome Res* 2002, **12**:1827-1836.

48. Fairbrother WG, Yeh RF, Sharp PA, Burge CB: **Predictive**
• **identification of exonic splicing enhancers in human genes**. *Science* 2002, **297**:1007-1013.
This paper describes a statistical analysis of over-represented oligonucleotides in the sequence of exons with weak splice signals. The analysis led to the identification of ten exonic splicing enhancers, which were subsequently found to display enhancer activity *in vivo*.

49. Weir M, Rice M: **Ordered partitioning reveals extended splice-**
• **site consensus information**. *Genome Res* 2004, **14**:67-78.
This paper reports that long introns have stronger splice site consensus signals than short introns. Incorporating this relationship into gene prediction models may help to overcome the difficulty in predicting long introns, particularly in combination with better models of the intron length distribution itself.

50. Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns**. *Proc Natl Acad Sci USA* 2001, **98**:11193-11198.

51. Wang J, Li S, Zhang Y, Zheng H, Xu Z, Ye J, Yu J, Wong GK: **Vertebrate gene predictions and the problem of large genes**. *Nat Rev Genet* 2003, **4**:741-749.

52. Stanke M, Waack S: **Gene prediction with a hidden Markov**
• **model and a new intron submodel**. *Bioinformatics* 2003, **19(suppl 2)**:II215-II225.
This paper describes a single-genome *de novo* gene prediction program that is based on an HMM model with number of elegant features that have not been previously described. The most significant is a new compromise between speed and accuracy that allows relatively accurate modeling of the distribution of intron lengths to be used without imposing an unacceptably long running time.

53. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M *et al.*: **The transcriptional activity of human chromosome 22**. *Genes Dev* 2003, **17**:529-540.

54. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, McDonagh PD, Loerch PM, Leonardson A, Lum PY, Cavet G *et al.*: **Experimental annotation of the human genome using microarray technology**. *Nature* 2001, **409**:922-927.

55. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR: **Large-scale transcriptional activity in chromosomes 21 and 22**. *Science* 2002, **296**:916-919.

56. Suzuki Y, Yamashita R, Nakai K, Sugano S: **DBTSS: database of human transcriptional start sites and full-length cDNAs**. *Nucleic Acids Res* 2002, **30**:328-331.

57. Bajic VB, Seah SH: **Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units**. *Genome Res* 2003, **13**:1923-1929.

58. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nat Rev Genet* 2004, **5**:276-287.

59. Eyras E, Caccamo M, Curwen V, Clamp M: **ESTGenes: alternative splicing from ESTs in Ensembl**. *Genome Res* 2004, in press.

60. Thanaraj TA, Clark F, Muilu J: **Conservation of human alternative splice events in mouse**. *Nucleic Acids Res* 2003, **31**:1-9.

61. Burge C: **Identification of genes in human genomic DNA [PhD Thesis]**. Stanford, CA: Stanford University: 1997.

62. Cawley SL, Pachter L: **HMM sampling and applications to gene finding and alternative splicing**. *Bioinformatics* 2003, **19(suppl 2)**:II36-II41.

63. Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigo R, Gladyshev VN: **Characterization of mammalian selenoproteomes**. *Science* 2003, **300**:1439-1443.

64. Castellano S, Novoselov SV, Kryukov GV, Lescure A, Blanco E, Krol A, Gladyshev VN, Guigo R: **Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution**. *EMBO Rep* 2004, **5**:71-77.

65. Burge CB, Tuschl T, Sharp PS: **Splicing precursors to mRNAs by the spliceosomes**. In *The RNA World*. Edited by Gesteland RF, Cech TR, Atkins J. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 1999:chapter 20.