Steady progress and recent breakthroughs in the accuracy of automated genome annotation

Michael R. Brent

Abstract | The sequencing of large, complex genomes has become routine, but understanding how sequences relate to biological function is less straightforward. Although much attention is focused on how to annotate genomic features such as developmental enhancers and non-coding RNAs, there is still no higher eukaryote for which we know the correct exon–intron structure of at least one ORF for each gene. Despite this uncomfortable truth, genome annotation has made remarkable progress since the first drafts of the human genome were analysed. By combining several computational and experimental methods, we are now closer to producing complete and accurate gene catalogues than ever before.

cDNA library
A collection of clones that propagate and amplify copies of diverse (usually random) cDNA sequences.

Cis alignment
The alignment of a cDNA sequence to the locus that matches it best in its source genome — the presumed template for its transcription.

Trans alignment
The alignment of a cDNA or protein sequence to a homologous locus other than the one from which it was transcribed.

De novo gene prediction
An approach to gene prediction in which the only inputs are genome sequences; no evidence derived from RNA is used.

Center for Genome Sciences, Campus BOX 8510, Washington University, 4444 Forest Park Blvd, Saint Louis, Missouri 63108, USA.
e-mail: brent@cse.wustl.edu
doi:10.1038/nrg2220

The aim of genome annotation efforts is to determine the biochemical and biological function, if any, of each nucleotide in a genome. A complete annotation would include the exon–intron structures of all RNA products, the coding regions of those that encode proteins and the *cis* and *trans* factors that control their transcription. The techniques used for annotating transcriptional control sites, non-coding RNAs and mRNAs are distinct, and each deserves its own Review. This article focuses on recent progress in the automated annotation of protein-coding genes. Henceforth, genome annotation refers to the annotation of the exon–intron structures of the coding portions of protein-coding genes (ORF structures). Ten years ago, automated genome annotation had a justifiable reputation for inaccuracy, but steady progress has brought us to the point where automated annotation is more reliable than it has ever been.

Genome annotation is best carried out by combining several methods. For example, genes that are highly expressed in several tissues can be annotated easily by sequencing randomly selected clones from cDNA libraries; these sequences can then be aligned to the most similar region of the genome (*cis* alignment) (BOX 1). However, such alignments typically leave approximately 20–40% of genes without any complete, annotated ORF structures (for human annotations, see REF. 1). Some of this remainder can be accurately annotated by aligning cDNA, either from the same organism or from another, to homologous loci rather than the ones from which they were transcribed (*trans* alignment); other genes are best annotated by methods that exploit patterns in genomic sequences rather than evidence derived from RNA (*de novo* gene prediction). Automated systems can also be trained to apply the best combination of evidence types to determine the location of each exon and intron. The strengths and weaknesses of each of these approaches, and how their accuracy depends on the sequence resources that are available for the target genome and its phylogenetic neighbours, are discussed in this Review.

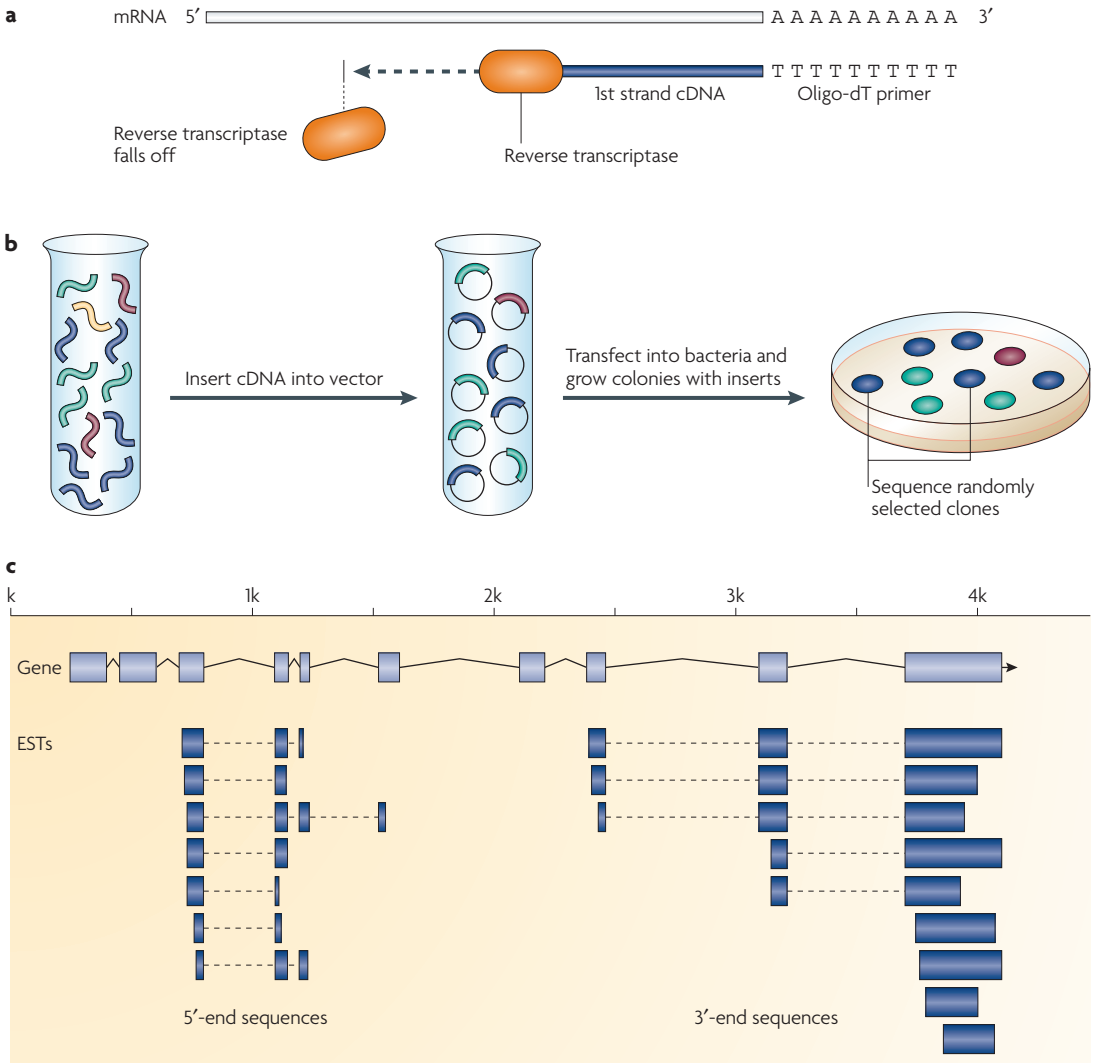
The rapidly increasing number and phylogenetic density of sequenced genomes is increasing the potential power of genome comparisons, but developing methods to effectively exploit them to improve the accuracy of *de novo* gene predictors has repeatedly proven to be more challenging than expected. Several systems that have overcome that challenge, leading to successive improvements in accuracy over the past 5 years, are reviewed here. This year has seen the publication of the first three *de novo* gene predictors based on a new modelling framework known as conditional random fields[2–4]. All three represent progress in some aspect of gene prediction, and one yielded a significant breakthrough in overall predictive accuracy[4].

Finally, I describe one scenario for integrating these approaches to create a state-of-the-art annotation system. This system includes the specific amplification and direct sequencing of predicted cDNAs. Aligning the resulting cDNA sequence to the genome determines the details of the exon–intron structure, which can confirm or correct the prediction.

Target genome
The genome to be annotated, as opposed to informant genomes or other supporting sequences. In gene prediction, informant genomes are genome sequences that are aligned to the target genome and used as auxiliary information for annotating it.

Conditional random field
A type of discriminative model that is used for assigning probabilities to possible annotations of a sequence. A discriminative model is a probability model in which the most likely values of hidden variables (for example, annotations of DNA segments) are calculated directly from the observed variable values (for example, the DNA sequences) without using the probability of the observed values.

Box 1 | Use of cDNA and EST sequences for genome annotation



cDNA is created by reverse transcription of RNA to DNA, a process that often terminates before reaching the 5′ end of the RNA (shown in part **a** in the figure). Currently, the most abundant type of cDNA sequence in databases is obtained by creating a cDNA library and selecting clones at random for sequencing. This often results in high-copy-number mRNAs being overrepresented whereas low-copy-number mRNAs are missed entirely[1] (shown in part **b** in the figure; each cDNA colour represents a different gene). Within this category of 'random clone' cDNA, most of the available sequences are ESTs — single sequencing reads of typically 500–700 nucleotides that are taken from one end of the clone insert. Clones can be sequenced from both ends, but even two end-reads might not cover the entire insert (shown in part **c** in the figure).

To use cDNA sequences for genome annotation, one must first align them to the genome. The term '*cis* alignment' describes the alignment of a cDNA sequence to the genomic template from which it was transcribed. Most *cis* alignments of high-quality cDNA sequences are correct, although for some sequences alignment ambiguities are unavoidable (see REF. 60 for an example). However, there is no way to know for certain whether the alignment covers the entire transcription unit. Furthermore, sequencing of randomly selected cDNA clones has not, so far, yielded anything like a complete catalogue of transcripts for any eukaryote. In a typical cDNA sequencing project, roughly 20–40% of transcripts are sequenced incompletely or not at all. These include many of the transcripts that are in low abundance, expressed only under specific conditions or very long. Nonetheless, when a *cis* alignment of cDNA sequence from a single, full-length clone is available, the information it provides is usually accurate.

New, ultra-high-throughput sequencing technologies may provide complete cDNA coverage of more ORFs, both because they do not require cDNA molecules to be cloned before sequencing and because they can sequence a cDNA mixture more deeply than was possible using previous methods. However, they provide much shorter reads than are obtained by Sanger sequencing. Thus, although ultra-high-throughput sequencing will yield reads from more mRNAs, complete transcripts will have to be assembled from many short reads. This poses challenges for determining the complete structure of individual transcript isoforms in the presence of many alternative splices.

Recently, it has become clear that — in mammals at least — the relationship between genes, transcripts and proteins is more complex than was previously thought[5], leading to a re-examination of the term 'gene'[6]. For this Review, it is assumed that a gene is a contiguous stretch of DNA that serves as the template for a set of overlapping transcripts encoding one or more proteins that share most of their coding sequence. In light of the numerous non-coding transcripts and UTR variants in mammalian genomes, it is important to keep in mind that there is currently no definitive, high-throughput experimental method for determining whether a transcript is translated and, if it is, what the exact sequence of its protein product is. Some proteins can be identified by shotgun mass spectrometry but, at the time of writing, they are a clear minority. Until some new technology comes along, the gold standard of annotation will continue to be a full-length cDNA sequence with a computationally inferred ORF.

Cis alignments

Currently, the gold standard for annotating exon–intron structures is *cis* alignment — the alignment of full-length cDNA sequences to their source gene (BOX 1). Many programs are able to align spliced cDNA to a genome with reasonable accuracy; these include EST_GENOME[7], which has a simple probability model and does not use heuristic methods, and GMAP[8], which is extremely fast.

The term 'full length' is sometimes used to describe a sequence that covers the entire cloned cDNA insert, but in this Review it is used to describe a sequence that covers at least the translated region of an mRNA. One cannot know for certain whether an mRNA sequence is full length without independent experimental confirmation of the translated region, but certain library construction methods enrich for complete transcripts[9,10]. Because systematically sequencing entire clone inserts is much more costly than taking single reads from each end — commonly called ESTs (BOX 1c) — most cDNA sequencing projects produce only ESTs. More than half of all transcripts are too long to be covered by ESTs from a single full-length clone, but different clones may have different 5′ ends due to variable degradation of mRNA and variable processivity of reverse transcriptase (BOX 1). Thus, it can be useful to assemble ESTs from different clones[11]; however, there is a risk that ESTs from two clones represent parts of distinct splice forms that are never found in a single molecule.

The most important limitation of cDNA sequences produced from randomly selected clones, however, is the poor representation of sequences that are expressed at low levels or under specialized circumstances. To remedy this shortcoming, such sequences can be supplemented by sequences from cDNAs that have been specifically amplified by using RT-PCR. This method requires a predicted cDNA sequence that can be used for designing PCR primers. Traditionally, RT-PCR has been applied to one or a few genes at a time, but in the last 4 years it has been scaled up to the point where it can now be used as part of a largely automated, high-throughput annotation process[12–16].

Trans alignments

For a large fraction of genes it is difficult to obtain full-length cDNA sequences, even with a concerted effort (BOX 1), and there are many sequenced genomes for which no concerted effort is made. One approach to annotating genes for which there is little or no cDNA sequence is *trans* alignment — aligning cDNAs from homologous genes in the same species or another species. A common way to do this is to align the proteins that are inferred from cDNA sequences (rather than the sequences themselves) in a process that is often called 'protein alignment': each cDNA is conceptually translated into a protein sequence, which is then aligned to the genome by translating it in all three frames on both strands. Various programs are available for this purpose, including BLAT[17], Exonerate[18] and GeneWise[19].

In the following sections, GeneWise is emphasized because it lies at the heart of the ENSEMBL[20] automated annotation system, the annotations of which have been used for the initial analyses of most vertebrate genomes and many other studies. GeneWise is also designed to be robust — it models genome sequencing errors, including frameshifts, so that it can maintain the reading frame of the aligned protein when such errors occur. However, programs such as Exonerate run much faster and have similar accuracy[18].

Strengths and limitations. Annotations based on both *trans* and *cis* alignments often cover many more loci than those that are based on *cis* alignments alone, especially when extensive cDNA sequencing has not been done. *Trans* alignments that cover nearly the entire aligned protein with high identity yield few false-positive annotations — this is probably the most conservative way to annotate loci for which no *cis* alignment is available (FIG. 1, bars and left scale). When only protein alignments that extend to within five amino acids of both ends of the protein are considered (NFL or nearly full-length protein alignments), accuracy drops slowly but steadily as the percent identity of the alignment decreases. Accuracy in the 90–95% identity range is comparable with that of the best multi-genome *de novo* gene predictors (see the discussion on conditional random fields below). The number of loci with high-identity NFL matches of >90% will vary with the evolutionary distance of the nearest organism for which extensive cDNA sequencing has been done. For comparisons among mammals, mice and humans represent an example in which the evolutionary distance is large, but the number of sequenced cDNAs is also large. Approximately 30% of known human genes in the sample analysed in FIG. 1 (line and right scale) had NFL alignments to mouse proteins with >90% identity. As the number of available cDNA sequences increases, the percent identity of the best match will also tend to increase, with cDNAs from less diverged organisms having a greater impact than those from more diverged organisms.

New approaches and discoveries. Several new approaches may improve the accuracy of *trans* alignments. The initial and terminal coding regions (those that are

Shotgun mass spectrometry
A method for simultaneously identifying many of the protein species present in a complex mixture by fragmenting them and precisely measuring the charge-to-mass ratios of the fragments in a mass spectrometer.

Processivity
The tendency of a polymerase to continue to move along a template molecule rather than falling off prematurely.

Robustness
The ability to function well in difficult circumstances or in unexpected circumstances for which it was not designed.

Nearly full-length (NFL) protein alignment
Alignment of a protein sequence to a genome in which the alignment extends to the ends of the protein, or nearly so.

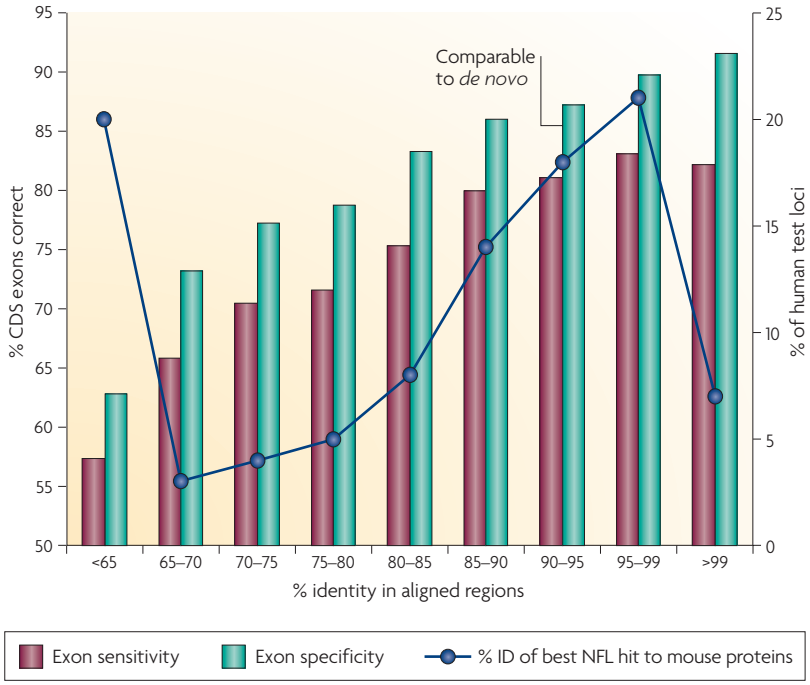


Figure 1 | **Performance of GeneWise, a *trans*-alignment program.** A graph showing that GeneWise is highly accurate for proteins that align to the human genome with at least 90% identity to within five amino acids of both ends (bars). In the 90–95% identity range such nearly full-length (NFL) alignments are roughly comparable in accuracy with *de novo* gene predictors using the mouse genome as an informant. About 30% of known human genes have NFL alignments to mouse proteins with ≥ 90% identity: this is indicated by the blue line, which shows the distribution of percent identity (% ID) when known human ORFs are aligned to the best NFL match among the 64,999 mouse proteins in UNIPROT (a repository of protein sequences). Specificity (represented as green bars) is the fraction of predicted coding exons that exactly matched known human coding exons (CDS); sensitivity (represented as maroon bars) is the fraction of known human coding exons that were predicted correctly. These data are based on coding exons in human ORFs on which HAVANA (the Sanger Institute's manual annotation team) agreed with ENSEMBL annotations (4,726 transcripts in 3,902 loci). Known ORFs were excised from the genome with 5 kb of flanking sequence on each side and searched against UNIPROT by using BLASTP. All matching proteins from *Caenorhabditis elegans*, *Drosophila melanogaster*, *Tetraodon nigrovirides*, zebrafish, frogs, chickens, mice, *Macaca mulatta* and orangutans were extracted. Proteins that aligned to within five amino acids of both their ends were binned by % ID (in aligned regions) and realigned with GeneWise. Only known ORFs that had an NFL protein match in a given bin were included when scoring that bin, so the maroon bars represent sensitivity given that there is a database protein with an NFL alignment. Analysis carried out by M. J. van Baren and M.R.B. using a target gene set provided by M. Schuster and E. Birney (European Bioinformatics Institute, Hinxton, UK).

adjacent to the start and stop codons) present particular problems because they can be short — sometimes just one codon or even part of a codon. Short sequences are difficult to align correctly because they match at many places in the genome. To deal with these cases more effectively, ENSEMBL now aligns both translated cDNAs and the complete nucleotide sequences of the same cDNAs, including the untranslated portions of the exons that contain the start and stop codon. Even when the translated portions of these exons are short, the exons as a whole have a typical exon length. Once the entire exon is located, it is relatively easy to find the coding portion.

Profile hidden Markov model
A mathematical model that represents the conserved elements of an entire family of related proteins or a family of conserved functional domains.

In a similar vein, the locations of introns in a cDNA sequence can be inferred from *cis* alignments and, because they are highly conserved, they can be used to constrain *trans* alignments of the same cDNA (or its translation) to other loci. A preliminary version of this method has been implemented in GeneWise 2.4 (see also REF. 21). It is also possible to *trans* align a profile hidden Markov model representing an entire protein family. This approach promises greater sensitivity for identifying novel family members.

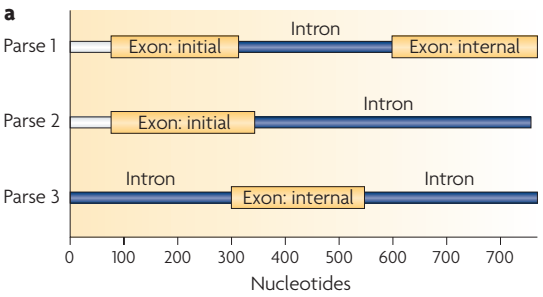
Best practices. *Trans* alignments are reliable for annotating genes that are similar to sequenced, full-length cDNAs. Because some genes are highly conserved, there is a core eukaryotic proteome that can be *trans* aligned to genome sequences nearly anywhere on the evolutionary tree[22]. High-identity NFL *trans* alignments should be used wherever they are available and *cis* alignments are not. NFL alignments of >95% identity should always be used, those of 90–95% are usually at least as good as *de novo* predictions, especially when there are few other sequenced genomes at useful evolutionary distances, when the target genome is fragmentary or when appropriate training data for *de novo* systems is unavailable.

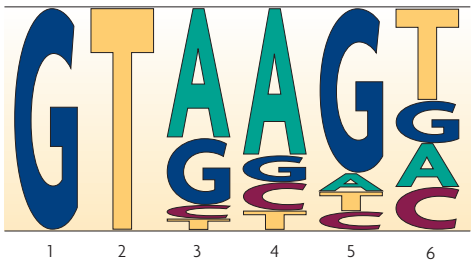
De novo gene prediction

De novo gene prediction is an approach in which the sequences of one or more genomes are the only inputs — no information derived from RNA or protein is used. It works primarily by recognizing genomic sequence patterns that are characteristic of splice donor and acceptor sites and translation initiation and termination sites. When such signal sequences are separated by genomic regions that do not contain any in-frame stop codons, they constitute potential coding exons. *De novo* gene predictors assign probability scores to many potential coding exons, and then join consecutive high-scoring exons with consistent reading frames to form high-scoring exon–intron structures (for a tutorial, see REF. 23) (BOX 2).

Single-genome de novo gene prediction. GENSCAN[24], which was released in 1997, represented a significant improvement in accuracy over previous gene predictors for eukaryotic genomes. GENSCAN is a 'single-genome *de novo* predictor', meaning that it takes the sequence of a single genome as its only input. GENSCAN was much more accurate and better suited to annotating whole genomes than its predecessors. This was partly because it was the first system designed to predict any number of complete ORF structures transcribed from either strand of the genome. Earlier programs were designed primarily to analyse a short segment of genomic DNA that was believed to contain a single gene of interest. GENSCAN was also highly robust — after training on 600 human genes, it was used successfully on *Drosophila melanogaster* and, with a change to the average intron length, on *Arabidopsis thaliana* (see REF. 25 and the GENSCAN parameter files). However, there are limits to this robustness, and because there was no software

Box 2 | **How single-genome *de novo* predictors work**





Graphical representation of a WMM model for intron starts

A DNA sequence can be annotated with many possible exon–intron structures, or parses, nearly all of them incorrect (panel **a** shows three examples). A gene predictor consists of a method for assigning a score to each possible parse, called the probability model, and a method for finding the highest-scoring parse, called the decoding algorithm.

Until late 2007, the most accurate *de novo* gene predictors were based on generalized hidden Markov models (GHMMs). GHMMs assign each parse a score that represents the probability that the parse is correct, given the sequence to be annotated: Pr(*parse* | *seq*). GHMMs are generative models, meaning that they compute this probability using Bayes' rule:

$$\Pr(parse \mid seq) = C \Pr(parse) \Pr(seq \mid parse) \quad (1)$$

where C is a constant that does not depend on which parse is being evaluated. Pr(*parse*) is the probability of the parse regardless of the sequence to which it is applied. The last component of equation 1, Pr(*seq* | *parse*) is the probability of the input sequence, given that the parse is correct. In a GHMM, this probability is computed by multiplying independent parts, one for each segment of the parse.

For example, for parse 1 in panel **a**:

$$\begin{aligned}\Pr(seq \mid \text{parse 1}) = {} & \Pr(seq(1,80) \mid \text{Intergenic}(1,80)) \times \\ & \Pr(seq(80,320) \mid \text{Exon:initial}(80,320)) \times \\ & \Pr(seq(320,600) \mid \text{Intron}(320,600)) \times \\ & \Pr(seq(600,800) \mid \text{Exon:internal}(600,800))\end{aligned}$$

where *seq*(1,80) is the first 80 nucleotides of the sequence and intergenic(1,80) is the assertion that the correct annotation of those 80 nucleotides is 'intergenic'. The model specifies how to compute each term. For example, Pr(*seq*(320,600) | intron(320,600)) would typically be decomposed into three components: the splice donor, the central region and the splice acceptor.

The donor probabilities could be represented by a weight matrix model (WMM)[61], which decomposes the probability of a sequence occurring as a donor site into the probabilities of each of its bases occurring in the corresponding position of a donor site (the probability of each base is proportional to the height occupied by the base letter in the WMM logo as shown in panel **b**). For example, ~99% of introns begin with GT. The third nucleotide is ~50% A and ~45% G, whereas the fourth is ~70% A. By compiling these statistics for the first six nucleotides of the intron and assuming that the base found in each position is independent of those found in other positions, one can estimate a simple probability model. This model can be used to calculate the probability of any hexamer occurring at the beginning of an intron.

Estimating probabilities from examples is called training the model. In a hypothetical donor-site training set containing GTAAGT (nine times), GTGGAG (two times), and GTGTTA (five times), AG never appears as the middle two nucleotides, yet a WMM that has been trained on these examples assigns a non-zero probability to GTAGGT because it has seen examples with A in position three and G in position four. Such generalization allows *de novo* gene predictors to identify genes with sequences that are different from those of the genes on which they were trained.

to train GENSCAN specifically for new genomes it was not widely used for annotation of single-celled fungi and parasites. For example, the annotation of the malaria parasite was done using GlimmerM[26] rather than GENSCAN.

After GENSCAN, the accuracy of *de novo* gene prediction for higher eukaryotes did not improve until the development of dual-genome *de novo* predictors: these pick up a signal from natural selection by considering an alignment between the genome to be annotated (the target) and the genome of a related organism (the informant) (FIG. 2).

Exploiting genome comparisons. The frequency and pattern of mutations in orthologous genomic sequences from two or more species provide valuable information about the function of the sequences. A sufficiently high degree of similarity suggests that natural selection is weeding out mutations through negative selection, which implies that the sequence has a biological function. Furthermore, the specific pattern of substitutions can provide information about the function itself.

Two of the most powerful signals of protein-coding function are: a concentration of substitutions in alignment columns that are separated by multiples of three (reflecting substitutions in the third position of the codon, many of which are silent) and insertions and deletions with lengths that are in multiples of three. When there is a frameshifting insertion or deletion, the reading frame is usually restored by another insertion or deletion nearby. One measure of this, called reading frame consistency (RFC), was pioneered for the analysis of *Saccharomyces cerevisiae* by comparison with three other budding yeasts[27]. RFC works particularly well in *S. cerevisiae* because <2% of genes are spliced, so detecting any part of an ORF implies that the protein continues until the next stop codon.

In mammals, the median length of a coding exon is <50 amino acids, and there is no definitive signal for the end of the exon analogous to the stop codon at the end of the ORF. Thus, conservation alone is not sufficient for accurately detecting exon–intron boundaries, nor for predicting which exons are spliced together to form a complete ORF. Accurate prediction of exon boundaries and protein products requires models of splice sites and splice-site conservation, among other sequence patterns. To accurately annotate mammalian genomes, the signal from natural selection must be combined with the signals picked up by *de novo* gene predictors such as GENSCAN. Programs that do this are known as dual- and multi-genome *de novo* gene predictors.

Dual-genome de novo gene prediction. When sequencing of the mouse genome began, the computational methods for using it to improve the accuracy of automated human gene annotation had not been developed. The intuition was that the coding exons would be conserved and nearly everything else would have mutated beyond recognition. Interestingly, it turned out that less than one-third of the sequences that have been under negative selection since the mouse–human

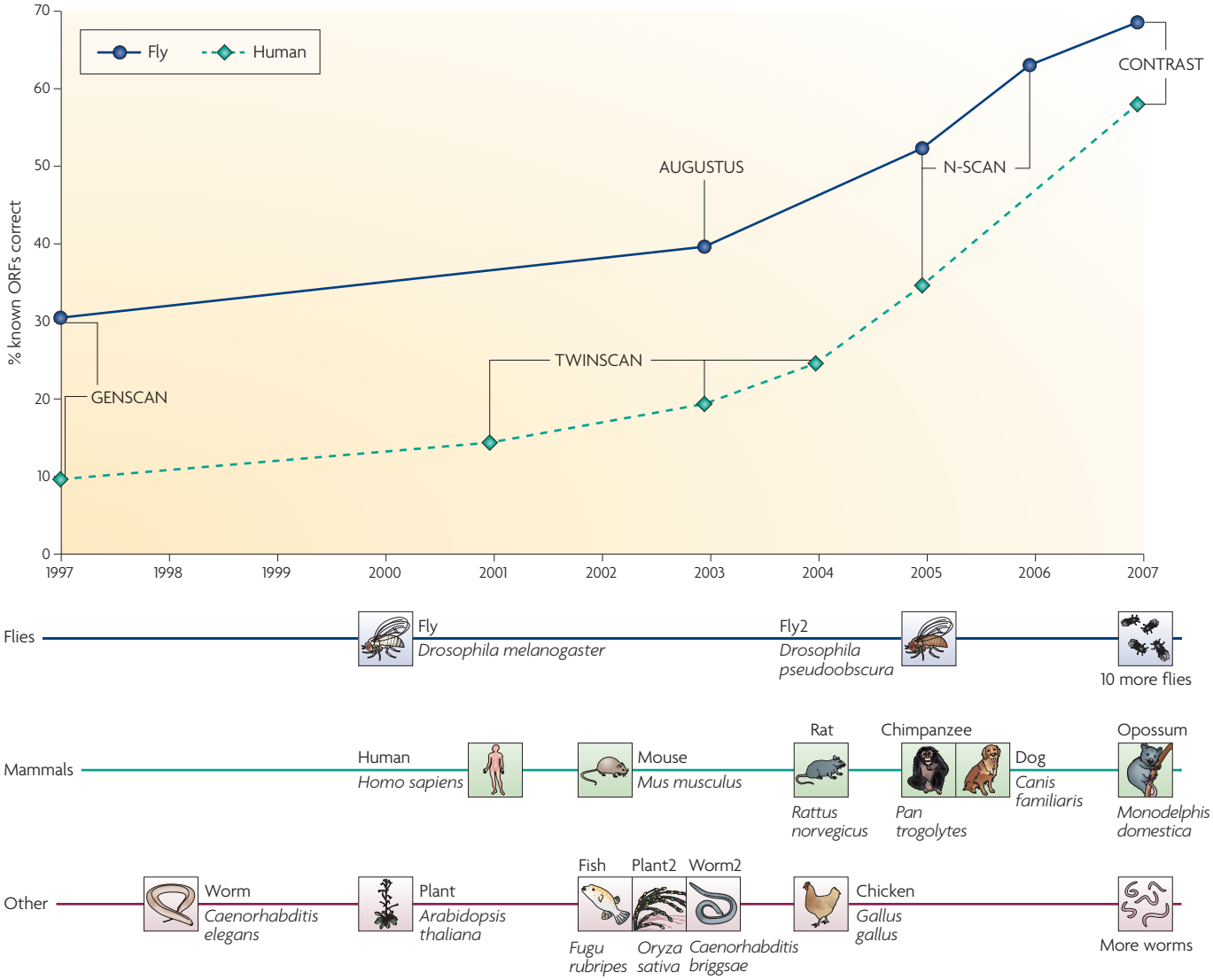


Figure 2 | **The steadily increasing accuracy of *de novo* gene prediction algorithms.** The graph shows the rise in the accuracy of *de novo* gene prediction programs since 1997 (when GENSCAN was introduced) and the dates on which genome sequences were first published. The measure of accuracy is ORF sensitivity — the fraction of known ORFs that are predicted exactly right, that is, yielding the correct protein. GENSCAN[24] and AUGUSTUS[62] use only the target genome, TWINSCAN[29] uses one informant and N-SCAN[39] and CONSTRAST[4] can use multiple informants. The graph reflects historical trends but is not a precise benchmarking of these programs on identical test data.

Training data
In *de novo* gene prediction, it is a set of known gene structures with the corresponding genomic sequence (and alignments to informant genomes, if available). Training data are used in specializing the probability model to fit the characteristics of a particular genome.

Parse
A segmentation of a string of letters together with a labelling of the segments.

Bayes' rule
A mathematical identity ($Pr(x|y) = Pr(y|x)\ Pr(x)/Pr(y)$) that allows one to swap variables in a conditional probability expression.

split encode amino acids[28]. Furthermore, although the average similarity of functional orthologous sequences is much higher than that of non-functional orthologous sequences, the two distributions overlap considerably[28].

By the year 2000, several groups had developed methods for combining information from mouse–human alignments with models of the DNA sequences that characterize splice donors and acceptors, start and stop codons and other biological features. The first programs to outperform GENSCAN by using mouse–human comparison were TWINSCAN[29,30] and SGP2 (REF. 31). Their success resulted, in part, from using genome alignments to modify the scoring schemes of successful single-genome *de novo* gene predictors (GENSCAN and GENE-ID[32], respectively). The biggest difference between them was that TWINSCAN included models of conservation in splice sites and start and stop codons, whereas SGP2 considered only the conservation in protein-coding regions. After training on known human genes with mouse alignments, the predictions of both programs were still influenced more by the patterns in the human DNA sequence than by the mouse alignments. For TWINSCAN, the primary effect of mouse–human alignments was to eliminate many of the false-positive genes and exons predicted by GENSCAN: TWINSCAN predicted 25,600 genes (versus approximately 45,000) and 198,000 exons (versus approximately 315,000). For comparison, current best estimates place the number of human protein-coding genes at 20,000–21,000 (REF. 33).

Empirical gene prediction studies on the human genome have consistently found that the mouse genome (at about 0.6 substitutions per synonymous site[28]) is close to having the optimal degree of divergence for comparison[4,34,35]. For *D. melanogaster*, the best single informant genome is *Drosophila ananassae* (R. Brown, personal communication), with approximately one substitution per synonymous site[36]. *Drosophila willistoni*, at approximately 1.2 substitutions per synonymous site[36], is much less useful (FIG. 3a). However, substitutions per synonymous site might not be a good predictor of the usefulness of informant genomes because the number of substitutions per synonymous site is calculated using only proteins that can be aligned. As a result, it does not account for the loss of alignability at greater evolutionary distances. For flies, a good predictor is the total number of mismatches (that is, observed substitutions) in the whole-genome alignment divided by the length of the target genome (FIG. 3b). When two genomes are too diverged to be useful, the number of mismatches is low because most of the sequence cannot be aligned; when they are too close to be useful, the number of mismatches is low because most of the sequence is unchanged. However, better models for the dependence of informant utility on divergence are needed.

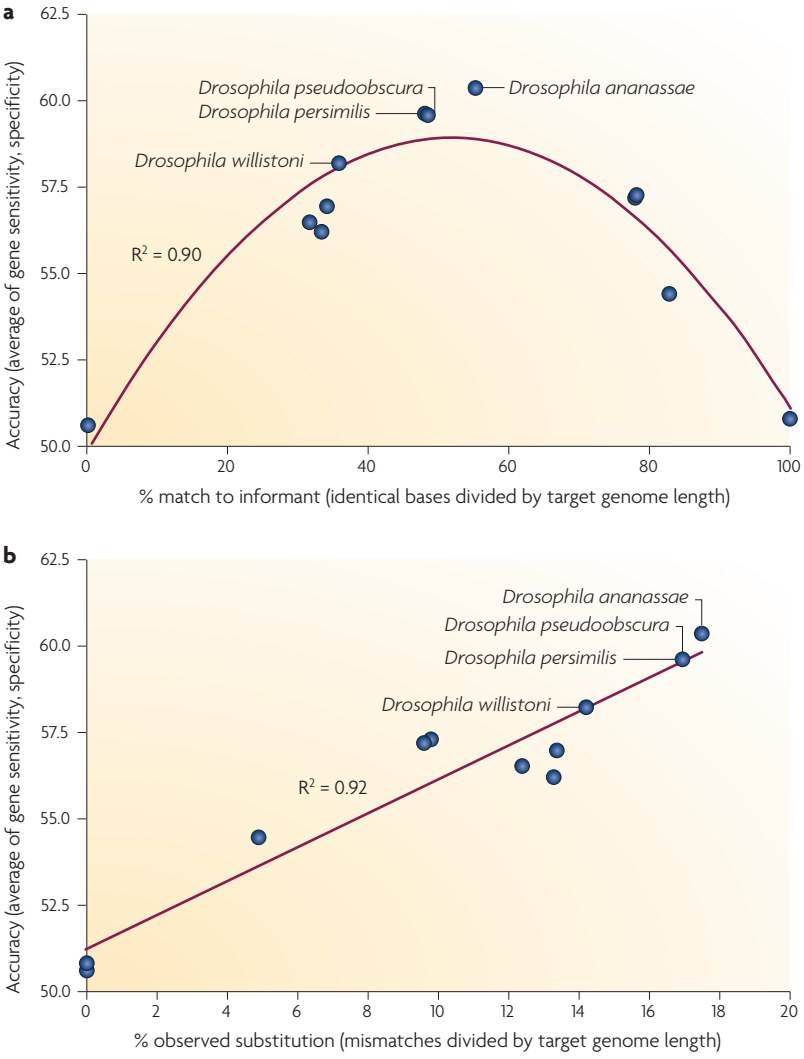


Figure 3 | **Criteria for selecting the best informant genome.** There is an optimal level of divergence between the target genome and a single informant, as shown here for gene prediction in *Drosophila melanogaster*. **a** | Accuracy of gene prediction by N-SCAN in *D. melanogaster* using a single informant genome as a function of genome similarity, with a quadratic regression line. Accuracy is the average of gene sensitivity (the percentage of annotated loci at which N-SCAN predicts the coding region of one transcript correctly) and specificity (the percentage of predicted transcripts that exactly match an annotated transcript throughout their coding regions). The best informant is *Drosophila ananassae* (55% match). The accuracy at 0% match was calculated by using a completely unaligned informant, and the accuracy at 100% match was calculated by using an identical informant base for each target base. R^2 is the percentage of variance accounted for by the regression line. This plot shows that accuracy does peak at an intermediate level of divergence as expected and that a quadratic model fits the data well. **b** | The same accuracy numbers plotted against the percent of the target genome nucleotides that are aligned to a different nucleotide or a small deletion gap in the informant, with a linear regression line. This plot shows that, at least for flies, accuracy can be predicted by an even simpler model (two parameters instead of three) when whole-genome mismatch percentage is used as the predictor variable.

Multi-genome de novo gene prediction. When multiple mammalian genomes became available, using them to improve on the state-of-the-art in *de novo* gene prediction proved more difficult than anticipated. An elegant program called EXONIPHY[38] achieved high specificity on individual exons, but did not attempt to link exons together. The first program that could make use of multiple informant genomes and could predict entire ORFs more accurately than TWINSCAN was N-SCAN[39]. However, N-SCAN was more accurate than TWINSCAN even when both programs used only the mouse genome as the informant. Furthermore, adding rat and chicken alignments left the accuracy of N-SCAN essentially unchanged. The situation was similar for *D. melanogaster*: using *Drosophila yakuba*, *Drosophila pseudoobscura* and *Anopheles gambiae* as informants was barely better than using *D. pseudoobscura* alone. It was not until multiple genomes at the right evolutionary distances became available that a combination of informants yielded a non-negligible accuracy improvement (approximately 4% gene sensitivity) over the best single informant (R. Brown, personal communication).

Recently, a program called CONTRAST has extracted bigger gains in human gene prediction from multi-genome alignments[4] (see the section below on conditional random fields). This work suggests that using both the mouse and the opossum, which is slightly more diverged, will give the best improvement over using the mouse alone.

Box 3 | **Generative versus discriminative models**

This box explains the difference between generative models, such as generalized hidden Markov models (GHMMs), and discriminative models, such as conditional random fields (CRFs).

Generative models

To assign a score to a parse, we want to know the probability that the parse is correct, given the DNA sequence to be annotated. This is written Pr(*parse* | *seq*), where *seq* is the input sequence to be annotated. Generative models consist of two parts that are combined by using Bayes' rule:

$$\Pr(parse \mid seq) = C \Pr(parse) \Pr(seq \mid parse) \quad (1)$$

where C is a constant that does not depend on which parse is being evaluated. Pr(*parse*) is the probability of the parse regardless of the actual sequence it is applied to. For example, if two-exon genes are rare, then all parses containing two-exon genes should receive low probabilities from this component.

The last component of equation 1, Pr(*seq* | *parse*), is the probability of the input DNA sequence, given that the parse is correct. In other words, of all genes with the exon–intron structure that is specified by *parse*, what fraction of them are expected to have a DNA sequence identical to *seq*? This number is normally very small, because many sequences can have the same exon–intron structure. In a GHMM, this number is computed by multiplying independent parts, each of which refers to a different segment of the sequence, as described in BOX 2.

Discriminative models

CRFs specify the probability of a parse given the input sequence directly using the formula:

$$\Pr(parse \mid seq) = Ce^{WF(seq,\, parse)} \quad (2)$$

where C is a constant that does not depend on which parse is being evaluated. The exponent, *WF*(*seq*, *parse*), is a weighted sum of feature functions:

$$WF(seq, parse) = \sum_j W_j F_j(seq, parse) \quad (3)$$

The predefined function F_j can depend on any part of the input sequence. For example, some functions might count the occurrences of specific splice enhancers with no concern about whether these occurrences overlap. This is possible because the probability of the sequence, Pr(*seq* | *parse*), does not appear in any of the CRF formulas. The weight W_j for each function is learned from examples of correct parses during training.

Negative selection
Sequences are under negative selection when mutations are deleterious to fitness and hence tend to be weeded out over time.

Substitutions per synonymous site
An estimate of evolutionary distance that makes use of silent substitutions in protein-coding regions, similar to the rate of substitutions in fourfold degenerate sites.

Generative model
A probability model in which, to calculate the most likely values of hidden variables (annotations of DNA segments), one must also calculate the probability of the observed variable values (the DNA sequence).

Limitations of comparative genomics. So far, groups that have succeeded in improving accuracy by using multiple informants have consistently reported diminishing returns as the number of informants increases[3,4,39], although it is now possible to achieve small improvements by adding third, fourth and fifth informants[4]. Phylogenetic trees for vertebrates and flies show that these diminishing returns are not primarily due to diminishing increases in total branch length.

Three possible reasons for the diminishing returns are sequencing error, alignment error and change of nucleotide function. Sequencing errors make the affected genomic regions appear to be changing faster than they really are. Alignment errors tend to eliminate what ought to be gaps and mismatches, making the affected regions appear to be changing more slowly than they really are. Finally, biological function is not necessarily conserved — a splice donor in one species might be aligned to an orthologous sequence in a related species that does not have the same function. As the number of aligned genomes grows, so does the noise from sequencing error, alignment error and change of function.

Conditional random fields. For the past 10 years, the most accurate gene prediction programs have all been based on a type of generative model called a generalized hidden Markov model (GHMM)[23] (BOX 2). Recently, a new modelling framework, conditional random fields (CRFs) (BOX 3), has generated a great deal of excitement in computational linguistics and has been applied to several problems in biological sequence analysis[40].

After years of development, CRF-based gene finders burst onto the scene in 2007 (REFS 2–4). The first published CRF gene finder, CRAIG[2], used only the target genome sequence. It was more accurate on mammalian sequences than previous single-genome *de novo* predictors, but it did not beat the best dual-genome predictors. The second CRF gene finder, CONRAD[3], used multiple informant genomes and was more accurate than all previous *de novo* predictors on the genomes of two fungi, *Cryptococcus neoformans* and *Aspergillus nidulans*. However, its training procedure was not designed to be run on mammalian genomes, which are approximately 100 times larger. The third CRF gene finder, CONTRAST[4], surpassed all previous *de novo* systems on both the human and fly genomes. On the human genome, CONTRAST predicts a perfect ORF structure at a stunning 58% of all known protein-coding genes (FIG. 2).

Like previous gene-prediction programs, CONTRAST shows diminishing returns as more informant genomes are added. Unlike previous systems, its accuracy on human genes using 11 informants (one ORF structure exactly right for 58% of genes) is substantially better than its accuracy using the best single informant (50%). Furthermore, the way in which it uses the informants is unique. Multi-genome predictors based on generative models have all used phylogenetic trees to model the patterns found in the columns of multi-genome alignments. This is important for generative models because the phylogenetic tree allows one to separate the probability of a given alignment column into a product that contains one independent factor per branch. In the absence of such a phylogenetic factorization, it is not feasible to compute the probabilities of alignment columns. This is because there is no model for the correlations among the sequences of informant genomes that diverged from one another more recently than they diverged from the target. The CRF framework makes it possible to exploit the information contained in multiple pairwise genome comparisons without assigning a probability to the alignment column.

Of the first three CRF gene finders, the model on which CONRAD is based is most similar to a standard GHMM, and many of its parameters are trained in the same way as GHMM parameters. This shows that it is possible to improve on the GHMM by adding a few weighting parameters and some additional conservation features. CRAIG also looks like a GHMM in many ways, but it includes multiple variants on the standard component models (BOX 2), and all of the parameters are trained in the discriminative framework. CONTRAST is the most distinctive, in that it does not use exon lengths and many of its submodels are quite different from those found in GHMM predictors. Each of these three systems takes a completely different approach to training

parameters, which is likely to remain an active research area for some time to come. However, the experience of the speech recognition community suggests that the training approach used by CONTRAST[41] may contribute significantly to its success.

Best practices. Currently, CONTRAST[4] is the best *de novo* gene predictor for mammals and flies, and it will probably perform well on worms, plants and fungi, although no benchmarks are available for these. CONRAD[3] has the best published results for fungi, and TWINSCAN[42,43] for *Caenorhabditis elegans*. For plants, FGENESH[44] is widely used, although there is evidence suggesting that TWINSCAN is quite accurate on plants[45,46]. It is always best to use at least one informant genome, even when the only choice has been shotgun sequenced to low coverage. Multiple informants are generally better than one, but three informants are nearly as good as ten. Combinations of informants at different distances from the target seem to be better than equally diverged informants. Within the useful range of divergence from the target, it is better to choose informants that are as diverged from one another as possible. When annotating genomes that contain substantial numbers of processed pseudogenes, a program such as PPFINDER[47] should be used to remove fragments of processed pseudogenes from predictions. Alternatively, one of several effective pseudogene detection programs can be used to mask pseudogenes before predicting functional genes[48,49].

Integrating information

Manual integration. The most accurate annotations are obtained by combining information from several of the sources described above. The traditional approach to integration is to present the results of computational analyses — such as *cis* alignments, *trans* alignments, genomic alignments and *de novo* predictions — to human annotators, who can accept or modify the gene structures that are presented. One of the largest manual annotation groups is the HAVANA (Human and Vertebrate Analysis and Annotation) group at the Wellcome Trust Sanger Institute, Cambridge, UK. RT-PCR experiments carried out on 30 Mb of the human genome suggest that the HAVANA annotation includes nearly all of the protein-coding gene loci within the 30 Mb, although the low confirmation rates for its least confident categories suggest that they contain some false positives[50]. Although manual annotation can be effective, it is too expensive to be applied beyond a few key genomes.

Automated integration using evidence hierarchy. The simplest way to automatically combine evidence is by picking the best source of evidence at each genomic site. The first step in this process is to define a hierarchy of sources, from most reliable to least reliable. For example, Pairagon+N-SCAN_EST[52] is a simple system that *cis* aligns cDNA sequences (from RefSeq[51]), and then fills in the gaps between these alignments with predictions. ENSEMBL[20], perhaps the best known and most widely used automated annotation system, is based on an evidence hierarchy.

The limitation of this approach is that it does not consider the quality of any particular alignment or prediction when selecting the best annotation at a given locus. For example, a particular *de novo* gene prediction might be especially strong, on the basis of its conservation pattern and splice sites, whereas a cDNA alignment in the same locus might be weak owing to unusual splice sites or multiple mismatches near the splice sites. An evidence hierarchy cannot weigh these factors in choosing between the *de novo* prediction and the cDNA alignment.

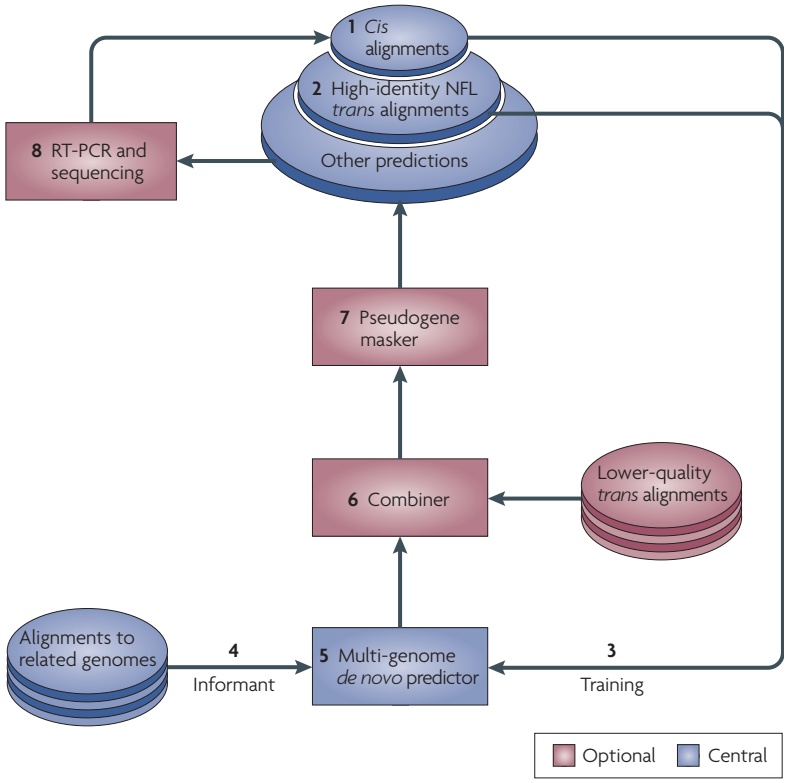
Automated integration using weighted evidence. There are two broad categories of systems that do weigh the quality of evidence at each locus: joint-probability models and combiners. Joint models weigh the evidence from each source before ruling out any possibilities — every possible annotation is considered in light of all the evidence. For example, dual- and multi-genome *de novo* predictors use joint-probability models to combine evidence from the target sequence with evidence from aligned informant genomes. Recently, joint models have been used to incorporate 'hints' such as EST and protein alignments within otherwise *de novo* prediction programs[43,53,54].

Unlike joint models, combiners do not weigh the evidence for every possible annotation; they consider only exon–intron structures that are produced by other programs, including *de novo* and cDNA-based systems. They either choose among these structures or combine elements, such as splice sites, from several different structures. Combiners (such as GAZE[55], GLEAN[56], Combiner[46] and its successors Jigsaw[57]and Genomix[58]) typically use probabilistic models of the relative accuracies of the evidence sources they are combining. The accuracy of each source can be set manually, trained by comparison to trusted annotations or trained without trusted annotations by using the principle that the system that generates predictions that are most often echoed by another system is likely to be the most accurate. This principle is supported by the observation that exons on which multiple predictions agree are nearly always more accurate than those that are predicted by only a single program[14]. Combiners can leverage this effect, creating a single 'centroid' annotation that agrees with the evidence sources more often than the sources agree with each other.

It is not clear why combining the predictions of multiple *de novo* programs nearly always produces better results than any of the programs does on its own. One hypothesis is that each prediction set includes 'random' errors and, because the number of possible errors is large, it is rare for two programs to produce the same error. Alternatively, it could be that each program has some components that are intrinsically better than the corresponding components of the other programs, so that explicitly recombining the best components would produce a single *de novo* predictor that is as good as the consensus predictions. Even if the latter possibility were the case, we do not yet know how to recognize and recombine the best components.

Generalized hidden Markov model
A type of generative model that is used for assigning probabilities to possible annotations of a sequence. Generalized hidden Markov models are preferred over ordinary hidden Markov models for gene prediction because they make it possible to model the distribution of exon lengths.

Box 4 | **How to annotate your genome**



The figure illustrates one approach to annotating a genome. First, cDNA sequences from the target genome are aligned to the location that they match best, yielding a core set of *cis* alignments (step 1). Next, databases are searched for proteins that align to the target genome with high identity across their entire length, or nearly so (NFL, nearly full length; step 2). A training set for a *de novo* predictor is then constructed out of the *cis* alignments and as many of the best *trans* alignments as are needed to make the set large enough (step 3). If whole-genome alignments are available, they should also be fed into the *de novo* predictor (step 4). Once *de novo* predictions have been made (step 5), *trans* alignments with lengths or percent identities that were insufficient for defining gene structures on their own can be included at this point by using a 'combiner' program (step 6), or alternatively by using a *de novo* program with a joint model. If the genome has significant numbers of processed pseudogenes, software for removing pseudogene fragments from the annotation should be used (step 7).

If possible, selected predictions should be targeted for RT-PCR and sequencing (step 8). When successful, these experiments augment the set of *cis* alignments. If the number or the length of *cis* alignments increases significantly, *de novo* training can be repeated. Predictions that fail to amplify should not necessarily be discarded, but they can be demoted to a lower confidence level.

The contribution of each component in the diagram depends on the available sequence resources. The contribution of *cis* alignments is determined by the number and diversity of available cDNAs from the target species. The contribution of *trans* alignments depends on the number and diversity of available cDNAs from related species, as well as the divergence of those species from the target. The identity threshold for *trans* alignments should be chosen so that alignments above the threshold are expected to be more accurate than *de novo* predictions, whereas those below the threshold are expected to be less accurate. The accuracy of *de novo* gene prediction depends on the characteristics of the target genome (intron length and count, for example), the number and divergence of available informant genomes, and the quantity and quality of training data. If there are not enough *cis* and high-identity *trans* alignments to serve as training examples for a *de novo* predictor, it may be possible to use parameters that have been trained on another genome. The quality of the target genome sequence is also a factor — *trans* alignments are more robust than *de novo* predictions when faced with a fragmented target genome that contains many insertion and deletion errors.

Best practices. A recent comparison of several automated systems for integrating evidence, including Pairagon+N-SCAN_EST, ENSEMBL and JIGSAW, found JIGSAW to be the most accurate genome annotation system by a small margin[59]. As with all combiners, results from JIGSAW will depend on the other prediction sets that are available for combination. Thus, a combiner does not eliminate the need for high-quality predictions produced by several other methods.

How to annotate your genome

BOX 4 describes one approach to annotating a newly sequenced genome. At one level, this is an 'evidence hierarchy' approach, represented by the three-layered 'cake' shown at the top of the figure in BOX 4. The bottom layer of the cake consists of predictions that are not supported by full-length *cis* alignments or high-identity, NFL *trans* alignments, which are represented by the top two tiers of the cake. The predictions in the bottom layer may be partially supported or they may be pure *de novo* predictions with no RNA-derived support.

The minimal machinery that is needed to construct this important bottom layer is a state-of-the-art, multi-genome *de novo* predictor. In addition, I recommend combining *de novo* predictions with any *cis* and *trans* alignments that are not of high enough quality to be included in the top two layers of the hierarchy. A combiner program can also fold in multiple sets of *de novo* gene predictions. Depending on how much time and expertise is available, a great deal of effort can be devoted to producing more inputs for the combiner, but combining two good *de novo* gene predictors with *cis*-aligned ESTs can be expected to produce reasonably good results.

Once a gene set for the bottom layer has been produced, it can be used to specifically amplify predicted cDNAs and sequence them by using RT-PCR and direct sequencing. The resulting sequences might verify or indicate modifications to the details of the prediction. It is now practical to carry out thousands of RT-PCR and sequencing reactions[16], which can significantly increase the number of high-confidence annotations in the top layer.

Summary and conclusions

Automated methods for annotating protein-coding genes in complex genomes are continually improving. As we sequence more cDNAs, a larger percentage of genes in each new genome will have high-identity NFL alignments in another genome, so more genes will be accurately annotated by *trans* alignments. New *de novo* gene finders based on CRFs are more accurate than any before them, continuing the trend of increasing accuracy. Furthermore, these systems gain accuracy from multiple informant genomes; so, as more genomes are sequenced, we can expect increasingly accurate *de novo* predictions. The CRF framework also promises to make it easier to encode new biological observations and exploit them for improved accuracy. Finally, automated systems for removing processed pseudogenes from gene prediction sets have helped to eliminate false positives from both *trans* alignments and *de novo* gene predictions.

The main challenge for the annotation of protein-coding genes is identifying the non-coding portions of the transcripts. In mammals, a large fraction of genes have multiple alternative transcription start sites and extensive alternative splicing, especially in the UTRs. Nonetheless, the great improvements we have seen in the annotation of protein-coding genes, together with the increase in publicly available cDNA and genome sequences, have brought us to the threshold of realizing the promise of the genomic era. In the next few years, high-throughput annotation systems will produce accurate and essentially complete catalogues of molecular parts, allowing us to turn our full attention to understanding the dynamics of their interactions.

1. The MGC Project Team. The status, quality, and expansion of the NIH full-length cDNA project: the mammalian gene collection (MGC). *Genome Res.* **14**, 2121–2127 (2004).
2. Bernal, A., Crammer, K., Hatzigeorgiou, A. & Pereira, F. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput. Biol.* **3**, e54 (2007).
This paper presents CRAIG, a CRF-based, single-genome *de novo* gene predictor with the best published accuracy for the human genome among programs that do not use comparison with related genome sequences.
3. Decaprio, D. *et al.* CONRAD: gene prediction using conditional random fields. *Genome Res.* **17**, 1389–1398 (2007).
This paper presents CONRAD, a CRF-based, multi-genome *de novo* gene predictor with the best published benchmark accuracy on fungal genomes.
4. Gross, S. S., Do, C. B., Sirota, M. & Batzoglou, S. CONTRAST: a discriminative, phylogeny-free approach to multiple informant *de novo* gene prediction. *Genome Biol.* (in the press).
This paper presents CONTRAST, a CRF-based, multi-genome *de novo* gene predictor that is currently the most accurate predictor, at least for mammals and flies. CONTRAST is also likely to work well on other complex eukaryotic genomes.
5. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
6. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17**, 669–681 (2007).
7. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13**, 477–478 (1997).
8. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
9. Shibata, Y. *et al.* Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method. *Biotechniques* **30**, 1250–1254 (2001).
10. Suzuki, Y. *et al.* Statistical analysis of the 5′ untranslated region of human mRNA using 'oligo-capped' cDNA libraries. *Genomics* **64**, 286–297 (2000).
11. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
12. Guigó, R. *et al.* Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl Acad. Sci. USA* **100**, 1140–1145 (2003).
13. Wu, J. Q., Shteynberg, D., Arumugam, M., Gibbs, R. A. & Brent, M. R. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. *Genome Res.* **14**, 665–671 (2004).
14. Eyras, E. *et al.* Gene finding in the chicken genome. *BMC Bioinformatics* **6**, 131 (2005).
15. Denoeud, F. *et al.* Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* **17**, 746–759 (2007).
16. Siepel, A. *et al.* Targeted discovery of novel human exons by comparative genomics. *Genome Res.* **17**, 1763–1773 (2007).
This paper shows that *de novo* gene prediction followed by RT-PCR and direct sequencing can be used to elucidate many novel exons and introns even in a genome as thoroughly studied as the human genome.
17. Kent, W. J. BLAT — the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
18. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
19. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
20. Birney, E. *et al.* An overview of ENSEMBL. *Genome Res.* **14**, 925–928 (2004).
21. Meyer, I. M. & Durbin, R. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res.* **32**, 776–783 (2004).
22. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
23. Brent, M. R. How does eukaryotic gene prediction work? *Nature Biotechnol.* **25**, 883–885 (2007).
24. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
25. Pavy, N. *et al.* Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics* **15**, 887–899 (1999).
26. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31 (1999).
27. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
This paper presents the RFC method of identifying protein-coding regions using only multi-genome alignments.
28. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
29. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**, S140–S148 (2001).
30. Flicek, P. & Brent, M. R. Using several pair-wise informant sequences for *de novo* prediction of alternatively spliced transcripts. *Genome Biol.* **7**, S8 (2006).
31. Parra, G. *et al.* Comparative gene prediction in human and mouse. *Genome Res.* **13**, 108–117 (2003).
32. Parra, G., Blanco, E. & Guigo, R. GeneID in *Drosophila*. *Genome Res.* **10**, 511–515 (2000).
33. Clamp, M. *et al.* Distinguishing protein-coding and non-coding genes in the human genome. *Proc. Natl Acad. Sci. USA* (in the press).
34. Wang, M., Buhler, J. & Brent, M. R. in *The Genome of Homo Sapiens* (eds Stillman, B. & Stewart, D.) 125–130 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2004).
35. Zhang, L., Pavlovic, V., Cantor, C. R. & Kasif, S. Human–mouse gene identification by comparative evidence integration and evolutionary analysis. *Genome Res.* **13**, 1190–1202 (2003).
36. Clark, A. G. *et al.* Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218 (2007).
37. Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res.* **13**, 46–54 (2003).
This paper shows that unassembled sequencing reads representing three- to fourfold coverage of an informant genome are almost as useful as a high-coverage informant assembly for *de novo* gene prediction.
38. Siepel, A. C. & Haussler, D. in *RECOMB* (ACM, San Diego, 2004).
39. Gross, S. S. & Brent, M. R. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13**, 379–393 (2006).
This paper presents N-SCAN, a multi-genome *de novo* gene predictor that was the most accurate program for animal genomes until CONTRAST was introduced.
40. Do, C. B., Woods, D. A. & Batzoglou, S. CONRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90–e98 (2006).
41. Gross, S. S., Russakovsky, O., Do, C. B. & Batzoglou, S. Training conditional random fields for maximum labelwise accuracy. *Adv. Neural Inf. Process. Syst.* **19**, (Neural Information Processing Systems Foundation, 2006).
42. Wei, C. *et al.* Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res.* **15**, 577–582 (2005).
43. Wei, C. & Brent, M. R. Using ESTs to improve the accuracy of *de novo* gene prediction. *BMC Bioinformatics* **7**, 327 (2006).
44. Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
45. Moskal, W. A. Jr. *et al.* Experimental validation of novel genes predicted in the un-annotated regions of the *Arabidopsis* genome. *BMC Genomics* **8**, 18 (2007).
46. Allen, J. E., Pertea, M. & Salzberg, S. L. Computational gene prediction using multiple sources of evidence. *Genome Res.* **14**, 142–148 (2004).
47. van Baren, M. J. & Brent, M. R. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res.* **16**, 678–685 (2006).
This paper presents PPFINDER, a program that can remove processed pseudogene fragments from gene predictions even when there is no database of previously known functional genes.
48. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. A genome-wide survey of human pseudogenes. *Genome Res.* **13**, 2559–2567 (2003).
49. Zhang, Z. & Gerstein, M. Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.* **14**, 328–335 (2004).
50. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7**, S4 (2006).
This paper provides useful insights into a modern manual annotation effort and how it compares with both automated annotation and experimental verification.
51. Pruitt, K., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **1**, 501–504 (2005).
52. Arumugam, M., Wei, C., Brown, R. H. & Brent, M. R. Pairagon + N-SCAN_EST: a model-based gene annotation pipeline. *Genome Biol.* **7**, S5 (2006).
53. Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
54. Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.* **7**, S11 (2006).
55. Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res.* **12**, 1418–1427 (2002).
56. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
57. Allen, J. E. & Salzberg, S. L. Jigsaw: integration of multiple sources of evidence for gene prediction. *Bioinformatics* **21**, 3596–3603 (2005).
This paper presents Jigsaw, a highly accurate system for combining predictions that are produced by other methods.

58. Coghlan, A. & Durbin, R. Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron–exon structure. *Bioinformatics* **23**, 1468–1475 (2007).
59. Guigo, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7**, S2 (2006).
This paper describes detailed benchmarks on the accuracy of several gene prediction programs that use a range of methods and evaluating them on 30 Mb of the human genome.
60. Brent, M. R. Genome annotation past, present and future: how to define an ORF at each locus. *Genome Res.* **15**, 1777–1786 (2005).
61. D'Haeseleer, P. What are DNA sequence motifs? *Nature Biotechnol.* **24**, 423–425 (2006).
62. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
This paper presents AUGUSTUS, currently the most accurate GHMM-based, single-genome *de novo* predictor for flies. AUGUSTUS uses innovative splice-site and intron-length models.

Acknowledgements
I am deeply grateful to M. J. van Baren, M. Schuster and E. Birney for help with the GeneWise analysis, R. Brown for analysis of informant genome utility, M. J. van Baren and S. Gross for comments on the manuscript, and L. Kyro and L. Langton for help with figures. M.R.B. is supported in part by grants from the National Institutes of Health (HG002278, HG003700, HG004271) and Monsanto.

FURTHER INFORMATION

Michael R. Brent's homepage: http://cse.wustl.edu/~brent
AUGUSTUS: http://augustus.gobics.de
CONRAD: http://www.broad.mit.edu/annotation/conrad
CONTRAST: http://contra.stanford.edu/contrast
CRAIG: http://alliance.seas.upenn.edu/~strctlrn/craig.html
ENSEMBL: http://www.ensembl.org
EST_GENOME: http://www.well.ox.ac.uk/~rmott/ESTGENOME/est_genome.shtml
Exonerate: http://www.ebi.ac.uk/~guy/exonerate
EXONIPHY: http://compgen.bscb.cornell.edu/~acs/software.html
GAZE: http://www.sanger.ac.uk/Software/analysis/GAZE
GeneID: http://www1.imim.es/software/sgp2
GeneWise: http://www.ebi.ac.uk/~birney/wise2
Genomix: http://www.sanger.ac.uk/Software/analysis/genomix
GENSCAN: http://genes.mit.edu/GENSCAN.html
GENSCAN parameter files: http://genes.mit.edu/license.html
GLEAN: http://sourceforge.net/projects/glean-gene
GlimmerM: http://www.tigr.org/tdb/glimmerm/glmr_form.html
GMAP: http://www.gene.com/share/gmap
HAVANA: http://www.sanger.ac.uk/HGP/havana
Jigsaw: http://cbcb.umd.edu/software/jigsaw
N-SCAN: http://mblab.wustl.edu/nscan
PPFINDER: http://mblab.wustl.edu/software/ppfinder
SGP2: http://www1.imim.es/software/sgp2
TWINSCAN: http://mblab.wustl.edu/nscan
UNIPROT: http://www.ebi.uniprot.org/index.shtml

ALL LINKS ARE ACTIVE IN THE ONLINE PDF