Chapter 1

Methods for the computational identification of gene structures

This short review outlines some of the current approaches to the computational identification of gene structures in genomic DNA. It focuses on the very specific problem of the prediction of the complete structures of protein-coding genes in the sequences of eukaryotic organisms. Gene prediction in prokaryotes is traditionally viewed as less challenging due to high gene density and lack of introns (although genes with overlapping coding regions are far more common in prokaryotes than in plants and animals). The prediction of non-protein-coding genes is on the other hand considered by most to be a harder problem and is currently attracting considerable research interest in its own right (see section 1.5).

I aim here to summarise the main issues rather than attempting to be comprehensive; there are many existing reviews on this subject, notably by Fickett [39], Claverie [26], Guigo [51], Burge and Karlin [22], Stormo [108], and recently by Zhang [122] and Mathe *et.al.* [77]. Many of the issues summarised here are discussed in further detail in later chapters of this thesis.

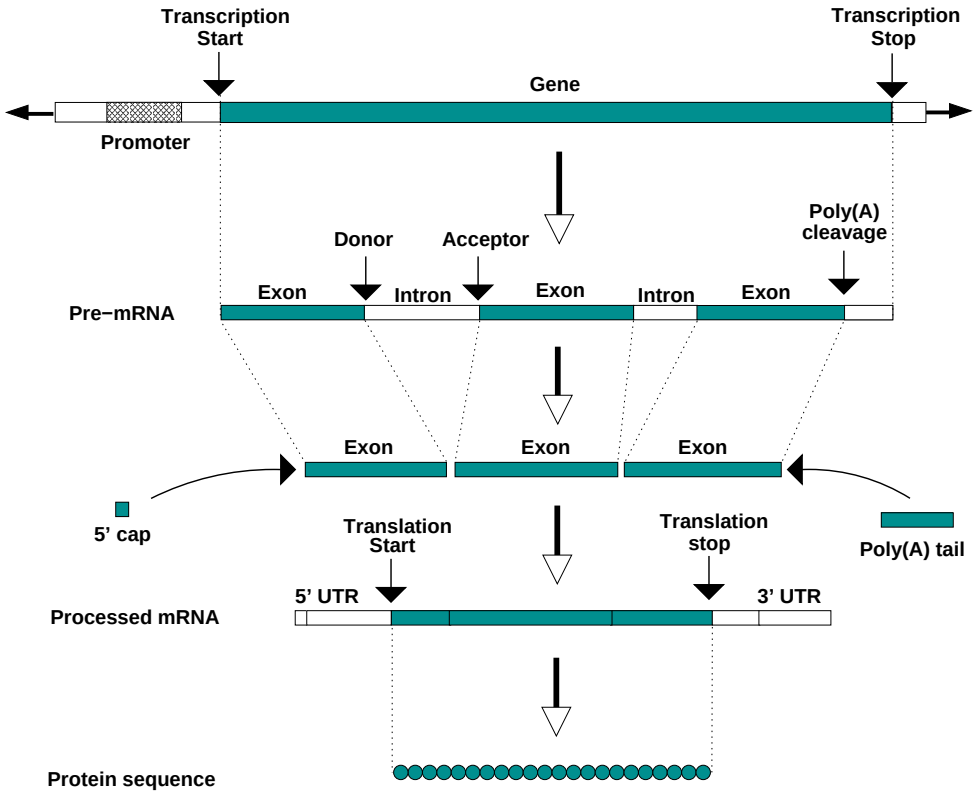

Figure 1.1: The primary components of a typical eukaryotic protein-coding gene, showing its transcription, the processing of the transcribed RNA, and the translation of the processed RNA.

1.1 Identifying the elements of gene structure

Figure 1.1 depicts the classic view of how a eukaryotic protein-coding gene specifies its protein sequence [115] [1]. With this model, the cellular processes of transcription, RNA processing and translation can be viewed abstractly as a series of segmentations: firstly of the genomic DNA sequence into gene and non-gene regions, each gene giving rise to one or more primary transcripts; secondly of the transcript sequence into exonic and intronic regions, the exons forming a processed transcript; and thirdly of the processed transcript into translated and untranslated regions, the translated region giving rise to a protein. The problem of gene prediction in this context can therefore be described as identifying the precise regions of the genomic

DNA that correspond with translated regions in processed RNAs, allowing at least in theory the inference of the proteome from the genome.

At present, our understanding of the mechanisms of transcription, RNA processing and translation is insufficient to be able to model them precisely *in silico*. Until our knowledge reaches a level where a deterministic computational model is realistic, we must supplement what we know with a variety of statistical methods.

1.1.1 The recognition of gene structural elements

The localised sequence signals that form the basis for the cellular recognition of the boundaries between functional regions of gene structure are most intuitively expressed as consensus sequences. The fact that these signals are neither unique to genes nor completely conserved across all genes makes their computational detection non-trivial.

Of the three processes of transcription, RNA processing and translation, it is the sequence signals of the first that are the most poorly understood. In attempting to define the site of transcription initiation, the AT-rich TATA-box signal located around 30 base-pairs upstream forms the basis for the many of detection techniques [40], but only around 70% of human promoters (for example) have this characteristic [19]. The association of promoters with CpG islands (short regions where suppression of methylation leads to a higher-than-average proportion of the dinucleotide CG [8]) provides another useful signal, but one that can give only low resolution mapping of the transcription start site. The site of transcription termination is even less well characterised, and most methods instead focus on trying to identify the upstream site of the transcript cleavage that occurs prior to polyadenylation, the poly(A) site. In many organisms, this is characterised by the presence of the hexanucleotide AATAAA around 30 nucleotides upstream, but in humans the signal is absent from around half of all genes [26], and in *C. elegans* there are as many as 23 one or two base variants documented [14].

Although the sites of *translation* initiation and termination are difficult to iden-

tify in isolation, they do have elements that are completely conserved: the translated regions of all eukaryotic genes begin in the processed mRNA with the triplet corresponding to the start codon, almost always ATG (with a slightly less well conserved context upstream of the ATG known as the Kozak sequence [65]), and end with the first in-frame occurrence of a triplet corresponding to one of the three possible stop codons, TAA, TGA, and TAG. This information however is insufficient to be able to to identify these sites in the genome with any accuracy at all. The degeneracy of the translation initiation site can be explained in part by the fact that the ribosome need "search" only a relatively small stretch of processed mRNA for a place to bind rather than a far longer stretch of genomic DNA. We still do not understand enough about translation however to identify the initiation site even in a processed mRNA with certainty.

As with translation, searching the genomic DNA for the signals important in RNA processing provides a more difficult problem than that encountered by the spliceosome, which is faced only with an orders-of-magnitude smaller RNA molecule. Even so, our understanding of the signals involved in the splicing process, although far from comprehensive, provides the most useful information for the identification of gene structures directly in the genomic DNA. The Donor and Acceptor splice sites have the consensuses AG—**GT**RAGT and YYYYYYYYYYNC**AG**—G respectively [122], with the GT and AG defining the beginning and end of virtually all introns. In vertebrates, the acceptor signal is often extended upstream into the intron to capture the completely conserved A residue that occurs at the branch-point [20].

Although these signals of transcription, translation and RNA-splicing are quite different, the same kinds of techniques are used to identify features of all of them. Most correspond to variations on the idea of a generalised consensus sequence for the signal, using a probability distribution over the possible residues occurring at each position. The widely used Weight Matrix method (WMM) [105] treats each position in the consensus as independent. Where it is believed that there are linear dependencies between the positions, the Weight Array method (WAM) is often used,

where the probability of residue s appearing in position i depends not only on the position (as with the WMM), but also on the residues appearing at positions $i-1, i-2, \ldots, i-k$ (where k is chosen according to the amount of training data available). Where the dependencies between columns are known but non-adjacent, Maximal Dependence Decomposition (MDD) [21] can be used, and multi-layered artificial networks can model complex dependencies between positions even when the dependencies themselves are poorly understood [17]. Finally, profile Hidden Markov Models are able to model insertions and deletions with respect to the generalised consensus [36].

In practice, the degeneracy of the short sequence-motifs characteristic of these gene features makes these so called *signal sensor* methods most useful when used in combination with *content sensor* methods for assessing the likelihood of the longer more extensive regions between the features, such as intron and exons (see below). However, there are many examples of where these methods are employed directly, for example, in the detection of transcription start sites [34], polyadenylation sites [111], translation start sites [55] and splice sites [86].

1.1.2 The recognition of gene regions

A protein coding region in the genomic DNA consists of a string of triplets, each of which corresponds to a codon that will be translated into an amino acid. Biases in the usage of amino acids in proteins, in the usage of codons for a single amino acid, and the absence of stop codons provide the basis for the majority of methods for differentiating coding regions from non-coding regions.

Inspired by the survey of coding statistics conducted by Fickett and Tung [41], most techniques for assessing the *coding potential* of a sequence are based on the relative frequency of occurrence of frame-specific hexamers (runs of 6 nucleotides) in coding regions compared with non-coding regions (usually introns). These differences can be expressed probabilistically as a kth order *Markov* model [36] which is a stochastic model that assumes that the probability of a base at a given position

depends only upon the previous k bases. The simple weight matrices above are therefore related to 0th-order Markov models, with a different probability distribution for each column, whereas the weight array matrices are related to kth order Markov models. Using hexamer frequencies therefore corresponds to a 5th order Markov model. In practice, protein-coding regions are modelled by using a separate 5th order Markov model for each of the three positions in a codon, as in the GENMARK program [15]. More recently, Interpolated Markov models (IMMs) have been utilised for eukaryotic gene finding [97], in which the probability of a base at a given position is a weighted average of the probabilities according to several Markov models of different orders (e.g. between 0 and 8).

Because these content sensing methods are poor at identifying the precise boundaries of the coding regions, they are most often used in combination with the signal sensing methods described earlier. A pair of detected signals of appropriate type defines a candidate gene region; for example, a translation start site followed by a donor splice site defines a candidate initial protein coding region of an exon, the likelihood of which can be judged by using a content sensor such as the Markov models outlined above. This approach is appropriate for detecting all types of gene region. For example, a candidate intron can be defined by upstream donor and downstream acceptor splice sites, and an exon by a transcription start site / acceptor splice site upstream followed by a polyadenylation site / donor splice site downstream.

In order to be able to discriminate between true and false candidate regions defined in this way, a *score* for the region is commonly computed by combining the information from many sources, typically (i) the signal scores of the defining boundary elements; (ii) content scores for the intervening sequence; (iii) a score obtained from a probability distribution over the possible lengths of the region. If the scores correspond to (log) probabilities and we assume independence, they can be combined most naturally by (addition) multiplication. More generally, it is necessary for the score components to be weighted and combined appropriately [109]. In the pioneering GRAIL system for the detection of protein-coding exons, the scores

reported by several sensors were combined with an artificial neural network [113]. An alternative approach results from viewing each of the n region discriminators as the axis of a high dimensional space, in which case candidate regions can be described by a point in this space. Techniques such as linear and quadratic discriminant analysis can be used to find the surface in this space that optimally separates true and false examples. Of the signals discussed earlier, it is the donor and acceptor splice sites that are perhaps the best characterised, so these techniques have both been employed most successfully for the identification of protein-coding *internal* exons [103] [120]. More recently quadratic discriminant analysis has been used for the recognition of complete initial [29] and terminal [110] exons.

1.2 Identifying complete gene structures

The prediction of complete gene structures is seemingly a more difficult problem than that of the prediction of localised gene features and regions. However, the accuracy of identification of such features and regions can be improved by considering not only their local properties, but also their relationship to other more distant gene features. At the very simplest level, the protein-coding regions of exons almost never overlap in eukaryotes, successive protein-coding regions on the same strand in the same gene must be frame compatible, and the length of the coding regions of the complete gene must be a multiple of three. Such constraints are the basis for most methods for the identification of complete gene structures. The widely used techniques can be broadly divided into two categories: gene fragment assembly, and Hidden Markov Models.

1.2.1 Gene fragment assembly methods

The majority of early programs for the prediction of complete gene structures conceptually broke the problem into two distinct sub-problems: (a) identify a set of candidate gene fragments (e.g. coding exons) using the methods such as those de-

scribed in the previous section; (b) of the many possible complete gene structures assemblies implied by these candidate gene fragments, eliminate *illegal* assemblies, i.e. those that do not adhere to the constraints of gene structure. Then, of those remaining, identify the assembly that is optimal according to some scoring function.

The constraints that determine the legality of candidate assemblies are conveniently expressed as *rules* that dictate which pairs of gene features or regions are allowed to appear next to each other in a legal gene structure. More formally, the rules can be expressed in the form of a grammar [33]. Early programs assumed that the input sequence contained exactly one complete gene on the forward strand of the sequence. Generalisation of the assumed gene structure constraints has given the majority of the more recent programs the ability to predict multiple genes, genes on both strands, and partial genes at the ends of the sequence. It is still the case however that most of these assumed gene models make no allowances for specific peculiarities of certain gene structures, such as the presence of introns that occur completely within the non-coding region of an mRNA (so-called UTR introns), and the presence of genes within an intron of another gene [57].

Disregarding assemblies that do not adhere to the constraints of an assumed gene structure model reduces the number of possible assemblies, but typically many will remain. The first programs worked by iterating though all legal gene structures explicitly [42] [47], but as the number of possible legal assemblies typically grows exponentially with the number of gene fragments [49] [116], this approach was only practicable for small sequences. Subsequent research addressed this problem by using heuristics to reduce the search-space, an example being the original version of the GENEID program which used series of hierarchical rules to filter out unlikely exons before the assembly stage [89]. However, such methods are not guaranteed to identify the optimal assembly. The application of *Dynamic Programming* [6] for the efficient identification of the optimal assembly therefore represented a significant advance. Dynamic Programming is a generic name for a family of recursive optimisation techniques that work by the progressive construction of a solution from

simpler sub-solutions. The technique allows the exploration of a search space that grows exponentially with the number of gene fragments in time that grows only polynomially ($O(n^2)$ or even $O(n)$). For this reason, dynamic programming algorithms are employed in most of the popular gene fragment-assembly based programs [101] [118] [84].

1.2.2 Hidden Markov models

Hidden Markov models (HMMs) provide a convenient framework for the representation of the signal and content displayed by gene features and regions and the constraints of gene structure, in one unified probabilistic model. An HMM can be thought of a probabilistic finite state automaton for generating a sequence from left to right according to an underlying *hidden* state path. At each stage the model (i) emits a single residue according to an *emission* probability distribution over residues that is dependent upon the current state, and (ii) moves to a new state (possibly the same state) according to a *transition* probability distribution over states that is dependent upon the current state.

HMMs can naturally model sequences that are partitioned into regions of different types, for example genes. At the very simplest level, we might have a state for 'exon' with two transitions, one into an 'intron' state, and another looping back into the 'exon' state. These two states can have different emission distributions, and can thus represent inherent compositional differences in introns and exons. Each path through the HMM therefore corresponds to a gene structure, and inference of the underlying state path (given the sequence) corresponds to a prediction of gene structure.

The transition and emission probabilities of an HMM amount to a joint probability distribution over pairs of state-paths and sequences. Given a query sequence, established, efficient (linear-time) techniques can then be used to obtain (i) the most likely state-path given the sequence (using the *Viterbi* dynamic programming algorithm [114]); (ii) the full probability of the sequence considering all state paths

(using the forward or backward algorithm [90]); and (iii) the posterior probability that residue i was generated by state k (using the forward-backward algorithm [90]). One of the key advantages of HMMs is that given a *training* set of sequences with known gene structure, maximum likelihood parameters for the model (i.e. emission and transition probabilities) can be obtained by counting.

HMMs as described generate a region of a particular function class by successively looping to the same state, jumping to a different state at the end of the region. This imposes a geometric distribution on the duration of the states, which is an unrealistic model for most of the functional regions in genes, particularly protein-coding exons [20] [112]. One of two enhancements to the standard HMM architecture is therefore usually employed to represent the non-geometric nature of the length of protein-coding exons.

The first is to model the sequence as a list of classes, with each residue being *labelled* as belonging to a particular class, and being generated by one of a *group* of states labelled as belonging to the same class. Such a formalism is known as a *Class*-HMM [67]. By chaining more than one state together with the label "exon" and then obtaining the most likely *labelling* by summing over all paths with the same labelling, a length distribution resembling that displayed by protein-coding exons is observed. This is the approach taken by HMMGENE [68], which uses a novel hybrid of the Viterbi and forward algorithms to identify the most likely labelling.

The second enhancement often employed is to allow the emission of an entire sequence region from each state, the length of which is chosen according to an explicit length probability distribution for the state. Such models are known as *semi*-Hidden Markov models [21] or *Generalised* HMMs (GHMMs) [72]. In theory, the worst-case run-time of the Viterbi, forward and backward algorithms when applied to GHMMs grows quadratically in the length of the sequence, making their computation prohibitively expensive for true genomic sequence fragments which may be hundreds of kilobases long. In practice, this problem is often addressed by pre-processing the sequence for a list of candidate gene features, alleviating the need to consider every

base of the sequence during the computationally intensive dynamic programming algorithms [72]. This does not address the problem of the quadratic growth of the run-time however. In practice, the run-time does not grow truly quadratically, because the length of coding regions is limited naturally by the fact that they cannot extend past an in-frame stop codon. The GENSCAN program [21] restricts the semi-Markov property to these protein-coding regions only, with non-coding regions being generated by standard self-looping transitions. Although imposing a geometric distribution on the length of such regions, the assumption allows the run-time of the program to grow effectively linearly with sequence length.

1.3 Using similarity to other sequences

1.3.1 Expressed sequences

Despite progressive advances in the *ab initio* gene prediction methods discussed above, it is still the case that the most reliable way to identify the gene structure of a piece of genomic DNA is by aligning it to a corresponding expressed sequence, either cDNA (complementary DNA, a DNA copy of an expressed mRNA) or homologous protein.

A cDNA can be aligned to genomic sequence with a standard pairwise comparison tool like BLASTN [2]. However, in most cases this will not identify the intron-exon boundaries precisely, since similarities often by chance extend partially into the introns flanking an exon. A variety of "spliced" alignment programs address this problem by incorporating knowledge of the donor and acceptor splice site consensuses, scoring long gaps that start with a GT and end with an AG in the genome more favourably [79] [44] [61]. These methods can be applied equally well to full-length cDNA and the more abundant partial cDNAs such as Expressed Sequence Tags (ESTs), although the use of the latter will typically reveal only a portion of the gene structure.

Correspondingly with cDNA methods, the alignment of a homologous protein

to the genomic sequence using a tool such as BLASTX [2] is useful for discovering regions that are likely to be protein-coding, but will usually not reveal the intron-exon boundaries precisely. Spliced alignment programs also exist for the alignment of a protein to genomic DNA. PROCRUSTES [48] for example first obtains a list of candidate internal exons that are delimited by the AG and GT acceptor and donor splice consensuses, and then obtains the assembly of the these exons for which the implied translation is most consistent with the given protein. GENEWISE [10] on the other hand aligns a profile Hidden Markov model constructed from the given protein sequence (or multiple sequence alignment) directly to the genomic DNA, implicitly considering all possible gene predictions. Both of these methods are extremely accurate when the protein is a close homolog to that encoded in the genomic sequence [53].

Techniques that integrate methods of *ab initio* and similarity based gene prediction approaches have the theoretical advantage of maximising accuracy where a similar sequence exists, without compromising the ability to predict novel gene structures where one does not. One common way to do this is to use the alignments resulting from a database search to modify the scores of the appropriate gene structural elements used in an *ab initio* program, e.g. using BLASTX matches to boost the scores of protein-coding exons [71] [69] [119]. As one would hope, these approaches can outperform traditional *ab initio* methods, although the margin of this improvement is limited by the quality of the sequences in the public databases used for prediction. The high-throughput, error-prone nature of ESTs in particular makes their alignment to the genome and subsequent use by an integrated gene prediction method non-trivial [91].

1.3.2 The sequences of other genomes

There have recently been published a number of gene prediction techniques that make use of the genome sequences of more than one organism. These methods are based on the hypothesis that the coding regions of the genomes of a group of

organisms that share a common ancestor should be under greater selection, and therefore be more conserved, than the non-coding regions. Such a comparative approach to gene prediction is appealing for many reasons, not least of which is the ability to discover truly novel genes that share neither the sequence feature characteristics of known genes, nor any observable similarity to expressed sequences in the public databases. In addition, it provides the possibility of identifying conserved non-coding regions that are also under selective pressure, for example those involved in gene regulation.

The most obvious way to look for conserved coding regions between genomes is to use a similarity search tool such as TBLASTX [2] which performs a heuristic-based pair-wise alignment of each possible translation of one sequence to each possible translation of the other. This will typically return a list of small, localised candidate exon-pairs. Programs such as WABA [62] and GLASS [4] on the other hand perform the global alignment of longer syntenic genomic regions, classifying parts of the alignment as likely coding or non-coding.

Although neither these local nor global alignment methods provide direct predictions of gene structures, they form the first stage of many programs that do. CEM [3] for example identifies the optimal assembly of candidate exon pairs reported by TBLASTX into complete pairs of gene structures. ROSETTA [4] on the other hand uses the longer alignments of GLASS and identifies the most likely *parse* of the alignment into intronic, exonic and intergenic regions. Both of these programs simultaneously predict gene structures in both sequences. TWINSCAN [64] is a generalisation of GENSCAN that uses BLASTN alignments of a query sequence to an "Informant" genome to predict gene structures in the query sequence only. The BLASTN alignments are used to construct a "conservation" sequence mirroring the query that classifies each position as "match", "mismatch" or "unaligned", and the predicted gene structure is that which maximises the joint probability of the query sequence and the conservation sequence.

Finally, it is worth mentioning two recent methods based on *pair* Hidden Markov

models [36]. DOUBLESCAN [78] models exon fusion, splitting, insertion and deletion in one sequence with respect to the other. SLAM [82] on the other hand assumes that the number of exons in the homologous gene structures is the same, but also models conserved non-coding regions explicitly, giving it the potential to identity homologous regulatory regions. Both simultaneously predict gene structures in both sequences. The advantage of this approach is that it is unnecessary for a genomic alignment (or set of alignments) to be produced in advance; gene prediction and alignment are performed simultaneously.

The accuracy of comparative gene prediction methods, although representing an advance over traditional single-sequence *ab initio* methods, has so far fallen short of what one might expect to be achievable by such an approach. One reason for this is that when considering only a pair of genomes (as current methods do), conservation will occur in many non-coding regions (about 50% of the conserved regions between human and mouse are non-coding [122]). The use of more than two genomes would be expected to improve performance, and many of the methods discussed above generalise naturally to multiple genomes. However, the increase in computational complexity that results from adding more genomes unfortunately makes this approach impracticle at present. Future research in this area might therefore be towards the use of heuristics to reduce the explosion of the search-space that results from adding more genomes, perhaps using ideas from classic multiple sequence alignment.

1.4 Assessing gene prediction accuracy

In the field of gene prediction, a new method is generally not considered an advancement unless it is shown to be as least as accurate as existing methods. In this section, I briefly discuss some of the issues involved in assessing the "accuracy" of a gene prediction technique.

Surveys of available techniques and their accuracies appear periodically (see for example refs. [23], [85], [94], and http://predict.sanger.ac.uk/th/brca2) although of course progressive improvement in the techniques can mean the specific figures

become dated. The survey of Burset and Guigo [23] however remains one of the most cited articles in the field because it laid down a standard for the way in which gene prediction accuracy should be assessed.

The biggest problem in providing a fair comparison between the accuracy of different gene prediction techniques is the choice of a test-set. The set of 570 vertebrate sequences constructed by Burset and Guigo has gained widespread use as a benchmark for the assessment of vertebrate gene finders. Its usefulness is limited in two ways however. Firstly, since each sequence contains exactly one complete gene on the forward strand, it provides no way to judge the accuracy of programs that can predict multiple gene on both strands of a sequence. Secondly, the utility of the set diminishes with time, since many researchers will use some (or all) of these sequences in the development and parameterisation of their methods.

For comparisons between gene prediction techniques to be fair, researchers must have a clear and unified idea of how accuracy is to be measured. To this end, Burset and Guigo proposed a series of metrics, each summarising a separate aspect of accuracy as a single floating point number. A number of these metrics are made use of in the remainder of this dissertation, and are outlined next.

1.4.1 Gene prediction accuracy metrics

Accuracy is classically presented at three levels: at the level of individual nucleotides or base-pairs; at the level of whole exons; and at the level of complete gene structures.

Base-level accuracy

Accuracy at the base-pair level is described in terms of sensitivity (Sn), which is the proportion of nucleotides annotated as coding that are predicted as coding; and specificity (Sp), the proportion of nucleotides predicted as coding that are annotated as coding. More formally, we can describe these quantities in terms of (i) the number of nucleotides predicted as coding that are actually coding (true positives, TP); (ii) the number of nucleotides predicted as coding that are actually non-coding

(false positives, FP); (iii) the number of nucleotides predicted as non-coding that are actually non-coding (true negatives, TN); and (iv) the number of nucleotides predicted as non-coding that are actually coding (false negatives, FN). Sensitivity and specificity are then calculated as:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TP}{TP + FP}$$

Each of these measures in isolation gives a poor indication of global accuracy, since a method can be highly sensitive by predicting every base as coding, or highly specific by predicting every base as non-coding. Burset and Guigo therefore defined the Correlation Coefficient (CC) that represents aspects of both sensitivity and specificity and acts as a measure of global accuracy at the nucleotide level:

$$CC = \frac{(TP)(TN) - (FN)(FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Exon and gene-level accuracy

At the exon level, sensitivity and specificity are slightly less obviously defined because we must be clear what it means for an annotated exon to be predicted correctly (and for a predicted exon to be correct). The standard laid down by Burset and Guigo states that a predicted exon is correct if it matches precisely the boundaries of a single, annotated exon. The same criterion is used to judge whether an annotated exon has been predicted correctly. Sensitivity is therefore defined as the proportion of annotated exons that have been predicted correctly, and specificity the proportion of predicted exons that are correct. The Correlation Coefficient is less useful as a summary of these two quantities in this case, so the average of sensitivity and specificity is often quoted as a global measure of accuracy at the exon level.

Requiring that exon boundaries match precisely does not allow for the fact that many predicted and annotated exons will overlap inexactly. The proportion of annotated exons that are completely missed by the prediction (missing exons, ME) and

the proportion of predicted exons that share no overlap with any annotated exon (wrong exons, WE) are therefore also often reported.

Although Burset and Guigo did not define any measures for assessing the accuracy of prediction of whole genes in their article, the exon-level measures are naturally applied in the same way, defining a predicted gene to be correct if it precisely matches the intron-exon structure of an annotated gene, and an annotated gene to be correctly predicted if its intron-exon structure is precisely matched by that of a predicted gene. Missing genes (MG) and Wrong genes (WG) are defined in a similar way to exons.

Reese and colleagues defined two additional gene-level measures to express the accuracy with which a method delineates a sequence containing several genes into its constituent genic regions [91]. In their scheme, an annotated gene is considered *Split* if it is overlapped by more than one predicted gene. Likewise a predicted gene is considered *Joined* if it overlaps more than one annotated gene. Split genes (SG) is defined as the number of predicted genes having some overlap to an annotated gene (i.e. total predicted genes minus number of wrong genes), divided by the number of annotated genes having some overlap to at least one predicted gene (i.e. total annotated genes minus number of missed genes). Joined genes (JG) is the number of annotated genes having some overlap to a predicted gene (i.e. total annotated genes minus number of missed genes) divided by the number of predicted genes having some overlap with an annotated gene (i.e. total predicted genes minus number of wrong genes). These measures are naturally only useful for assessing the accuracy of programs that are capable of predicting several genes in a sequence.

1.5 Other issues

One issue that complicates gene prediction somewhat is that the structural and compositional features of genes can vary according to the particular species. A brief comparison of the genes of a simple animal (the nematode worm, *C. elegans*) and a warm-blooded mammal (human) highlights some of these differences. The propor-

tion of the genome that is protein coding is far higher in the worm than in human: around 25% in the worm compared with under 5% in the human genome. The gene density is also uniformly higher, with the worm having only half as many genes as human, but compressed into a thirtieth of the genomic space. Worm genes also characteristically have longer exons and shorter introns than human genes [112]. These factors give a higher signal-to-noise ratio in *C.elegans* sequences, facilitating gene prediction. On other hand, worm introns lack a detectable branch-site consensus [13], reducing the sequence signal available for the detection of acceptor splice sites. Also many *C.elegans* genes have atypical structural organisation, being transcribed as part of a multi-gene transcript called an operon [12], or *trans*-spliced [66] (see chapter 3 for more details).

These organism-specific properties are usually addressed by determining separate parameter sets for each organism (although the degree to which this approach can reflect structural as well as compositional differences is limited). The properties of genes can also differ markedly within a single species as well as between species. The human genome, and the genomes of other warm-blooded vertebrates tend to exhibit long-range variations in certain base compositional properties (e.g. C+G content) which has been explained in terms of "isochores" [7]. It has been shown that certain structural properties of human genes, for example intron-length and intergenic distance, vary according to C+G content of the background genomic DNA [37] [20]. This makes it difficult to arrive at a set of parameters that work well uniformly across the whole genome. Many programs therefore have distinct sets of parameters for different C+G% strata, and this has been shown to improve performance [102] [21].

One of the peculiarities of RNA processing not addressed by the majority of current techniques is that of alternative splicing, where two (or more) transcripts from the same gene are spliced in different ways, often giving rise to distinct proteins. It has been estimated that over half of human genes give rise to products that are alternatively spliced [112]. The most reliable way to identify such events is by

manual inspection of a collection of cDNA or EST alignments and making gene structures consistent with this set by hand. There have been recent attempts at the automation of this process, including a promising method used in the ENSEMBL automatic annotation system [60], which identifies the minimal set of gene structures that "explains" a collection of EST alignments to a region of the genome [E. Eyras, pers. comm.]. Our understanding of the signals involved in alternative splicing is still insufficient for their *ab initio* computational prediction directly [73], but it has been shown that *sub-optimal* exons (i.e. those with significant probabilities that are not part of the single most-likely gene structure) are sometimes involved in alternative splice forms of the gene transcript [20].

To end, it is worth mentioning that many transcribed RNAs are not translated into proteins but assume some other role in the cell. These include for example ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and small nuclear RNAs (snRNAs). There is considerable interest in computational methods for the identification of genomic regions that give rise to these non-coding RNAs, or ncRNAs [38]. In the absence of the normal signals present in protein-coding genes, the majority of methods work on the assumption of a consensus secondary structure for ncRNAs of a particular type. Mathematical representations of the consensus called *Stochastic Context Free Grammars* (SCFGs) [36] provide a basis for the modelling of base-paired stems that occur in RNA secondary structure, and can be aligned to a genomic sequence providing simultaneous prediction of a ncRNA and its secondary structure. This technique has been applied extremely successfully to the prediction of transfer RNAs in the human genome [75]. More recently, *pair*-SCFGs have been used to represent the long range compensating mutations that are observed in the alignment of homologous ncRNAs, allowing them to be discriminated from conserved coding regions [93].