Finding Genes by Computer: Probabilistic and Discriminative Approaches

Victor V. Solovyev

The more comprehensive and accurate initial computational analysis performed for new genomic sequences, the less time-consuming and costly experimental work will have to be done to determine their functions. For this reason, computational gene identification is an issue of obvious importance as a tool of identifying biologically relevant features (protein coding sequences) that often cannot be found by the traditional sequence database searching technique. This chapter describes statistically based methods for the recognition of eukaryotic genes. We review the structure and significant characteristics of gene components, and discuss recent advances and open problems in gene-finding methodology and its application to sequence annotation of long genomic sequences. Finally, we consider the application of gene expression data for large-scale verification of predicted genes.

9.1 General Features of Eukaryotic Genes

Genes carry and express the hereditary information encoded by segments of nucleic sequence involved in producing protein or RNA molecules. Genetic organization and packaging of eukaryotic genes is fundamentally different from those of prokaryotes. The major differences are large proportion of noncoding DNA (regulatory sequences, introns, repeats, pseudogenes) and the existence of interruptions (introns) that separate different parts of protein coding region in DNA. A typical DNA fragment of protein coding gene includes noncoding regulatory sequences, exons, and introns (figure 9.1).

9.1.1 Gene Expression Steps

The gene is expressed by a several stage process comprising transcription and translation (figure 9.1). Transcription (or pre-mRNA synthesis on DNA template) involves initiation, elongation, and termination steps. RNA polymerase catalyzing RNA synthesis binds a special region (promoter) at the start of the gene and moves along the template, synthesizing RNA, until it reaches a terminator sequence. Post-transcriptional processing of messenger RNA precursors includes capping, 3'polyadenylation, and splicing. The processing events of mRNA capping and polyA addition take place before pre-mRNA splicing and result in producing the mature mRNA. The mRNA consist of sequences (called exons) that encode the protein product (according to the rules of the genetic code). The gene sequence often includes noncoding regions, called introns, that are removed from the primary tran-
Figure 9.1
Expression stages and structural organization of typical eukaryotic protein-coding gene including associated regulatory regions.

Transcription, 5'-Capping and 3'-polyadenilation

Pre-mRNA

Splicing (removing of intron sequences)

mRNA

Translation

Protein

script during RNA splicing. Eukaryotic pre-mRNA is processed in the nucleus and then transported to the cytoplasm for translation. The sequence of mRNA contains a series of triplet codons that interact with the anticodons of aminoacyl-tRNAs (carrying the amino acids) so that the corresponding series of amino acids is incorporated into a polypeptide chain. The small subunit of eukaryotic ribosome binds to the 5'-end of mRNA and then migrates to the special sequence on mRNA (preceding to the start codon) called the ribosome binding site, where it is joined by a large ribosome subunit, forming a complete ribosome. The ribosome initiates protein synthesis at the start codon (AUG in eukaryotes) and moves along the mRNA synthesizing polypeptide chain until it reaches a stop codon sequence (TAA, TGA, or TAG), where release of polypeptide and dissociation the ribosome from the mRNA take place. After that, many proteins undergo post-translational processing (i.e., covalent modifications such as proteolytic cleavage, attachment of carbohydrates and phosphates) before they become functional.

9.1.2 Structural Characteristics

Information about gene structure is accumulated in GenBank and EMBL nucleotide sequence databases. These databases contain annotations of contiguous sequences;
Finding Genes by Computer

Figure 9.2
InfoGene Java viewer (Seledzov and Solovyev 1999) presentation of Homo sapiens gene PACE4. This gene has several alternative forms and is described in 17 records of GenBank. Continues sequences regions corresponding different GenBank entries are separated by the vertical bars.

therefore, one gene can be described in dozens of entries with partially sequenced gene regions, alternative splicing forms, or mRNA. The gene-centric database InfoGene (Solovyev and Salamov 1999) contains descriptions of known genes and their basic functional signals extracted from GenBank (Benson et al. 1999). InfoGene also includes all predicted genes for human and Drosophila draft genomes and several chromosomes of the Arabidopsis genome. InfoGene is realized under JAVA interactive environment system (Seledzov and Solovyev 1999) that provides visual analysis of known information about complex gene structure (figure 9.2) and search for different gene component and signals. The database is currently available at http://www.softberry.com/infodb.html. A similar project, ENSEMBL, was started
Table 9.1
Structural characteristics of human genes deposited in GenBank (Release 119)

<table>
<thead>
<tr>
<th>Gene features</th>
<th>Numbers from the Infogen database</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS/partially sequenced CDS</td>
<td>53435/29404</td>
</tr>
<tr>
<td>Exons/partial sequenced exons</td>
<td>83488/21142</td>
</tr>
<tr>
<td>Genes/partially sequenced genes</td>
<td>20791/16141</td>
</tr>
<tr>
<td>Alternative splicing</td>
<td>2167, 10.4%</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>8.5%</td>
</tr>
<tr>
<td>Genes without introns</td>
<td>1552, 7.4%</td>
</tr>
<tr>
<td>Number of exons (maximal, average)</td>
<td>117, 5.7</td>
</tr>
<tr>
<td>Exon length (range, average)</td>
<td>1–1088, 201.6</td>
</tr>
<tr>
<td>Intron length (maximal, average)</td>
<td>259776, 2203.5</td>
</tr>
<tr>
<td>Gene length (maximal, average)</td>
<td>401910, 9033</td>
</tr>
<tr>
<td>Repeats in genome</td>
<td>41% of total DNA</td>
</tr>
<tr>
<td>DNA occupied by coding exons</td>
<td>3%</td>
</tr>
<tr>
<td>Donor sites</td>
<td>58707, 98.0%</td>
</tr>
<tr>
<td>Acceptor sites</td>
<td>58112, 98.33%</td>
</tr>
</tbody>
</table>

Statistics based on InforGene records.

as a collaboration between the Sanger center and European Bioinformatics Institute (http://www.ensembl.org/).

Major organisms are presented in the InfoGene separate divisions. The human division (based on GenBank 119 release) contains about 21,000 genes, 53,000 coding regions, 83,000 exons, and about 58,000 donor and acceptor splice sites. Table 9.1 shows the major structural characteristics of human genes.

About 41 percent of sequenced human DNA consists of different kinds of repeats. Only about 3 percent of the genome sequence contains protein coding exon sequences. Table 9.2 presents the characteristics of genes in major model organisms such as mouse, D. melanogaster, C. elegans, S. cerevisiae, and Arabidopsis.

The gene sizes are often larger in vertebrates, and especially in primates. The average size of an exon is about 190 bp, which is close to the DNA length associated with the nucleosome particle. Human exon sizes are significantly smaller than the gene sizes. There are many exons as short as several bases.

Computational identification of small exons (1–20 bp) cannot be done using the composition based methods that were successful for predicting prokaryote coding regions. Eukaryotic gene prediction approaches should be based on recognition of functional signals encoded in the DNA sequence.

Figure 9.3 illustrates how the same DNA sequences may code several different proteins due to alternative promoters or terminators and alternative splicing. These processes also can significantly complicate computational gene finding.
Table 9.2  
Structural characteristics of genes in eukaryotic model organisms

<table>
<thead>
<tr>
<th></th>
<th><em>Drosophila melanogaster</em></th>
<th><em>C. elegans</em></th>
<th><em>S. cerevisiae</em></th>
<th><em>Arabidopsis thaliana</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS/partial</td>
<td>20314/1510 (20622)</td>
<td>20634/526</td>
<td>12635/1016</td>
<td>1194/1461</td>
</tr>
<tr>
<td>Exons/partial</td>
<td>66990/19343 (79935)</td>
<td>12951/38293</td>
<td>13572/13127</td>
<td>14594/42844</td>
</tr>
<tr>
<td>Genes/partial</td>
<td>17435/1154 (20622)</td>
<td>19655/1263</td>
<td>12513/1096</td>
<td>28346/1023</td>
</tr>
<tr>
<td>Alt. splicing</td>
<td>1785, 10%</td>
<td>1194, 6.1%</td>
<td>598, 4.7%</td>
<td>227, 0.1%</td>
</tr>
<tr>
<td>No introns genes</td>
<td>3583, 20%</td>
<td>669, 3%</td>
<td>11070, 88.5%</td>
<td>5776, 19.7%</td>
</tr>
<tr>
<td>Number of exons</td>
<td>50, 3.88</td>
<td>52, 6.1</td>
<td>3, 1.03</td>
<td>78, 5.1</td>
</tr>
<tr>
<td>Intron length</td>
<td>6-10785, 419.5</td>
<td>1-14975, 221.5</td>
<td>1-7471, 1500.0</td>
<td>3-75916, 192.0</td>
</tr>
<tr>
<td>Gene size</td>
<td>155515, 2854</td>
<td>45315, 2624</td>
<td>14733, 1462</td>
<td>117658, 99.2%</td>
</tr>
<tr>
<td>Donor sites</td>
<td>49892, 98.0%</td>
<td>102872, 99.5%</td>
<td>471, 93.0%</td>
<td>475, 95.6%</td>
</tr>
<tr>
<td>Acceptor sites</td>
<td>49602, 97.9%</td>
<td>102933, 99.7%</td>
<td>475, 95.6%</td>
<td>121917, 96.9%</td>
</tr>
</tbody>
</table>

Description of features is given in table 9.1. For *Drosophila* genes, the numbers in () are taken from *ab initio* computer annotation of *Drosophila* genome by Softberry Inc. (http://www.softberry.com/int/dros_an.html).
9.2 Functional Signals Description and Identification

In this section, we focus on several approaches for gene functional signal recognition and some features of these signals used in gene identification. We describe the application of different weight matrices, which usually contain more information about the structure of functional signal than the corresponding consensus sequences, and later elaborate their implementation gene prediction approaches to score potential functional signals.

9.2.1 Position Specific Discrimination

The consensus sequence consists of the most common base at each position of an alignment of binding sites of a particular type. Often it uses a special letters (IUPAC) to indicate the potential presence of more than one nucleotide at a given position. Position weight matrices usually provide better representation of functional signals including quantitative information (Staden 1984; Zhang and Marr 1993; Burge
1997). We can consider the weight matrix as a simple model based on a set of position-specific probability distributions \( \{ p_i^f \} \), that give the probability of observing a particular type of nucleotide in a particular position of a functional signal (S) sequence. The probability of generating the signal sequence sequence \( X(x_1, \ldots, x_k) \) under this model is

\[
P(X/S) = \sum_{i=1}^{k} p_{i}^f \tag{9.1}
\]

where nucleotides of the signal are generated independently. The corresponding model can be constructed for nonsite (N) sequences, \( \{ \pi_i^f \} \), with the same probability distribution in each position. A discriminative function based on these models is the log likelihood ratio:

\[
\text{LLR}(X) = \log P(X/S)/P(X/N) \tag{9.2}
\]

It can be written in weight matrix notation, \( w(i, s) = \{ \log(p_i^f/\pi_i^f) \} \), and

\[
\text{Score} = \text{LLR}(X) = \frac{1}{k} \sum_{i=1}^{k} w(i, x_i) \tag{9.3}
\]

The other types of weight functions can be used to score the sequence of signal. For example, weights can be generated by some optimization procedures such as perceptron or neural network (Stormo 1982); different position-specific probability distributions \( \{ p_i^f \} \) might also be considered.

More general types of weight matrix uses position-specific probability distributions \( \{ p_i^f \} \) of oligonucleotides (instead of nucleotides). Oligonucleotide frequencies are successfully used in Markov chain models, where the probability to generate a particular nucleotide \( x_i \) of the signal sequence depends on \( k_0 - 1 \) previous bases (i.e., depends on oligonucleotide \( [k_0 - 1 \text{ base long}] \) ended at the position \( i - 1 \)). Then the probability of generating the signal sequence \( X \) is:

\[
P(X/S) = p_0 \prod_{i=k_0}^{k} p_{x_i, i}^{f-1, i} \tag{9.4}
\]

where \( p_{x_i, i}^{f-1, i} \) is the conditional probability of generating nucleotide \( x_i \) in position \( i \) given that oligonucleotide \( s_{i-1} \) ends at position \( i - 1 \); \( p_0 \) is the probability of generating oligonucleotide \( x_1 \ldots x_{k_0-1} \). A simple weight matrix represents an independent mononucleotide model (or 0th Markov chain), where \( k_0 = 1, p_0 = 1 \) and
When we use dinucleotides (first order Markov chain) $k_0 = 2$, $p_0 = p_{\text{di}}$, and $p_{i-1,i}$ is the conditional probability of generating nucleotide $x_i$ in position $i$ given nucleotide $x_{i-1}$ at position $i - 1$. The conditional probability can be estimated from the ratio of observed frequency of oligonucleotide $k_0$ bases long ($k_0 > 1$) ending at position $i(s_{i-1}, x_i)$ to the frequency of the oligonucleotide $k_0 - 1$ bases long ending at position $i - 1(s_{i-1})$ in a set of aligned sequences of some functional signal:

$$p_{i-1,i} = \frac{f(s_{i-1}, x_i)}{f(s_{i-1})}$$

A model for nonsite sequences for computing $P(X/N)$ is usually based on a 0-order Markov chain with genomic base frequencies (or even equal frequencies $[0.25]$).

A log likelihood ratio (9.3) with Markov chains was used in a description of promoter, splice sites, and start and stop of translation signals in gene finding programs such as Genscan (Burge and Karlin 1997), Fgenesh (Salamov and Solovyev 1998, 2000) and GeneFinder (Green and Hillier 1998).

A useful discriminative measure taking into account a priori knowledge can be based on computing Bayesian probabilities as components of position-specific distributions $\{p_i\}$:

$$P(S/o_i) = P(o_i/S)P(S)/(P(o_i/S)P(S) + P(o_i/N)P(N)) \quad (9.5)$$

where $P(o_i/S)$ and $P(o_i/N)$ can be estimated as position specific frequencies of oligonucleotides $o_i$ in the set of aligned sites and nonsites; $P(S)$ and $P(N)$ are the a priori probabilities of site and nonsite sequences, respectively, and $o_i$ is a type of the oligonucleotide starting (or ending) in $i$th position (Solovyev and Lawrence 1993).

The probability of a sequence $X$ to belong to a signal, if one assumes independence of oligonucleotides in different positions, is

$$P(S/X) = \sum_{i=1}^{k} P(S/o_i)$$

Another empirical discriminator, called "Preference," uses average positional probability to belong to a signal:

$$Pr(S/X) = 1/k \sum_{i=1}^{k} P(S/o_i) \quad (9.6)$$

This measure was used in constructing discriminant functions for the Fgenes gene finding program (Solovyev 1998).
9.2.2 Content Specific Discrimination

To take into account general oligonucleotide composition of a functional region (such as GC-rich promoter sequences) we can use probability distributions and their estimations by oligonucleotide frequencies computed on the whole set of functional signal sequences. Then the Markov chain based probability formula (9.4) of generating the signal sequence X is:

\[ P(X/S) = p_0 \sum_{\lambda \in k} p_{N,1,\lambda} \]  

(9.7)

9.2.3 Frame Specific Discrimination

The best discrimination of coding and noncoding sequences in gene prediction approaches was achieved by frame specific recognizers (Claverie and Bougueleret 1986; Claverie et al. 1991; Fickett and Tung 1992). The coding sequence is a sequence of triplets (codons) read continuously from a fixed starting point. Three different reading frames with different codons are possible for any nucleotide sequence (six, if a complementary chain is also considered). It was noted that nucleotides are distributed very unevenly relative to the positions within codons. Therefore, the probability of observing a specific oligonucleotide in coding sequences depends on its position relative to the coding frame (three possible variants), as well as on neighboring nucleotides (Shepherd 1981; Borodovskiy et al. 1986; Borodovskiy and McIninch 1993). Asymmetry in base composition between codon positions arises due to uneven usage of amino acids and synonymous codons, in addition to the particular structure of genetic code (Guigo 1999). In Markov chain approaches, the frame dependent probabilities \( p_{N,1,\lambda} \) (\( f \in \{1, 2, 3\} \)) are used to model coding regions. The probability of generating a protein coding sequence X is:

\[ P(X/C) = p_0 \sum_{\lambda \in k} p_{N,1,\lambda} \]  

(9.8)

where \( f \) is equal 1, 2, or 3 for oligonucleotides ending at codon position 1, 2, or 3, respectively.

9.2.4 Prediction Performance Measures

Sensitivity and specificity measures are widely used to characterize the accuracy of an algorithm or a recognition function (Fickett and Tung 1993; Snyder and Stormo 1993;
1994; Dong and Searls 1994). Let us have $S$ sites (positive examples) and $N$ nonsites (negative examples). By applying the recognition function, we identify correctly $T_p$ sites (true positives) and $T_n$ nonsites (true negatives). At the same time, $F_p$ (false positives) sites are wrongly classified as nonsites and $F_n$ (false negative) nonsites are wrongly classified as sites. Note that $S = T_p + F_n$ and $N = T_n + F_p$. Sensitivity ($S_p$) measures the fraction of the true positive examples that are correctly predicted: $S_p = T_p/(T_p + F_n)$. Specificity ($S_n$) measures the fraction of the predicted examples that are correct: $S_n = T_p/(T_p + F_p)$. When we see only one value of accuracy estimation, it means the average accuracy of sites and nonsites is a true prediction: $AC = 0.5(T_p/S + T_n/N)$. The more general single measure (correlation coefficient) takes into account a possible difference in the sizes of site and nonsite sets (Matthews 1975):

$$CC = (T_p T_n - F_p F_n)/\sqrt{(T_p + F_p)(T_n + F_n)(T_p + F_n)(T_n + F_p)}$$

### 9.2.5 Fisher's Linear Discriminant

The linear discriminant analysis approach provides a method to select a “best” set of an objects features and combine them in a discriminant function that yields an output that is an estimate of the class membership of this object. We assume that each given sequence fragment can be described by a vector $x$ of $p$ characteristics $(x_1, x_2, \ldots, x_p)$, which can be measured. The procedure of linear discriminant analysis is to find a linear combination of the measures (called the linear discriminant function or LDF) that provides maximum discrimination between sites sequences (class 1) and nonsite examples (class 2). The LDF

$$Z = \sum_{i=1}^{p} a_i x_i$$

classifies $(X)$ into class 1 if $Z > c$ and into class 2 if $Z < c$. The vector of coefficients $(a_1, a_2, \ldots, a_p)$ and the threshold constant $c$ are derived from the training set by maximizing the ratio of the between-class variation of $z$ to within-class variation (or minimizing expected probability of misclassification) and are equal to (Duda and Hart 1973; Affifi and Azen 1979)

$$\hat{a} = s^{-1}(\hat{m}_1 - \hat{m}_2)$$

$$c = \hat{a}(\hat{m}_1 + \hat{m}_2)/2 + \ln \frac{p_0^0}{p_1}$$
where \( \bar{m}_i \) are the sample mean vectors of characteristics for class 1 and class 2, respectively; \( s \) is the pooled covariance matrix of characteristics,

\[
s = \frac{1}{n_1 + n_2 - 2}(s_1 + s_2)
\]

\( s_i \) is the covariance matrix, \( p_i^0 \) is the prior probability, and \( n_i \) is the sample size of class \( i \). Based on these equations, we can calculate the coefficients of LDF and the threshold constant \( c \) using the characteristics of site and nonsite sequences from the training sets, and we can then test the accuracy of LDF on the test set data. This classification actually assigns a feature vector \( \hat{x} \) to the category of the nearest mean measure, the squared Mahalanobis distance

\[
D^2 = (\hat{x} - \bar{m}_i)s^{-1}(\hat{x} - \bar{m}_i)
\]

from \( \hat{x} \) to each of the mean vectors \( m_i \). The significance of a given characteristic or a set of characteristics can be estimated by the Mahalonobis distance between two classes:

\[
D^2 = (\bar{m}_1 - \bar{m}_2)s^{-1}(\bar{m}_1 - \bar{m}_2)
\]

which is computed based on values of the characteristics in the training sequences of classes 1 and 2. To find discriminating sequence features, many possible characteristics, such as score of weigh matrices, distances, oligonucleotide preference at different subregions, and so on, are generated. Selection of the subset of significant characteristics \( q \) (among the tested \( p \)) is performed by a step-wise discriminant procedure including only characteristics, which significantly increases the Mahalonobis distance. The procedure to test this significance uses the fact that the quantity:

\[
F = \frac{n_1 + n_2 - p - 1}{p - q} \frac{n_1n_2(D^2_p - D^2_q)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1n_2D^2_q}
\]

has an \( F(\ p - q, n_1 + n_2 - p - 1) \) distribution when testing hypothesis \( H_0: \Delta^2_p = \Delta^2_q \), where \( \Delta^2_q \) is the population Mahalonobis distance based on \( q \) variables (Affi and Aizen 1979).

### 9.2.6 Quadratic Discriminant Analysis

Classical linear discriminant analysis often assumes the probability model in which the observations for classes have different means, but a common covariation matrix. The feature space is partitioned by hyperplane optimally separating observations of different classes. To classify groups having different covariation matrices, one can use
the quadratic discriminant analysis (QDA). Quadratic discriminant analysis provides a curved boundary in multidimensional feature space. Maximum discrimination between the two classes is achieved with the quadratic discriminant function QDF:

$$QDF = \log \frac{p_i}{p_j} - \frac{1}{2} (D_i^2 - D_j^2) - \frac{1}{2} \log \frac{|S_i|}{|S_j|}$$

where $D_i^2$ is Mahalanobis distance from an object to the mean and $S_i$ is the covariance matrix of class $i$ ($i = 1, 2$). Quadratic discriminant function might provide a more effective discrimination, but will require a larger learning set of observations to accurately define its larger set of parameters. Such an approach was used in exon prediction method developed by Zhang (1997) to improve the accuracy of the linear discriminant exon predictor (Solovyev et al. 1994).

### 9.2.7 Splice Sites Conservative Features

The precise removal of introns from mRNA precursors is mainly defined by the highly conserved sequences near the ends of introns (Breathnach and Chambon 1981; Wieringa et al. 1983). The donor (or 5'-splice site) is characterized by a sequence of eight nucleotides AG|GTRAGT. The acceptor (or 3'-splice site) possesses a sequence of four nucleotides preceded by a pyrimidine rich region: YYTTYYYYYNCGAGG (Senapathy et al. 1990). The third less conserved intron sequence (branch site), of about 5–8 nucleotides and containing an adenosine residue, usually lies between 10 and 50 nucleotides upstream of the acceptor splice site.

The vast majority of introns contains invariant GT and AG dinucleotides at their termini excised from pre-mRNA by the spliceosome, including U1, U2, U4/U6, and U5 snRNPs (Breathnach et al. 1978; Breathnach and Chambon 1981; Nilsen 1994). A rare type of splice pair, the AT-AC, has also been discovered. It is processed by related but different splicing machinery (Jackson 1991; Hall and Puddey 1994). For the AT-AC group, different conserved positions have been noticed: [ATATCCTTT for the donor site and YAC] for the acceptor site (Deitrich et al. 1997; Sharp and Burge 1997; Wu and Krainer 1997).

Burset et al. (2000) have recently done a comprehensive investigation of canonical and noncanonical splice sites. They applied ESTs and high-throughput genomic (HTG) sequences to analyze 43,437 pairs of exon-intron boundaries and their sequences from InfoGene (Seledtsov and Solovyev 1999) database, including all annotated genes in mammalian genomic sequences. Of the 43,437 pairs of donor and acceptor splice sites (splice pairs), 1,215 were annotated as nonstandard donor sites (2.80 percent), and 1,027 were annotated as nonstandard acceptor sites (2.36 percent). Forty-one thousand seven hundred and sixty-seven splice pairs (96.18 percent)
Finding Genes by Computer

Table 9.3
Splice sites sequences presented in the SpliceDB (Burset, Seledtsov, and Solovyev 2001)

<table>
<thead>
<tr>
<th>Sequences of splice pairs</th>
<th>Canonical</th>
<th>Noncanonical</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mammals</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original from GenBank</td>
<td>41722 (96.27%)</td>
<td>1615 (3.73%)</td>
</tr>
<tr>
<td>EST supported</td>
<td>22374 (98.07%)</td>
<td>441 (1.93%)</td>
</tr>
<tr>
<td>EST supported + corrected</td>
<td>22199 (98.71%)</td>
<td>290 (1.29%)</td>
</tr>
<tr>
<td><strong>Human</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original from GenBank</td>
<td>27486 (96.55%)</td>
<td>982 (3.45%)</td>
</tr>
<tr>
<td>EST supported</td>
<td>15384 (98.33%)</td>
<td>261 (1.67%)</td>
</tr>
<tr>
<td>EST supported + corrected</td>
<td>15263 (98.89%)</td>
<td>171 (1.11%)</td>
</tr>
<tr>
<td>HTG supported</td>
<td></td>
<td>156</td>
</tr>
</tbody>
</table>

contained the standard splice site pair GT-AG. Analysis showed that of 1,615 noncanonical pairs, 441 were supported by EST (27.3 percent) and just 290 (18 percent) were supported by EST after removing potential annotation errors and examples with ambiguity in the position of the splice junction (table 9.3).

Analysis of human noncanonical splice pairs that have corresponding EST and HTG sequences shows that all human EST-supported GC-AG cases having HTS matches were supported (39 cases). Thirty-one errors were found damaging the standard splice pairs (seven cases had one or both intronic GenBank sequences completely unsupported by HTS, whereas eight cases had intronic GenBank sequences supported; there was a gap between exonic and intronic parts, and in the end, 16 cases had small errors, such as insertions, deletions, or substitutions). Three of five observed AT-AC pairs were correctly annotated in the original noncanonical set; two were recovered from errors. Two more cases were annotated as introns, but in HTS, the exonic parts were continuous (accession numbers: U70997 and M13300).

This analysis shows that the overwhelming majority of splice sites contain conserved dinucleotides GT-AG (99.2 percent). The other major group includes GC-AG pairs (0.62 percent), the alternative splicing machine group AC-AT (about 0.08 percent), and a very small number of other noncanonical splice sites (about 0.03 percent). Therefore, gene-finding approaches using just standard GT-AG splice sites can potentially correctly predict 97 percent genes (if we assume four exons per gene, on average). Including the GC-AG splice pair will increase this level to 99 percent. Twenty-two thousand two hundred and fifty-three verified examples of canonical splice pairs were presented in a SpliceDB database, which is available at http://genomic.sanger.ac.uk (Burset et al. 2000). It also includes 1,615 annotated and 292 EST-supported and shift-verified noncanonical pairs. The weight matrices and consensus sequences for the major group of splice sites are presented in figure 9.4.
a) GT-AG group (canonical splice sites): 22199 examples

\[ M_{T70A_{50}G_{80}|GTR_{95}A_{71}G_{81}|T_{46}} \]

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>34.0</td>
<td>66.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>26.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>C</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.8</td>
</tr>
<tr>
<td>G</td>
<td>18.3</td>
<td>32.5</td>
<td>86.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>U</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.2</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

\[ Y_{73}Y_{72}Y_{71}Y_{70}Y_{69}Y_{68}Y_{67}Y_{66}Y_{65}|NC_{71}AG|G_{52} \]

|   | -14 | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 1 |
|---|-----|-----|-----|-----|-----|----|----|----|----|----|----|----|----|----|----|----|
| A | 9.0 | 8.4 | 7.6 | 6.8 | 8.0 | 9.7 | 9.2 | 7.6 | 7.8 | 23.7 | 4.2 | 100 | 0.0 | 23.3 |
| C | 31.0 | 31.6 | 30.7 | 29.3 | 32.6 | 33.0 | 37.3 | 38.5 | 41.0 | 35.2 | 30.9 | 70.8 | 0.0 | 0.0 | 13.8 |
| G | 12.5 | 11.5 | 10.6 | 10.4 | 11.9 | 11.3 | 11.3 | 8.5 | 6.6 | 6.4 | 21.2 | 0.3 | 0.0 | 100 | 52.6 |
| U | 42.3 | 44.6 | 47.0 | 49.4 | 47.1 | 46.3 | 40.8 | 42.9 | 44.5 | 50.4 | 24.0 | 24.6 | 0.0 | 0.0 | 10.4 |

b) GC-AG group: 126 examples

\[ M_{A_{83}A_{88}G_{98}|GCA_{87}A_{84}G_{97}|T_{71}} \]

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>40.5</td>
<td>88.9</td>
<td>1.6</td>
<td>0.0</td>
<td>0.0</td>
<td>87.3</td>
<td>84.1</td>
<td>1.6</td>
<td>7.9</td>
</tr>
<tr>
<td>C</td>
<td>42.1</td>
<td>0.8</td>
<td>6.8</td>
<td>0.0</td>
<td>100</td>
<td>0.0</td>
<td>3.2</td>
<td>0.8</td>
<td>11.9</td>
</tr>
<tr>
<td>G</td>
<td>15.9</td>
<td>1.6</td>
<td>97.6</td>
<td>100</td>
<td>0.0</td>
<td>12.7</td>
<td>6.3</td>
<td>96.8</td>
<td>9.5</td>
</tr>
<tr>
<td>U</td>
<td>1.6</td>
<td>8.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>6.3</td>
<td>0.8</td>
<td>70.6</td>
</tr>
</tbody>
</table>

\[ Y_{75}Y_{74}Y_{73}Y_{72}Y_{71}Y_{70}Y_{69}Y_{68}Y_{67}Y_{66}Y_{65}|NC_{65}AG|G_{63} \]

|   | -14 | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 1 |
|---|-----|-----|-----|-----|-----|----|----|----|----|----|----|----|----|----|----|----|
| A | 11.1 | 12.7 | 3.2 | 4.8 | 12.7 | 8.7 | 16.7 | 16.7 | 12.7 | 9.5 | 26.2 | 6.3 | 100 | 0.0 | 21.4 |
| C | 36.5 | 30.9 | 19.1 | 23.0 | 34.9 | 39.7 | 34.9 | 40.5 | 40.5 | 36.5 | 33.3 | 68.2 | 0.0 | 0.0 | 7.9 |
| G | 9.8 | 10.3 | 15.1 | 22.7 | 8.7 | 9.8 | 16.7 | 6.8 | 2.4 | 6.3 | 13.5 | 0.0 | 0.0 | 100 | 42.7 |
| U | 38.9 | 41.3 | 52.7 | 55.6 | 42.2 | 40.5 | 30.9 | 37.3 | 44.4 | 41.6 | 27.0 | 25.4 | 0.0 | 0.0 | 7.9 |

c) AT-AC group: 8 annotated examples + 2 examples recovered from annotation errors

\[ S_{90}|ATA_{100}T_{100}C_{100}C_{100}T_{100}|T_{90}T_{70} \]

\[ T_{70}G_{60}C_{70}|NC_{60}AC|A_{60}T_{60} \]

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AC002397</td>
<td>TGCCAGATG</td>
<td>atatccttggt</td>
<td>ctctgtctcac</td>
<td>CTGAGAGAAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC004976</td>
<td>GAAAGACCA</td>
<td>atatccttctgg</td>
<td>actctgtcatac</td>
<td>GAAGAGCCTCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF1346129</td>
<td>TATGCTAGA</td>
<td>atatcctctact</td>
<td>acgttttgcag</td>
<td>ATGACCAAAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL021578</td>
<td>AGCTGACCC</td>
<td>atatccttggg</td>
<td>ttaacgccac</td>
<td>TGCCCAAGCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L102995</td>
<td>ATGCTGAGA</td>
<td>atatccttttag</td>
<td>aatgtctatac</td>
<td>ATGCTGACCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U39692</td>
<td>AGATTAGAGA</td>
<td>atatcccttcct</td>
<td>aactcgacagac</td>
<td>ATTTGTCGAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U47924</td>
<td>TGCCAGATG</td>
<td>atatcctttgc</td>
<td>aacctctccac</td>
<td>CTGAGAGAAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U53004</td>
<td>GAAAGACCA</td>
<td>atatccttctgg</td>
<td>aacctctccac</td>
<td>GAAGAGCCTCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 9.4

Consensus sequences and weight matrices for major groups of splice site pairs. Numbering splice site positions is provided relative to the splice junction along the gene sequence.
Table 9.4
Significance of various characteristics for discrimination of donor splice sites

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual $D^2$</td>
<td>9.3</td>
<td>2.6</td>
<td>2.5</td>
<td>0.1</td>
<td>1.5</td>
<td>0.01</td>
<td>0.4</td>
</tr>
<tr>
<td>Combined $D^2$</td>
<td>9.3</td>
<td>11.8</td>
<td>13.6</td>
<td>14.9</td>
<td>15.5</td>
<td>16.6</td>
<td>16.8</td>
</tr>
</tbody>
</table>

1, 2, 3 are the triplet preference (13) of consensus ($\ldots$ $\ldots$ $\ldots$), intron G rich (+7 $\ldots$ +50) and coding regions ($\ldots$ $\ldots$ $\ldots$), respectively. 4 is the number of significant triplets in the consensus region. 5 and 6 are the octanucleotide preference for being coding 54 bp region on the left and for being intron 54 bp region on the right of donor splice site junction. 7 is the number of G bases, GG doublets, and GGGG triplets in +8 $\ldots$ +50 intron G rich region.

9.2.8 Computational Recognition of Splice Sites

Computational analysis of splice site sequences demonstrates that their consensus are slightly specific for distant classes of organisms (Senapathy et al. 1990; Mount 1993) and that some important information is encoded by the sequences outside the short conserved regions. There were several attempts to develop accurate splice site identification algorithms based on consensus sequences or weight matrices, which take into account the information about open reading frames, free energy of base pairing of RNA with snRNA, and other sequence features. These approaches reached the accuracy of about 80 percent for the prediction splice site positions (Nakata et al. 1985; Gelfand 1989). More accurate prediction was achieved by neural network algorithms (Lapedes et al. 1988; Brunak et al. 1991). To demonstrate what sequence features can help identify authentic splice sites, we will describe a simple method using a linear discriminant function (Solovyev and Lawrence 1993; Solovyev et al. 1994).

Donor Splice Site Recognition To test the significance of different sequence features by liner discriminant approach described in section 9.2.5, seven characteristics were selected for donor splice site identification. In table 9.4, we can see the Mahalonobis distances showing the significance of each characteristic. The strongest characteristic of donor sites is triplet composition in consensus region ($D^2 = 9.3$), in the adjacent intron region ($D^2 = 2.6$), and in coding region ($D^2 = 2.5$). Other significant characteristics are a number of significant triplets in conserved consensus region; the number of G bases, GG doublets, and GGGG triplets; and the octaplet composition of the coding and intron regions.

A rigorous testing of several splice site prediction programs on the same sets of new data demonstrated that the linear discriminant function (implemented in SPL program: http://www.genomic.sanger.ac.uk) provides the most accurate local donor site recognizer (table 9.5) (Milanesi and Rogozin 1998).
Table 9.5
Comparing the prediction accuracy of local donor splice site recognizers

<table>
<thead>
<tr>
<th>Method/program</th>
<th>False positives</th>
<th>False negatives</th>
<th>CC</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight matrix</td>
<td>5.0%</td>
<td>20%</td>
<td>0.22</td>
<td>Guigo et al. 1991</td>
</tr>
<tr>
<td>Neural network (NETGENE)</td>
<td>16.3%</td>
<td>6.7%</td>
<td>0.35</td>
<td>Brunak et al. 1991</td>
</tr>
<tr>
<td>Discriminant analysis (SPL)</td>
<td>22.0%</td>
<td>2.3%</td>
<td>0.51</td>
<td>Solovyev et al. 1994</td>
</tr>
</tbody>
</table>

The accuracy is averaged for three tested sets.

A single weight matrix provides less accurate recognition than more sophisticated approaches, but it can be easily recomputed for new organisms and is very convenient to use in probabilistic HMM gene prediction methods. Using maximal dependence decomposition procedure (Burge 1998), we constructed five donor recognition weight matrices for different subsets of splice site sequences. The subclassification of donor signals and the matrices constructed, based on 22,306 EST supported splice sites, are presented in figure 9.5. Performance of these matrices comparing with the other methods was estimated on the Burset and Guigo (1996) data set (figure 9.6). It shows that several weight matrices provide better splice site discrimination than just one. However, their discriminative power is similar to the triplet matrix and lower for most levels of sensitivity than the linear discriminant function of the SPL program.

Acceptor Splice Site Recognition The performance of acceptor site recognition (Rogozin and Milanesi 1997) by different computational methods is presented in table 9.6. We can see that acceptor site recognition accuracy is lower than the accuracy of predicting donor sites. The linear discriminant function (Solovyev et al. 1994) implemented in the SPL program demonstrates the higher accuracy.

Burge (1998) demonstrated that the first order Markov chain model formula (9.11) based on dinucleotide frequencies of [−20, +3] acceptor site region gives slightly better discrimination than the simple weight matrix model. Such a model was implemented in the Genscan gene prediction method (Burge and Karlin 1997). Recently, Thanaraj (2000) evaluated several splice site recognitions. Among them, the SPL program remains the best local recognizer. Of course, complex gene prediction systems (HMM gene, Genscan, Fgenes, Fgenesh, and some intermediate approaches such as NetGene2) using a lot of global information about optimal exon (or splice site) combination will have a better accuracy level. However, they cannot be applied to study possible alternative splice sites in a particular gene. Local recognizers might be useful for such tasks.
9.2.9 Poll Promoter Recognition

Because each eukaryotic polymerase II promoter has a unique selection and arrangement of regulatory elements providing a unique program of gene expression, the computational identification of promoter sequences in genomic DNA is an extremely difficult problem (see chapter 2). Here we consider a version of promoter recognition program TSSW (Solovyev and Salamov 1997), several modules of which are implemented in the gene prediction program FGENES (Solovyev 1997). In the last version of TSSW, it was assumed that TATA+ and TATA− promoters have very different sequence features, so these groups were analyzed separately. Potential
Figure 9.6
Comparing accuracy of donor splice site recognizers: single weight matrix, five weight matrices, matrix of triplets, linear discriminant function.

Table 9.6
Comparing the prediction accuracy of local acceptor splice site recognizers

<table>
<thead>
<tr>
<th>Method/program</th>
<th>False positives</th>
<th>False negatives</th>
<th>CC</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight matrix</td>
<td>2.3%</td>
<td>53%</td>
<td>0.13</td>
<td>Guigo et al. 1992</td>
</tr>
<tr>
<td>Consensus MAG/GURAGU</td>
<td>6.0%</td>
<td>18%</td>
<td>0.27</td>
<td>Mount 1982</td>
</tr>
<tr>
<td>Consensus MAG/GURAGU</td>
<td>6.0%</td>
<td>18%</td>
<td>0.27</td>
<td>Mount 1982</td>
</tr>
<tr>
<td>Five consensuses</td>
<td>4.2%</td>
<td>15%</td>
<td>0.31</td>
<td>Rogozin and Milanesi 1997</td>
</tr>
<tr>
<td>Neural network (NETGENE)</td>
<td>25.0%</td>
<td>2.7%</td>
<td>0.51</td>
<td>Brumak et al. 1991</td>
</tr>
<tr>
<td>Discriminant analysis (SPL)</td>
<td>10.0%</td>
<td>3.0%</td>
<td>0.56</td>
<td>Solovyev et al. 1994</td>
</tr>
</tbody>
</table>

The accuracy is averaged for three tested sets.

TATA+ promoter sequences were selected by the value of the score of Bucher TATA box weight matrix (Bucher 1990) with the threshold close to the minimal score value for the TATA+ promoters in the learning set. Such a threshold divides the learning sets of known promoters into approximately equal groups. Selected significant characteristics of both groups found by discriminant analysis are presented in table 9.7. This analysis demonstrated that TATA− promoters have much weaker general features than TATA+ promoters. Probably TATA− promoters possess more gene specific structure; they will be extremely difficult to predict by any general-purpose methods.
Table 9.7  
Characteristics of promoter sequences used by TSSW programs for identification of TATA+ and TATA− promoters

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>D² for TATA+ promoters</th>
<th>D² for TATA− promoters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hexaplets −200−45</td>
<td>2.6</td>
<td>1.4 (−100−1)</td>
</tr>
<tr>
<td>TATA box score</td>
<td>3.4</td>
<td>0.9</td>
</tr>
<tr>
<td>Triplets around TSS</td>
<td>4.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Hexaplets +1−140</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Sp1-motif content</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>TATA fixed location</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Cpg content</td>
<td>1.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Similarity −200−100</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>Missense Density (MD)</td>
<td>4.5</td>
<td>3.2</td>
</tr>
<tr>
<td>Direct/Inverted MD +100−1</td>
<td>4.0</td>
<td>3.3 (−100−1)</td>
</tr>
<tr>
<td>Total Mahalonobis distance</td>
<td>11.2</td>
<td>4.3</td>
</tr>
<tr>
<td>Number promoters/nonpromoters</td>
<td>203/4000</td>
<td>193/74000</td>
</tr>
</tbody>
</table>

For each position of a given sequence, the TSSW program evaluates the occurrence of TSS using two linear discriminant functions (for TATA+ and TATA− promoters) with characteristics computed in the (−200, +50) region around the given position. If we find a TATA-box (using TATA-box weight matrix) in this region, then we compute the value of LDF for TATA+ promoters, otherwise the value of LDF for TATA-less. Only one prediction with the highest score of LDF and greater than some threshold will be selected within any 300-bp region. If we observe a lower scoring promoter predicted by the TATA− LDF near a higher scoring promoter predicted by TATA+ LDF, then the first prediction is also displayed as a potential enhancer region.

Figure 9.7 shows an example of TSSW program results for the sequence of the human connexin 32 (GJB1) gene (GenBank accession number L47127). The TSSW predicts one enhancer at position 246 and one potential TSS at position 428, with corresponding TATA-box at the position 388. GenBank annotation based on experimental data shows real TATA-signal in positions 389–394. TSSW also optionally lists all potential TF binding sites around predicted promoters or enhancers (figure 9.7). It outputs the position, the strand (+/−), TRANSFAC identifier, and the consensus sequences of found sites. The information about these sites may be of interest for researchers studying the transcription of a particular gene.

Due to a high false positive rate of promoter prediction in long genomic sequences, they are more useful when we can remove some predictions about the positions of coding regions. The TSSW was additionally tested on the several GenBank entries that have information about experimentally verified TSS and were not included in the
tssw. Mod Dec 20 20:02:08 EST 2000
> Homo sapiens connexin 32 (GJB1) gene
Length of sequence: 950
Thresholds for TATA+ promoters - 0.45, for TATA-/enhancers - 3.70
2 promoter/enhancer(s) are predicted
Enhancer Pos: 246 LDF - 14.25
Promoter Pos: 428 LDF - 0.64 TATA box at 388 18.68
Transcription factor binding sites:
for position - 246
179 (+) CHICKSACRA CCGCCC
241 (-) CHICKSACRA CCGCCC
226 (+) CHICKSACRA CCGCCC
188 (+) MAIZEADH1 COTOG
26 (+) YSADH2_01 TCTCC
168 (+) YSADH2_01 TCTCC
42 (-) YSADH2_01 TCTCC
215 (+) HSFAPOE_08 GGGCGG
236 (+) HSFAPOE_08 GGGCGG
184 (+) HSFAPOE_08 GGGCGG
243 (-) HSFAPOE_09 GGGCGGG
146 (+) MOUSEHMK occaaACCTGCtctgctgaccc
t87 (-) RATTSAF_01 GTCGG
14 (+) RATTSAF_01 GTCTCT
50 (+) YSAG11_10 AGCCCT
116 (+) YSAG11_10 AGCCCT
226 (-) HSVEG_01 ccaACCGGg
214 (+) CHICKSACRA GGGGGGCGG
183 (+) CHICKSACRA GGGGGGCGG

for promoter at position - 428
179 (+) CHICKSACRA CCGCCC
423 (+) CHICKSACRA CCGCCC
312 (-) CHICKSACRA CCGCCC
241 (-) CHICKSACRA CCGCCC
226 (+) CHICKSACRA CCGCCC
289 (+) CHICKSACRA TATATA
188 (+) MAIZEADH1 COTOG
168 (+) YSADH2_01 TCTCC
215 (+) HSFAPOE_08 GGGCGG
236 (+) HSFAPOE_08 GGGCGG
307 (+) HSFAPOE_08 GGGCGG
418 (+) HSFAPOE_08 GGGCGG
184 (-) HSFAPOE_08 GGGCGG
243 (-) HSFAPOE_09 GGGCGG
357 (+) BPV15BPV1 ACCgtgtaCCTGT
368 (+) BPV15BPV1 ACCgtgtaCCTGT
357 (+) BPV15BPV1 ACCgtgtaCCTGT
368 (+) BPV15BPV1 ACCgtgtaCCTGT
357 (+) BPV15BPV1 ACCgtgtaCCTGT
368 (+) BPV15BPV1 ACCgtgtaCCTGT
417 (+) MOUSEHMK gcccagcctctccATTTGGgcagag
146 (+) MOUSEHMK cccaaACCTGCtctgctgaccc
381 (+) ADSE3_06 ggccagggTATAatctacctga

Figure 9.7
Results of promoter prediction by TSSW program in human connexin 32 (GJB1) gene (GenBank accession number L47127).
The lengths of sequences varied from 950 to 28,438 bp, with a median length of 2,938 bp. All true TSS in these sequences can be considered as correctly predicted, with an average of 1.5 false positives per sequence or 1 false positive per 3,340 bp. The distances between true TSS and the correctly predicted ones varied from exact matching to 196 bp, with the median deviation of about 15 bp, which means that half of the predictions are close to the experimental mapping of TSS with the estimated precision of ±5 bp (Perier et al. 2000).

The above prediction algorithm uses the propensities of each TF binding site independently, not taking into account their mutual orientation and positioning. At the same time, the transcription regulation is a highly cooperative process, involving the simultaneous binding of several transcription factors to their corresponding sites. In future algorithms we should analyze patterns of regulatory sequences, where mutual orientation and location of individual regulatory elements are necessary.

### 9.2.10 Recognition of PolyA Signals

A 3'-untranslated region (3'UTR) has a diversity of cytoplasmic functions affecting the localization, stability, and translation of mRNAs (Decker and Parker 1995). Practically all eukaryotic mRNAs undergo 3'-end processing, which involves endonucleotide cleavage followed by polyadenylation of the upstream cleavage product (Wahle 1995; Manley 1995). The formation of large RNA-protein complexes is essential for 3'-end processing (Wilusz et al. 1990). RNA sequences directing the binding of specific proteins are usually poorly conserved and often recognized in a cooperative fashion (Wahle 1995). Therefore, the approaches for poly-A signal identification use statistical characteristics of poly-A regions, which can reflect some unknown functional sequences.
There are three types of basic RNA sequences defining a 3'-processing site (WaWe 1995; Proudfoot 1991) (figure 9.8). The most conserved is the hexamer signal AAUAAA (polyA signal), situated 10–30 nucleotides upstream of the 3'-cleavage site. About 90 percent of sequenced mRNAs have a perfect copy of this signal. Two other types, the upstream and the downstream elements, are poorly conserved and characterized. Downstream elements are frequently located within approximately 50 nucleotides 3' of the cleavage site and often GU or U rich (Wahle and Keller 1992).

Comparing their sequences, McLachlan et al. (1985) suggest a possible consensus of one downstream element: YGUGUUYY. The efficiency of polyadenylation in a number of genes can be also increased by generally U-rich sequences upstream of AAUAAA (Wahle 1995).

A few computer programs were developed to identify 3'-processing. Yada et al. (1994) analyzed human DNA sequences in the vicinity of the poly-A signal, trying to distinguish them from other AATAAA sequences nonactive in polyadenylation (pseudo polyA signals). They found that C frequently appears on the upstream side of the AATAAA signal and T or C often appears on the downstream side, generating an extended consensus of poly-A signal: CAATAAA(T/C). Kondrakhin et al. (1994) constructed a generalized consensus matrix using 63 sequences of cleavage/polyadenylation sites in vertebrate pre-mRNA. The matrix elements were the absolute frequencies of triplets at each site position. Using this matrix for recognition of polyadenylation regions produces a very high number of false positive predictions.

A LDF recognition function for poly-A signal was developed by Salamov and Solovyev (1997). The prediction algorithm was realized in the POLYAH program. It searches for the AATAAA pattern by using weight matrix. After it finds the pattern, it computes the value of the linear discriminant function, defined by seven sequence characteristics around this position. The POLYAH program has been tested on the sequence of Ad2 genome, where for eight correctly identified sites, it predicts only four false sites.

Further improvement of poly-A recognition was reached in using a pair of quadratic discriminant function in the Polyadq program (Tabaska and Zhang 1999).
This program outperformed the POLYAH detection method and is the first that can detect significant numbers of ATTAAA-type signals.

9.3 ORF, Exon, and Single Gene Prediction Approaches

The first generation of computational gene finding programs searched for open reading frames with organism-specific codon usage (Staden and McLachlan 1982). These approaches worked successfully for bacterial genes (Staden 1984; Borodovsky et al. 1986), but short eukaryotic exons and spliced eukaryotic genes require algorithms taking into account additional information about functional signals. One application of such approaches is useful for predicting coding regions or coding ORF in partially or completely sequenced mRNA(EST) sequences. Several HMM-based predictors developed recently, such as BESTORF (Solovyev and Salamov 1999a) and ESTscan (Iseli et al. 1999), significantly improve the accuracy of earlier approaches.

The internal exon prediction program SORFIND (Hutchinson and Hayden 1992) was designed based on codon usage and Berg and von Hippel (1987) discrimination energy for intron-exon boundary recognition. The accuracy of exact internal exons prediction (at both 5′- and 3′-splice junctions and in the correct reading frame) by the SORFIND program reaches 59 percent, with a specificity of 20 percent. Snyder and Stormo (1993) applied a dynamic programming approach (an alternative to the rule-based approach) to internal exon prediction in the GeneParser algorithm. It recognized 76 percent of internal exons, but the structure of only 46 percent of the exons was exactly predicted when tested using entire GenBank entry sequences. The HEXON (Human EXON) program (Solovyev et al. 1994a), based on linear discriminant analysis, was one of the most accurate in exact internal exon prediction. It was recently upgraded to predict all type of exons (and renamed FEX—find exon) (Solovyev et al. 1994b). The FEX program can be useful to analyze a possible set of alternatively spliced exons in a given sequence in addition to the optimal variant of gene structure produced by exon assembling programs. In an effort to improve the accuracy of exon prediction, Zhang (1997) applied quadratic discriminant technique (in the MZEF program) as a direct extension of the classical liner discriminant approach used in the HEXON program. The statistical evaluation of MZEF predictions on 473 genes (partially included in MZEF training) demonstrated a better performance than the HEXON program.

Later, a number of single-gene prediction programs were developed to assemble potential eukaryotic coding regions into translatable mRNA sequences, selecting optimal combinations of compatible exons (Fields and Soderlund 1990; Gelfand 1990; Guigo et al. 1992; Dong and Searls 1994). Dynamic programming was suggested as
a fast method to find an optimal combination of pre-selected exons (Gelfand and Roytberg 1993; Solovyev and Lawrence 1993b; Xu et al. 1994), which is different from the approach in the GeneParser algorithm suggested by Snyder and Stormo (1993) to recursively search for exon-intron boundary positions. The FGENEH (Find GENE in Human) algorithm incorporated 5'-internal, and 3'-exon identification linear discriminant functions and a dynamic programming approach (Solovyev et al. 1994, 1995). Burset and Guigo (1996) conducted a comprehensive test for it and the other gene finding algorithms; the FGENEH program was one of the best in the tested group, having an exon prediction accuracy 10 percent higher than the others and the best level of accuracy on the protein level. A novel step in gene prediction approaches was the application of generalized hidden Markov models implemented in the Genie algorithm (Kulp et al. 1996). Genie is similar in design to GeneParser, but is based on a rigorous probabilistic framework. It is similar to FGENEH in performance (Kulp et al. 1996).

9.4 Multiple Gene Prediction by Discriminative and Probabilistic Approaches

Whole genome sequencing projects were initiated for a number of organisms, from bacteria to higher eukaryotes. They require gene-finding approaches that are able to identify many genes encoded in the genomic sequences. The most accurate multiple gene prediction programs include such HMM-based probabilistic approaches as Genescan (Burge and Karlin 1997) and Fgenes (Salamov and Solovyev 2000), Fgenes (discriminative approach) (Solovyev 1997), and Genie (generalized HMM with neural network splice site detectors) (Reese et al. 2000). In the next section, we will describe a general scheme of HMM-based gene prediction that was initially realized in the works of Dr. Haussler’s group (Krogh et al. 1994; Kulp et al. 1996). This pattern-based approach can also be considered as a variant in which transition probabilities are not taken into account.

9.4.1 HMM-Based Eukaryotic Gene Identification

Exons, introns, 5'-, and 3'-untranslated regions are different components (states) of gene structure that occupy k non-overlapping subsequences of a sequence \( X = \bigcup_{i=1,k} X_i \). There are 35 states constituting an eukaryotic gene model, considering direct and reverse chains as possible gene locations (figure 9.9). The absence of protein coding characteristics reduces significantly prediction accuracy of noncoding 5'- and 3'-exons (and introns; therefore, they are not considered in the current gene prediction algorithms. The other 27 states consist of six exon states (first, last, single, and three types of internal exons due to three possible reading frames) and seven non-
Figure 9.9
Different states and transitions in eukaryotic gene model.

coding states (three intron, noncoding 5' and 3', promoter, and polyA) in each chain, plus the noncoding intergenic region.

A gene structure can be considered as the ordered set of state/subsequence pairs, \( \phi = \{(q_1, x_1), (q_2, x_2), \ldots, (q_k, x_k)\} \), called the parse. We call the predicted gene structure to be such a parse \( \phi \) that the probability \( P(X, \phi) \) of generating \( X \) according to \( \phi \) is maximal over all possible parses (or when some score is optimal in some meaningful sense, i.e., best explains the observations [Rabiner 1989]). This probability can be computed using statistical parameters describing a particular state and generated from the training set of known gene structures and sequences.

Successive states of this HMM model are generated according to the Markov process with the inclusion of explicit state duration density. A simple technique based on the dynamic programming method for finding the optimal parse (or the single best state sequence) is called the Viterbi algorithm (Forney 1973). The algorithm requires \( O(N^2D^2L) \) calculations, where \( N \) is the number of states, \( D \) is the longest duration, and \( L \) is the sequence length (Rabiner and Juang 1993). Burge (1997) introduced a helpful technique to reduce the number of states and simplify computations by
modeling noncoding state length with a geometrical distribution. We will shortly consider the algorithm of gene finding using these technique, which was initially implemented in the Genscan program (Burge 1977; Burge and Karlin, 1997) and used later in the Fgenesh program (Salamov and Solovyev 2000). As any valid parse will consist of only alternating series of noncoding and coding states, NCNCNC, . . ., NCN, we need only 11 variables, corresponding to the different types of N states. For each sequence position (starting from 1), we select the maximum joint probability to continue the current state or to move to another noncoding state defined by a coding state (from a pre-computed list of possible coding states) that terminates at the analyzed sequence position. The parse probability is

\[
P(X, \delta) = P(q_1) \sum_{i=1}^{k-1} P(x_i | l(x_i), q_i) P(l(x_i) | q_i) P(l(x_i) | q_{i+1}, q_i) P(x_i | l(x_i), q_k) P(l(x_k) | q_k)
\]

where \(P(q_1)\) denotes the initial state probability; \(P(x_i | l(x_i), q_i) P(l(x_i) | q_i)\) and \(P(q_{i+1}, q_i)\) are the independent joint probabilities of generation of the subsequence \(x_i\) of length \(l\) in the state \(q_i\) and transitioning to the \(q_{i+1}\) state. \(P(x_i | l(x_i), q_i) P(l(x_i) | q_i)\) is a production of a probability of generation \(l\)-length sequence \(x_i\) and the probability of observing such an \(l\)-length sequence in the state \(q_i\), which are computed using the sequence of \(x_i\) and the statistical data from a training set of known genes. To compute \(P(x_i | l(x_i), q_i)\) for an internal exon state, we use donor and acceptor site models based on position specific weight matrices and frame-specific Markov models based on hexaplet frequencies in exons and introns.

**Search for Optimal Parse**  Let us define the best score (the highest joint probability \(\gamma_j[j]\) of the optimal parse of the subsequence \(S_{i,j} = [s_1, s_2, \ldots, s_j]\), which ends in state \(q_i\) at position \(j\)). Assume a set \(A_j\) of coding states \(\{c_k\}\) of lengths \(\{d_k\}\), starting at positions \(\{m_k\}\) and ending at position \(j\), which have the previous states \(\{b_i\}\). The length distribution of state \(c_k\) is denoted by \(f_{c_k}(d)\). The searching procedure can be stated as follows:

**INITIALIZATION:**

\[
\gamma_i(1) = \pi_i P_i(s_1) p_i \quad \text{and} \quad Z_i(1) = 0, \quad i = 1, \ldots, 11
\]

**RECURSION:**

\[
\gamma_i(j+1) = \max \left\{ \gamma_i(j) p_i P_i(s_{j+1}), \max_{a_i | d_i} \left\{ \gamma_i(m_k - 1)(1 - p_{n_i}) f_{c_k}(d_k) P(S_{m_k,i}) \times t_{c_k}(p_i P_i(s_{j+1})) \right\} \right\} \quad i = 1, \ldots, 11, \quad j = 1, \ldots, L - 1.
\]
Finding Genes by Computer

\[ \gamma_i(L+1) = \max \left\{ \gamma_i(L), \max_{\alpha \in L} \{ \gamma_i(m_k - 1)(1 - p_{th_i}) \theta_{b_i \alpha} f_{c_i}(d_k) P(S_{m_{a_i}}) t_{c_i, i} \} \right\} \]

On each step we record the location and type of transition maximizing the functional to restore the optimal set of states (gene structure) by a backtracking procedure. Most parameters of these equations can be calculated from the learning set of known gene structures. Instead of scores of coding states \( P(S_{m_{a_i}}) \), it is better to use log-likelihood ratios, which do not produce scores below the limits of computer precision.

This technique to predict multiple eukaryotic genes was initially implemented in the Genscan algorithm (Burge and Karlin 1997). Several other HMM-based gene prediction programs were developed later: Veil (Hederson et al. 1997), HMMgene (Krogh 1997), Fgenes (Salamov and Solovyev 1999, 2000), a variant of GeneMark (Kulp et al. 1996), and GeneMark (Lukashin and Borodovsky 1998). Fggenes (Find GENES using Hmm) is currently the most accurate program. It is different from Genscan in computing the coding scores of potential exons, where a priori probabilities of exons were taken into account according to the Bayes theorem. As a result, the coding scores of potential exons are generally lower than in Genscan. Some minor differences exist in the functional signal description and preparing of training sets to compute specific parameters for each model organism, such as human, Drosophila, nematode, yeast, Arabidopsis, monocotyledons, and so on. Coding potentials were calculated separately for four isochores (human) and for two isochores (other species). The run time of Fgenes is practically linear; the current version has no practical limit on the length of analyzed sequence. Prediction of about one thousand genes in 34.5 MB of chromosome 22 sequence takes about 1.5 minutes with a Dec-alpha processor EV6.

9.4.2 Discriminant Analysis–Based Gene Prediction

The Fgenes (Find GENES) program predicts multiple genes using dynamic programming and discriminant classifiers to generate a set of exon candidates. The following major steps describe analysis of genomic sequences by the Fgenes algorithm:

1. Create a list of potential exons by selecting: ATG ... GT, AG-GT, AG. Stop sequence regions having exons scores higher than the specific thresholds depending on GC content (four groups);
2. Find a set of compatible exons with the maximal total score. Guigo (1999) described an effective procedure for finding such a set. Fgenes uses a simpler variant of
Table 9.9
Significance of internal exon characteristics selected by LDA

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a Individual $D^2$</td>
<td>15.0</td>
<td>12.1</td>
<td>0.4</td>
<td>0.2</td>
<td>1.5</td>
</tr>
<tr>
<td>b Combined $D^2$</td>
<td>15.0</td>
<td>25.3</td>
<td>25.8</td>
<td>25.8</td>
<td>25.9</td>
</tr>
</tbody>
</table>

Characteristics 1 and 2 are the values of donor and acceptor site LD recognition functions. 3 gives the octanucleotide preference for being coding of potential exon region. 4 gives the octanucleotide preference for being intron 70-bp region on the left and 70-bp region on the right of potential exon region.

this algorithm: we order all exon candidates according to their 3'-end positions. Then, going from the first to the last exon, we select for each exon the maximal score path (compatible exons combination) terminated by this exon using a dynamic programming approach. Include in the optimal set either this exon or an exon with the same 3'-splicing pattern ending at the same position or earlier, whichever has the higher maximal score path.

3. Take into account promoter or polyA scores (if predicted) in terminal exon scores. The run time of the algorithm grows approximately linearly with the sequence length.

Fgenes is based on usage linear discriminant functions developed for identification of splice sites, exons, promoter, and polyA sites (Solovyev et al. 1994; Solovyev and Salamov 1997). We consider these functions for internal exons to demonstrate what sequence features are important to exon identification.

**Internal Exon Recognition** We consider as potential internal exons all open reading frames in a given sequence flanked by AG (on the left) and GT (on the right). The structure of such exons is presented in figure 9.10. The values of five exon characteristics were calculated for 952 authentic exons and for 690,714 pseudo-exon training sequences from the set. Table 9.9 gives the Mahalonobis distances, showing the significance of each characteristic. We can see that the strongest characteristics are the recognition functions of flanking donor and acceptor splice sites ($D^2 = 15.04$ and $D^2 = 12.06$, respectively). The preference of ORF as a coding region has $D^2 = 1.47$ and adjacent left intron region has $D^2 = 0.41$ and right intron region has $D^2 = 0.18$.

The accuracy of the discriminant function based on these characteristics was estimated on the recognition of 451 exon and 246,693 pseudo-exon sequences from the test set. The sensitivity of exact internal exon prediction is 77 percent, with a specificity of 79 percent. At the level of individual nucleotides, the sensitivity of exon prediction is 89 percent, with a specificity of 89 percent; the sensitivity of intron positions prediction is 98 percent, with a specificity of 98 percent. This accuracy is better than that demonstrated by dynamic programming and neural network based methods.
Finding Genes by Computer

![Diagram of gene structure](image)

**Figure 9.10**
Different functional regions of the first (a), internal (b), last (c) corresponding to components of recognition functions. Single exons include left and right characteristics of first and last exons, respectively.

(Sayder and Stormo 1993), which have a 75 percent sensitivity of the exact internal exons prediction, with a specificity of 67 percent.

5'-coding exon recognition LDF uses the average value of positional triplet preference in the (-15, 10) region around ATG codon (instead of donor splice site score). 3'-exon coding region recognition LDF includes the average value of positional triplet preference in the (-10, 30) region around the stop codon (instead of the acceptor site score). The recognition function of single exons combines corresponding characteristics of 5'- and 3'-exons (figure 9.10) (Solovyev et al. 1994; Solovyev and Salamov 1997). Features describing sequences near initial and stop codons have much less discriminative power than the splice site characteristics; therefore, terminal and short single exons have a lower accuracy of recognition.

### 9.5 Accuracy of Gene Identification Programs

Burset and Guigo (1996) specially selected a set of 570 single-gene sequences of mammalian genes, which they used to evaluate the performance of many gene finding approaches. The results of this test are presented in table 9.10. Of course, some of these data have only a historical value to show the progress in gene finding development, and some of these programs have been improved since the test. We can see that the best programs on average predict accurately 93 percent of exon nucleotides...
Table 9.10
Characteristics of accuracy for the gene prediction programs on single gene sequences of Burset and Guigo 1996 dataset

<table>
<thead>
<tr>
<th>Algorithm/ program</th>
<th>Sn (exons)</th>
<th>Sp (exons)</th>
<th>Sn nucleotides</th>
<th>Sp nucleotides</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fgenesh</td>
<td>0.84</td>
<td>0.86</td>
<td>0.94</td>
<td>0.95</td>
<td>Salamov &amp; Solovyev 1998</td>
</tr>
<tr>
<td>Fgenes</td>
<td>0.83</td>
<td>0.82</td>
<td>0.93</td>
<td>0.93</td>
<td>Solovyev 1997</td>
</tr>
<tr>
<td>Genscan</td>
<td>0.78</td>
<td>0.81</td>
<td>0.93</td>
<td>0.93</td>
<td>Burge &amp; Karlin 1997</td>
</tr>
<tr>
<td>Fgeneh</td>
<td>0.61</td>
<td>0.64</td>
<td>0.77</td>
<td>0.88</td>
<td>Solovyev et al. 1995</td>
</tr>
<tr>
<td>Morgan</td>
<td>0.58</td>
<td>0.51</td>
<td>0.83</td>
<td>0.79</td>
<td>Salsberg et al. 1998</td>
</tr>
<tr>
<td>Veil</td>
<td>0.53</td>
<td>0.49</td>
<td>0.83</td>
<td>0.79</td>
<td>Henderson et al. 1997</td>
</tr>
<tr>
<td>Genie</td>
<td>0.55</td>
<td>0.48</td>
<td>0.76</td>
<td>0.77</td>
<td>Kulp et al. 1996</td>
</tr>
<tr>
<td>GenLang</td>
<td>0.51</td>
<td>0.52</td>
<td>0.72</td>
<td>0.79</td>
<td>Dong &amp; Sears 1994</td>
</tr>
<tr>
<td>Sorfind</td>
<td>0.42</td>
<td>0.47</td>
<td>0.71</td>
<td>0.85</td>
<td>Hutchinson &amp; Hyden 1992</td>
</tr>
<tr>
<td>GenelD</td>
<td>0.44</td>
<td>0.46</td>
<td>0.63</td>
<td>0.81</td>
<td>Guigo et al. 1992</td>
</tr>
<tr>
<td>Grail2</td>
<td>0.36</td>
<td>0.43</td>
<td>0.72</td>
<td>0.87</td>
<td>Xu et al. 1994</td>
</tr>
<tr>
<td>GeneParser2</td>
<td>0.35</td>
<td>0.40</td>
<td>0.66</td>
<td>0.79</td>
<td>Snyder &amp; Stormo 1995</td>
</tr>
<tr>
<td>Xpound</td>
<td>0.15</td>
<td>0.18</td>
<td>0.61</td>
<td>0.87</td>
<td>Thomas &amp; Skolnick 1994</td>
</tr>
</tbody>
</table>

Sn (sensitivity) = number of exactly predicted exons/number of true exons (or nucleotide); Sp (specificity) = number of exactly predicted exons/number of all predicted exons. Accuracy data for programs developed before 1996 were estimated by Burset and Guigo (1996). The other data were produced by the authors of the corresponding programs.

(Sn = 0.93), with just 7 percent false positive predictions. However, the accuracy on the nucleotide level does not completely reflect the quality of gene structure prediction because missing small exons and the imperfect location of exon ends will not much affect its value. Therefore, it is important to provide the accuracy of exact exon prediction level, which is usually lower than at the nucleotide level.

The table clearly demonstrates that the recent multiple gene prediction programs such as Fgenesh, Fgenes, and Genscan significantly outperform the older approaches. The exon identification rate is actually even higher than the presented data because the overlapped exons were not counted as true predictions in exact exon accuracy evaluation. Yet there is still room for significant improvement. The accuracy of exact gene prediction is only 59 percent for Fgenesh, 56 percent for Fgenes, and 45 percent for Genscan programs computed on this relatively simple test with single gene sequences.

A more practical task is to identify multiple genes in long genomic sequences containing genes in both DNA strands. We selected a test set of 19 long genomic sequences of 26,000–240,000 bp and 19 multigene sequences with 2–6 genes from GenBank to compare performance of gene-finding programs in analyzing genomic DNA. Table 9.11 demonstrates the results of gene prediction for these data. The results show that...
the accuracy is still pretty good on the nucleotide and exon level, but exact gene
prediction is lower than for the test with short single gene sequences. Sensitivity for
exact internal exon prediction is 85–90 percent, but 5'-, 3', and single exons have a
prediction sensitivity of about 50–75 percent, which can partially explain relatively
low level of exact gene prediction. As a result, we observe the splitting of some actual
genes and/or joining some other multiple genes into a single one.

Another limitation of current gene-finding programs is that they cannot detect the
nested genes, that is, genes located inside introns of other genes. This is one of the
future directions for improvement of gene-finding software. Although this is probably
a rare event for the human genome, for organisms like Drosophila, it presents a real
problem. For example, annotators identified 17 examples of such cases in the Adh
region (Ashburner et al. 1999). Masking repeats is important. It significantly increases
the specificity of prediction.

<table>
<thead>
<tr>
<th>Program</th>
<th>Sequences/Genes</th>
<th>Accuracy per nucleotide</th>
<th>Accuracy per exon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sn</td>
<td>Sp</td>
<td>CC</td>
</tr>
<tr>
<td>Fgenes</td>
<td>38/77 M_r</td>
<td>0.94</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>0.94</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>GenScan</td>
<td>38/77 M_r</td>
<td>0.93</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.70</td>
<td>0.79</td>
</tr>
<tr>
<td>Fgenes</td>
<td>38/77 M_r</td>
<td>0.91</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.76</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Me, missing exons; WE, wrong exons. M_r lines provide predictions on sequences with masked repeats. Sn_o, exon prediction accuracy including overlapping exons.

9.6 Knowledge of Similar Protein or EST Can Improve Gene Prediction

Automatic gene prediction approaches can take into account some information about
exon similarity with a known protein or EST (Gelfand et al. 1996; Xu and Uberbacher
is a modification of the Fgenes algorithm, which uses additional information from
available similar proteins. These proteins can be acquired by running Fgenes on a
given sequence. Then the predicted proteins (or amino acid fragments translated
from predicted exons) are used to select similar proteins in some protein database.
After that, we can use selected proteins to improve prediction accuracy. Fgenes+
reads the protein homolog sequence and aligns all predicted potential exons with that
protein using the Smith-Waterman algorithm, as implemented in the Sim program
(Huang et al. 1990) or the Lial (local iterative alignments) algorithm developed by
Seledsov and Solovyev (1999). To improve the computational time, all overlapped
exons in the same reading frame are combined into one sequence and aligned only
once.

Fgenesh+ includes two major additions to the Fgenesh algorithm: augmentation
of the exon score (for exons having detected similarity) by an additional term pro-
portional to the alignment score and imposing a penalty for the adjacent exons in a
dynamic programming procedure, when the distance between their corresponding
protein segments is significantly different from the distance between the correspond-
ing fragments of a similar protein. We tested Fgenesh+ on a set of 61 GenBank
human sequences, which have imperfect ab initio Fgenesh predictions and known
protein homologs from other organisms with identity varying from 99 percent to 40
percent. The results of applying Fgenesh+ to these sequences show (table 9.12) that
when the alignment covers the entire lengths of both proteins, the accuracy increases
(relative to Fgenesh) and the improvement does not depend significantly on the level
of percent identity (for ID > 40 percent). This feature makes valuable the proteins
from distant organisms for improving the accuracy of gene identification. Having a
sequence of the human genome, we can find where in the genomic sequence a given
protein is located using Blast-like search in all predicted proteins of this genome.
Then we use Fgenesh+ for a prediction of the full-length mRNA (its coding part) for
a given protein using its sequence and the selected genomic sequence. Recently we
have developed a Fgenesh++ script, which initially predicts genes using the Fgenesh
program and then selects from NR (nonredundant protein database) similar proteins
for predicted genes using the Dbscan program (a Blast-like program, but about 10
times faster). The Fgenesh++ script uses found protein sequences to improve initial
gene prediction and can automatically generate annotation of the entire chromosome.
Similar to the Fgenesh+ algorithm scheme of exploiting known EST/cDNA information to improve accuracy of gene identification is the program Fgenesh_c (Salamov and Solovyev 2000a). Fgenesh+ and Fgenesh_c are very fast programs. For example, gene prediction by Fgenesh+ for a sequence of 80,000 bp with a protein of eight hundred amino acids takes about 1 second on an EV6 processor in a Dec-alpha computer.

9.7 Annotation of Genomic Sequences

GenBank (Benson et al. 1999) and EMBL (Stoesser et al. 1999) databases have for many years collected information about sequences of different genomes. A sequence-based structure of these databases often produces annotations of one gene in many different records when several gene fragments are sequenced independently. Last year, the vast amount of sequence information was produced by genome sequencing projects. Absence of experimental information about genes in a major part of these sequences makes valuable a presentation of computationally identified genes to provide positional cloners, gene hunters, and others with the gene candidates contained in finished and unfinished genomic sequences. Using these predictions, the scientific community can experimentally work with most real genes, because gene finding programs usually predict correctly most exons in a gene sequence.

9.7.1 Gene-Centered InfoGene Database

The InfoGene database database (http://www.softberry.com/inf/infodb.html) is created to collect and interactively work with information about known and predicted gene structures of human and other model genomes. Known genes are presented in 17 separate divisions (including human, mouse, Drosophila, nematode, Arabidopsis, rice, maize, and wheat), which contain records uniting available information about a particular gene from many GenBank (Release 119) entries. The human InfoGene division, for example, contains about 20,791 genes (including 16,141 partially sequenced genes), 54,558 coding regions, 83,488 exons, and about 58,000 donor and acceptor splice sites. This information can be applied to create different sets of functional gene components for extraction of their significant characteristics as used in gene prediction systems.

The interactive Java Viewer of Gene Structures has been designed by Igor Seledtsov and Victor Solovyev (1999) to visually inspect the gene structure of Infogene entries of known genes and predicted genes and to use for analysis of different gene prediction algorithms in annotating genomic sequences from genome sequencing projects.
The viewer has four main panels (see figures 9.2, 9.12): General View, Detail View, Locus Selection Panel, and the Output Message Panel. Both the General View Panel and the Detail View Panel have horizontal zoom scroll bars at the bottom of the windows. You can also zoom in and out by entering a scaling value from a pull-down menu: Action > Set Horizontal Scaling Factor.

9.7.2 General View Panel

The General View Panel shows all genes found in a given locus. If the locus contains overlapping genes, such as alternatively spliced ones, every such gene is displayed on its own line. As an example, let us look at the InfoGene locus HSAB001898. This locus should be automatically displayed in the General View Window if you have chosen the default setting from the InfoGene page (the button “Show data” on this page). This example is also shown on the picture above. Genes are shown as red bars. When the mouse cursor points to one of the genes, the gene’s name is displayed in the Output Message Panel. If you press and hold the right mouse button, detailed information about the gene is displayed in a separate temporary window. When the button is released, the temporary window disappears. The same operation performed with the Shift key leaves the temporary window open after the release of the mouse button. The number of temporary information windows that can be opened at the same time is unlimited. You can mark genes or groups of genes in General View Panel by pressing and dragging the left mouse button. Marked gene(s) will then be displayed in the Detail View Window, replacing its previous content. If you press the Shift key at the same time as pressing and dragging the left mouse button, you can add new marked regions to the current ones. Then all marked regions will be displayed one under another in the Detail View Panel.

9.7.3 Detail View Panel

This panel offers detailed view of selected genes. Each selected gene or region is drawn on a separate line, the number of which is unlimited. Dark gray bars represent genes, yellow bars show exons, red bars show coding regions, and green bars show gene regions that are not included in a transcript. Symbols < and > at the end of exon or coding region mean that their exact boundaries are unknown. White separators that cut through genes separate the unsequenced regions. Colored triangles above genes represent functional signals: black—CAAT box, blue—TATA box, green—transcription start point, red—PolyA signal, pink—polyA site. When the mouse cursor points to any bar or triangle, information about this object is displayed in the Output Message Panel. If the mouse cursor goes across a gene, the gene name is displayed in the Output Message Panel.
9.7.4 Locus Selection Panel

This panel has five fields: Selection List, Locus Info Button, Back Button, Forward Button, and Search Input Window. Selection List shows the list of sequence identifiers that satisfy current criteria, set forth in the Divisions, Options, and Search Fields menus of the Viewer. The Divisions menu allows the user to choose the source of data: Genbank/Infogen known genes for several taxonomic groups, or predicted genes for several organisms. The Options menu allows the user to choose a field to search in Infogen loci, GenBank Identifiers, GenBank Accession Codes, or Context. The latter option allows a search through all fields, performed by typing a search string (the wildcard * is allowed) into the Search Input Window and pressing Enter. Entries that satisfy search criteria are displayed in a Selection List. To display an entry from the list in the General View panel, double-click on an entry or select it and press the Enter key. The Forward and Back buttons display the next/previous thousand entries in selection list. Pressing the Locus Info button opens a separate window with detailed information on a given locus.

InfoGene exon-intron gene structures can be visualized by Dbscan program (http://www.softberry.com/scan.html), which searches for conserved regions in two sequences. This tool is useful to compare the localization of conservative regions with the localization of corresponding exon sequences or gene regulatory signals. In figure 9.11, we present the results of searching similar regions for mouse mRNA in a database of known human genes. We can see that all exons shown by red boxes (in the second window) have corresponding conserved regions (in the first window). In this way we can see exon boundaries in the mouse RNA. If we use predicted gene database and a given mRNA sequence, we can verify the corresponding predicted exons using this tool.

9.7.5 Predicted Genes in the *Drosophila* Genome

The Predicted Genes division includes an annotation of a draft of the *Drosophila* genome and a draft human genome sequence. The nucleotide sequence of nearly all euchromatic portion of the *Drosophila* genome (~120 MB) has been determined (Adams et al. 2000). These sequences were annotated by predicting genes with the Fgenes program and checking exon similarity with PfamA domains (Bateman et al. 2000). The results of this analysis are shown in table 9.13. In this table, in addition to computer predicted genes, also shows the results of removing (filtering out) most unreliable genes. Two criteria were used: (1) remove genes with the total length of protein coding region less than 30 amino acids; and (2) remove genes with total score of exons < 15. Such filtering proved useful to improve the accuracy of gene predic-
Figure 9.11
Dbscan visualization of results of searching similar regions for a mouse mRNA in a database of known human genes. We can see that all exons shown by red boxes (in the second window) have corresponding conserved regions (in the first window). By this we can detect exon boundaries in the mouse RNA. In third window we can display alignment by marking some conserved region in the first window.

Table 9.13
Summary of predicted genes and proteins in *Drosophila* genome sequences

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>2L</th>
<th>2R</th>
<th>3L</th>
<th>3R</th>
<th>4</th>
<th>Y</th>
<th>Unknown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (MB)</td>
<td>22.2</td>
<td>23.0</td>
<td>21.4</td>
<td>24.1</td>
<td>28.3</td>
<td>1.2</td>
<td>0.02</td>
<td>4.6</td>
<td>124.8</td>
</tr>
<tr>
<td>All genes</td>
<td>4071</td>
<td>4610</td>
<td>4573</td>
<td>4851</td>
<td>5954</td>
<td>133</td>
<td>1</td>
<td>691</td>
<td>24884</td>
</tr>
<tr>
<td>filtered</td>
<td>3349</td>
<td>3768</td>
<td>3915</td>
<td>4017</td>
<td>4962</td>
<td>105</td>
<td>1</td>
<td>504</td>
<td>20622</td>
</tr>
<tr>
<td>All exons</td>
<td>13036</td>
<td>15215</td>
<td>16310</td>
<td>16047</td>
<td>20382</td>
<td>679</td>
<td>10</td>
<td>1804</td>
<td>83463</td>
</tr>
<tr>
<td>filtered</td>
<td>11767</td>
<td>13713</td>
<td>15138</td>
<td>14561</td>
<td>18654</td>
<td>625</td>
<td>10</td>
<td>1467</td>
<td>75935</td>
</tr>
<tr>
<td>Exons-PfamA</td>
<td>1932</td>
<td>2148</td>
<td>2348</td>
<td>2130</td>
<td>2919</td>
<td>109</td>
<td>0</td>
<td>159</td>
<td>11745</td>
</tr>
<tr>
<td>filtered</td>
<td>1925</td>
<td>2141</td>
<td>2341</td>
<td>2126</td>
<td>2916</td>
<td>105</td>
<td>0</td>
<td>147</td>
<td>11701</td>
</tr>
<tr>
<td>United Pfam</td>
<td>1138</td>
<td>1193</td>
<td>1287</td>
<td>1216</td>
<td>1654</td>
<td>58</td>
<td>0</td>
<td>76</td>
<td>6622</td>
</tr>
<tr>
<td>Pfam types</td>
<td>431</td>
<td>475</td>
<td>499</td>
<td>460</td>
<td>546</td>
<td>43</td>
<td>0</td>
<td>40</td>
<td>1017</td>
</tr>
</tbody>
</table>
Figure 9.12
InfoGene viewer representation of Fgenes annotation of chromosome 4 of *Drosophila melanogaster*. Genes marked in the upper panel are presented in the lower panel. Coding exons are marked by red and introns by dark color. The triangles show the starts of transcription and the poly-A signals. Underlined red genes have similarity with Pfam domains. Pointing with the mouse to the first exon, we can see in down information line the similarity with Src domain; the exon in reverse chain (marked by 3) has a similarity with EGF-like domain.

We should note that 20,622 genes include some pseudogenes and genes of mobile elements. The sequences of exons and gene annotation data can be copied from http://www.softberry.com/inf/dro_ann.html for using them locally or to create microarray oligos.

The predicted genes and proteins for each human chromosome can be seen in figure 9.12, and used for further investigation at http://www.softberry.com/inf/infodb.html.

9.7.6 Predicted Genes in the Human Genome

The nucleotide sequence of nearly 90 percent of the human genome (3 GB) has been determined by an international sequencing effort. Assembly of the current draft of the human genome was done by Prof. Haussler’s Human Genome Project Team at UC Santa Cruz. Half of this sequence is occupied by repeat sequences and undefined
Table 9.14

<table>
<thead>
<tr>
<th>Number</th>
<th>PfamA short name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>467</td>
<td>Pkinase</td>
<td>Eukaryotic protein kinase domain</td>
</tr>
<tr>
<td>372</td>
<td>7tm_I</td>
<td>7 transmembrane receptor (rhodopsin family)</td>
</tr>
<tr>
<td>308</td>
<td>Myc_N_term</td>
<td>Myc amino-terminal region</td>
</tr>
<tr>
<td>256</td>
<td>Topoisomerase_I</td>
<td>Eukaryotic DNA topoisomerase I</td>
</tr>
<tr>
<td>224</td>
<td>Ig</td>
<td>Immunoglobulin domain</td>
</tr>
<tr>
<td>183</td>
<td>Rrm</td>
<td>RNA recognition motif</td>
</tr>
<tr>
<td>182</td>
<td>PH</td>
<td>PH domain</td>
</tr>
<tr>
<td>180</td>
<td>Myosin_tail</td>
<td>Myosin tail</td>
</tr>
<tr>
<td>166</td>
<td>EGF</td>
<td>EGF-like domain</td>
</tr>
<tr>
<td>159</td>
<td>Filament</td>
<td>Intermediate filament proteins</td>
</tr>
<tr>
<td>154</td>
<td>Syndecan</td>
<td>Syndecan domain</td>
</tr>
<tr>
<td>143</td>
<td>Ras</td>
<td>Ras family</td>
</tr>
</tbody>
</table>

Domain of the same type localized in neighboring exons were counted only once.

nucleotides inserted during assembling. The Fgenesh program was used on this sequence (with masked repeats) to predict exons and assemble predicted genes. Annotation of similarity of each exon with the PfamA protein domain database (Bateman et al. 2000) was produced by the Blast program (Altschul et al. 1997). A total of 49,171 genes and 282,378 coding exons were predicted. On average, one gene was found per about 68,623 bp, and one exon per 11,949 bp. Complete summary of this analysis including the gene and exon numbers in different chromosomes, is presented at http://www.softberry.com/inf/humd_an.html and can be viewed in the InfoGene database. Sequences of predicted exons and gene annotation data can also be copied from this site. One thousand one hundred and fifty-four types of PfamA different domains were found in the predicted proteins. The top part of the domain list is presented in table 9.14.

9.8 Using Expression Data for Characterization and Verification of Predicted Genes

Large-scale functional analysis of predicted, as well as known, genes might be done using expression micro array technology, which gives us the possibility of presenting all human genes on one or several Affymetrix type GeneChips. Traditionally genes are presented on the chips by unique oligonucleotides close to the 3’-end of the mRNA, but there are a lot of predicted new genes that have no known corresponding EST sequences. However, the expression of such genes could be studied using predicted exon sequences. We can present all predicted human exons (about 300,000) on
a few chips and use expression profiling across many tissues to verify the predicted exons, observing if they are expressed in some of them.

Moreover, with this approach, we can verify the structure of genes (identify a subset of predicted exons that really belong to the same gene) based on the similar expression behavior of exons from the same gene in a set of tested tissues. Exons wrongly included in a predicted gene will have different expression patterns, and exons wrongly excluded by prediction will have similar expression patterns (figure 9.13). It is interesting that such gene verification on a large scale can be done in parallel with identification of disease (tissue) specific drug target candidates. The recent chip designed by EOS Biotechnology included all predicted by Fgenesh and Genescan exons from chromosome 22, as well as predicted exons from human genomic sequences of phase 2 and 3. It was found that the predicted exon sequences present a good alternative to EST sequences, which opens a possibility of working with predicted genes on a large scale.

An example of expression behavior of three exons of the myoglobin gene in different tissues is presented in figure 9.14 (expression data were received in EOS Biotechnology Inc.). Tissue-specific expression of this gene is clearly seen with the major peaks located in skeletal muscle, heart, and diaphragm tissues. The level of expression in these tissues is 10–100 times higher than the level of signals for other tissues, as well as the average level of expression signal for randomly chosen exons. We can observe that for specific tissues, all three exons demonstrate such a high level (with correlation coefficient 0.99; for random exons it is about 0.06). These exons were predicted correctly by the Fgenesh program and were used for selection of oligonucleotide probabilities. From this result we can conclude that the predicted exons can be used as a gene representatives. An additional application of expression data is the
Expression of 3 Myoglobin exons from Chr 2

Figure 9.14
Coordinative expression of three exons of human myoglobin gene from chromosome 22 (exons were predicted by Fgenesh program and used to design EOS Biotechnology Human genome chip). The high level of expression is observed only in several specific tissues.

Figure 9.15
Exon representation can be used to characterize alternatively spliced variants of genes. Oligonucleotides selected in 3'-end of mRNA/EST sequences will not be selective for different gene variants. We will observe the sum of two genes' signal and we can miss cancer specificity of three exons' gene structure. Using oligonucleotides derived from exon sequences, we can detect different expressions of these two forms.
Table 9.15  
Web servers for eukaryotic gene and functional signal prediction

<table>
<thead>
<tr>
<th>Program/task</th>
<th>WWW address</th>
</tr>
</thead>
</table>
| FgeneHMM-based gene prediction (human, Drosophila, dicots, monocots, C. elegans, S. pombe) | [http://genomic.sanger.ac.uk/gf/gf.shtml](http://genomic.sanger.ac.uk/gf/gf.shtml)  
[http://genes.mit.edu/GENSCAN.html](http://genes.mit.edu/GENSCAN.html) |
| Fgenes Discriminative gene prediction (human) | [http://genomic.sanger.ac.uk/gf/gf.shtml](http://genomic.sanger.ac.uk/gf/gf.shtml)  
| FgenesM Prediction of alternative gene structures (human) | [http://genomic.sanger.ac.uk/gf/gf.shtml](http://genomic.sanger.ac.uk/gf/gf.shtml)  
| Fgenes-2 gene prediction with the help of similar protein/EST | [http://genomic.sanger.ac.uk/gf/gf.shtml](http://genomic.sanger.ac.uk/gf/gf.shtml)  
[http://genomic.sanger.ac.uk/gf/gf.shtml](http://genomic.sanger.ac.uk/gf/gf.shtml)  
[http://argen.cshl.org/genefinder/]  
[http://www.cbs.dtu.dk/services/Promoter/](http://www.cbs.dtu.dk/services/Promoter/) |
| Mad internal exon prediction (human, mouse, Arabidopsis) | [http://searchlauncher.bcm.tmc.edu:9331/seq-search/gene-search.html](http://searchlauncher.bcm.tmc.edu:9331/seq-search/gene-search.html)  
[http://genomic.sanger.ac.uk/gf/gf.shtml](http://genomic.sanger.ac.uk/gf/gf.shtml)  
[http://argen.cshl.org/genefinder/]  
[http://www.cbs.dtu.dk/services/Promoter/](http://www.cbs.dtu.dk/services/Promoter/)  
[http://argen.cshl.org/genefinder/]  
[http://www.cbs.dtu.dk/services/Promoter/](http://www.cbs.dtu.dk/services/Promoter/)  
[http://genomic.sanger.ac.uk/gf/gf.shtml](http://genomic.sanger.ac.uk/gf/gf.shtml)  
| Promoter 2.0 promoter prediction | [http://www.cbs.dtu.dk/services/Promoter/](http://www.cbs.dtu.dk/services/Promoter/)  
| SpiSite/splice site prediction (human, Drosophila, plants) | [http://genomic.sanger.ac.uk/gf/gf.shtml](http://genomic.sanger.ac.uk/gf/gf.shtml)  
functional analysis and identification of alternatively spliced genes (exons), when in particular tissues some exons (or their parts) have very different expression intensities compared to the other exons from the same gene. Moreover, sometimes 3'-EST generated probabilities cannot be used for the identification of disease-specific gene variants in contrast with the using exon representation of a gene (figure 9.15).

9.9 Internet Resources for Gene Finding and Functional Site Prediction

Prediction of genes, ORF, promoter, and splice sites finding by the methods described above is available on the World Wide Web. Table 9.15 presents just a few useful programs. It does not provide a comprehensive list.

Acknowledgements

I am grateful to Asaf Salamov and Igor Seledtsov for their collaboration in the development of gene-finding and other algorithms discussed here.

References


