4

Statistical Approaches in Eukaryotic Gene Prediction

V. Solovyev

Department of Computer Science, University of London, Surrey, UK

Finding genes in genomic DNA is a foremost problem of molecular biology. With the ongoing genome sequencing projects producing large quantities of sequence data, computational gene prediction is the major instrument for the identification of new genes. Usually, gene-finding programs accurately predict most coding exons in analyzed sequences, while producing a complete set of exact gene structures in any genome is still unsolved and difficult task, complicated by the large amount of gene variants generated by alternative splicing, alternation promoters and alternative polyadenylation sites. Nevertheless using gene prediction, the scientific community is now able to start experimental work with the majority of genes in dozens of sequenced genomes. Therefore, computational methods of gene identification have attracted significant attention of the genomics and bioinformatics communities. This chapter presents a comprehensive description of advanced probabilistic and discriminative gene-prediction approaches such as Hidden-Markov Models and pattern-based algorithms. We have described the structure of functional signals and significant gene features incorporated into the programs to recognize protein-coding genes. We have presented comparative performance data for a variety of gene structure identification programs and discussed some experiences in annotation of sequences from genome sequencing projects. A complex approach for finding promoters and pseudogenes have been considered as well as evaluation of their accuracy in annotation of human genome sequences. Finally, we described structural features and expression of miRNA genes and some computational methods for miRNA gene identification in genomic sequences as well as computational methods of finding miRNA targets.

4.1 STRUCTURAL ORGANIZATION AND EXPRESSION OF EUKARYOTIC GENES

The gene is the unit of inheritance encoded by a segment of nucleic sequence that carries the information representing a particular polypeptide or RNA molecule. A two-stage process comprising transcription and translation makes use of this information.

Handbook of Statistical Genetics, Third Edition. Edited by D. J. Balding, M. Bishop and C. Cannings.
© 2007 John Wiley & Sons, Ltd. ISBN: 978-0-470-05830-5

Transcription (or pre-mRNA synthesis on a 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. Posttranscriptional processing of mRNA precursors includes capping, 3′-polyadenylation and splicing. The processing events of mRNA capping and polyA addition take place before pre-mRNA splicing finally produces the mature mRNA. The mature mRNA includes sequences that correspond exactly to the protein product according to the rules of the genetic codes, called *exons*. The genomic gene sequence often includes noncoding regions called *introns* that are removed from the primary transcript during RNA splicing. Eukaryotic pre-mRNA is processed in the nucleus and then transported to the cytoplasm for translation (protein synthesis). 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 the ribosome binds to the 5′-end of mRNA and then migrates to the special sequence on mRNA (prior 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 the polypeptide chain, until it reaches a stop codon sequence (TAA, TGA or TAG), where release of polypeptide and dissociation of the ribosome from the mRNA take place. Many proteins undergo posttranslational processing (i.e. covalent modifications such as proteolytic cleavage, attachment of carbohydrates and phosphates) before they become functional. The expression stages and the structural organization of a typical eukaryotic protein-coding gene including associated regulatory regions is shown in Figure 4.1. Figure 4.2 illustrates how one DNA sequence may code for multiple proteins due to alternative promoters or terminators and alternative splicing. These processes significantly complicate *ab initio* computational gene finding.

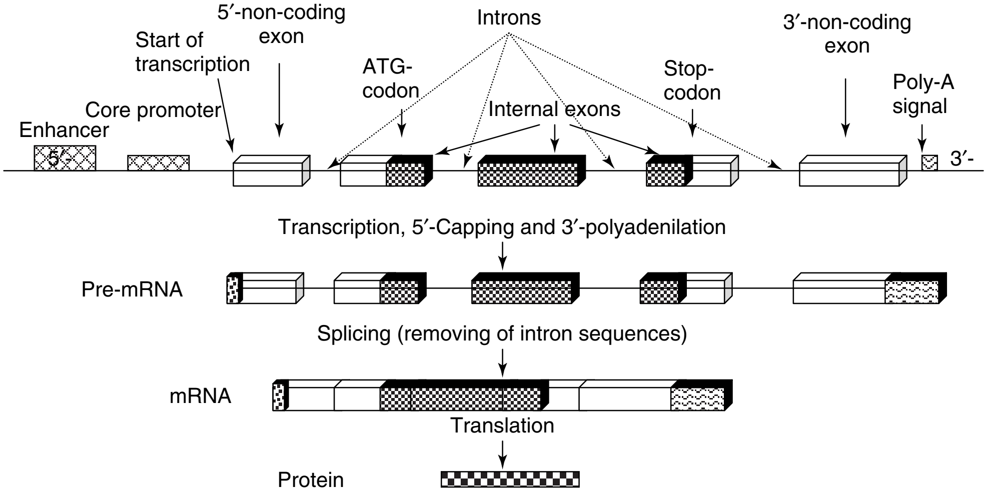


Figure 4.1 Expression stages and structural organization of typical eukaryotic protein-coding gene including its regulatory regions.

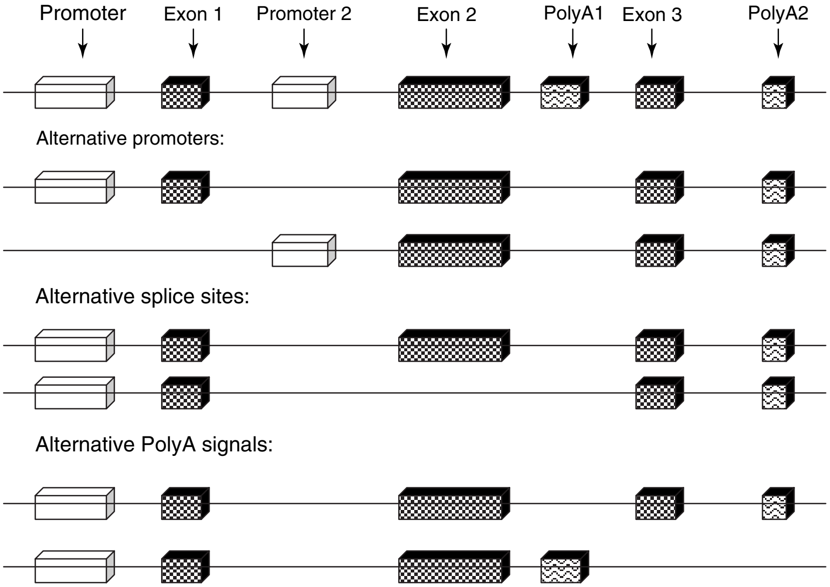


Figure 4.2 Alternative gene products coded by the same DNA region.

Information describing structural gene characteristics is accumulated in the GenBank (Benson *et al.*, 1999) and the EMBL (Kanz *et al.*, 2005) nucleotide sequence databases. However, these databases mostly contain annotations of genome fragments. Therefore, one gene can be described in dozens of entries containing partially sequenced gene regions or alternative splicing forms of its mRNA. The value of public availability of predicted genes was recognized during human genome sequencing by creating the InfoGene database (Solovyev and Salamov, 1999), which contained descriptions of known and predicted genes and their basic functional signals. Later, a few big research groups developed powerful specialized WEB accessible recourses, where various annotations of different genomes are stored and can be interactively analyzed: University of California Santa Cruz (USSC) Genome Browser (Kent *et al.*, 2002) and Ensembl genome database (Hubbard *et al.*, 2002). Some big sequencing centers such as The Institute for Genomic Research (TIGR) and Joint Genomic Institute (JGI) produce and present annotation of specific genomes at their Web servers.

Table 4.1 shows the major structural characteristics of human genes deposited to GenBank (Release 116).

41 % of sequenced human DNA consists of different kinds of repeats. Only ~3 % of the genome sequence contains protein-coding exon sequences. Characteristics of genes in major model organisms such as mouse, *Drosophila melanogaster, Caenorhabditis elegans, S. cerevisiae* and *Arabidopsis thaliana* are presented in Table 4.2.

In general, there is no big difference in the size of protein-coding mRNAs in different types of organisms, but the gene sizes are often higher in vertebrates and especially in primates. Human coding exons are significantly shorter than the sizes of genes. The average size of the exons is about 190 bp that is close to the DNA length associated with the nucleosome particle. There are many exons as short as a few bases. For example, the human pleiotrophin gene (*HUMPLEIOT*) includes a 1-bp exon and one of the alternative

Table 4.1 Structural characteristics of human genes.*

Gene features	Numbers from Infogen
CDS/partially sequenced CDS	48 088/26 584
CDS length (minimal, maximal, average)	15, 80 781, 1482
Exons/partially sequenced exons	72 488/19 392
Genes/partially sequenced genes	18 429/14 385
Alternative splicing	12 %
Pseudogenes	8.5 %
Genes without introns	8 %
Number of exons (maximal, average)	117, 5.4
Exon length (range, average)	1–10 088, 195
Intron length (maximal, average)	185 838, 2010
Gene length (maximal, average)	401 910, 7865
Repeats in genome	41 % of total DNA
DNA occupied by coding exons	3 %

*The numbers reflect genes described in GenBank, which might deviate from the average parameters for the organism. Gene numbers were calculated for DNA loci only. Many long genes have partially sequenced introns, therefore average sizes of genes and introns are actually bigger. The average numbers of exons and gene lengths were calculated for completely sequenced genes only.
CDS: protein CoDing Sequences.

forms of the human folate receptor (*HSU20391*) gene contains a 3-bp exon. Coding exons can also be very short.

The human myosin-binding protein gene (*HSMYBPC3*) includes 2 exons that are 3-bp long. At the same time we observe a coding region about 90 000 bp for the titin gene (NM_003319) and an exon 8000 bp in human gene encoding microtubule-associated protein 1a (HSU38291). Very often protein-coding exons occupy a small percent of the gene size. The human fragile X mental retardation gene (*HUMFMR1S*) presents a typical example: 17 exons (40–60-bp long) occupy just 3 % of 67 000 bp gene sequence.

The structural characteristics of eukaryotic genes, discussed above, create difficult problems in computational gene identification. Low density of coding regions (3 % in human DNA) will generate a lot of false-positive predictions in fragments of noncoding DNA. The number of these false positives might be even comparable with the true number of exons. Recognition of small exons (1–20 bp) cannot be achieved using any composition-based method that is relatively successful for identification of prokaryote coding regions. It is necessary to develop gene-prediction approaches that rely significantly on the recognition of functional signals encoded in the DNA sequence.

4.2 METHODS OF FUNCTIONAL SIGNAL RECOGNITION

In this paragraph, we will describe several approaches for gene functional signal recognition and some features of these signals used in gene identification. The simplest way to find functional sequences is based on a consensus sequence or weight matrix reflecting conservative bases of the signal. Using a consensus sequence or weight matrix we can scan a given sequence and select high scoring regions as potential functional signals.

Table 4.2 Structural characteristics of genes in eukaryotic model organisms.*

	Mus musculus	*D. melanogaster*	*C. elegans*	*S. cerevisiae*	*A.thaliana*
CDS/partial	20 340/11 192	5095/1057(13 601)	18 146/377	12 629/1040	14 590/1076
Exons/partial	14 940/7812	8661/3694 (56 673)	108 934/33 821	13 444/13 028	62 151/19 505
Genes/partial	5571/4077	2802/948 (13601)	17 336/1003	12 495/1024	12 221/616
Alternative splicing	11 %	15 % (7 %)	5.6 %	5.8 %	1.4 %
No introns genes	10 %	20 %	4 %	90 %	20 %
Number of exons	64, 4.59	26, 3.1 (4.2)	48, 6.1	3, 1.03	70, 5.1
Exon length	1–6642, 199.1	6–9462, 454 (425)	1–14 975, 22 125	1–7471, 1500.0	2–5130, 190.8
Intron length	29 382, 678.4	73650, 457 (488)	19 397, 244.0	7317, 300	118 637, 186.39
Gene size	8020, 3388	74 691, 2150	45 315, 2496	14 733, 1462	170 191, 2040

*Description of features is given in Table 4.1. For Drosophila genes, the numbers in () are taken from computer and manual annotation of Drosophila genome (Adams *et al.*, 2000).

To select only significant matches, a statistical method for estimating the significance of similarity between the consensus of functional signal and a sequence fragment was developed (Shahmuradov *et al.*, 1986; Solovyev and Kolchanov, 1994). Computation of this statistic was implemented in the program NSITE (`http://www.softberry.com/berry.phtml?topic=nsite&group=programs&subgroup=promoter`) (Shahmuradov and Solovyev, 1999) that identifies nonrandom similarity between fragments of a given sequence and consensuses of regulatory motifs from various databases such as Transcription Factor Database (TFD) (Ghosh, 2000), Transfac (Wingender *et al.*, 1996), RegsiteAnimal and RegsitePlant (Solovyev *et al.*, 2003). Here we briefly describe application of weight matrices, which usually contain more information about the structure of functional signal than consensus sequences. Procedures using weight matrices are implemented in many modern gene-prediction approaches that score potential functional signals.

4.2.1 Position-specific Measures

Weight matrices are typically used for functional signal description (Staden, 1984a; Zhang and Marr, 1993; Burge, 1997). We can consider weight matrix as a simple model based on a set of position-specific probability distributions $\{p_s^i\}$, that provide probabilities of observing a particular type of nucleotide in a particular position of functional signal sequence (S). The probability of generating a sequence $X(x_1, \ldots, x_k)$ under this model is

$$P(X/S) = \prod_{i=1}^{k} p_{x_i}^i, \tag{4.1}$$

where each position of the signal is considered to be independent. A corresponding model can also be constructed for a sequence having no functional signal (N): $\{\pi_s^i\}$. An appropriate discriminative score based on these models is the log likelihood ratio:

$$\text{LLR(X)} = \log \frac{P(X/S)}{P(X/N)}. \tag{4.2}$$

To evaluate a given sequence fragment, a score can be computed as an average sum of weights of observed nucleotides using the corresponding weight matrix $w_{(i,s)} = \{log(p_s^i/\pi_s^i)\}$:

$$\text{Score} = \text{LLR}(X) = \frac{1}{k}\sum_{i=1}^{k} w(i, x_i). \tag{4.3}$$

Different weight functions have been used to score the sequence, for example, weights can be obtained by some optimization procedures such as a perceptron or neural network (Stormo *et al.*, 1982). Different position-specific probability distributions $\{p_s^i\}$ can also be considered.

A generalization of the weight matrix uses position-specific probability distributions $\{p_s^i\}$ of oligonucleotides (instead of single nucleotides). Another approach is to exploit Markov chain models, where the probability of generating a particular nucleotide x_i of the signal sequence depends on the $k_0 - 1$ previous bases (i.e. it depends on an oligonucleotide ($k_0 - 1$ base long) ending at the position $i - 1$). Then the probability of generating the signal sequence X is:

$$P(X/S) = p_0 \prod_{i=k_0}^{k} p^{i-1,i}_{s_{i-1},x_i} : \tag{4.4}$$

where $p^{i-1,i}_{s_{i-1},x_i}$ is the conditional probability of generating nucleotide x_i in position i given oligonucleotide s_{i-1} ended at position $i-1$, p_0 is the probability of generating oligonucleotide $x_{1\ldots}x_{k0-1.}$ For example, a simple weight matrix represents independent mononucleotide model (or 0-order Markov chain), where $k_0 = 1$, $p_0 = 1$ and $p^{i-1,i}_{x_{i-1},x_i} = p^i_{x_i}$. When we use dinucleotides (1st order Markov chain) $k_0 = 2$, $p_0 = p^1_{x_1}$, and $p^{i-1,i}_{x_{i-1},x_i}$ 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 to the frequency of the oligonucleotide $k_0 - 1$ bases long ending at position $i-1$ in a set of aligned sequences of some functional signal.

$$p^{i-1,i}_{s_{i-1},x_i} = \frac{\mathrm{f}(s_{i-1}, x_i)}{\mathrm{f}(s_{i-1})}.$$

Using the same procedure we can construct a model for nonsite sequences for computing $P(X/N)$, where often 0-order Markov chain with genomic base frequencies (or even equal frequencies (0.25)) is used.

A log likelihood ratio (3) with Markov chains was applied to select CpG island regions (Durbin *et al.*, 1998). The same approach was used in a description of promoters, splice sites and start and stop of translation in gene-finding programs such as Genscan (Burge and Karlin, 1997), Fgenesh (Find GENES Hmm) (Salamov and Solovyev, 2000) and GeneFinder (Green and Hillier, 1998).

A useful discriminative measure taking into account *a priori* knowledge is based on the computation of Bayesian probabilities as components of position-specific distributions $\{p^i_s\}$:

$$P(S/o^i_s) = \frac{P(o^i_s/S)P(S)}{\big(P(o^i_s/S)P(S) + P(o^i_s/N)P(N)\big)}, \tag{4.5}$$

where $P(o^i_s/S)$ and $P(o^i_s/N)$ can be estimated as position-specific frequencies of oligonucleotides o^i_s in the set of aligned sites and nonsites; $P(s)$ and $P(N)$ are the *a priori* probabilities of site and nonsite sequences, respectively. S is a type of the oligonucleotide starting (or ending) in ith position (Solovyev and Lawrence, 1993a). The probability that a sequence X belongs to a signal, if one assumes independence of oligonucleotides in different positions, is:

$$P(S/X) = \prod_{i=1}^{k} P(S/o^i_m).$$

Another empirical discriminator called *preference* uses the average positional probability of belonging to a signal:

$$Pr(S/X) = \frac{1}{k}\sum_{i=1}^{k} P(S/o^i_m). \tag{4.6}$$

This measure was used in constructing discriminant functions for the Fgenes gene-finding program (Solovyev *et al.*, 1995). It can be more stable than the previous measure on short sequences and has simple interpretation: if the $Pr > 0.5$, then our sequence is more likely to belong to a signal than to a nonsignal sequence.

4.2.2 Content-specific Measures

Some functional signal sequences have a distinctive general oligonucleotide composition. For example, many eukaryotic promoters are found in GC-rich chromosome fragments. We can characterize these regions by applying similar methods to the above scoring functions, but using probability distributions and their estimates by oligonucleotide frequencies computed on the whole sequence of the functional signal. For example, the Markov-chain-based probability of generating the signal sequence X will be:

$$P(X/S) = p_0 \prod_{i=k_0}^{k} p_{s_{i-1}}, x_i. \tag{4.7}$$

4.2.3 Frame-specific Measures

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 (or 6 if the complementary chain is also considered). The nucleotides are distributed unevenly relative to the positions within codons. Therefore the probability of observing a specific oligonucleotide in coding sequence depends on its position relative to the coding frame (three possible variants) as well as on neighboring nucleotides (Shepherd, 1981; Borodovskii *et al.*, 1986; Borodovsky and McIninch, 1993). Asymmetry in base composition between codon positions arises because of uneven usage of amino acids and synonymous codons, as well as the specific nature of the genetic code (Guigo, 1999). Fickett and Tung (1992) did a comprehensive assessment of the various protein-coding measures. They estimated the quality of more than 20 measures and showed that the most powerful is 'in phase hexanucleotide composition'. In Markov chain approaches, the frame-dependent probabilities $p^f_{s_{i-1},x_i} (f = \{1,2,3\})$ are used to model coding regions. The probability of generating a protein-coding sequence X is

$$P(X/C) = p_0 \prod_{i=k_0}^{k} p^f_{s_{i-1},x_i}, \tag{4.8}$$

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

4.2.4 Performance Measures

Several measures to estimate the accuracy of a recognition function were introduced in genomic research (Fickett and Tung, 1992; Snyder and Stormo, 1993; Dong and Searls, 1994). Consider that we have S sites (positive examples) and N nonsites (negative examples). By applying the recognition function, we correctly identify Tp 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. $T_p + F_n = S$ and $T_n + F_n = N$. Sensitivity (S_n) measures the fraction of the true positive examples that are correctly predicted: $S_n = T_P/(T_P + F_n)$. Specificity (S_p) measures the fraction of the predicted sites that are correct amongst those predicted: $S_p = T_P/(T_P + F_P)$. Note that the definition of S_p used in gene-prediction research is different from the usual $S_p = T_n/(T_n + F_P)$. Only the simultaneous consideration of both S_n and S_p values makes sense when we provide some accuracy information. Using only one value of accuracy estimation means that the average accuracy of prediction of true sites and nonsites is AC $= 0.5(T_P/S + T_n/N)$. However, this measure does not take into account the possible difference in sizes of site and nonsites sets. A more correct single measure (correlation coefficient) takes the relation between correctly predictive positives and negatives as well as false positives and negatives into account (Matthews, 1975):

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

4.3 LINEAR DISCRIMINANT ANALYSIS

Different features of a functional signal may have different significance for recognition and may not be independent. Classical linear discriminant analysis provides a method to combine such features in a discriminant function. A discriminant function, when applied to a pattern, yields an output that is an estimate of the class membership of this pattern. The discriminative technique provides minimization of the error rate of classification (Afifi and Azen, 1979). Let us assume that each given sequence can be described by vector X of p characteristics $(x_1, x_2, \ldots, x_p)$, that can be measured. The linear discriminant analysis procedure finds a linear combination of the measures (called the *linear discriminant function* or *LDF*), that provides maximum discrimination between site sequences (class 1) and nonsite examples (class 2). The LDF classifies (X) into class 1 if $Z > c$ and into class 2 if $Z > c$. The vector of coefficients $(\alpha_1, \alpha_2, \ldots, \alpha_p)$ and threshold constant c are derived from the training set by maximizing the ratio of the between-class variation of z to within-class variation and are equal to (Afifi and Aizen, 1979):

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

and

$$\vec{c} = \vec{a}(\vec{m}_1 - \vec{m}_2)/2,$$

where $\vec{m}_i$ 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 covariation matrix, and n_i is the sample size of class i. On the basis of these equations, we can calculate the coefficients of LDF and threshold constant c using the values of characteristics of site and nonsite sequences from the training sets and then test the accuracy of LDF on the test set data. Significance of a given characteristic or a set of characteristics can be estimated by the generalized distance between two classes (called the *Mahalonobis distance* or D^2):

$$\vec{D}^2 = (\vec{m}_1 - \vec{m}_2)s^{-1}(\vec{m}_1 - \vec{m}_2),$$

that is computed on the basis of values of the characteristics in the training sequences of classes 1 and 2. To find sequence features a lot of possible characteristics as score of weigh matrices, distances, oligonucleotide preferences at different subregions are generated. Selection of the subset of significant characteristics (among those tested) is performed by a stepwise discriminant procedure including only those characteristics that significantly increase the Mahalonobis distance (Afifi and Aizen, 1979).

4.4 PREDICTION OF DONOR AND ACCEPTOR SPLICE JUNCTIONS

4.4.1 Splice-sites Characteristics

Splice-site patterns are mainly defined by nucleotides at the ends of introns, because deletions of large parts of intron do not affect their selection (Breathnach and Chambon, 1981; Wieringa *et al.*, 1984). A sequence of eight nucleotides is highly conserved at the boundary between an exon and an intron (donor or 5′-splice site). This is AG|GTRAGT and a sequence of 4 nucleotides, preceded by a pyrimidine rich region, is also highly conserved between an exon and an intron (acceptor or 3′-splice site): YYTTYYYYYYNC|AGG (Senapathy *et al.*, 1990). The third less-conserved sequence of about 5–8 nucleotides, and containing an adenosine residue, lies within the intron, usually between 10 and 50 nucleotides upstream of the acceptor splice site (branch site). These sequences provide specific molecular signals by which the RNA splicing machinery can select the splice sites with precision.

Two very conservative dinucleotides are observed in practically all introns. The donor site has GT just after the point where the spliceosomes cut the 5′-end of intron sequences and the acceptor site has AG just before the point where the spliceosomes cut the 3′-end of intron sequences (Breathnach *et al.*, 1978; Breathnach and Chambon, 1981).

Additionally, a rare type of splice pair AT–AC has been discovered. It is processed by related but different splicing machinery (Jackson, 1991; Hall and Padget, 1994). Introns flanked by the standard GT–AG pairs excised from pre-mRNA by the spliceosome including U1, U2, U4/U6 and U5 snRNPs (Nilsen, 1994). A novel type of spliceosome composed of snRNPs U11, U12, U4atac/U6atac and U5 (Hall and Padgett, 1996, Tarn and Steitz, 1996a; 1996b; 1997) excises AT–AC introns. For AT–AC group a different conserved positions have been noticed: |ATATCCTTT for donor site and YAC| for acceptor site (Dietrich *et al.*, 1997; Sharp and Burge, 1997; Wu and Krainer, 1997).

Burset *et al.* (2000) have done a comprehensive investigation of canonical and noncanonical splice sites. They have extracted 43 427 pairs of exon–intron boundaries and their sequences from the InfoGene (Solovyev and Salamov, 1999) database including all the annotated genes in mammalian genomic sequences. Annotation errors present a real problem in getting accurate information about eukaryotic gene functional signals from nucleotide sequence databases, such as GenBank or EMBL (Benson *et al.*, 1999; Kanz *et al.*, 2005). The authors generated a spliced construct for every splice pair combining 40 nt. of the left exon and 40 nt. of the right exon producing the same sequence as the splicing machinery generated by removing intron region (Figure 4.3). To verify the extracted splice sites, the alignments of splice constructs with known mammalian expressed sequence tags (ESTs (Boguski *et al.*, 1993) were used. For 43 427 pairs of donor and acceptor splice sites (splice pairs), 1215 were annotated as nonstandard donor sites (2.80 %) and 1027 were annotated as nonstandard acceptor sites (2.36 %). 41 767 splice pairs (96.18 %) contained the standard splice-site pair GT–AG. As a result of the analysis, from 1660 noncanonical pairs, 441 were supported by ESTs (27.35 %) and just 292 (18 %) were supported by ESTs after removing potential annotation errors and cases with ambiguities in the position of the splice junction (Table 4.3).

It is interesting to note that the EST-supported rate is clearly higher for canonical splice pairs. There were 22 374 out of 42 212 canonical pairs supported by ESTs (53.63 %) and just 27.3 % of noncanonical pairs. About a half (43.15 %) of all noncanonical splice pairs belongs to the GC–AG group (126). The next biggest in size of the noncanonical group GG–AG contains significantly less cases (12). There were many other groups in the same size range, including those processed by the special splicing machinery, the AT–AC group. Weight matrices for GT–AG and GC–AG pairs are presented in Table 4.4 and a consensus sequence for AT–AC pair in Figure 4.4.

Most other noncanonical splice pairs have a canonical conserved dinucleotide shifted by one base from the annotated splice junction. For example, for 12 EST-supported GG–AG pairs, 10 have a shifted canonical donor splice site with the GT dinucleotide, 1 major noncanonical site has the GC dinucleotide, and one GA–AG case (Figure 4.5).

One is prompted to explain the observations with the shifted canonical dinucleotides by an annotation error of inserting/deleting one nucleotide that is actually absent/present in real genomic sequence. This hypothesis was tested by comparing human gene sequences deposited to GenBank earlier with the sequences of the same region obtained in high throughput genome sequencing projects. Several examples of clear annotation and sequencing errors identified by the comparison are presented in Figure 4.6. We found 88

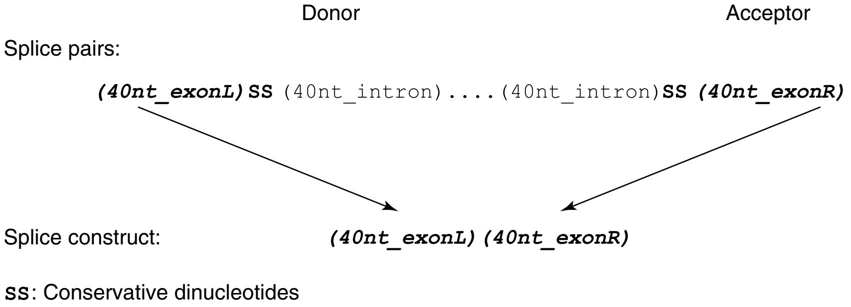

Donor
Acceptor
Splice pairs:
(40nt_exonL)SS (40nt_intron)....(40nt_intron)SS (40nt_exonR)
Splice construct:
(40nt_exonL)(40nt_exonR)
SS: Conservative dinucleotides

Figure 4.3 Structure of spliced constructs. Two sequence regions of a splice pair (marked as Donor and Acceptor) with the corresponding splice-site dinucleotides surrounded by 40 nt. of gene sequence at each side. Joining exon part of donor (40nt_ exonL) and exon part of acceptor (40nt_exonR) produce a sequence of splice construct to be verified by ESTs.

Table 4.3 Canonical and noncanonical splice sites in mammalian genomes.

(a) Annotated splice pairs.

Splice sites	Donors	Acceptors	Pairs
Canonical	42160 (97.28 %)	42344 (97.71 %)	41722 (96.27 %)
Noncanonical	1177 (2.72 %)	993 (2.29 %)	1615 (3.73 %)
EST - supported canonical	22437 (98.34 %)	22568 (98.92 %)	22374 (98.07 %)
EST - supported noncanonical	378 (1.66 %)	247 (1.08 %)	441 (1.93 %)
EST - supported canonical after correction	22306 (**98.94 %**)	22441 (**99.54 %**)	22253 (**98.70 %**)
EST - supported noncanonical after correction	239 (**1.06 %**)	104 (**0.46 %**)	292 (**1.30 %**)

(b) Generalization of analysis of human noncanonical splice pairs.

GT–AG	22310	99.20 %
GC–AG	140	0.62 %
AT–AC	18	0.08 %
Other Noncanonical	7	0.03 %
Errors	14	0.06 %
TOTAL	**22489**	**100 %**

AT-AC group:

```
AC002397     TGCCAAGATG| atatccttgtgt    ctgtctgctcac |CTTGGAGAAG
AC004976     GAAAGAACCC| atatcctttctg    actacttcatac |AAAACAGTCA
AF1 36179    TATGGTAGAG| atatcctttact    actgtttcggac |ATTGACCAAA
AL021578     ACGCTGAACC| atatcctttggg    ttaaccgctcac |TGGCCCAGCT
L10295       ATTGGTGAAG| atatcctttttag   aatcattactac |ATGTGAATCC
U39892       AGATTAGAGA| atatcctttctt    aactgccagcac |ATTTTGTCAG
U47924       TGCCAAGATG| atatccttctgc    aaccctcctcac |CTTGGAGAAG
U53004       GGAAGTGGTC| atatccttcctg    aactctgcacac |GAAGCTCACG
```

Consensus of donor site:

$G_{50}\,|\,A_{100}\,T_{100}\,A_{100}\,T_{100}\,C_{100}\,C_{100}\,T_{100}\,T_{100}\,T_{62}$

Consensus of acceptor site:

$C_{62}T_{75}T_{37}C_{75}T_{37}C_{62}A_{100}C_{100}\,|\,A_{50}T_{62}$

Figure 4.4 Consensus sequences for the AT–AC pair of the alternative splicing machinery.

examples of independent gene sequencing with sequences overlapping splice junctions. All human EST-supported GC–AG cases having HTS matches were supported by them (39 cases). 31 errors damaging the standard splice pairs were found. 7 cases had one or both intronic GenBank sequences completely unsupported by HTS, 8 cases had intronic GenBank sequences supported, but there was a gap between exonic and intronic parts and finally 16 cases had small errors as some insertions, deletions or substitutions. 5 AT–AC pairs (3 pairs were correctly annotated in original noncanonical set and 2 were recovered from errors) were identified. In additition, 2 cases were annotated as introns, but in HTS the exonic parts were continuous (accession numbers: U70997 and M13300). 7 cases of HTS were themselves GenBank sequences and for this reason they were excluded from the analysis.

Table 4.4 Characteristics of major splice-pair groups.

GT–AG group. Number of supported cases: 22 268

Donor frequency matrix

A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

Acceptor frequency matrix

A	9.0	8.4	7.5	6.8	7.6	8.0	9.7	9.2	7.6	7.8	23.7	4.2	100	0.0	23.9
C	31.0	31.0	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	11.0	11.3	11.3	8.5	6.6	6.4	21.2	0.3	0.0	100	52.0
U	42.3	44.0	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

GC–AG group. Number of supported cases: 126

Donor frequency matrix

A	40.5	88.9	1.6	0.0	0.0	87.3	84.1	1.6	7.9
C	42.1	0.8	0.8	0.0	100	0.0	3.2	0.8	11.9
G	15.9	1.6	97.6	100	0.0	12.7	6.3	96.8	9.5
U	1.6	8.7	0.0	0.0	0.0	0.0	6.3	0.8	70.6

Acceptor frequency matrix

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	9.5	10.3	15.1	12.7	8.7	9.5	16.7	4.8	2.4	6.3	13.5	0.0	0.0	100	62.7
U	38.9	41.3	58.7	55.6	42.1	40.5	30.9	37.3	44.4	47.6	27.0	25.4	0.0	0.0	7.9

	Donor	Acceptor	Donor + 1 shift
U07083	**AAGGG|** ggtaagg	ctttaag **|GGTGT**	GG ⇒ GT
X98208	**CGGCA|** ggtcaga	aatgcag **|GTGTA**	GG ⇒ GT
L43831	**CAAAG|** ggtactg	tctgcag **|CTTTG**	GG ⇒ GT
U37431	**AAACA|** ggtcagt	gccccag **|GGGAA**	GG ⇒ GT
U02978	**AGGCC|** ggtgagt	gggccag **|GGGTC**	GG ⇒ GT
AJ000060	**AGTAT|** ggtaagg	tttccag **|GGAGA**	GG ⇒ GT
U12599	**GCTGG|** ggtaagt	tccccag **|TCATA**	GG ⇒ GT
U01247	**TCACA|** ggtatgc	attctag **|GAGAA**	GG ⇒ GT
U28721	**CGCAG|** ggcaagg	ctaacag **|GTCTA**	GG ⇒ **GC**
M20214	**AACAG|** ggaaggc	acgctag **|GGAAA**	GG ⇒ **GA**
M62601	**TGCAG|** ggtatac	cctttag **|ACAAT**	GG ⇒ GT
U66878	**TAGTG|** ggtgagt	ccttcag **|GAGTG**	GG ⇒ GT

Figure 4.5 Shifted splice sites. Example for GG–AG verified splice sites (12 cases). In donor, exactly after the cut point was always found a GG pair. To obtain which splicing pair are characteristic to this donor we should produce a shift of 1 nucleotide downstream. After this we reclassify sites as 10 GT–AG canonical splice sites, 1 GC–AG site and 1 apparently strange GA–AG site.

By generalizing these results we conclude that the overwhelming majority of splice sites contain the conserved dinucleotides GT–AG (99.2 %). The other major group includes GC–AG pairs (0.62 %), the alternative splicing mechanism group AC–AT (about

```
Sequences of homeodomain protein, HOXA9EC (AF010258)
                      Donor                       Acceptor
Genbank:          CGATCCCAAT| aa-tgtctcct         cccgcagaat  |AACCCAGCAG
High throughput:  CGATCCCA| gtaagtgtctcct         cccgcag |AT-AACCCAGCAG

Sequences of  poly(A) binding protein II, PABP2 (AF026029)
                      Donor                       Acceptor
Genbank:          TCCAGGCAAT| gctgagtaac          tttcctgata  |GCTGGCCCGG
High throughput:  TCCAGGCAATG|  gtgagtaac         tttcctgatag  |CTGGCCCGG
```

Figure 4.6 Errors found by comparing GenBank and the human high-throughput sequences for several annotated noncanonical splice sites.

0.08 %) and a very small number of other noncanonical splice sites (about 0.03 %) (Table 4.3.d). Therefore, gene-finding approaches using only standard GT–AG splice sites can potentially predict 97 % genes correctly (if we assume 4 exons per gene, on average). Including the GC–AG splice pair will increase this level to 99 %. 22 253 verified examples of canonical splice pairs were presented in a database (SpliceDB), which is available for public use through the www (`http://www.softberry.com/berry.phtml?topic=splicedb&group=data&subgroup=spldb`) (Burset *et al.*, 2000). It also includes 1615 annotated and 292 EST-supported and shift-verified noncanonical pairs. This set can be used to investigate the reality of these sites as well as to further understand the splicing machinery.

Analysis of splice-site sequences demonstrated that their consensuses are somewhat specific for different classes of organisms (Senapathy *et al.*, 1990; Mount, 1993) and some important information is encoded by the sequences outside the short conserved regions. Scoring schemes 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 peculiarities, give an accuracy of about 80 % for the prediction splice-site positions (Nakata *et al.*, 1985; Gelfand, 1989). More accurate prediction is produced by neural network algorithms (Lapedes *et al.*, 1988; Brunak *et al.*, 1991; Farber *et al.*, 1992). The integral view on the difference of triplet composition in splice and pseudosplice sequences is shown in Figure 4.7. This figure demonstrates the various functional parts of splice sites. We can see that the only short regions around splice junctions have a great difference in triplet composition. Their consensus sequences are usually used as determinants of donor or acceptor splice-site positions. However, dissimilarity in many other regions can also be seen. For the donor site – coding region, a G-rich intron region may be distinguished. For acceptor sites – a G-rich intron region, a branch point region, a polyT/C tract and coding sequence. Splice-site prediction methods using a linear function that combines several of such features is described below (Solovyev and Lawrence, 1993a; Solovyev *et al.*, 1994).

4.4.2 Donor Splice-site Characteristics

Seven characteristics were selected for donor splice-site identification. Their values were calculated for 1375 authentic donor site and for 60 532 pseudosite sequences from the learning set. The Mahalonobis distances showing the significance of each characteristic are given in Table 4.5. The strongest characteristic for donor sites is a triplet composition in consensus region ($D^2 = 9.3$) followed by the adjacent intron region ($D^2 = 2.6$) and the

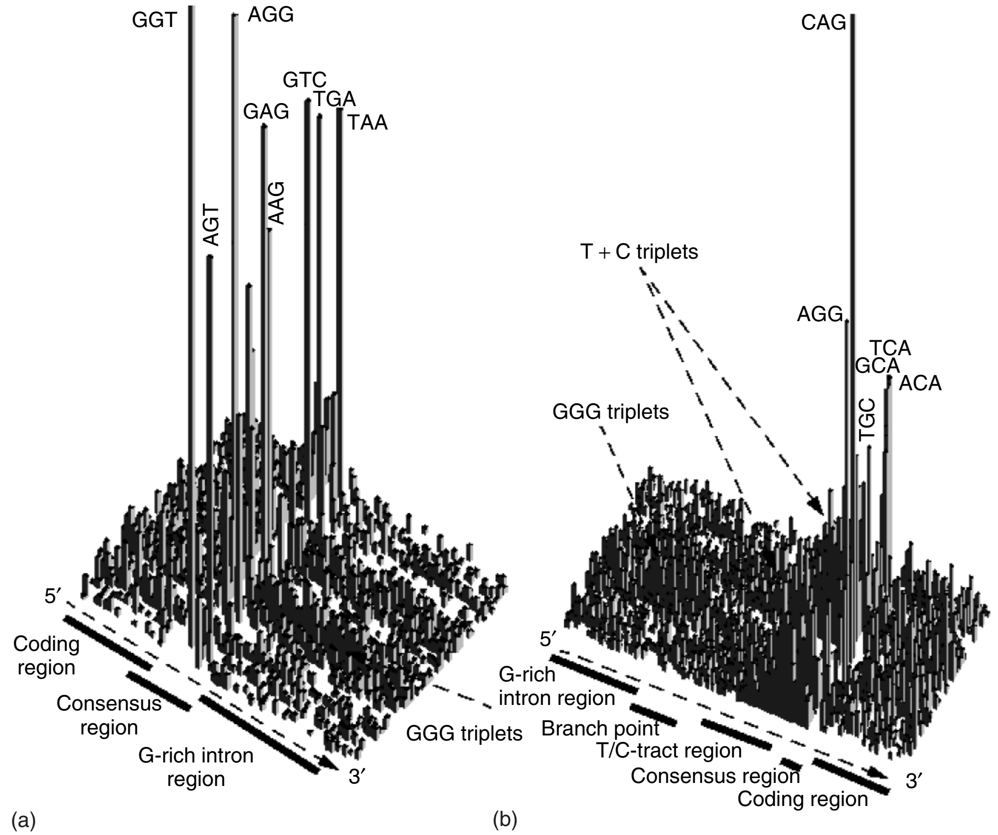


Figure 4.7 Difference of the triplet composition around donor and GT-containing non-donor sequences (a); around acceptor and AG-containing non acceptor sequences (b) in 692 human genes. Each column presents the difference of specific triplet numbers between sites and pseudosites in a specific position. For comparison the numbers were calculated for equal quantities of sites and pseudosites.

Table 4.5 Significance of various characteristics of donor splice sites.

Characteristics	1	2	3	4	5	6	7
Individual D^2	9.3	2.6	2.5	0.01	1.5	0.01	0.4
Combined D^2	9.3	11.8	13.6	14.9	15.5	16.6	16.8

1, 2, 3 are the triplet preferences (13) of consensus (−4 to +6), intron G-rich (+7 to +50) and coding regions (−30 to −5), respectively. 4 is the number of significant triplets in the consensus region. 5 and 6 are the octanucleotide preferences 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 GGG triplets in +6 to +50 intron G-rich region.

coding region ($D^2 = 2.5$). Other significant characteristics are the number of significant triplets in conserved consensus region; the number of G bases, GG doublets and GGG triplets; and the quality of the coding and intron regions.

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.softberry.com/berry.phtml?topic=spl&group=programs&subgroup=gfind`) provides the most accurate local donor site recognizer (Table 4.6) (Milanesi and Rogozin, 1998).

Although a simple weight matrix provides less accurate recognition than more sophisticated approaches, it can be easily recomputed for new organisms and is very convenient to use in probabilistic HMM-gene-prediction methods. An interesting extension of this approach was suggested on the basis of analysis of dependencies between splice-site positions (Burge and Karlin, 1997). Using a maximal dependence decomposition procedure (Burge, 1998), 5 weight matrices corresponding to different subsets of splice-site sequences were generated. The subclassification of donor signals and the matrices constructed based on 22 306 EST-supported splice sites are presented in Figure 4.8. Performance of these matrices compared with other methods was evaluated on the Burset and Guigo (1996) data set (Figure 4.9). We can observe that several weight matrices definitely provide better splice-site discrimination than just one. However, their discriminatory power is similar to that of the matrix of triplets and lower than that of the linear discriminant function described above.

4.4.3 Acceptor Splice-site Recognition

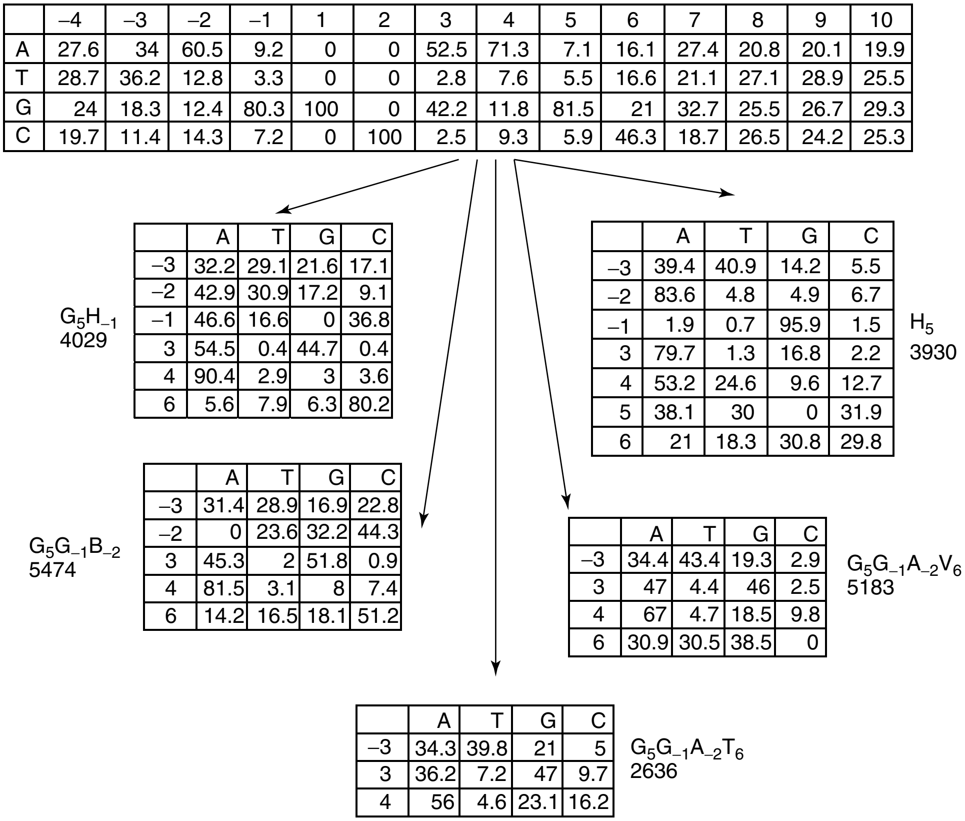
Seven characteristics were selected for acceptor splice-sites recognition. Their values were calculated for 1386 authentic acceptor site and 89 791 pseudosite sequences from the learning set. The Mahalonobis distances showing the individual significance for each characteristic are given in Table 4.7. The strongest characteristics for acceptor sites are the triplet composition in the polyT/C tract region ($D^2 = 5.1$); consensus region ($D^2 = 2.7$); adjacent coding region ($D^2 = 2.3$); and branch point region ($D^2 = 1.0$). Some significance is found using the number of T and C in the adjacent intron region ($D^2 = 2.4$) and the quality of the coding region ($D^2 = 2.6$).

Table 4.8 illustrates the performance of different methods for acceptor site recognition (Milanesi and Rogozin, 1998). The linear discriminant function described above provides the best accuracy. Also, we can observe that acceptor site recognition accuracy is lower than that for donor sites.

It was shown that the first-order Markov chain model (11) based on dinucleotide frequencies of [−20, +3] acceptor site region gives slightly better discrimination than the simple weight matrix model (Burge, 1998). Such a model was incorporated in

Table 4.6 Comparing the accuracy of local donor splice-site recognizers. The accuracy is averaged for 3 tested sets.

Method	False positives (%)	False negatives (%)	CC	Reference
Weight matrix	2.3	53	0.13	Guigo *et al.* (1992)
Consensus MAG/GURAGU	6.0	18	0.27	Mount (1982)
Five consensuses	4.2	15	0.31	Milanesi and Rogozin (1998)
Neural network	25.0	2.7	0.51	Brunak *et al.* (1991)
Discriminant analysis	10.0	3.0	0.56	Solovyev *et al.* (1994)



	−4	−3	−2	−1	1	2	3	4	5	6	7	8	9	10
A	27.6	34	60.5	9.2	0	0	52.5	71.3	7.1	16.1	27.4	20.8	20.1	19.9
T	28.7	36.2	12.8	3.3	0	0	2.8	7.6	5.5	16.6	21.1	27.1	28.9	25.5
G	24	18.3	12.4	80.3	100	0	42.2	11.8	81.5	21	32.7	25.5	26.7	29.3
C	19.7	11.4	14.3	7.2	0	100	2.5	9.3	5.9	46.3	18.7	26.5	24.2	25.3

G_5H_{-1}
4029

	A	T	G	C
−3	32.2	29.1	21.6	17.1
−2	42.9	30.9	17.2	9.1
−1	46.6	16.6	0	36.8
3	54.5	0.4	44.7	0.4
4	90.4	2.9	3	3.6
6	5.6	7.9	6.3	80.2

H_5
3930

	A	T	G	C
−3	39.4	40.9	14.2	5.5
−2	83.6	4.8	4.9	6.7
−1	1.9	0.7	95.9	1.5
3	79.7	1.3	16.8	2.2
4	53.2	24.6	9.6	12.7
5	38.1	30	0	31.9
6	21	18.3	30.8	29.8

$G_5G_{-1}B_{-2}$
5474

	A	T	G	C
−3	31.4	28.9	16.9	22.8
−2	0	23.6	32.2	44.3
3	45.3	2	51.8	0.9
4	81.5	3.1	8	7.4
6	14.2	16.5	18.1	51.2

$G_5G_{-1}A_{-2}V_6$
5183

	A	T	G	C
−3	34.4	43.4	19.3	2.9
3	47	4.4	46	2.5
4	67	4.7	18.5	9.8
6	30.9	30.5	38.5	0

$G_5G_{-1}A_{-2}T_6$
2636

	A	T	G	C
−3	34.3	39.8	21	5
3	36.2	7.2	47	9.7
4	56	4.6	23.1	16.2

Figure 4.8 Classification of donor splice sites by several weight matrices reflecting different splice-site groups (Burge and Karlin, 1997).

Genscan gene-prediction method (Burge and Karlin, 1997). Thanaraj (2000) performed a comprehensive analysis of computational splice-site identification. The HSPL program remains the best local recognizer. Of course, most complex gene-prediction systems use a lot of other information about optimal exon (or splice site) combinations that provides a better level of accuracy. However it cannot be applied to study the possible spectrum of all alternative splice sites for a particular gene. Local recognizers seem useful for such tasks.

4.5 IDENTIFICATION OF PROMOTER REGIONS IN HUMAN DNA

Computational recognition of eukaryotic polymerase II (PolII) promoter sequences in genomic DNA is an extremely difficult problem. Promoter 5′-flanking regions may contain dozens of short motifs (5–10 bases) that serve as recognition sites for proteins providing initiation of transcription as well as specific regulation of gene expression. Each promoter has its own composition and arrangement of these elements providing a unique regime

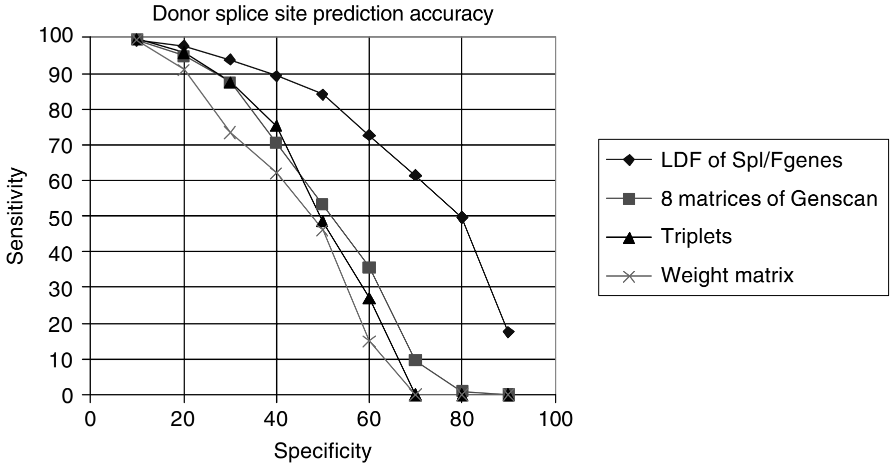


Figure 4.9 Comparison of the accuracy of donor splice-site recognizers: single weight matrix, five matrices suggested by Burge and Karlin (1997), matrix of triplets, linear discriminant function.

Table 4.7 Significance of various characteristics of acceptor splice sites.

Characteristics	1	2	3	4	5	6	7
Individual D^2	5.1	2.6	2.7	2.3	0.01	1.05	2.4
Combined D^2	5.1	8.1	10.0	11.3	12.5	12.8	13.6

1, 3, 4, 6 are the triplet preferences (13) of (−33 to −7) polyT/C tract, consensus (−6 to +5), coding (+6 to +30) and branch point (−48 to −34) regions, respectively. 7 is the number of T and C in intron polyT/C tract region. 2 and 5 are the octanucleotide preferences for being coding 54 bp region on the left and 54 bp region for being intron on the right side of donor splice-site junction.

Table 4.8 Comparing the accuracy of local acceptor splice-site recognizers. The accuracy is averaged for 3 tested sets.

Method	False positives (%)	False negatives (%)	CC	Reference
Weight matrix	5.0	20	0.22	Guigo *et al.* (1992)
Neural network	16.3	6.7	0.35	Brunak *et al.* (1991)
Discriminant analysis	22.0	2.3	0.51	Solovyev *et al.* (1994)

of gene expression. Here we will consider some general features of PolII promoters that can be exploited in promoter prediction programs.

The minimal promoter region that is capable of initiating basal transcription is referred to as the core promoter. It contains a transcription start site (TSS), often located in initiator region (Inr) and typically spans from −60 to +40 bp relative to TSS. About 30–50 % of all known promoters also contain a TATA box at a position about 30 bp upstream from TSS. The TATA box is apparently the most conserved functional signal in eukaryotic promoters. In some cases it can direct accurate transcription initiation by PolII, even in the absence of other control elements. Many highly expressed genes contain a strong TATA box in their core promoter. However for large groups of genes, like most housekeeping genes, some oncogenes and growth factor genes, a TATA box is absent and the corresponding promoters are referred as *TATA-less promoters*. In these promoters, Inr may control the exact position of the transcription start point or the recently found downstream promoter element (DPE), usually located 30 bp downstream of TSS. Many human genes are transcribed from multiple promoters often producing alternative first exons. Moreover, transcription initiation appears to be less precise than initially assumed. In the human genome, it is not uncommon that the 50 ends of mRNAs transcribed from the same promoter region are spread over dozens or hundreds bp (Suzuki *et al.*, 2001; Cooper *et al.*, 2006; Schmid *et al.*, 2006).

The region 200–300 bp immediately upstream of the core promoter constitutes the proximal promoter. The proximal promoter usually contains multiple transcription factor binding sites that are responsible for transcription regulation. Further upstream is the distal part of promoter that may also contain transcription factor binding sites as well as enhancer elements. The typical organization of PolII promoters is shown in Figure 4.10. Because the distal part is usually the most variable region of promoters and generally poorly described, most current promoter recognition tools use the characteristics of only the core and/or proximal regions. Comprehensive reviews of eukaryotic promoters, specifically written from the prediction point of view, have appeared in the literature (Werner, 1999; Pedersen *et al.*, 1999; Cooper *et al.*, 2006).

A collection of experimentally mapped TSSs and surrounding sequences called *Eukaryotic Promoter Database* (EPD) was created by Bucher and Trifonov (1986). In 2000, this database comprised about 800 independent promoter sequences including about 150 human promoters. There is no information about specific regulatory features in this database (Perier *et al.*, 2000). Up to release 72 (October 2002) EPD was a manually compiled database, relying exclusively on experimental evidence published in scientific journals. With release 73, they started to exploit 5′-ESTs from full-length cDNA clones

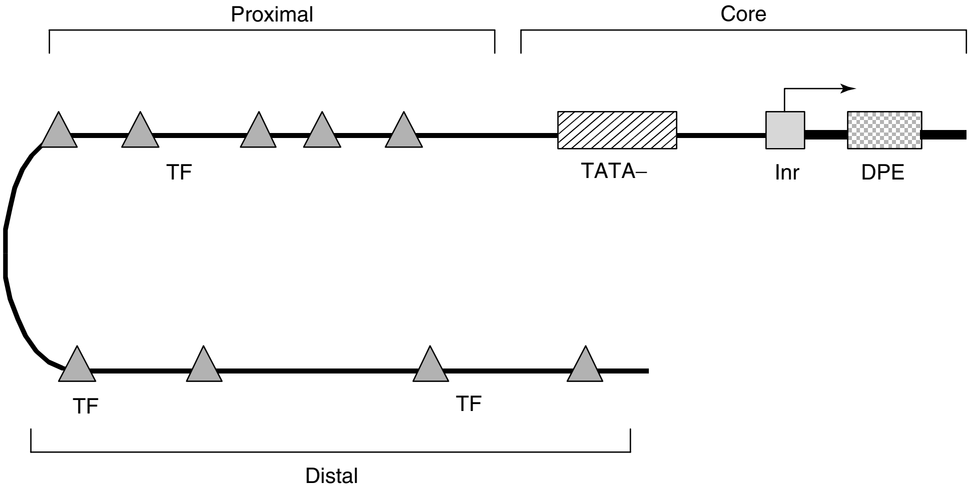


Figure 4.10 Schematic organization of polymerase II promoter. Inr – initiator region, usually containing TSS. DPE – downstream promoter element, often appearing in TATA-less promoters, TF binding sites – transcription factor binding sites.

as a new resource for defining promoters. These data are automatically processed by computer programs and already a year after the introduction of this new method, more than half of the EPD entries (1634) are based on 5′-EST sequences (Schmid *et al.*, 2004). EPD is not the only database providing information about experimentally mapped TSSs. DBTSS (Suzuki *et al.*, 2004) and PromoSer (Halees *et al.*, 2003) are large collections of mammalian promoters based on clustering of EST and full-length cDNA sequences. These databases define the TSS as the furthest 50 position in the genome which can be aligned with the 50 end of a cDNA from the corresponding gene. In contrast, EPD considers the most frequent cDNA 50 end as the TSS and further applies a specialized algorithm to infer multiple promoters for a given gene (Schmid *et al.*, 2006). There is a plant-specific PlantProm (Shahmuradov *et al.*, 2003) database of promoters based on published TSS mapping data.

Regulatory promoter elements are relatively short sequence motifs (typically 5–15 bp in length) (Wingender, 1988; Tjian, 1995; Fickett and Hatzigeorgiou, 1997). A relational TFD including collection of regulatory factors and their binding sites was created by Ghosh (1990; 2000). Over 7900 sequences of transcriptional elements have been described in TRANSFAC database (Wingender *et al.*, 1996; Matys *et al.*, 2006). It gives information about localization and sequence of individual regulatory elements within gene and transcription factors, which bind to them. RegsiteDB (Plant) contains about 1300 various regulatory motifs of plant genes and detail descriptions of their functional properties (`http://www.softberry.com/berry.phtml?topic=regsite`). Practically it is difficult to use most of these motifs to annotate long genomic sequences, because of their short length and degenerate nature. For example, even using well described the TATA box weight matrix there exists one false positive every 120–130 bp (Prestridge and Burks, 1993). Nevertheless, such resources are invaluable for detail analysis of gene regulation and interpretation of experimental data.

To generate regulatory diversity of gene expression, combinations of simple motifs can be used. Transcription factors and regulatory sequences are composed of modular components to achieve the high level of specificity by a relatively small number of different transcription factors (Tjian and Maniatis, 1994). Therefore, to understand gene function we should concentrate our attention on patterns of regulatory sequences rather than on single elements. Searching for such patterns should be much more effective in annotation of new sequences compared to the poor recognition of single motifs. The simplest examples of regulatory patterns are observed in composite regulatory elements (CE) of vertebrate promoters. Composite elements are modular arrangements of contiguous or overlapping binding sites for various distinct factors, raising the possibility that the bound regulatory factors may interact directly, producing novel patterns of regulation. For example, the composite element of proliferin promoter comprises glucocorticoid receptor (GR) and AP-1 factor binding sites. Both GR and AP-1 are expressed in most cell types, but the composite element demonstrated remarkable cell specificity: the hormone–receptor complex repressed the reporter gene expression in CV-1 cells, but enhanced its expression in HeLa cells and had no effect in the F9 cell (Diamond *et al.*, 1990). The database of composite elements (COMPEL) was set up as a common effort by the groups of Wingender (Germany) and Kolchanov (Russia) (Kel *et al.*, 1995). Currently the compilation contains information about several hundred experimentally identified composite elements, where each element consists of two functionally linked sites.

Development of the Transcription Regulatory Regions Database (TRRD), which describes observed regulatory elements in gene regulatory regions, was started in the Kolchanov group (Russia) in 1994 (Kolchanov *et al.*, 2000). TRRD was created by scanning the literature and covers just a fraction of genes taking into account the rather limited resources and very complex nature of the problem of comprehensive and accurate annotation. We are faced with exponentially growing data on transcriptional control owing to the advancement of experimental technologies. This requires us to unite efforts and expertise in creating knowledge databases in this field.

In one of the first attempts to predict eukaryotic promoters, Prestridge (1995) used the density of specific transcription factor binding sites in combination with the TATA box weight matrix. The program PROMOTERSCAN uses the promoter preferences for each binding site listed in TFD (Ghosh, 1990) previously calculated on the set of promoter and nonpromoter sequences. The other general-purpose promoter recognition tools take into account the oligonucleotide content of promoter sequences (Hutchinson, 1996; Audic and Claverie, 1997; Knudsen, 1999; Ohler *et al.*, 1999). In an earlier version of the linear discriminant recognizer, the signal-specific (TATA box weight matrix, binding site preferences) and content-specific characteristics (hexamer preferences) were combined for recognition of TSS (Solovyev and Salamov, 1997).

Fickett and Hatzigeorgiou (1997) presented a performance review of many general-purpose promoter prediction programs. Among these were oligonucleotide content-based (Hutchinson, 1996; Audic and Claverie, 1997), neural network (Guigo *et al.*, 1992; Reese *et al.*, 1996) and the linear discriminant approaches (Solovyev and Salamov, 1997). Although several problems were identified through the relatively small test set (18 sequences) (Ohler *et al.*, 1999), the results demonstrated that the programs can recognize just 50 % of promoters with false-positive rate about 1 per 700–1000 bp. If the average size of a human gene is more than 7000 bases and many genes occupy hundreds of kilobases, then we will expect significantly more false-positive predictions than the number of real promoters. However, these programs can be used to find promoter position (start of transcription and TATA box) in a given 5′-region or to help selecting the correct 5′-exons in gene-prediction approaches.

We will describe a current version of the promoter recognition program TSSW (Transcription Start Site, W stands for using functional motifs from the Wingender *et al.* (1996) database) (Solovyev and Salamov, 1997) to show sequence features that can be used to identify eukaryotic promoter regions. In this version, it was suggested that TATA+ and TATA− promoters have very different sequence features and these groups were analyzed separately. Potential TATA+ promoter sequences were selected by the value of score computed using the TATA box weight matrix (Bucher, 1990) with the threshold closed 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 parts. Significant characteristics of both groups found by discriminant analysis are presented in Table 4.9. This analysis demonstrated that TATA− promoters have much weaker general features compared with TATA+ promoters. Probably TATA− promoters possess more gene-specific structure and they will be extremely difficult to predict by any general-purpose method.

The TSSW program classifies each position of a given sequence as TSS or non-TSS based on two linear discriminant functions (for TATA+ and TATA− promoters) with characteristics calculated in the (−200, +50) region around a given position. If the TATA

Table 4.9 Significance of characteristics of promoter sequences used by TSSW programs for identification of TATA+ and TATA− promoters.

Characteristics	D^2 for TATA+ promoters	D^2 for TATA− promoters
Hexaplets −200 to −45	2.6	1.4 (−100 to −1)
TATA box score	3.4	0.9
Triplets around TSS	4.1	0.7
Hexaplets +1 to +40		0.9
Sp1-motif content		0.9
TATA fixed location	0.7	
CpG content	1.4	0.7
Similarity −200 to −100	0.3	0.7
Motif Density(MD) −200 to +1	4.5	3.2
Direct/Inverted MD −100 to +1	4.0	3.3 (−100 to −1)
Total Mahalonobis distance	11.2	4.3
Number promoters/nonpromoters	203/4000	193/74 000

box weight matrix gives a score higher than some threshold, then the position is classified based on LDF for TATA+ promoters, otherwise the LDF for TATA-less promoters is used. Only one prediction with the highest LDF score is retained within any 300 bp region. If we observe a lower scoring promoter predicted by the TATA-less LDF near a higher scoring promoter predicted by TATA+ LDF, then the first prediction is also retained as a potential enhancer region.

The recognition quality of the program was tested on 200 promoters, which were not included in the learning set. We provide the accuracy values for different levels of true predicted promoters in Table 4.10. The data demonstrate a poor quality of TATA− promoter recognition on long sequences and show that their recognition function can provide relatively unambiguous predictions within regions less than 500 bp. Contrarily, 90 % of TATA+ promoters can be identified within the range 0–2000 bp that makes their incorporation into gene-finding programs valid.

Ohler *et al.* (1999) used interpolated Markov chains in their approach and slightly improved the previous results. They identified 50 % in Fickett and Hatzigeorgiou (1997) promoter set, while having one false-positive prediction every 849 bp. Knudsen (1999), applying a combination of neural networks and genetic algorithms, designed another

Table 4.10 Performance of promoter identification by TSSW program.

Type of promoter	Number of test sites	True predicted (%)	1 false positive per bp
TATA+	101	98	1000
		90	2200
		75	3400
		52	6100
TATA−	96	52	500
		40	1000

program (Promoter2.0). Promoter 2.0 was tested on a complete Adenovirus genome 35 937 bases long. The program predicted all 5 known promoter sites on the plus strand and 30 false-positive promoters. The average distance between a real and the closest predicted promoter is about 115 bp. The TSSW program with the threshold to predict all 5 promoters produced 35 false positives. It gives an average distance between predicted TSS and real promoter of just 4 bp (2 predicted exactly, 1 with 1 bp shift, 1 with 5 bp shift and the weakest promoter was predicted with 15 bp shift).

Figure 4.11 shows an example of the results of the TSSW program for the sequence of human laminin beta 2 chain (GenBank accession number Z68155). The structure of this gene including its promoter region has been extensively studied. The length of gene is 11 986 bp, the first 1724 bp of which constitute a promoter region. TSSW predicts one enhancer at position 931 and one potential TSS at position 1197 with corresponding TATA box at the position 1167. Although both the predicted sites fall inside the designated promoter region, the second prediction is probably a false positive, because the

```
>H.sapiens LAMB2 gene for    laminin beta 2 chain
 Length of sequence    -     11986bp
 Thresholds for TATA+ promoters     -  0.45, for TATA-/enhancers -  3.70

 2 promoter/enhancer(s) are predicted

 Enhancer Pos  :    931 LDF score -  3.78
 Promoter Pos  :   1197 LDF score -  1.13  TATA box at   1167   Score -
18.96

 Transcription factor binding sites:
 for enhancer  at position -     931
   874 (+) CHICK$ACRA   CCGCCC
   778 (-) Y$ADH2_01    TCTCC
   631 (-) Y$ADH2_01    TCTCC
   831 (+) RAT$ANTEN_   ccacagttgggatttCCCAACctgaccag
   842 (+) RAT$ANTEN_   ccacagttgggatttCCCAACctgaccag
   879 (-) HS$APOE_08   GGGCGG
   876 (-) Y$CYC1_09    ctcatttggcgagcGTTGGt
   865 (-) Y$CYC1_09    ctcatttggcgagcGTTGGt
   842 (-) Y$CYC1_09    ctcatttggcgagcGTTGGt
   657 (+) AD$E2L_04    TGACgcA
   833 (+) AD$E2L_04    TGACgcA
   835 (-) HS$EGFR_15   TCAAT
   840 (-) RAT$EAI_09   GTCAG
   649 (+) Y$GAL1_10    AGCCT
   929 (-) MOUSE$AAG_   gcaacTGATAaggat
   928 (-) MOUSE$AAG_   cctgTGATAagga
   841 (+) HS$BG_01     ccaCACCCg
   852 (+) HS$BG_01     ccaCACCCg
   863 (+) HS$BG_01     ccaCACCCg
   907 (-) HS$BG_01     ccaCACCCg
   773 (-) HS$BG_01     ccaCACCCg
   661 (-) HS$BG_01     ccaCACCCg
```

Figure 4.11 An example of output of the TSSW program for the sequence of human laminin beta 2 chain (GenBank accession number Z68155).

predicted TATA box is located far upstream (500 bp) from the experimentally determined beginning of the 5′-UTR. TSSW also optionally lists all potential TF binding sites around the predicted promoters or enhancers (Figure 4.11). It outputs the position, the strand (+/−), the TRANSFAC identifier and the consensus sequences of sites found. The information about these sites may be of interest for researchers studying the transcription of a particular gene.

There is a high false-positive rate of promoter prediction in long genomic sequences. It is more useful to remove some false-positive predictions using knowledge of the positions of the coding regions. TSSW was additionally tested on the several GenBank entries that have information about experimentally verified TSS and were not included in the learning set (Table 4.11). The lengths of the sequences varied from 950 to 28 438 bp with a median length of 2938 bp. According to the criteria defined by Fickett and Hatzigeorgiou (1997), all true TSS in these sequences can be considered as correctly predicted, with an average 1.5 false positives per sequence or 1 false positive per 3340 bp. The distances between true TSS and those correctly predicted varied from exact matching to 196 bp, with the median deviation of 9 bp. This can be considered to be quite a good prediction taking into account that experimental mapping of TSS has an estimated precision of +/− 5 bp (Perier *et al.*, 2000).

Accurate prediction of promoters is fundamental to understanding gene expression patterns, where confidence estimation of the produced predictions is one of the main requirements for many applications. Using recently developed transductive confidence machine (TCM) techniques, we developed a new program TSSP-TCM (Shahmuradov *et al.*, 2005) for the prediction of plant promoters that also provides confidence of the prediction. The method presented in the paper identifies ~85 % of tested promoters with one false positive per ~5000 bp. It allows us not only to make predictions, but more importantly, it also gives valid measures of confidence in the predictions for each individual example in the test set. Validity in our method means that if we set up a confidence level, say, 95 %, then we can guarantee that we are not going to have more than 5 errors out of 100 examples.

Recently there was an attempt to make a critical assessment of the promoter prediction accuracy in its current state relative to the manual Havana gene annotation (Bajic *et al.*, 2006). There were only 4 programs in this EGASP project: 2 variants of McPromoter

Table 4.11 Results of TSSW predictions on some GenBank entries with experimentally verified TSS.

Gene	GenBank accession number	Length (bp)	True TSS	Predicted TSS	Number of false positives
CXCR4	AJ224869	8747	2632	2631	4
HOX3D	X61755	4968	2280	2278	2
DAF	M64356	2003	733	744	1
GJB1	L47127	950	404	428	0
SFRS7	L41887	8213	< 415	417	4
ID4	AF030295	1473	1066	1081	1
C inhibitor	M68516	15 571	2200	2004	4
MBD1	AJ132338	2951	1964	1876	1
Id-3	X73428	2481	665	663	0

program (Ohler *et al.*, 1999; 2002), N-scan (Arumugam *et al.*, 2006) and Fprom (Solovyev *et al.*, 2006), which is a modification of TSSW program described above that use TFD transcriptional motif database (Ghosh, 2000). McPromoter and Fprom derived its predictions from a sequence of the genome under analysis; N-scan used corresponding sequences of several genomes (such as human, mouse and chicken). When the maximum allowed mismatch of the prediction from the reference TSS for counting true positive predictions on test sequences was 1000 bp, the N-scan produced ∼3 % higher accuracy than the next most accurate predictor Fprom, but for the distance criterion 250 bp Fprom shows the best performance on most prediction accuracy measures (Bajic *et al.*, 2006). We should note that the sensitivity of computational promoter predictions was only 30–50 % (relatively 5′-gene ends of Havana annotation), but we should take into account that TSS annotation from two experimentally derived databases also produced a sensitivity of only 48–58 %. This leaves the open issue to create a reliable TSS reference dataset. The lesson from this EGASP experiment relative to promoter predictions is that it is beneficial to combine the TSS/promoter predictions with gene-finding programs as was done in generating N-scan or Fprom predictions.

Despite recent improvements in promoter prediction programs, their current accuracy is still not enough for their successful implementation as independent submodules in gene-recognition software tools. The rather small amount of experimentally verified promoters in databases such as EPD and GenBank hindered computational promoter identification progress, while now there is an order more promoter data became available (Seki *et al.*, 2002) generated by CapTrapper technique (Carninci *et al.*, 1996) that provided a relatively reliable method for promoter identification. Many of the above-mentioned promoter prediction algorithms use the propensities of each TF binding site independently and does not take into account their mutual orientation and positioning. It is well known that the transcriptional regulation is a highly cooperative process, involving simultaneous binding of several transcription factors to their corresponding sites. Specific groups of promoters may have specific patterns of regulatory sequences, where mutual orientation and location of individual regulatory elements are necessary requirements for successful transcription initiation or regulation.

4.6 RECOGNITION OF POLYA SITES

Another functionally important signal of eukaryotic transcripts is the 3′-untranslated region (3′UTR), which has a diversity of cytoplasmic functions affecting the localization, stability and translation of mRNAs (Decker and Parker, 1995). Almost all eukaryotic mRNAs undergo 3′-end processing which involves endonucleotide cleavage followed by the polyadenylation of the upstream cleavage product (Wahle, 1995; Manley, 1995). The essential sequences are involved in the formation of several large RNA-protein complexes (Wilusz *et al.*, 1990). RNA sequences directing binding of specific proteins are frequently poorly conserved and often recognized in a cooperative fashion (Wahle, 1995). Therefore we have been forced to use statistical characteristics of the polyA regions that may involve some unknown significant sequence elements.

Numerous experiments have revealed three types of RNA sequences defining a 3′-processing site (Wahle, 1995; Proudfoot, 1991) (Figure 4.12). The most conserved

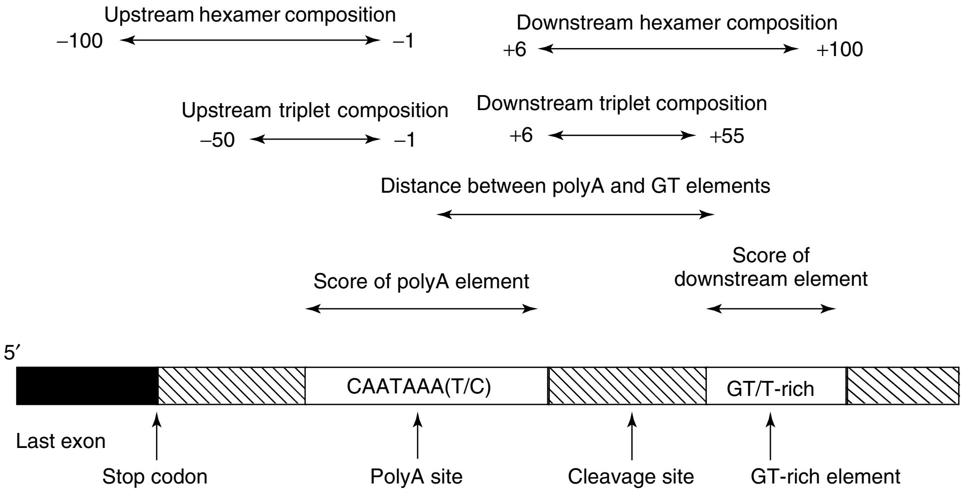

Upstream hexamer composition
−100
−1
Downstream hexamer composition
+6
+100
Upstream triplet composition
−50
−1
Downstream triplet composition
+6
+55
Distance between polyA and GT elements
Score of polyA element
Score of downstream element
5′
CAATAAA(T/C)
GT/T-rich
Last exon
Stop codon
PolyA site
Cleavage site
GT-rich element

Figure 4.12 Characteristics of polyA signal sequences.

is the hexamer signal AAUAAA (polyA signal), situated 10–30 nucleotides upstream of the 3′-cleavage site. About 90 % of sequenced mRNAs have a perfect copy of this signal. Two other types, the upstream and the downstream elements, are degenerate and have not been properly characterized. Downstream elements are frequently located within approximately 50 nucleotides 3′ of the cleavage site (Wahle and Keller, 1992). These elements are often GU- or U-rich, although they may have various base compositions and locations. On the basis of sequence comparisons McLauchlan *et al.* (1985) have suggested that one of the possible consensuses of the downstream element is YGUGUUYY. The efficiency of polyadenylation in a number of genes can be also increased by sequences upstream of AAUAAA, which are generally U-rich (Wahle, 1995). All these RNA sequences serve as nucleation sites for a multicomponent protein complex catalyzing the polyadenylation reaction.

There have been a few attempts to predict 3′-processing sites by computational methods. Yada *et al.* (1994) conducted a statistical analysis of human DNA sequences in the vicinity of polyA signal in order to distinguish them from AATAAA sequences that are not active in polyadenylation (pseudo polyA signals). They found that a base C frequently appears on the upstream side of the AATAAA signal and a base T or C often appears on the downstream side, implying that CAATAAA(T/C) can be regarded as a consensus of the polyA signal. Kondrakhin *et al.* (1994) constructed a generalized consensus matrix using 63 sequences of cleavage/polyadenylation sites in vertebrate pre-mRNA. The elements of the matrix were absolute frequencies of triplets at each site position. Using this matrix, they have provided a multiplicative measure for recognition of polyadenylation regions. However this method has a very high false-positive rate.

Salamov and Solovyev (1997) developed LDF recognition function for polyA signal. The data sets for 3′-processing sites and 'pseudo' polyA signals were extracted from GenBank (Version 82). 3′-processing sites were taken from the human DNA entries, containing a description of the polyA signal in the feature table. Pseudosites were taken out of human genes as the sequences comprising (−100, +200) around the patterns revealed by polyA weight matrix (see below), but not assigned to polyA sites in the feature table.

4.7 CHARACTERISTICS FOR RECOGNITION OF 3′-PROCESSING SITES

As the hexamer AATAAA is the most conservative element of 3′-processing sites, it was considered as the main block in our complex recognition function. Although the hexamer is highly conserved, other variants of this signal were observed. For example, in the training set, 43 out of 248 polyA sites had hexamer variants of AATAAA with one mismatch. To consider such variants the position weight matrix for recognizing this signal has been used. The other characteristics such as content statistics of hexanucleotides and positional triplets in the upstream and downstream regions were defined relative to the position of the conservative hexamer sequence (Figure 4.12).

1. Position weight matrix for scoring of polyA signal $[-1, +7]$.
2. Position weight matrix [8] for scoring of downstream GT/T-rich element.
3. Distance between polyA signal and predicted downstream GT/T-rich element.
4. Hexanucleotide composition of downstream $(+6, +100)$ region.
5. Hexanucleotide composition of upstream $(-100, -1)$ region.
6. Positional triplet composition of downstream $(+6, +55)$ region.
7. Positional triplet composition of upstream $(-50, -1)$ region.
8. Positional triplet composition of the GT/T-rich downstream element

In Table 4.12, the Mahalonobis distances for each characteristic calculated on the training set are given. The most significant characteristic is the score of AATAAA pattern (estimated by the position weight matrix) that indicates the importance of occurrences almost perfect polyA signal (AATAAA). The second valuable characteristic is the hexanucleotide preferences of the downstream $(+6, +100)$ region. Although the discriminating ability of GT-rich downstream element itself (characteristic 2) is very weak, combining it with the other characteristics significantly increases the total Mahalonobis distance.

Kondrakhin *et al.* (1994) reported the error rates of their method at different thresholds for polyA signal selection. If the threshold is set to predict 8 of 9 real sites, their function also predicts 968 additional false sites. The algorithm-based LDF for 3′-processing site identification is implemented in the POLYAH program (`http://genomic.sanger.ac.uk`). First, it searches for the pattern similar to AATAAA using the weight matrix and, if the pattern is found, it computes the value of the linear discriminant function defined by the characteristics around this position. A polyA site is predicted if the value of this function is greater than an empirically selected threshold. The method demonstrates $S_n = 0.86$ and $S_p = 0.63$ when applied to a set of 131 positive and 1466 negative examples

Table 4.12 Significance of various characteristics of polyA signal.

Characteristics	1	4	2	5	3	6	8	7
Individual D^2	7.61	3.46	0.01	2.27	0.44	1.61	0.16	0.17
Combined D^2	7.61	10.78	11.67	12.36	12.68	12.97	13.09	13.1

that were not used in the training. The POLYAH program has been tested also on the sequence of the Ad2 genome, where for 8 correctly identified sites it predicts only 4 false sites.

4.8 IDENTIFICATION OF MULTIPLE GENES IN GENOMIC SEQUENCES

Computational gene finding started a long time ago with looking for open reading frames with an organism-specific codon usage (Staden and McLachlan, 1982). The approach worked well for bacterial genes (Staden, 1984b; Borodovskii *et al.*, 1986), but short eukaryotic exons and spliced genes require algorithms combining information about functional signals and the regularities of coding and intron regions. Several internal exon-predicting algorithms have been developed. The program SORFIND (Hutchinson and Hayden, 1992) was designed to predict internal exons based on codon usage plus Berg and von Hippel (1987) suggested discrimination energy for intron–exon boundary recognition. The accuracy of exact internal exon prediction (at both 5′- and 3-′ splice junctions and in the correct reading frame) by the SORFIND program reaches 59 % with a specificity of 20 %. Snyder and Stormo (1993) applied a dynamic programming approach (alternative to the rule-based approach) to internal exon prediction in GeneParser algorithm. It recognized 76 % of internal exons, but the structure of only 46 % exons was exactly predicted when tested on the entire GenBank sequence entries. HEXON (Human EXON) program (Solovyev *et al.*, 1994) based on linear discriminant analysis was the most accurate in exact internal exon prediction at that time.

Later a number of single gene-prediction programs has been developed to assemble potential eukaryotic coding regions into translatable mRNA sequence 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 of finding an optimal combination of preselected exons (Gelfand and Roytberg, 1993; Solovyev and Lawrence, 1993b; Xu *et al.*, 1994). This is different from the approach suggested by Snyder and Stormo (1993) in the GeneParser algorithm to recursively search for exon–intron boundary positions. 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) have made a comprehensive test of gene-finding algorithms. The FGENEH program was one of the best in the tested group having the exact exon prediction accuracy 10 % higher than the other programs and the best level of accuracy at the protein level. A novel step in gene-prediction approaches was application of generalized Hidden Markov Models implemented in Genie algorithm. It was similar in design to GeneParser, but was based on a rigorous probabilistic framework (Kulp *et al.*, 1996). The algorithm demonstrated similar performance to FGENEH.

4.9 DISCRIMINATIVE AND PROBABILISTIC APPROACHES FOR MULTIPLE GENE PREDICTION

Genome sequencing projects require gene-finding approaches able to identify many genes encoded in the transcribed sequences. The value of sequence information for the biomedical community is strongly dependent on the availability of candidate genes that are computationally predicted. The best multiple gene-prediction programs involve HMM-based probabilistic approaches as implemented in Genscan (Burge and Karlin, 1997) and Fgenesh (Salamov and Solovyev, 2000), Fgenes (discriminative approach) (Solovyev *et al.*, 1995) and Genie (Generalized HMM with neural network splice-site detectors) (Reese *et al.*, 2000). Initially we will describe a general scheme for HMM-based gene prediction (Stormo and Haussler, 1994) (first implemented by the Haussler group (Krogh *et al.*, 1994; Kulp *et al.*, 1996)) as the most general description of the gene model. A pattern-based approach can be considered as a particular case of this approach, where transition probabilities are not taken into account.

4.9.1 HMM-based Multiple Gene Prediction

Different components (states) of gene structure such as exons, introns and 5′-untranslated regions occupy k subsequences of a sequence X:$X = \bigcup_{i=1,k} x_i$. There are 35 states that describe eukaryotic gene model considering direct and reverse strands as possible gene locations (Figure 4.13). However, in the current gene-prediction approaches, noncoding 5′- and 3′-exons (and introns) are not considered because the absence of protein-coding characteristics makes their prediction less accurate. In addition, the major practical goal of gene prediction is to identify the protein-coding sequences. The remaining 27 states include 6 exon states (first, last single and 3 types of internal exons in 3 possible reading frame) and 7 noncoding states (3 intron, noncoding 5′- and 3′-, promoter and polyA) in each strand plus the noncoding intergenic region.

The predicted gene structure can be considered as the ordered set of states/sequence pairs, $\phi = \{(q_1, x_1), (q_2, x_2), \ldots, (q_k, x_k)\}$, called the *parse*, such that the probability $P(X, \phi)$ of generating X according to ϕ is maximal over all possible parses (or a score is optimal in some meaningful sense that best explains the observations (Rabiner, 1989)):

$$P(X,\phi) = P(q_1)\left(\prod_{i=1}^{k-1} P(x_i|l(x_i), q_i)P(l(x_i)|q_i)(P(q_{i+1}, q_i)\right)P(x_i|l(x_k), q_k)P(l(x_k)|q_k),$$

where $P(q_1)$ denotes the initial state probabilities; $P(x_i|l(x_i), q_i)P(l(x_i)|q_i)$ and $P(q_{i+1}, q_i)$ are the independent joint probabilities of generating the subsequence x_i of length l in the state q_i and transitioning to $i + 1$ state.

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 on the order of 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). A useful technique was introduced by Burge (1997) to reduce the number of states and simplify computations by modeling noncoding state length by a geometrical distribution. The algorithm for gene finding using this technique was initially implemented in the Genscan program (Burge and Karlin, 1997) and used later in Fgenesh program (Salamov and Solovyev, 2000). Since any valid parse will consist only of an alternating series of Noncoding and Coding states: NCNCNC, . . , NCN, we need only 11 variables, corresponding to the different types of N states. At each step corresponding to some sequence position, we select the maximum joint probability to continue the current state or to move to another noncoding

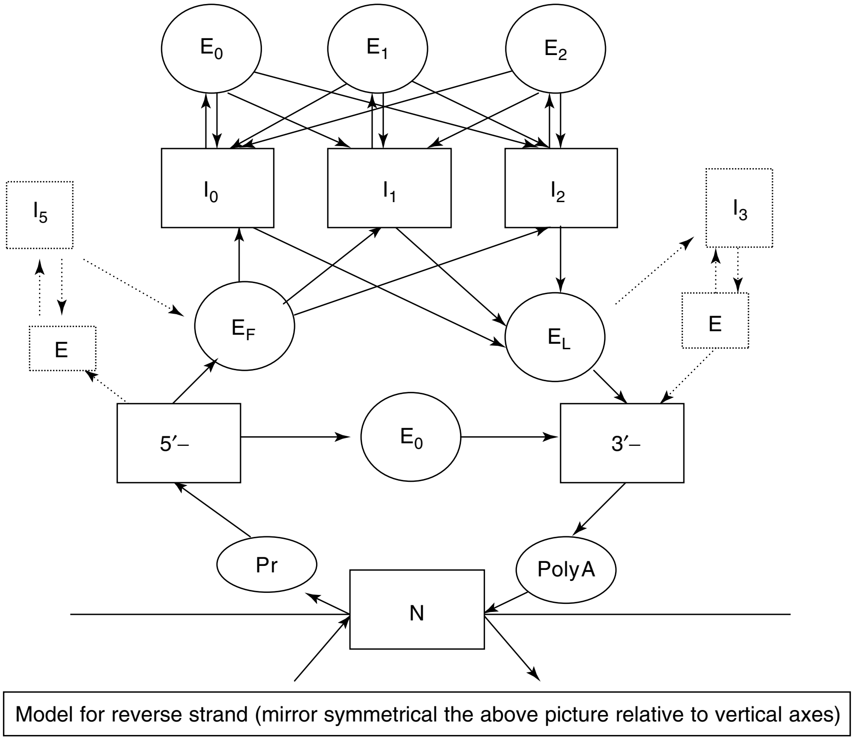

E$_0$
E$_1$
E$_2$
I$_0$
I$_1$
I$_2$
I$_5$
I$_3$
E
E$_F$
E$_L$
E
5′–
E$_0$
3′–
Pr
PolyA
N
Model for reverse strand (mirror symmetrical the above picture relative to vertical axes)

Figure 4.13 Different states and transitions in eukaryotic HMM genes model. E$_i$ and I$_i$ are different exon and intron states, respectively (i = 0,1,2 reflect 3 possible different ORF). E marks noncoding exons and I5/I3 are 5′- and 3′-introns adjacent to noncoding exons.

state defined by a coding state (from a precomputed list of possible coding states) that ends in analyzed sequence position.

Define the best score (highest joint probability) $\gamma_i(j)$ of optimal parse of the subsequence $s_{1,j}$, which ends in state q_i at position j. We have 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 $\{b_k\}$. The length distribution of state c_k is denoted by $f_{c_k}(d)$. The searching procedure can be stated as follows:

Initialization:

$$\gamma_i(1) = \pi_i P_i(S_1) p_i, \; i = 1, \ldots 11.$$

Recursion:

$$\begin{aligned}\gamma_i(j+1) = \mathbf{max}\{&\gamma_i(j) p_i P_i(S_{j+1}),\\ &\mathbf{max}_{c_k \in A_j}\{\gamma_i(m_k - 1)(1 - p_{b_k}) t_{b_k,c_k} f_{c_k}(d_k) P(S_{m_k,j}) t_{c_k,i} p_i P_i(S_{j+1})\}\}\\ i = 1, \ldots 11,\; &j = 1, \ldots, L - 1.\end{aligned}$$

Termination:

$$\begin{aligned}\gamma_i(L+1) = \mathbf{max}\{&\gamma_i(L),\\ &\mathop{\mathbf{max}}_{c_k \in A_j}\{\gamma_i(m_k - 1)(1 - p_{b_k})t_{b_k,c_k} f_{c_k}(d_k) P(S_{m_k,j}) t_{c_k,i}\}\}\\ i &= 1, \ldots 11.\end{aligned}$$

At 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 on the learning set of known gene structures. Instead of scores of coding states $P(S_{m_k,j})$ it is better to use log likelihood ratios, which do not produce scores below the limits of computer precision.

Genscan (Burge and Karlin, 1997) was the first published algorithm to predict multiple eukaryotic genes. Several HMM-based gene-prediction programs were developed later: Veil (Henderson *et al.*, 1997), HMMgene (Krogh, 1997), Fgenesh (Salamov and Solovyev, 2000), a variant of Genie (Kulp *et al.*, 1996) and GeneMark (Lukashin and Borodovsky, 1998). Fgenesh is currently one of the most accurate programs. It is different from Genscan because, in the model of gene structure, a signal term (such as splice site or start site score) has some advantage over a content term (such as coding potential). In log likelihood terms, the splice sites and other exon functional signals have an additional score, depending on the environments of the sites. Also, in computing the coding scores of potential exons, *a priori* probabilities of exons are taken into account according to Bayes theorem. As a result, the coding scores of potential exons are generally lower than in Genscan. Fgenesh works with separately trained parameters for each model organism such as human, drosophila, chicken, nematode, dicot and monocot plants, and dozen yeast/fungi (currently using known genes or predicted protein-supported genes, Fgenesh gene-finding parameters has been computed for about 40 various organisms: `http://sun1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind)`. Coding potentials were calculated separately for 4 isochores (human) and for 2 isochores (other species). The run time of Fgenesh is practically linear and the current version has no practical limit on the length of analyzed sequence. Prediction of about 800 genes in 34 MB of Chromosome 22 sequence takes about 1.5 minutes of Dec-alpha processor EV6 for the latest Fgenesh version.

4.9.2 Pattern-based Multiple Gene-prediction Approach

FGENES (Solovyev, 1997) is the multiple gene-prediction program based on dynamic programming. It uses discriminant classifiers to generate a set of exon candidates. Similar discriminant functions were developed initially in Fexh (Find Exon), Fgeneh (Find GENE) program (h stands for version to analyze human genes) and described in details earlier (Solovyev and Lawrence, 1993a; Solovyev *et al.*, 1995, Solovyev and Salamov, 1997).

The following major steps describe the analysis of genomic sequences by the Fgenes algorithm:

1. Create a list of potential exons, selecting all ORFs: ATG.. GT, AG–GT, AG.. Stop with exons scores higher than the specific thresholds depending on GC content (4 groups);

2. Find the set of compatible exons with maximal total score. Guigo (1998) described an effective algorithm for finding such set. Fgenes uses a simpler variant of a similar algorithm: Order all exon candidates according to their 3′-end positions; Going from the first to the last exon select for each exon the maximal score path (compatible exons combination) terminated by this exon using the dynamic programming approach. Include in the optimal gene structure either this exon or the exon with the same 3′-splicing pattern ending at the same position or earlier (which has the higher maximal score path).
3. Take into account promoter or polyA scores (if predicted) in the terminal exon scores.

The run time of the algorithm grows approximately linearly with the sequence length. Fgenes is based on the linear discriminant functions developed earlier for the identification of splice sites, exons, promoter and polyA sites (Solovyev *et al.*, 1994; Salamov and Solovyev, 1997). We consider these functions in the following sections to see what sequence features are important in exon prediction.

4.10 INTERNAL EXON RECOGNITION

For internal intron prediction, we consider all open reading frames in a given sequence that are flanked by AG (on the left) and by GT (on the right) as potential internal exons. The structure of such exons is presented in Figure 4.14. The values of 5 exon characteristics were calculated for 952 authentic exons and for 690 714 pseudoexon training sequences from the set. The Mahalonobis distances showing significance of each characteristic are given in Table 4.13. We can see that the strongest characteristics for exons are the values of the recognition functions for the flanking donor and acceptor splice sites ($D^2 = 15.04$ and $D^2 = 12.06$, respectively). The preference of an ORF being a coding region has $D^2 = 1.47$ and adjacent left intron region has $D^2 = 0.41$ and right intron region has $D^2 = 0.18$.

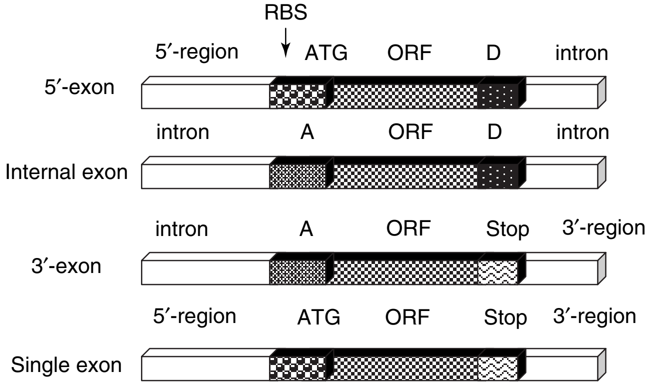


Figure 4.14 Different functional regions of the first, internal, last and single exons corresponding to components of recognition functions.

Table 4.13 Significance of internal exon characteristics.

	Characteristics	1	2	3	4	5
a	Individual D^2	15.0	12.1	0.4	0.2	1.5
b	Combined D^2	15.0	25.3	25.8	25.8	25.9

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

The performance of the discriminant function based on these characteristics was estimated using 451 exon and 246 693 pseudoexon sequences from the test set. The general accuracy of the exact internal exon prediction is 77 % with specificity 79 %. At the level of individual nucleotides, the sensitivity of exon prediction is 89 % with specificity 89 %; and the sensitivity of prediction of intron positions is 98 % with specificity 98 %. This accuracy is better than in the most accurate dynamic programming and neural network-based methods (Snyder and Stormo, 1993), which have 75 % accuracy of the exact internal exon prediction with specificity 67 %. The method has 12 % less false exon assignments with a better level of true exon prediction.

4.11 RECOGNITION OF FLANKING EXONS

Figure 4.15 shows the 3-dimensional histograms reflecting the oligonucleotide composition of the gene flanking regions based on a graphical fractal representation of nucleotide sequences (Jeffrey, 1990; Solovyev *et al.*, 1991; Solovyev, 1993). The clear differences in compositions were exploited to develop of recognizers of these regions.

4.11.1 5′-terminal Exon-coding Region Recognition

For 5′-exon prediction, all the open reading frames in a given sequence starting with the ATG codon and ending with the GT dinucleotide were considered as potential first exons. The structure of such exons is presented in Figure 4.14. The exon characteristics and their Mahalonobis distances are given in Table 4.14. The accuracy of the discriminant function based on these characteristics was computed using the recognition of 312 first exons and 246 693 pseudoexon sequences. The gene sequences were scanned and the 5′ exon with the maximal weight was selected for each of them. The accuracy of the prediction of the true first coding exon is 59 %. Competition with the internal exons was not considered in this test.

4.11.2 3′-exon-coding Region Recognition

All ORF regions that are flanked by GT (on the left) and finish with a stop codon were considered as potential last exons. The structure of such exons is presented in Figure 4.14. The characteristics of the discriminant functions and their Mahalonobis distances are presented in Table 4.15. The accuracy of the discriminant function was tested on the

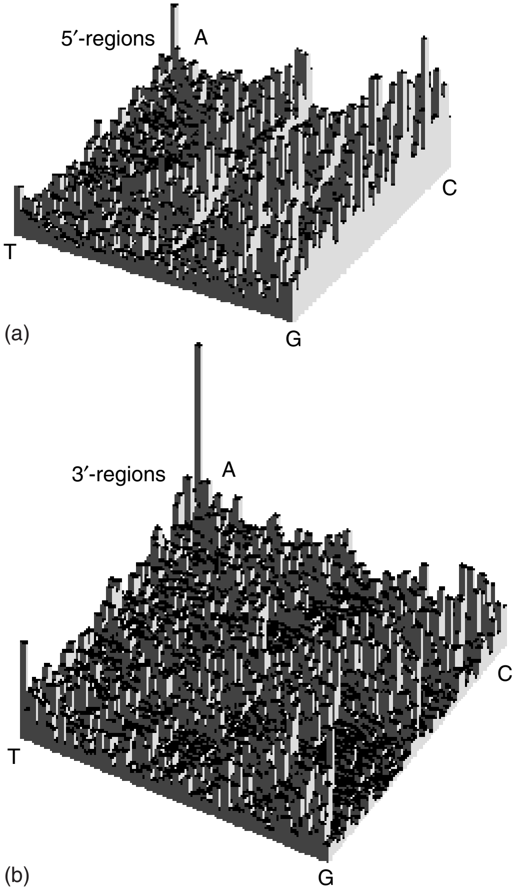


Figure 4.15 Graphical representation of the number of different oligonucleotides 6 bp long in 5’ (a) and 3’ (b) gene regions. Each colon is the number of a particular oligonucleotide in the set of sequences.

Table 4.14 Significance of 5′- exon characteristics.

	Characteristics	1	2	3	4	5	6	7
a	Individual D^2	5.1	2.6	2.7	2.3	0.01	1.05	2.4
b	Combined D^2	5.1	8.1	10.0	11.3	12.5	12.8	13.6

Characteristic 1 is the value of donor site recognition function. 2 is the average value of positional triplet preferences in −15 to +10 region around ATG codon. 4 gives the octanucleotide preferences for being intron in 70 bp region on the right of potential exon. 3, 5 and 7 are the hexanucleotide preferences in −150 to −101 bp, −100 to −51 bp and −50 to −1 bp regions on the left of potential exon, respectively; 6 is the octanucleotide preferences for being coding in exon region.

recognition of 322 last exons and 2 47 644 pseudoexon sequences. The gene sequences were scanned and the 3′ exon with the maximal weight was selected for each of them. The function can identify 60 % of annotated last exons.

Table 4.15 Significance of 3′-exon characteristics.

	Characteristics	1	2	3	4	5	6	7
a	Individual D^2	10.0	3.2	0.8	2.2	1.2	0.2	1.6
b	Combined D^2	10.0	11.4	12.0	13.8	14.3	14.5	14.6

Characteristic 1 is the value of acceptor site recognition. Characteristic 2 is the octanucleotide preferences for being coding of ORF region. 3, 5 and 7 are the hexanucleotide preferences in +100 to 150 bp, +50 to +100 bp and +1 to +50 bp regions on the left of coding region, respectively. 4 is the average value of positional triplet preferences in −10 to +30 region around the stop codon. 6 is the octanucleotide preferences for being intron in 70 bp region on the left of exon sequence.

The recognition function for single exons combines the corresponding characteristics of 5′- and 3′-exons.

4.12 PERFORMANCE OF GENE IDENTIFICATION PROGRAMS

Most gene-recognition programs were tested on a specially selected set of 570 single gene sequences (Burset and Guigo, 1996) of mammalian genes (Table 4.16). The best programs predict accurately on average 93 % of the exon nucleotides ($S_n = 0.93$) with just 7 % of false-positive predictions. Because the most difficult task is to predict small exons and exactly identify exon 5′ and 3′ ends, the accuracy at the exon level is usually lower than at the nucleotide level.

The table demonstrates that the modern multiple gene-prediction programs as Fgenesh, Fgenes and Genscan significantly outperform the older approaches. The exon identification rate is actually even higher than the data presented as the overlapped exons were not

Table 4.16 Accuracy of the best gene-prediction programs for single gene sequences from (Burset and Guigo, 1996) data set. S_n (sensitivity) = number of exactly predicted exons/number of true exons (or nucleotide); S_p (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 received by authors of programs.

Algorithm	S_n (exons)	S_p (exons)	S_n nucleotides	S_p nucleotides	Authors/year
Fgenesh	0.84	0.86	0.94	0.95	Solovyev and Salamov (1999)
Fgenes	0.83	0.82	0.93	0.93	Solovyev (1997)
Genscan	0.78	0.81	0.93	0.93	Burge and Karlin (1997)
Fgeneh	0.61	0.64	0.77	0.88	Solovyev *et al.* (1995)
Morgan	0.58	0.51	0.83	0.79	Salsberg *et al.* (1998)
Veil	0.53	0.49	0.83	0.79	Henderson *et al.* (1997)
Genie	0.55	0.48	0.76	0.77	Kulp *et al.* (1996)
GenLang	0.51	0.52	0.72	0.79	Dong and Searls (1994)
Sorfind	0.42	0.47	0.71	0.85	Hutchinson and Hyden (1992)
GeneID	0.44	0.46	0.63	0.81	Guigo *et al.* (1992)
Grail2	0.36	0.43	0.72	0.87	Xu *et al.* (1994)
GeneParser2	0.35	0.40	0.66	0.79	Snyder and Stormo (1995)
Xpound	0.15	0.18	0.61	0.87	Thomas and Skolnick (1994)

counted in exact exon predictions. However, there is a lot of room for future improvement. The accuracy at the level of exact gene prediction is only 59 % for Fgenesh, 56 % for Fgenes and 45 % for the Genscan program even on this relatively simple test set.

The real challenge for *ab initio* gene identification is to find multiple genes in long genomic sequences containing genes on both DNA strands. Often, there is no complete information about the real genes in such sequences. One example studied experimentally at the Sanger Centre (UK) is the human BRACA2 region (1.4 MB) that contains eight genes and 169 experimentally verified exons. This region is one of the worse cases for genome annotation because it has genes with many exons and almost all genes show no similarity of their products with known proteins. Moreover, it contains four pseudogenes and at least two of the genes have alternative splicing variants. The results of gene prediction initially provided by T. Hubbard and R. Bruskiewich (Sanger Centre Genome Annotation Group) are shown in Table 4.17.

Fgenesh predicts 20 % less false-positive exons in this region than the Genscan approach, with the same level of true predicted exons. Even for such difficult region about 80 % of exons were identified exactly by *ab initio* approaches.

The accuracy of the gene-finding programs depends not only on underlying algorithm, but also strongly affected by parameter file computed on the learning set of known genes. While Fgenesh and Genscan demonstrate similar performance for human gene prediction, Fgenesh has shown significantly better accuracy than many other tested gene-finders (including Genscan) in predicting rice genes (Yu *et al.*, 2002).

4.13 USING PROTEIN SIMILARITY INFORMATION TO IMPROVE GENE PREDICTION

The lessons from manual annotations show that it is often advantageous to take into account all the available information to improve gene identification. Automatic gene-prediction approaches can take into account the information about exon similarity with

Table 4.17 Accuracy of gene-prediction programs for BRACA2 1.4 MB human genomic sequence. Masked is prediction when repeats have been defined by RepeatMasker (Smit and Green, 1997) program in the analyzed sequence and excluded from potential exon location during prediction. The region consisted of 20 sequences with 8 verified genes, 4 pseudogenes and 169 exons. Later one sequence was constructed and three additional exons were identified. The results of prediction on this sequence marked bold.

	CC	$S_{\rm nb}$	$S_{\rm pb}$	$P_{\rm e}$	$C_{\rm e,ov}$	$S_{\rm ne}$	$S_{\rm n,ov}$	$S_{\rm pe}/\boldsymbol{S}_{\rm pe,ov}$
Genscan	0.68	90	53	271/**271**	109/**131**	65	80/**76**	40/**49**
Fgenesh	0.80	89	73	188/**195**	115/**131**	69	80/**76**	61/**67**
Fgenes	0.69	79	62	298/**281**	110/**136**	66	86/**78**	37/**48**
Genscan masked	0.76	90	66	217	109	65	80	50
Fgenesh masked	0.84	89	82	172/**168**	114/**131**	68	79/**73**	66/**76**
Fgenes masked	0.73	80	68	257/**228**	107/**133**	64	85/**75**	42/**58**

CC is the correlation coefficient reflecting the accuracy of prediction at the nucleotide level. $S_{\rm nb}$, $S_{\rm pb}$ – sensitivity and specificity at the base level (in %), $P_{\rm e}$ – number of predicted exons, $C_{\rm e}$ –number of correctly predicted exons, $S_{\rm ne}$, $S_{\rm pe}$ – sensitivity and specificity at the exon level, $S_{\rm nep}$ – exon sensitivity, including partially correct predicted exons (in %). Ov is including overlapped exons.

known proteins or ESTs (Gelfand *et al.*, 1996; Krogh, 2000). Fgenesh+ (Salamov and Solovyev, 2000) is a version of Fgenesh which uses additional information from the available protein homologs. When exons initially predicted by Fgenesh show high similarity to a protein from the database, it is often advantageous to use this information to improve the accuracy of prediction. Fgenesh+ requires an additional file with protein homolog, and aligns all predicted potential exons with the protein homolog using own alignment algorithm. To decrease the computational time, all overlapped exons in the same reading frame are combined into one sequence and align only once.

The main additions to the algorithm, relative to Fgenesh, include:

1. Augmentation of the scores of exons with detected similarity by an additional term proportional to the alignment score.
2. An additional penalty included for the adjacent exons in the dynamic programming (Viterbi algorithm), if the corresponding aligned protein segments are not close in the corresponding protein.

Fgenesh+ was tested on the selected set of 61 GenBank human sequences, for which Fgenesh predictions were not accurate (correlation coefficient 0.0 <= CC < 0.90) and which have protein homologs from another organism. The percentage identity between the encoded proteins and their homologs varied from 99 % to 40 %. The prediction accuracy using this set is presented in Table 4.18. The results show that if the alignment covers the whole length of both proteins, then Fgenesh+ usually increases the accuracy relative to Fgenesh and does not depend significantly on the level of identity (for ID> 0 %). This result makes knowledge of proteins from distant organisms valuable for improving the accuracy of gene identification. A similar approach exploiting known EST/cDNA information was implemented in Fgenesh_c program (Salamov and Solovyev, 2000).

4.13.1 Components of Fgenesh++ Gene-prediction Pipeline

Ab initio gene-prediction program such as Fgenesh predicts ~93 % of all coding exon bases and exactly predicts ~80 % of human exons when applied to single gene sequences (Table 4.16). Analysis of multigene, long genomic sequences is a more complicated task. A program can erroneously join neighboring genes or split a gene into two or more. To improve automatic annotation accuracy, we developed a pipeline Fgenesh++, which can take into account available supporting data such as mRNA or homologous protein sequences. Fgenesh++ is a pipeline for automatic, without human modification of results, prediction of genes in eukaryotic genomes. It uses thefollowing sequence analysis software.

Table 4.18 Comparison of accuracy of Fgenesh and Fgenesh+ on the set of 'difficult' human genes with known protein homologs from another organism.

	CG	S_{ne}	S_{pe}	S_{nb}	S_{pb}	*CC*
Fgenesh	0	63	68	86	83	0.74
Fgenesh+	46	82	85	96	98	0.95

The set contains 61 genes and 370 exons. *CG* – percent of correctly predicted genes; S_{ne}, S_{pe}–sensitivity and specificity at the exon level (in %); S_{nb}, S_{pb} – sensitivity and specificity at the base level (in %); *CC* – correlation coefficient.

Fgenesh++ script to execute the pipeline;
Fgenesh: HMM-based *ab initio* gene-prediction program;
Fgenesh+: gene-prediction program that uses homologous protein sequence to improve performance;
Est_map: a program for mapping known mRNAs/ESTs to a genome, producing genome alignment with splice sites identification;
Prot_map: a program for mapping a protein database to genomic sequence.

Est_map can map a set of mRNAs/ESTs to a chromosome sequence. For example, 11 000 full-length mRNA sequences from NCBI reference set were mapped to 52 MB unmasked Y chromosome fragment in ~20 minutes **Est_map** takes into account statistical features of splice sites for more accurate mapping. **Prot_map** uses a genomic sequence and a set of protein sequences as its input data, and reconstructs gene structure based on protein identity or homology, in contrast to a set of unordered alignment fragments generated by Blast (Altschul *et al.*, 1997). The program is very fast and produces gene structures with similar accuracy to those of relatively slow GeneWise program (Birney and Durbin, 2000), but does not require knowledge of protein genomic location. The accuracy of gene reconstruction can be significantly improved further using Fgenesh+ program on output of Prot_map, that is, using a fragment of genomic sequence (where prot_map found a gene) and the cooreponding protein sequence mapped to it.

Comparison of accuracy of gene prediction by *ab initio* Fgenesh and gene prediction with protein support by Fgenesh+ or GeneWise (Birney and Durbin, 2000) and Prot_ map was performed on a large set of human genes with homologous proteins from mouse or drosophila. We can see that Fgenesh+ shows the best performance with mouse proteins (Table 4.19). With Drosophila proteins, *ab initio* gene prediction by Fgenesh works better than GeneWise for all ranges of similarity and Fgenesh+ is the best predictor if similarity is higher than 60 % (Table 4.20).

Table 4.19 Accuracy of human gene prediction using similar mouse proteins.

(a) Similarity of mouse protein > 90 % in 921 sequences *

	Sn_{ex}	Sp_{ex}	Sn_{nuc}	Sp_{nuc}	*CC*	*%CG*
Fgenesh	86.2	88.6	93.9	93.4	0.9334	34
Genwise	93.9	95.9	99.0	99.6	0.9926	66
Fgenesh+	97.3	98.0	99.1	99.6	0.9936	81
Prot_map	95.9	96.9	99.1	99.5	0.9924	73

(b) 80 %¡ similarity of mouse protein < 90 % in 1441 sequences

	Sn_{ex}	Sp_{ex}	Sn_{nuc}	Sp_{nuc}	*CC*	*%CG*
Fgenesh	85.8	87.7	94.0	93.4	0.9334	30
Genewise	92.6	94.1	98.9	99.5	0.9912	58
Fgenesh+	96.8	97.2	99.1	99.5	0.9929	77
Prot_map	93.9	94.1	98.9	99.3	0.9898	60

* Sn_{ex}, Sensitivity on exon level (exact exon predictions); Sno_{ex}, sensitivity with exon overlap; Sp_{ex}, specificity, exon level; Sn_{nuc}, sensitivity, nucleotides; Sp_{nuc}, specificity, nucleotides; *CC*, correlation coefficient; *%CG*, percent of genes predicted completely correctly (no missing and no extra exons, and all exon boundaries are predicted exactly correctly).

Table 4.20 Accuracy of gene prediction using similar Drosphilla pfroteins.

(a) Similarity of Drosophila protein > 80 % – 66 sequences.

	Sn_{ex}	Sp_{ex}	Sn_{nuc}	Sp_{nuc}	CC	CG %
Fgenesh	90.5	95.1	97.9	96.9	0.950	55
Genewise	79.3	86.8	97.3	99.5	0.985	23
Fgenesh+	95.1	97.0	98.9	99.5	0.9914	70
Prot_map	86.4	88.1	97.6	99.0	0.982	41

(b) 60 % < similarity of Drosophila protein < 80 % – 290 sequences.

	Sn_{ex}	Sp_{ex}	Sn_{nuc}	Sp_{nuc}	CC	CG %
Fgenesh	88.6	90.8	94.9	93.8	0.941	34
Genewise	76.3	82.9	92.8	99.4	0.959	7
Fgenesh+	89.2	92.7	95.5	98.5	0.968	44
Prot_map	75.1	74.9	91.4	97.5	0.941	10

In addition to the programs listed above, Fgenesh++ package also includes files with gene-finding parameters for specific genome, configuration files for programs and a number of Perl scripts. In addition, Fgenesh++ pipeline uses the following public software and data: BLAST executables blastall and bl2seq (Altschul *et al.*, 1997), NCBI NR database (nonredundant protein database) formatted for BLAST, and NCBI RefSeq database (Pruitt *et al.*, 2005).

Fgenesh++ analyzes genome sequences and, optionally, same sequences with repeats masked by N. Sequences can be either complete chromosomes or their fragments such as scaffolds, contigs, etc. When preparing repeats-masked sequences, it is recommended not to mask low complexity regions and simple repeats, as they can be parts of coding sequences.

There are three main steps in running the pipeline:

1. mapping known mRNAs/cDNAs (e.g. from RefSeq) to genomic sequences;
2. prediction of genes based on homology to known proteins (e.g. from NR);
3. *ab initio* gene prediction in regions having neither mapped mRNAs nor genes predicted on the basis of protein homology.

The output of the pipeline consists of predicted gene structures and corresponding proteins. It also indicates whether particular gene structure was assigned on the basis of mRNA mapping, protein homology, or *ab initio* gene prediction.

4.14 GENOME ANNOTATION ASSESSMENT PROJECT (EGASP)

NHGRI (The National Human Genome Research Institute) has initiated the ENCODE project to discover all human genome functional elements (The ENCODE Project Consortium, 2004). Its pilot phase is focused on performance evaluation of different techniques of genome annotation, including computational analysis, on a specified 30 MB of human genome sequence. The community experiment (EGASP05) was organized (Guigo and Reese, 2005) to evaluate how well automatic annotation methods are able to reproduce manual annotations. The best performance in most categories has been demonstrated by predictors that used the most sources of available information. Some of them included conservation of corresponding coding regions in several available genomes: Augustus (Stanke *et al.*, 2006), Jigsaw (Allen *et al.*, 2006) and Paragon (Arumugam *et al.*, 2006). The sensitivity of Fgenesh++ pipeline (which uses one genome information) is similar with them, but the above multigenome programs demonstrated better specificity (Guigo *et al.*, 2006). Performance of Fgenesh++ pipeline for mRNA or protein-supported predictions and *ab initio* predictions (in sequence regions where similar mRNA/protein were not found) is presented in Table 4.21 (Solovyev *et al.*, 2006). All the above-mentioned pipelines demonstrate ~90–95 % sensitivity on the nucleotide level and 75–80 % on the exact exon prediction level. They can exactly predict ~70 % genes (when we count at least one transcript per locus predicted exactly) and ~50 % of all annotated transcripts. No annotation strategy produces perfect gene predictions even using a lot of supportive information that is available for human genome. It is worthwhile to note that the human genes are the most difficult to predict (due to regular occurrences ofvery short exons and very long intron sequences), while the accuracy on simpler organisms is usually much better. While human genes present the most difficult case, the other sequenced genomes have much less available experimental information.

4.15 ANNOTATION OF SEQUENCES FROM GENOME SEQUENCING PROJECTS

Knowledge of gene sequences has opened a new way of performing biological studies called *functional genomics* and the major challenge is to find out what all the newly discovered genes do, how they interact and how they are regulated (Wadman, 1998). Comparisons between genes from different genomes can provide additional insights into the details of gene structure and function.

The successful completion of the Human Genome Project has demonstrated that large-scale sequencing projects can generate high-quality data at a reasonable cost. In addition to human genome, researchers have already sequenced the genomes of a number of important model organisms that are commonly used as test beds in studying human biology. These are the chimpanzee, the mouse, the rat, two puffer fish, two fruit flies, two sea squirts, two roundworms and baker's yeast. Recently, sequencing centers completed working drafts of the chicken, the dog, the honey bee, the sea urchin and a set of four

Table 4.21 Performance data for annotating 44 ENCODE sequences by either mRNA and protein-supported predictions or by *ab initio* predictions.

	mRNA + protein- supported, Sn and Sp (%)	*Ab initio*, Sn and Sp (%)
nucleotide level	91.14	88.44
	89.54	74.46
CDS EXACT	77.19	67.54
	86.48	64.22
CDS OVERLAP	90.60	85.00
	91.4	71.71

fungi. A variety of other genomes are currently in the sequencing pipelines. Many new genomes lack such rich experimental information as the human genome, and therefore their initial computational annotation is even more important as a starting point for further research to uncover their biology. The more comprehensive and accurate such computational analysis is performed, the less time consuming and costly experimental work will have to be done to determine all functional elements in new genomes. Using computational predictions, the scientific community can get at least partial knowledge of majority of real genes because usually gene-finding programs correctly predict most exons of each gene. Fgenesh++ gene-prediction software has been used in annotation of dozen new genomes such as human, rice, Medicago, silkworm, many yeast genomes, bee and sea urchin (see for example, Sodergren *et al.*, 2006; The Honeybee Genome Sequencing Consortium, 2006). Annotation of many genomes is quite a complex procedure. For example, five gene lists were combined to produce bee genome master gene set. Three of them present gene predictors from NCBI, Fgenesh++ and ENSEMBL. Two others comprised evolutionary conserved gene set and Drosophila ortologs. These gene sets merged by special procedure (GLEAN), that construct consensus prediction based on combination of evidences provided by the five gene lists. The Glean set of 10 157 genes is considered as based on experimental evidence, the official *ab initio* gene set comprised 15 500 gene models that did not overlap models of the Glean set (The Honeybee Genome Sequencing Consortium, 2006).

4.15.1 Finding Pseudogenes

Pseudogenes prediction can use two types of initial information (Solovyev *et al.*, 2006). One type contains exon–intron structures of annotated genes and their protein sequences for a genome under analysis. To get such information, we can execute a gene-finding pipeline such as Fgenesh++. In this case, we run Prot_ map program with a set of protein sequences to find possible significant genome–protein alignments that do not correspond to a location of a gene for mapped protein. Other type of initial data can be a set of known proteins for a given organism. Having such data, we can restore gene structure of a given protein using Prot_ map program. For each mapped protein, we can select the best scoring mapping and the computed exon/intron structure as the 'parent' gene structure of this protein. If the alignment of a protein with its own parent has obvious internal stop codons or frameshifts, this locus could be included in the list of potential pseudogenes, but we need to keep in mind more trivial explanations like sequencing errors. Such loci cannot be analyzed on their K_a/K_s or checked for intron losses. In any case, for each of two approaches we have a set of protein sequences, their parent gene structures, and protein–genome alignments for further analysis to identify pseudogenes. Most other pseudogene-finding methods do not include gene-finding and rely on the available protein databases (Harrison *et al.*, 2003) or search only for processed pseudogenes (Baren and Brent, 2006). Example of two types of pseudogenes, processed and nonprocessed, and their characteristics are presented in Figures 4.16 and 4.17.

4.15.2 Selecting Potential Pseudogenes

Using genome–protein alignments generated by Prot_ map program, PSF program produces a list of alignments possessing the following properties for each protein.

Alignment vs. protein encoded by the parent gene.

Identity:	83.7 %
Coverage of protein sequence:	93.9 %
Number of internal stop codons:	2
Number of frameshifts:	1
K_a/K_s:	0.484

```
[DD] Sequence: 11931(1), S: 21.993, L:99 C14000887 chr14 2 exon (s) 75425067 - 75425530 ORF: 1 - 297
98 aa, chain + ## BY PROTMAP: gi|18597373|ref|XP_090893.1| similar to 60S acidic ribosomal protein


1  58970658  58970665  58970695  58970725  58970755  58970785  58970815  58970835
nnnnnnn(..)ccgcgcc?[MASVSELACIY*ALILHDDEVTVTEDKINALIKAAGVNIEPF*PGLFAKAtggtcNVNIGSLICSVEAGG
.......(..).......  |||7|||||||0||||||5||||||0||2|||||||||7|||0|||||||.....||||0||||5|0|||
-------(..)-------  MASISELACIYSALILHDNEVTVTEYKIKALIKAAGVNVEPFRPGLFAKAp---aNVNIRSLICNVGAGG
          1         1         1        11        21        31        41        51       58

          58970865  58970889  58970919  58970947  58970956  63811645
          AAP--AEEKKVEAKKEESEDGDDDMRFGLtttcactga]acctctt(..)nnnnnnn
          0||..|||||5||||||0||2||||0|||......... .......(..).......
          PAPaaAEEKKMEAKKEEFEDSDDDMGFGLsd*------ -------(..)-------
         68        78        88        98       100       100
```

Figure 4.16 An example of processed pseudogene.

Alignment vs. protein encoded by the parent gene.

Identity:	86.4 %
Coverage of protein sequence:	97.6 %
Number of internal stop codons:	3
Number of frameshifts:	4
K_a/K_s:	0.594

```
[RD] Sequence: 35522(1), S: 50.463, L:423 C7000711 chr7 3 exon (s) 51197888 - 51195897 ORF: 1 - 1269
     422 aa, chain - ## BY PROTMAP: gi|27481026|ref|XP_209794.1| similar to hypothetical protein DKFZp43


       1  63659329  63659336  63659366  63659385  63659392  63659422  63659452  63659472
       nnnnnnn(..)tacagtc?[PTSASQQILHAQcatctac(..)gtggaccPQAKLPTFQQLLHTQLPPASGLFRPatggggcSFLTTAFP
       .......(..).......  |2|||50|||||.......(..).......||5|0|2|50|022||0||||||||.......||||||||
       -------(..)-----mg  PASASQRTLHAQlala---(..)---slrpPQSKAPAFRPLRQAQLLPASGLFRP------sSFLTTAFP
       1         1         3        13        19        23        33        43        48

63659498  63659528  63659558  63659588  63659618  63659648  63659678  63659708  63659738
       GPVFPFRRPLRAQNLLKSASPDPLAPSGRSLRAQLFFLVGSPGPIPASQQPLWTQCLPISWRPWSAHSFLKPSSPGPGQASRWPLQDELL
       ||7|||5|||5||||||0|||0|||||||0|5||||2022||||0||||||||||||||||||||||||||||||||||||||||||||6||
       GPIFPFQRPLQAQNLLKLASPGPLAPSGRPLQAQLFLPAASPGPTPASQQPLWTQCLPISWRPWSAHSFLKPSSPGPGQASRWPLQDQLL
      57        67        77        87        97       107       117       127       137

63659768  63659798  63659828  63659858  63659888  63659907  63659952  63659971  63660001
       PSDGISRPQMVSGRWAPPRQGWASRRLPQAQVVLKSGSPGPASQQ]gtaagca(..)tttgtag[APNFLQPSSEGPPPASWWPVQF*HW
       ||||7||||||||||||||02|||||00||||||||2|||||||| .......(..)....... |||||||||2||||||0||||000|
       PSDGVSRPQMVSGRWAPPRPAWASRRLPQAQVVLKSASPGPASQQ -------(..)------- APNFLQPSSSGPPPASRWPVQAQLW
     147       157       167       177       187       192       192       197       207

63660031  63660061  63660089  63660119  63660147  63662724  63662731  63662748  63662766
       LENSLCRPRPCLPgGPLQAQLLPPRRPPGAKSLPASQQPgc]gtgcggc(..)tctccag[gPDSGccgactccagVPTTSLDSAPAQLP
       |||||||||0|||.|||||||0||5|||||||||||5||.. .......(..)....... .||||..........5|00||||||||||
       LENSLCRPRSCLP-GPLQAQLSPPQRPPGAKSLPASRQP-- -------(..)------- aPDSG----------LPIRSLDSAPAQLP
     217       227       236       246       255       255       255       260       264

63662796  63662826  63662856  63662884  63662914  63662944  63662974  63663004  63663034
       AALVGPQLP*AKLPRPSSGLAVASPGSAPgALR*HLQAPNGLRSVGSSRPSLGLPAASAGPNRPEVSLSRLSSSLPAASAGPSRPQVGLE
       |||||||||0||||||||||2||||||||.|||0||||||||||||||||||||||||||||||||2|||0||2|||||||0||||||||
       AALVGPQLPEAKLPRPSSGLTVASPGSAP-ALRRHLQAPNGLRSVGSSRPSLGLPAASAGPNRPEVGLSRPSSGLPAASAGLSRPQVGLE
     274       284       294       303       313       323       333       343       353

63663064  63663094  63663124  63663154  63663184  63663214  63663244  63811645
       VGLEEQQVGLPGPSSVLSTASPGAKLPRVSLSRPSSSCLPVASFSPAQLMALGGLRRPCF*]cttttgg(..)nnnnnnn
       |||||0||||||||||||2|||||||||||||||||||||||||2||||||||2|0||0|| .......(..).......
       VGLEELQVGLPGPSSVLSAASPGAKLPRVSLSRPSSSCLPVASFGPAQLMALGSLPRPRF* -------(..)-------
     363       373       383       393       403       413       423       424
```

Figure 4.17 An example of not processed pseudogene.

1. Identity in blocks of alignment exceeds certain value
2. Substantial portion of protein sequence is included in the alignment
3. Genomic location of alignment differs from that of parent gene
4. At least one of four events is observed:

   i. *Damage to ORF.* There is one or more frameshifts or internal stop codons;

   ii. *Single exon with close PolyA site.* PolyA site is too close to a 3′-end of an alignment, while C-terminus of protein sequence is aligned to the last amino acid, and a single exon covers 95 % of protein sequence.

   iii. *Loss of introns.* Protein coverage by alignment is at least 95 %, and a number of exons is fewer than in parent gene by a certain number.

   iv. *Protein sequence is not preserved.* The ratio of nonsynonymous to synonymous replacements exceeds certain threshold ($K_a/K_s > 0.5$). K_a/K_s is calculated relative to a parent gene by method presented by Nei and Gojobori (1986).

4.15.3 Selecting a Reliable Part of Alignment

The procedures described above apply to a so-called reliable part of alignment. Necessity of introducing this concept is caused by imperfections in aligning a protein against a chromosome sequence. There are complex cases where accurate alignment cannot be produced, such as very short (1–3 bp) exons separated by a large intron, or some errors in protein or genome draft sequence that prevent perfect alignment. For instance, if a protein as a whole is well aligned to a chromosome, but ~20 amino acids on its 5′-end cannot be aligned in one continuous block, Prot_ map will most likely try to align these 20 amino acids by scattering them along several short blocks. Most likely, these blocks will not have any relation to a gene or a pseudogene. Therefore, in search for pseudogenes, we remove short insignificant trailer blocks. The rest of alignment is considered as its reliable part. To find a reliable part of alignment, we evaluate the quality of alignment blocks (exons). For each exon found by Prot_map, we calculate the number of aligned amino acid (M), number of nonaligned amino acids (AI) and nucleotides (NI) within an exon, number of aligned amino acids (AO) and nucleotides (NO) located outside of exon region to the left and to the right side of an exon. We also compute the 'correctness' of splice sites conserved dinucleotides (SSC) that flank an exon. If an exon is N- or C-terminal one, we also compute 'correctness' of corresponding start or stop codons. The length of an intron (IL) that separates an exon from nearest exon in the direction of the longest mapped exon is also computed. The empirical 'quality' measure is defined by the following formula:

$$Q = M - P_{\mathrm{AI}}(AI) - P_{\mathrm{NI}}(NI) - P_{\mathrm{AO}}(AO) - P_{\mathrm{NO}}(NO) + \mathrm{B}_{\mathrm{SSC}}(SSC) - \mathrm{P}_{\mathrm{IL}}(IL).$$

Where P_{AI}, P_{NI}, P_{AO}, P_{NO} are the penalties for the internal and external unaligned amino acids and nucleotides, BSSC is a bonus for correctness of splice sites or start/stop codons, and P_{IL} is a penalty for high intron length. The reliable part of alignment consists of neighboring exons alignments that each have $Q > 5$. After Prot_map mapping, many loci on a chromosome include alignments to more than one protein. In such cases, we choose only one most reliable alignment, based on a sum of included exon's qualities.

The PSF (pseudogene finding) approach described above has been applied to identify pseudogenes in 44 ENCODE sequences (Solovyev *et al.*, 2006). As a result, it was found 181 potential pseudogenes, 118 of which had a significant overlap with annotated 145 HAVANA pseudogenes. 68 (58 %) of these 118 pseudogenes had only one exon and could be classified as processed pseudogenes: 58 had the parent gene with more than one exon and seven others had polyA tail. 106 (90 %) of 118 pseudogenes had one or more defects in their ORFs. Among the remaining 12, there are four pseudogenes with a single exon (while their parents have four or more exons), four contain both polyA signal and polyA tract, four have only a polyA tract, and two have only high K_a/K_s ratios (0.59 and 1.04). The PSF has not found 27 HAVANA annotated pseudogenes. Three of them were not reported because they are located in introns of larger pseudogenes (AC006326.4-001, AC006326.2-001 and AL162151.3-001). The other ten represent fragments of some human proteins and are missing stop codons or frameshifts. We did not include pseudogenes corresponding to fragments of proteins in our pseudogene set. The remaining 14 HAVANA pseudogenes were not found probably due to some limitation of our program and the used datasets of predicted genes and known proteins. Missed pseudogenes might have parent genes that were absent from our initial protein set compiled by Fgenesh++ gene-prediction pipeline. Some of 63 pseudogenes that have been predicted by PSF but were absent from HAVANA set might have appeared because of imperfect predictions by the pipeline, which produced frameshifts when a pseudogene candidate and its parent gene were aligned. However, some of these 'over-predicted' pseudogenes might be actual pseudogenes missed by HAVANA annotators (for example, see Figure 4.18).

To summarize, PSF pseudogene prediction program has found 81 % of annotated pseudogenes. Its quality can further be improved by improving the quality of parent gene/protein sets.

4.16 CHARACTERISTICS AND COMPUTATIONAL IDENTIFICATION OF miRNA GENES

MicroRNA (miRNA) are a class of small (~22 nt), noncoding RNAs that can regulate gene expression by directing mRNA degradation or inhibiting productive translation (Mallory and Vaucheret., 2004). They are sequence-specific regulators of posttranscriptional gene expression in many eukaryotes. Some components of miRNA machinery have been found even in archaea and eubacteria, revealing their very ancient origination. They are believed to control the expression of thousands of target mRNAs, with each mRNA possible targeted by multiple miRNAs (Pillai, 2005). miRNA discovery by molecular cloning has been supplemented by computational approaches that identify evolutionary conserved miRNA genes by searching for patterns of sequence and secondary structure conservation These approaches indicate that miRNA constitute nearly 1–3 % of all identified genes in nematodes, flies and mammalians (Jones-Rhoades and Bartel, 2004). Only in humans the latest miRNA count exceeded 800 genes (Pillai, 2005). The first two miRNA genes (lin-4 and let-7) were discovered in *C. elegans*, where their mutations cause defects in the temporal regulation of larval stage-specific programs of cell divisions. These miRNAs affect by base pairing to partially complementary sites in the 3′ untranslated region (UTR) of their target mRNAs and repressing their translation (Lee *et al.*, 1993; Reinhart *et al.*, 2000).

```
[DD] Sequence: 622(1), S: 27.323, L:153 C6000781 chr6 6 exon (s) 840966 - 845318 ORF: 1 - 459  152 aa,
chain + ## gi|6755368|ref|NP_035426.1| ribosomal protein S18 [Mus musculus] gi|11968182|ref ## 152


 1    151509    151516    151546    151576    151606    151636    151664    151694    151724
 caaannn(..)tcctgct?[MSLVIPEKFQRILRILNSNINGQQKIGFAITAIKDVG*QYTHaVLRKADVDLTKWAGELTEDEMERVMTIM
 .......(..).......  ||||||||||2|||7||5||5|55||2|||||||0||05|2|.||||||7||||0||||||||5|||5|||
 -------(..)-------  MSLVIPEKFQHILRVLNTNIDGRRKIAFAITAIKGVGRRYAHvVLRKADIDLTKRAGELTEDEVERVITIM
 1         1         1         11        21        31        41        51        61        71

   151754    151784    151814    151844    151874    151904    151934    151964
QNPCQYKIPDWFLNRRKDVKDGKYSQVLASGLDKKLRADVERLKKIQAHRGPHHFWGLRVRGQHTKTTGHHGCTMGGSKKK*]gtctgca(..)aaaataa
|||0|||||||||||5|||||||||||||5|||2|||0|5||||||5||||02||||||||||||||||22|0|5|0|||| .......(..).......
QNPRQYKIPDWFLNRQKDVKDGKYSQVLANGLDNKLREDLERLKKIRAHRGLRHFWGLRVRGQHTKTTGRRGRTVGVSKKK* -------(..)-------
        81        91       101       111       121       131       141       151
```

Figure 4.18 A pseudogene in ENm004 sequence that is absent in the manual HAVANA annotation. The alignment has a stop codon close to position 151636.

The majority of miRNA genes are located in intergenic regions or in antisence orientation to annotated genes, indication that they for independent transcriptional units. Most of the other miRNA genes are found in intronic regions, which may be transcribed as part of the annotated gene. Independent miRNA genes are initially transcribed by RNA PolII (Lee *et al.*, 2004) as part of a long primary transcript, which contain the mature miRNA as part of a predicted RNA hairpin. This transcript is cropped into the hairpin-shaped pre-miRNAs by nuclear RNaseIII Drosha (Lee *et al.*, 2003). The hairpin RNAs of approximately 70 nt bearing the 2-nucleotide 3′ overhang are exported to the cytoplasm by a Ran dependent nuclear transport receptor family. Once in the cytoplasm, pre-miRNAs are subsequently cleaved by cytoplasmic RNase III Dicer into ~22 nt miRNA duplex, one strand of which is degraded by a nuclease, while the other strand remains as a mature miRNA (Lee *et al.*, 2004; Denli *et al.*, 2004). A typical structure of miRNA gene and its processing is presented in Figure 4.19.

Despite the plenty of miRNAs that have identified from cloning, such technique is likely to be far from saturated, as it is biased to abundant miRNA. Therefore, computational approaches have been developed that predict miRNAs encoded in animal and plant genomes (Grad *et al.*, 2003; Jones-Rhoades and Bartel, 2004; Ohler *et al.*, 2004). There are several variations of these methods: one is based on analysis of sequence and secondary structure properties of typical pre-miRNA. However, the short length and high degree of sequence and structure variation limit the accuracy of computational predictions based on such characteristics along. To decrease the number of false-positive predictions, the candidate miRNAs are selected to be conserved across species (the presence in two or more genomes of very similar sequences embedded in the same stems of predicted hairpins). A flowchart of computational selection miRNA candidates for plant miRNA predictions is presented in Figure 4.20 (Jones-Rhoades and Bartel, 2004). Another algorithm is based on the search for possible homologs (including hairpin selection and Smith–Waterman sequence alignment) of a few hundreds of known miRNAs cloned from *C. elegans, D. melanogaster, M. musculus and H. sapiens* (for identification of miRNAs in animal genomes) (Grad *et al.*, 2003). Recently, using similar approaches the FindmiRNA and the TargetmiRNA programs were developed

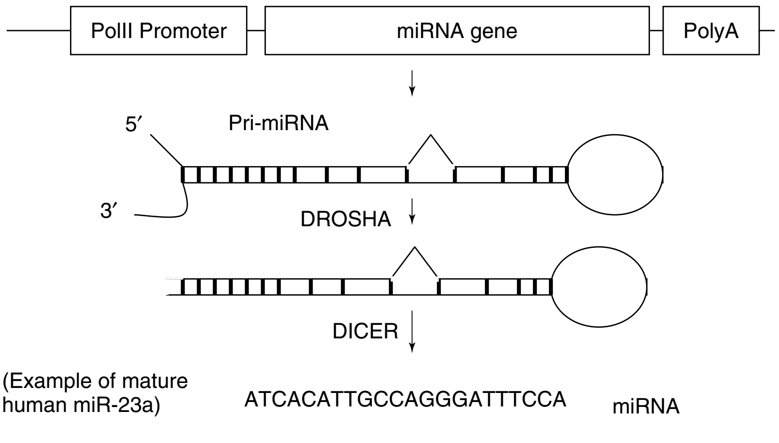


Figure 4.19 A model of expression of miRNA gene and processing of miRNA.

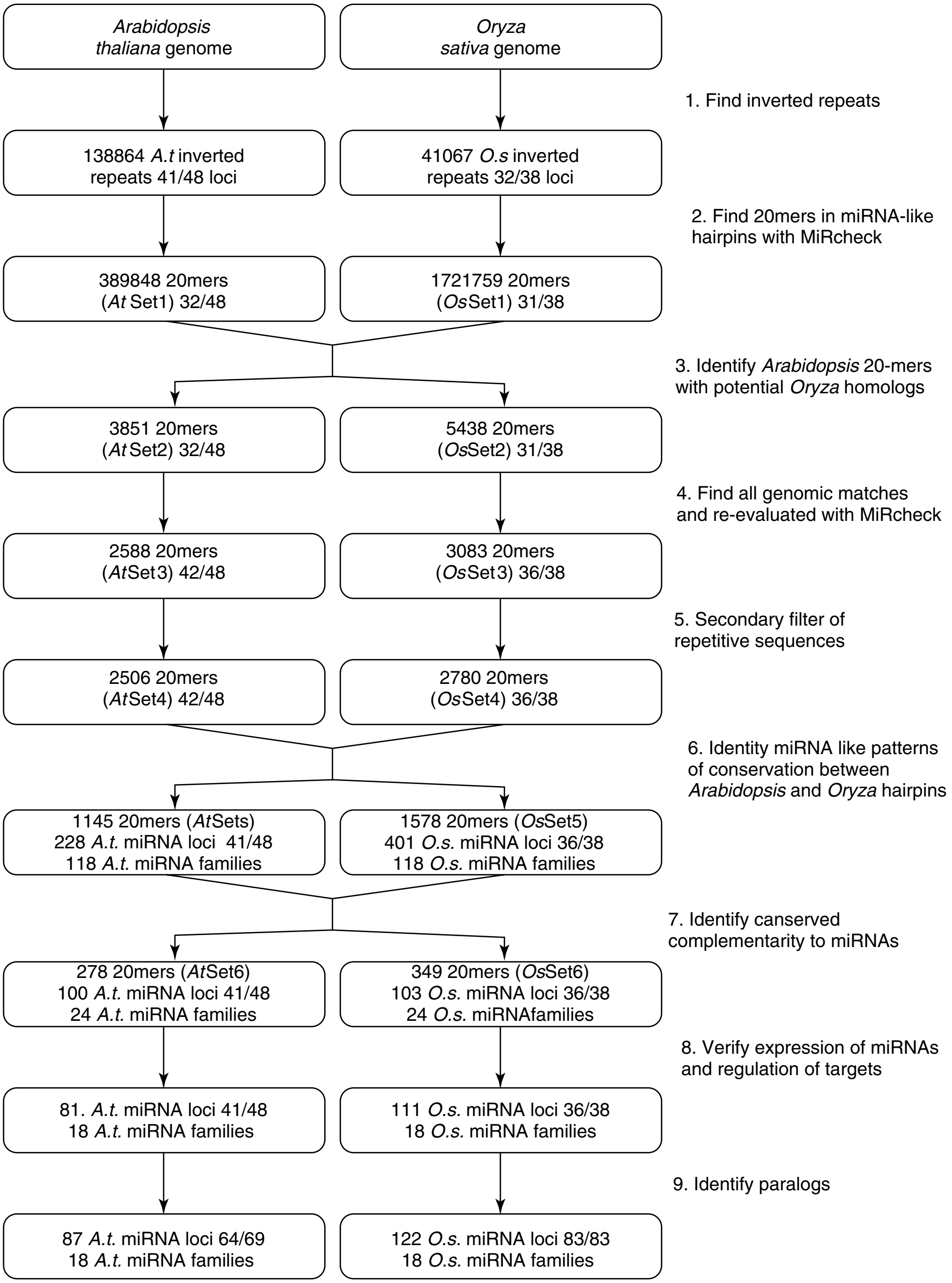
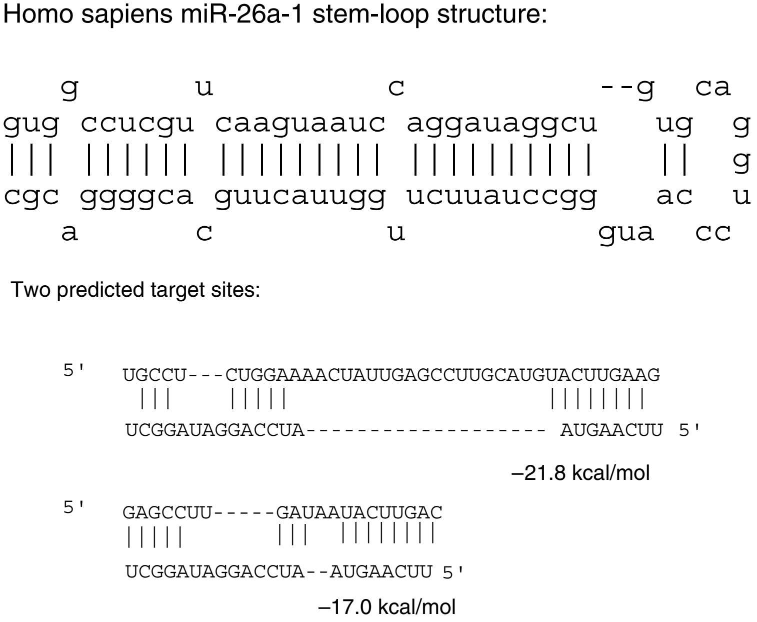

Arabidopsis thaliana genome
Oryza sativa genome
1. Find inverted repeats
138864 *A.t* inverted repeats 41/48 loci
41067 *O.s* inverted repeats 32/38 loci
2. Find 20mers in miRNA-like hairpins with MiRcheck
389848 20mers (*At* Set1) 32/48
1721759 20mers (*Os*Set1) 31/38
3. Identify *Arabidopsis* 20-mers with potential *Oryza* homologs
3851 20mers (*At* Set2) 32/48
5438 20mers (*Os*Set2) 31/38
4. Find all genomic matches and re-evaluated with MiRcheck
2588 20mers (*At*Set3) 42/48
3083 20mers (*Os*Set3) 36/38
5. Secondary filter of repetitive sequences
2506 20mers (*At*Set4) 42/48
2780 20mers (*Os*Set4) 36/38
6. Identity miRNA like patterns of conservation between *Arabidopsis* and *Oryza* hairpins
1145 20mers (*At*Sets) 228 *A.t.* miRNA loci 41/48 118 *A.t.* miRNA families
1578 20mers (*Os*Set5) 401 *O.s.* miRNA loci 36/38 118 *O.s.* miRNA families
7. Identify canserved complementarity to miRNAs
278 20mers (*At*Set6) 100 *A.t.* miRNA loci 41/48 24 *A.t.* miRNA families
349 20mers (*Os*Set6) 103 *O.s.* miRNA loci 36/38 24 *O.s.* miRNAfamilies
8. Verify expression of miRNAs and regulation of targets
81. *A.t.* miRNA loci 41/48 18 *A.t.* miRNA families
111 *O.s.* miRNA loci 36/38 18 *O.s.* miRNA families
9. Identify paralogs
87 *A.t.* miRNA loci 64/69 18 *A.t.* miRNA families
122 *O.s.* miRNA loci 83/83 18 *O.s.* miRNA families

Figure 4.20 Flowchart of the miRNA prediction approach using two plant genomes. (Reprinted from Jones-Rhoades, M.W. and Bartel, D.P. (2004) Computational identification of plant microRNAs and their targets, including a stressinduced miRNA. *Mol. Cell* **14**: 787-799, with permission form Elsevier.)

to search for miRNA and their targets in sequences of a range of model eukaryotic organisms (`http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=rnastruct`).

4.17 PREDICTION OF microRNA TARGETS

While hundreds of miRNAs have been deposited in the databases, their regulatory targets have not been established or predicted for many of them. Finding regulatory targets for plant miRNA is simply performed by looking for near-perfect complementarity to the mRNAs. For example, in a search for the targets of 13 *Arabodopsis* miRNA families, 49 unique targets were found with just a few false predictions (Rhoades *et al.*, 2002). However, animal miRNA targets have complementarity to the miRNAs only in the 'seed' sequence (usually 2–8 nucleotides numbered from the 5′ end) and often have multiple regions of complementarity, therefore more sophisticated search methods considering these features have recently been published (Stark *et al.*, 2003; Enright *et al.*, 2003; Lewis *et al.*, 2003; Rehmsmeier *et al.*, 2004). In general, miRNA, target genes are selected on the basis of three properties: sequence complementarity using a position-weighted local alignment algorithm, free energies of RNA–RNA duplexes, and conservation of target sites in related genomes. Lewis *et al.* (2003) in their TargetScan software took into account multiple miRNA–mRNA UTR complementary regions summing Z-scores (exp(-G/T)) produced by each such region in evaluating a potential target mRNA, where G



Homo sapiens miR-26a-1 stem-loop structure:

```
   g      u         c          --g  ca
gug ccucgu caaguaauc aggauaggcu   ug  g
||| |||||| ||||||||| ||||||||||   ||  g
cgc ggggca guucauugg ucuuauccgg   ac  u
   a      c         u          gua  cc
```

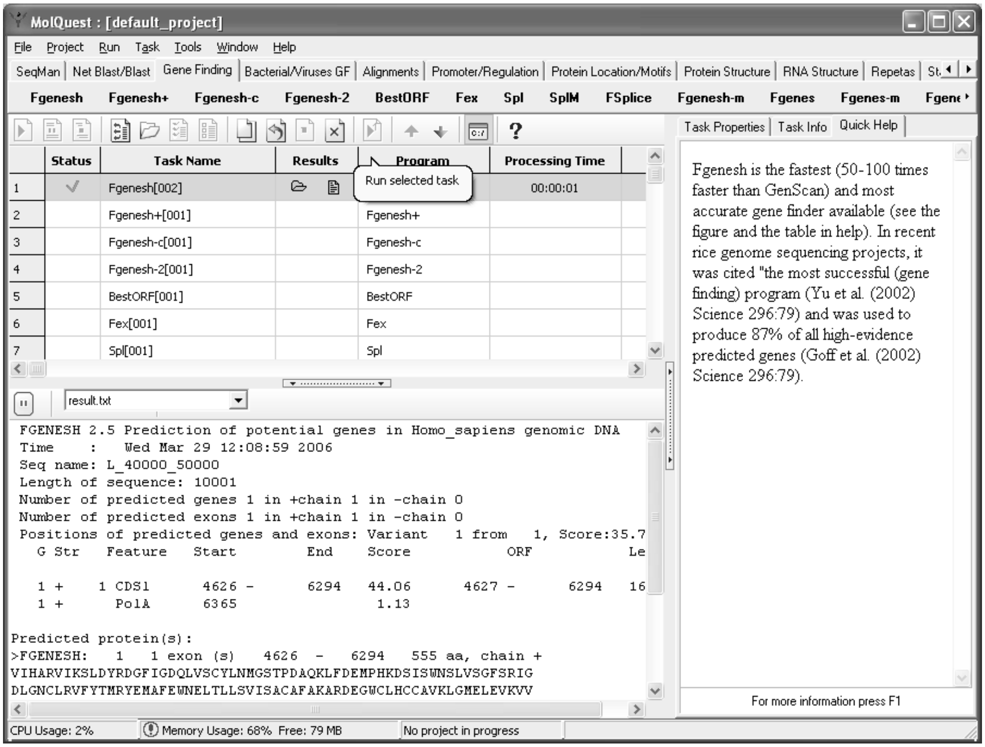
Two predicted target sites:

```
5'  UGCCU---CUGGAAAACUAUUGAGCCUUGCAUGUACUUGAAG
     |||    |||||                    ||||||||
    UCGGAUAGGACCUA------------------- AUGAACUU 5'

                               –21.8 kcal/mol

5'  GAGCCUU-----GAUAAUACUUGAC
    |||||       |||  ||||||||
    UCGGAUAGGACCUA--AUGAACUU 5'

                  –17.0 kcal/mol
```

Figure 4.21 An example of stem-loop structure and predicted target sites for miR26a in human SMAD1 gene.

Table 4.22 Web server software for eukaryotic gene and functional signals prediction.

Program/task	WWW address
Fgenesh/HMM-based gene prediction (Human, Drosophila, Dicots, Monocots, *C.elegans*, *S. pombe* and etc.)	http://sun1.softberry.com/berry.phtml?topic= fgenesh &group=programs&subgroup=gfind
Genscan/HMM-based gene prediction (Human, Arabidipsis, Maize)	`http://genes.mit.edu/GENSCAN.html`
HMM-gene/HMM-based gene prediction (Human, C.elegans)	`http://www.cbs.dtu.dk/services/HMMgene/`
Fgenes/Disciminative gene prediction (Human)	`http://sun1.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind`
Fgenesh-M/Prediction of alternative gene structures (Human)	`http://sun1.softberry.com/berry.phtml?topic=fgenesh-m&group=programs&subgroup=gfind`
Fgenesh+/Fgenesh_c/ gene prediction with the help of similar protein/EST	`http://sun1.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfind`
Fgenesh-2/gene prediction using 2 sequences of close species	`http://sun1.softberry.com/berry.phtml?topic=fgenes_c&group=programs&subgroup=gfs`
BESTORF/Finding best CDS/ORF in EST (Human, Plants, Drosophila)	`http://sun1.softberry.com/berry.phtml?topic=bestorf&group=programs&subgroup=gfind`
FgenesB/gene, operon, promoter and terminator prediction in bacterial sequences	`http://sun1.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfindb`
Mzef/internal exon prediction (Human, Mouse, Arabidopsis, Yeast	`http://rulai.cshl.org/tools/genefinder/`
FPROM/TSSP/ promoter prediction	`http://sun1.softberry.com/berry.phtml?topic=index&group=programs&subgroup=promoter`
NSITE/search for functional motifs	
Promoter 2.0/promoter prediction	`http://www.cbs.dtu.dk/services/Promoter/`
CorePromoter/promoter prediction	`http://rulai.cshl.org/tools/genefinder/CPROMOTER/index.htm`
SPL/SPLM/splice-site prediction (Human, Drosophila, Plants nd etc.)	`http://www.softberry.com/berry.phtml?topic=spl&group=programs&subgroup=gfind`
NetGene2/NetPGene/splice-site prediction (Human, C.elegans, Plants)	`http://www.cbs.dtu.dk/services/NetPGene/`
Scan2 searching for similarity in genomic sequences and its visualization	`http://sun1.softberry.com/berry.phtml?topic=scan2&group=programs&subgroup=scanh`
RNAhybrid prediction of microRNA target duplexes	`http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/`


MolQuest : [default_project]
File Project Run Task Tools Window Help
SeqMan Net Blast/Blast Gene Finding Bacterial/Viruses GF Alignments Promoter/Regulation Protein Location/Motifs Protein Structure RNA Structure Repetas St
Fgenesh Fgenesh+ Fgenesh-c Fgenesh-2 BestORF Fex Spl SplM FSplice Fgenesh-m Fgenes Fgenes-m Fgene
Task Properties Task Info Quick Help
Run selected task
result.txt
For more information press F1
CPU Usage: 2%
Memory Usage: 68% Free: 79 MB
No project in progress

	Status	Task Name	Results	Program	Processing Time
1	✓	Fgenesh[002]			00:00:01
2		Fgenesh+[001]		Fgenesh+	
3		Fgenesh-c[001]		Fgenesh-c	
4		Fgenesh-2[001]		Fgenesh-2	
5		BestORF[001]		BestORF	
6		Fex[001]		Fex	
7		Spl[001]		Spl	

```
 FGENESH 2.5 Prediction of potential genes in Homo_sapiens genomic DNA
 Time    :   Wed Mar 29 12:08:59 2006
 Seq name: L_40000_50000
 Length of sequence: 10001
 Number of predicted genes 1 in +chain 1 in -chain 0
 Number of predicted exons 1 in +chain 1 in -chain 0
 Positions of predicted genes and exons: Variant   1 from   1, Score:35.7
   G Str   Feature   Start        End    Score           ORF            Le

   1 +    1 CDS1      4626 -      6294   44.06      4627 -      6294   16
   1 +      PolA      6365                1.13

Predicted protein(s):
>FGENESH:   1   1 exon (s)   4626  -   6294   555 aa, chain +
VIHARVIKSLDYRDGFIGDQLVSCYLNMGSTPDAQKLFDEMPHKDSISWNSLVSGFSRIG
DLGNCLRVFYTMRYEMAFEWNELTLLSVISACAFAKARDEGWCLHCCAVKLGMELEVKVV
```

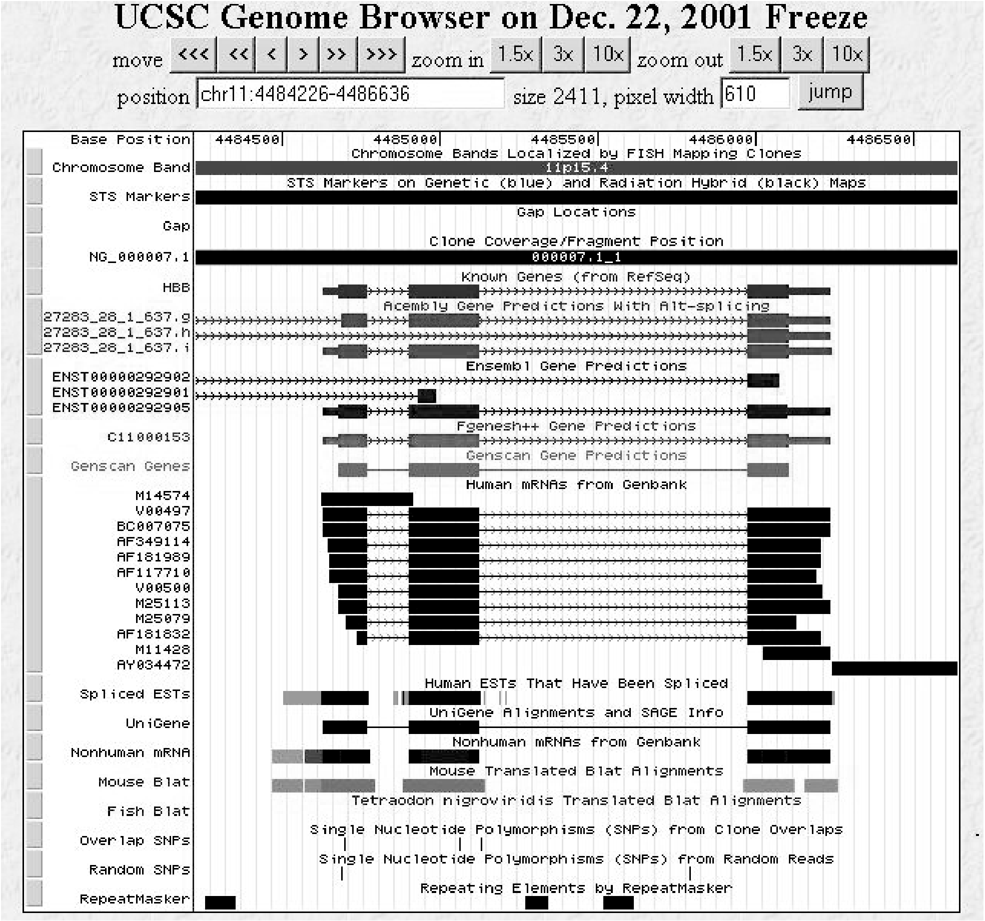
Fgenesh is the fastest (50-100 times faster than GenScan) and most accurate gene finder available (see the figure and the table in help). In recent rice genome sequencing projects, it was cited "the most successful (gene finding) program (Yu et al. (2002) Science 296:79) and was used to produce 87% of all high-evidence predicted genes (Goff et al. (2002) Science 296:79).

Figure 4.22 A user interface of **MolQuest** comprehensive desktop package for gene finding, sequence analysis and molecular biology data management.

is the free energy of miRNA:target site interaction. An example of stem-loop structure and predicted target sites for miR26a in human *SMAD1* gene is presented in Figure 4.21. Using TargetScan ~400 regulatory target genes have been predicted for the conserved vertebrate miRNAs. Eleven predicted targets (out of 15 tested) were supported experimentally (Lee *et al.*, 2003).

4.18 INTERNET RESOURCES FOR GENE FINDING AND FUNCTIONAL SITE PREDICTION

Prediction of genes, ORF, promoter, splice sites finding by the methods described in the preceding text is mostly available via World Wide Web. Table 4.22 presents WEB addresses of some of them. Many of these programs can be used within window-based **Molquest** computer package (www.molquest.com). It is the most comprehensive, easy-to-use desktop application for desktop sequence analysis (see Figure 4.22). The package includes gene-finders family (fgenesh/fgenesh+) programs for many organisms as well as pipelines (fgenesh++ and fgeneshb_annotator) that often used for fully automatic annotation eukaryotic and bacterial genomes (or genome



UCSC Genome Browser on Dec. 22, 2001 Freeze

position chr11:4484226-4486636 size 2411, pixel width 610

move <<< << < > >> >>> zoom in 1.5x 3x 10x zoom out 1.5x 3x 10x
jump
Base Position
4484500
4485000
4485500
4486000
4486500
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
11p15.4
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
Clone Coverage/Fragment Position
NG_000007.1
000007.1_1
Known Genes (from RefSeq)
HBB
Acembly Gene Predictions With Alt-splicing
27283_28_1_637.g
27283_28_1_637.h
27283_28_1_637.i
Ensembl Gene Predictions
ENST00000292902
ENST00000292901
ENST00000292905
Fgenesh++ Gene Predictions
C11000153
Genscan Gene Predictions
Genscan Genes
Human mRNAs from Genbank
M14574
V00497
BC007075
AF349114
AF181989
AF117710
V00500
M25113
M25079
AF181832
M11428
AY034472
Human ESTs That Have Been Spliced
Spliced ESTs
UniGene Alignments and SAGE Info
UniGene
Nonhuman mRNAs from Genbank
Nonhuman mRNA
Mouse Translated Blat Alignments
Mouse Blat
Tetraodon nigroviridis Translated Blat Alignments
Fish Blat
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Repeating Elements by RepeatMasker
RepeatMasker

Figure 4.23 A screenshot of UCSC Genome Browser displaying gene predictions computed by various approaches.

communities) (The Honeybee Genome Sequencing Consortium, 2006; Tyson *et al.*, 2004). The package provides a user-friendly interface for sequence editing, primer design, internet database searches, gene prediction, promoter identification, regulatory elements mapping, patterns discovery protein analysis, multiple sequence alignment, phylogenetic reconstruction, and a wide variety of other functions. A lot of information generated during new genomes annotations (including gene predictions) is available through various genome browsers. A screenshot of popular UCSC Genome Browser (`http://genome.ucsc.edu/`) is presented in Figure 4.23. Another such interactive tool Genome Explorer (`http://sun1.softberry.com/berry.phtml?topic=human&group=genomexp`) can show a graph of expression data for selected genes (Figure 4.24). Its version for annotations of bacterial genomes is demonstrated in Figure 4.25. These web browsers provide search of numerous genome elements, visualization and retrieval of gene and protein sequences and fast comparison with auser-provided

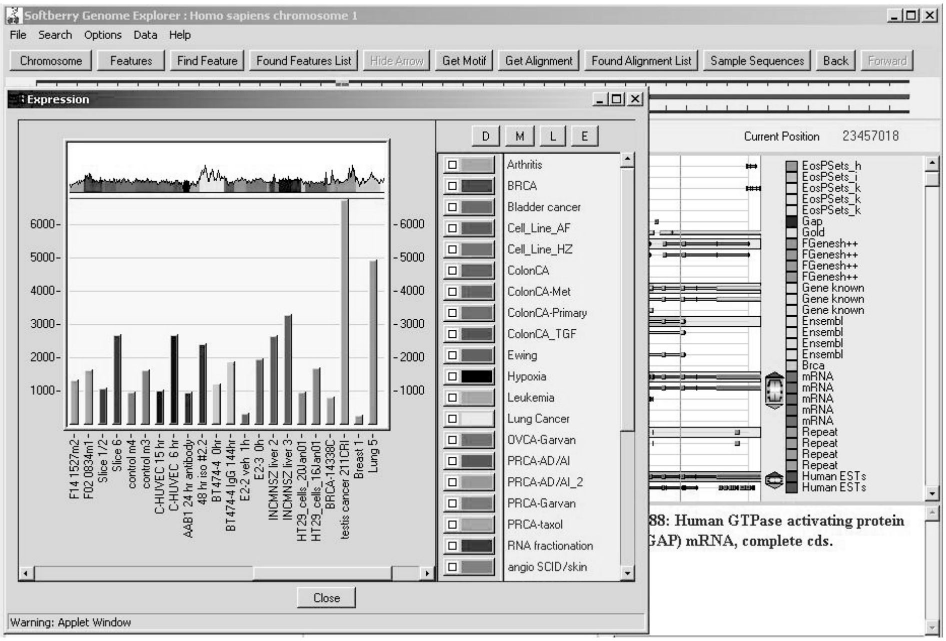
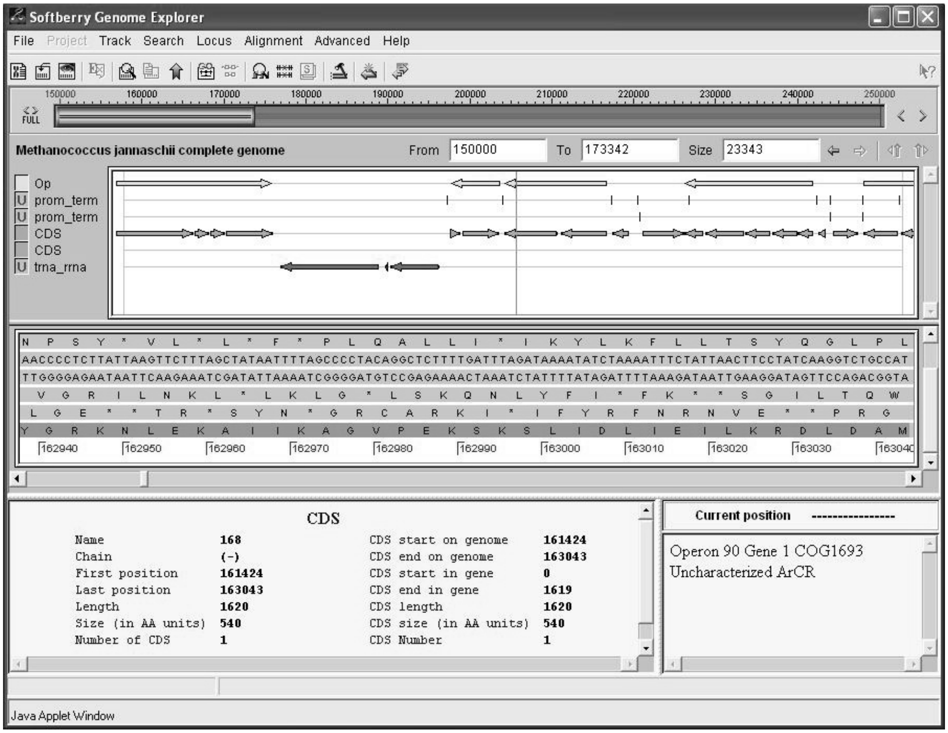


Figure 4.24 The Genome Explorer annotation browser showing graph of expression data for selected genes.


Softberry Genome Explorer
File Project Track Search Locus Alignment Advanced Help
150000
160000
170000
180000
190000
200000
210000
220000
230000
240000
250000
FULL
Methanococcus jannaschii complete genome
From 150000
To 173342
Size 23343
Op
prom_term
prom_term
CDS
CDS
trna_rrna
Current position ----------------
Java Applet Window

```
N  P  S  Y  *  V  L  *  L  *  F  *  P  L  Q  A  L  L  I  *  I  K  Y  L  K  F  L  L  T  S  Y  Q  G  L  P  L
AACCCCTCTTATTAAGTTCTTTAGCTATAATTTTAGCCCCTACAGGCTCTTTTGATTTAGATAAAATATCTAAAATTTCTATTAACTTCCTATCAAGGTCTGCCAT
TTGGGGAGAATAATTCAAGAAATCGATATTAAAATCGGGGATGTCCGAGAAAACTAAATCTATTTTATAGATTTTAAAGATAATTGAAGGATAGTTCCAGACGGTA
  V  G  R  I  L  N  K  L  *  L  K  L  G  *  L  S  K  Q  N  L  Y  F  I  *  F  K  *  *  S  G  I  L  T  Q  W
 L  G  E  *  *  T  R  *  S  Y  N  *  G  R  C  A  R  K  I  *  I  F  Y  R  F  N  R  N  V  E  *  *  P  R  G
Y  G  R  K  N  L  E  K  A  I  I  K  A  G  V  P  E  K  S  K  S  L  I  D  L  I  E  I  L  K  R  D  L  D  A  M
  162940    162950    162960    162970    162980    162990    163000    163010    163020    163030    16304C
```

CDS

Name	**168**	CDS start on genome	**161424**
Chain	**(-)**	CDS end on genome	**163043**
First position	**161424**	CDS start in gene	**0**
Last position	**163043**	CDS end in gene	**1619**
Length	**1620**	CDS length	**1620**
Size (in AA units)	**540**	CDS size (in AA units)	**540**
Number of CDS	**1**	CDS Number	**1**

Operon 90 Gene 1 COG1693
Uncharacterized ArCR

Figure 4.25 The Bacterial genome explorer displaying predicted operons, genes, promoters and terminators.

sequences. They are actively used not only by academic research community but also by many drug discovery and biotechnology companies for identification of drug candidates.

Acknowledgments

I would like to gratefully acknowledge collaboration with Asaf Salamov who produced several gene-finding algorithms and many results of this chapter. The paragraph about analysis of canonical and noncanonical splice sites presents our work with Moises Burset and Igor Seledtsov. Peter Kosarev and Oleg Fokin actively participated in the development of Fgenesh++ pipeline as well as in Moquest package development.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D., Scheeler, F., Shen, H., Shue, B.C., Siden-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., WoodageT, Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., Venter and J.C. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Afifi, A.A. and Azen, S.P. (1979). *Statistical Analysis. A Computer Oriented Approach*. Academic Press, New York.

Allen, J.E., Majoros, W.H., Pertea, M. and Salzberg, S.L. (2006). JIGSAW, GeneZilla and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biology* **7**(Suppl. 1), S9.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17), 3389–3402.

Arumugam, M., Wei, C., Brown, R.H. and Brent, M.R. (2006). Pairagon+N-SCAN_EST: a model-based gene annotation pipeline. *Genome Biology* **7**(Suppl. 1), S5.1–S5.10.

Audic, S. and Claverie, J. (1997). Detection of eukaryotic promoters using Markov transition matrices. *Computers and Chemistry* **21**, 223–227.

Bajic, V., Brent, M., Brown, R., Frankish, A., Harrow, J., Ohler, U., Solovyev, V. and Tan, S. (2006). Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biology* **7**(Suppl. 1), S3.1–S3.13.

Baren, M. and Brent, M. (2006). Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Research* **16**, 678–685.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A. and Wheeler, D.L. (1999). GenBank. *Nucleic Acids Research* **27**(1), 12–17.

Berg, O.G. and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology* **193**, 723–750.

Birney, E. and Durbin, R. (2000). Using GeneWise in the *Drosophila* annotation experiment. *Genome Research* **10**, 547–548.

Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993). dbEST–database for "expressed sequence tags". *Nature Genetics* **4**(4), 332–333.

Borodovsky, M. and McIninch, J. (1993). GENMARK: parallel gene recognition for both DNA strands. *Computers and Chemistry* **17**, 123–133.

Borodovskii, M., Sprizhitskii, Yu., Golovanov, E. and Alexandrov, N. (1986). Statistical patterns in the primary structures of functional regions of the genome in *Escherichia coli. II.* nonuniform Markov models. *Molekulyarnaya Biologia* **20**, 1114–1123.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978). Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proceedings of the National Academy of Sciences* **75**(10), 4853–4857.

Breathnach, R. and Chambon, P. (1981). Organization and expression of eukaryotic split genes coding for proteins. *Annual Review of Biochemistry* **50**, 349–393.

Brunak, S., Engelbrecht, J. and Knudsen, S. (1991). Prediction of Human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology* **220**, 49–65.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology* **212**, 563–578.

Bucher, P. and Trifonov, E. (1986). Compilation and analysis of eukaryotic PolII promoter sequences. *Nucleic Acids Research* **14**, 10009–10026.

Burge, C. (1997). Identification of genes in human genomic DNA. Ph.D. Thesis, Stanford pp 152.

Burge, C. (1998). Modelling dependencies in pre-mRNA splicing signals. In *Computational Methods in Molecular Biology*, S. Salzberg, D. Searls, and S. Kasif, eds. Elsevier, 129–164.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78–94.

Burset, M. and Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics* **34**(3), 353–367.

Burset, M., Seledtsov, I. and Solovyev, V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research* **28**(21), 4364–4375.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N. and Izawa, M., (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**(3), 327–336.

Cooper, S., Trinklein, N., Anton, E., Nguyen, L. and Myers, R. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1 % of the human genome. *Genome Research* **16**, 1–10.

Decker, C.J. and Parker, R. (1995). Diversity of cytopasmatic functions for the 3′-untranslated region of eukaryotic transcripts. *Current Opinions in Cell Biology* **7**, 386–392.

Diamond, M., Miner, J., Yoshinaga, S. and Yamamoto, K. (1990). Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science* **249**, 1266–1272.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F. and Hannon, G.J. (2004). Processing of primary microRNAs by the microprocessor complex. *Nature* **432**, 231–235.

Dietrich, R., Incorvaia, R. and Padgett, R. (1997). Terminal intron dinucleotides sequences do not distinguish between U2- and U12-dependent introns. *Molecular Cell* **1**, 151–160.

Dong, S. and Searls, D. (1994). Gene structure prediction by linguistic methods. *Genomics* **23**, 540–551.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, pp 344.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D.S. (2003). MicroRNA targets in Drosophila. *Genome Biology* **5**, R1.

Farber, R., Lapedes, A. and Sirotkin, K. (1992). Determination of eukaryotic protein coding regions using neural networks and information theory. *Journal of Molecular Biology* **226**, 471–479.

Fickett, J. and Hatzigeorgiou, A. (1997). Eukaryotic promoter recognition. *Genome Research* **7**, 861–878.

Fickett, J.W. and Tung, C.S. (1992). Assesment of protein coding measures. *Nucleic Acids Research* **20**, 6441–6450.

Fields, C. and Soderlund, C. (1990). GM: a practical tool for automating DNA sequence analysis. *CABIOS* **6**, 263–270.

Forney, G.D. (1973). The Viterbi algorithm. *Proceedings of the IEEE* **61**, 268–278.

Gelfand, M. (1989). Statistical analysis of mammalian pre-mRNA splicing sites. *Nucleic Acids Research* **17**, 6369–6382.

Gelfand, M. (1990). Global methods for the computer prediction of protein-coding regions in nucleotide sequences. *Biotechnology Software* **7**, 3–11.

Gelfand, M. and Roytberg, M. (1993). Prediction of the exon-intron structure by a dynamic programming approach. *BioSystems* **30**(1–3), 173–182.

Gelfand, M., Mironov, A. and Pevzner, P. (1996). Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 9061–9066.

Ghosh, D. (1990). A relational database of transcription factors. *Nucleic Acids Research* **18**, 1749–1756.

Ghosh, D. (2000). Object-oriented transcriptional factors database (ooTFD). *Nucleic Acids Research* **28**, 308–310.

Grad, Y., Aach, J., Hayes, G.D., Reinhart, B.J., Church, G.M., Ruvkun, G. and Kim, J. (2003). Computational and experimental identification of C.elegans microRNAs. *Molecular Cell* **11**, 1253–1263.

Green, P., Hillier, L. (1998). *Genefinder, unpublished software*. It is still unpublished.

Guigo, R. (1998). Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology* **5**, 681–702.

Guigo, R. (1999). DNA composition, codon usage and exon prediction. In *Genetics Databases*, Academic Press, pp. 54–80.

Guigo, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T.R., Harrow, J., Hubbard, T., Lewis, S.E. and Reese, M.G. (2006). EGASP: the human ENCODE genome annotation assessment project. *Genome Biology* **7**(Suppl. 1), S2-1–S2-31.

Guigo, R., Knudsen, S., Drake, N. and Smith, T. (1992). Prediction of gene structure. *Journal of Molecular Biology* **226**, 141–157.

Guigo, R. and Reese, M.G. (2005). EGASP collaboration through competition to find human genes. *Nature Methods* **2**(8), 577.

Halees, A.S., Leyfer, D. and Weng, Z. (2003). PromoSer: a large-scale mammalian promoter and transcription start site identification service. *Nucleic Acids Research* **31**, 3554–3559.

Hall, S.L. and Padgett, R.A. (1994). Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *Journal of Molecular Biology* **239**(3), 357–365.

Hall, S.L. and Padgett, R.A. (1996). Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science* **271**, 1716–1718.

Harrison, P., Milburn, D., Zhang, Z., Bertone, P. and Gerstein, M. (2003). Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Research* **31**(3), 1033–1037.

Henderson, J., Salzberg, S. and Fasman, K. (1997). Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology* **4**, 127–141.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research* **30**(1), 38–41.

Hutchinson, G. (1996). The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Computer Applications in the Biosciences* **12**, 391–398.

Hutchinson, G.B. and Hayden, M.R. (1992). The prediction of exons through an analysis of splicible open reading frames. *Nucleic Acids Research* **20**, 3453–3462.

Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Research* **19**(14), 3795–3798.

Jeffrey, H.J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research* **18**, 2163–2170.

Jones-Rhoades, M.W. and Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stressinduced miRNA. *Molecular Cell* **14**, 787–799.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2005). The EMBL nucleotide sequence database. *Nucleic Acids Research* **33**, D29–D33.

Kel, O., Romaschenko, A., Kel, A., Wingender, E. and Kolchanov, N. (1995). A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Research* **23**, 4097–4103.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research* **12**(6), 996–1006.

Knudsen, S. (1999). Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics* **15**, 356–361.

Kolchanov, N.A., Podkolodnaya, O.A., Ananko, E.A., Ignatieva, E.V., Stepanenko, I.L., Kel-Margoulis, O.V., Kel, A.E., Merkulova, T.I. and Goryachkovskaya, T.N. (2000). Transcription regulatory regions regions database (TRRD): its status in 2000. *Nucleic Acids Research* **28**, 298–301.

Kondrakhin, Y.V., Shamin, V.V. and Kolchanov, N.A. (1994). Construction of a generalized consensus matrix for recognition of vertebrate pre-mRNA 3′-terminal processing sites. *Computer Applications in the Biosciences* **10**, 597–603.

Krogh, A. (1997). Two methods for improving performance of an HMM and their application for gene finding. *Intelligent Systems in Molecular Biology* **5**, 179–186.

Krogh, A. (2000). Using database matches with HMMgene for automated gene detection in *Drosophila*. *Genome Research* **4**, 523–528.

Krogh, A., Mian, I.S. and Haussler, D. (1994). A hidden Markov Model that finds genes in *Escherichia coli* DNA. *Nucleic Acids Research* **22**, 4768–4778.

Kulp, D., Haussler, D., Rees, M. and Eeckman, F. (1996). A generalized hidden Markov model or the recognition of human genes in DNA. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, D. States, P. Agarwal, T. Gaasterland, L. Hunter and R. Smith, eds. AAAI Press, St. Louis, MO, pp. 134–142.

Lapedes, A., Barnes, C., Burks, C., Farber, R. and Sirotkin, K. (1988). Application of neural network and other machine learning algorithms to DNA sequence analysis. *In Proceedings Santa Fe Institute* **7**, 157–182.

Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S. and Kim, V.N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**(6956), 415–419.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H. and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO Journal* **23**, 4051–4060.

Lewis, B.P., Shih, I., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003). Prediction of mammalian microRNA targets. *Cell* **115**, 787–798.

Lukashin, A.V. and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research* **26**, 1107–1115.

Mallory, A.C. and Vaucheret, H. (2004). MicroRNAs: something important between the genes. *Current Opinion in Plant Biology* **7**, 120–125.

Manley, J.L. (1995). A complex protein assembly catalyzes polyadenylation of mRNA precursors. *Current Opinion in Genetics and Development* **5**, 222–228.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta* **405**, 442–451.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E. and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**, D108–D110.

McLauchlan, J., Gaffney, D., Whitton, J.L. and Clements, J.B. (1985). The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3′ termini. *Nucleic Acids Research* **13**, 1347–1367.

Milanesi, L. and Rogozin, I.B. (1998). Prediction of human gene structure. In *Guide to Human Genome Computing*, 2nd edition. M.J. Bishop, ed. Academic Press, London, pp. 215–259.

Mount, S. (1982). A catalogue of splice junction sequences. *Nucleic Acids Research* **10**, 459–472.

Mount, S.M. (1993). Messenger RNA splicing signal in Drosophila genes. In *An Atlas of Drosophila Genes*, G. Maroni. Oxford University Press, Oxford.

Nakata, K., Kanehisa, M. and DeLisi, C. (1985). Prediction of splice junctions in mRNA sequences. *Nucleic Acids Research* **13**, 5327–5340.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.

Nilsen, T.W. (1994). RNA-RNA interactions in the spliceosome: unraveling the ties that bind. *Cell* **78**, 1–4.

Ohler, U., Harbeck, S., Niemann, H., Noth, E. and Reese, M. (1999). Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics* **15**, 362–369.

Ohler, U., Liao, G.C., Niemann, H. and Rubin, G.M. (2002). Computational analysis of core promoters in the *Drosophila* genome. *Genome Biology*, **3**(12), research0087.1–research0087.12.

Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P. and Burge, C.B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* **10**, 1309–1322.

Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. (1999). The biology of eukaryotic promoter prediction – a review. *Computers and Chemistry* **23**, 191–207.

Perier, C.R., Praz, V., Junier, T., Bonnard, C. and Bucher, P. (2000). The eukaryotic promoter database (EPD). *Nucleic Acids Research* **28**, 302–303.

Pillai, R. (2005). MicroRNA function: multiple mechanisms for a tiny RNA? *RNA* **11**, 1753–1761.

Prestridge, D. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. *Journal of Molecular Biology* **249**, 923–932.

Prestridge, D. and Burks, C. (1993). The density of transcriptional elements in promoter and non-promoter sequences. *Human Molecular Genetics* **2**, 1449–1453.

Proudfoot, N.J. (1991). Poly(A) signals. *Cell* **64**, 617–674.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**(1), D501–D504.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2), 257–285.

Rabiner, L., Juang, B. (1993). *Fundamentals of speech recognition*. Prentice Hall, New Jersey, p. 507.

Reese, M.G., Harris, N.L. and Eeckman, F.H. (1996). *Large Scale Sequencing Specific Neural Networks for Promoter and Splice Site Recognition. Biocomputing: Proceedings of the 1996 Pacific Symposium*, L. Hunter and T. Klein, eds. World Scientific Publishing Company, Singapore.

Reese, M., Kulp, D., Tammana, H. and Haussler, D. (2000). Genie – Gene finding in *Drosophila melanogaster*. *Genome Research* **10**, 529–538.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R. and Ruvkun, G. (2000). The 21- nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901–906.

Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507–1517.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B. and Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell* **110**, 513–520.

Salamov, A.A. and Solovyev, V.V. (1997). Recognition of 3′-end cleavage and polyadenilation region of human mRNA precursors. *CABIOS* **13**(1), 23–28.

Salamov, A. and Solovyev, V. (2000). Ab initio gene finding in Drosophila genomic DNA. *Genome Research* **10**, 516–522.

Salsberg, S., Delcher, A., Fasman, K. and Henderson, J. (1998). A decision tree system for finding genes in DNA. *Journal of Computational Biology* **5**, 667–680.

Schmid, C.D., Perier, R., Praz, V. and Bucher, P. (2006). EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Research* **34**, D82–D85.

Schmid, C.D., Praz, V., Delorenzi, M., Perier, R. and Bucher, P. (2004). The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Research* **32**, D82–D85.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A. and Shinozaki, K. (2002). Functional annotation of a full-length Arabidopsis cDNA collection. *Science* **296**, 141–145.

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M. and Solovyev, V.V. (2003). PlantProm: a database of plant promoter sequences. *Nucleic Acids Research* **31**, 114–117.

Shahmuradov, I., Solovyev, V. and Gammerman, A. (2005). Plant promoter prediction with confidence estimation. *Nucleic Acids Research* **33**(3), 1069–1076.

Shahmuradov, I.A., Kolchanov, N.A., Solovyev, V.V. and Ratner, V.A. (1986). Enhancer-like structures in middle repetitive sequences of the eukaryotic genomes. *Genetics (Russ)* **22**, 357–368.

Shahmuradov, I.A., Solovyev, V.V. (1999). NSITE program for identification of functional motifs with estimation of their statistical significance `http://sun1.softberry.com/berry.phtml?topic=nsite&group=programs&subgroup=promoter`.

Sharp, P.A. and Burge, C.B. (1997). Classification of introns: U2-type or U12-type. *Cell* **91**, 875–879.

Senapathy, P., Sahpiro, M. and Harris, N. (1990). Splice junctions, brunch point sites, and exons: sequence statistics, identification, and application to genome project. *Methods in Enzymology* **183**, 252–278.

Shepherd, J.C.W. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and ist possible evolutionary justification. *Proceedings of the National Academy of Sciences of the United States of America* **78**, 1596–1600.

Smit, A. and Green, (1997). RepeatMasker Web server: `http:// repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker`.

Snyder, E.E. and Stormo, G.D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Research* **21**, 607–613.

Snyder, E. and Stormo, G. (1995). Identification of protein coding regions in genomic DNA. *Journal of Molecular Biology* **21**, 1–18.

Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C., Angerer, L.M., Arnone, M.I., Burgess, D.R., Burke, R.D., Coffman, J.A., Dean, M., Elphick, M.R., Ettensohn, C.A., Foltz, K.R., Hamdoun, A., Hynes, R.O., Klein, W.H., Marzluff, W., McClay, D.R., Morris, R.L., Mushegian, A., Rast, J.P., Smith, L.C., Thorndyke, M.C., Vacquier, V.D., Wessel, G.M., Wray, G., Zhang, L., Elsik, C.G., Ermolaeva, O., Hlavina, W., Hofmann, G., Kitts, P., Landrum, M.J., Mackey, A.J., Maglott, D., Panopoulou, G., Poustka, A.J., Pruitt, K., Sapojnikov, V., Song, X., Souvorov, A., Solovyev, V., Wei, Z., Whittaker, C.A., Worley, K., Durbin, K.J., Shen, Y., Fedrigo, O., Garfield, D., Haygood, R., Primus, A., Satija, R., Severson, T., Gonzalez-Garay, M.L., Jackson, A.R., Milosavljevic, A., Tong, M., Killian, C.E., Livingston, B.T., Wilt, F.H., Adams, N., Belle, R., Carbonneau, S., Cheung, R., Cormier, P., Cosson, B., Croce, J., Fernandez-Guerra, A., Geneviere, A.M., Goel, M., Kelkar, H., Morales, J., Mulner-Lorillon, O., Robertson, A.J., Goldstone, J.V., Cole, B., Epel, D., Gold, B., Hahn, M.E., Howard-Ashby, M., Scally, M., Stegeman, J.J., Allgood, E.L., Cool, J., Judkins, K.M., McCafferty, S.S., Musante, A.M., Obar, R.A., Rawson, A.P., Rossetti, B.J., Gibbons, I.R., Hoffman, M.P., Leone, A., Istrail, S., Materna, S.C., Samanta, M.P., Stolc, V., Tongprasit, W., Tu, Q., Bergeron, K.F., Brandhorst, B.P., Whittle, J., Berney, K., Bottjer, D.J., Calestani, C., Peterson, K., Chow, E., Yuan, Q.A., Elhaik, E., Graur, D., Reese, J.T., Bosdet, I., Heesun, S., Marra, M.A., Schein, J., Anderson, M.K., Brockton, V., Buckley, K.M., Cohen, A.H., Fugmann, S.D., Hibino, T., Loza-Coll, M., Majeske, A.J., Messier, C., Nair, S.V., Pancer, Z., Terwilliger, D.P., Agca, C., Arboleda, E., Chen, N., Churcher, A.M., Hallbook, F., Humphrey, G.W., Idris, M.M., Kiyama, T., Liang, S., Mellott, D., Mu, X., Murray, G., Olinski, R.P., Raible, F., Rowe, M., Taylor, J.S., Tessmar-Raible, K., Wang, D., Wilson, K.H., Yaguchi, S., Gaasterland, T., Galindo, B.E., Gunaratne, H.J., Juliano, C., Kinukawa, M., Moy, G.W., Neill, A.T., Nomura, M., Raisch, M., Reade, A., Roux, M.M., Song, J.L., Su, Y.H., Townley, I.K., Voronina, E., Wong, J.L., Amore, G., Branno, M., Brown, E.R., Cavalieri, V., Duboc, V., Duloquin, L., Flytzanis, C., Gache, C., Lapraz, F., Lepage, T., Locascio, A., Martinez, P., Matassi, G., Matranga, V., Range, R., Rizzo, F., Rottinger, E., Beane, W., Bradham, C., Byrum, C., Glenn, T., Hussain, S., Manning, F.G., Miranda, E., Thomason, R., Walton, K., Wikramanayke, A., Wu, S.Y., Xu, R., Brown, C.T., Chen, L., Gray, R.F., Lee, P.Y., Nam, J., Oliveri, P., Smith, J., Muzny, D., Bell, S., Chacko, J., Cree, A., Curry, S., Davis, C., Dinh, H., Dugan-Rocha, S., Fowler, J., Gill, R., Hamilton, C., Hernandez, J., Hines, S., Hume, J., Jackson, L., Jolivet, A., Kovar, C., Lee, S., Lewis, L., Miner, G., Morgan, M., Nazareth, L.V., Okwuonu, G., Parker, D., Pu, L.L., Thorn, R. and Wright, R. (2006). The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941–952.

Solovyev, V.V. (1993). Fractal graphical representation and analysis of DNA and Protein sequences. *BioSystems* **30**, 137–160.

Solovyev, V. (1997). Fgenes – Pattern based finding multiple genes in human genome sequences. `http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind`.

Solovyev, V., Kolchanov, N. (1994). Search for functional sites using consensus. In *Computer Analysis of Genetic Macromolecules. Structure, Function and Evolution*. N.A. Kolchanov and H.A. Lim, eds. World Scientific, pp. 16–21.

Solovyev, V.V., Korolev, S.V., Tumanyan, V.G. and Lim, H.A. (1991). A new approach to classification of DNA regions based on fractal representation of functionally similar sequences. *Proceedings of the National Academy of Sciences of USSR (Russ) (Biochemistry)* **319**(6), 1496–1500.

Solovyev, V., Kosarev, P., Seledsov, I. and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology* **7**(Suppl. 1), 10-1–10-12.

Solovyev, V.V., Lawrence, C.B. (1993a). Identification of Human gene functional regions based on oligonucleotide composition. In *Proceedings of First International Conference on Intelligent System for Molecular Biology*, L. Hunter, D. Searls and J. Shalvic, eds. AAAI Press, Menlo Park, Californiya, pp. 371–379.

Solovyev, V., Lawrence, C. (1993b). Prediction of human gene structure using dynamic programming and oligonucleotide composition. In *Abstracts of the 4th Annual Keck Symposium*, Pittsburgh, PA, p. 47.

Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research* **22**, 6156–5153.

Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1995). Prediction of human gene structure using linear discriminant functions and dynamic programming. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, C. Rawling, D. Clark, R. Altman, L. Hunter, T. Lengauer and S. Wodak, eds. AAAI Press, Cambridge, pp. 367–375.

Solovyev, V.V. and Salamov, A.A. (1997). The Gene-Finder computer tools for analysis of human and model organisms' genome sequences. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, C. Rawling, D. Clark, R. Altman, L. Hunter, T. Lengauer and S. Wodak, eds. AAAI Press, Halkidiki, pp. 294–302.

Solovyev, V.V. and Salamov, A.A. (1999). INFOGENE: a database of known gene structures and predicted genes and proteins in sequences of genome sequencing projects. *Nucleic Acids Research* **27**(1), 248–250.

Solovyev, V., Shahmuradov, I., Akbarova, Y. (2003). The RegsiteDB: A database of transcription regulatory motifs of animal and plant eukaryotic genes: `http://www.softberry.com/berry.phtml?topic=regsite`.

Staden, R. (1984a). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research* **12**, 505–519.

Staden, R. (1984b). Mesurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Research* **12**, 551–567.

Staden, R. and McLachlan, A. (1982). Codon prefernce and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research* **10**, 141–156.

Stanke, M., Tzvetkova, A. and Morgenstern, B. (2006). AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology* **7**(Suppl. 1), S11.

Stark, A., Brennecke, J., Russell, R.B. and Cohen, S.M. (2003). Identification of Drosophila microRNA targets. *PLoS Biology* **1**, 1–13.

Stormo, G.D. and Haussler, D. (1994). Optimally parsing a sequence into different classes based on multiple types of evidence. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 47–55.

Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *Escherichia coli*. *Nucleic Acids Research* **10**, 2997–3011.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S., Okubo, K., Sakaki, Y., Nakamura, Y., Suyama, A. and Sugano, S. (2001). Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Reports*, **2**, 388–393.

Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K. (2004). DBTSS, database of transcriptional start sites: progress report 2004. *Nucleic Acids Research* **32**, D78–D81.

Tarn, W.Y. and Steitz, J.A. (1996a). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84**(5), 801–811.

Tarn, W.Y. and Steitz, J.A. (1996b). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science* **273**, 1824–1832.

Tarn, W.Y. and Steitz, J.A. (1997). Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends in Biochemical Sciences* **22**(4), 132–137.

The ENCODE Project Consortium, (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–639.

Tjian, R. (1995). Molecular machines that control genes. *Scientific American* **272**, 54–61.

Tjian, R. and Maniatis, T. (1994). Transcriptional activation: a complex puzzle with few easy pieces. *Cell* **77**, 5–8.

Thanaraj, T.A. (2000). Positional characterization of false positives from computational prediction of human splice sites. *Nucleic Acids Research* **28**, 744–754.

The Honeybee Genome Sequencing Consortium. (2006). Insights into social insects from the genome of the honey bee Apis mellifer. *Nature* **433**(7114), 931–949.

Thomas, A. and Skolnick, M. (1994). A probabilistic model for detecting coding regions in DNA sequences. *Ima Journal of Mathematics Applied in Medicine and Biology* **11**, 149–160.

Tyson, G., Chapman, J., Hugenholtz, P., Allen, E., Ram, R.J., Richardson, P., Solovyev, V., Rubin, E., Rokhsar, D. and Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43.

Wadman, M. (1998). Rough draft' of human genome wins researchers' backing. *Nature* **393**, 399–400.

Wahle, E. (1995). 3′-end cleavage and polyadelanytion of mRNA precursor. *Biochimica et Biophysica Acta* **1261**, 183–194.

Wahle, E. and Keller, W. (1992). The biochemistry of the 3′-end cleavage and polyadenylation of mRNA precursors. *Annual Review of Biochemistry* **61**, 419–440.

Werner, T. (1999). Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome* **10**, 168–175.

Wieringa, B., Hofer, E. and Weissmann, C. (1984). A minimal intron length but no specific internal sequence is required for splicing the large rabbit Bglobin intron. *Cell* **37**, 915–925.

Wilusz, J., Shenk, T., Takagaki, Y. and Manley, J.L. (1990). A multicomponent complex is required for the AAUAAA-dependent cross-linking of a 64-kilodalton protein to polyadenylation substrates. *Molecular and Cellular Biology* **10**, 1244–1248.

Wingender, E. (1988). Compilation of transcription regulating proteins. *Nucleic Acids Research* **16**, 1879–1902.

Wingender, E., Dietze, P., Karas, H. and Knuppel, R. (1996). TRANSFAC: a database of transcription factors and their binding sites. *Nucleic Acids Research* **24**, 238–241.

Wu, Q. and Krainer, A.R. (1997). Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. *RNA* **3:6**, 586–601.

Xu, Y., Einstein, J.R., Mural, R.J., Shah, M. and Uberbacher, E.C. (1994). An improved system for exon recognition and gene modeling in human DNA sequences. In *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*,

R. Altman, D. Brutlag, P. Karp, R. Lathrop and D. Searls, eds, , Menlo Park, Californiya, pp. 376–383.

Yada, T., Ishikawa, M., Totoki, Y., Okubo, K. (1994). Statistical analysis of human DNA sequences in the vicinity of poly(A) signal. Technical Report TR-876, Institute for New Generation Computer Technology.

Yu, J., Hu, S., Wang, J., Wong, G., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y. and Zhang, X. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296**, 79–92.

Zhang, M. and Marr, T. (1993). A weight array method for splicing signal analysis. *Computer Applications in the Biosciences* **9**, 499–509.