The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames

Victor V. Solovyev, Asaf A. Salamov and Charles B. Lawrence

Department of Cell Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030
solovyev@cmb.bcm.tmc.edu

Abstract

Discriminant analysis is applied to the problem of recognition 5'-, internal and 3'-exons in human DNA sequences. Specific recognition functions were developed for revealing exons of particular types.The method based on a splice site prediction algorithm that uses the linear Fisher discriminant to combine the information about significant triplet frequencies of various functional parts of splice site regions and preferences of oligonucleotides in protein coding and intron regions (Solovyev,Lawrence,1994). The accuracy of our splice site recognition function is about 97%. A discriminant function for 5'-exon prediction includes hexanucleotide composition of upstream region, triplet composition around the ATG codon, ORF coding potential, donor splice site potential and composition of downstream intron region. For internal exon prediction, we combine in a discriminant function the characteristics describing the 5'-intron region, donor splice site, coding region, acceptor splice site and 3'-intron region for each open reading frame flanked by GT and AG base pairs. The accuracy of precise internal exon recognition on a test set of 451 exon and 246693 pseudoexon sequences is 77% with a specificity of 79% and a level of pseudoexon ORF prediction of 99.96%. The recognition quality computed at the level of individual nucleotides is 89% for exon sequences and 98% for intron sequences. A discriminant function for 3'-exon prediction includes octanucleotide composition of upstream intron region, triplet composition around the stop codon, ORF coding potential, acceptor splice site potential and hexanucleotide composition of downstream region. We unite these three discriminant functions in exon predicting program ***FEX*** (find exons). ***FEX*** exactly predicts 70% of 1016 exons from the test of 181 complete genes with specificity 73%, and 89% exons are exactly or partially predicted. On the average, 85% of nucleotides were predicted accurately with specificity 91%.

Keywords: *splice site, exon prediction, human genes*

Introduction

One of the challenging problems in analyzing newly sequenced DNA is to develop of reliable gene identification method. A number of complex systems for predicting gene structure have been developed (Fields and Soderlund,1990; Uberbacher, Mural,1991; Guigo et al.,1992; Hutchinson,Hayden,1992; Snyder,Stormo,1993). These systems analyze information about functional signals and some characteristics of coding or intron regions. On this basis, potential first, internal and terminal exons can be revealed and the top ranking combination of them usually will present the predicted gene structure. The program ***SORFIND*** (Hutchinson, Hayden,1992) shows only the positions of candidate exons and do not attempt to produce assembled genes. The accuracy of exact internal exons prediction by ***SORFIND*** program reaches about 59%. To date, ***GeneModeler*** (Fields and Soderlund,1990), ***GeneID*** (Guigo et al.,1992), ***GRAIL*** (Uberbacher et al.,1993) and ***GeneParser*** (Snyder,Stormo,1993) are the valuable integrated packages that predict gene structure from genomic DNA. The first two methods rely on revealing of the potential functional motifs such as start and stop codons, splice sites and poly(A) signals and then on sequential filtering evaluation of the assembled combination gene component. ***GeneID*** can predict the true gene structure as a top ranking structure in only 14% cases of tested vertebrate gene sequences and in only 54% cases identify the correct exons with correct splice boundaries (Guigo et al.,1992). A dynamic programming approach (alternative to the rule-based approach) was suggested by Snyder and Stormo (1993). It accomplishes an exhaustive and mathematically rigorous search for the globally optimal solution. Weights for the various classification procedures are determined by training a feed-forward neural network to maximize the number of correct predictions. ***GeneParser*** precisely identifies 74% of the internal exons (with a specificity of 62%), but the structure of only 17% test genes were exactly predicted. The prediction quality decreases dramatically for terminal exons, which seems require a special consideration (Snyder,Stormo,1993).

The goal of this work is to develop a computational approach of revealing human exon regions, which is based on our improved splice site recognition method. Special discriminant functions for internal, 5'- and 3'-exons have been constructed and combined in exon prediction program ***FEX*** (find exons). ***FEX*** compares favourably with other programs currently used to predict protein-coding regions. A test set of 181 human sequences containing complete genes (all coding regions and introns, flanked by150 nucleotides before start codon and after stop codon) was used for estimation the accuracy of exon prediction. ***FEX*** exactly predicts 70% of 1016 exons with specificity 73%, and 89% exons are exactly or partially predicted. On the average, 85% of nucleotides were predicted accurately with specificity 91%. Results for ***GRAIL-2*** email server prediction for the same sequences show 39% of exact exon prediction

with specificity 47% and 76% partially exon prediction; 77% accuracy on the nucleotide level with specificity 87%.

The Methods

Discriminant analysis

We have applied the technique of discriminant analysis to relate the given region to one of two alternative classes, W_1 (sites or exons) or W_2 (pseudosites or pseudoexons) (Afifi,Azen, 1979). The procedure of linear discriminant analysis is to find a linear combination of the measures (or 'characteristics') $x_1, \ldots, x_p$, such that the distributions for the two groups will possess minimal overlap. The linear discriminant function:

$$z = \sum_{i=1}^{p} \alpha_i x_i \qquad \text{(EQ 1)}$$

classifies $\vec{x}$ into W_1 if $z \geq c$ and $\vec{x}$ into W_2 if $z < c$. The optimal selection of $\vec{\alpha} = S^{-1}(\vec{\mu}_1 - \vec{\mu}_2)$ and $c = \frac{\vec{\alpha}(\vec{\mu}_1 - \vec{\mu}_2)}{2}$ maximizes the ratio of between-class variation to within-class variation. This ratio with the optimal $\vec{\alpha}$ parameters presents so called Mahalonobis distance $\mathbf{D}^2$

$$D^2 = (\vec{\mu}_1 - \vec{\mu}_2)' S^{-1} (\vec{\mu}_1 - \vec{\mu}_2) \qquad \text{(EQ 2)}$$

where; $\vec{\mu}_i$ are the sample mean vectors for W_i; and $S = \frac{1}{n_1 + n_2 - 2}(s_1 + s_2)$ is the pooled covariance matrix (s_i are the covariance matrices for W_i classes, and n_i are the sample sizes) (Afifi,Azen, 1979). $\mathbf{D}^2$ is a good measure of the "distance" between the two populations. It can characterize the classification power of a particular characteristic as well as any of their combination.

Positional triplet preferences method

Base, deplete or triplet composition of sequences adjacent to a particular site positions is a good discriminant of these sites (Staden,1990; Mural, Mann, Uberbacher,1990). We will use it as a characteristic of functional regions defining a particular signal such as splicing site or start of translation. We tabulate the frequency of triplets, in the (L,R) window around a site, where L is the number position to the 5'-side, and R is the number position to the 3'-side of the exon-intron (or intron-exon) boundary. The triplet frequencies are stored in a matrix (L+R, 64) size.

Let $F^i_{s,k}$, $F^i_{p,k}$ be the frequencies of a specific triplet (the triplet type marked by k, where k=1,2,...,64) in the learning site and pseudosite sets of sequences in i-th position of a (L,R) window, respectively. As a pseudosite we will consider any sequence where analyzing site is absent, but it can have some of its features. Then the preference of a given triplet {k} in i-th position belongs to a site sequence can be defined as:

$$P(i) = \frac{F^i_{s,k}}{F^i_{s,k} + F^i_{p,k}} \qquad \text{(EQ 3)}$$

For example, for donor splice sites discrimination we use the mean preference index obtained by averaging the preferences in the *(L,R)* window around any GT dinucleotide of a sequence under analysis (eqn.4), where j is the splice site position, corresponding to the G base of the conserved dinucleotide. Only a subset of all possible triplets can influence site selection. Therefore, the discrimination function is modified to take into account only those triplets which have a significant difference in their occurrence between site and pseudosite site regions. If triplets are equally present in both types of regions, P(i,k) will be equal 0.5. For computing only significant triplets we calculate the following function

$$P_{sp}(j) = \frac{1}{m}\left(\sum_{i=L}^{R} P(i) \right) \qquad \text{(EQ 4)}$$

The summation is made if $(P(i) - 0.5) > \alpha$, where α is some threshold value for considering only significant triplets, and m is the number of significant triplets. Pseudosites may be localized in intron as well as in exon regions. The significant difference of triplet composition in intron and coding regions is clear, therefore recognition function have to be more sensitive if we will not represent the triplet composition of both cases in a single table. Two separate tables of triplet frequencies around pseudosite junctions that localized either in intron $F_{pi,k}$ or in coding $F_{pc,k}$ region may be calculated. For discrimination of a given sequences the average value of eqn.4 computed with each of these tables is used.

Oligonucleotide preferences method

As a characteristics distinguishing 5'-,3'-, intron and coding regions we use an oligonucleotide composition statistics. This method was described in details and tested on human sequences earlier (Solovyev, Lawrence,1993a). Here we outline only its main equations.

If we have the sequence S :

$$S = n_1 n_2 n_3 \ldots n_N; \{n_i \in A, C, G, T; i = 1, \ldots, N\}$$

Then

$$s = n_1 n_2 n_3 \ldots n_L; \{n_i \in A, C, G, T; i = 1, \ldots, L < N\}$$

describes an oligonucleotide of length L.

For discrimination of functional (F) and nonfunctional (N) regions we can use the probability of oligonucleotide s_k belong to a functional region as estimated by the Bayesian method:

$$P(F|s_k) = \frac{P(s_k|F)\,P(F)}{P(s_k|F)\,P(F) + P(s_k|N)\,P(N)} = \frac{F_c(s_k)}{F_c(s_k) + F_n(s_k)} \qquad \text{(EQ 5)}$$

where $P(s_k|F)$, $P(s_k|N)$ are the *a posteriori* probabilities for s_k to occur in functional and nonfunctional regions; and $P(F)$, $P(N)$ are the *a priori* probabilities of a functional or nonfunctional region. We assume that $P(F) = P(N)$ and $F_c(s_k)$, $F_n(s_k)$ are the frequencies of s_k in coding and noncoding sets, respectively.

If we consider oligonucleotides only in phase with coding regions a discriminant function analogous to Eq.5 based on such in-phase oligonucleotides:

$$P^1(C|s_k) = \frac{F^1(s_k|C)}{F^1(s_k) + F(s_k|N)} \qquad \text{(EQ 6)}$$

A simple discriminant index for revealing a functional region is the average of Eq. 5 or Eq.6 along a sequence window W:

$$P_\alpha(F|W) = \frac{1}{m}\left(\sum_{i=1}^{n} P(i)\right), i = 1, s+1, 2s+1, \ldots \qquad \text{(EQ 7)}$$

where $P(i)$ is $P(F|s_k)$ or $P^1(C|s_k)$ and $s=1$ or $s=3$; s_k is the oligonucleotide starting in the i-th position of the sequence, and m is the number of summed oligonucleotides.

RESULTS AND DISCUSSION

Splice site prediction

Splice site prediction method using the linear function that combine triplet preferences around splice junction and preferences to be coding and intron of adjacent regions have been suggested earlier (Solovyev, Lawrence,1993). We improved this approach to separate in discriminant function triplet preferences of different functional parts of splice site regions and applied the discriminant analysis search for optimal coefficients of the discriminant function. The exon prediction method mainly based on accurate splice site prediction, therefore, we outline the significant feature of splice site prediction approach (Solovyev,Salamov, Lawrence,1994).

We extract from GenBank (Release 72) (Cinkosky et al., 1991) 692 sequences with 2037 donor splice sites and 2054 acceptor splice sites having the GT and AG conserved dinucleotide in flanking intron positions. Also, 89417 pseudodonor and 134150 pseudoacceptor sites that contain either a GT or AG base pair (and are not annotated as splice sites) were extracted from these sequences. The characteristics of sequences around splice sites and pseudosites were used for developing and testing human splice site recognition function to distinguish them. A training set including 2/3 of all sequences, and a test set containing the remaining ones. The data set for computing octanucleotide preferences in coding and intron regions included 4074593 bp of coding regions and 1797572 bp of intron sequences.

The difference of triplet composition of particular regions around splice site junctions is clearly observed from the figure 1. We combine the characteristics of marked parts of splice site regions (Figure 1) in a linear discriminant function. The characteristics used for classifying donor site are: the triplet preferences (eqn.3,4) in the potential coding region (-30 -- -5); conserved consensus region (-4 -- +6) and G-rich region (+7 -- +50); the number of significant triplets in conserved consensus region (α=0.15 in the eqn.4); octanucleotide preferences (eqn.6) for being coding in the (-60 to -1) region and being intron in the (+1 to +54) region; the number of G-bases, GG-doublets and GGG-triplets in +6 -- +50 region. The values of these 6 characteristics of donor site were calculated for 1375 authentic donor site and for 60532 pseudosite sequences from the learning set. The Mahalonobis distances showing significance of each characteristics are given in Table 1a. We can see that the strongest characteristic for donor sites is triplet composition in consensus region (D^2=9.3) and then the adjacent intron region (D^2=2.6) and 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; the quality of the coding and intron regions.

The characteristics for acceptor splice sites are: the triplet preferences (eqn.3,4) in the branch point region (-48 -- -34); poly(T/C)-tract region (-33 -- -7); conserved consensus region (-6 -- +5); coding region (+6 -- + 30); and octanucleotide preferences (eqn.6) of being coding in the (+1 to +54) region and in the (-1 to -54) region; and the number of T and C in poly(T/C)-tract region.

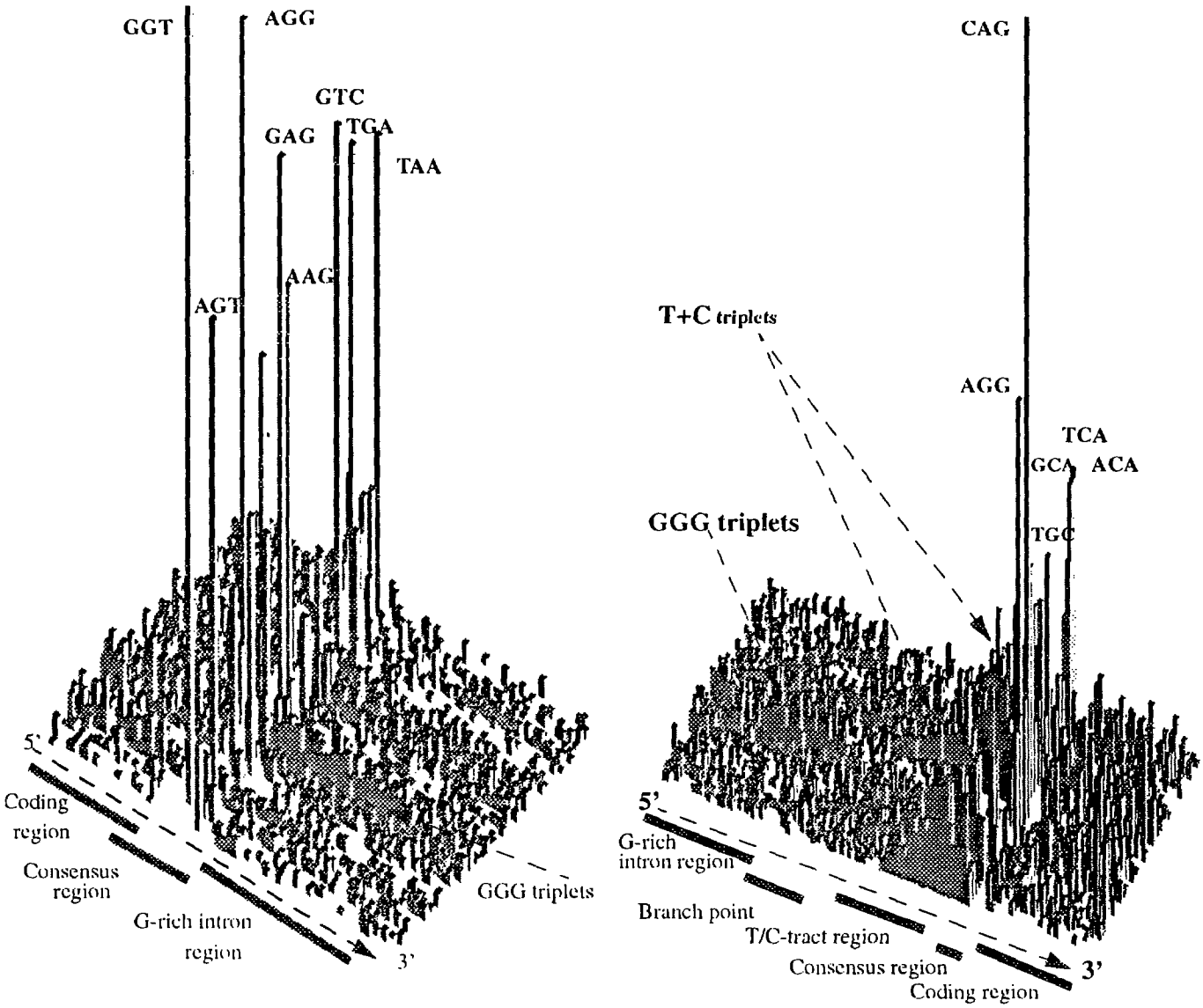

FIGURE 1. Difference of the triplet composition around donor and GT-containing pseudodonor sites (left); around acceptor and AG-containing pseudoacceptor sites (right) in 462 sequences of human genes from the training. Each column presents the difference of specific triplet numbers between sites and pseudosites in a specific position. For comparing the numbers for equal quantities of site and pseudosites were calculated.

The accuracy of the discriminant function based on these characteristics was tested on the recognition of 662 donor sites and 28855 pseudosite sequences. The general accuracy of donor site prediction is 97%. This accuracy is better than in the neural network-based method, which has C=0.61 at 95% accuracy (Brunak et al., 1991), comparing to C=0.63 using the approach reported here. C is an important accuracy criterion (correlation coefficient) that takes the relation between correctly predictive positives and negatives as well as false positives and negatives into account (Matthews 1975).

The values of 7 characteristics of acceptor sites were calculated for 1386 authentic acceptor site and 89791 pseudosite sequences from the learning set. The Mahalonobis distances showing individual significance for each characteristic are given in Table 2a. Table 2b shows the increasing combined Mahalonobis distance with the subsequent addition of each characteristic. We can see that strongest characteristics for acceptor sites are: the triplet composition in poly(T/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

TABLE 1. Significance of various characteristics of donor splice sites

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

a. 1, 2, 3 are the triplet preferences of consensus, intron G-rich and coding regions, 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-depletes and GGG-triplets in intron G-rich region.

TABLE 2. Significance of various characteristics of acceptor splice sites

Characteristics[a]	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

a. 1, 3, 4, 6 are the triplet preferences of poly(T/C)-tract, consensus, coding and branch point regions, respectively; 7 is the number of T and C in intron poly(T/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 of donor splice site junction.

intron region (D^2=2.4); and the quality of the coding region (D^2=2.6).

The accuracy of the discriminant function based on the other significant characteristics was tested on recognition 666 acceptor sites and 43726 pseudosite sequences. The general accuracy of acceptor site prediction is 96% (C=0.47).

To our knowledge, this accuracy of donor and acceptor splice site prediction is better than any other splice site prediction method yet described, that permit us to apply it for internal exon recognition.

Internal exon recognition

For internal intron prediction we consider all open reading frames in a given sequence that flanked AG (on the left) and GT (on the right) base pairs as potential internal exons. The structure of such exons are presented in Figure 2b. As components of internal exon recognition function we take the octanucleotide composition preferences for intron 70 bp of the left to the potential intron region; the value of the acceptor splice site recognition function, the octanucleotide composition preferences for coding ORF, the value of the donor splice site recognition function and the octanucleotide composition preferences for intron 70 bp to the right of potential intron region.

The data set for distinguishing human exon and pseudoexon ORF sequences contains 952 exons and 528480

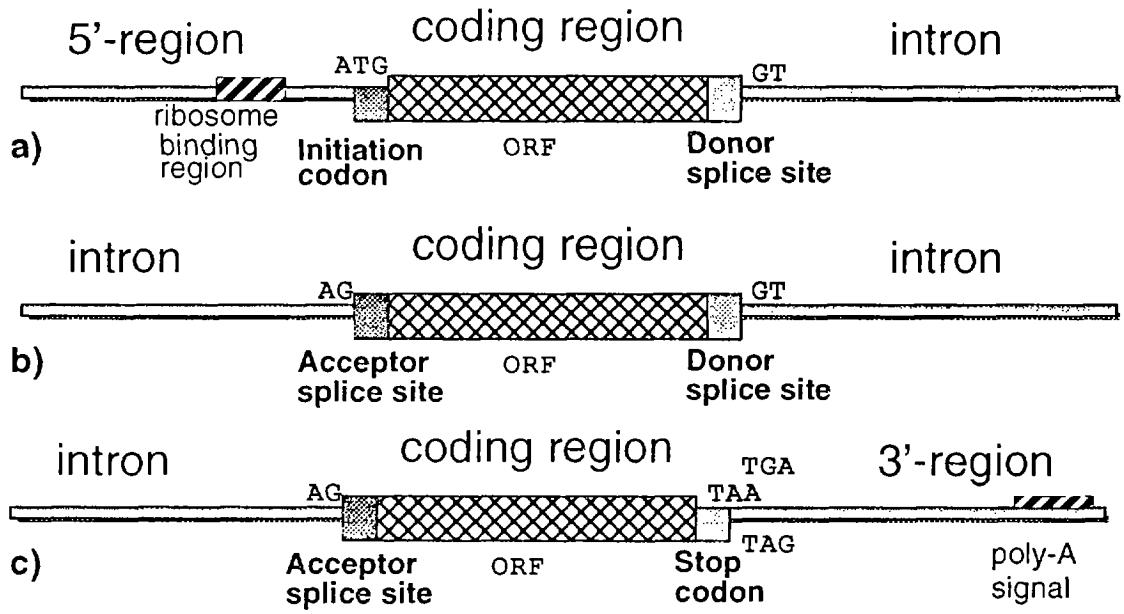

Figure 2. Different functional regions of the first (a), internal (b) and the last (c) exons corresponding to components of recognition functions.

pseudoexons in the training set, and 451 exons and 246693 pseudoexons in the test set.

The values of for 5' exon characteristics were calculated for 952 authentic exons and for 690714 pseudoexon training sequences from the set. The Mahalonobis distances showing significance of each characteristic are given in Table 3a. Table 3b shows the increasing combined Mahalonobis distance on subsequently adding each characteristic. We can see that the strongest characteristics for exons are the values of recognition functions of flanking donor and acceptor splice sites (D^2=15.04 and D^2=12.06, respectively). The preference of 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 .

The accuracy of the discriminant function based on these characteristics was calculated on the recognition of 451 exon and 246693 pseudoexon sequences from the test set. The general accuracy of exact internal exon prediction is 77% with specificity 79%. If the results are analyzed at the level of individual nucleotides, the accuracy of exon prediction is 89% with specificity 89%; and intron positions prediction is 98% with specificity 98%. This accuracy is better than in the most accurate described dynamic programming and neural network-based method (Snyder,Stormo,1993), which has 75% accuracy of the exact internal exons prediction with specificity 67%. Our method has 12% less false exon assignments with the better level of true exon prediction.

TABLE 3. Significance of internal exon characteristics.

Characteristics[a]	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

a. 1 and 2 are the values of donor and acceptor site recognition functions; 3 is the octanucleotide preferences for being coding of potential exon region; 4 are the octanucleotide preferences for being intron 70 bp region on the left and 70 bp region on the right of potential exon region; 5 .

TABLE 4. Significance of 5'- exon characteristics.

Characteristics[a]	1	2	3	4	5	6	7
(a) Individual D^2	16.3	3.8	2.5	0.01	1.9	1.6	2.5
(b) Combined D^2	16.9	18.2	18.8	19.1	19.1	19.1	19.2

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

5'-terminal exon coding region recognition

For 5'-exon prediction, we consider open reading frames in a given sequence that starting wit an ATG codon and ending with a GT dinucleotide as potential first exons. The structure of such exons are presented in Figure 2a. As components of the 5'-exon recognition function we take the hexanucleotide composition preferences for 5'-regions -150 - -101 bp, -100 - -51 bp, -50 - -1 bp to the left of the potential coding region; the average value of positional triplet preferences in the -15 - +10 region around ATG codon; octanucleotide composition preferences to be coding region of ORF, the value of donor splice site recognition function and the octanucleotide composition preferences for intron 70 bp to the right potential intron region.

The data set for distinguishing the first exon and pseudoexon ORF sequences contains 312 exons and 76611 pseudoexons from human genes.

The values of 7 first exon characteristics were calculated for theis data set. The Mahalonobis distances showing significance of each characteristic are given in Table 4a. Table 4b shows the increasing combined Mahalonobis distance subsequently adding each characteristic.

The accuracy of the discriminant function based on these characteristics was calculated on the recognition of 312 the first exon and 246693 pseudoexon sequences from the training set. We scan all gene sequences and select the 5'-exon with maximal weight for each of them. The accuracy of exact the first exon coding region prediction is 59% . It must be noted that the competition with internal exons was not considered in this test.

3'-terminal exon coding region recognition

We consider all ORF regions that flanked coding of potential exon region; GT (on the left) base pair and finished with a stop codon as potential last exons. The structure of such exona are presented in Figure 2c. As components of the 3'-exon recognition function we take the octanucleotide composition preferences for intron 70 bp to the left potential intron region; the value of the donor splice site recognition function; octanucleotide composition preferences to be coding region of ORF, hexanucleotide composition preferences for 3'-region +1-+50 bp, +51 - +100 bp, +101 - +151 bp to the right of the potential coding region; the average value of positional triplet preferences in the -10 - +30 region around the stop codon.

TABLE 5. Significance of 3'- exon characteristics.

Characteristics[a]	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

a. 1 is the value of acceptor site recognition function; 2 is the octanucleotide preferences for being coding of ORF region; 3 is the hexanucleotide preferences to be 3'-region +100 - 150 bp of the left potential coding region; 4 is the average value of positional triplet preferences in the -10 - +30 region around the stop codon; 5 is the hexanucleotide preferences to be 3'-region +50 - +100 bp of the left potential coding region; 6 is the octanucleotide preferences for being intron 70 bp region on the left of potential exon region; 7 is the hexanucleotide preferences to be 3'-region +1- + 50 bp of the left potential coding region.

The data set for distinguishing the last exon and pseudoexon ORF sequences contains 322 exons and 247644 pseudoexons from human genes.

The values of 7 characteristics of the last exon were calculated for these exons and pseudoexon sequences. The Mahalonobis distances showing significance of each characteristics are given in Table 4a. Table 4b shows the increasing combined Mahalonobis distance subsequently adding each characteristic.

The accuracy of the discriminant function based on these characteristics was calculated on the recognition of the last 322 exon and 247644 pseudoexon sequences from the training set. We scan all gene sequences and select the 3'-exon with maximal weight for each of them.The accuracy of exact the last exon coding region prediction is 60%. It must be noted that the competition with internal exons was not considered in this test.

Combined prediction the first, internal and the last exons in human genes

We have developed a computer program ***FEX*** which predicts coding regions in a given sequence. The program initially predicts internal exons based on internal exon discriminant function. Then we search for 5'-coding region starting from the beginning of the sequence until the end of the first predicted internal exon. In this region the 5'-coding exon with the maximal weight of the first exon discriminant function is selected. After that we search for 3'-coding region starting from the beginning of the last predicted internal exon until the end of the sequence. In this region the 3'-coding exon with the maximal weight of the last exon discriminant function is selected.

Two scoring schemes were used to evaluate the performance of ***FEX***. In scheme 1, only the sequences of complete human genes between -150 bp (before the first coding region) and +150 bp (after the last coding region) were considered. This data set includes the first complete 181 gene sequences of GenBank human genes. Scheme 2 analyzes the entire GenBank sequences. We also test the performance of GRAIL-2 using email server for these data sets.

At the level of complete exon sequences under the scoring scheme 1, ***FEX*** precisely identifies 709 of 1016 exons (70%) with specificity 73% and partially identifies 89% exons. The accuracy at the level of individual nucleotides is 85% with specificity 91% and correlation coefficient (calculated as in Brunak et al., 1991) equal to 0.84. ***GRAIL-2*** precisely predicts 39% exons with specificity 47% and partially predicts 76% exons. It has an accuracy of 77% at the nucleotide level with specificity 87% and correlation coefficient equal to 0.76. The accuracy for scheme 2, is slightly less for ***FEX*** (correlation coefficient equals 0.78) as well as for ***GRAIL-2*** (correlation coefficient equals 0.66).

Discriminant function for splice site position recognition in cDNA

Recognition of splice site position in cDNA may be very useful for gene mapping. Accurate prediction of splice site positions improve the possibility to select primers in internal exon sequence.

A simple approach to reveal splice site position is using remaining in mRNA parts of donor (**MAG/GURAGU**) and acceptor (**YAG/G**) consensus sequences (Senapathy et al.,1990; Mount,1993); i.e. **MAG/G** sequence. However this consensus is found only in 25% of splice site positions and at the same time per one such consensus belonging to authentic splice site we will predict about 15 false splice site positions. We try to use some information from adjacent splice site position sequences to reduce this enormous false site prediction.

The recognition discriminant function taking into account two components: triplet preferences within the consensus region (-4 - +3), triplet preferences adjacent to the splice site consensus (-20 - -5 and +4 - +20 bp) was developed.

The triplet preferences were calculated for three types of consensus sequences: **AGG** (that found in 28% of splice site position), **AGG** with **1** mismatch (70.41%) and **AGG** with **2** mismatch (95%) were considered. Triplet preferences (eqn.3,4) were computed using triplet frequencies of mRNA regions around authentic splice site positions and non-splice site positions of mRNA that contain the mentioned above consensuses.

In this case initially we find one of the consensuses in a given mRNA sequence and then estimate each of them using the corresponding triplet preferences.

Splice site position in cDNA prediction

The values of 2 characteristics were calculated for 1123 splice site positions and 262264 other positions in cDNA of human gene sequences from the training set. The Mahalonobis distance of the first characteristic (triplet preferences in close to consensus region) is 3.5 and of the second characteristic (triplet preferences in the right and left adjacent to consensus regions) is 3.2. The combined Mahalonobis distance of the both characteristics is 6.1. This result shows that some information about splice sites remains in mRNA sequence and may be used for predicting their positions. However this information is much less than we observed in pre-mRNA, where the Mahalonobis distance of splice site discrimination is about 16. Therefore we can expect many false splice site position predictions in cDNA analysis. We compared the quality of our discriminant function with prediction of splice site position using some consensus sequences: **MAGG, AGG, MAGG** with **1** mismatch, and **AGG** with **1** mismatch (Table 6). For the recognition function the level of false prediction was calculated with the level of true prediction the same as for a given consensus sequence. We can see that for the level of true prediction corresponding to the sensitivity of a consensus sequence, the first discriminant function has 2-3 times and the second discriminant function has 2.5 -20 times less the number of false predictions as compared with the consensus sequences. On the basis of the discriminant functions, it is possible to create a profile of probability (Lawrence, Solovyev, 1994) to be a splice site position for any position in a given cDNA sequence and then to select primer subsequences in the regions with minimal values of these probabilities.

Summary

Improved accuracy of splice site and human exon recognition using a combined linear classification scheme have been demonstrated. Using discriminant analysis we show relative significance of these regions for recognition. To our knowledge, this accuracy is better than any other splice site prediction method yet described. One of the advantages of our approach is that we can easily recalculate the tables of triplets to obtain increasingly reliable statistics as the size of the sequence data base increases.

Some of predicted pseudoexon ORFs can be further removed in a gene structure predictive system because only a subset of them will have an uninterrupted open reading frame through the entire gene. The first variant of such a system, ***FGENE***, has been developed (Solovyev, Lawrence, 1993b). This system takes into account the oligonucleotide composition of all key gene components (5'-region, exons, introns, 3'-region and noncoding regions) and the recognition of these components based on the functions similar to eqns 5 and 6. Dynamic programming is applied to search for a combination of splice sites with the maximal weight for the tested gene components. Testing the system on 200 human gene sequences shows that ***FGENE*** can predict precisely 80% exons with specificity 70% and 96% exons are predicted partially. The detailed description of this method will be published elsewhere. The algorithm for prediction of splice site position in cDNA may significantly increase of the effectiveness of primer selection for gene mapping by PCR reaction.

Analysis based on the methods for splice site, internal exon prediction and construction of a profile for the probability splice site positions in cDNA will be available through a network server by sending the file containing the sequence to *service@theory.bchs.uh.edu* with subject lines *hspl, hexon* or *fexh*.

TABLE 6. Prediction of splice site position in cDNA.

		Consensus	Discriminant function
	Sn	Number of false predictions per one correct prediction	
MAGG	25%	14	0.7
AGG	29%	19	0.9
MAGG*	59%	50	15
AGG*[a]	70%	68	27

a. * means that the consensus can have 1 mismatch; Sn is the percent of true prediction (sensitivity).

Acknowledgments

This work was supported by the W.M. Keck Center for Computation Biology, a grant to C.B.L. from the National Library of Medicine and an award from National center for human genome research (NIH) to V.V.S. Authors are grateful to Dr. N. Goodman for attraction their attention to primer selection problem.

References

Afifi A.A., Azen S.P. (1979) Statistical analysis. A computer oriented approach. Academic Press, New York.

Brunak, S.; Engelbrecht J.; Knudsen S. 1991. Prediction of Human mRNA donor and acceptor sites from the DNA sequence. *J. Mol.Biol.* 220: 49-65.

Cinkosky M.J.; Fickett J.W.; Gilna P.; Burks C. 1991. Electronic Data Publishing and GenBank. *Science* 252: 1273-1277.

Fields C.; Soderlund C.A. 1990. gm:a practical tool for automating DNA sequence analysis. *CABIOS* 6: 263-270.

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

Hutchinson G.B., Hayden M.R. 1992. The prediction of exons through an analysis of specialize open reading frames. Nucl.Acids Res. 20:3453-3462.

Lawrence C.B. , Solovyev V.V. 1994. Assignment of position-specific error probability to primary DNA sequence data. *Nucl. Acids Res., 22, N 7.*

Matthews B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem.Biophys.Acta* 405: 442-451.

Mount,S.M. (1993) Messenger RNA splicing signals in Drosophila genes. In An Atlas of Drosophila genes. (ed. Maroni G.), Oxford, 333-358.

Mural,R.J., Mann,R.C., Uberbacher,E.C. (1990) Pattern recognition in DNA sequences: The intron-exon junction problem. In: The first International Conference on Electrophoresis, Supercomputing and the Human Genome. (Cantor C.R., Lim H.A. eds). World Scientific, London, 164-172.

Senapathy P.; Shapiro M.B.; Harris N.L. 1990. Splice junctions, Branch point sites, and Exons. *Methods of Enzymology* (ed. R.F. Doolittle) 183: 252-280.

Snyder E.E.,Stormo G.D. (1993) Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nu cl.Acids Res., 21:607-613.

Solovyev,V., Lawrence,C. (1993a) Identification of human gene functional regions based on oligonucleotide composition. In: The First International conference on Intelligent systems for Molecular Biology (eds. Hunter L., Searls D., Shavlic J.), NLM IIII, Bethesda, 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, 47.

Solovyev V.V.; Lawrence C. 1994. Prediction of human mRNA donor and acceptor splice sites based on oligonucleotide composition, CABIOS (submitted).

Staden R. 1990 Finding protein coding regions in genomic sequences. In *Methods of Enzymology* (ed. R.F. Doolittle) **183**, 163-180.

Uberbacher E.C.; Mural R.J. 1991. Locating protein coding regions in human DNA sequences using a multiple sensor - neural net approach. *Proc.Natl.Acad.Sci.USA* 88: 11261-11265.