5156–5163 Nucleic Acids Research, 1994, Vol. 22, No. 24

Predicting internal 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, USA

Received November 4, 1994; Revised and Accepted November 11, 1994

ABSTRACT

A new method which predicts internal exon sequences in human DNA has been developed. The method is based on a splice site prediction algorithm that uses the linear discriminant function to combine information about significant triplet frequencies of various functional parts of splice site regions and preferences of oligonucleotides in protein coding and intron regions. The accuracy of our splice site recognition function is 97% for donor splice sites and 96% for acceptor splice sites. For 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%. The recognition quality computed at the level of individual nucleotides is 89% for exon sequences and 98% for intron sequences. This corresponds to a correlation coefficient for exon prediction of 0.87. The precision of this approach is better than other methods and has been tested on a larger data set. We have also developed a means for predicting exon – exon junctions in cDNA sequences, which can be useful for selecting optimal PCR primers.

INTRODUCTION

Prediction of coding and intron regions within large regions of uncharacterized genomic DNA is one of the challenging problems in analyzing newly sequenced DNA. Although the intermediates, products and reaction mechanisms of splicing were characterized some years ago, pre-mRNA structural features that are important for this process are just now being investigated (1). This is one of the principal motivations for using statistical methods for exon recognition.

Many methods and algorithms have been suggested for recognizing the components of gene structure. Currently there are two general approaches used for finding protein coding regions [see reviews by Stormo and Staden (1,2)]. The global approach (gene search by content) uses one or more coding measures, a function that calculates, for any window of sequence, a number or vector that estimates the protein-coding potential of these regions. The local approach (gene search by signal) is the identification of promoters, splice sites, translation initiating and terminating sites, poly(A)-signals, that surround coding regions. A comprehensive assessment of various protein coding measures was done by Fickett and Tung (3). They estimate the quality of more than 20 measures and showed that the most powerful—such as 'in-phase hexanucleotide composition', codon and amino acids usage—can result in up to 81% accuracy as coding region recognition functions on 54 base windows. Combining 'fourier', 'run', 'ORF' and 'in-phase hexamer' measures gave 82.4% accuracy on phase-coding human 54 base windows and 87.8% on 108 base windows. Accurate recognizers of coding gene regions based on neural network approaches have also been demonstrated recently (4–5).

A typical way of finding a functional signal is to search for similarity to a consensus sequence or by measuring the fit using a weight or neural network matrix. Such matrices can present the information about the sequences from both sides of the conserved dinucleotide (AG for the acceptor sites and GT for the donor sites) (6–8,20). Lapedes *et al.* (7) using the neural network matrices achived accuracy of 94% and 91% for predicting donor and acceptor splice sites, respectively. The most accurate splice site prediction using a neural network approach shows (8) that 95% of the true donor and acceptor sites can be detected. It means that on average there are one and a half false donor sites per true donor site and six false acceptor sites per true acceptor site. The neural network method performed better than the weight matrix method of Staden, and a network that combined the detection of coding/noncoding regions and splice junction detection significantly reduced the number of false positive splice junction predictions (8).

During the last few years, several complex systems for predicting gene structure have been developed (4,9–13). These systems combine information about functional signals and regularities of coding or intron regions. On this basis, potential first, internal and terminal exons can be predicted and the top ranking combination of them will present a model of gene

*To whom correspondence should be addressed

structure. The programs *GRAIL* (4) and *SORFIND* (11) show only the positions of candidate exons and do not attempt to produce assembled genes. *SORFIND* reaches 59% accuracy of internal exons prediction (at both 5′ and 3′ splice junctions and in the correct reading frame). To date, *GeneModeler* (9), *GeneID* (10), *GeneParser* (12) and *XGRAIL* (13) are the often used integrated packages that predict gene structure from genomic DNA. The first two methods initially identify potential functional motifs, including start and stop codons, splice sites and poly(A) signals, then evaluate the assembled combination of the gene components by sequential filtering. *GeneID* can predict the true gene structure as a top ranking structure in only 14% of tested vertebrate gene sequences, and in only 54% identify the correct exons with correct splice boundaries (10). A dynamic programming approach (alternative to the rule-based approach) was suggested by Snyder and Stormo (12). It accomplishes an exhaustive and mathematically rigorous search for the globally optimal solution. A sequence is divided into exons and introns by finding the best internally consistent set of high-scoring exon and intron subsequences. Weights for the various classification procedures are determined by training a feed-forward neural nerwork 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% of the test genes were exactly predicted. The prediction quality decreases dramatically for terminal exons, which seems to require special consideration (12). However, accurate prediction of internal exons is very important, because a cloned genomic DNA fragment rarely contains the entire sequence of a gene.

The goal of our work is to develop a simple computational approach for revealing internal exon regions, based on an improved splice site recognition method.

METHODS

The data

The data set was taken from GenBank (Release 72) (14). It includes all DNA fragments of human genes that contain introns and have unambiguous exon assignments in the feature table. This set contains 692 sequences with 2037 donor splice sites and 2054 acceptor splice sites having the GT and AG conserved dinucleotide in flanking intron positions. We use only GT- and AG-containing splice sites because the occurrence of other dinucleotides in conservative splice site positions is very rare (5 of 794) (15). Limiting our data set to these sites probably reduces the influence of annotation error in our study. Analysis of 50 cases of 'GT-AG rule' exceptions shows that 21 of them to be due to annotation error and in 17 other cases the intron−exon assignment appeared to be essentially putative (15). Moreover, the striking similarity among the rare splice junctions that do not contain AG or GT indicates the existence of special mechanisms to recognize them (16). 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 number of pseudosites is 44 and 65 times more than the number of real donor and acceptor splice sites, respectively. The characteristics of sequences around splice sites and pseudosites were used for developing and testing human splice site recognition functions. The data were divided into 2 parts, a training set including 2/3 of all sequences, and a test set containing the remaining ones. The selected data is 5 times larger than that used in the analysis of Brunak *et al.* (8). We did not remove the homologous fragments that could increase prediction values, but we also did not remove groups of genes with alternative splicing sites [as in Brunak *et al.* (8)], that may decrease values of predictive accuracy because only the major way of splicing (and corresponding splice site positions) is usually annotated for each sequence.

The data set for computing octanucleotide preferences in coding and intron regions included all human gene sequences from GenBank. From this data corpus, the regions of genes that are in the test splice site data set were removed. The remaining data set includes 4074593 bases of coding regions and 1797572 bases of intron sequences.

The data set for distinguishing potential exon and pseudoexon ORF sequences contains 952 exons and 528480 pseudoexons in the training set and 451 exons and 246693 pseudoexons in the test set.

The data set for distinguishing exon−exon junction positions in cDNA includes 1123 exon−exon junctions and 262264 other positions in the training set; and 517 exon−exon junctions and 137438 other positions in the test human mRNA sequences.

Discriminant analysis

Recognition of authentic splice sites (from the population of all AG or GT base dinucleotides) or authentic exons (from all open reading frames (ORF) beginning with AG and terminating by GT dinucleotides) was performed by the technique of linear discriminant analysis. Characteristics of different functional parts of splice site and exon sequences may have different weights in the recognition function reflecting their relative significance for recognition. We have applied the technique of discriminant analysis to relate a given sequence which has particular values of the p measures (or 'characteristics') $x_1, \ldots x_p$ to one of two alternative classes: class 1 (sites or exons) or class 2 (pseudosites or pseudoexons) (14). The term 'pseudo' is applied to all GT and AG dinucleotides, which are not splice sites. Analogously, all open reading frame fragments flanked by GT and AG dinucleotides which are not real exons are refered to as pseudoexons. The procedure of linear discriminant analysis is to find a linear combination of the measures (called the linear discriminant function or LDF), that provides maximum discrimination between real and pseudosites (or pseudoexons). The linear discriminant function:

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

classifies $x(x_1, \ldots, x_p)$ into class 1 if $z \geq c$ and x into class 2 if $z < c$. The vector of coefficients α $(a_1, a_2, \ldots a_p)$ and threshold constant c are derived from the training set by maximizing the ratio of between-class variation of z to within-class variation and are equal to (17):

$$\bar{\alpha} = S^{-1}(\bar{\mu}^1 - \bar{\mu}^2) \qquad \text{(EQ 2)}$$

and,

$$c = \frac{\bar{\alpha}(\bar{\mu}^1 - \bar{\mu}^2)}{2}, \qquad \text{(EQ 3)}$$

where μ^1 and μ^2 are the sample mean vectors of characteristics for class 1 and class 2, respectively:

$$\mu_i^l = \frac{1}{n_l}\sum_{k=1}^{n_l} x_{ik}^l . \qquad \text{(EQ 4)}$$

$S\{s_{ij}\}$ is the pooled covariance matrix of characteristics:

$$s_{ij} = \frac{1}{n_1 + n_2 - 2}\sum_{l=1}^{2}\sum_{k=1}^{n_l}(x_{ik}^l - \mu_i^l)(x_{ik}^l - \mu_j^l) \qquad \text{(EQ 5)}$$

(i,j = 1,..,p); $x^l ik$ is the value of characteristic i of k-th sequence in class l and n_l is the sample size of class l. Thus, based on these equations we can calculate the coefficients of LDF ($a_1, a_2, \ldots a_p$) and threshold constant c using the values of characteristics of site and pseudosite sequences from the training sets and then to test the accuracy of LDF on the test set data. Significance of a given characteristic or set of characteristics can be estimated by the generalized 'distance' between two classes (called the Mahalonobis distance or D^2) (17):

$$\bar{D}^2 = (\bar{\mu}^1 - \bar{\mu}^2)s^{-1}(\bar{\mu}^1 - \bar{\mu}^2), \qquad \text{(EQ 6)}$$

which is computed based on values of the characteristics in the training sequences of class 1 and 2.

Splice site recognition

The triplet composition of sequences adjacent to splice site positions is a good discriminant of splice sites (18–19). We can use it as a characteristic of the particular regions that define a splice signal.

The integral view on the difference of triplet composition in splice and pseudosplice sequences is shown with the 3-dimensional histograms in Figure 1. This illustration allows the visualization of the occurence of significant triplets in the flanking regions of the major splice consensus sequences.

We calculate the number of triplets in the (L,R) window around a conserved AG or GT dinucleotides, where L is the number position to the 5′-side, and R is the number position to the 3′-side of dinucleotides (L = 30, R = 50 bp for donor sites and pseudosites; and L = 80, R = 30 for acceptor sites and pseudosites). Each column in the figure presents the difference in the frequency for a specific triplet between authentic splice sites and pseudosites in a specific position relative to the conserved dinucleotide. We see that only short regions around splice junctions have a big difference in triplet composition between sites and pseudosites. However, dissimilarity in many other regions can also be seen: for donor sites, the coding region and G-rich intron region may be distinguished; for acceptor sites, the intron G-rich region, branch point region, poly(T/C)-tract, and coding region show significant difference between splice site and pseudosplice regions. We have applied some sequence characteristics of these regions for splice site recognition (21,22).

We tabulate the frequency of triplets, in the (L,R) window around a splice 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) in size. This matrix is computed for 1375 donor splice sites and for 60532 GT-containing pseudosites from the training set. The same is done for 1386 acceptor splice sites and 89791 AG-containing pseudosites from the training set.

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 (s) and pseudosite (p) sets of sequences in i-th position of a (L,R) window, respectively. Then the preference of a given triplet {k} in the i-th position of a splice site can be defined as:

$$P(i) = \frac{F_{s,k}^i}{F_{s,k}^i + F_{p,k}^i} \qquad \text{(EQ 7)}$$

For splice site discrimination we use the mean preference index obtained by averaging the preferences in the (L,R) window around any GT (for donor) or AG (for acceptor) dinucleotide of a sequence under analysis (eqn. 8), where j is the potential splice site position, corresponding to the G base of the GT or AG dinucleotide; m is the total number triplets and i is the position of a triplet within the (L,R) window:

$$P_{sp}(j) = \frac{1}{m}\left(\sum_{i=L}^{R} P(i)\right) \qquad \text{(EQ 8)}$$

Only a subset of all possible triplets is useful for splice site prediction. Therefore, the discrimination function is modified to take into account only those triplets which have a significant difference in the occurrence between splice sites and pseudosplice sites. If triplets are equally present in both types of regions, P(i) will be equal 0.5. For computing significant triplets, the summation in eq.8 is made if $(P(i)-0.5) > \alpha$, where α is a threshold value for significance, and m is the number of the significant triplets.

Pseudosites may be localized in intron as well as in exon regions. The significant difference of triplet composition between intron and coding regions is clear, therefore the recognition function will be more sensitive if the triplet composition of both cases is not represented in a single table. Two separate tables of triplet frequencies around pseudosplice junctions localized in either intron $F_{pi,k}$ or in coding $F_{pc,k}$ regions were calculated. For discrimination, the average value of eqn. 8 computed with each of these tables is used.

For the characteristics of intron and coding regions adjacent to splice sites, we use octanucleotide composition statistics (19). If the sequence S is defined as:

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

then

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

describes an oligonucleotide of length L.

For discriminating coding and noncoding regions, we can use the probability that oligonucleotide s_k is coding as estimated by the Bayesian method:

$$P(C|s_k) = \frac{P(s_k|C)P(C)}{P(s_k|C)P(C) + P(s_k|N)P(N)} =$$

$$\frac{F_c(s_k)}{F_c(s_k) + F_n(s_k)} \qquad \text{(EQ 9)}$$

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

We can consider oligonucleotides only in phase with coding regions (during learning on coding sequences), i.e. consider the oligonucleotides beginning with the first position of codons. A discriminant function analogous to Eq. 9 based on such in-phase oligonucleotides is:

$$P^1(C|s_k) = \frac{F^1(s_k|C)}{F^1(s_k) + F(s_k|N)} \qquad \text{(EQ 10)}$$

The simplest discriminant index for predicting a coding region is the average of Eq. 9 or Eq.10 along a sequence window W:

$$P(C|W) = \frac{1}{m}\left(\sum_{i=1}^{n} P(i)\right), i = 1,\ s+1,\ 2s+1,\ \ldots \qquad \text{(EQ 11)}$$

where $P(i)$ is $P(C|s_k)$ 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.

Discriminant function for splice site recognition

We combine the characteristics of various parts of splice site regions (Figure 1) in a linear discriminant function. The characteristics used for classifying donor sites are: (1) the average triplet preferences (eqn. 8) in the potential coding region (−30 to −5), (2) conserved consensus region (−4 to +6), and (3) G-rich region (+7 to +50); (4) the number of significant triplets in the conserved consensus region ($a = 0.15$ in the eqn. 8); (5) the octanucleotide preferences (eqn. 10) for being coding in the (−60 to −1) region and (6) being intron in the (+1 to +54) region; and (7) the number of G-bases, GG-doublets and GGG-triplets in +6 to +50 region.

The characteristics used for classifying acceptor splice sites are: (1) the average triplet preferences (eqn. 8) in the branch point region (−48 to −34), (2) poly(T/C)-tract region (−33 to −7), (3) conserved consensus region (−6 to +5), and (4) coding region (+6 to + 30); (5) the octanucleotide preferences (eqn. 6) of being coding in the (+1 to +54) region and (6) in the (−1 to −54) region; and (7) the number of T and C in poly(T/C)-tract region. The values of these characteristics were calculated for the training set and the parameters $\bar{\alpha}$ of the discriminant function were computed based on them. Then the accuracy of the discriminant function was estimated on the test data set. The search for splice site positions starts from finding GT or AG dinucleotides and the discriminant functions estimate assigning them to donor or acceptor splice sites, respectively.

Discriminant function for internal exons recognition

We consider all open reading frames in a given sequence that are flanked by AG (on the left) and GT (on the right) dinucleotides as potential internal exons. The structure of such exons is presented in Figure 2. As components of the internal exon discriminant function we take the octanucleotide composition

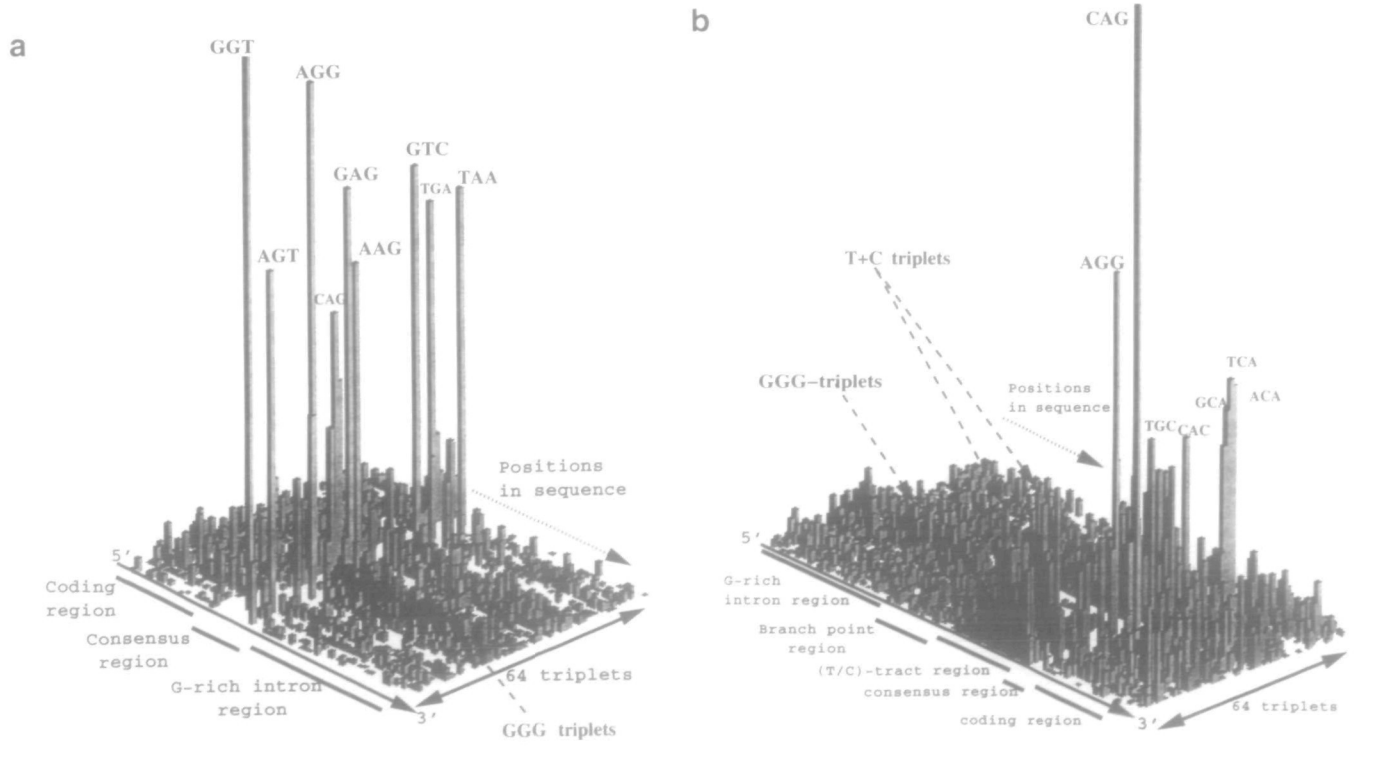

Figure 1. Difference of the triplet composition around donor and GT-containing pseudodonor sites (**left**); around acceptor and AG-containing pseudoacceptor sites (**right**) in 692 sequences of human genes. Each column presents the difference of specific triplet frequences between authentic sities and pseudosites in a specific position.

preferences for the intron 70 bp to the left of the potential acceptor site, the value of the acceptor splice site recognition function, the octanucleotide composition preferences for coding of the ORF, the value of the donor splice site recognition function, and the octanucleotide composition preferences for intron 70 bp to the right of the potential donor site.

The values of these characteristics were calculated for the training set and the parameters $\vec{\alpha}$ of the discriminant function were computed based on them. Then the accuracy of the discriminant function was estimated on the test data set.

Discriminant function for exon−exon junction recognition in cDNA

Recognition of exon−exon junctions in cDNA may be very useful for gene sequencing when starting with a sequence of cDNA clone. The essence of the mapping strategy is as follows. In a given cDNA sequence one selects sites for PCR primers that (hopefully) lie in adjacent exons. Then PCR used to amplify the intron that lies between the sites. The amplified DNA is sequenced (i) to confirm that we have, in fact, amplified the expected product, and (ii) to allow selection of a second primer set. Accurate prediction of exon−exon junctions in cDNA improves primer selection in internal exon sequence.

An approach for identifying exon−exon junctions is to look for the remnants of donor (**MAG/GURAGU**) and acceptor (**YAG/G**) consensus sequences that remain in the mRNA (16,23); i.e. **MAG/G** sequence. This consensus is only found in 25% of authentic exon−exon junctions, and at the same time per each consensus belonging to an authentic junction, we will predict about 15 false ones. We use the information about adjacent to consensus sequences to reduce these false predictions.

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

Another variant of the recognition function uses the same components of triplet preferences, but they are computed based on three matrices of triplet composition depending on occurrences of junction conserved triplets. **AG/G** (that found in 28% of exon−exon junctions), **AG/G** with **1** mismatch (70.41%) and **AG/G** with **2** mismatch (95%) were considered. Triplet preferences (eqn. 7) were computed using triplet frequencies of mRNA regions around authentic junction positions and non-junction positions of mRNA that contain the triplet variants describing above.

The search for exon−exon junctions begins with finding a junction triplet variant in a given mRNA sequence, and then estimating the exon−exon junction triplet preferences.

The values of these characteristics were calculated for the training set and the parameters $\vec{\alpha}$ of the discriminant function were computed based on them. Then the accuracy of the discriminant function was estimated on the test data set.

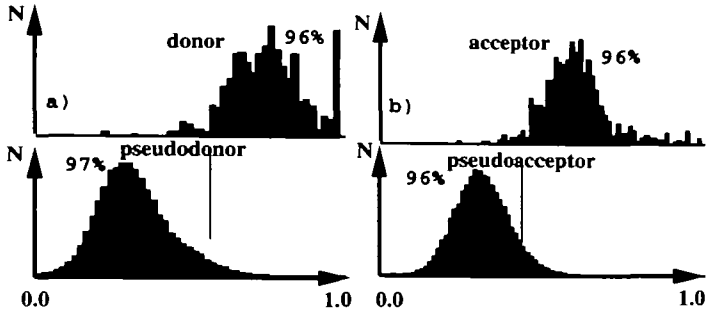

Figure 2. Various functional regions of internal exon corresponding to components of the recognition function.

RESULTS AND DISCUSSION

Splice site prediction

We applied linear discriminant analysis to the development of a splice site recognition function. The values of 6 characteristics of donor site were calculated for 1375 authentic donor sites and 60532 pseudosite sequences from the learning set. The Mahalonobis distances showing the significance of each characteristic are given in Table 1a. The most significant characteristic, as measured by the value of D^2, is combined separately with each of the remaining characteristics to yield a new combined D^2. The additional characteristics giving the largest D^2 are then included in the selected set. The cycle is repeated with the remaining characteristics. Following this procedure, characteristics were included in the discriminant function in the order presented in Table 1b, which also shows the increase of the combined D^2 with subsequent addition of each characteristic. The strongest characteristic for donor sites is triplet composition in the consensus region (D^2 = 9.3) following by 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 in intron G-rich region; the quality of the coding and intron regions. Each of the last four characteristics increases the total D^2 of discrimination between sites and pseudosites by about 0.5.

The accuracy of the discriminant function based on these characteristics was tested on the recognition of 662 donor sites and 28855 pseudosite sequences not included in the training set.

Table 1. Significance of various characteristics of donor sites

Characteristics[a]	1	2	3	4	5	6	7
(a) Individual D^2	9 25	2.64	2.47	0.01	1.53	0.01	0.41
(b) Combined D^2	9.25	11.79	13.55	14.92	15.49	16.56	16.78

[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 54bp region on the right of donor splice site junction; 7 is the number of G bases, GG-doublets and GGG-triplets in intron G-rich region (see more detail description of the characteristics in Methods).

Table 2. Significance of various characteristics of acceptor sites

Characteristics[a]	1	2	3	4	5	6	7
(a) Individual D^2	5.14	2.63	2.71	2.34	0.01	1.05	2.41
(b) Combined D^2	5.14	8.08	9.98	11.33	12.50	12.82	13.64

[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 for being intron 54bp region on the right of donor splice site junction; (see more detail description of the characteristics in Methods).

The histograms of the LDF function value distribution are shown in figure 3a. The general accuracy of donor site prediction is 97% (C = 0.63). The neural network-based method has C = 0.61 at 95% accuracy (8). C is an important accuracy criterion (correlation coefficient) that takes into account the relation between true positives and negatives as well as false positives and negatives predictions (24):

$$C(X) = \frac{(P_t N_t - P_t^f N_t^f)}{\sqrt{(N_t + N_t^f)(P_t + P_t^f)(N_t + P_t^f)(N_t^f + N_t)}} \qquad \text{(EQ 12)}$$

Here P_x and N_x are the correctly predicted positives and negatives, and $P^f x$ and $N^f x$ are similarly the incorrectly predicted positives and negatives.

The values of 7 acceptor site characteristics of were calculated for 1386 authentic acceptor site and 89791 pseudosite sequences from the learning set. The D^2 showing the individual significance for each characteristic are given in Table 2a. Table 2b shows the increase of the combined D^2 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 for 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). The triplet composition of the G-rich region before the branch point position did not significantly increase the quality of acceptor sites recognition, and we did not include it in discriminant function. The accuracy of the discriminant function based on these significant characteristics was tested on the recognition of 666 acceptor sites and 43726 pseudosite sequences not included in the training set. The histograms of the LDF function value distribution are shown in Figure 2b. The general accuracy of acceptor site prediction is 96% (C = 0.47). This accuracy is better than in the neural network-based method, which has C < 0.40 at this level of acceptor site recognition (8).

We have demonstrated improved accuracy of splice site recognition by using a combined classification scheme that considers both characteristics of various regions of authentic splice site sequences as well as adjacent protein coding and intron sequences. Using discriminant analysis we have shown the relative significance of these regions for recognition and applied it to exon recognition.

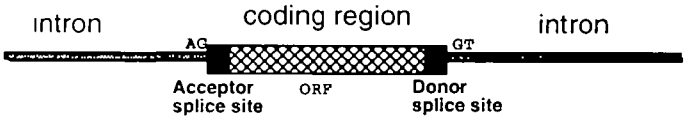

Figure 3. The histograms of combined discriminant fuction values distribution for 662 donor sites and 28885 pseudosites (a) and for 668 acceptor sites and 44359 pseudosites from sequences. (b). The vertical axis is the number of sequences with a specific weight, the horizontal axis is the weight values.

Table 3. Significance of various characteristics of internal exon regions

Characteristics[a]	1	2	3	4	5
(a) Individual D^2	15.04	12.06	0.41	0.18	1.47
(b) Combined D^2	15.04	25.32	25.77	25.82	25.89

[a]1 and 2 are the values of donor and acceptor site recognition functions; 3 and 4 are the octanucleotide preferences for being intron 70bp region on the left and 70bp region on the right of potential exon region; 5 is the the octanucleotide preferences for being coding of potential exon region (see more detail description of the characteristics in Methods).

Internal exon prediction

We have applied linear discriminant analysis to the development of an internal exon recognition function. The values of 5 exon characteristics were calculated for 952 authentic exon and 528480 pseudoexon sequences from the learning set. The D^2 showing the significance of each characteristic are given in Table 3a. Table 3b shows the increase in the combined Mahalanobis distance by subsequently adding each characteristic. 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 last three characteristics do not significantly increase recognition suggesting that splice site sequences play the main role in exon recognition.The accuracy of the discriminant function based on these characteristics was calculated from 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 results are analyzed at individual nucleotides level, the accuracy of exon prediction is 89% with specificity 89%; intron position prediction is 98% with specificity 98%. The combined dynamic programming and neural network-based method (12) described earlier have 75% accuracy of the exact internal exons prediction with specificity 67%. Our method has 17% less false exon assignments with the better level of true exon prediction.

Table 4. Accuracy of internal exon prediction for some genes of the test set

GenBank name of gene	Number of annotated exons	Number of correctly predicted exons	Number of partially predicted exons
HUMA1ADG	9	5	3
HUMALDC	6	6	0
HUMALPI	9	7	0
HUMCAD	10	7	1
HUMC1IA1P	5	5	0
HUMCOL2A1G	12	11	1
HUMCP21OH	8	8	0
HUMCYP2DG	7	6	1
HUMCYPIIE	7	5	0
HUMEF1A	5	4	1

Table 5. Prediction of splice site position in cDNA

	Sn (%)	Consensus	LDF[a] 1	LDF 2
		Number of false positive predictions per one correct prediction		
MAGG	25	14	5	0.7
AGG	29	19	8	0.9
MAGG*	59	50	28	15
AGG*[b]	70	68	42	27

[a]LDF—linear discriminant function
[b]*—means that the consensus can have 1 mismatch;
Sn—percent of true prediction (sensitivity).

Another method specially designed for internal exon prediction (11) shows 59% accuracy of exact exon prediction with specificity of 34.5%. This result was obtained on a test set of 80 internal exons. We have observed an accuracy (77%/79%) analysing 451 test exon sequences. The prediction of internal exons for a sample set of genes using our recognition functions are shown in Table 4.

Exon–exon junction prediction in cDNA

The values of 2 characteristics were calculated for 1123 exon–exon junctions and 262264 other positions in cDNA of human gene sequences from the training set. The D^2 of the first characteristic (triplet preferences in 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 D^2 of the both characteristics is 6.1. This result shows that some information about exon–exon junctions remains in the mRNA sequence and may be used for predicting their positions. However the information content of mRNA is much less than observed for pre-mRNA, where the Mahalonobis distance of splice site discrimination is about 16. Therefore, many false splice site position predictions can be expected in cDNA analysis. We compared the quality of our discriminant function with prediction of exon–exon junctions using some consensus sequences: MAGG, AGG, MAGG with 1 mismatch, and AGG with 1 mismatch (Table 5). For the discriminant functions 1 and 2, the level of false prediction was calculated with the level of true prediction the same as for a given consensus sequence. For a particular level of true prediction (that corresponds to 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 values of our discriminant functions, it is possible to create a profile of probability (25) of being an exon–exon junction for any position in a given cDNA sequence. Primer subsequences can be selected in the regions with minimal values of these probabilities. An example of such a profile for the cDNA of ribosomal protein gene (GenBank entry HSRPS7) is shown in Figure 4. We can see that the primer sequences can be selected within the regions with the probability about zero and such primers will not overlap an adjacent exons pair.

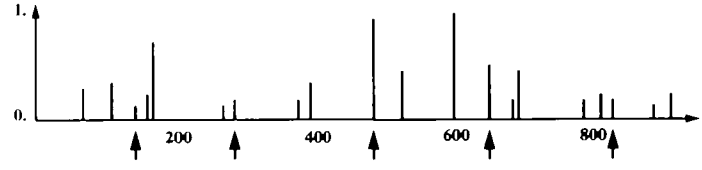

Figure 4. Profile of the probability belong to a exon–exon junction for the cDNA of ribosomal protein gene HSRP7. The vertical axis is the value of probability of a particular position belong to a exon–exon junction; the horizontal axis is the position in the analysed sequence. The positions of true exon–exon junction are marked by arrows.

SUMMARY

We have demonstrated improved accuracy of splice site and internal exon recognition by using a combined classification scheme that considers both characteristics of various regions of authentic splice site sequences as well as adjacent protein coding and intron sequences. Using discriminant analysis we have shown the relative significance of these regions for recognition. One of the advantages of our approach is that we can recalculate the

tables of triplets to obtain increasingly reliable statistics as the size of the sequence data base increases. Also, the triplet tables may be calculated for each class of organisms and will be used for splice site selection of their genes.

The approach described in this article has also been applied to developing 5′- and 3′-exon discriminant functions (22).

Some of the 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 version of such a system has been developed (26). 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 9 and 11. 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 212 complete human gene sequences shows that it can predict 80% of all exons with 70% specificity and that 96% of the exons are predicted partially. The detailed description of this method will be published elsewhere. A test of the performance of the latest versions of the most successful gene prediction programs, *GeneModeler*, *GeneId*, *Grail* and *GeneParser*, shows that they have an accuracy of correct exon prediction of 0.02, 0.33−0.42, 0.31−0.52 and 0.47, respectively (27). This shows that the gene prediction problem requires further investigation.

The algorithm for prediction of exon−exon junctions in cDNA may increase the effectiveness of primer selection for gene mapping by PCR reaction.

Analysis of uncharacterized human sequences based on our methods for splice site (*HSPL*), internal exons (*HEXON*), all type of exons (*FEXH*) and gene structure (*FGENEH*) prediction is available using a network server by sending the file containing a sequence (the sequence name in the first string) to *service@.bchs.uh.edu* with the subject line *hspl, hexon, fexh or fgeneh*.

ACKNOWLEDGEMENTS

This work was supported by the W.M.Keck Center for Computation Biology, a grant to C.B.L. from the Department of Energy and an award from National Center for Human Genome Research (NIH) to V.V.S. The authors are grateful to N.Goodman for bringing their attention to the primer selection problem. S.Honda and the referees for very helpful comments on the first version of this manuscript.

REFERENCES

1. Moore, M.J., Query, C.C., and Sharp, P.A. (1993) In *The RNA world* (R. Gesteland and J. Atkins, eds.), New York: Cold Spring Harbor Laboratory Press, 303−357.
2. Stormo, G.D. (1987) In *Nucleic acid and protein sequence analysis* (Bishop M.J. and Rawlings C.J. eds) IRL Press, Oxford, 231−258.
3. Fickett, J.W.; Tung, C.S. 1992. *Nucl. Acids Res.* **20**, 6441−6450.
4. Uberbacher, E.C.; Mural, R.J. (1991) *Proc. Natl. Acad. Sci.USA* **88**, 11260−11265.
5. Farber, R.; Lapedes, A.; Sirotkin, K. (1992) *J. Mol.Biol.* **226**, 471−479.
6. Nakata, K.; Kanehisa, M.; DeLisi, C. (1985) *Nucl.Acids Res.* **13**, 5327−5340.
7. Lapedes, A.; Barnes, C.; Burks, C.; Farber, R.; Sirotkin K. 1988. *Application of neural network and other machine learning algorithms to DNA sequence analysis*. In Proceedings Santa Fe Institute 7: 157−182.
8. Brunak, S.; Engelbreht, J.; Knudsen, S. (1991) *J. Mol.Biol.* **220**, 49−65.
9. Fields, C.; Soderlund, C.A. (1990) *CABIOS* 6: 263−270.
10. Guigo, R.; Knudsen, S.; Drake, N.; Smith, T. (1992) *J. Mol.Biol.* **226**, 141−157.
11. Hutchinson, G.B., Hayden, M.R. (1992) *Nucl.Acids Res.* **20**, 3453−3462.
12. Snyder, E.E., Stormo, G.D. (1993) *Nucl. Acids Res.* **21**, 607−613.
13. Uberbacher, E.C., Einstein, J.R., Guan, X., Mural, R.J. (1993) In: *The second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*. (Lim, H., Fickett, J., Cantor, C.R., Robbins, R.J., eds). World Scientific, London, 465−476.
14. Cinkosky, M.J.; Fickett, J.W.; Gilna, P.; Burks C. (1991) *Science* **252**, 1273−1277.
15. Penotti, F.E. (1991) *J. Theor.Biol.* **150**, 385−420.
16. Senapathy, P.; Shapiro, M.B.; Harris, N.L. (1990) *Methods of Enzymology* (ed. R.F. Doolittle) **183**, 252−280.
17. Afifi, A.A., Azen, S.P. (1979) *Statistical analysis. A computer oriented approach*. Academic Press, New York.
18. Mural, R.J.,Mann, R.C., Uberbacher, E.C. (1990) In: *The first International Conference on Electrophoresis, Supercomputing and the Human Genome*. (Cantor C.R., Lim H.A. eds). World Scientific, London, 164−172.
19. Solovyev, V., Lawrence,C. (1993) In: *The First International conference on Intelligent systems for Molecular Biology* (eds. Hunter L., Searls D., Shavlic J.), NLM HIH, Bethesda, 371−379.
20. Stephens, R., Schneider, T.D. *J. Mol.Biol.* **228**, 1124−1136.
21. Solovyev, V.V.; Lawrence, C. (1994) *CABIOS* (submitted).
22. Solovyev, V., Salamov, A.A., Lawrence,C. (1994) In: *The Second International conference on Intelligent systems for Molecular Biology* (eds. Altman, R., Brutlag, D., Karp, P., Lathrop, R., Searls D.), Stanford Univ., Stanford, CA, 354−362.
23. Mount, S.M. (1993) In *An Atlas of Drosophila genes*. (ed. Maroni G.), Oxford, 333−358.
24. Matthews, B.W. (1975) *Biochem.Biophys.Acta* **405**, 442−451.
25. Lawrence, C.B., Solovyev, V.V. (1994) *Nucl. Acids Res.*, **22**, N 7, 1272−1280.
26. Solovyev, V., Lawrence, C. (1993) In: *Abstracts of the 4th annual Keck symposium*. Pittsburgh, 47.
27. Snyder, E.E., Stormo, G.D. 1994. In *Nucleic Acid and Protein sequence analysis: A practical Approach*, Second edition (in press).
28. Staden R. (1990) In *Methods of Enzymology* (ed. R.F.Doolittle), **183**, 163−180.