Identification of Human Gene Functional Regions Based on Oligonucleotide Composition

Victor V. Solovyev and Charles B. Lawrence

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

Abstract

Accurate recognition of coding and intron regions within large regions of uncharacterized genomic DNA is an unsolved problem. A data base of more than 4240791 bp coding and 7790682 bp noncoding human sequences was extracted from GenBank to develop a function for locating coding regions in anonymous sequences. Several coding measures based on oligonucleotide preferences were tested on a control set that including 1/3 of all extracted sequences. An accuracy of separation of coding/noncoding regions is 87% for 9 bp oligonucleotides on 54 bp windows and 91% on 108 bp windows, respectively. For separation of coding/ intron regions the accuracy is 89-90% for 8 bp oligonucleotides on 54 bp windows and up to 95% on 108 bp windows. Using the information about preferences of octanucleotides in protein coding and intron regions and significant triplet frequencies as a function of position near splice junctions, a joint splice site prediction scheme was developed. The accuracy of the joint scheme for predicting splice site positions on the test set was about 96-97%, which exceeds the accuracy of the previously reported splice site selection method based on a more complex artificial neural network approach. A model of splicing using poly-G(C) rich exon flanking sequences is suggested. A remarkable difference of oligonucleotide composition 5'- and 3'- gene regions is displayed and applied in a gene structure predictive system.

Introduction

Large scale genome sequencing projects have given rise to a unique set of problems connected with the analysis of large quantities of genetic sequence data. The prediction of exon-intron gene structure and discovering functional motifs is one of the most important roles of the computer analysis of the newly sequenced DNA. Many methods and algorithms have been suggested for recognizing of the components of gene structure. Currently there are two general approaches used for finding protein coding regions (see reviews by Stormo 1987; Staden 1990). 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. Comprehensive assessment of various protein coding measures was done by Fickett and Tung (1992). They estimate the quality of more than 20 measures and showed that most powerful - such as 'in phase hexanucleotide composition', codon and amino acids usage - can give up to 81% accuracy as coding region recognition functions on 54 base windows. Also, an interesting result was the higher accuracy of a recognition function based on using Penrose or classical linear discriminant analysis than an extraction of information from the coding measure in some other way. 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 (Uberbacher, Mural 1991; Farber *et al.* 1992).

A typical way to find functional signals is with consensus sequences, a consensus matrix or neural network matrix. The most favorable accuracy of donor and acceptor splice junction recognition was based on a neural network matrix that used information about a splice junction surrounding a sequence of specified length from both sides of the highly conserved dinucleotide (AG for the acceptor sites and GT for the donor sites) (Nakata *et al.* 1985; Lapedes *et al.* 1990; Brunak, Engelbreht & Knudsen 1991). An accuracy of 94% was achieved for 11-21 bp windows surrounding the conserved dinucleotides for predicting donor sites and 91% for predicting acceptor sites using windows 41 bp length (Lapedes *et al.* 1990). It must be mentioned that the problem is complicated by the possibility of alternative splicing (and splicing sites). Careful analysis of the splice site prediction by the neural network approach shows (Brunak *et al.* 1991) that 95% of the true donor and acceptor sites detected 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. Although the neural network method performed better than the weight matrix method of Staden, but only a network that combined the detection of coding/noncoding regions and splice junction detection reduced the number of false positive splice juction predictions (Brunak *et al.* 1991).

During last few years, some complex systems for predicting gene structure have been developed (Fields and Soderlund 1990; Uberbacher, Mural 1992; Guigo *et al.* 1992). 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 revealed and the top ranking combination of them will present the predicted gene structure. However, the problem requires further investigation. For example, the testing of the latter algorithm (Guigo *et al.* 1992) on an independent data set shows that in only 15 cases (54%) were correct exons with correct splice boundaries were predicted (Guigo *et al.* 1992). It must be mentioned that using this system 84% of the predicted coding region is actually coding.

The goal of our work is to develop a computational approach of revealing different human gene regions based on oligonucleotide composition, which may be simply for updated with new sequence data and applied to other species of organisms.

Recognition of Protein Coding Regions

Fickett and Tung (1992) showed that based on the linear discriminant function the in phase hexanucleotide composition of coding and noncoding regions is one of most powerful measures by which we can make the coding/noncoding decision. Although the authors discussed that applying LDA (or Penrose discriminant) give the higher accuracy than the extraction of information from the measure some other way, we supposed that LDA is probably useful for combining some different measures and carried out the analysis of various measures of oligonucleotide composition itself.

The Data

For estimating the usefulness of coding measures, fully coding and fully noncoding human sequences were taken from GenBank (release 72) (Cinkosky *et al.* 1991). The data corpus was divided into 2 parts, the part used for training included 2/3 of all sequences, and the part used for testing included the remaining ones. Our measures were tested on 54 and 108 bp. windows. The noncoding sequences were used as a whole and introns were used as separate set itself. Table 1 shows the numbers of human sequences used as training and test sets.

The Methods

We define the sequence S by:

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

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 coding and noncoding regions we can use the probability of oligonucleotide s_k being coding as estimated by the Bayesian method:

$$P(C|s_k) = \frac{P(s_k|C)P(C)}{P(s_k|C)P(C) + P(s_k|N)P(N)} = \frac{F_c(s_k)}{F_c(s_k) + F_n(s_k)} \quad \text{(EQ 1)}$$

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. 1 based on such in-phase oligonucleotides:

$$P^1(C|s_k) = \frac{F^1(s_k|C)}{F^1(s_k) + F(s_k|N)} \quad \text{(EQ 2)}$$

The most simple discriminant index for revealing a coding region is the average of Eq. 1 or Eq. 2 along a sequence window W:

$$P_\alpha(C|W) = \frac{1}{m}\left(\sum_{i=1}^{n} P(i)\right), i = 1, s+1, 2s+1, \ldots \quad \text{(EQ 3)}$$

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 .

The probability of a sequence window W being coding if one assumes that the probabilities of a oligonucleotide appearing in i-th position is independent of the oligonucleotides appearing in other positions in the sequence:

$$P_p(C|W) = \left(\prod_{i=1}^{n} P(i)\right), i = 1, s+1, 2s+1, \ldots \quad \text{(EQ 4)}$$

Sequence Set	# in set	Number of base pairs in set: Training	Testing	Total	# 54bp windows
Noncoding	11,007	5,221,893	2,568,789	7,790,682	
Coding	3917	2,826,946	1,413,845	4,240,791	25,575
Introns	4963	2,880,822	1,494,204	4,375,026	26,003

TABLE 1. Summary of data in human training and testing sets (coding sequences were taken in 5' to 3' orientation; introns include direct as well complementary sequences; noncoding sequences include intergenic regions and introns)

The second discriminant index we define as:

$$P_C(C|W) = \frac{P_p(C|W)}{P_p(C|W) + P_p(N|W)}. \quad \text{(EQ 5)}$$

The third discriminant index is constructed based on a likelihood ratio. The likelihood ratio "in favor of coding region" is the ratio

$$L(s_k) = \frac{P^1(C|s_k)}{P(N|s_k)},$$

For discrimination we use the mean value of $ln(L_k(i))$ obtained by averaging in L size window:

$$L(W) = \frac{1}{L}\left(\sum_{i=1}^{L} ln(L(i))\right), i = 1, s+1, 2s+1... \quad \text{(EQ 6)}$$

Discriminant indexes similar to Eq. 1 and Eq. 2 were used in previous investigations (Claverie and Bougueleret 1986; Claverie et al. 1990; Farber et al. 1992), but were not tested on the same data set.

Coding Region Classification

Oligonucleotide frequencies were calculated for coding, noncoding and intron sequences. The accuracy of coding and noncoding region recognition using Eqs. 1 and 3 and training set oligonucleotides frequencies is shown in Table 2.

Window size	Oligonucleotide size		
	6	7	8
54	70.5%	72%	74%

TABLE 2. Accuracy of recognition of coding and noncoding regions. Accuracy was computed as average true prediction of coding and noncoding sequences.

One can see that Eq. 3 using only oligonucleotide frequencies gave the same accuracy for hexamers as Penrose discriminant function (PD) for the best measure tested by Fickett and Tung (1992). The accuracy increases with increasing oligonucleotide length. As the most powerful recognition was based on in-phase oligonucleotides of a coding region (Claverie and Bougueleret 1990; Fickett and Tung 1992), the further analysis was done with the frequencies of oligonucleotides in-phase computed based on the coding region train set.

The accuracy of the classification of coding and noncoding regions based on Eqs. 2 and 3 are shown in Table 3.

The result of classification using oligonucleotides of 8 and 9 bp length is better then the accuracy for the combined six most powerful measures and LDA function in the Fickett and Tung (1992) investigations. They obtained 82.4% accuracy on phase-coding human 54 base windows and 87.8% accuracy on phase-coding human 108 base windows.

Window size	Oligonucleotide size	
	8	9
54	85.5%	87%
108	91%	91%

TABLE 3. Accuracy of recognition of coding and noncoding regions using in-phase oligonucleotides.

We tested these measures on the separation of coding and intron sequences, because it has a great value when we localize the exon and intron boundaries within a gene sequence.The frequencies of oligonucleotides then were computed based on the intron learning set. The results of classification using oligonucleotides 8 and 9 bp length are presented in Table 4.

Window size	Oligonucleotide size	
	8	9
54	89%	90%
108	93%	93.5%

TABLE 4. Recognition of coding and intron regions using in-phase oligonucleotides and Eq.4.

The accuracy of recognition for oligonucleotides 8 and 9 bp. length is approximately the same. We suppose that 8-9 bp lengths of oligonucleotides are the upper limits for reliable statistics of their frequencies on the current size of the data base. It is interesting that if we use the intron oligonucleotide frequencies for classifying coding and noncoding region sets, the accuracy is the same as using noncoding region oligonucleotide frequencies. This may support the suggestion that most errors in classification appear due to the occurrence junk parts of coding regions in introns.

The results of testing Eqs. 3, 5 and 6 for 54 and 108 bp windows are presented in Figure 1. If we apply Eq. 5 for the classification, the separation of coding and noncoding regions has approximately the same accuracy, but the distribution of coding and noncoding window weights seems more reliable for recognition (Figure 1b). We can see that almost all coding regions have high weights far from the majority of noncoding region weights. The likelihood function (Eq. 6) gives a similar result (figure 1c). It is possible that some noncoding regions contain junk parts of coding gene regions that appeared as a result of evolutionary rearrangements during formation of gene structure. Such parts we can see as the tail of noncoding region weights distribution (Figure 1 b,c). For this reason, there are may be an upper limit for separating coding and noncoding regions based on any coding measure without considerating the functional signals of a gene (as splice sites, etc.). Another reason for the misclassification can be due to errors of information about some coding regions represented in GenBank. On the left corner of Figure 1 (b,c) we can see a number of coding regions with very low weights. Some of them may refer to incorrectly classified coding regions.

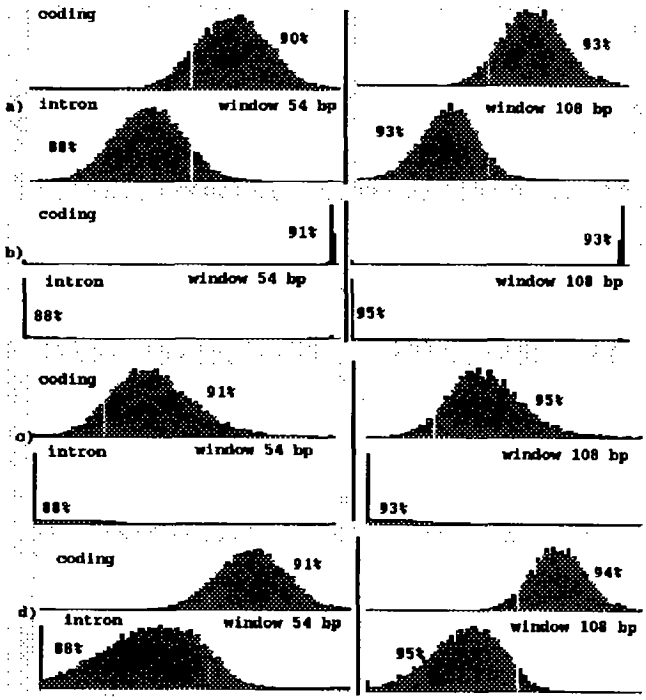

FIGURE 1. The distribution of windows weights for coding and noncoding regions 54 and 108 bp long based on 8 bp in phase oligonucleotides and functions a)(3); b)(5); c)(6); d) (3) with the weight of stop codon containing oligonucleotides is -0.7. Here and in susequent figures the horizontal axis is the weight of a window and the vertical axis is the number of windows with a specific weight.

Slightly higher accuracy of prediction for Eq. 5 and Eq. 6 seems due to more sensitivity to stop codons. If we assign the oligonucleotides that are absent in coding regions (containing stop codons) weight equal to 0.7, Eq. 3 will have the same accuracy as Eqs. 5 and 6 (Figure 1d).

We suppose that the improved accuracy of this approach is based on using statistics for longer oligonucleotides instead of hexamer as in previous studies. Also, we did not include complementary chains of coding region in noncoding sequences, which slightly increases recognition accuracy. This seems reasonable because the main task is to recognize and separate coding and noncoding (intron) regions along a given DNA chain.

Joint Function for Splice Site Prediction

Prediction of splice sites based on oligonucleotide composition is described in detail (Solovyev and Lawrence 1993). Here we briefly give only the main points of this scheme. We tabulate the frequency of oligonucleotides (duplets, triplets, *etc.*) in *(L,R)* window around a splice site, *L* is the number position on the left, *R* is the number position on the right from the exon-intron (or intron-exon) boundary. For triplets, their frequencies are written down in a matrix *(L+R, 64)* size. These matrices are computed for 1375 donor splice sites and for 60532 GT-containing pseudosites from the learning set. The same is done for 1386 acceptor splice sites and 89791 AG-containing pseudosites from the learning set. Let $F^i_{s,k}, F^i_{p,k}$ be the frequencies of k type oligonucleotides in the learning site and pseudosite sets of sequences in i-th position of the (L,R) window, then the preference of a given oligonucleotide k in i-th position belonging to a splice site can be defined as:

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

For splice sites discrimination we use the mean preference index obtained by averaging the preferences in the *(L,R)* window:

$$P_{sp}(W) = \frac{1}{m}\left(\sum_{i=L}^{R} P(i)\right) \qquad \text{(EQ 7)}$$

The summation is made if $(P(i) - 0.5) > \alpha$, where α is some threshold value for considering only significant oligonucleotides, and m is the number of significant oligonucleotides. We take 2 separate tables of triplet occurrences around pseudosplice junctions in intron $(F_{pi,k})$ and coding $(F_{pc,k})$ regions and compute the average value (Eq. 7) based on these 2 tables. For selection of the donor splice site in the i-th position of a given sequence from a pseudosplice region the following linear function is used:

$$F_d(i) = \gamma_1 \times P_{sp}(i) + \gamma_2 \times P_c(W_c) + \gamma_3 \times P_i(W_i), \qquad \text{(EQ 8)}$$

where P_{sp} is from Eq. 7, P_c is from Eq. 3, $P_i(W_i)$ is similar to Eq. 3, but calculated for average preference for being intron; W_c and W_i are windows on the right and on the left of the donor splice site position; i is the position of the first base in the intron; γ_2, γ_3 are the coefficients that give optimal prediction on test set; γ_1 is the coefficient that gives the function values between 0 and 1. A similar function is used for acceptor splice site selection with its own optimal parameters. Histograms of Eq. 8 value distribution with optimal parameters for 662 donor and 28885 pseudodonor sites from test sets (not including in learning sets) are shown in Figure 2a. The accuracy of prediction for the optimal threshold is 97% (C=0.61). This accuracy is the same as for the more complex neural network method (Brunak *et*

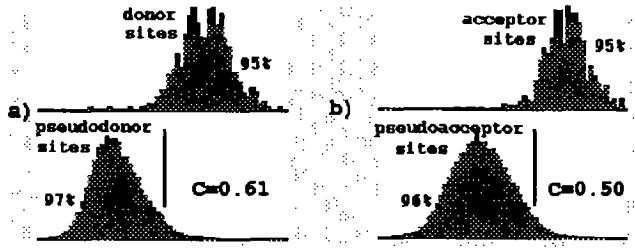

FIGURE 2. The histograms of joint-sheme weights distribution for 662 donor sites and 28885 pseudosites (a) and for 668 donor sites and 44359 pseudosites from testing sequences (b) using 2 matries with triplet composition of pseudosites in coding regions and introns.

al. 1991). For comparing our results with this work we compute an accuracy criterion (C) that takes the relation between correctly predicted positives and negatives as well as false positives and negatives into account (Mathews 1975):

$$C(X) = \frac{(P_x N_x - P_x^f N_x^f)}{\sqrt{(N_x + N_x^f)\,(P_x + P_x^f)\,(N_x + P_x^f)\,(N_x^f + P_x)}}. \quad \text{(EQ 9)}$$

Here P_x and N_x are the correctly predicted positives and negatives, and P_x^f and N_x^f are similarly the incorrectly predicted positives and negatives. The results of the joint approach for prediction of 668 acceptor splice sites and 44358 pseudosites from test sets are shown in Figure 2b. The accuracy of prediction is 96% (C=0.50). This accuracy is better than for the complex neural network method, where for 95% level of acceptor sites prediction C=0.40 (Brunak *et al.* 1991).

Detailed analysis of significant triplets around splice junctions will be published elswhere. Here we only mention that the most interesting new characteristic of authentic splice sites is the ocurrence of poly-G or poly-C oligonucleotides in the intron region flanking the donor site (at +7 to +45) and in the region flanking the acceptor site (at -65 to -27, upstream from the polypyrimidine tract). The numbers of GGG (or CCC) triplets in these regions are approximately 3-4 times higher than the expected values, especially taking into account the uncommonness of G(C)-rich subsequences in introns. We suggest that some RNA binding proteins have afiinity to themself and recognize G(C)-rich sequences near exon boundaries and bring together the ends of an intron (or introns) to create spatial proximity of splice signals (Figure 3). We cannot exclude that some complementary interactions between poly-G and poly-C sequences of different ends of an exon might take part in this process.

Thus, the results of this part of our work give the recognition functions for coding regions and splice sites that have a higher accuracy than found in previous studies. Developing statistics for them is much simpler than for neural network recognition functions or LDA functions. For further development of coding recognition methods we plan to use these functions as components of an automatic gene structure recognition system that will take into account the splice sites signal features.

Analysis of Oligonucleotide Composition of Various Gene Regions

Different oligonucleotides are specific to different gene regions (Volinia *et al.* 1989). We use the fractal graphical representation (Jeffrey, 1990; Solovyev *et al.* 1991; 1992; Solovyev 1991) for analysis of oligonucleotide composition of 5'-, exon, intron and 3'-regions of human genes.

Graphic Representation of Nucleotide Sequences

If we label the four corners of a square with the characters of the nucleotide alphabet: A, T, G, and C (Figure 4a) and plot the points corresponding to the consecutive nucleotides in a DNA sequence as follows:

1. The first base of the sequence is plotted halfway between the center of the square and the corner corresponding to the alphabet;
2. The next nucleotide is plotted halfway between the point just plotted and the corresponding corner;
3. Step 2 is repeated for all subsequent nucleotides in the DNA fragment.

If the fractal representation is constructed in a square with side 1, two sequences having k identical terminal

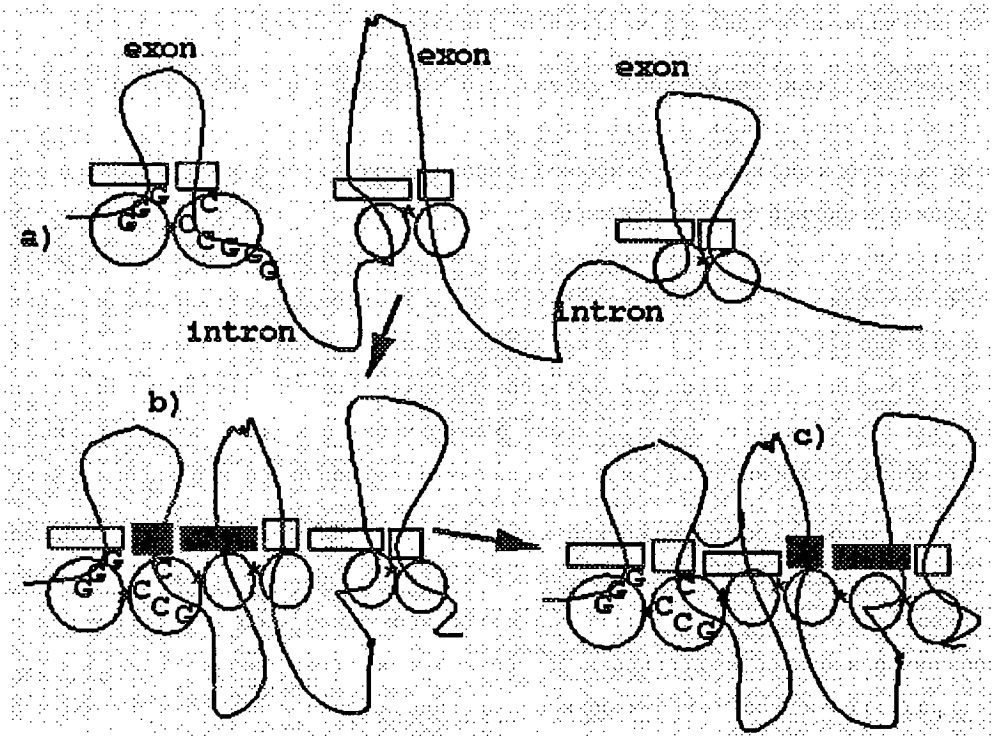

FIGURE 3. Model of splicing. a) The G(C)-rich sequences flanking a particular exon are usually closer in space and they could be bound at the first stage of splicing and create the basis for the exon defenition model (Robertson,Cote,Berget 1990); b) In the second stage, neighbour exons could be grouped altogether by G(C)-affinity proteins and exon ligation may take place.

nucleotides will always be in a subsquare of length 2^{-k} . (Jeffrey,1990).

Solovyev *et al.* (1991) suggested drawing all sequences having a similar function on the square and using this image for recognition of the function. The plot is characterized by an uneven density of points. A high density should be observed in conservative regions of the sequences while a low density should be observed in nonspecific oligonucleotide regions. If the square is subdivided into 4 cells (n=2), each cell will contain the dots corresponding to a particular nucleotide. For $n = 4, 8, \ldots 2^k$, each subsquare will contain the dots corresponding to a dinucleotide, trinucleotide, tetranucleotide,... oligonucleotide of k nucleotides long, respectively (Figure 4b). As some subsquares correspond to similar oligonucleotides in a set of sequences when the square is subdivided into cells, the more abundant an oligonucleotide, the higher the number dots P in the corresponding cell. So it is natural to plot P as a third coordinate (we will call such plot a fractal representation of oligonucleotide composition (FRC)). Maxima in such a plot will correspond to highly represented oligonucleotides for the set, while minima correspond to the reverse.

FRC's of intron and coding human regions from Table 1 are shown in Figure 5, where each column corresponds to the number of specific oligonucleotides in these sets. We can see the essential differences of the composition in these sets, which was used for their classification.

In Figure 6, the FRC of 100 bp on the left from first CDS regions (5'-regions of human genes) and 100 bp on the right from the last CDS regions (3'-regions of human genes) were represented. The essential differences of oligonucleotide composition of them can be applied for recognition of these regions.

The FRC for 5'-flanking regions before the point of transcription initiation of human genes are shown in Figure 7 a,b,c. We can see that the -300 to -100 bp region, the -200 to -100 bp region and the -100 to -1 region have a similar composition, but the concentration of specific oligonucleotides increases toward the promoter (-100 to -1) region. Such a structure of the 5'-region may have a functional value, providing the specific organization of chromatin or the concentration of regulatory proteins in this regions. The trend in the number of some oligonucleotides can provide effective interaction of the region with universal protein complex, providing the transcription initiation (Solovyev 1990). Among the oligonucleotides typical for these 5'-regions include sites of transcription initiation: TATA-box, CCAAT-box, GC-box and poly-G(C) sequences which could be important for first exon splicing according to our model (Figure 3). This trend (Figure 7) can be used for the prediction of 5'-regions of eukaryotic genes.

Graphical System for Gene Structure Analysis

Now we will realize the simplest variant of a program for revealing coding regions based on oligonucleotide composition of various gene regions. For the discrimination between coding and intron regions Eqs. 2 and 3 are used. For the discrimination of the 5'- and 3'-regions, we used Eqs. 1 and 3 where the frequencies of octanucleotides of 5'- or 3'-regions, were used instead oligonucleotide frequencies of coding regions.

The window of fixed length is slid along the sequence of interest and the value F is calculated for each window position using fractal matrices of exons, introns, 5'- and 3'-regions. Figure 8 presents an example of the analysis of the

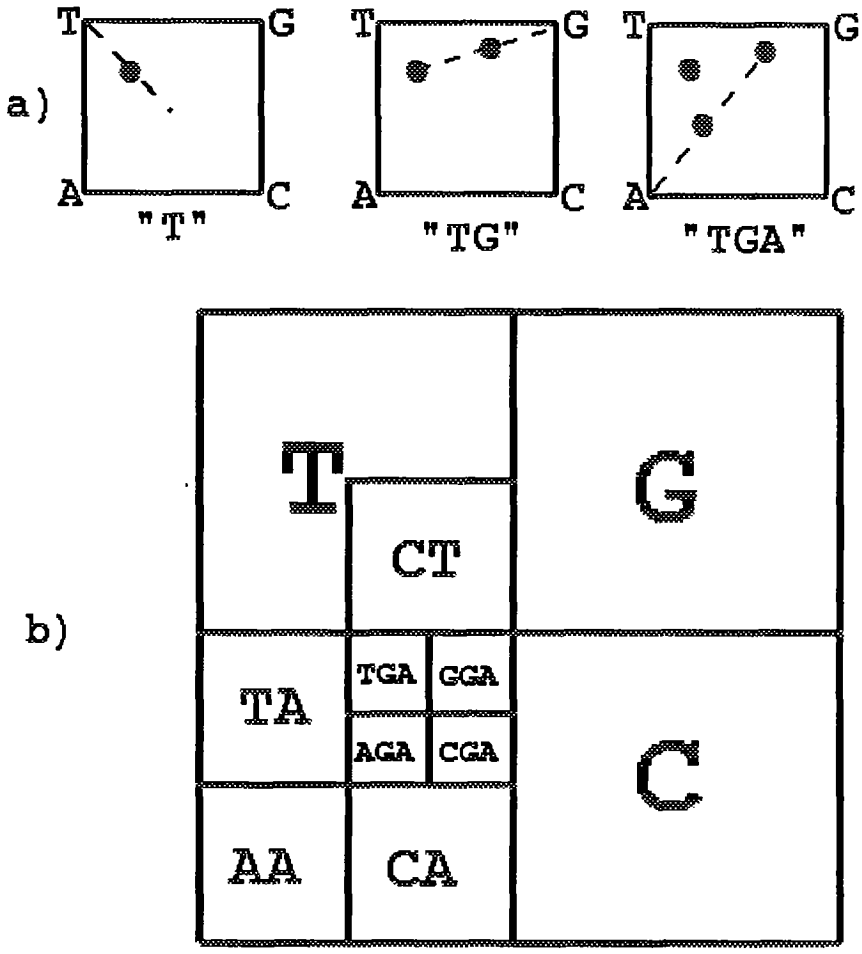

FIGURE 4. (a) The fractal graphic representation of the sequence "TGA". The lines are fictious and are used to show how the sequence "TGA" plotted; (b) Correspondence of oligonucleotides and subsquares. Subdividing the primitive square into quadrants of mononucleotides and subquadrants of dinucleotides and trinucleotides. The process can be continued to get subquadrants of trinucleotides, tetranucleotides and so on.

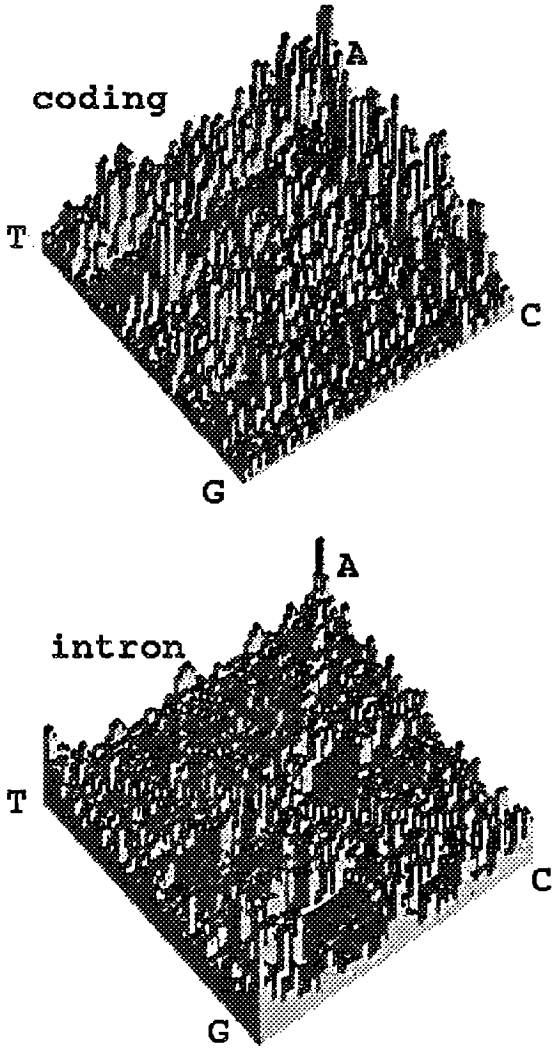

FIGURE 5. Graphical representation of the number of different oligonucleotides in-phase with coding frame 6 bp long in exon (top) and intron (bottom) regions.

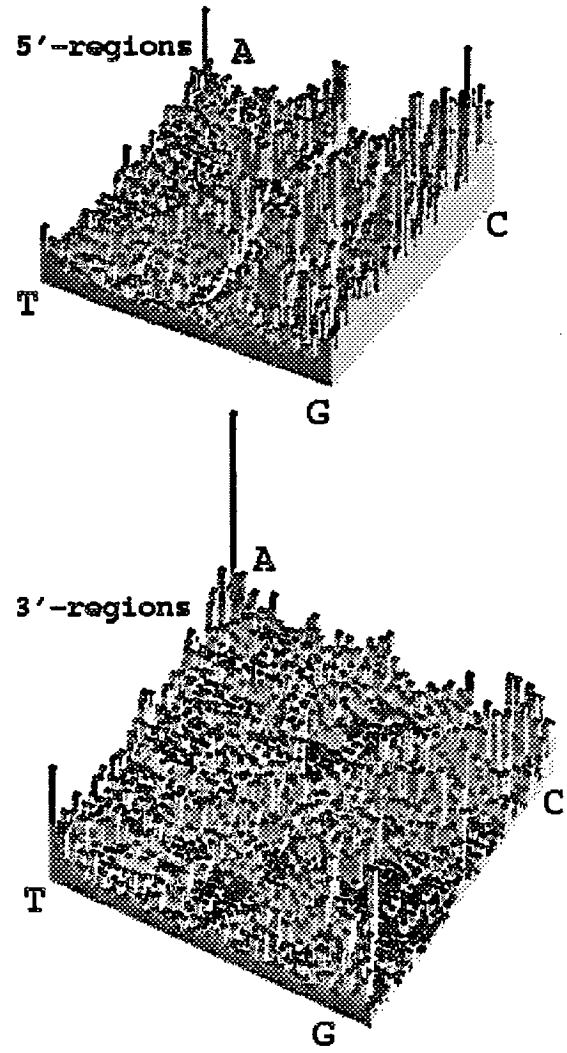

FIGURE 6. Graphical representation of the number of different oligonucleotides 6 bp long in 5'- (top) and 3'- (bottom) gene regions.

human alkaline phosphatase gene. The real coding regions correspond to peaks of discriminant function as well as 5'- and 3'-region. All splice sites are founded. Also, some overpredicted splice sites could be further removed in gene structure predictive system because probably only the authentic set of splice sites will have an open reading frame with a high score. The first variant of the system has been developed. It takes into account the weights based on oligonucleotide composition of all gene components: 5'-region, exons, introns, 3'-region and noncoding regions and then using a dynamic programming method searches for a combination of splice sites with maximal weight of these gene components. Testing this system on a few examples shows that if all authentic splice sites are among the predicted, then we can calculate precisely the positions of all exons and introns. This system and discriminant functions will be the basis of an automatic identification of gene structure algorithm. The graphical representation could be useful for gene structure analysis as some alternative splicing variants and gene structure may occur.

Acknowledgments

This work was supported by the W. M. Keck Center for Computational Biology and a grant to C.B.L. from the National Library of Medicine.

References

Brunak, S.; Engelbreht 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 Pablishing and GenBank. *Science* 252: 1273-1277.

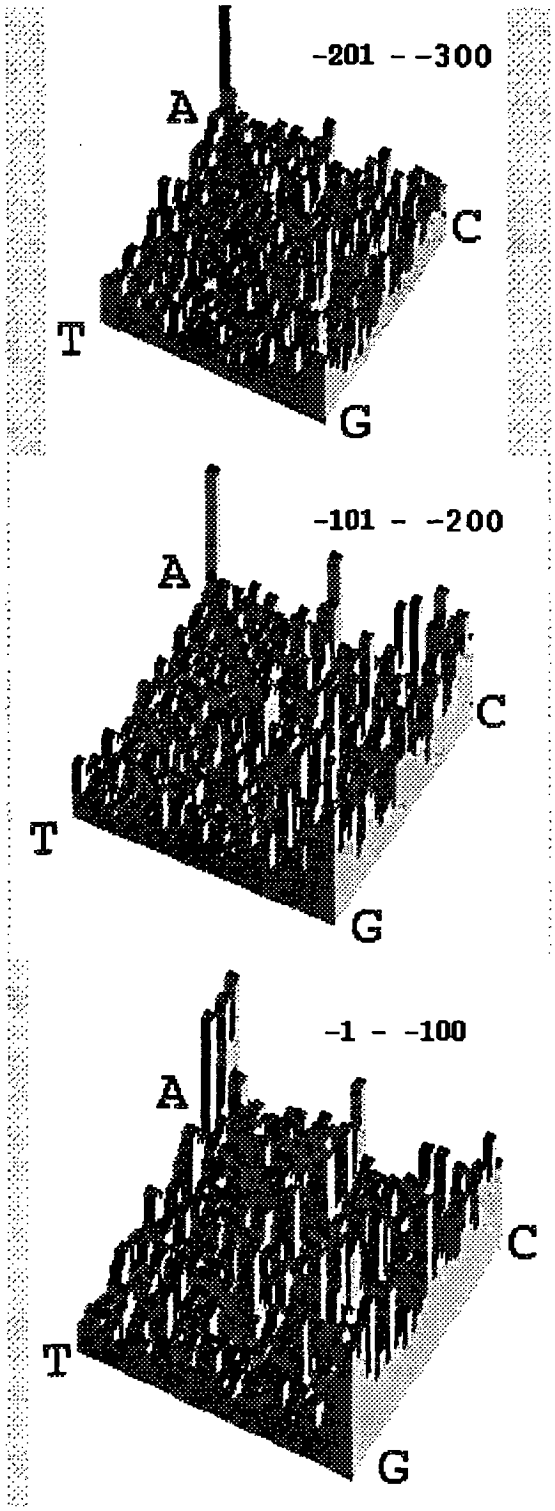

FIGURE 7. FRC for 5'-flanking regions before the point of transcription initiating of human genes were shown. We can see that -300 - -100 bp region (a), -200 - -100 bp region (b) and -100 - 1 region (c) have the similar composition, but the concentration of definite oligonucleotides increases toward to promoter (-100 - 1) region.

Claverie J.M.; Bougueleret L. 1986. Heuristic information analysis of sequences. *Nucl.Acids Res.* 14: 179-196.

Claverie J.M.; Sauvaget I.; Bougueleret L. 1990. K-typle sequences analysis: from exon-intron discrimination to T-cell epitope mapping. In *Methods of Enzymology* (ed. R.F. Doolittle) 183: 237-252.

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

Fickett J.W.; Tung C.S. 1992. Assesment of Protein Coding Measures. *Nucl. Acids Res.* 20: 6441-6450.

Farber R.; Lapedes A.; Sirotkin K. 1992. Determination of eukaryotic protein coding regions using neural networks and information theory. *J.Mol.Biol.*, 226: 471-479.

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

Jeffrey H.J. 1990. Chaos game representation of gene structure. *Nucl. Acids Res.* 18: 2163-2170.

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.

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

Mount S.M. 1982. A catalogue of splice junction sequences. *Nucl.Acids Res.* 10: 459-472.

Nakata k.; Kanehisa M.; DeLisi C. 1985. Prediction of splice junctions in mRNA sequences. *Nucl.Acids Res.* 13: 5327-5340.

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.

Robberson B.L; Cote G.J; Berget S.M. 1990. Exon defenition may facilitate splice selection in RNAs with multiple exons. *Mol.Cell. Biol.* 10: 84-94.

Solovyev V.V. 1991. Graphic methods of representation and analysis of the DNA and protein sequeences. The Institute of Cytology & Genetics USSR Academy of Sciences, Novosibirsk.

Solovyev V.V.; Korolev S.V.; V.G. Tumanyan; Lim H.A. 1991. A new approach to classification of DNA regions based on fractal representation of functionally similar sequences. *Proceedings of the USSR Academy of Science* 319: 1496-1500.

Solovyev V.V.; Korolev S.V.; Lim H.A. (1993) A new approach for the classification of functional regions of DNA sequences based on fractal representation. *Int. J. Genome Res.* 1: 109-127.

Solovyev V.V.; Lawrence C. 1993. Prediction of human mRNA donor and acceptor splice sites based on oligonucleotide composition (submitted).

Stormo G.D. 1987. Identifying coding sequences. In *Nucleic acid and protein sequence analysis* (Bishop M.J. and Rawlings C.J. eds) IRL Press, Oxford, 231-258.

Staden R. Finding protein coding regions in genomic sequences. 1990. 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 sen-

sor - neural net approache. *Proc.Natl.Acad.Sci.USA* 88: 11261-11265.

Volinia S.; Gambari R.; Bernardi F.; Barrai I. 1989. The frequency of oligonucleotides in mammalian genic regions.*CABIOS* 5: 33-40.

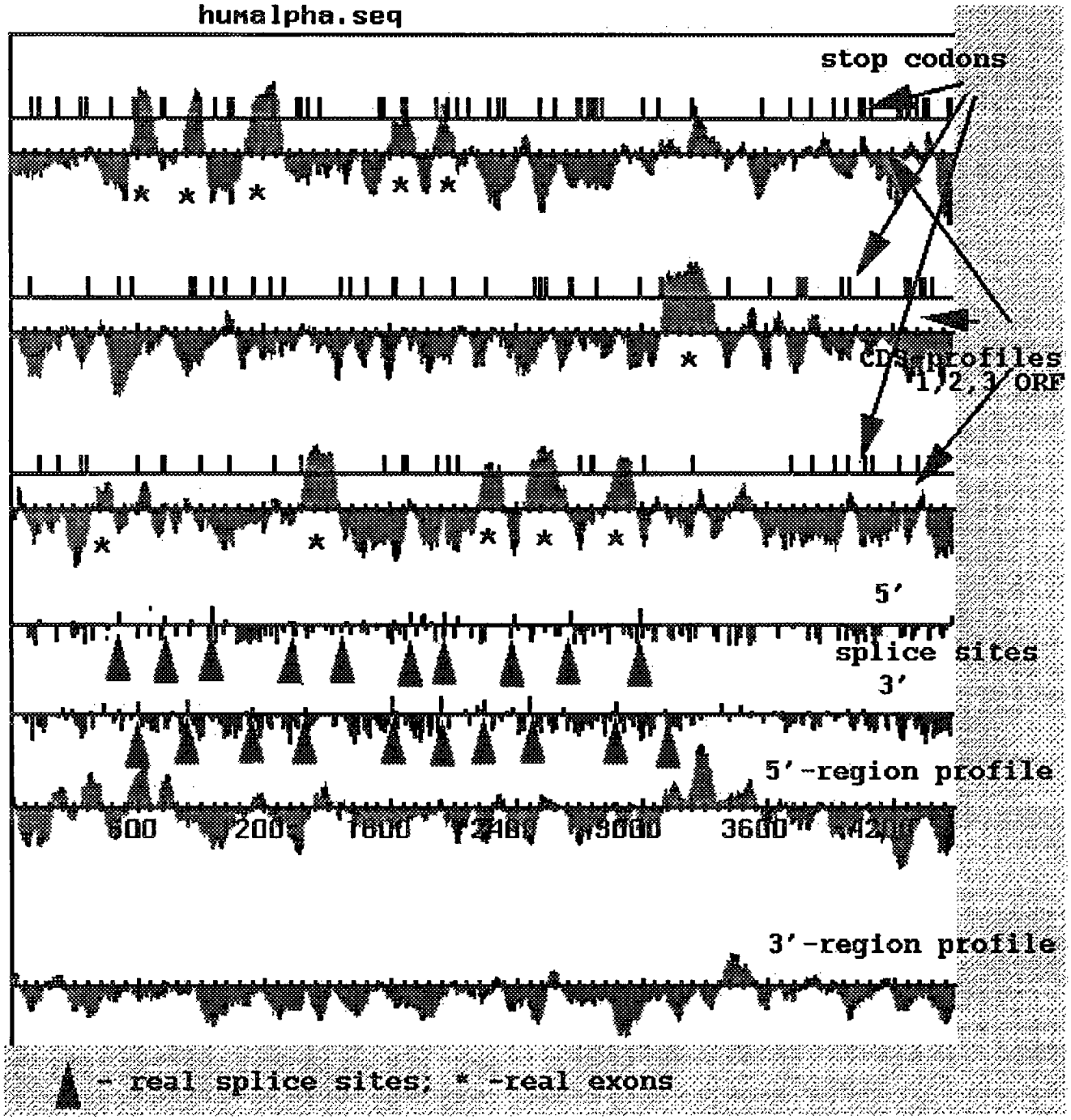

FIGURE 8. Graphical representation of the profiles of coding potential weights of 1,2,3 -d ORFs, splicing sites and potential 5'- and 3'- gene region weights for human alkaline phosphatase gene with 11 exons was not included in train set.