CABIOS

Vol.9 no.6. 1993
Pages 617 – 627

SITEVIDEO: a computer system for functional site analysis and recognition. Investigation of the human splice sites

A.E.Kel, M.P.Ponomarenko, E.A.Likhachev, Yu.L.Orlov, I.V.Ischenko, L.Milanesi[1] and N.A.Kolchanov[2]

Abstract

We developed the computer system SITEVIDEO for analysis and recognition of the functional sites in DNA and RNA molecules. It reveals contextual features essential for site function and thus enable the user to design efficient methods for recognition of the functional sites. We mainly considered only quantitative characteristics reflecting the uneven distribution of oligonucleotides in the sequences of functional sites of interest. The approach suggested makes use of available information about the hierarchical organization of the functional sites, and ensures highly precise prediction of the sites. The present analysis is concerned with the human donor and acceptor splice sites. A method for recognizing these sites in the sequences with an accuracy of ~90% was developed.

Introduction

Recognition of the gene coding regions is one of the significant problems in computer analysis of the human genome. The eukaryotic genes have a mosaic structure and, hence, they are much more complex than the prokaryotic. They consist of exons and introns, their boundaries being determined by the location of the splice sites.

Thus, to ensure successful search for the gene coding regions in the human genome, development of highly precise methods for recognition of the splice sites is needed.

Many functional signals in RNA sequences involved in the splicing process are known, and are presently under study. These include 3′- and 5′-cutting RNA sites of the splicing (donor and acceptor sites) (Padgett *et al.*, 1986), the polypyrimidine tract near the 3′-splice site (Shapiro and Senapathy, 1987), the branch site of the 3′-region of the intron (Green, 1986). However, many other contextual factors affecting the splicing process remain unclear. There is circumstantial experimental evidence that contextual factors are present in the intron and exon regions of the genes adjacent to the splice sites and have an effect on the selection of the correct splice sites (Reed and Maniatis, 1986; Smith *et al.*, 1986; Gelfand, 1992). The presence of these factors is provided by the emergence of cryptic splice sites in the neighboring regions in the case of mutation or deletion in the authentic site (Padgett *et al.*, 1986). The normal process of splicing and correct intron excision seems to be provided by a certain contextual correlation between the donor and acceptor splice sites and the intron and exon regions surrounding the splice sites.

Among the established approaches to functional site analysis and recognition, the following deserve particular mention: the consensus; the matrix of nucleotide frequencies (Shapiro and Senapathy, 1987); the perceptron function (Iida, 1988); discrimination energy (Berg and von Hippel, 1988); neural networks (Holley and Karplus, 1991); methods of functional site classification (Kudo *et al.*, 1992; Rogozin *et al.*, 1993). All these approaches reveal various contextual features important for the activity (as well as for the recognition) of the functional sites. In the majority of cases, these methods are applied in analysis of rather narrow regions (not exceeding 10–20 bp) spanning functional sites. However, a wide range of contextual characteristics of long flanking sequences around functional sites is also important for functional site activity, such as frequencies of oligonucleotide usage, the physical, stereochemical or statistical properties of nucleotides, and their location in the primary structures, among others.

This paper describes an approach that allows one to build new recognition methods by means of such quantitative characteristics. This approach is based on the theory of utility (Fishburn, 1970). For each quantitative characteristic the value of utility (U) defined by certain statistical expert rules is calculated. The greater the U is, the more efficient the characteristic is for recognition. Relying on this criterion, only characteristics with high recognition capacity of the functional site of interest are chosen.

Very short consensus sequences with just two or three conserved positions are characteristic of the donor and acceptor splice sites (Gelfand, 1989). The location of these conserved sequences is quite invariable, and they have all been proven to be essential for normal site function. Contextual factors such as highly conserved consensus sequences will be referred to as *obligatory* features.

Additional contextual factors not involved in the consensus ensure the normal function of the sites. These may be regions complementary to snRNA, or different RNA hairpin structures, which, probably, may be responsible for the close proximity of the ends of the adjacent exons during splicing (Gelfand, 1989). Such additional contextual factors will be referred to as *facultative* features.

Institute of Cytology and Genetics of the Russian Academy of Sciences, pr. Lavrentyeva 10, Novosibirsk, 630090, Russia and [1]Istituto di Tecnologie Biomediche Avanzate, Consiglio Nazionale Delle Ricerche, Milano, Italy

[2]To whom reprint requests should be sent

The consensus of the acceptor and donor splice sites have been previously examined (Padgett *et al.*, 1986). No attempts have been made to analyze thoroughly the facultative contextual factors of the splice sites. This may be one of the reasons why no efficient methods have so far been elaborated for recognition of the splice sites by means of the above approach relying on calculated recognition utility.

It was assumed that a normal consecutive run of splicing may be ensured by different contextual factors of the introns, exons, those of the intron–exon boundaries, as well as by their interactions. These interactions provide the specific hierarchical structural–functional organization of the splice sites. We developed the computer SITEVIDEO system for analysis of the hierarchical structure of the contextual features of the functional sites. This system uses the recursive approach (Kel *et al.*, 1989) to describe the hierarchical structure. A set of contextual features in the nucleotide sequences of the functional sites were examined, and their interactions and effects on the splice site function were revealed. Based on this analysis, programs for recognition of the donor and acceptor splice sites were developed.

System and methods

The sequences used were drawn from the EMBL Data Library (Tables 1 and 4). The human splicing sites were taken from the 5′- and 3′-boundaries of the introns of the genes in compliance with the 'FEATURE TABLE' of sequences of the bank. Two samples were set up to build the consensuses of the narrow regions near the RNA cutting points: 844 donor sites of 22 bp (12 bp upstream and 10 bp downstream of the cutting point) and the 826 acceptor sites of the 23 bp (19 bp upstream and 4 bp downstream of the point).

Furthermore, to analyze the extensive flanking regions of the splicing sites, we set up mRNA fragment samples evenly representing the various evolutionary gene families and also the possible variants of the location of the splicing sites in the exon–intron structure of the genes. The former sample contained 134 human RNA fragments with the donor sites; the latter sample consisted of 128 fragments with the acceptor sites. Only the splicing sites with adjacent exons and introns not shorter than 50 bp were included in the sample. The length of the sequences containing the splicing sites was 100 bp (50 bp upstream and 50 bp downstream of the exon–intron or intron–exon boundary). Training data for building programs recognizing the splicing sites containing 25 donor and 25 acceptor sites were derived. The remaining 103 donor and 97 acceptor sites from the two samples were used as control data to test the recognition programs. Furthermore, samples containing nucleotide sequences 100 bp long from the central regions of exons and introns of certain human genes were set up.

The major tool for analysis of the samples and development of the recognition programs was the specifically designed SITEVIDEO system. The computer system was implemented in Turbo C (v. 2.0) on an IBM PC/AT computer running under the DOS operating system and in a VMS operating system environment on the VAX computer.

Table I. List of the genes with acceptor and donor splicing sites that were included in training samples

Acceptor splice site		Donor splice site	
EMBL identifier	Location	EMBL identifier	Location
HSA1ATP	9362–9461	HSA1ATP	7913–8008
HSACHG7	438–537	HSACHG7	249–344
HSALBGC	5990–6089	HSALBGC	11 033–11 128
HSALDB2	732–831	HSALDB2	323–418
HSALPHA	1348–1447	HSALPHA	1273–1368
HSATPSY1	5217–5316	HSATPSY1	4647–4742
HSC1A1	6615–6714	HSC1A1	2327–2422
HSC4AB	427–526	HSC4AB	335–430
HSCFVII	8256–8355	HSCFVII	6543–6638
HSERYA	1858–1957	HSERYA	1751–1846
HSFBRGG	4594–4693	HSFBRGG	4293–4388
HSFESFPS	5750–5849	HSFESFPS	5549–5644
HSGHVA	720–819	HSGHVA	469–564
HSGRP78	2124–2223	HSGRP78	1491–1586
HSGSTPIG	3462–3561	HSGSTPIG	2602–2697
HSHLIC	1691–1790	HSHLIC	1106–1201
HSHSC70	2046–2145	HSHSC70	1724–1819
HSIG05	563–662	HSIG05	886–981
HSIGJ2	774–873	HSIGJ2	105–200
HSIL1B	5823–5922	HSIL1B	5278–5373
HSINT1G	2915–3014	HSINT1G	2455–2550
HSKER65C	405–504	HSKER65C	47–142
HSLCATG	1909–2008	HSLCATG	1832–1927
HSMETIF	1454–1553	HSMETIF	1124–1219
HSPLPSPC	1809–1908	HSPLPSPC	1466–1561

Algorithm

This section presents a description of the computer system SITEVIDEO designed for analysis and recognition of the functional sites in the DNA and RNA molecules on the basis of their primary structures. A general scheme of the system is given in Figure 1. Let us consider the system in more detail.

Contrasting samples

Contrasting samples of functional sites are used as input of the system. Each sample consists of two parts: YES, a set of functional sites under study having a common functional property; NO, a set of nucleotide sequences not having this property.

Generators of simple contextual characteristics

The block is used to reveal numerous characteristics reflecting various properties of the nucleotide context in the functional sites under consideration. The block consists of a set of

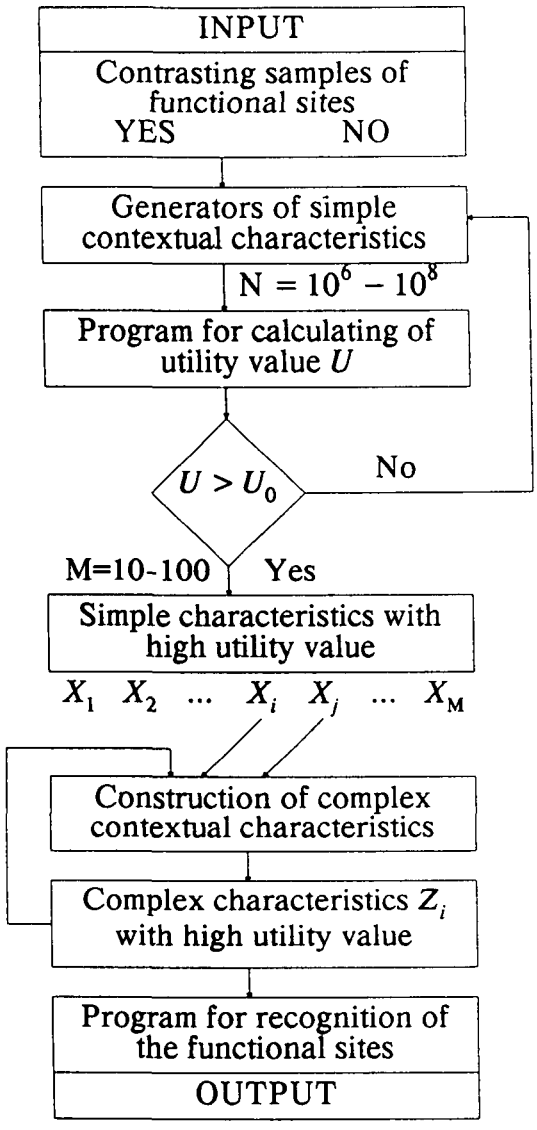

Fig. 1. A general scheme of the system SITEVIDEO for analysis and recognition of the functional sites. N is the number of generated simple contextual characteristics; M is the number of chosen simple characteristics.

programs, generators calculating quantitative characteristics such as: (i) statistical properties—characteristics related with the frequencies of mono- or oligonucleotides and their location in the functional sites; (ii) physical properties—charge, stacking energy, mass, volume, polarity, hydropathy; and (iii) chemical properties—presence of certain atom groups in the nucleotides and certain other distinctive features of their chemical structure.

Let us consider the functional site S of length L, $S = \{s_1, s_2, \ldots, s_L\}$. Let A_k be definite oligonucleotides of length m in a 15 letter code. This code allows one to classify all the nucleotides according to their common structural features. For example, the letter M represents two nucleotides {A, C} possessing a common property such as the presence of an amide group in their chemical structures.

The occurrence profile of oligonucleotides of the given A type can be built along a sequence:

$$P(S, A_k) = (P_1, P_2, \ldots, P_{1-m+1})$$

where

$$P_i = \begin{cases} 1, \text{ if } s_i, s_{i+1}, \ldots, s_{i+m-1} = A_k \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

The majority of statistical characteristics calculated by the generators in the described block of the SITEVIDEO system is derived from analysis of the occurrence profile of the oligonucleotides.

The profile of the physicochemical properties of the oligonucleotide composition of the functional sites is built as follows. Let $H = \{H(A_k)\}_{k=1,15^m}$ be the set of values of a definite physicochemical property for all the oligonucleotides A_k of the length m. For example, it can be the stacking energy for all the dinucleotides in the DNA structure or some other property. Then the property profile for the functional site S is determined as follows:

$$P(S,H) = [H(s_1 \ldots s_m), H(s_2 \ldots s_{m+1}), \ldots, H(s_{L-m+1} \ldots s_L)] \quad (2)$$

The profiles $P(S, H)$ and $P(S, A)$ contain abundant information on different properties of the functional site under consideration. The SITEVIDEO system has a set of procedures to analyze these profiles and also to reveal important information concerning the structure of the functional sites, i.e. smoothing and averaging procedures.

The smoothing procedure is as follows. Let us consider the window of the size j. The profile P_i can be transformed into the smoothed profile G_i using the following expression:

$$G_i = \sum_{k=0}^{j-1} \alpha_k P_{i+k} \quad (3)$$

Here α_k is the set of weight coefficients describing the relative contribution of positions within the window. By means of this procedure we simulate the local interactions of the oligonucleotides at distance j from each other. Another variant of the smoothing procedure describes the importance of the extreme features within the sliding window of the size j:

$$G_i = \min (P_i, \ldots, P_{i+j-1})$$
$$G_i = \max (P_i, \ldots, P_{i+j-1}). \quad (4)$$

The next step of analysis is determination of the mean value of a given contextual feature X for the functional site as a whole:

$$X = \sum_{i=1}^{L} G_i \times F_i \quad (5)$$

Here F_i is the function of position significance. When analyzing the functional sites, we consider their different positions as not equivalent in terms of their functional importance. To describe the different importance of position to site function, the set of functions F_i is used, differing in the shape and location of the maximum (minimum) of the function

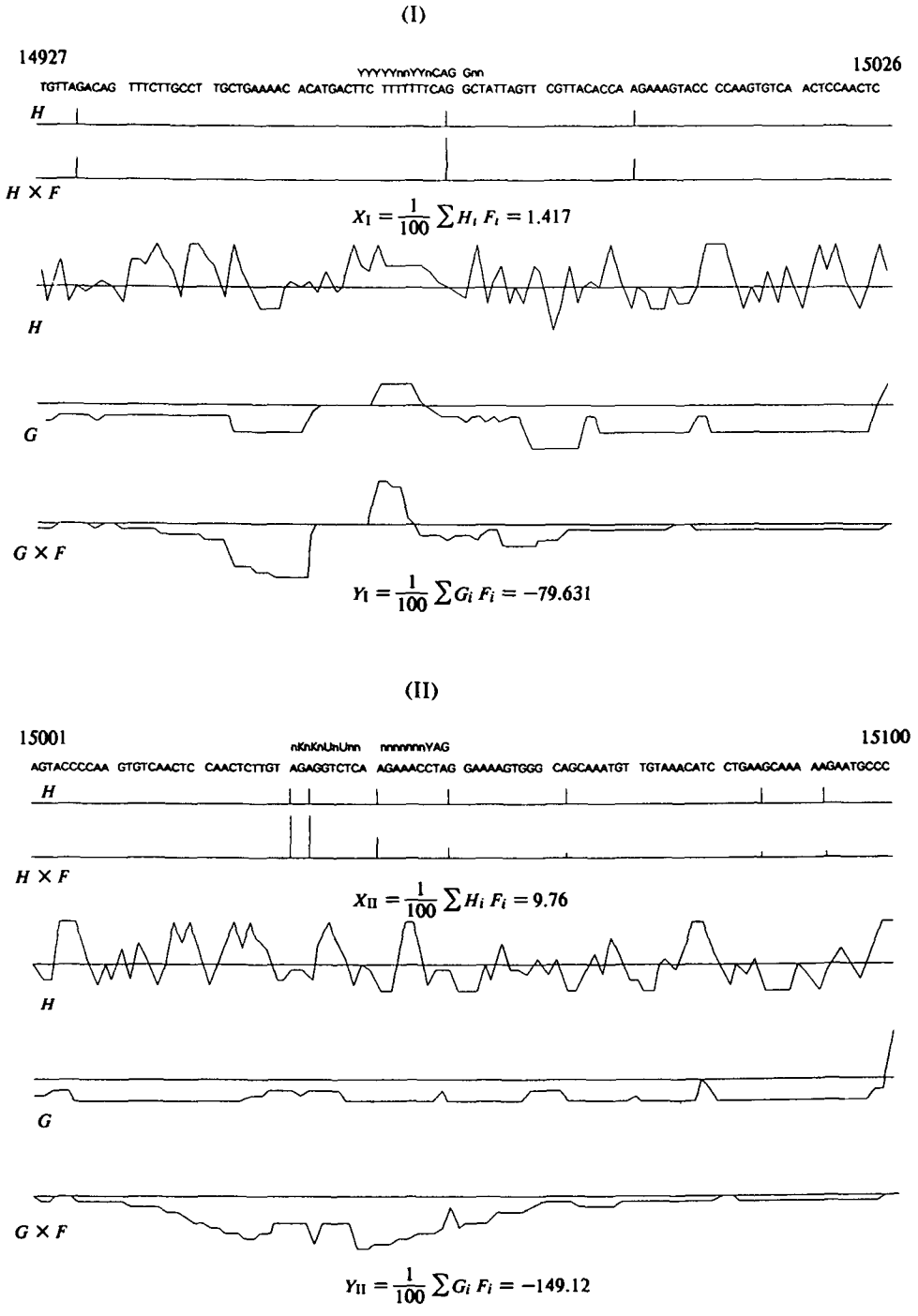

Fig. 2. Two sequences exemplifying the computation of contextual characteristics. The computation steps of the two characteristics X and Y for the two sequences are shown. Computation of the characteristic X is done in three steps. The occurrence profile of the trinucleotide AGV along the sequence is computed first (H). The multiplication procedure by the preference function (see Figure 4) follows $H \times F$). The profile is then averaged. In computation of the characteristic Y, the building of the profile is based on the dinucleotide scale given in Figure 5 (H). This is followed by smoothing of the profile (G): at window size $j = 7$, $G_i = \min (P_i, \ldots, P_{i+6})$ for each position i. Averaging of the profile with the preference function (see Figure 5) is performed ($G \times F$). The values of the characteristics X and Y for both sequences are given under the profiles. One of the possible consensuses for the acceptor site with which the given sequence matches is indicated above each sequence.

within the site. Obviously, the more the profile G corresponds to the function of the position significance F, the greater values the characteristic X assumes.

Let us examine two regions of the human serum albumin gene to see how the system works. One region (I) is 100 bp long, and it covers the acceptor splice site from the 10th intron of the gene. The second region (II) is of the same length, and it is from the 11th exon not containing the acceptor site (see Figure 2). To highlight the capacity of the system to reject 'false' splicing signals, a sequence region (II) is chosen so that the AG dinucleotide falls in its central part. Figure 2 demonstrates the computation procedure of the two characteristics X and Y: X reflects the distribution of oligonucleotides AGV in the two sequences, and Y is a reflection of the preferential distribution of oligonucleotides YY in conformance with the dinucleotide scale given in Figure 5. Clearly, $X_I < X_{II}$, whereas $Y_I > Y_{II}$ and, therefore, the contexts in the two examined sequences differ significantly.

Thus, for each functional site the generators calculate an enormous number of different contextual characteristics. This is achieved by varying over a large range the type and length of oligonucleotides, their physicochemical properties, and the smoothing and averaging procedures. This diversity reflects the various aspects of the distribution of a large class of oligonucleotides and their physicochemial properties in the examined sequences of the functional sites.

The given block also contains generators for analysis of the blockwise structure of the functional sites by dividing the sequences into numbers of non-overlapping intervals and also for analysis of the mutual location of pairs of certain oligonucleotides within the functional sites.

Program to estimate the predicting capacity of the contextual characteristics

Of all the simple generated characteristics, only a small portion is specific to the given functional sites, and they can be used for recognition. To select such characteristics, a special block is designed—a program to estimate the predicting capacity of the contextual characteristics.

The choice of the characteristics with high recognition capacity is based on the utility theory for decision making (Fishburn, 1970). It relies on analysis of contrasting YES and NO samples. For the quantitative characteristic X its value can be calculated for all the sequences of classes YES and NO. Then we obtain a utility value which is an integral quantitative estimate of the recognizing capacity of the given contextual characteristic X via analysis of the distributions of the values X for the sequences of sites (class YES) and non-sites (class NO).

The utility value is calculated by applying a set of expert rules based on three groups of statistical criteria (tests of statistical significance of the differences between the mean value of a given characteristic of the two contrasting samples; tests of closeness of the characteristic distribution to the normal distribution; tests of stability of recognition results).

Utility value lies within the interval $[-1, +1]$. If $0 < U < 0.5$, the given X characteristic may be used for discrimination of classes YES and NO, however, only in conjunction with other characteristics. If $0.5 < U < 1$, the given X characteristic is useful itself for discrimination of the classes YES and NO (for a more detailed treatment of utility, see the Appendix).

Construction of complex contextual characteristics

Based on the simple contextual features with high recognizing capacity that are revealed, SITEVIDEO allows construction programs for functional site recognition. The recognition programs are constructed step by step by pairwise combination of simple contextual characteristics into a more complex characteristic. At each step of analysis, the system uses a set of standard two-dimensional discriminating methods (linear and non-linear) such as perceptron (Duda and Hart, 1973), Fisher's linear discriminant (Bolch and Huang, 1974), majority votes (Mlinsky, 1975), typological discriminant (Duda and Hart, 1973), among others. The complex characteristic Z can be calculated on the basis of the two characteristics X and Y by the linear or non-linear functionals in accordance with the chosen discriminating methods. For example, in the case of the linear functional:

$$Z(X,Y) = \cos(b_0)(X - b_1) + \sin(b_0)(Y - b_2) \qquad (6)$$

Here b_0, b_1, b_2 are the free coefficients that are calculated by the chosen discriminating method. The coefficients yield the position of the right line $Z(X,Y) = 0$ on the plane given by the characteristics X and Y and separating the sample YES from NO (b_0 is the angle formed by the line; b_1 and b_2 are the values of deviation from the center of the coordinates).

Figure 3 schematically represents how the block works. The deviation of the complex characteristic Z from the two characteristics X and Y considered in the previous example is now done for the acceptor site (I) and exon region (II). The plane with axes X and Y is given (Figure 3). The sequences of the acceptor sites (sample YES) and those of the random sequences (sample NO) are represented by points on the plane. The two points corresponding to the sequences (I) and (II) are also indicated. The optimum position of the line separating sample YES from NO was defined by the Fisher's linear discriminant method (Bolch and Huang, 1974). The value of Z was calculated using the formula (6) with $Z_I = 40.69 > 0$ and $Z_{II} = -13.45 < 0$. The characteristic Z thus produced is sufficient for recognizing the acceptor sites. If $Z_S > 0$, there is good reason for referring the sequence S to the acceptor sites.

In addition to the automatic discriminating method, manual methods are provided to allow correction of the automatic perfomance. This ensures more accurate division of the samples.

The recognition capacity of the calculated complex characteristic is estimated by the utility value. In the case when the complex characteristic $Z(X,Y)$ has a utility higher than that of the characteristics X or Y, it is used for recognition program construction.

Synthesis of the program for functional site recognition

At the final step of the analysis, the text of the recognition program is synthesized in automatic mode by SITEVIDEO on the basis of the most complex characteristic Z^* with the highest utility level. For the chosen sequence, if $Z^* > 0$, the program will recognize it as a functional site, otherwise as a non-functional site.

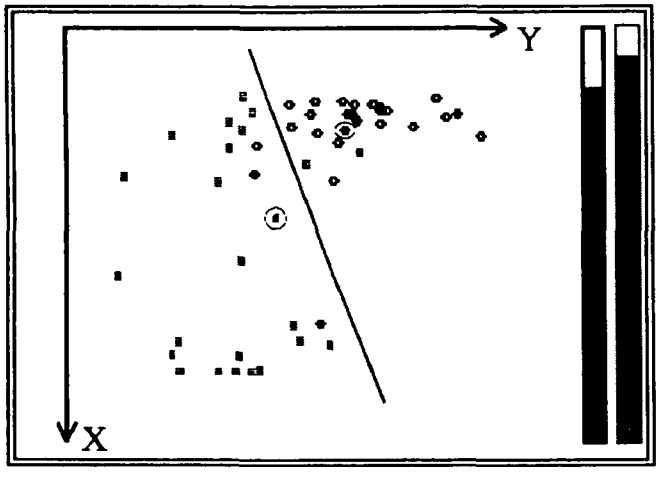

Fig. 3. An illustrative example of computation of the complex characteristic Z derived from two simple characteristics X and Y (Figure 2). A plane with axes X and Y is shown. The sequences from the contrasting samples are indicated by points (circles are the acceptor splice sites, hatched squares are the random sequences). The two points in the circles correspond to the acceptor site (I) and exon region (II) (Figure 2). The optimal position of the straight line separating the two samples is defined by Fisher's linear discriminant method (Bolch and Huang, 1974). According to the position of the line $Z = -1.98X + 0.54Y + 86.6$. Based on the values of the characteristics X and Y for sequences I and II, $Z = 40.7$ and $Z = -13.5$.

Analysis of splice sites

The first step of the splice site analysis is an examination of contrasting samples which takes into account the hierarchy of the main groups of the contextual features important for the function of the donor and acceptor splice sites.

Let us first consider the acceptor splicing sites. Here we consider the following groups of functional elements.

The first group (PR) represents contextual features responsible for the primary recognition of the acceptor splice sites by the different protein and RNP molecules involved in the splicing process. This group includes contextual features that distinguish the acceptor site from RNA regions where these sites should not be detected, i.e. the 5'- and 3'-ends of the primary RNA transcript, the central parts of the exons and the introns, and the donor splice sites. To reveal such contextual features, the samples CPE, CPI and DAS are examined (Table II). Here, class YES in all three samples contains the acceptor splice sites; class NO in the CPE contains the central parts of the exons of several human genes; class NO in the CPI contains the central fragments of introns; and class NO contains the donor sites in the DAS (Table II).

After the primary recognition of an acceptor splice site, there follows a definite local interaction of the spliceosome with the RNA region. The group of contextual features responsible for the local interaction consists of (i) the obligatory features of the acceptor site needed for normal splicing such as AG dinucleotide on the intron–exon boundary; and (ii) the

Table II. The number of facultative characteristics revealed in analysis of training samples of acceptor splice sites

No.	Block	Class	Samples	Location	No. of revealed characteristics	Mean utility	Utility of complex characteristic A
1	RS	YES	acceptor splice sites	−50/+50[a]	21	0.67	
		NO	random sequences	100[b]			
2	3'IF	YES	3'-ends of introns	−50/0	23	0.71	
		NO	random sequences	50			
3	5'EF	YES	5'-ends of exons	0/+50	34	0.67	
		NO	random sequences	50			
4	BF	YES	3'-ends of introns	−50/0	35	0.76	
		NO	5'-ends of introns	0/+50			0.85
5	CPE	YES	acceptor splice sites	−50/+50	29	0.61	
		NO	central parts of exons	100			
6	CPI	YES	acceptor splice sites	−50/+50	34	0.63	
		NO	central parts of exons	100			
7	DAS	YES	acceptor splice sites	−50/+50	23	0.75	
		NO	donor splice sites	−50/+50			

[a]Indicated is the location of the sequence with respect to the RNA cutting point at splicing. Left of '/' is the fragment length upstream of the cutting point; right, downstream of the cutting point.
[b]Sequence length.

facultative features of the acceptor site which modulate the efficiency of splicing.

The acceptor splice sites involve the 5′-ends of the exon coding for certain protein fragments. Thus it may be assumed that the gene primary structures and frequency of codon usage specific to different gene groups may be responsible for some facultative contextual features at the 5′-end of the exons which can affect the efficiency of splicing. In addition, it may be suggested that some contextual features located at the 3′-end of the introns may also modulate splicing. To analyze these features, the contrasting samples 5′EF and 3′IF are examined with classes YES containing 50 nucleotide fragments of the 5′-ends of the exons and 3′-ends of the introns of the same length and with classes NO containing random sequences (Table II).

To analyze the contextual features in terms of the entire acceptor site (i.e. including both the 3′-end of the intron and the 5′-end of the exon), a contrasting sample RS is examined (Table II). Here class YES contains acceptor site sequences, and class NO is composed of the random sequences generated by nucleotide frequencies of the acceptor sites.

The sample BF (boundary features) is also examined (Table II). BF is composed of the fragments of the introns (class YES) and the exons (class NO) along the exon−intron boundary.

At the next step of analysis, the recognizing capacities for $\sim 10^8$ simple contextual characteristics are automatically estimated for each contrasting sample. Then, the most significant characteristics are chosen out of the whole array, i.e. those with maximal utility for the splice site recognition.

In analysis of the sample RS, we identify the simple characteristic reflecting the distribution of oligonucleotides AGV near the cutting point (see Figure 4). We thus revealed that low concentration of the AGV oligonucleotides upstream of the cutting point is a distinctive feature of the acceptor splice site. This appears noteworthy because all the acceptor sites have the dinucleotide AG adjacent to the RNA cutting point. Thus, it is our view that lower concentration of AG-containing oligonucleotides is needed to avoid the presence of 'false' AG-signals in the acceptor sites. This characteristic has a high utility value (0.70) and, hence, it is important for acceptor site recognition. Figure 4 shows the distribution of the characteristic for the acceptor splice site and also presents the curve of positional significance for this characteristic.

Another example of a contextual characteristic with high utility is given in Figure 5. Figure 5(a) shows the dinucleotide scale reflecting the relative importance of different dinucleotides. This scale was computed by the generator program for the RS contrasting sample. Figure 5(b) shows the function of position significance that indicates the region where these dinucleotides are most important in terms of site function. The great importance of the YY dinucleotides in the revealed dinucleotide scale probably indicates the presence of the known poly(Y) tracts before the cutting point of the acceptor site. The characteristic has a high utility $U = 0.84$.

The number of simple characteristics detected in analysis of the given contrasting samples is given in Table II. The mean utility values are also given. The highest mean utility (0.76) is obtained for the contrasting sample BF, distinguishing the 3′ intron region of the acceptor site from the 5′ exon region of this site. This demonstrates that changes have taken place

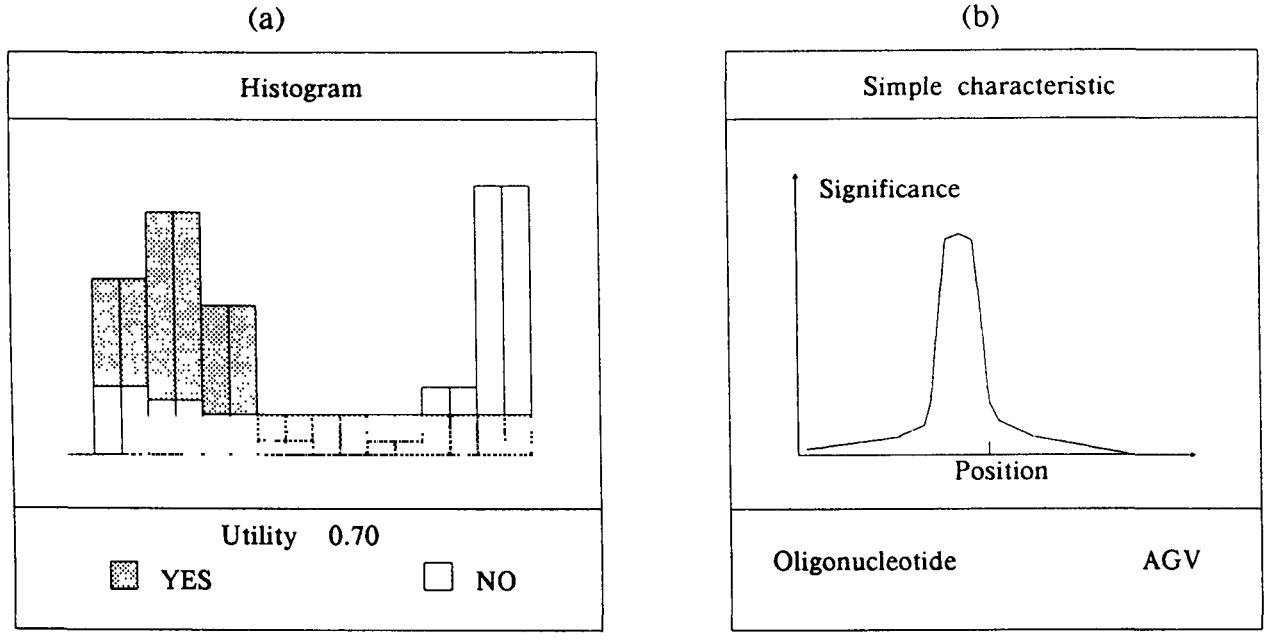

Fig. 4. The characteristics reflect the distribution of the oligonucleotide AGV along the splice site: (**a**) the histogram for the acceptor sites (hatched column) and random sequence (open column); (**b**) function of position significance.

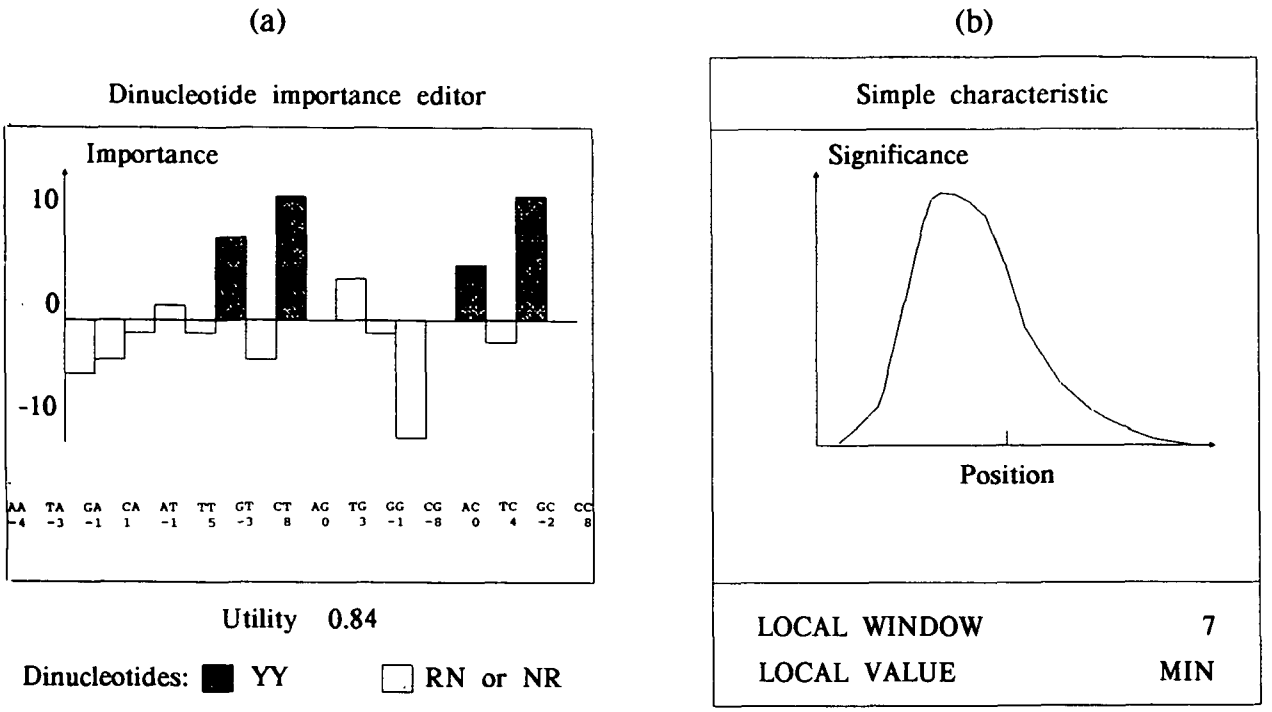

Fig. 5. The characteristic reflects the preferable location of certain dinucleotides within the narrow regions of the acceptor site. The dinucleotide YY (hatched columns) has maximum significance.

in the context of the nucleotide sequences along the intron–exon boundaries of the genes. A high mean utility of 0.75 is obtained for the DAS sample (the acceptor versus the donor sites). This is good evidence indicating that in the course of evolution the acceptor sites might have given rise to contextual features different, in principle, from those of the donor sites.

Programs for recognition of the acceptor and donor sites

Using the contextual characteristics revealed, a program for the recognition of the acceptor splice sites is designed via recursive construction of the complex characteristic A incorporating all the revealed contextual features of the acceptor splice sites. We proceed from building a complex characteristic of each of the contrasting samples; these are then combined into the most complex characteristic A (see Table II).

Figure 6(a) gives the distribution of the A value for the training sample of the acceptor sites (black columns) and random sequences (open columns). These two distributions virtually do not overlap.

The complex characteristic D, which combines the revealed contextual features of the donor splice sites, is constructed in the same way. Figure 6(b) shows the distribution of D for the training sample of the donor splice sites and random sequences. The final utilities are very high: 0.85 for A, and 0.84 for D.

With the use of these characteristics A and D, we construct the subroutine FACULT for recognition of the acceptor and donor splice sites relying on the revealed facultative contextual features. This subroutine works as follows. For a nucleotide sequence S of 100 bp, if $A(S) > 0$, the sequence S is classified as a potential acceptor site, if $D(S) > 0$, the sequence S is classified as a potential donor site, otherwise it is classified as a random sequence. This subroutine is the first part of the final program SPLICE.

The second part of the program is the subroutine OBLIGAT for recognition of the splice sites on the basis of obligatory features such as consensuses. To construct the donor–acceptor site consensuses, we use the classification procedure previously described (Rogozin *et al.*, 1993). A sample of 844 donor sites yielded a set of 20 different consensuses that covered all the sequences of the sample; 27 consensuses were produced from the sample of 826 acceptor sites. Examples of the most frequent revealed consensuses are given in Table III.

The program SPLICE combines the above two subroutines. For the sequence S of length L (from 10 bp up to 10 000 bp) the program recognizes the location of the potential donor and acceptor sites. At the first step, the subroutine OBLIGAT works so as to find the potential splice sites by matching with any consensus revealed. The regions of the sequence S with revealed matching with a consensus are further tested with the subroutine FACULT. The region taken to be tested is 100 bp long and surrounds the potential splicing site. When FACULT confirms the presence of an acceptor or donor site, SPLICE outputs the position of the recognized site in sequence S.

Two regions of the human serum albumin gene will serve as examples of what each block of the program does with a sequence. The program OBLIGAT has referred the two regions

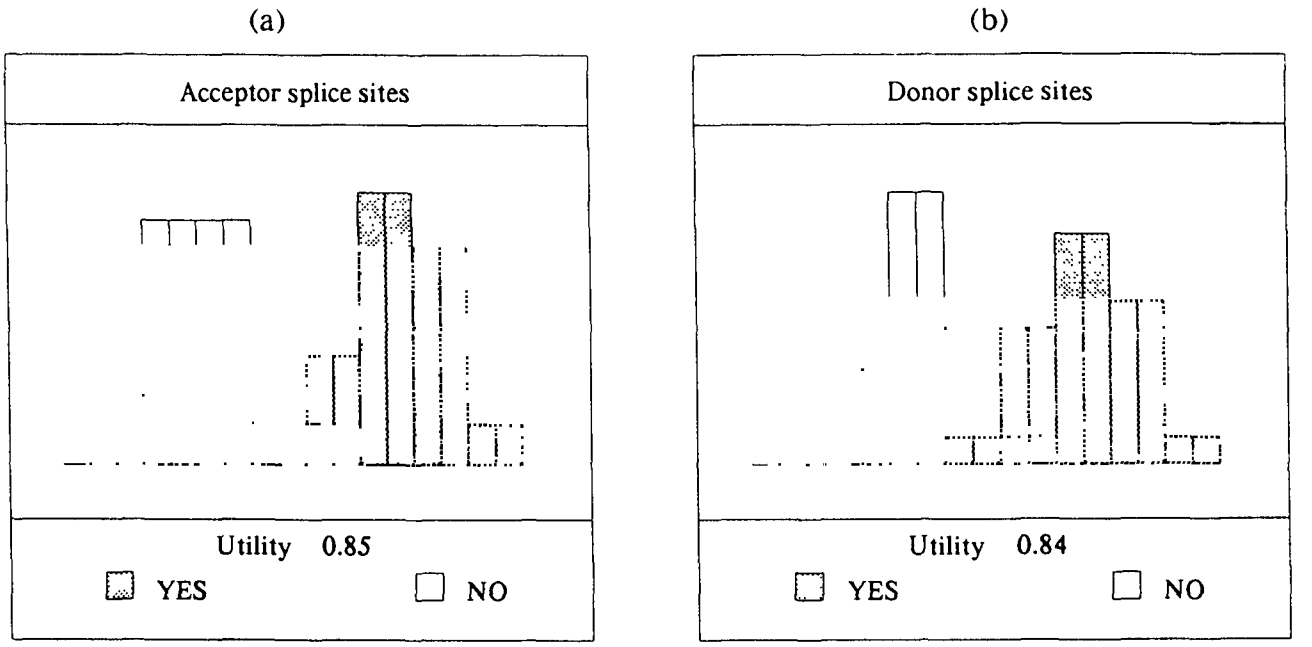

Fig. 6. The final complex characteristics used for recognition of the acceptor and donor sites: (**a**) a histogram for the distribution of the values of the complex characteristic A of the sequences from the training sample—the acceptor sites (hatched column) and random sequences (open column); (**b**) the same distribution for the characteristic D of the donor sites and random sequences.

to the acceptor sites because both match with one of the consensuses given in Table III (see Figure 3). This is correct for region I and incorrect for II. However, when additional contextual factors are taken into account, these sequences can be recognized by FACULT. The value of the complex characteristic $A(\text{I}) = 13.59 > 0$, whereas $A(\text{II}) = -9.38 < 0$, and, therefore, FACULT recognizes region I as an acceptor site and region II as a non-acceptor site.

We tested SPLICE and OBLIGAT on a set of control data containing splice sites that were not included in the training sets. As additional controls, we used sequences of 20 human genes with known exon–intron structure and total length of 127 862 bp. The results of these two control tests are given in Table V. As shown in Table V, consideration of the facultative contextual factors leads to an obvious improvement in the prediction results. A slight increase in type I error (underprediction for donor sites), decreases type II error (overprediction) by 1.5 times for the donor sites and by 2 times for the acceptor sites. Table V also shows that the method yields better results in the sequence regions that correspond to the exons. Type II error for the acceptor sites is then 0.37% and for the donor site 0.44%.

Thus, the methods described demonstrate the high accuracy of the recognition on independent data.

Table III. Consensuses of donor and acceptor splice sites

Splice sites	Consensus	No. of sites corresponding to a particular consensus
Donor	nnnnnnnnnnAG/GURAnnnnnn	255
	nnnnnnnnnnnG/GURAGnnnnn	182
	nnnnnnnnnnnn/GURAGUnnnn	93
	nnnnSnnnnnRG/GUASnSnnGn	8
	MKnYnnSnVnRG/GUnnnnnnnn	10
Acceptor	nnnnnnYYYYYnnYYnCAG/nnnn	193
	nnnnnYnnnnnYnYYnCAG/Gnnn	104
	nnnnYnYnYnYnnnCAG/nnnn	94
	nnnnBnnYnnnBCUnnYAG/nnnn	13
	nYnUnnYnnnnYBYnnHAG/Vnnn	12
	KnKnUnUnnnnnnnnnTAG/nnnn	6

Discussion

The present method is novel in treating functional sites as objects with complex structural organization. Also considered are different quantitative features modulating to different extents the activities of the sites. The method makes it possible to take into account both multistep function and the hierarchical relationships between the various regions of these sites.

In our investigation of the splice sites, we assumed the presence of the following two elements: obligatory features characterized by their fixed location in the sequence, and facultative features whose location in the sequences is not invariable. The obligatory features seem indispensable for the basal level of site activity; the facultative features modulate the site function. The results provide evidence indicating the important role of these elements for the whole organization of the splice sites. In addition to the features known to date not to be invariable in a sequence—e.g. the poly(Y) tract in the

Table IV. List of the genes used for control testing of the program SPLICE

EMBL identifier	Accession no.	Length (bp)	No. of exons
HSAPOE3	M10065	5515	4
HSAPOC2B	J02698	4057	4
HSAPOA2	X04898	3360	4
HSALPHA	J03252	4556	11
HSALDA1	X06351	3586	4
HSALBGC	M12523	19 002	14
HSALBG	M11518	7619	4
HSAFPCP	M16110	22 166	14
HSACTH	V01510	8658	3
HSACTGA	M19283	3583	6
HSACHRB	X02508	2318	2
HSACCYBB	M10277	3646	5
HSA1ATP	K02212	12 222	4
HSA1AR2	M19684	3758	3
HSALDB1	M15657	10 239	5
HSBSF2	Y00081	5961	5
HSC1A1	M20789	7616	25

Table V. Results of the acceptor and of the donor splice sites tested on control data

Program	Underprediction error (%)	Overprediction error (%)			Mean sequence length with a single 'false' signal (bp)
		Exons	Introns	Total	
Acceptor sites					
OBLIGAT	8.6	0.80	1.12	1.02	98
SPLICE	7.7	0.37	0.58	0.50	202
Donor sites					
OBLIGAT	5.5	0.62	0.78	0.74	136
SPLICE	7.8	0.44	0.46	0.47	214

acceptor sites—we succeeded in revealing ~100 other different facultative characteristics correlated with the distribution of mono-, di-, tri- and tetranucleotides and certain physicochemical properties of the donor and acceptor splice sites. Again, the suggested system is able to consider such features while building the methods of splice site recognition and thereby improving accuracy.

Also of interest is the presence of a set of correlated contextual features. Here, absence of one feature can be compensated by the presence of others. For example, inhibition of 'false' AG-signals near the cutting point of an acceptor site is ensured by lowering of the concentration of the AG-containing oligonucleotides within the region. The system takes into account such compensating effects and reveals complex characteristics in the given sites reflecting some general physicochemcial properties.

The involvement of gene regions adjacent to the splice sites in many other molecular and genetic processes is another noteworthy feature. They contain portions of the gene-encoding regions involved in translation. These processes (translation and splicing) set certain mutual restrictions on the structure of the sequences. And consideration of such correlations in the recognition methods is ensured through analysis of contrasting samples. With regard to the splice sites, contrasting samples involving regions of the exons or introns immediately adjacent to the cutting points are analyzed. The differences revealed are all taken into account in the methods of recognition with the recursive tools, thereby ensuring precise prediction of the donor and acceptor splice sites.

Testing of the program designed to predict the splicing sites assured us that consideration of the facultative contextual features yields improved prediction results compared with those obtained when relying on obligatory features. Type I errors, in fact, are reduced by a factor of 2. This decrease in overprediction error can greatly simplify the prediction of the exon–intron structure of the genes. In fact, a single acceptor site is falsely recognized per 98 nucleotides by OBLIGAT and per 202 nucleotides by SPLICE, which combines both obligatory and facultative features (Table V). The calculated mean length of the exons of the human gene is 161 nucleotides, and >80% of the exons are not longer than 200 bp (these calculations are based on data from the EMBL Data Library, 1990). This means that for the majority of the exons SPLICE will not find 'false' sites, thereby greatly facilitating the prediction of the exon–intron pattern in the genes.

In this way, our approach makes it possible to use all the available information about the structure of the functional sites. We believe that the method is advantageous in cases when thorough analysis of known sequences, as well as highly precise prediction of a certain functional site, are needed. And in this sense the suggested system is not just a tool for tackling technical problems. It allows one to examine thoroughly the internal structure of a site and thus to obtain new information.

Acknowledgements

This research was supported by the Russian National Programme 'The Human Genome' and 'Genome Engineering' and 'Bioinformatics' projects, Italy. The authors are grateful to Mrs A.Fadeeva for translating this paper into English.

References

Berg,O.G. and von Hippel,P.H. (1988) Selection of DNA binding sites by regulatory proteins. *J. Mol. Biol.*, **193**, 723–750.

Bolch,B.W. and Huang,C.J (1974) *Multivariative Statistical Methods for Business and Economics*. Prentice Hall, Englewood Cliffs, NJ.

Duda,R.O. and Hart,P.E. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

Fishburn,P.C. (1970) *Utility Theory for Decision Making*. Wiley, New York.

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

Gelfand,M. (1992) Statistical analysis and prediction of the exonic structure of human genes. *J. Mol. Evol.*, **35**, 239–252.

Green,M.R. (1986) Pre-mRNA splicing. *Annu. Rev. Genet.*, **20**, 671–708.

Holley,H. and Karplus,M. (1991) Neural networks for protein structure prediction. *Methods Enzymol.*, **202**, 204–224.

Iida,Y. (1988) Categorical discriminant analysis of 3′-splice signals of mRNA precursors in higher eukaryote genes. *J. Mol. Biol.*, **135**, 109–118.

Kel,A.E., Ponomarenko,M.P., Orlov,Yu.L., Mischenko,T.M. and Kolchanov,N.A. (1989) Computer system for functional sites analysis in polynucleotide sequences. In *Computer Analysis of Structure, Functions and*

Evolution of Genetic Macromolecules, IC&G SB AS USSR, Novosibirsk, pp. 221–242 (in Russian).
Kudo,M., Kitamura-Abe,S., Shimbo,M. and Iida,Y. (1992) Analysis of context of 5'-splice site sequences in mammalian mRNA precursors by subclass method. *Comput. Applic. Biosci.*, **8**, 367–376.
Mlinsky,M.A. (1975) *Frames for Representing Knowledge*. Energy, Moscow.
Padgett,R.A., Grabowski,P.J., Konarska,M.M., Seller,S. and Sharp,P.A. (1986) Splicing of messenger RNA precursors. *Annu. Rev. Biochem.*, **55**, 1119–1150.
Reed,R. and Maniatis,T. (1986) A role for exon sequences and splice-site proximity in splice-site selection. *Cell*, **46**, 681–690.
Rogozin,I.B., Kolchanov,N.A. and Milanesi,L. (1993) Classification of human splice sites. *Int. J. Genome Res.*, **1**.
Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.
Smith,C.W.J., Porro,E.B., Patton,J.C. and Nadal-Ginard,G. (1989) Scanning from independently specified branch point defines the 3' splice site of mammalian introns. *Nature*, **342**, 243–347.

Received on February 24, 1992; accepted on May 31, 1993

Appendix: utility of a quantitative characteristic

Let $X(s)$ be the quantitative characteristic for the sequences s. The distribution of the $X(s)$ value can be constructed for all the sequences of classes YES (s^+) and NO (s^-) (see Figure 4). We then obtain a utility value that is an integral quantitative estimate of the recognizing capacity of the given contextual characteristic X via analysis of distributions of the values $X(s^+)$ and $X(s^-)$ for the sequences of sites (class YES) and non-sites (class NO). The utility value is calculated by applying a set of expert rules based on three groups of statistical criteria.

Specificity

The specificity of the contextual feature A with respect to the functional site sequences may be estimated by analysis of the overlapping of the distributions of $X(s^+)$ and $X(s^-)$ values. Based on the overlapping values, the potential recognition error of the sequences of classes YES and NO is estimated. The smaller this overlapping is, the more useful is the contextual feature X in functional site recognition. In addition, the significance of the differences between the mean value of a given characteristic for the two contrasting samples is tested. Partial utility U_s is estimated. This estimate integrates all the values of the criteria derived from the expert rules we provided the system with.

Normality

In the theory of pattern recognition, as a rule, the distributions of the considered characteristics X throughout the samples of interest are assumed to be normal. Normality of distribution, as a rule, ensures reproducibility of results on the control samples. For this reason, characteristics with distribution close to normal are preferable. Similarity to the normal distribution is subjected to the χ^2 test. As a result, partial utility U_N is calculated.

Stability

The results of recognition should be stable; variations in the learning data should not affect them (samples YES and NO). To estimate this stability, the method of repeated usage of learning data was used. Stability of the prediction results for each tested characteristic is estimated as follows. Samples YES and NO are many times and randomly divided into two parts (YES_1), (NO_1) and (YES_2), (NO_2). Each time, based on the samples obtained after such division, the threshold value of the characteristic R is calculated: $R = (M_{YES} + M_{NO})/2$, where M_{YES} and M_{NO} are the mean values of the characteristic X for the sample YES and NO respectively. Then, using this threshold R, all the sequences of the sample YES_2 and NO_2 are classified as site and non-site sequences. If the mean value of the characteristic X for the sequence exceeds the calculated threshold R, the given sequence is classified as a functional site, if it does not exceed the R, it is classified as a non-site. If the results of the given classification do not agree with the real classification according to the classes YES and NO, the average number of underpredicted regions E_1 (in the case of wrong reference of sequences of functional sites to non-sites), and overpredicted regions E_2 (in the case of wrong recognition of non-site sequences as functional sites) are evaluated. With this multiple random division, the calculated average values E_1, E_2 characterize the 'stability' of the considered characteristic. Based on the results of the tests, the partial utility U_T is calculated. In such case the smaller the error of recognition is, the higher is the corresponding utility value.

Then, based on the additive theory of utility (Fishburn, 1970), an integral quantitative estimate of the utility U can be obtained through weighted averaging of the partial utilities:

$$U(X) = \alpha_1 U_S + \alpha_2 U_N + \alpha_3 U_T \qquad (7)$$

Here α_j is the weight coefficient of the partial utility ($\Sigma\ \alpha_i = 1$). The U is interpreted as follows: (a) $-1 \leq U \leq 1$; (b if $U(X) \leq 0$, the contextual feature X cannot be used for recognition of the considered type of functional sites; (c) if $U(X) > 0$, the contextual feature X can be used for recognition; (d) if $U(X) > U(Y) > 0$, the contextual feature X is preferable for recognition to Y.