Published online February 11, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 3 ***1131–1142***
DOI: 10.1093/nar/gkh273

Analysis and recognition of 5′ UTR intron splice sites in human pre-mRNA

E. Eden and S. Brunak*

Center for Biological Sequence Analysis, Biocentrum-DTU Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark

Received July 4, 2003; Revised October 21, 2003; Accepted December 16, 2003

ABSTRACT

Prediction of splice sites in non-coding regions of genes is one of the most challenging aspects of gene structure recognition. We perform a rigorous analysis of such splice sites embedded in human 5′ untranslated regions (UTRs), and investigate correlations between this class of splice sites and other features found in the adjacent exons and introns. By restricting the training of neural network algorithms to 'pure' UTRs (not extending partially into protein coding regions), we for the first time investigate the predictive power of the splicing signal proper, in contrast to conventional splice site prediction, which typically relies on the change in sequence at the transition from protein coding to non-coding. By doing so, the algorithms were able to pick up subtler splicing signals that were otherwise masked by 'coding' noise, thus enhancing significantly the prediction of 5′ UTR splice sites. For example, the non-coding splice site predicting networks pick up compositional and positional bias in the 3′ ends of non-coding exons and 5′ non-coding intron ends, where cytosine and guanine are over-represented. This compositional bias at the true UTR donor sites is also visible in the synaptic weights of the neural networks trained to identify UTR donor sites. Conventional splice site prediction methods perform poorly in UTRs because the reading frame pattern is absent. The NetUTR method presented here performs 2–3-fold better compared with NetGene2 and GenScan in 5′ UTRs. We also tested the 5′ UTR trained method on protein coding regions, and discovered, surprisingly, that it works quite well (although it cannot compete with NetGene2). This indicates that the local splicing pattern in UTRs and coding regions is largely the same. The NetUTR method is made publicly available at www.cbs.dtu.dk/services/NetUTR.

INTRODUCTION

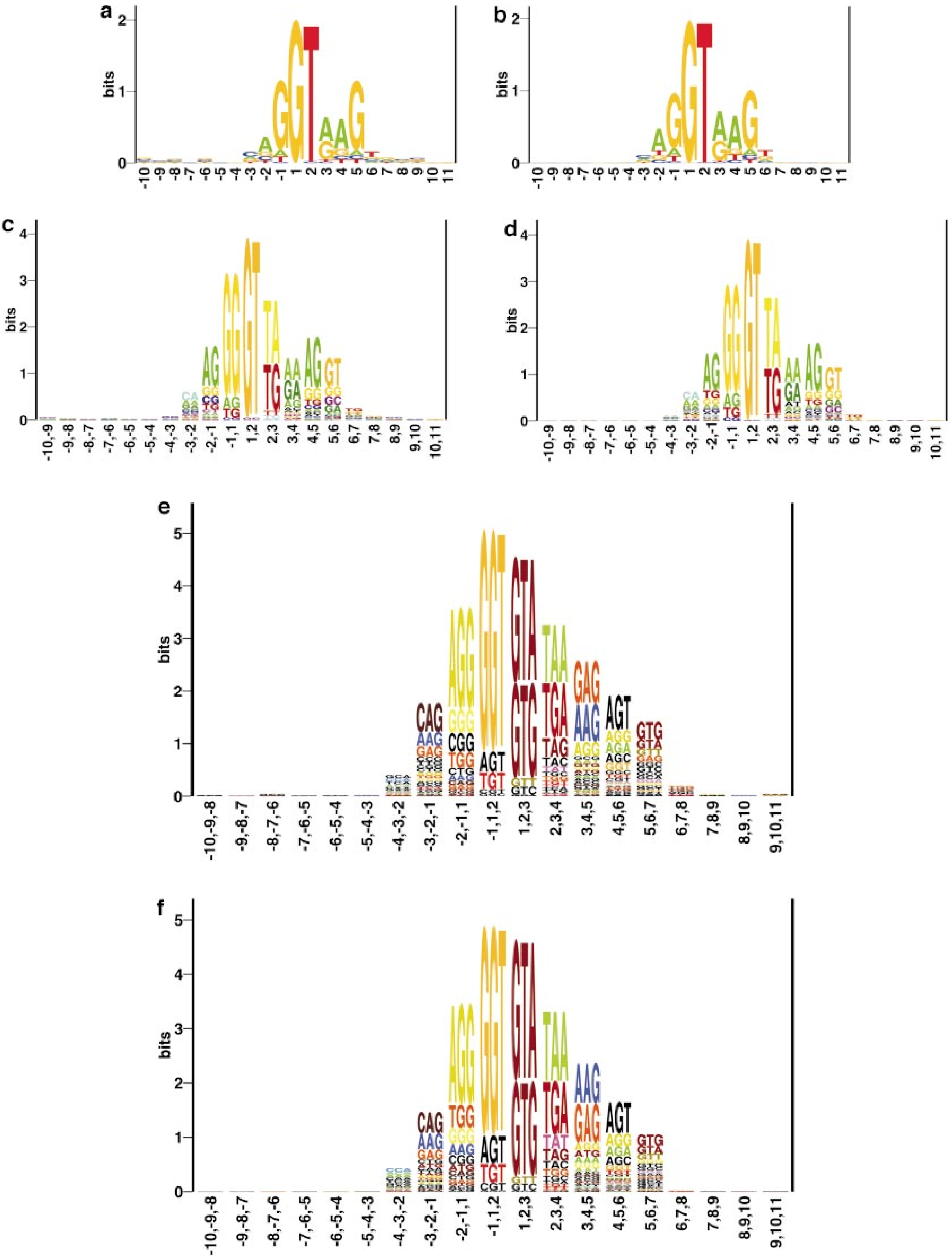
After the completion of human genome sequencing, more and more effort is being put into understanding the regulatory untranslated regions (UTRs) of genes. Various features of UTRs in general and 5′ UTRs in particular have been investigated (1–3). Among the important features are the characteristics of intron splice sites that reside in the 5′ UTRs.

Although programs for splice site and gene structure recognition [including GeneSplicer (4), NetGene2 (5), TWINSCAN (6), GenScan (7)] have reached a high level of performance on internal coding exons, with specificity and sensitivity above 90% at the nucleotide level (8) [~45% at the exon level (9)], predicting splice sites in 5′ UTRs still remains a challenge (9–11).

One of the major difficulties of splice site recognition in 5′ UTRs, as compared with internal exon splice site recognition, is that one cannot rely on the change in sequence when going from protein coding to non-coding. Irrespective of how the intron cuts the reading frame, the transition from protein coding to non-coding DNA is not that difficult to pick up by most methodologies, and it therefore eases the identification of the splice site location (5,12). This transition is of course absent in the 5′ UTR thus reducing the predictive performance of all methods that rely on differences between coding reading frames and non-coding sequence. Furthermore, alignment-based algorithms that exploit sequence similarity between a DNA sequence and a target protein cannot be applied when detecting UTR embedded splice sites either (10).

FIRSTEF, a new powerful tool for finding promoters and first exons, both coding and non-coding, was recently developed (8) with 86% accuracy for true positives and 17% for false positives. This method is novel in the sense that it exploits conserved non-coding patterns in the UTR, e.g. the CpG level in the region from –500 to +500 around the transcriptional start site. However, while FIRSTEF deals with the prediction of the first donor site, it does not predict the location of the first acceptor site, and since at least 40% of the 5′ UTRs cover more than one exon (8), these first acceptor sites are not well predicted using the currently available splice site prediction tools. Furthermore, our data show that of these 40% of 5′ UTRs, at least 9% have a second non-coding exon and 3% a third non-coding exon [we also found one case, AF135187, with a fourth completely non-coding exon (13)]. In all these cases both the donor and the acceptor splice sites are placed at the transition between introns and non-coding

*To whom correspondence should be addressed. Tel: +45 45252477; Fax: +45 45931585; Email: brunak@cbs.dtu.dk


a
bits
0
1
2
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11
b
bits
0
1
2
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11
c
bits
0
1
2
3
4
-10,-9 -9,-8 -8,-7 -7,-6 -6,-5 -5,-4 -4,-3 -3,-2 -2,-1 -1,1 1,2 2,3 3,4 4,5 5,6 6,7 7,8 8,9 9,10 10,11
d
bits
0
1
2
3
4
-10,-9 -9,-8 -8,-7 -7,-6 -6,-5 -5,-4 -4,-3 -3,-2 -2,-1 -1,1 1,2 2,3 3,4 4,5 5,6 6,7 7,8 8,9 9,10 10,11
e
bits
0
1
2
3
4
5
-10,-9,-8 -9,-8,-7 -8,-7,-6 -7,-6,-5 -6,-5,-4 -5,-4,-3 -4,-3,-2 -3,-2,-1 -2,-1,1 -1,1,2 1,2,3 2,3,4 3,4,5 4,5,6 5,6,7 6,7,8 7,8,9 8,9,10 9,10,11
f
bits
0
1
2
3
4
5
-10,-9,-8 -9,-8,-7 -8,-7,-6 -7,-6,-5 -6,-5,-4 -5,-4,-3 -4,-3,-2 -3,-2,-1 -2,-1,1 -1,1,2 1,2,3 2,3,4 3,4,5 4,5,6 5,6,7 6,7,8 7,8,9 8,9,10 9,10,11

exons making the task of identifying them much harder than for conventional translated region splice sites.

To deal with this problem we present a new data driven algorithm, NetUTR, for the prediction of all splice sites in the 5′ UTR. Using experimentally validated data extracted from GenBank, we created a high quality data set of 5′ UTRs, which extends through more than one exon. A large part of the available data was discarded due to conflicting splice site assignments and other errors (12). We then made a rigorous analysis of 5′ UTR splice sites and compared them with internal region splice sites. Based on this analysis we designed the neural networks training scheme, which was restricted not to contain any coding region sequence. The performance of NetUTR was then compared with that of other splice site predictors. Finally, we tested NetUTR on the coding gene regions. It was interesting to find that a method trained on local splice site information not contaminated by translated sequence was able to recognize splice sites embedded in the coding parts of pre-mRNA with only a small reduction in predictive power.

MATERIALS AND METHODS

Characteristics of the data set

Since the success of the neural network prediction is highly dependent on the size and quality of the data set, a significant effort was devoted to the creation of a high quality data set. Initially, 589 genes were extracted from GenBank version 128.0. We were interested in examining 5′ UTR embedded splice sites; it follows that the criterion for extraction was the existence of at least one 5′ end, entirely non-coding exon. Genes containing known alternative splicing were discarded.

The following ‘sanity’ checks were performed in order to validate the correctness of the data and omit possible GenBank errors (12,14): (i) the reading frame did not lack a start codon at the beginning or a stop codon at the end; (ii) the reading frame did not contain an abnormal number of stop codons, that is no more than one (the entry AL137800 contained no fewer then 19 stop codons in the reading frame); (iii) the reading frame size should modulo 3 give 0; (iv) introns did have a minimal functional length of 58 (12); (v) no logical conflicts existed, e.g. donor/acceptor site in the middle of an intron, genes with no coding region. As a result of the above checks we discarded 25 genes.

The non-consensus introns, i.e. those that differ from the canonical dinucleotides GT for donor sites and AG for acceptor sites, were extracted and manually inspected. Out of the 34 non-consensus introns that were found, eight were GC-AG introns and one was a CT-AG intron. No AT-AC introns were found. Taking into account the size of the data set and the frequency of the rest of the non-consensus introns (14), i.e. those that are non-GC-AG, -CT-AG or -AT-AC, it is highly unlikely that these introns had correct splice sites and they were therefore discarded. For four of the GC-AG introns as well as the CT-AG intron we found shifts of +3 and –3 in the reading frame that transformed the non-consensus introns into a perfect consensus GT-AG intron. These corrective shifts in frame might suggest an interpretation shift or typing error; to be on the safe side we discarded these genes as well. After the removal of the ‘suspicious’ non-consensus introns, four GC-AG introns remained. The percentage and type of the non-consensus introns in our 5′ UTR data set is in agreement with the non-consensus frequency for internal regions (8,14).

Redundancy reduction

In order to assess the predictive capability of the neural network it is paramount to ensure that the training and test sets are sufficiently different. Therefore, redundancy reduction was performed on the remaining 530 genes. Using BLAST local alignment at the amino acid level and a 10^{-6} threshold for random match a match list was created [low complexity regions were filtered out using the SEG package (15)]. The Hobohm algorithm (16) was then applied to assign a cutoff similarity value, and to select a maximal set of non-similar sequences (data not shown). As a result, more than half the data set was discarded resulting in the final data set, Set I, consisting of 233 genes containing 274 splice sites of each type.

Data set division

Set I was divided into two subsets. (i) A training set consisting of the first 194 genes with 232 splice sites of each type, used as positive learning examples for the neural networks. We used all the GT and AG dinucleotides that reside in the 5′ UTR and are not annotated donor and acceptor splice sites, as negative examples. The training set contained 46 125 negative GT examples and 62 231 negative AG examples. (ii) A test set consisting of the remaining 39 genes with 42 splice sites of each type, 10 060 negative GTs and 12 944 negative AGs, was used for the evaluation of the neural network prediction performance. Data Set I was also used to perform cross-validation runs where the set was divided into five parts. To enable comparison with coding embedded splice sites, an additional data set, Set II, was extracted from GenBank version 128.0, containing 2590 genes, which had a 5′ UTR that resided entirely in the first exon. Potentially faulty entries were discarded as for Set I.

Architecture of neural networks

The networks used in this study were of the multi-layer error back propagation type (17). The networks consisted of three layers: an input layer, one hidden layer and an output layer, all fully connected. The sequence input encoded using the sparse encoding scheme that has been described elsewhere (17). The output layer consisted of one neuron, giving a value between 0.0 and 1.0, interpreted as category assignment for either the central nucleotide in symmetrical input windows or a predefined nucleotide in unsymmetrical windows. In most cases a cutoff value of 0.5 was chosen. A value larger than 0.5 was interpreted as a splice site prediction, while a value lower than 0.5 was interpreted as non-splice site prediction.

Figure 1. Single nucleotide, dinucleotide and trinucleotide logo plots for donor splice sites that reside in the 5′ UTR (**a**, **c** and **e**) compared with the corresponding coding region donor sites (**b**, **d** and **f**). Only slight differences are found suggesting a dominance of the splice site signals at the nucleotide level over the amino acid coding constraints.

The maximal correlation coeffcient, *C*, for the test set was used in order to quantify the neural network performance using early stopping (17):

$$C(X) = \frac{(P_x N_x) - (N_x^f P_X^f)}{\sqrt{(N_x + N_x^f)(N_x + P_x^f)(P_x + N_x^f)(P_x + P_x^f)}} \qquad \mathbf{1}$$

where X can be the categories I^D (intron donor splice sites) or I^A (intron acceptor splice sites). P_X are correctly predicted positives, i.e. annotated splice sites, while N_X are the correctly predicted negatives, i.e. GT and AG dinucleotides that are not annotated as donor and acceptor splice sites, respectively. Similarly, $P_X{}^f$ and $N_X{}^f$ are the incorrectly predicted positives and negatives. A perfect prediction gives $C\ (X) = 1$, whereas a truly imperfect one gives $C\ (X) = -1$. We used the sensitivity:

$$S_n = P_x/(P_x + N_x^f) \qquad \mathbf{2}$$

and the specificity

$$S_p = P_x/(P_x + P_x^f) \qquad \mathbf{3}$$

as means of quantifying and comparing the neural network performance with other prediction methods.

RESULTS

Comparison of UTR donor and acceptor sites with coding region splice sites

Correlations in the 5′ UTR donor splice sites were analyzed qualitatively and visualized using sequence logos (18) made at the single nucleotide, dinucleotide and trinucleotide levels and compared with that of translated region splice sites (Fig. 1a–f). We found a striking similarity between the 5′ UTR donor consensus patterns and the patterns in translated region donor sites, the only difference being an increased tendency for G and C at positions –6 to –10 and 7 to 9, respectively, at the 5′ UTR donor sites. As expected, both types of splice sites are complementary to the 5′ end of the U1 snRNA. This similarity in local splice signal indicates that the constraints on splicing at the nucleotide level dominate entirely over the amino acid coding constraints at exon–exon junctions.

The 5′ UTR acceptor splice sites were also visualized using logos (Fig. 2a–f). In addition to the strong well known human consensus at the acceptor site, the 5′ UTR acceptor site pyrimidine tract typically extends through position –3 and gradually fades until position –26, where it stops. It has a weaker bias for cytosine at position –3 and slightly stronger bias at position –4 and 4 than that of coding region acceptor splice sites. The bias for thymine is stronger at several positions including –5, –6 and –12.

Exon and intron length distributions

The length distributions of 5′ UTR exons (Table 1) and introns (Table 2) embedded in non-coding sequence were compared with those of translated regions. Entirely non-coding exons seem to be slightly shorter than internal coding exons. Among all the exon categories in the data set, the entirely non-coding exons had the lowest average length of 130 nt, against the average for the fully coding exons of 145 nt. Partially coding exons were significantly larger with an average of 232 nt. These numbers are in agreement with the data from Davuluri *et al.* (8). For the introns the median is a more appropriate means of comparison than the average because introns lengths vary considerably and therefore a few particularly large introns may increase the average significantly. The 5′ UTR introns have a higher median (1368 nt) than introns embedded in coding regions (929 nt). The initial coding introns, contained in the coding region, were also examined and were found to have a median (1035 nt) shorter than 5′ UTR introns, but higher than coding region introns.

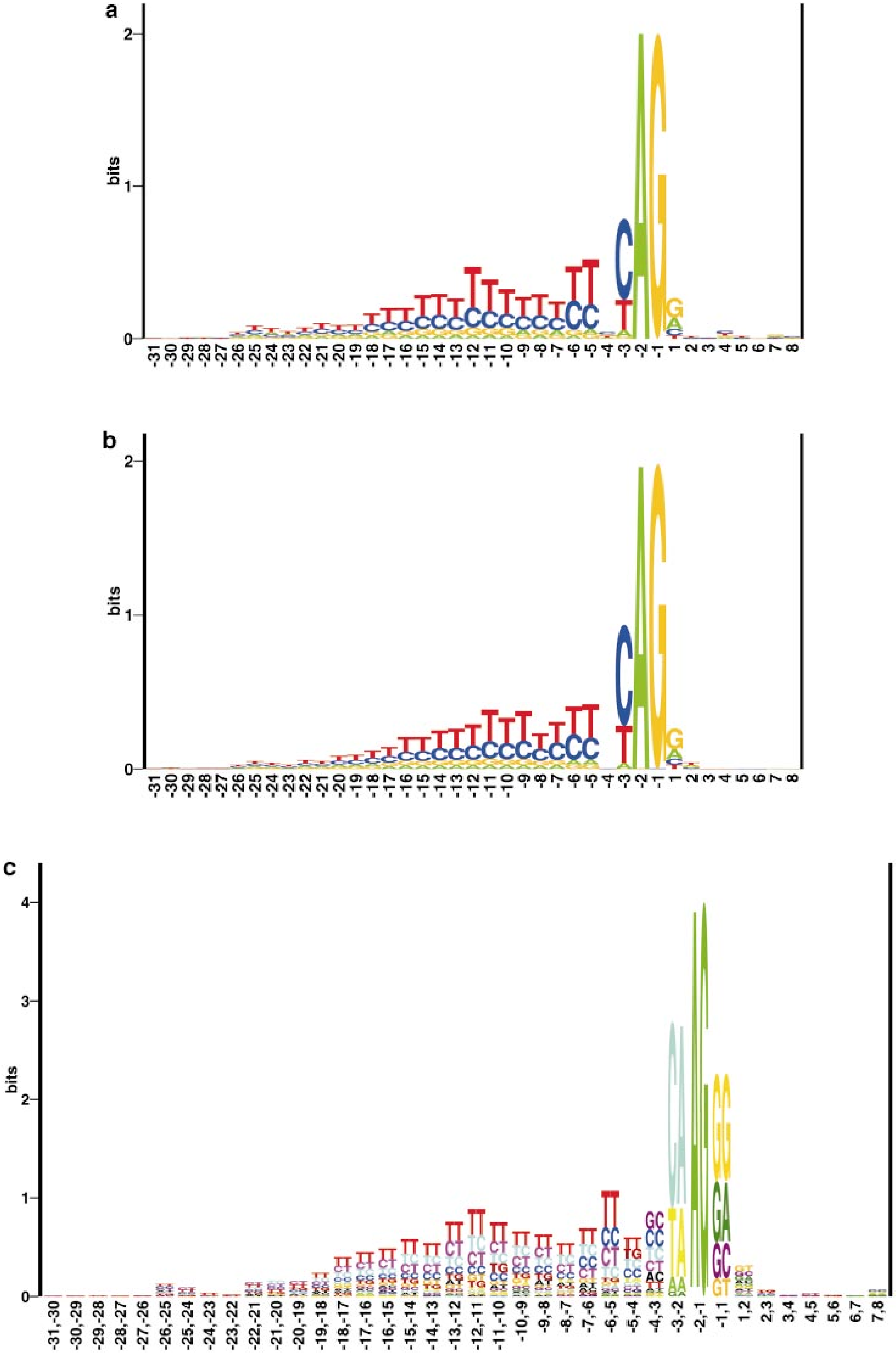
The average length of the 5′ UTRs in data set I is 296 nt. A division of the 5′ UTRs into three classes according to the number of non-coding exons that reside within them revealed a substantial difference in the length. The 5′ UTRs that extended through two exons (an entirely non-coding and a partially coding) had an average of 215 with a 789 SD. Those that extended through three exons had an average of 377 with a 181 SD and those that extended through four exons had an average of 420 with a 243 SD.

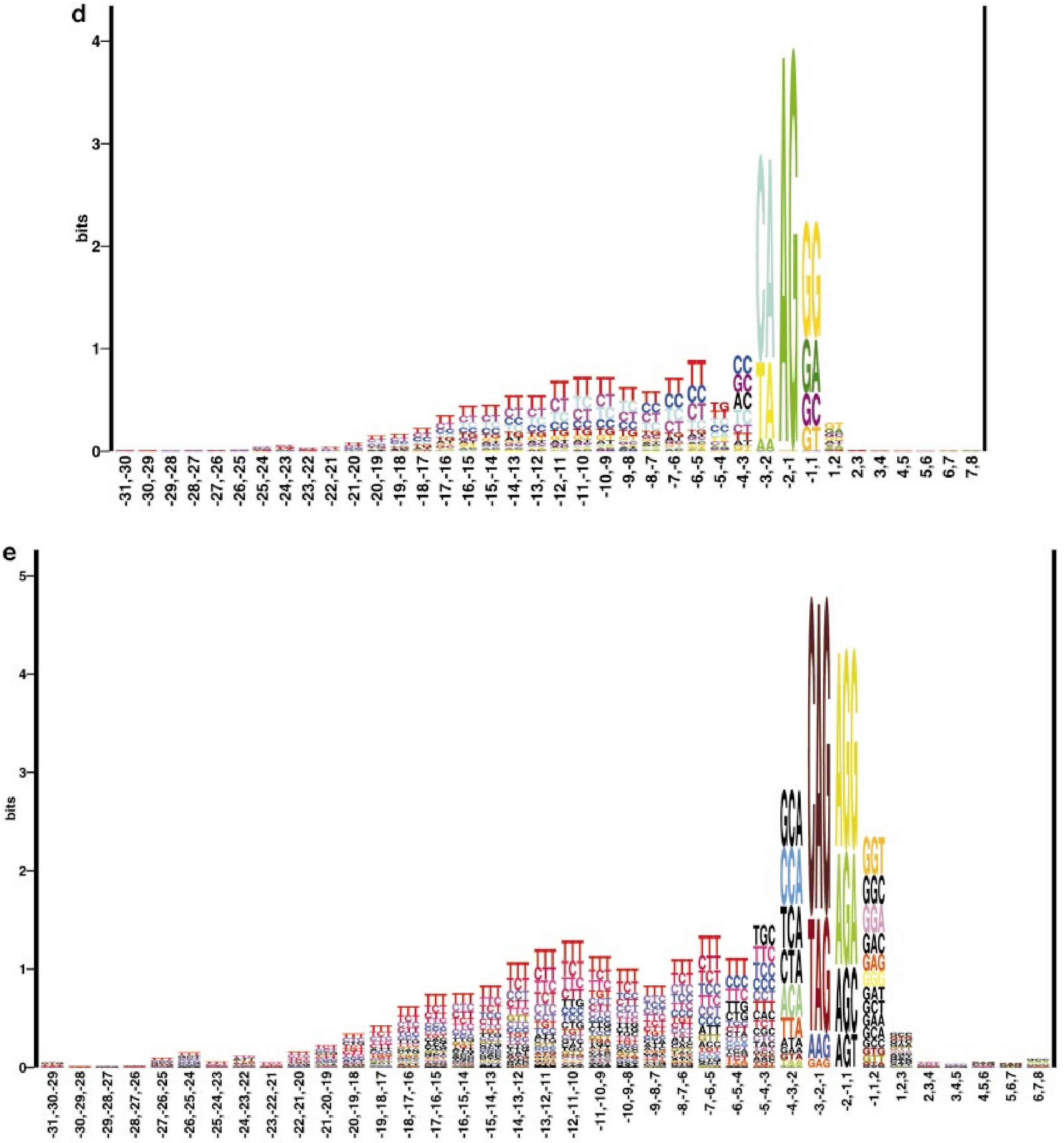
Nucleotide composition

We compared the nucleotide distribution of 5′ UTR exons and introns with exons and introns embedded in coding regions (Table 3). The base distribution in the two types of introns is quite similar, while the bias for C and G over A and T in the 5′ UTR exons is stronger than the bias in coding region exons. Exons in coding regions have a lower percentage of observed stop codons than the percentage expected by mere single nucleotide frequencies. This is caused by the suppression of the stop codons in the reading frame. One would not expect such suppression in 5′ UTR exons. We analyzed the data to test if the situation is the same here (Table 3). This was done by calculating the stop codon trinucleotide frequency in all three reading frames and comparing it with the expected stop codon percentage, which was calculated by multiplying the three single nucleotide frequencies. We used the ratio between expected and observed to quantify the degree of stop codon suppression. Introns from both coding and non-coding regions had the same suppression ratio with an observed stop codon frequency close to the expected. As anticipated, the coding region exons had a considerable suppression of stop codons. However, it was interesting to find that 5′ UTR exons also had some suppression of stop codons, stronger than the one found in introns, yet weaker than the one found in coding exons. This may be explained by the fact that some of the sequence regions currently annotated as non-coding are in fact alternatively spliced and actually coding. We report further evidence to support this claim in the next section. We checked the non-consensus splice site percentage (0.4%) and composition in our data set and found them to be identical to that of the internal introns (14,19) (for details see Materials and Methods).

Splice sites predicting networks

UTR donor sites. In order to find the optimal neural network which can extract the local splice site sequence pattern, various training methods were tested including a sliding


a
bits
2
1
0
-31 -30 -29 -28 -27 -26 -25 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8
b
bits
2
1
0
-31 -30 -29 -28 -27 -26 -25 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8
c
bits
4
3
2
1
0
-31,-30 -30,-29 -29,-28 -28,-27 -27,-26 -26,-25 -25,-24 -24,-23 -23,-22 -22,-21 -21,-20 -20,-19 -19,-18 -18,-17 -17,-16 -16,-15 -15,-14 -14,-13 -13,-12 -12,-11 -11,-10 -10,-9 -9,-8 -8,-7 -7,-6 -6,-5 -5,-4 -4,-3 -3,-2 -2,-1 -1,1 1,2 2,3 3,4 4,5 5,6 6,7 7,8



window that scans the entire sequence and searches for splice sites at every position in the 5′ UTR as well as a window that only inspects the environment around GT dinucleotides that constitute splice site candidates. We used a balanced training scheme where true positives were introduced at the same frequency as true negatives. Since our data set contains more negative examples than positive examples, in each epoch the neural network was shown all the positive examples and the same number of randomly selected negative examples. Other unbalanced training schemes were tested as well. Note that the balancing only affects the training and the presentation of the training data to the network, and not the test data, and hence not the evaluation of the predictive performance.

The inspection of GT splice site candidates using a balanced training scheme yielded far better results. We then tested a wide range of architectures with symmetrical input windows ranging from 5 to 43 nt, and hidden units ranging from 0 to 25. An architecture with a 21 nt window size and 20 hidden units

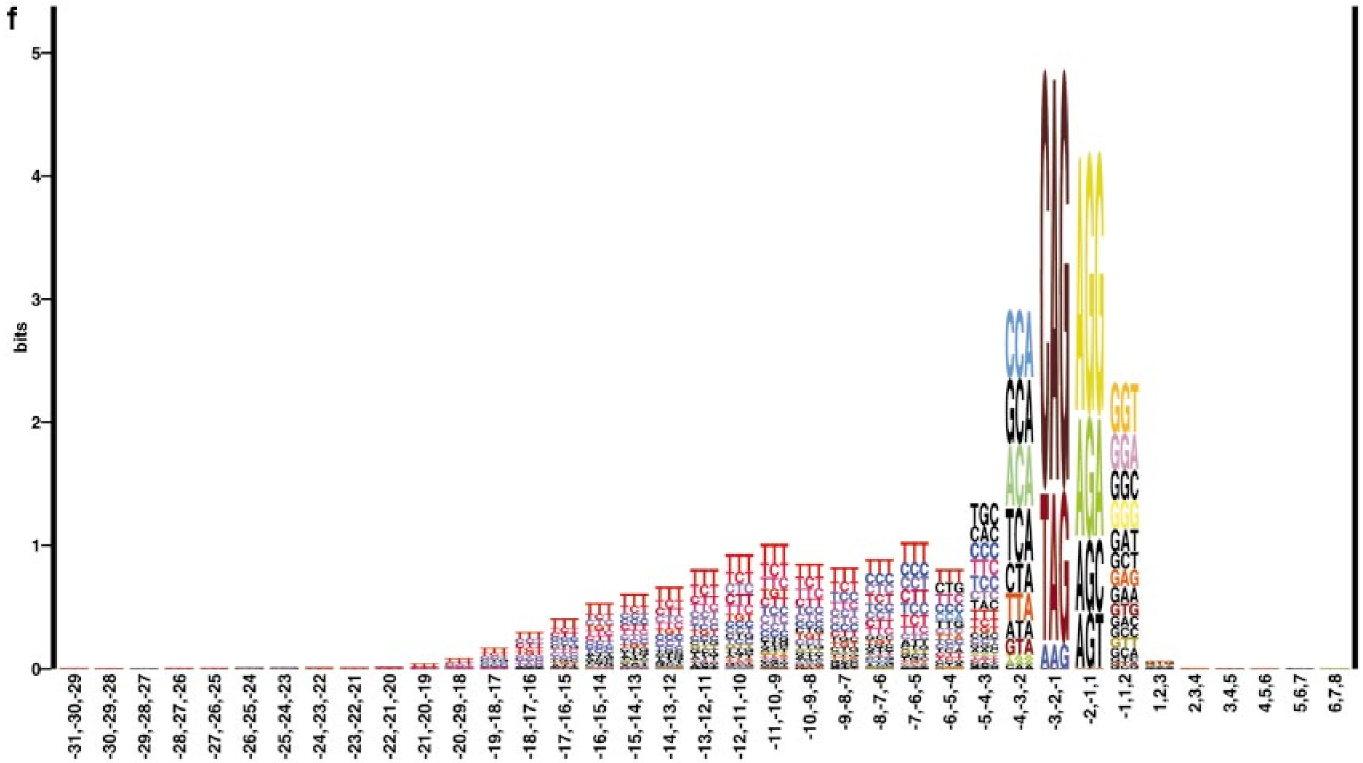


Figure 2. Single nucleotide, dinucleotide and trinucleotide logo plots for acceptor splice sites that reside in the 5′ UTR (**a**, **c** and **e**) compared with the corresponding coding region acceptor sites (**b**, **d** and **f**). The 5′ UTR embedded acceptor splice sites have weaker bias for cytosine at position –3 and slightly stronger bias at positions –4 and 4 than that of coding region acceptor splice sites. The bias for thymine is stronger at several positions including –5, –6 and –12.

gave the highest correlation coefficient (Fig. 3). This window size is 6 nt larger than the optimal 15 nt window size proposed in earlier work (5) for recognition of internal donor sites in human pre-mRNA.

This increase in window size is interesting (the performance is better by ~25% in terms of correlation coefficient). The absence of the otherwise dominant coding to non-coding transition pattern apparently enables the NetUTR network to pick up more remote and subtle signals correlated to splicing. A combination with a network with an even larger window of 35 enhanced the performance further and yielded a 0.45 correlation coefficient with 68% sensitivity and 30% specificity (ROC curve not shown). The results were similar when performing a 5-fold cross validation, training on 80% (randomly selected) of the redundancy reduced data set and testing on the remaining 20%.

An analysis of the nucleotide frequency of the false negatives and true positives assigned by the neural network was made (data not shown). The false negatives context had a *[G; A]* | *GT [G; A]X [G; T]* consensus while that of the true positives was *G* | *GT [G; A]AG*, the main difference being preference for adenine at position +4 and a preference for guanine at position –1 and +5, in the true positives. Furthermore, we found that the majority of the true positives were embedded in a C/G-rich environment that stretched more than 20 nt both up- and downstream of the donor splice sites, while the majority of the false negatives were embedded in a context with a bias for thymine. This observation is discussed in the section ‘CpG-specific splice site prediction’ below.

It is interesting to ask what precisely the network is picking up in the 21 nt (–10, +10) window extending three ‘extra positions’ nucleotides up- and downstream from the conventional 15 (–7, +7) nt window earlier found to be optimal in coding regions (5). By analyzing the weights of the trained neural networks it became clear that the network picks up additional signal both in the exon and intron parts of the local donor site region. In Figure 4 the weights connecting the input layer and the hidden layer are displayed as a so-called weight logo (21). The size of the symbols reflects the size and sign of the input-to-hidden weights weighted (multiplied) by the corresponding hidden-to-output weights. If negative, the letters are shown upside-down. The weight logo can be used to identify features which may influence the network to make a positive UTR donor site prediction. The central GT is upside-down reflecting the fact that the true GT donor sites are outnumbered by non-true GT sites—a GT in itself is therefore given a negative influence on a positive prediction. A true donor GT should therefore contain sequence context that will train the network into making a positive prediction. We analyzed in particular the additional signal in the outermost three positions (left and right) in the weight logo shown in Figure 4. The main trend in these six positions is that the presence of T and A inhibits a donor site prediction, while C and G mostly have the opposite effect. These preferences reflect a positional bias in favor of G/C and against A/T nucleotides at the corresponding positions in the 5′ UTR donor site context (see the logo in Fig. 1a and b). For example, at position 9, A and T inhibit and C excites. Interestingly, the strength of the compositional bias in terms of information content upstream of the non-coding exon/non-coding intron junction is comparable with the strength of the bias generated by the protein coding reading frame upstream of the conventional donor site (when all three reading frame interruption modes are combined in one logo).

UTR acceptor sites. As for donor sites, various training schemes were tested and again the inspection of the environment around AG dinucleotides that constitutes acceptor splice site candidates using a balanced training scheme proved to be the most efficient. When designing the network architecture we had to take into consideration the following limitations. The training should be 'clean' in the sense that the network should not be influenced by coding region sequence, i.e. the sliding window should halt before it penetrated into the coding region. At the same time the segment of the window extending downstream of the acceptor site had to be large enough in order to capture splicing signals that reside within the exon (of course no constraints existed on the upstream segment of the window since there was no fear it would penetrate the coding region). These two limitations are conflicting in cases where the 3′-most acceptor site resides close to the start codon. In these cases we had to decide whether to stop the training upstream of the most 3′ end acceptor site without the network being trained on that acceptor site as a classification position or let the network window partially overlap with the coding region. In order to decide on a distance from the start codon where the training should halt (a distance which fulfills the above constraints in most cases, and at the same time does not result in losing too a large portion of the data), we examined the distance distribution between the 3′-most acceptor site and the beginning of translation (20). This distance had an average of 41.4 nt with 61.9 SD. By restricting the distance upstream of the start codon to at least 8 nt, we were able to create a 'safety' zone that would substantially reduce the cases where the sliding window would partially penetrate into the coding region and at the same time only resulted in 24% loss of positive examples of acceptor sites and 0.23% loss of the negative AG examples.

We then tested many network architectures with symmetrical and unsymmetrical window sizes ranging from 19 to 81 nt and 5 to 25 hidden units. An architecture with a 53 nt window (–26, +26) and 20 hidden units produced the highest correlation coefficient (Fig. 5). Once again, the fact that the optimal window size extended 6 nt further upstream than the optimal window size of 41 nt found in previous work (5), may suggest that the network detects more subtle splicing signals, when trained in the absence of the strong non-coding/coding transition patterns. The final performance was enhanced by a linear combination of the above network with a 26 nt unsymmetrical window network (–17, +8)—the classification nucleotide being at position 0 (20 hidden units), resulting in a correlation coefficient of 0.36 with 59% sensitivity and 23% specificity.

The false negatives were characterized by a low cytosine frequency in the pyrimidine tract and a very weak pyrimidine signal in positions –5 to –8 upstream of the acceptor site. Given the big difference in ratio between true positives and true negatives (1:309) it was interesting to find that in 41% of the cases where the network combination missed one annotated acceptor site it assigned exactly one false acceptor site. In the Discussion we elaborate on this issue in relation to alternative splicing, which seems to be much more frequent in the 5′ UTR.

CpG-specific splice site prediction. It has been suggested that promoters and first exons fall into two classes: CpG related and non-CpG related (8). We tested whether this assertion holds for splice sites embedded in the 5′ UTR and, if so, whether our prediction method could benefit from a segregation into two CpG categories.

For each of the 5′ UTRs in our data set we calculated the CpG score using the Zhang method (8) in the following way: we used a 201 nt long window starting 500 nt upstream of the 5′ UTR and sliding 1 nt at a time until it reaches the 3′ end of the 5′ UTR (the window does not penetrate at all into the coding region). At each window position the CpG dinucleotide percentage was calculated and the window with the highest CpG percentage was defined as the CpG score of that 5′ UTR. In Figure 6 we show a histogram of the CpG scores in 5′ UTRs. The distribution is somewhat different from the earlier reported score distribution, as high CpG scores are more rare in 5′ UTRs (8). The difference may be explained by the fact that the earlier CpG score calculation was influenced by coding region nucleotide composition, i.e. the human coding region codons have a bias for C/G at the last codon position and a preference for G at the first (17) which significantly increases the sum of the CpG scores for regions with many subsequent codons.

In order to examine a CpG-specific prediction approach, the data set was then divided into two: CpG-related 5′ UTRs (UTRs with a CpG score <5%) and non-CpG-related 5′ UTRs (UTRs with a CpG score >7%) containing 70 and 121 genes, respectively.

We retrained the neural networks on the (smaller) CpG-specific data set. The prediction correlation coefficient for donor sites was ~0.4, lower than the performance of the neural networks trained on the original data set. After further analysis, we concluded that this decrease in predictive performance is caused by the reduction in data set size. We compared the decrease in prediction with the decrease that occurs when training a network on a random subset of the original data set with the same size as the CpG-related data subset. The result was a reduced prediction correlation coefficient of 0.37, caused once again by the data set reduction, only this time the decrease in performance was even higher. This means that, provided the data sets are large enough, the CpG-specific prediction approach may eventually outperform the approach we present here. This is not an uncommon situation in machine learning bioinformatics applications, where, for example, an organism-specific prediction method may be outperformed by a method trained on a broader set of sequences, where the larger data set then plays an essential role (22).

Comparison of NetUTR with NetGene2 and GenScan

We tested the performance of NetGene2 (5) on the entire UTR data set (Set I) and found that NetUTR was 2–3-fold better in terms of correlation coefficient (Table 4). The correlation coefficient for the donor site prediction of NetGene2 was only 0.274 with a sensitivity of 81% and a specificity of 1%. The acceptor site prediction of NetGene2 resulted in a correlation coefficient of 0.112 with a sensitivity of 46% and a specificity of 3%. The big difference in performance of NetUTR over NetGene2 is due to the difference in the level of false positives. While NetGene2 predicts 8.3 false positives for each true positive donor site, NetUTR predicts only 2.3 false positives for each true positive. The difference in performance is even bigger for acceptor sites, where NetGene2 predicts 32 false positives for each true one, while NetUTR predicts 3.2 false positives for each true positive. We think the reason for this high rate of false positive predictions in NetGene2 is the following. NetGene2 was trained to predict coding region splice sites, where the local splice site networks (corresponding to the NetUTR networks) are combined with a global coding exon/coding intron network, that changes the assignment threshold for each potential splice site. The hidden units in the conventional NetGene2 coding exon/coding intron predictor are strongly influenced by the compositional jump at the intron/exon and exon/intron boundaries. In fact, such feature detectors develop automatically in the hidden layer (in addition to reading frame feature detectors). It turns out that while coding and non-coding introns are quite similar in nucleotide composition (as reported in the section above on nucleotide composition), the UTR exons have a stronger nucleotide bias in the same direction as coding exons that makes the jump much stronger. This strong bias is not compatible with the threshold of the local NetGene2 network predictor in the vicinity of the exon–intron junctions and is therefore likely to trigger many NetGene2 splice site predictions, both true and false, in UTRs.

GenScan was run on the 5′ UTR data set using default parameters. It received entire genes as input (which is the normal input for GenScan). Only the splice site predictions in the 5′ UTR, both true and false, were taken into account in the comparison. The general performance was similar to NetGene2, with NetUTR having a donor site sensitivity a factor of 8.5 better, and a factor of 4 better for acceptor sites. As for NetGene2, it is not surprising that GenScan performs poorly, as during training, it only has been shown each UTR as one single non-coding region. It is perhaps surprising that GenScan does detect some true splice sites in the UTR. This mostly occurs when GenScan wrongly predicts a false start codon upstream of the real start codon and then it actually predicts some of the correct splice sites downstream and annotates them as if they were flanking coding exons (possibly for the same compositional reasons as described above). Another explanation may be unannotated alternative splicing in the 5′ UTR.

Table 1. Lengths of entirely non-coding exons, partially coding exons and coding exons

	Non-coding exons	Partially coding exons	Coding exons
Average	130	232	145
Standard deviation	197	266	140

Table 2. Length characteristics of introns embedded in the 5′ UTR and in coding regions

	5′ UTR introns	Initial introns in coding region	Coding region introns
Median	1368	1035	929
Average	3679	2862	1875
Standard deviation	6411	6045	3486

Translated region testing

No experimental evidence suggests that the biological splicing mechanism in the 5′ UTR is any different from that of the translated region. We tested NetUTR on introns within coding regions (data set II, as described in Materials and Methods) in order to examine whether or not our method supports this assumption. The performance of the combined neural networks for donor site predication yielded a 0.4 correlation coefficient with 44% sensitivity and 37% specificity. The results for acceptor sites were a 0.27 correlation coefficient with 28% sensitivity and 25% specificity. For both donor and acceptor sites prediction we noticed a decrease in sensitivity and an increase in specificity relative to prediction in 5′ UTR. The fact that the neural network performance only slightly decreased when predicting splice sites in the translated region—although it was trained on splice sites in the 5′ UTR—supports the claim that the splicing mechanism is similar in both types of regions and that it in the same way relies on local sequence information.

DISCUSSION

We analyzed splice sites embedded in 5′ UTRs and compared them with splice sites embedded in translated regions. Besides some minor differences in base frequencies both classes of splice sites were similar. By examining the 5′ UTR donor sites we were able to study splicing signals that were not constrained by reading frame patterns. The fact that we found the splicing signals in the 5′ UTR donor sites to be highly similar to the translated region donor sites means that the splicing signals are not altered when put in a coding context, thus indicating the dominance of splicing over that of amino acid coding.

Artificial neural networks were then trained to predict splice sites embedded in the 5′ UTR. To ensure a conservative

Table 3. The nucleotide composition of 5′ UTR exons and introns versus coding region exons and introns

	A	C	G	T	Observed stop codon	Expected stop codon	Ratio
5′ UTR introns	25.17	22.49	24.09	28.25	4.50	5.21	0.86
5′ UTR exons	21.44	28.89	28.98	20.70	2.88	3.52	0.81
Coding region introns	26.48	21.51	22.46	20.70	4.87	5.59	0.87
Coding region introns	25.05	26.07	26.65	22.23	3.09	4.36	0.71

The ratio between the expected and observed percentage of stop codons is used as a means of evaluating the stop codon suppression.

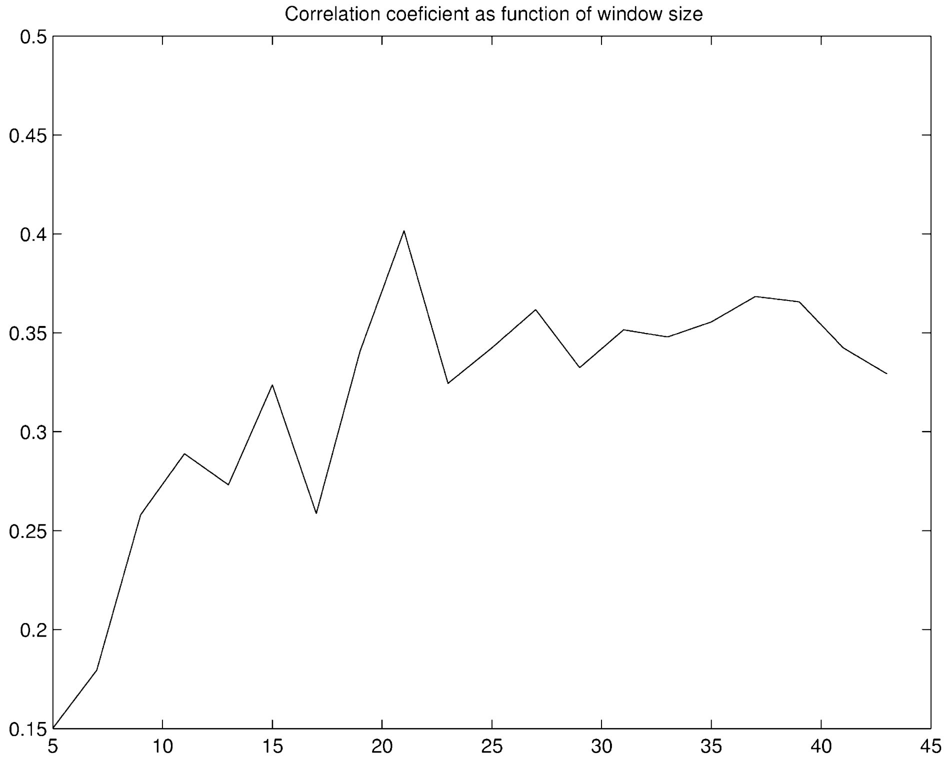


Figure 3. The maximal correlation coefficient for the prediction of 5′ UTR donor sites in the test set as a function of the neural network window size.

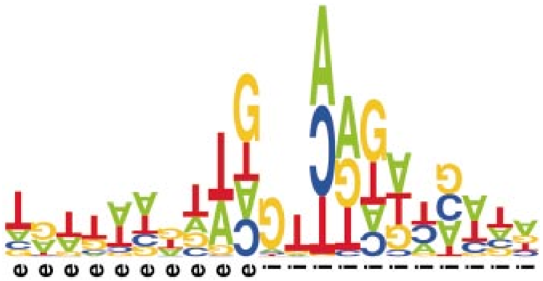


Figure 4. Visualization of the relative size and sign of weights in a neural network trained to identify donor sites in 5′ UTRs. The network window has 21 positions, and the symbol sizes in the weight logo indicate the position-specific sizes and signs of the input-to-hidden weights weighted (multiplied) by the corresponding hidden-to-output weights. If negative, the symbols are shown upside-down. The weight logo shows the 'contrast' between true GT UTR donor sites and other UTR GTs. The numbering in the window has been replaced by *e* and *i* indicating where the corresponding signal is found in the actual sequence.

estimation of our method's performance we used nearly 20% of the data for testing, and made sure that the sequence similarity of the training and test sets was low. Our motivation for restraining the neural network training to the 5′ UTR only was dual: first, we wanted to make sure that the neural network did not rely on the transition between coding and non-coding patterns, which is inherent to conventional methods that train on translated region splice sites; secondly, by eliminating the coding/non-coding 'noise' the neural network was apparently able to pick up subtler signals that may be involved in the biological splicing mechanism. This may explain the fact that our optimal window sizes both for donor and acceptor sites are larger than those described in earlier work, indicating the existence of splicing signals, which are remoter from the cleavage site. Our combined neural networks performed better on the 5′ UTR embedded donor sites than on the 5′ UTR embedded acceptor sites, yielding the final result of a correlation coefficient of 0.45 with 68% sensitivity and 30% specificity for the donor sites and a 0.36 correlation coefficient with 59% sensitivity and 23% specificity for acceptor sites.

Comparison with the NetGene2 method showed that the approach presented here is 2–3-fold better mainly due to a substantially lower amount of false positives. When inspecting the true positives and the false negatives it did appear that the true positives had a tendency to be embedded in more GC-rich regions, in contrast to the false negatives. This observation is in agreement with the work of the Zhang group, where promoters and first exons are classified as CpG related and non-CpG related, with the last category being hardest to detect (8). This would suggest that a UTR splice site prediction method could potentially benefit from a segregation of the data into two classes leading to two different types of predictors. However, when testing this idea it became clear that the reduction in data set size 'eats up' the gain when training CpG-specific networks. With more data this is most likely an easy way to enhance the performance.

Given the big difference in ratio between true positives and true negatives (1:309) it was interesting to find that in 41% of the cases, where the network combination missed one true acceptor site, it did instead assign exactly one false acceptor site. This could indicate that alternative splicing could be more common in UTRs. A number of papers confirm that this seems to be true (23–25). It has been suggested that alternative splicing in the 5′ UTR is coupled to differences in

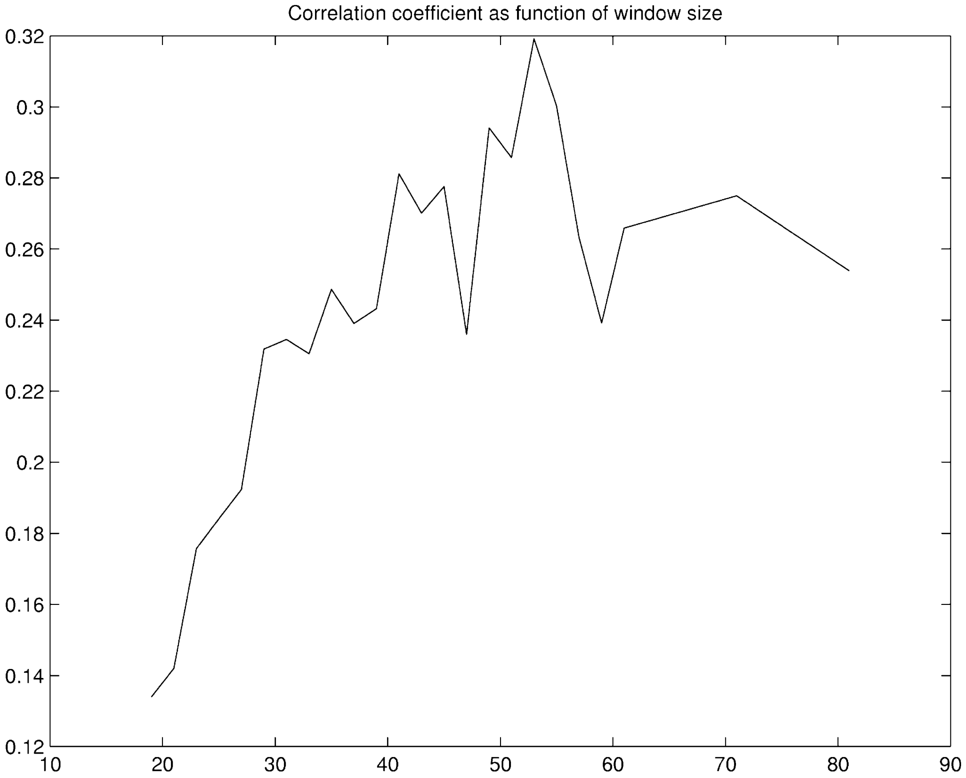


Figure 5. The maximal correlation coefficient for the prediction of 5′ UTR acceptor sites in the test set as a function of neural network window size.

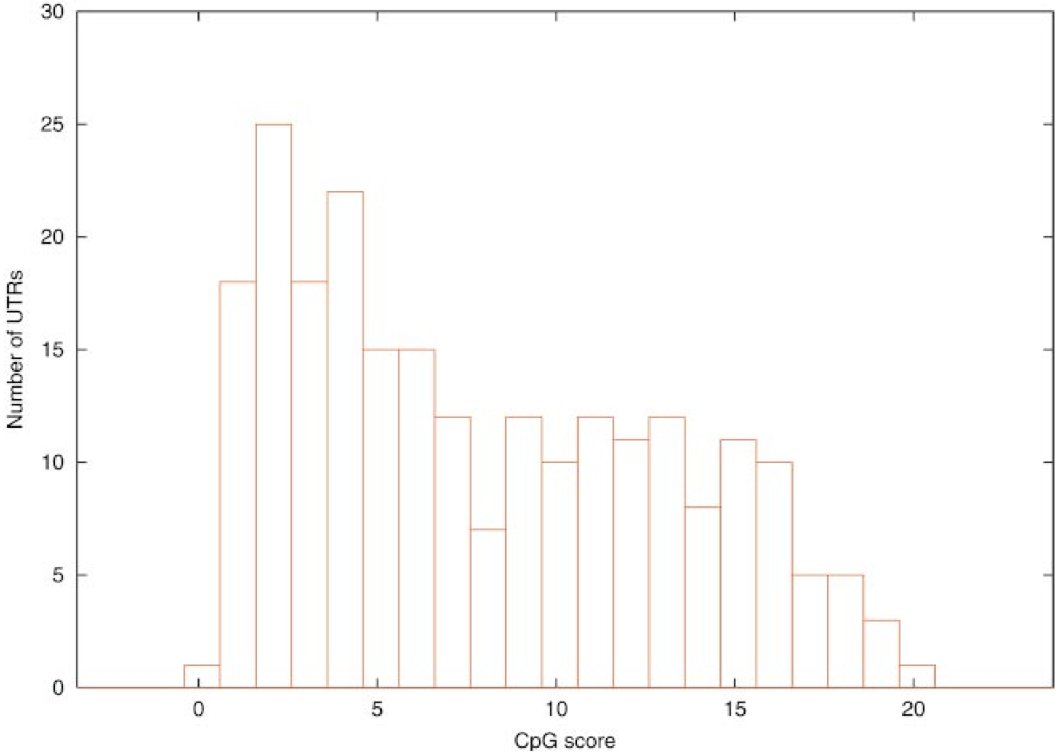


Figure 6. A histogram of CpG scores for 5′ UTRs. The CpG score was calculated using a 201 nt long sliding window that starts 500 nt upstream of the 5′ UTR. The window slided 1 nt at a time and for each window the CpG percentage was calculated. The CpG window with the maximal percentage was defined as the CpG score of that 5′ UTR.

Table 4. The predictions of 5′ UTR embedded splice sites were compared between NetUTR, NetGene2 and GenScan

	NetUTR	NetGene2	GenScan
Donor site			
Correlation coefficient	0.45	0.27	0.15
Sensitivity	0.68	0.80	0.08
Specificity	0.30	0.01	0.27
Acceptor site			
Correlation coefficient	0.36	0.11	0.18
Sensitivity	0.59	0.46	0.13
Specificity	0.23	0.03	0.25

The quantitative results are given at the nucleotide level.

transcriptional initiation points and a mechanism that allows cells to use several differently regulated promoters for the same gene (23). In other work (24) it was also found that alternative splicing seemed to be more frequent in the UTR than in the coding region, and that alternative splicing often skips the known 5′ or 3′ terminal by inserting a new intron or by expanding an existing intron (24). A UTR reconstruction algorithm based on EST data was able to reconstruct 3′ UTRs with a 72% success; however, it encountered significant problems in 5′ UTR reconstruction, where only 15% of the regions were successfully reconstructed (25). This decrease in performance is attributed to low EST coverage as well as to the high frequency of alternative splicing occurring in the 5′ UTR (23–25).

The pattern in the false positive predictions from NetGene2 in 5′ UTRs indicates that the NetUTR method (that is based on local splice site information) would benefit from a combination with a global predictor discriminating between non-coding exons and non-coding introns. This approach reduces the threshold for splice site assignment near exon–intron boundaries and increases it otherwise in regions of no, or small change, in 'exon-ness' (5). Such a network would then, as in NetGene2, be able to control the threshold for splice site assignment, thereby substantially reducing the level of false positives produced by NetUTR. We are currently working on an extension of the method along these lines.

Finally, the test of the NetUTR method in translated regions yielded a slight decrease in the overall performance that was characterized by a decrease in sensitivity and a balancing increase in specificity; the neural networks predicted more negatives, both true and false. The fact that the method did not lose much of its predictive performance when tested in the translated region, although it was trained only on 5′ UTRs, suggests that the splicing mechanism is similar in both regions and does not depend differently on the local splice site pattern.

Internet access

The prediction method can be accessed at www.cbs.dtu.dk/services/NetUTR or via email by sending the word 'help' to NetUTR@cbs.dtu.dk.

ACKNOWLEDGEMENTS

We thank Kristoffer Rapacki and Hans-Henrik Stærfeldt for their competent computer assistance, and the Danish National Research Foundation for support.

REFERENCES

1. Davuluri,R.V., Suzuki,Y., Sugano,S. and Zhang,M.Q. (2000) CART classification of human 5′ UTR sequences. *Genome Res.*, **10**, 1807–1816.
2. Kozak,M. (2001) Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**, 187–208.
3. Meijer,H.A. and Thomas,A.A.M. (2002) Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. *Biochem. J.*, **367**, 1–11.
4. Pertea,M., Lin,X. and Salzberg,S.L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, **29**, 1185–1190.
5. Brunak,S. Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.
6. Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.
7. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
8. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in human genome. *Nature Genet.*, **29**, 412–417.
9. Zhang,M.Q. (2002) Computational prediction of eukaryotic protein coding genes. *Nature Rev. Genet.*, **3**, 698–709.
10. Mathe,C., Sagot,M-F., Schiex,T. and Rouze,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, 4103–4117.
11. Gelfand,M., Mironov,A.A. and Pevzner,P.A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci. USA*, **93**, 9061–9066.
12. Korning,P.G., Hebsgaard,S.M., Rouze,P. and Brunak,S. (1996) Cleaning the GenBank, *Arabidopsis thailana* data set. *Nucleic Acids Res.*, **24**, 316–320.
13. Tazi-Ahnini,R., di Giovine,F.S., McDonagh,A.J., Messenger,A.G., Amadou,C., Cox,A., Du,G.W. and Cork,M.J. (2000) Structure and polymorphism of the human gene for the interferon-induced p78 protein (MX1): evidence of association with alopecia areata in the Down syndrome region. *Hum. Genet.*, **106**, 639–645.
14. Burset,M., Seledstov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.
15. Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
16. Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Enlarged representative set of protein structures. *Protein Sci.*, **1**, 409–417.
17. Baldi,P. and Brunak,S. (2001) *Bioinformatics—The Machine Learning Approach*, 2nd Edn. MIT Press, Cambridge, MA.
18. Thomas,D., Schneider,R. and Stephens,M. (1990) Sequnce Logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
19. Thanaraj,T.A. and Clark,F. (2001) Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon position. *Nucleic Acids Res.*, **29**, 2581–2593.
20. Tolstrup,N., Rouze,P. and Brunak,S. (1997) A branch point consensus from *Arabidopsis* found by non-circular analysis allows better prediction of acceptor sites. *Nucleic Acids Res.*, **25**, 3159–3163.
21. Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouze,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.
22. Nielsen,H., Brunak,S., Engelbrecht,J. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
23. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
24. Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (2001) Gene Structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889–900.
25. Kan,Z., Gish,W.R., Rouchka,E.C., Glasscock,J. and States,D.J. (2000) UTR reconstruction and analysis using genomically aligned EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 218–227.