© 1996 Oxford University Press Nucleic Acids Research, 1996, Vol. 24, No. 17 **3439–3452**

# Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information

**Stefan M. Hebsgaard, Peter G. Korning, Niels Tolstrup, Jacob Engelbrecht+, Pierre Rouzé[1] and Søren Brunak***

Center for Biological Sequence Analysis, The Technical University of Denmark, Building 206, DK-2800 Lyngby, Denmark and [1]Laboratoire Associé de l'INRA, V.I.B., University of Gent, K. L. Ledeganckstraat 35, B-9000 Gent, Belgium

Received April 12, 1996; Revised and Accepted July 12, 1996

## ABSTRACT

**Artificial neural networks have been combined with a rule based system to predict intron splice sites in the dicot plant *Arabidopsis thaliana*. A two step prediction scheme, where a global prediction of the coding potential regulates a cutoff level for a local prediction of splice sites, is refined by rules based on splice site confidence values, prediction scores, coding context and distances between potential splice sites. In this approach, the prediction of splice sites mutually affect each other in a non-local manner. The combined approach drastically reduces the large amount of false positive splice sites normally haunting splice site prediction. An analysis of the errors made by the networks in the first step of the method revealed a previously unknown feature, a frequent T-tract prolongation containing cryptic acceptor sites in the 5′ end of exons. The method presented here has been compared with three other approaches, GeneFinder, GeneMark and Grail. Overall the method presented here is an order of magnitude better. We show that the new method is able to find a donor site in the coding sequence for the jelly fish Green Fluorescent Protein, exactly at the position that was experimentally observed in *A.thaliana* transformants. Predictions for alternatively spliced genes are also presented, together with examples of genes from other dicots, monocots and algae. The method has been made available through electronic mail (NetPlantGene@cbs.dtu.dk), or the WWW at http://www.cbs.dtu.dk/NetPlantGene.html**

## INTRODUCTION

The biochemistry of splicing and the processing of introns in nuclear pre-mRNA in plants has not been understood to the same degree as in mammals and yeast (1,2). In virtually all organisms there has been much experimental evidence indicating that the selection of splice sites in pre-mRNA is based on information from different length scales in the nucleotide sequence (1,3). In plants the bias in the nucleotide composition of exons and introns has in particular been assigned an important role for the correct recognition of splice sites. Very often the high AU content of dicot introns is stressed (2,4,5). It has been claimed, based on experiments with synthetic introns, that appropriate splice site consensus sequences together with the elevated AU level are the principal requirements of the pre-mRNA to be spliced correctly (6). It was found, that the splicing ability of synthetic introns varies with infusions of AU-rich sequences. The latter may compensate for the complete lack of the polypyrimidine tract found in mammalian introns as suggested by work showing that soybean pre-mRNA cannot be spliced correctly by human HeLa cells (7). Even among monocot and dicot plants there seem to be large differences in splicing features, as experiments show that the pre-mRNA of monocots can only be poorly spliced in dicot cells (8,9). Also the more or less non-existent branch point consensus sequence, which seems to be reduced to a single adenine nucleotide in dicots differs markedly from the clear consensus sequence found in yeast. In plant genes, the presence of a strong donor site helps the recognition of a matching acceptor site (and vice versa), which would otherwise remain cryptic (10).

Identification of active splice sites from local sequence analysis is difficult due to the presence of a large number of false but consensus-like splice sites. This holds true for sequence analysis, but is likely to be true for the splice site selection *in vivo* as well. It is therefore important to use non-local information to filter out false positives. It is unclear precisely how this filtering works *in vivo*, but a number of computational methods and rules for removing false positives can be constructed. These include use of protein coding potential, predicted exon and intron length, and strength of neighboring splice sites. Furthermore some non-sequence specific knowledge can be used. This includes the exon and intron length distributions and the average GC content.

We present a data driven algorithmic approach for the recognition of splice sites based on the experimental evidence in the GenBank entries. The approach is a further development of

[*] To whom correspondence should be addressed

+Present address: Novo Nordisk, Scientific Computing, Building 9M1, Novo Alle, DK-2800 Bagsværd, Denmark the NetGene method (11), which is founded on the earlier observation of complementarity between splice site strength and the strength of the associated exon. Small exons (or long exons with a weak coding potential) tend to have strong consensus splice sites, while strong exons allow for weaker splice sites (11). A large part of the GenBank entries has been discarded because many of them are of surprisingly low quality, as they contain numerous false and conflicting splice site assignments (12). Many of these errors stem from incorrect interpretation of sequence data by the experimentalists.

## MATERIALS AND METHODS

### The data set

Considerable effort went into the preparation of a high quality data set which does not contain errors of the type previously found in GenBank (12). The main criteria for the genes extracted were: no missing exons; contains at least two introns; low sequence identity between genes. A detailed description of the methods used for extracting and correcting the data set is given elsewhere (12). The data set contains 146 genes extracted from GenBank (rel.87) comprising 764 donor sites and 766 acceptor sites.

The data was divided into two parts, where the first part was used for network training, and the second for testing the generalization ability of the final method. The training set contains 109 genes and two times 539 splice sites. The test set contains 37 genes, 225 donor sites and 227 acceptor sites. The imbalance of splice sites stems from two entries which start in the middle of an intron located in the 5′ UTR. In order to compensate for the fact that some GenBank entries contain parts of adjacent genes without annotation, each entry was reduced such that only 150 nucleotides (nt) before and after the transcribed part of the sequence were included.

The data set was divided into two parts, such that the first 109 genes constituted the training set and the remaining 37 genes constituted the test set (the complete set was kept in lexicographical order according to their GenBank LOCUS; 12). Pairwise comparison between the two sets was performed in order to check that no pair of closely related genes were present in both parts. This was done to ensure that the prediction method will extract general information about the splice sites rather than just memorizing the training set.

Two sequences were removed from the test set after the final evaluation of the splice site prediction system due to a seemingly wrongly placed exon in ATSUCSYN (X60987) and one wrong and one very suspicious acceptor site in ATU08315 (U08315).

ATU08315: from homology with z18242 it appears that the third intron is misplaced and should be six positions ahead (2043/2044 instead of 2049/2050). Homology with U20502 and z35108 confirms this. Furthermore, in the absence of a cDNA homolog, the borders of the last intron (which is located in a poorly conserved region) remain uncertain. We have therefore discarded the entry.

ATSUCSYN: the entire first exon shows no homology to other sucrose synthases, albeit these proteins are highly conserved. However, conserved sequence elements can be found, with the initiator ATG located at position 585 (instead of 464) and a possible donor site at 671/672. This produces a frameshift in the downstream exon, suggesting that there are likely sequencing errors in that area (maybe two Ts are lacking between position 664 and 666, that would give the canonical ending of the exon: SLFSR, and give the correct frame for the splice sites). Moreover, a poly-T tract would appear that again may present a sequencing problem for the determination of the precise number of Ts. Based on these considerations we have discarded the entry from the dataset.

### Neural network algorithms

The networks used in this study are of the multi-layer error-back-propagation type (13). They are fully connected and have three layers: an input layer, one hidden layer and an output layer. The network input is a segment of nucleotides from the nucleotide sequence. The sequence of nucleotides is sparsely encoded: A as (1000), C as (0100), G as (0010) and T as (0001) to avoid algebraic dependencies between nucleotides in the encoding. The output consists of one unit, giving a real valued output between 0.0 and 1.0. Using a threshold this number is interpreted as a category assignment for the middle nucleotide in the input window.

The networks were trained by standard error backpropagation (13) on two different tasks: (i) detection of coding nucleotides (versus non-coding nucleotides), and (ii) the prediction of splice sites (defined as the first and last intron nucleotide, respectively).

We used the correlation coefficient (14) to quantify the performance and stop the training of the coding/predicting networks:

$$C = \frac{(PN)-(N^{f}P^{f})}{\sqrt{(N + N^{f})(N + P^{f})(P + N^{f})(P + P^{f})}} \qquad \mathbf{1}$$

Here $P$ is the number of correctly predicted coding nucleotides (true positives), $N$ is the number of correctly predicted non-coding nucleotides (true negatives), $P^{f}$ is the number of incorrectly predicted coding nucleotides (false positives) and $N^{f}$ is the number of incorrectly predicted non-coding nucleotides (false negatives). Output activities larger than a threshold of 0.5 are interpreted as coding predictions, while output activities ≤0.5 represent non-coding predictions. A perfect prediction gives $C = 1.0$ whereas a truly imperfect prediction gives $C = -1.0$, which is actually just as good. A random prediction gives a value of $C$ close to zero. Networks that have been stopped with a maximal correlation coefficient have a balanced prediction of coding and non-coding nucleotides. A balanced prediction gives more information about the coding properties of the pre-mRNA than a biased prediction.

A different measure is used to evaluate and stop the training of the splice site predicting networks. The network training was stopped when the false positive rate at a sensitivity level of 95% was minimal. The false positive rate is given by

$$F = \frac{P^{f}}{N + P^{f}} \qquad \mathbf{2}$$

where $P^{f}$ is the number of incorrectly predicted splice sites and $N + P^{f}$ the total number of non-splice sites, while the sensitivity, or true positive rate, is given by

$$S = \frac{P}{P + N^{f}} \qquad \mathbf{3}$$

To keep the sensitivity level at 95%, the threshold separating the splice site predictions from the non-splice site ones cannot be kept at 0.5, but must be adjusted until $S = 95\%$. The virtue of this criterion is that a large number of true splice sites are detected. This is essential for the subsequent application of rules because the system only selects splice sites among the predictions from the splice site detecting network, see below.

### Information content in the splice site sequence context

The local sequence information available to the networks can be visualized using sequence logos (15), which are based on Shannon's information measure (16,17). The donor (or acceptor) sites are aligned, and for each column *i*, *R(i)* is computed

$$R(i) = 2.0 + \sum_{\alpha=A}^{T} P_i^{\alpha} \cdot \log_2(P_i^{\alpha}) \tag{4}$$

where $P_i^{\alpha}$ is the probability of finding nucleotide $\alpha$, $\alpha \in \{A,C,G,T\}$, at position *i*. In each column the four nucleotide letters have heights corresponding to their frequency (15).

## RESULTS

### Exon and intron lengths

The length distributions of *Arabidopsis thaliana* exons and introns were compared with the case of human genes. The average length of internal exons was 179 nt, which is longer than the average for human exons (≈150 nt). The bulk of the exons (74%) are between 40 and 200 nt long, the smallest is 9 nt, while the longest exon is 2151 nt. This compares very well with the length distribution of human exons (data not shown), and suggests that similar evolutionary mechanisms govern the internal exon lengths in plants and mammals. The intron length distribution (Fig. 1) differs from the length distribution of human introns. The average length of *A.thaliana* introns is 146 nt, while the average for human introns is much longer at 740 nt (Tolstrup, Dalsgaard, Engelbrecht and Brunak, manuscript in preparation). However, both distributions peak at an intron length between 80 and 90 nt, but where 84% of all *A.thaliana* introns are between 65 and 200 nt long, only 31% of human introns are found in this length interval, most of them are longer. This indicates that an intron length of 80–90 nt is favorable in both organisms. The longest occurring intron in the *A.thaliana* data set is 1242 nt long. A minimum intron length of 70–73 nt in dicots has been postulated earlier (6). Our data set contains four introns below this size. In ATHATCC1A:M85523 the shortest intron (second of two) is 59 nt long, in ATHANSYNAB:M92354 the tenth intron (of 10) is 63 nt long, in ATU06745:U06745 the second intron (of 10) is 69 nt long, and in ATU12126:U12126 the sixth intron (of eight) is 69 nt long. A minimum functional length of 55–70 nt is perhaps more realistic, at least for *A.thaliana*. This length is slightly smaller than the minimum length of 64 nt given by Filipowicz *et al.* (2).

We have investigated the number of introns in *A.thaliana* genes and compared them with the number of introns in human genes (Fig. 2). The highest number of introns found in the *A.thaliana* genes is 30 (ATHACOACAR:L27074). The average number of introns is five for both organisms and the distributions are very similar. These findings indicate that larger genomes like the human genome do not have more introns than small genomes, but rather that the length of introns increase with genome size only.

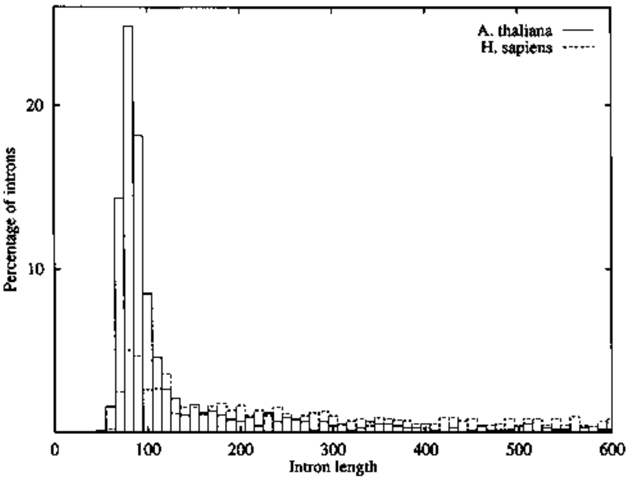


**Figure 1.** The two intron length distributions from *A.thaliana* (766 introns) and *Homo sapiens* (1573 introns) shown in one histogram. Only introns <600 nt are included.

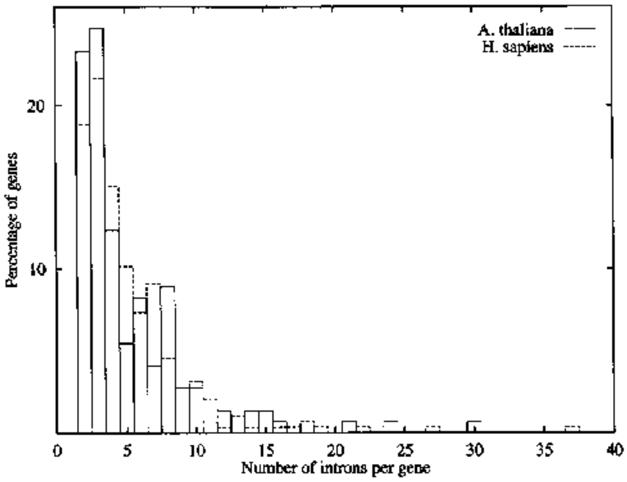


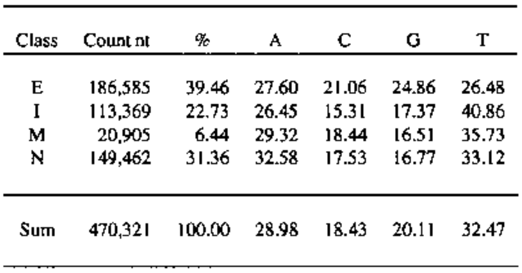
**Figure 2.** The number of introns per gene for *A.thaliana* (146 genes) and *H.sapiens* (286 genes). Only genes with at least two introns are included.

### Nucleotide frequencies

The nucleotide frequencies for exons and introns are given in Table 1. *Arabidopsis thaliana* genes have a high content of adenine and thymine, while the average frequencies of the four nucleotides is closer to 25% in human genes (data not shown). The introns contain less cytosine and guanine than exons and much more thymine, while the adenine content differs only by 1%. A similar tendency holds true for human genes although the absolute values differ.

In Table 2 the nucleotide frequencies of the average codon in *A.thaliana* is shown. The most frequent nucleotide(s) at position one is guanine, in position two adenine or thymine, and in position three it is thymine. The main difference from the reading frame of human genes is that the third position here is occupied preferably by cytosine or guanine.

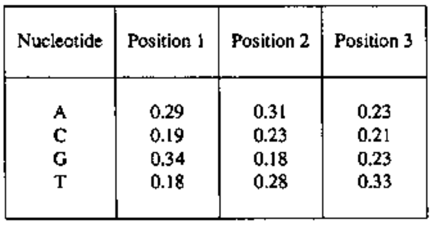
**Table 1.** The nucleotide distribution in the data set given for translated exon (E), intron (I), untranslated exon (M), and non-transcribed DNA (N)



| Class | Count nt | % | A | C | G | T |
|---|---|---|---|---|---|---|
| E | 186,585 | 39.46 | 27.60 | 21.06 | 24.86 | 26.48 |
| I | 113,369 | 22.73 | 26.45 | 15.31 | 17.37 | 40.86 |
| M | 20,905 | 6.44 | 29.32 | 18.44 | 16.51 | 35.73 |
| N | 149,462 | 31.36 | 32.58 | 17.53 | 16.77 | 33.12 |
| Sum | 470,321 | 100.00 | 28.98 | 18.43 | 20.11 | 32.47 |

Notice, in introns, the high presence of adenine and, especially, thymine.

**Table 2.** The nucleotide distribution at the three codon positions for the translated exon sequence in *A.thaliana*



| Nucleotide | Position 1 | Position 2 | Position 3 |
|---|---|---|---|
| A | 0.29 | 0.31 | 0.23 |
| C | 0.19 | 0.23 | 0.21 |
| G | 0.34 | 0.18 | 0.23 |
| T | 0.18 | 0.28 | 0.33 |

The non-organism specific reading frame pattern G/non-G on the two first codon positions is clearly visible (34).

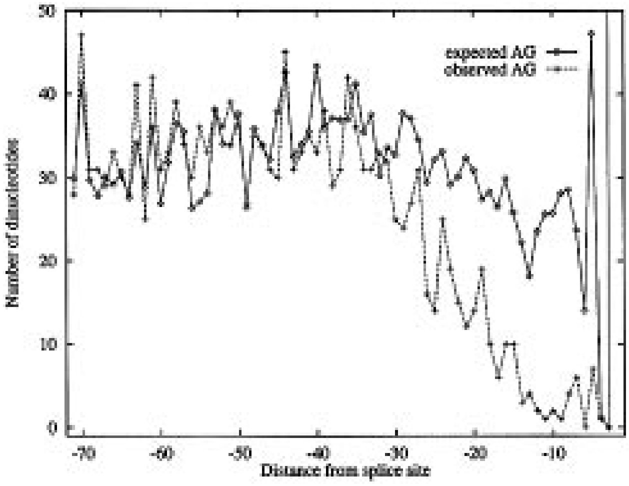


**Figure 3.** The expected and observed number of dinucleotides upstream from the acceptor site in the alignment of all acceptor sites in the data set. The expected number of AG dinucleotides (with the A at a given position) is the product of the frequency of A at that position and G at the next position multiplied by the total number of sequences (766).

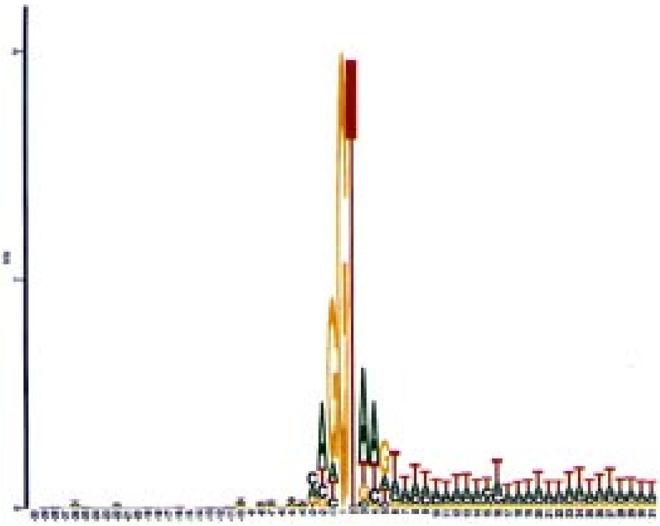


**Figure 4.** The sequence logo plot for the *A.thaliana* donor sites in the data set. The most frequent nucleotides correspond to the consensus sequence for dicot plant donor sites, AG|GTAAGT.

The dinucleotide frequencies and the ‘mutual information’ (17,18) of the exons and introns did correspond quite well to the frequencies found for dicots in earlier work (18). In the first 13 nt downstream from the donor site, there is generally a selection against the GT dinucleotide. Only at position five downstream can a positive selection for the GT dinucleotide be observed. Also upstream from the donor site GT is suppressed, and only 33 GT dinucleotides were found in the last 5 nt of the exons, while 75 instances were to be expected from the G and T frequencies. These findings support the view that the GT dinucleotide at intron position five is used for donor site recognition (19).

It has been proposed (20) that the scenario for localization of the acceptor site in mammals is the following: Once the lariat has been formed, the sequence from the branch point to the splice site, consisting of between 20 and 30 nt, is scanned, and the first AG dinucleotide is used as the splicing acceptor. To find out whether our data set supports this theory, we scanned for AG dinucleotides up to 70 nt upstream from the acceptor site and compared the result with the expected number of AG dinucleotides upstream from the acceptor site (Fig. 3). It is clear that there is a very strong selection against AG dinucleotides close to the acceptor site and 30 nt upstream into the intron. Only very few AG dinucleotides are found in this region consistent with the scanning hypothesis.

The sequence context of the splice sites has been visualized as logos (Figs 4 and 5). In the donor site logo (Fig. 4) we notice a lot of structure. The highest frequency nucleotides correspond to the well known consensus sequence for dicot plant donor sites (18), AG|GTAAGT. There is a lot more structure in the intron part than in the exon part, in particular, there is a high frequency of thymine in the introns (41%). In the exons the corresponding value is 26% (Table 1). In introns, adenine is the second most common nucleotide, 27%, while guanine and cytosine occur at 17% and 15%, respectively. According to Wiebauer *et al.* (7), the average thymine/adenine level for dicotyledonous plants is 73% in introns and 55% in exons. The *A.thaliana* genes examined in this paper have the percentages 67 for introns and 54 for exons. However, Goodall and Filipowicz (4) report that *A.thaliana* has the lowest known thymine/adenine level in dicots, namely 50.5%. This number is not confirmed by our analysis.

For the acceptor site logo (Fig. 5) we see much the same pattern with a lot of structure on the intron side and a high thymine level. There seems to be more structure on both the intron and the exon side of acceptor sites compared with donor sites. The dicot

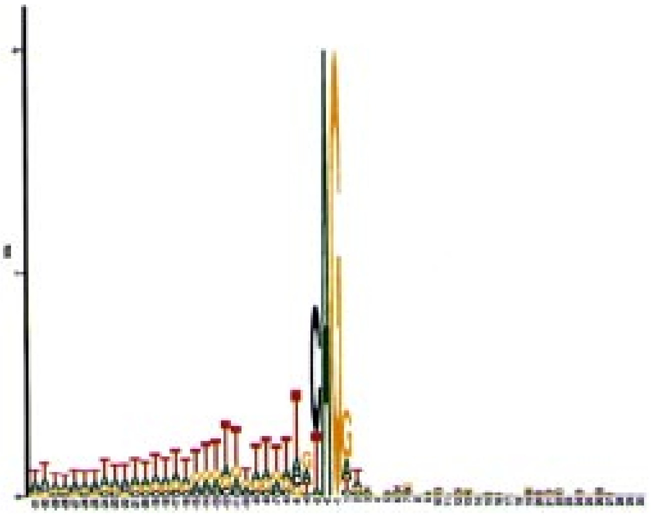


**Figure 5.** The sequence logo plot for the *A.thaliana* acceptor sites in the data set. The most frequent nucleotides correspond to the consensus sequence for dicot plant acceptor sites, TGYAG|GT.

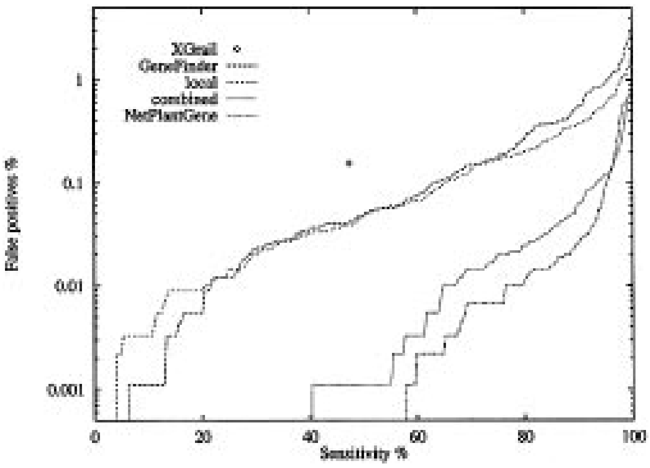


**Figure 6.** Percentages of false positive test set donor site predictions plotted against the sensitivity level for five different prediction methods. The line designated ‘local’ is the prediction of the ensemble of the local donor site predicting neural networks. The line designated ‘combined’ is the performance of the local network ensemble with a threshold controlled by the derivative of the coding prediction output. The NetPlantGene line is the final performance of the present method including the rule based system. The diamond is the performance of Xgrail, the sensitivity level is fixed for this method therefore only one data point appears on the plot.

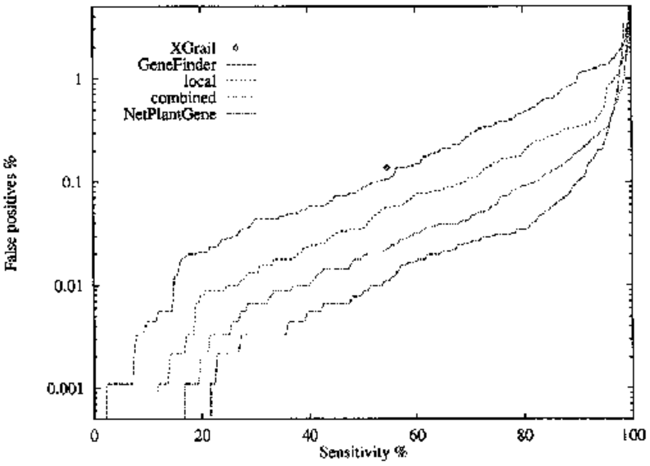


**Figure 7.** Percentages of false positive test set acceptor site predictions plotted against the sensitivity level for five different prediction methods. See legend to Figure 6 for details.

consensus sequence given by White *et al.* (18) is TGYAG|GT in agreement with the corresponding positions in the logo.

A separation of the splice sites according to their intersections with the triplet reading frame was also examined. While the resulting three logos differed somewhat in appearance, no informative pattern, was visible (data not shown). The ratio between the three possible intersections was 3:1:1, with the type of splice site that cuts the beginning (or the end) of the reading frame being the most common. In human genes the corresponding ratios are close to 2:1:1. It has been suggested that the weaker consensus sequences in plants, compared with humans, are somehow compensated by their large A and T content (21). Below, we return to the reading frame when we analyze the weights of the trained networks.

## Splice site predicting networks

To find an optimal network configuration for the donor site recognition problem, we did train and test a wide range of architectures. Networks with 3–71 nt in the input window and with 0, 2, 5, 10, 15 and 20 units in the hidden layer have been examined.

From these runs a network architecture with 23 nt visible in the input window and 10 hidden units was chosen. To further enhance the performance of the donor site recognition, 10 networks with this architecture initialized differently were trained. The average output of these networks was used as the result (a so called neural network ensemble). This ensemble was able to recognize 138 of the 225 test set donor sites with only 62 false positives, equivalent to a correlation coefficient of 0.65 (Fig. 6).

To get a view of the pattern of the false donor sites we have plotted the sequence logo for the alignment of all the test set non-donor sites that the network ensemble classifies as donor sites (data not shown). The false donor sites clearly follow the consensus of the *A.thaliana* donor sites. Also there is a clear overweight of thymine and adenine on the ‘intron’ side of the false splice sites. The fact that no network can make a better performance using local information, indicates that the selection may benefit from a combination of local and global sequence information.

We trained and tested the acceptor site networks on a lot of different network architectures. From these runs an ensemble of 10 networks with 61 nt present in the input window and 15 units in the hidden layer were chosen. The percentage of false positives as function of the sensitivity (true positive rate) is shown in Figure 7. The quality of the acceptor site prediction is very similar to the quality of the donor site prediction, showing that it is equally difficult to predict donor and acceptor sites from local information only.

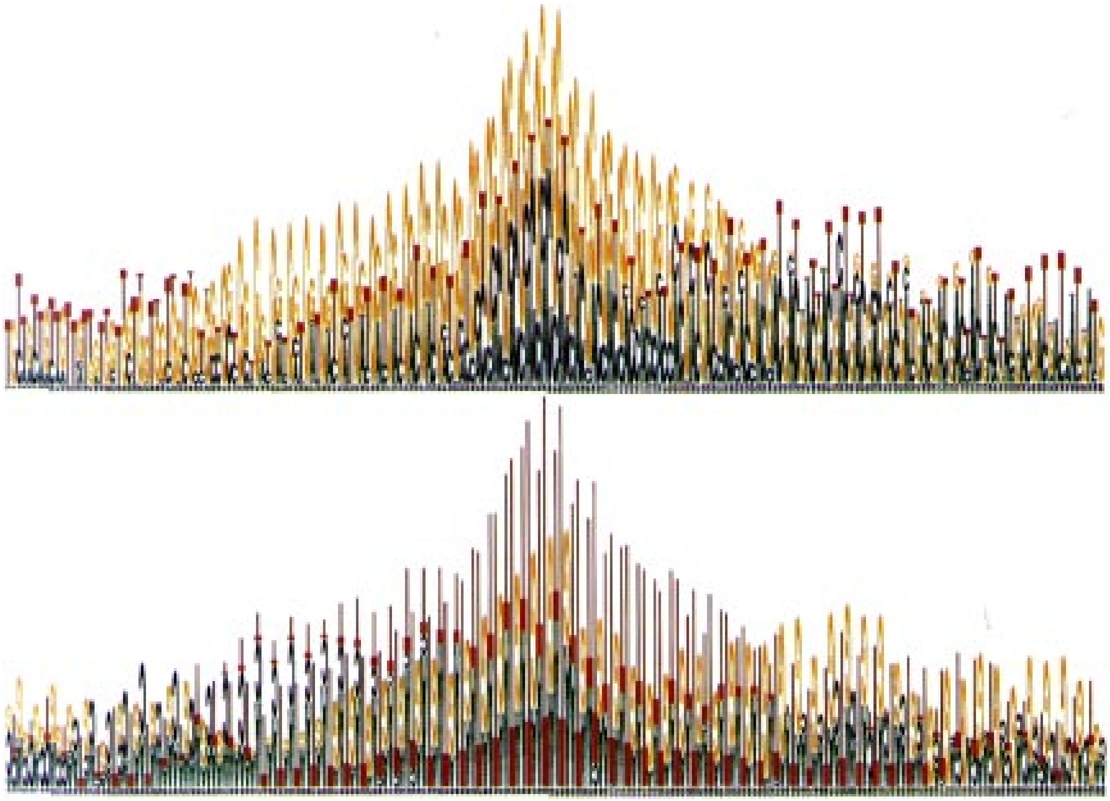


**Figure 8.** Combined weight logos of the positive (top) and negative (bottom) weights from the input to the three hidden units sensitive to the reading frame pattern in the window. For two of the weight vectors the components were shifted one position to the left and right, respectively. The logo therefore covers the input window from position 2 to position 200.

It is interesting that good acceptor site recognition requires a much larger window, 61, than donor site recognition, 23. However, we should not be too surprised that a difference exists, as in the cell different mechanisms in the spliceosome are used for the identification. Donor sites are recognized by base pairing to the U1 snRNA, while several less sequence specific elements seem to be involved in the recognition of the acceptor site.

As for donor sites we have plotted the information logo for the alignment of the false positive acceptor sites that the network assigned (data not shown). The false positive acceptor sites follow the consensus except in position –4 where the guanine is substituted by thymine. As with the false positive donor sites there is a clear overpopulation of thymine and adenine on the ‘intron’ side of the false acceptor sites.

## Analysis of the local network weights

It is highly interesting to understand as precisely as possible what sequence features the networks are looking for. These features are encoded in the weights, especially those connecting the input window positions and the hidden units. For each hidden unit its incoming weight vector will show the positions and nucleotide types that will excite or inhibit its activation.

Examination of the weights in the local network shows that they essentially learn what is present in the corresponding logos, together with negative weights of an anti-consensus sequence. For the donor site network the consensus sequence AG|GTAAGT can be identified as strong weights in the network. Donor sites are flanked by high frequencies of T in the intron part, but in the right part of the window the networks look for G deficiency instead.

The acceptor site network has strong weights for the consensus T(non-C)YAG|GNNG. This should be compared with the consensus read from the logo TGYAG|GT. In the network weights deficiency of C at position one in the logo is more significant than a large weight on G, and a strong weight for a G four positions into the exon can be observed as well. We assume that this G is part of the reading frame which is recognized in the exon by the acceptor site network.

## Recognition of coding DNA

In order to utilize global information which is available in the DNA sequence we have trained large window networks to discriminate between coding and non-coding nucleotides. When the middle nucleotide in the input window belongs to a translated exon the network will be trained to answer yes, otherwise no. We also trained networks to predict untranslated exons, but the prediction of untranslated exons proved to be very hard. Untranslated exons tend to be more intron- than exon-like. Their nucleotide frequencies correspond more to those of introns than those of exons (Table 1).

Networks with 101, 151, 201, 251, 301, 351 and 401 nt visible in the input window were examined, with different numbers of units in the hidden layer. The best network had a window of 201 nt, 15 hidden units, and a correlation coefficient of 0.75. This network was able to recognize 89.7% of the true coding nucleotides and 87.4% of the non-coding nucleotides. An

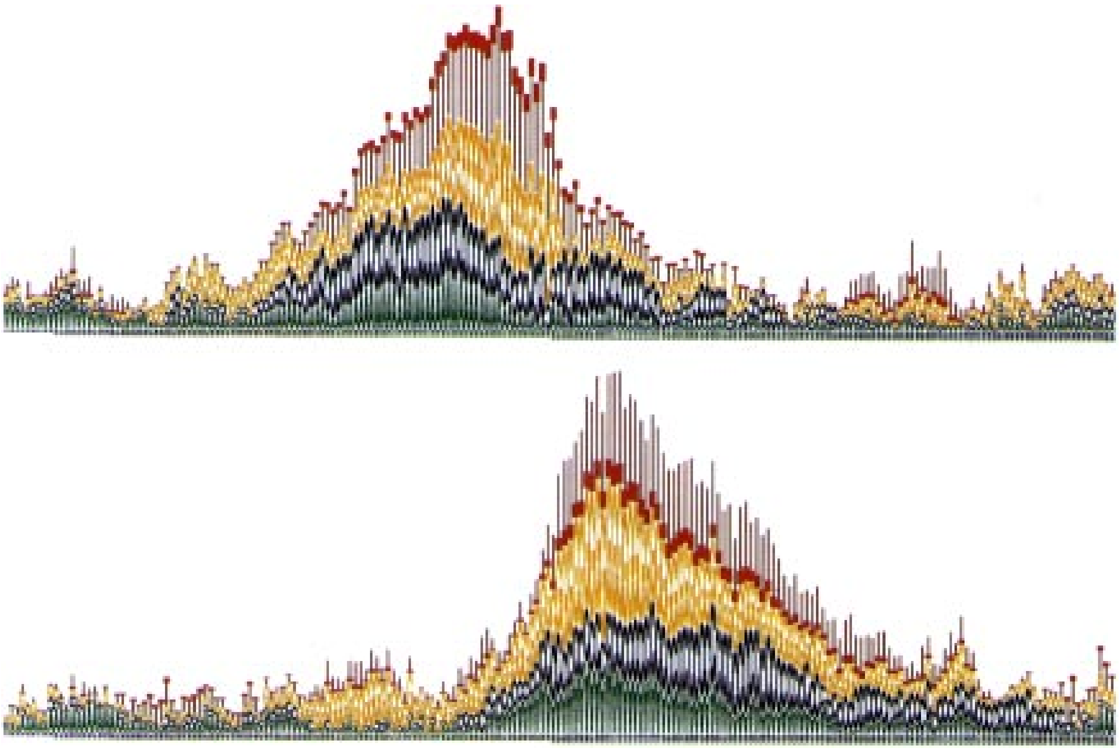


**Figure 9.** Weight logos for the three statistical hidden units which are sensitive to either intron- or exon-like DNA. The top shows the weights for one intron sensitive unit, while the bottom plot shows a combined logo for two exon sensitive units. Negative weights are indicated by upside-down symbols.

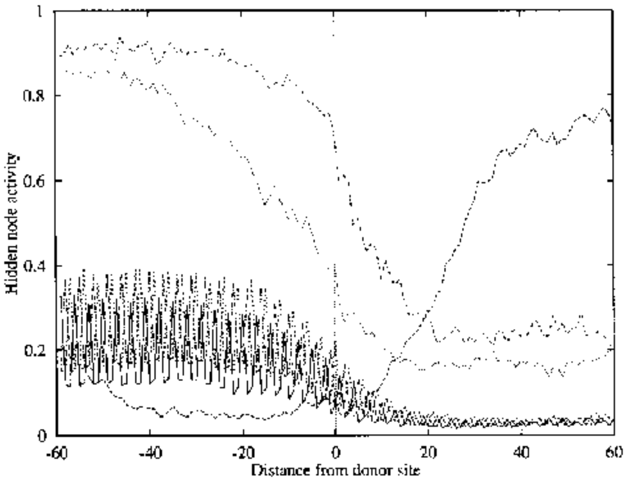


**Figure 10.** The average activity of the six hidden units at all donor sites in the test set. The qualitative differences of the reading frame units and the statistical transition-region recognizing units can be seen clearly.

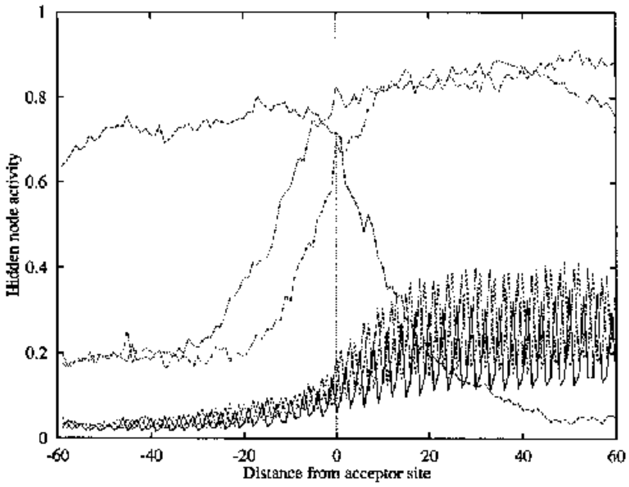


**Figure 11.** The average activity of the six hidden units at all acceptor sites in the test set. The qualitative differences of the reading frame units and the statistical transition-region recognizing units can be seen clearly.

ensemble of six networks, one with a window size of 101 nt, four with a window of 201 nt and one with a 251 nt window, was used in the final system. The reason for the use of networks with suboptimal window sizes in the ensemble is that these networks have a better performance for small or large exons, respectively. The joint correlation coefficient was 0.76, and the percentages 91.0% and 89.5%. The optimal window size for coding/non-coding networks trained on human genes (11) was 301 nt. This is a result of the difference in average intron length in human and *A.thaliana* genes.

## Analysis of the global coding/non-coding network weights

Networks with many different numbers of hidden units were analyzed, they all seemed to use similar detection principles in their internal working albeit with some smaller difference in their

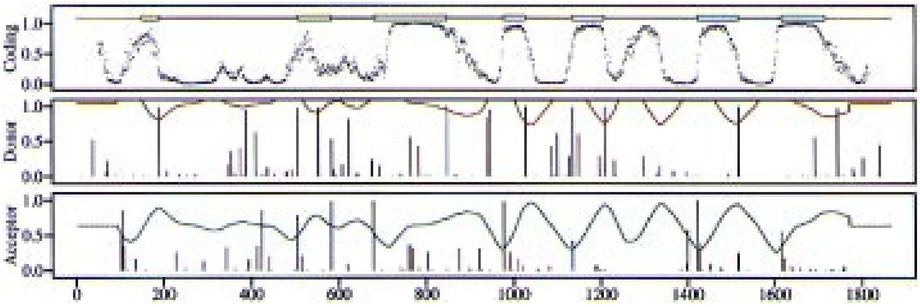


**Figure 12.** The NetPlantGene prediction for the test set sequence ATRAH1GNA. The correct exons are depicted as boxes in the top of the figure. The top plot designated 'Coding' shows the activity of an ensemble of coding predicting networks, values close to 0.0 indicate intron, while values close to 1.0 indicate exon. In the 'Donor' and 'Acceptor' panels the activity of the ensembles of the local splice site predicting networks are shown as impulses. An impulse with a height close to 1.0 indicates a strong *A.thaliana* splice site. A magenta impulse is a prediction that has been discarded during the refinement, and a cyan colored impulse is a prediction that has been changed by the rule based system. The variable threshold, computed from the coding predicting ensemble output, is used to select donor and acceptor site predictions. In this example 11 out of the 12 splice sites are predicted correctly at the cost of five false positive predictions. The donor site at position 584 is missed because it differs considerably from the consensus sequence, and because there is no clear transition between a coding and a non-coding region. This site is not recognized by the rule based system either, because there is another nearby candidate, with a strong splice site prediction in the vicinity of a transition region at position 553. At positions 1425 and 1429 two false acceptor site predictions are removed by the refinement, and at position 848 a donor site prediction is enhanced by the rule based system.

performance. If a network is deprived of resources (weights), the primary features will stand out more clearly.

Networks with an input window of 201 nt and at least six hidden units all had correlation coefficients >0.70. For simplicity we present a weight analysis of the smallest six unit network here only.

The six weights connecting the hidden units and the output unit had about the same numerical size. Five of the weights were positive, while one was negative, meaning that this unit will be pro-intron when activated. The weights were approximately one order of magnitude larger than the thresholds of the hidden units. The thresholds of the hidden units had almost been nullified by the training and were of no numerical importance for the function.

For the 201 × 6 input-to-hidden weights we found that three of the six hidden units were involved in checking the triplet reading frame. A weight logo (22) of the combined weights to these three hidden units can be seen in Figure 8, positive weights in the upper part and negative weights in the lower.

In Table 2 the nucleotide frequencies at each codon position can be seen. Position 100 in Figure 8a, for instance, corresponds to position one in the codons. The large G corresponds to the large guanine fraction of 0.34 at codon position one. The large negative T in Figure 8b corresponds to the low thymine fraction of 0.18 at codon position one. This is a general tendency: for each of the four nucleotides at each position in the codons, their size and sign relative to their size and sign at the two neighboring positions more or less mirror their frequencies in Table 2. If we again take position 100 as reference, this position corresponds to position one in the reading-frame in one of the hidden units, to position two in the reading-frame in another hidden unit, and to position three in the reading-frame in a third hidden unit. We expect that one of the three units will be active for a given input window where the central nucleotide belongs to an exon. Plots of the activities of the three reading frames checking hidden units on presentation of all windows in the test set (Figs 10 and 11) confirmed this. It also appears that the units are mostly inactive when inside an intron.

The three other hidden units are engaged in the recognition of intron-and exon-like DNA, their weight logos can be seen in Figure 9. One unit will be activated by a high adenine and thymine content combined with a low cytosine and guanine content in the left side of the input window. In other words, when the nucleotide frequencies correspond to those found in *A.thaliana* introns. This means that inside an intron, this hidden unit will be active and due to its hidden-to-output weight being negative it will suppress the output activity of the network. This suppression will level off when the input window enters a coding region.

Two units are activated by a high cytosine and guanine content combined with a low adenine and thymine content in the right side of the input window. These units will be deactivated by an intron-like nucleotide composition, while activated inside a coding region, note that the added weights from these two units are shown in Figure 9. Together with other features they recognize the GC content. One of the units gives the most accurate prediction of the coding to non-coding border at an acceptor site due to its weights being largest in the left part of the window, while the other gives a more accurate prediction at the donor site because the weights are largest in the right part of the window (data not shown). The pruning technique 'optimal brain damage' (23), which discards unessential weights, has previously been used with great success on networks trained on human genes (22).

## Combining local and global sequence information

From the weight analysis we know that the local and global networks exploit the nucleotide pattern of the reading frame, the transition between coding and non-coding DNA and the consensus-like sequence of the splice sites quite differently. The combined approach used here proceeds in two steps: a prediction step and a refinement step. The first step is equivalent to earlier work on human genes (11). The second step is based on rules found by investigating the mistakes of the first step.

In Figure 12 a typical output of the coding/non-coding network ensemble can be seen for the test sequence ATRAH1GNA. Several interesting features can be observed. The exon covering positions 979–1027 is predicted nicely by the ensemble. Strong splice site predictions are also found at the border of this region. In this unproblematic case, the splice sites could be determined either from the global prediction or from the local prediction alone. However, a conflicting situation is found further downstream. At this point the global network ensemble incorrectly predicts a coding region from position 1250 to 1350, but no corresponding splice sites are found by the local network ensembles in this region. The combined system should not predict splice sites here. At position 1134 a weak acceptor site is predicted and it is indeed at the start of a region predicted as being coding. This splice site should be predicted, while the donor site predicted at position 1135 is very strong, but in conflict with the prediction from the global network ensemble, and it should therefore not be predicted by the combination. How do we combine the predictions made by the global and local systems?

The combination is made by letting the signal from the coding/non-coding network ensemble control the threshold of the local splice site assigning network ensembles. In regions with abruptly decreasing output activity of the coding/non-coding networks, donor sites should be enhanced and acceptor sites should be suppressed. On the other hand regions with abruptly increasing output activity should enhance acceptor site assignment and suppress donor site assignment. Regions with a more or less uniform activity should demand a high confidence level to suppress false positives.

To obtain this we calculate an approximation to the first derivative of the output activity of the coding/non-coding network ensemble, Δ. This is done by summing up *n* output values to the left of the potential splice site and *n* values to the right of the splice site. The left side sum is subtracted from the right sum and the result is divided by the number of addends. For each output this gives a value between –0.5 and 0.5. The output of the local network ensemble, ($O_{\text{local}}$), is interpreted by using the following simple formula

$$O_{local} > \mathrm{a}\Delta + \mathrm{t} \qquad \mathbf{5}$$

Δ is the value calculated from the coding/non-coding network ensemble, a and t are constants. This means that if $O_{\text{local}}$ is greater than aΔ + t, the output should lead to the assignment of a splice site, otherwise a non-splice site.

We found the optimal values of a and t for all sensitivity levels. The maximal correlation coefficient for donor site prediction is 0.86 at a sensitivity level of 84%, while the best correlation coefficient for acceptor site prediction is 0.76 at 74% sensitivity (Figs 6 and 7).

## Post-prediction rule based filtering

The rule based filtering of splice sites is performed on the basis of predictions from the combined networks as described above. The combination can give predictions at different sensitivity levels, and it is not clear *a priori* what sensitivity level to choose for the refinement. Loosely speaking we want to extract the maximal information from the prediction data. A quantitative measure of gained information can actually be defined for cases like this (24), and an analysis showed that this measure did peak close to the 70% sensitivity level. This is also where the splice site correlation coefficients C(D) and C(A) peak, similar to the result obtained earlier on human genes (11).

We have designed a number of post-processing steps in order to (i) discard wrong splice site predictions, (ii) choose between two or more nearby equally strong predictions, and (iii) to enhance weak (or missing) predictions which must be preferred when viewing the prediction non-locally. Each step can be associated with a biological mechanism previously suggested in the splicing literature. The mechanisms may or may not be active in the cell, but their computational efficiency may be used as indications in this direction.

## Discarding splice sites in uniformly predicted regions

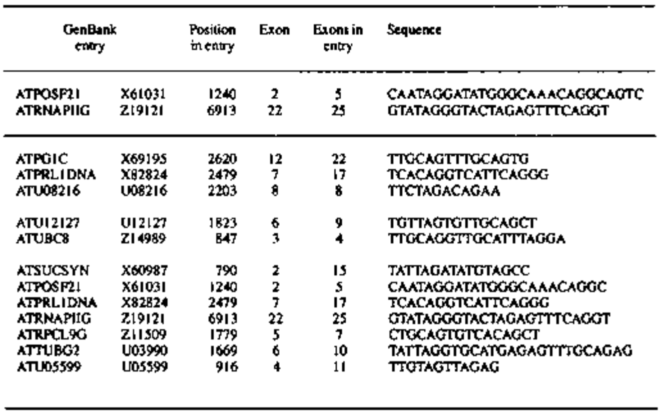
A fair amount of false sites are located in the middle of uniformly predicted strong coding, or non-coding regions. Consequently, splice sites may safely be discarded in these regions. In the cell nucleus, one can speculate that these ‘perfect’ sites are hidden due to the secondary structure of the pre-mRNA, and thus not available for incorporation in the spliceosome.

We then need to know when we are in a uniformly predicted region. The derivative of the coding prediction is not a good measure of strong uniform prediction as it can average to zero in regions of oscillating prediction as well. To estimate how flat a coding prediction around a true splice site can be in general, we extracted all splice sites from the training set with a flank of 1 nt to each side, with 2 nt to each side and so on up to 45 nt. For each flanking length and each splice site the maximal and minimal coding prediction values were found. If a splice site is found in a uniformly low coding region, the maximal coding prediction in this region will be close to zero. The minimal value of all the maximal values found for all splice sites with a given flanking length represent an upper bound for the flatness of the surroundings of a splice site with a low coding prediction. Likewise, an upper bound for the flatness of the surroundings of a splice site with a high coding prediction can be found. A table of flanking length and max/min values can immediately be used to filter out false splice sites without removing true splice sites from the training set (data not shown). To allow for values beyond those in the training set, a 20% margin was added, respectively subtracted, from the max/min values. These values show a close-to linear progression, and therefore in practice we used a linear approximation to the max/min curve. 610 out of 8818 potential donor sites from the test set with a non-zero score could be removed, and 819 out of 5708 potential acceptor sites without removing any true sites.

## Scanning procedure for acceptor site pairs in T-tract prolongation in 5′ exon ends

Figure 7 shows the result of the acceptor splice site network and of the combination. The detection of unambiguous acceptor sites is generally harder than the prediction of donor sites. To investigate this phenomenon further we have checked the false positives that arise when we have a recognition of 25%, 36% and 55% true positives. In this low sensitivity region the corresponding number of false splice sites in the entire test set is 3, 8 and 20, respectively. The majority of these false splice sites are found between 2 and 20 nt downstream from the correct splice site into the exon. These sites are characterized by having a strong consensus and by an elevated adenine and thymine content between the true splice site and the false. Moreover, the coding/non-coding network often shifts from intron to exon closer to the false splice site (data not shown). At 25% recognition of true acceptor sites 2 out of 3, at 36% 7 out of 8, and at 55% 14 out of 20 false positives were of this kind. Table 3 shows the 14 false splice sites and the sequence from the true acceptor site to the false one. The above mentioned false acceptor sites in the T-tract prolongation are consistent with an experimental observa-

**Table 3.** The false acceptor sites detected downstream from the true acceptor splice sites at a true site recognition level of 55% on the test set



| GenBank entry | | Position in entry | Exon | Exons in entry | Sequence |
|---|---|---|---|---|---|
| ATPOSF21 | X61031 | 1240 | 2 | 5 | CAATAGGATATGGGCAAACAGGCAGTC |
| ATRNAPIIG | Z19121 | 6913 | 22 | 25 | GTATAGGGTACTAGAGTTTCAGGT |
| ATPGIC | X69195 | 2620 | 12 | 22 | TTGCAGTTTGCAGTG |
| ATPRL1DNA | X82824 | 2479 | 7 | 17 | TCACAGGTCATTCAGGG |
| ATU08216 | U08216 | 2203 | 8 | 8 | TTCTAGACAGAA |
| ATU12127 | U12127 | 1823 | 6 | 9 | TGTTAGTGTTGCAGCT |
| ATUBC8 | Z14989 | 847 | 3 | 4 | TTGCAGGTTGCATTTAGGA |
| ATSUCSYN | X60987 | 790 | 2 | 15 | TATTAGATATGTAGCC |
| ATPOSF21 | X61031 | 1240 | 2 | 5 | CAATAGGATATGGGCAAACAGGC |
| ATPRL1DNA | X82824 | 2479 | 7 | 17 | TCACAGGTCATTCAGGG |
| ATRNAPIIG | Z19121 | 6913 | 22 | 25 | GTATAGGGTACTAGAGTTTCAGGT |
| ATRPCL9G | Z11509 | 1779 | 5 | 7 | CTGCAGTGTCACAGCT |
| ATTUBG2 | U03990 | 1669 | 6 | 10 | TATTAGGTGCATGAGAGTTTGCAGAG |
| ATU05599 | U05599 | 916 | 4 | 11 | TTGTAGTTAGAG |

The table is subdivided into three parts. The top two sites are already present with a recognition level of 25% true acceptor sites, the top seven are present with a recognition level of 36% and all 14 are present with a recognition level of 55%. For each false site the sequence is shown starting 6 nt upstream from the true splice site and ending 2 nt downstream from the false splice site. The third column from the left gives the position downstream of the G in the AG dinucleotide in the false acceptor site. The high A+T content in the exon sequence between the true and false acceptor sites can be clearly seen in some of the entries.

tion made by Lou *et al.* (9,21). In dicot plant nuclei, when the true acceptor site is eliminated, cryptic acceptor sites located downstream are preferentially selected over cryptic sites located upstream (in the intron).

The observation that a correctly predicted acceptor site is often followed by a weaker falsely predicted acceptor site, suggests a potential method for discarding false predictions. By identifying all instances of double predictions, where the leftmost prediction was strongest, and by removing all predictions to the right up to a distance of 20 nt, a further 632 out of the remaining 4889 potential acceptor sites could be discarded at the cost of two true sites from the sequence ATU08315. An investigation of the two sites in ATU08315 showed that they were highly suspicious (see Materials and Methods).

### Selection between nearby donor site predictions

Inspection of the donor predictions made it clear that the most strongly predicted donor site in a pair is normally the true donor site. We removed all weaker donor site predictions within 15 nt from each strongly predicted donor site thereby reducing the number of donor predictions by 5413 from 8208 to 2795. Two true sites from ATSUCSYN and one from ATPGIC were lost by this approach. The ATSUCSYN sequence was later discarded due to a wrongly annotated exon (see Materials and Methods).

### A model for scoring intron-exon pairs—coupling between splice sites

Experiments (10) indicate that close cooperation exists between donor and acceptor sites, and that such cooperation influences their mutual selection. A vertebrate acceptor site was spliced by a dicot plant splicing system, when a plant donor site was present. This was not the case when the plant donor site was substituted by a vertebrate donor site. This indicates that the splicing mechanisms utilize information beyond what is present in the local context of the splice sites and that splicing will not function properly without the availability of this information.

In non-alternatively spliced genes perfect predictions must obey a number of constraints, for example that donor and acceptor sites must come in alternating order to be correct. If a donor site prediction is followed not by an acceptor site prediction, but by yet another donor site prediction, something is definitely wrong. Either we missed a site or one of the donor sites is wrong and must be discarded. To detect a missing donor (acceptor) site between two acceptor (donor) sites, all potential donor (acceptor) sites must be evaluated and compared. This means that we would like to quantify the likelihood of consistently spliced pairs of exons and introns. We therefore assign scores to D-intron-A-exon-D and A-exon-D-intron-A objects. The score for each potential ‘middle’ splice site is obtained by multiplying a number of factors (including added combinations of them): the local prediction strength and confidence, scores from the exon/intron length distributions, the distance from the steepest transition in the coding/non-coding output, and maximal and minimal coding output in the ‘exon’ and ‘intron’ sequence surrounding the potential site.

The score $S(\mathrm{D}) = S_{\text{‘exon’}} \times S_{\text{‘intron’}}$ for the ‘middle’ donor splice site is obtained by computing $S_{\text{‘exon’}}$ and $S_{\text{‘intron’}}$ separately,

$$S_{\text{‘exon’}} = S_{\mathrm{local}} S_{\mathrm{elength}} S_{\mathrm{c-max}} \qquad \mathbf{6}$$

where $S_{\mathrm{local}}$ is a score quantifying the strength of the donor and acceptor sites, $S_{\mathrm{elength}}$ is a score derived from the exon length distribution and $S_{\mathrm{c-max}}$ is the maximal coding prediction found by the coding/non-coding network in the exon. These three factors are described in detail below. The intron score $S_{\text{‘intron’}}$ is computed similarly.

After all the $S(\mathrm{D})$ scores have been obtained for the A-exon-D-intron-A objects the best donor site is reported as a final prediction provided $S(\mathrm{D})$ exceeds a threshold of 0.3. If the best $S(\mathrm{D})$ value is below the threshold, one of the surrounding splice sites must be removed. We remove a site if its partner is <50 nt away, and at least 20% stronger in local network output. If we cannot find the missing site nor remove one of the splice sites, we cannot improve the prediction, but must leave things as they are. In this way the system is very conservative and produces very few errors. The acceptor sites are treated similarly.

### Scoring donor and acceptor sites

The strength of the donor and acceptor sites $S_{\mathrm{local}}$ is calculated from the confidence of the splice site predictions $S_{\mathrm{cnf}}$, the local network output $O_{\mathrm{local}}$, and the Δ value from the coding/non-coding prediction

$$S_{\mathrm{local}} = \frac{2S_{\mathrm{cnf}} + O_{\mathrm{local}}}{3} \frac{1}{1 + e^{-20\Delta}} \qquad \mathbf{7}$$

Here the confidence $S_{\mathrm{cnf}}$ of a site is equal to the specificity of the lowest sensitivity level that would accept the splice site. The specificity $Sp$ is defined as

$$Sp = \frac{P}{P + P^{\mathrm{f}}} \qquad \mathbf{8}$$

where $P$ is the number of correctly predicted splice sites (true positives), and $P + P^{\mathrm{f}}$ is the total number of splice site predictions. The sensitivity levels and the corresponding specificities have been determined on the test set. If the local network output for a site is zero, its confidence is also zero. If the site is predicted by the network combination at a sensitivity level close to 100% only, its confidence was empirically found to be ~0.5. Below a sensitivity level of 50%, the confidence was close to 1.0. Empirically we weight the confidence value and the local network output $O_{\text{local}}$ in the relation 1:2. The site closest to the largest change in coding value is usually preferred over its competing neighbors. This observation is implemented by the non-linear squashing function $1/(1 + e^{-20\Delta})$.

## Exon/intron length distributions

From the exon length distribution we observed that practically no exons are <20 or >3000 nt in length. Likewise, the intron lengths are found between 55 and 1500 nt. If the distance between a predicted donor and its neighboring acceptor site falls outside this range, one of the predictions is probably wrong, or a site in between has been missed. We can also discriminate between lengths inside these constraints, some being more likely than others. Exons with a length <45 nt are rare, and there are fewer long exons than short exons. Most introns have a length between 65 and 100 nt. From these observations exon $S_{\text{elength}}$ and intron length scores $S_{\text{ilength}}$ can be calculated, estimating how much we believe in a proposed length. Using simply the raw log-normal distribution of exon lengths, will not work because even though internal exons at length 300 nt are relatively rare, a downscaling to this probability level will be quite harmful in single cases. If one was supposed to make predictions for a large set of test genes, this general distribution could be used to regulate the level of larger exons. Instead, we use a piecewise linear candidate exon length score which is increasing from 0.0 to 0.98 for lengths between 0 and 20 nt, increasing to 1.0 at 45 nt and then decreasing to 0.98 at 3000 nt where it drops to zero. Likewise, the candidate intron length score was 0.0 for lengths <55 nt, increased linearly to 1.0 for lengths up to 65 nt, being 1.0 between 65 and 100 nt, and decreasing linearly to 0.97 between 100 and 1500 nt where it drops to zero.

## Maximal and minimal coding prediction in exons and introns

A final factor found to be of relevance is the maximal coding prediction for exons $S_{\text{c–max}}$ and the minimal coding prediction for introns. The predictions of the coding/non-coding network should at least once come close to 1.0 in a potential exon, othewise it is probably not a correct exon, at least not a coding exon. Likewise, we must assume introns to have at least one very low coding prediction to accept them. As the coding prediction for very short exons is known to be weak, a special correction for exons <30 nt was applied, where a value of 0.5 was added to the maximal exon prediction $S_{\text{c–max}}$. For introns $1 - S_{\text{c–min}}$ is used as factor instead of $S_{\text{c–max}}$.

## NetPlantGene performance

The final performance of our method, NetPlantGene, on the test set is shown in Figures 6 and 7. They show the false positives plotted against the number of true positives for donor and acceptor sites, respectively. The maximal correlation coefficient for donor sites was reached with a recognition level of 88.4% true positives and 0.02% false positives. The correlation coefficient $C$(D) was 0.90 with $a_{\text{donor}} = 0.45$ and $t_{\text{donor}} = 1.02$. The approach was able to detect 57.3% or more than half of the true donor sites without any false sites. When detecting 95% of the true donor sites, the combined approach makes 0.097% false donor site assignments.

The maximal correlation coefficient for acceptor sites was reached with a recognition level of 80.2% true positives and 0.034% false positives. The correlation coefficient $C$(A) is 0.83 with $a_{\text{acceptor}} = -1.75$ and $t_{\text{acceptor}} = 1.04$. When detecting 95% of the true acceptor sites, the combined approach makes 0.26% false acceptor site assignments. 21.1% of the true acceptor sites in the test set could be predicted without any false predictions.

## Comparison with GeneMark, GeneFinder and Grail

The prediction quality of the coding/non-coding network in the present study was compared with the prediction quality of GeneMark (25) (used with its *A.thaliana* matrices). As mentioned above, the overall performance of the coding/non-coding network ensemble on the test set is 0.76 in terms of the correlation coefficient. The overall performance of GeneMark reaches 0.55 only. The reason may be that the inhomogeneous Markov models of order 4 used by the program have problems in dealing with the often weak and irregular reading frame in *A.thaliana* genes. To investigate the prediction quality on protein-coding exons of different length, a set of 'partial' correlation coefficients was calculated for each method. The data used for the calculation of a partial correlation coefficient in a given length interval is all test set non-coding material and all protein-coding exons with lengths in that given interval. (This definition produces correlation coefficients which are generally lower than the overall performance correlation coefficients, so the values should not be regarded as an additional measure of the absolute quality of the prediction technique, but only as a fair means of comparison.) In every interval our coding/non-coding network is superior to GeneMark in prediction quality. While the former reaches a sustained performance on all exon lengths, the latter actually approaches a negative correlation coefficient for short exons, rendering GeneMark useless for exons <50 nt.

The prediction quality of the splice site assignment by our combined method, NetPlantGene, was also compared with that of an *A.thaliana* version of GeneFinder (26). We recalculated the weight matrices used by GeneFinder on our training set to give a fair comparison between the two methods. When assigning the same number of true splice sites GeneFinder assigns nearly an order of magnitude more false splice sites than NetPlantGene. At a recognition level of 90% true splice sites NetPlantGene assigns 24 false donor sites and 90 false acceptor sites. GeneFinder assigns 506 false donor sites and 812 false acceptor sites at the same level. The detailed comparison for all levels can be seen in Figures 6 and 7. The performance of the GeneFinder donor prediction is very similar to the performance of the local neural networks. The local neural network performance is better for high sensitivity levels and worse for low sensitivity levels. This is because we have pushed the networks to perform well at the high sensitivity levels at the cost of a slightly inferior performance at the low sensitivity levels by using the stopping criterion for the training. The GeneFinder performance on acceptor site prediction is significantly lower even when compared with the local neural networks. We think this is a result of the sequence window length used by GeneFinder. GeneFinder uses an asymmetric window of 31 nt (5 exon and 26 intron nt). This window size is significantly smaller than the window size of 61 nt found to be optimal for neural network acceptor site prediction. We believe that the performance of GeneFinder could be improved to the level of the local neural networks by changing the window size to 61 nt and recalculating the weight matrix.

Xgrail predicts exons, we have compared the quality of the exon/intron and intron/exon border prediction with our method (Figs 6 and 7). Xgrail predicts acceptor sites at a sensitivity level of 54% and produces 0.14% false positives. NetPlantGene comes up with 0.01% false positives at this sensitivity level, more than an order of magnitude improvement. The donor prediction of Xgrail has a sensitivity of 47% and produces 0.16% false positives. NetPlantGene did not come up with any false positive predictions at this sensitivity level. We conclude from this that the splice site prediction of NetPlantGene is significantly more accurate than both Xgrail and GeneFinder. As information on the training set used for constructing Xgrail is not available (Uberbacher, personal communication) the performance reported here for Xgrail must be viewed as an upper limit. We cannot exclude that several of the test set sequences used in this study were used for training the Xgrail method (Grail 2, version 1.3b).

## NetPlantGene and alternative splicing

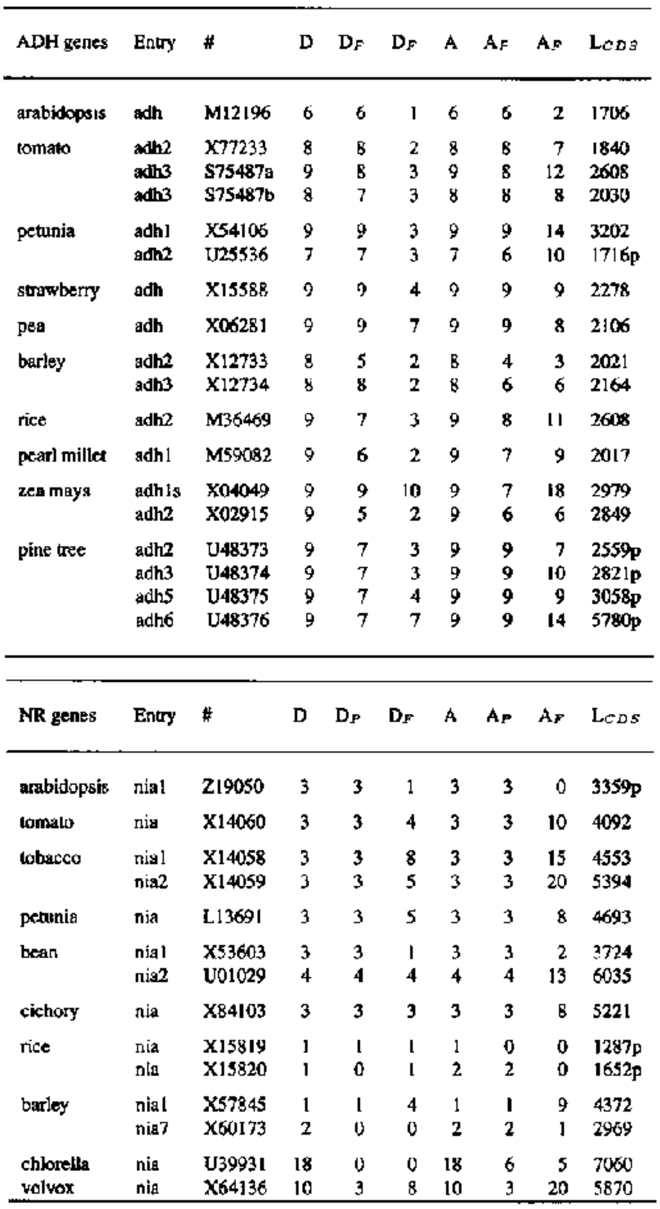
Although splicing efficiency appears to be low in plants, alternative splicing is rarely observed, compared with metazoa (2). NetPlantGene predictions for known cases of alternatively spliced coding sequences from dicots were investigated. In *A.thaliana* itself, the RuBisCO activase gene (M86720) contains six introns, the last one having two alternative acceptor sites (27). All sites are predicted by the method, apart from the first alternative acceptor site of intron 6. The first alternative, 11 nt upstream from the second, did indeed have a high network score but was later discarded by the rules.

The HprA gene from *Cucumis sativus* (X58542) contains 12 introns, the last one showing an alternative choice between two donor sites separated by 35 nt (28). NetPlantGene predicts each of the 24 sites with high score (>0.85) including the second alternative position from intron 12, which is the one utilized in most species. The first site is also predicted, but with a lower score (0.74) together with one false positive donor in the CDS.

The GdcsH gene from *Flaveria trinervia* is spliced by three introns, the first one having two alternative acceptor sites, separated by 6 nt (29). All introns are predicted, but only the second alternative acceptor site of the first intron is. Lastly, three genes coding for glycine-rich RNA-binding proteins from tobacco (D16204–D16206) contain one intron each in the CDS, with alternative donor sites for all of them (30). The upstream donor was predicted for one gene (D16205), while none of the other donors were, whatever the gene.

Alternative splicing was also reported in the gene encoding the large subunit of RNA polymerase II from *A.thaliana* and soybean. The intron is situated outside the CDS, in the 3′ trailer sequence. Both sites for this alternate intron were predicted in the soybean gene, while none of them were in *A.thaliana*. This tells us that at least some of the predictions made in the 3′ non-coding sequence are relevant, albeit NetPlantGene was trained for predictions inside the coding sequence. Interestingly, the two entries for the *A.thaliana* gene diverged from one another by several frameshifts and gaps in the sequence, and moreover by positioning of the introns, one of them is even missing in one entry, and was excluded from the data set for these reasons (12). In this specific case we find that NPG helps in searching for the correct features of this gene, locating the missing exon, and identifying the likely borders of two others with divergent locations.

**Table 4.** NetPlantGene predictions in other species, monocots, dicots, gymnosperms and algae, for two gene families: adh coding for alcohol dehydrogenase (ADH), and nia coding for nitrate reductase (NR)



| ADH genes | Entry | # | D | $D_P$ | $D_F$ | A | $A_P$ | $A_F$ | $L_{CDS}$ |
|---|---|---|---|---|---|---|---|---|---|
| arabidopsis | adh | M12196 | 6 | 6 | 1 | 6 | 6 | 2 | 1706 |
| tomato | adh2 | X77233 | 8 | 8 | 2 | 8 | 8 | 7 | 1840 |
| | adh3 | S75487a | 9 | 8 | 3 | 9 | 8 | 12 | 2608 |
| | adh3 | S75487b | 8 | 7 | 3 | 8 | 8 | 8 | 2030 |
| petunia | adh1 | X54106 | 9 | 9 | 3 | 9 | 9 | 14 | 3202 |
| | adh2 | U25536 | 7 | 7 | 3 | 7 | 6 | 10 | 1716p |
| strawberry | adh | X15588 | 9 | 9 | 4 | 9 | 9 | 9 | 2278 |
| pea | adh | X06281 | 9 | 9 | 7 | 9 | 9 | 8 | 2106 |
| barley | adh2 | X12733 | 8 | 5 | 2 | 8 | 4 | 3 | 2021 |
| | adh3 | X12734 | 8 | 8 | 2 | 8 | 6 | 6 | 2164 |
| rice | adh2 | M36469 | 9 | 7 | 3 | 9 | 8 | 11 | 2608 |
| pearl millet | adh1 | M59082 | 9 | 6 | 2 | 9 | 7 | 9 | 2017 |
| zea mays | adh1s | X04049 | 9 | 9 | 10 | 9 | 7 | 18 | 2979 |
| | adh2 | X02915 | 9 | 5 | 2 | 9 | 6 | 6 | 2849 |
| pine tree | adh2 | U48373 | 9 | 7 | 3 | 9 | 9 | 7 | 2559p |
| | adh3 | U48374 | 9 | 7 | 3 | 9 | 9 | 10 | 2821p |
| | adh5 | U48375 | 9 | 7 | 4 | 9 | 9 | 9 | 3058p |
| | adh6 | U48376 | 9 | 7 | 7 | 9 | 9 | 14 | 5780p |

| NR genes | Entry | # | D | $D_P$ | $D_F$ | A | $A_P$ | $A_F$ | $L_{CDS}$ |
|---|---|---|---|---|---|---|---|---|---|
| arabidopsis | nia1 | Z19050 | 3 | 3 | 1 | 3 | 3 | 0 | 3359p |
| tomato | nia | X14060 | 3 | 3 | 4 | 3 | 3 | 10 | 4092 |
| tobacco | nia1 | X14058 | 3 | 3 | 8 | 3 | 3 | 15 | 4553 |
| | nia2 | X14059 | 3 | 3 | 5 | 3 | 3 | 20 | 5394 |
| petunia | nia | L13691 | 3 | 3 | 5 | 3 | 3 | 8 | 4693 |
| bean | nia1 | X53603 | 3 | 3 | 1 | 3 | 3 | 2 | 3724 |
| | nia2 | U01029 | 4 | 4 | 4 | 4 | 4 | 13 | 6035 |
| cichory | nia | X84103 | 3 | 3 | 3 | 3 | 3 | 8 | 5221 |
| rice | nia | X15819 | 1 | 1 | 1 | 1 | 0 | 0 | 1287p |
| | nia | X15820 | 1 | 0 | 1 | 2 | 2 | 0 | 1652p |
| barley | nia1 | X57845 | 1 | 1 | 4 | 1 | 1 | 9 | 4372 |
| | nia7 | X60173 | 2 | 0 | 0 | 2 | 2 | 1 | 2969 |
| chlorella | nia | U39931 | 18 | 0 | 0 | 18 | 6 | 5 | 7060 |
| volvox | nia | X64136 | 10 | 3 | 8 | 10 | 3 | 20 | 5870 |

D means number of donor sites, $D_P$ the number of predicted sites, $D_F$ the number of false positives, and $L_{CDS}$ indicates the length of the CDS in the GenBank entry (p, partial).

## NetPlantGene performance in other plants

A preliminary test of the performance of NPG on various plant genes was done using two sets of genes, coding for the same proteins, respectively alcohol dehydrogenases (ADH) and nitrate reductases (NR). The results of this comparison (Table 4) shows that NPG keeps predicting nearly all of the splice sites in dicot genes, but works differently on monocot genes, predicting a

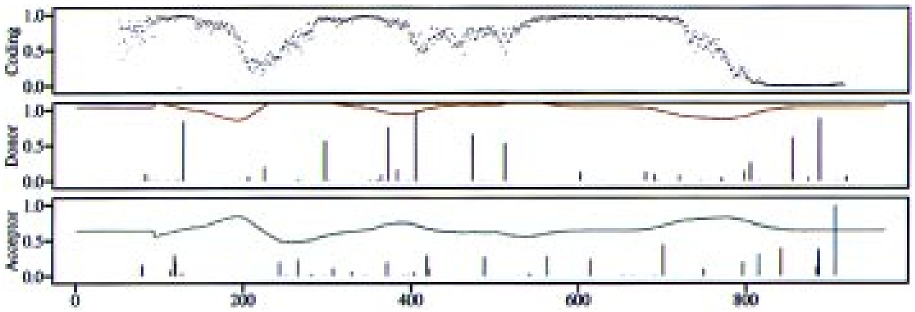


**Figure 13.** NetPlantGene splice site prediction for the cDNA from the *A.victoria* green fluorescent protein. A donor site is predicted at position 405 with a confidence of 0.94. The mutated version of the CDS does not result in any splice site predictions.

fraction of them only, always more than a half. The results on pine ADH genes is surprisingly good, owing to the phylogenetic distance with monocots. A dramatic fall in performance is observed with NR genes from green algae. Besides the varying capacity to recognize the true sites, according to phylogeny outside the dicots, NPG shows an increased level of false predictions compared with *A.thaliana*. The level of false predictions varies from species to species, even among dicots. One explanation for these variations is clearly the very different sizes of the plant genomes in the comparison. These observations would benefit from further investigations on a wider scale. As such, they fit with the observation of differences in splicing capacity between monocots and dicots, and point to the use of NPG as a way to anticipate how a gene from one species will be spliced when transferred into another plant species.

### Green fluorescent protein

The coding sequence for the green fluorescent protein from the jelly fish *Aequorea victoria* is used as a reporter gene in a number of organisms and experimental assays. If the gene is expressed, the organisms glow green. Expression of the gene in *A.thaliana* has proven unsuccessful because the gene is spliced at a cryptic splice site. A mutant has been made that is not spliced in *A.thaliana* (31). To test the performance of NetPlantGene, the cDNA encoding the green fluorescent protein and its mutated version were tested for potential splice sites. NetPlantGene correctly identifies the cryptic donor splice site in the wild type (Fig. 13). No site is predicted in the modified sequence.

## CONCLUSION

Neural networks have been trained to recognize splice sites in *A.thaliana* DNA. This task is not trivial for several reasons. First, the number of possible AG and GT dinucleotides are ~100 times larger than the number of true splice sites. Secondly, the fact that a portion of the AG and GT sites may have been active as splice sites once and therefore are very similar to real splice sites makes this a difficult task. Thirdly, it is not known for sure how much of the information needed for splicing is available directly in the DNA sequence and how much is contributed from other sources; e.g. the structure of the pre-mRNA and information contained in the spliceosome and other parts of the cell machinery. The last reason of course puts a potential theoretical upper limit on the possible quality of a prediction based on the nucleotides in the genomic DNA sequence alone.

To ensure a conservative estimate of the performance of the algorithm presented in this paper we have used a large part of the available data to test the performance (nearly 25% of the available splice sites have been used). Furthermore, we have made sure that the sequence similarity between the entries used for training the neural networks and the test sequences is low.

When comparing the results of the final algorithm with the results obtained using the local network only, it is clear that a lot is gained by combining the local and global networks. At 80% true donor site recognition the combination assigns 0.011% false positives only, while the local network alone assigns 0.20% false positives. At 95% recognition the numbers are 0.097% and 0.60%. For the acceptor site recognition the corresponding numbers are: at 80% recognition 0.034% and 0.20%; and at 95% recognition 0.26% and 0.56%. Furthermore, the combination was able to predict more than half of the true donor sites without false positives. Comparison with three other approaches, GeneMark, GeneFinder and Grail, showed that the method presented here has an order of magnitude fewer false sites at nearly all sensitivity levels.

One of the main criticisms of neural networks is their ‘black box’ status, meaning that one will gain no insight into the characteristics of the problem when using neural networks. In this study we have analyzed the inner workings of the trained local and global networks. This has led to the discovery of the main features of the algorithms used by the local and global networks. It is clear that the base distribution plays an important role when identifying transition region between coding and non-coding nucleotides in the global context. Especially, the elevated A and T content of the introns in *A.thaliana* is an important factor in identifying the transitions, but surprisingly a reading frame recognition scheme develops by itself through the training process.

The analysis of the coding/non-coding network together with the fact that the network combination has severe difficulties in identifying true acceptor sites without making false predictions as well, led us to the discovery that *A.thaliana* introns often have a prolongation of the T-tract ending in a cryptic acceptor splice site. This might explain why the splicing machinery prefers cryptic acceptor sites located downstream in the exon and not cryptic sites located upstream in the intron, when the true acceptor site is eliminated (9,21).

As global sequence information is essential for computational selection of splice sites at low levels of false positives, one may ask how global information influences spliceosome assembly in the cell nucleous? Inference of a too detailed model from this work would clearly be far too speculative, but it is interesting that the network, by training, develops detectors which correspond to experimentally observed features: the triplet reading frame and the AT-high to AT-low transition regions. Recently it has been shown that the reading frame in internal exons is scanned for potential stop codons (see ref. 32 for a review), and also that AT-richness plays a prominent functional role in the splicing of plant introns (4).

## ACKNOWLEDGEMENTS

This work was supported by a grant from the Danish National Research Foundation. P.R. is Research Director of INRA (Institut National de la Recherche Agronomique, France).

## REFERENCES

1 Carle-Urioste, J. C., Ko, C. H., Benito, M., and Walbot, V. (1994) *Plant Mol. Biol.* **26**, 1785–1795.
2 Filipowicz, W., Gniadkowski, M., Klahre, U., and Liu, H. (1994) chapter title: Pre-mRNA Splicing in Plants. In Lamond,A.I. (ed.) *Pre-mRNA Processing.* R.G. Landes Company, Austin, TX, USA pp. 65–77.
3 Robberson, B. L., Cote, G. J., and Berget, S. M. (1990) *Mol. Cell. Biol.* **10**, 84–94.
4 Goodall, G. J. and Filipowicz, W. (1989) *Cell* **58**, 473–483.
5 Csank, C., Taylor, F. M., and Martindale, D. W. (1990) *Nucleic Acids Res.* **18**, 5133–5141.
6 Goodall, G. J. and Filipowicz, W. (1990) *Plant Mol. Biol.* **14**, 727–733.
7 Wiebauer, K., Herrero, J., and Filipowicz, W. (1988) *Mol. Cell. Biol.* **8**, 2042–2051.
8 Keith, B. and Chua, N. H. (1986) *EMBO J.* **5**, 2419–2425.
9 Lou, H., McCullough, A. J., and Schuler, M. A. (1993) *Plant J.* **3**, 393–403.
10 Waigmann, E. and Bartha, A. (1992) *Nucleic Acids Res.* **20**, 75–81.
11 Brunak, S., Engelbrecht, J., and Knudsen, S. (1991) *J. Mol. Biol.* **220**, 49–65.
12 Korning, P. G., Hebsgaard, S. M., Rouze, P., and Brunak, S. (1996) *Nucleic Acids Res.* **24**, 316–320.
13 Hertz, J., Krogh, A., and Palmer, R. G. (1991) *Introduction to the Theory of Neural Computation,* Addison Wesley, Reading.
14 Mathews, B. W. (1975) *Biochim. Biophys. Acta* **405**, 442–451.
15 Schneider, T. D. and Stephens, R. M. (1990) *Nucleic Acids Res.* **18**, 6097–6100.
16 Shannon, C. E. (1948) *Bell System Tech. J.* **27**, 379–423/623–656.
17 Hamming, R. W. (1980) *Coding and Information Theory,* Prentice-Hall, Englewood Cliffs NJ.
18 White, O., Soderlund, C., Shanmugan, P., and Fields, C. (1992) *Plant Mol. Biol.* **19**, 1057–1064.
19 Umen, J. G. and Guthrie, C. (1995) *RNA* **1**, 869–885.
20 Smith, C. W., Chu, T. T., and Nadal-Ginard, B. (1993) *Mol. Cell. Biol.* **13**, 4939–52.
21 Lou, H., McCullough, A. J., and Schuler, M. A. (1993) *Mol. Cell. Biol.* **13**, 4485–4493.
22 Tolstrup, N. (1995) *Int. J. Neural Sys.* **6**, 31–42.
23 Cun, Y. L., Denker, J. S., and Solla, S. A. (1989) In *Advances in Neural Information Processing Systems II.* San Mateo Morgan Kaufmann. pp. 396–404.
24 Zhi-Wang (1994) *Nature Struct. Biol.* **1**, 144–145.
25 Borodovsky, M. and McIninch, J. D. (1993) *Computers Chem.* **17**, 123–133.
26 Green, P. (1995) Genefinder. Unpublished.
27 Werneke, J., Chatfield, J., and Ogren, W. (1989) *Plant Cell* **1**, 815–825.
28 Hayashi, M., Tsugeki, R., Kondo, M., Mori, H., and Nishimura, M. (1996) *Plant Mol. Biol.* **30**, 183–189.
29 Kopriva, S., Cossu, R., and Bauwe, H. (1995) *Plant J.* **8**, 435–441.
30 Hirose, T., Sugita, M., and Sugiura, M. (1993) *Nucleic Acids Res.* **21**, 3981–3987.
31 Haseloff, J. and Amos, B. (1995) *Trends Genet.* **11**, 328–329.
32 Maquat, L. E. (1995) *RNA* **1**, 453–465.
33 Trifonov, E. (1987) *J. Mol. Biol.* **194**, 643–652.