Prediction of human mRNA donor and acceptor sites from the DNA sequence

Søren Brunak, Jacob Engelbrecht and Steen Knudsen†
Department of Structural Properties of Materials
The Technical University of Denmark
DK-2800 Lyngby, Denmark

† Molecular Biology Computer Research Resource
Dana–Farber Cancer Institute
Harvard School of Public Health
44 Binney St., Boston, MA 02115, USA

Abstract

Artificial neural networks have been applied to the prediction of splice site location in human pre–mRNA. A joint prediction scheme where prediction of transition regions between introns and exons regulates a cutoff level for splice site assignment was able to predict splice site locations with confidence levels far better than previously reported in the literature. The problem of predicting donor and acceptor sites in human genes is hampered by the presence of numerous amounts of false positives — in the paper the distribution of these false splice sites is examined and linked to a possible scenario for the splicing mechanism *in vivo*. When the presented method detects 95% of the true donor and acceptor sites it makes less than 0.1% false donor site assignments and less than 0.4% false acceptor site assignments. For the large data set used in this study this means that on the average there are one and a half false donor sites per true donor site and six false acceptor sites per true acceptor site. With the joint assignment method more than a fifth of the true donor sites and around one fourth of the true acceptor sites could be detected without accompaniment of any false positive predictions. Highly confident splice sites could not be isolated with a widely used weight matrix method or by separate splice site networks. A complementary relation between the confidence levels of the coding/non–coding and the separate splice site networks was observed, with many weak splice sites having sharp transitions in the coding/non–coding signal and many stronger splice sites having more ill–defined transitions between coding and non–coding.

Subject category: Genes, under the sub–headings: expression, sequence and structure.
Keywords: Intron–splicing, human genes, exon selection, neural network, computer–prediction.

Journal of Molecular Biology, to appear.

1 Introduction

Splice site selection in vertebrate pre–mRNA's (see Green, 1986 for a review) involves small nuclear ribonucleoproteins (snRNP's) recognising sequences around the exon–intron borders; U1 snRNP's recognise 5' (donor) sites (Mount *et al.*, 1983; Zhuang & Weiner, 1986) and U2 and U5 snRNP's interact with the 3' end of the intron (Chabot *et al.*, 1985; Krainer & Maniatis, 1985; Tazi *et al.*, 1986) where a 3' consensus signal (acceptor site) and a branch point signal are located (Keller & Noon, 1984; Harris & Senapathy, 1990). Numerous experiments have addressed the question of necessary and sufficient sequence conditions for proper selection of splice sites. Deletion of large intron parts seems to have no effect on the selection (Wieringa *et al.*, 1984), whereas mutation or duplication of splice sites often alters splice site selection (Krainer *et al.*, 1990; Kuhne *et al.*, 1983). Mutation of the 3' splice site appears to have an effect only when the most conserved AG dinucleotide is involved (Reed & Maniatis, 1986). Mutations at the 5' site either activate a nearby cryptic site (Treisman *et al.*, 1983) or, if no suitable cryptic site can be located in close proximity, cause exon skipping by abolishing splicing at the 3' site of the upstream intron as well (Orkin, 1987). The latter observation of exon skipping seems to indicate some functional coordination in the recognition of the 3' and 5' sites that surround an exon. Truncation or elongation of an exon causes inhibition of splicing of the upstream intron by blocking splicing at the 3' site (Furdon & Kole, 1986; Furdon & Kole, 1988; Robberson *et al.*, 1990). Thus the recognition of 3' and 5' splice sites around internal exons may be a concerted process in which the exon length is an important factor. It has been demonstrated that position effects are involved in splice site selection (Nelson & Green, 1988). A model of exon recognition would imply recognition of initial and terminal exons (which lack the 3' and 5' site, respectively) by separate processes.

The apparent presence of consensus sequences at the 3' and 5' sites has prompted many efforts to predict splice site selection based on the primary DNA sequence (Ohshima & Gotoh, 1987). The consensus of the 5' site was found to be $\frac{C}{A}$AG/GT$\frac{G}{A}$AGT (Mount, 1982), whereas the 3' site and the branch point are weakly defined by a consensus sequence (Harris & Senapathy, 1990) and seem to be specified by an intervening polypyrimidine tract characterised by the strict absence of AG dinucleotides (Smith et al., 1989; Mengeritsky & Smith, 1989). Construction of a finite state automaton as a recogniser of 5' splice sites only resulted in 50-55% prediction on test genes (Kudo *et al.*, 1987). Scoring and ranking schemes of potential 3' and 5' sites based on nucleotide weight tables have been unable to locate the majority of exons in small sets of test genes (Shapiro & Senapathy, 1987; Staden, 1984a). If more information was included in the scoring scheme: open reading frames, free energy of base pairing between the snRNP and the mRNA, as well as coding statistics, the prediction of single introns did rise to about 80% (Nakata *et al.*, 1985). Simple neural network algorithms have been employed in the recognition of 3' and 5' sites as well as coding regions (Lapedes

et al., 1988). Predictions on test samples of true and false sites have been claimed to be correct to 91-95%, but the method was unable to correctly assign the exons of a test gene.

2 Materials and Methods

2.1 Characteristics of the data set

From GenBank Release 62.0 all annotated human sequences were extracted (5141 entries). A subset of these consisting of entries with IVS feature lines, and which did not contain the string "alternative splic" — in order to leave out alternative splice sites — contained 1824 sequences. From this set entries were discarded, 1) if they did not contain the string "complete cds" as well as the complete RNA transcript, 2) if they did contain one intron only, and 3) if they contained more than one gene, which left the 95 entries listed in Table 1. Criterion 2) ensured that all genes had internal exons, and 3) had the effect of removing two very long sequences with a large number of repeats. Statistics regarding base content for translated exon (E), intron (I), untranslated exon (M) and non–transcribed (N) parts of the sequences are given in Table 2. Partially coding exons were thus in part classified as (E) and in part as (M).

2.1.1 Exon length distribution

The data set contained 521 exons (translated and untranslated, categories E and M, respectively) with a total of 133372 nucleotides. The average length of these exons was thus approximately 256 bp. Their length distribution peaks in the interval 101-200 bp as shown in Figure 1.

The lengths of the internal and terminal exons were quite different, as can be seen in Figure 2. The average length of internal exons was 200 bp, whereas the average length of terminal exons was as large as 649 bp. This is in contrast to length distributions in genes of some vertebrates, insects, plants and fungi as reported in (Hawkins, 1988).

2.2 Data set division

The data set used for construction of prediction algorithms and for test contained the 95 genes listed in alphabetical order in Table 1. The total data set was divided into two parts, part I used for training of neural network algorithms included the first 65 entries, and part II used for testing the remaining 30 entries. The total set of 95 genes contained 449 donor sites (and 449 acceptor sites), with 331 donor/acceptor sites in part I, and 118 donor/acceptor sites in part II. The number of nucleotides in the first 65 entries was 385488, in the last 30 entries 190987.

The full data set was utilised for training and test in three different versions:

1) on the set of *complete* sequences.

GenBank Locus	Length	D	A	GenBank Locus	Length	D	A
HUMA1ATP	12222	4	4	HUMIL2	5737	3	3
HUMA1GLY2	4944	5	5	HUMIL2A	6684	3	3
HUMACCYBA	3657	5	5	HUMIL2B	5561	3	3
HUMACTGA	3583	5	5	HUMIL5	3230	3	3
HUMADAG	36741	11	11	HUMIL5A	3241	3	3
HUMAK1	12229	6	6	HUMINCP	3716	2	2
HUMALBGC	19002	14	14	HUMKAL2	6139	4	4
HUMALPHA	4556	10	10	HUMKEREP	5339	7	7
HUMALPI	5291	10	10	HUMLACTA	3310	3	3
HUMALPPD	4268	10	10	HUMLCATG	6901	5	5
HUMANFA	2710	2	2	HUMMETIF1	2076	2	2
HUMANT1	5768	3	3	HUMMHCP42	5141	9	9
HUMAPOAIT	2385	3	3	HUMMHCP51	3206	9	9
HUMAPOC2	4057	3	3	HUMMHCP52	3422	9	9
HUMAPOC2G	3840	3	3	HUMMIS	3100	4	4
HUMAPOCIA	5375	3	3	HUMMRP14A	4439	2	2
HUMAPOCIB	5507	3	3	HUMMRP8A	4195	2	2
HUMAPOCII	4340	3	3				
HUMAPOE4	5515	3	3	HUMMT2A	1922	2	2
HUMAPRTA	2956	4	4	HUMMYCC	8082	2	2
HUMCACY	3671	2	2	HUMMYCRT	3922	2	2
HUMCAPG	3734	4	4	HUMOPS	6953	4	4
HUMCFVII	12850	8	8	HUMOTNPI	1338	2	2
HUMCKMT	6896	8	8	HUMP45C17	8549	7	7
HUMCP21OH	4042	9	9	HUMPCNA	6340	5	5
HUMCP21OHC	4044	9	9	HUMPGAMMG	3771	2	2
HUMCS1	2301	4	4	HUMPIM1A	6113	5	5
HUMCSFGMA	3194	3	3	HUMPLAPL	1282	3	3
HUMCSPB	3393	4	4	HUMPNMTA	4174	2	2
HUMCYC1A	4622	6	6	HUMPOMC	8658	2	2
HUMCYCAA	3088	2	2	HUMPPPA	2775	3	3
HUMCYPIIE	14776	8	8	HUMPRCA	11725	7	7
HUMDES	8878	8	8	HUMPRPH1	4946	3	3
HUMEDHB17	4845	5	5	HUMPRPH2	4037	3	3
HUMEF1A	4695	7	7	HUMRPS14	5985	4	4
HUMERPA	3602	4	4	HUMRPS17A	4029	4	4
HUMFABP	5204	3	3	HUMSAA	3460	3	3
HUMG0S19A	2790	2	2	HUMSAACT	3778	6	6
HUMG0S19B	3023	2	2	HUMSPRO	5296	7	7
HUMGASTA	7739	2	2	HUMTFPB	13865	5	5
HUMGCB1	7604	10	10	HUMTHY1A	2806	2	2
HUMGG	6455	3	3	HUMTKRA	13500	6	6
HUMGHN	2657	4	4	HUMTNFA	3633	3	3
HUMGRP78	5470	7	7	HUMTPA	36594	13	13
HUMHMG14A	8882	5	5	HUMTUBAG	4087	3	3
HUMHSP90B	8210	11	11	HUMTUBBM	3284	3	3
HUMHST	6616	2	2	HUMUBILP	3583	3	3
HUMIL1B	7824	6	6	HUMVPNP	2500	2	2

Table 1: The 95 entries from GenBank Release 62.0 used in this study. After the name follows the length in nucleotides. The number of donor (D) and acceptor sites (A) is listed for each entry. The total number of donor and acceptor sites was 2×449. All entries contained one gene only. A single donor site (from HUMAPRTA) did not conform to the standard `GT` rule in that `C` was substituted for `T`. A blank line separates entries in part I from entries in part II defined in section 2.2.

Class	Bp count	%	A	C	G	T
E	81643	14.16	22.92	28.96	28.47	19.63
I	336684	58.40	23.95	24.58	25.66	25.79
M	51729	8.97	24.12	26.44	24.31	25.10
N	106404	18.45	24.30	25.54	26.24	23.91
S	576460	100.00	23.88	25.54	26.04	24.51

Table 2: The base distribution in the 95 entries listed in Table 1 given for translated exon (E), intron (I), untranslated exon (M) and non–transcribed (N) DNA separately. S denotes the complete sample. The 95 entries contained in all 576475 nucleotides of which 15 were unknown. Note that the nucleotides `C` and `G` are overabundant in the translated exon class, whereas `A` and `T` are underabundant.

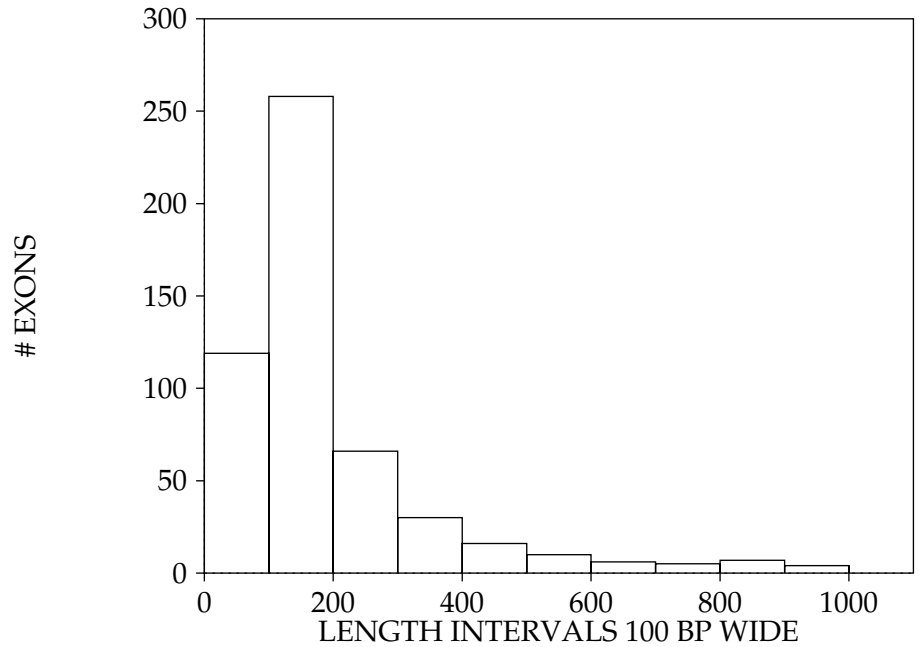

Figure 1: Distribution of lengths of the 521 exons in the total data set. Each channel has a length interval of 100 bp. The distribution peaks in the length interval 101-200 bp. The last interval 901-1000 includes all lengths $\geq$ 901 bp.

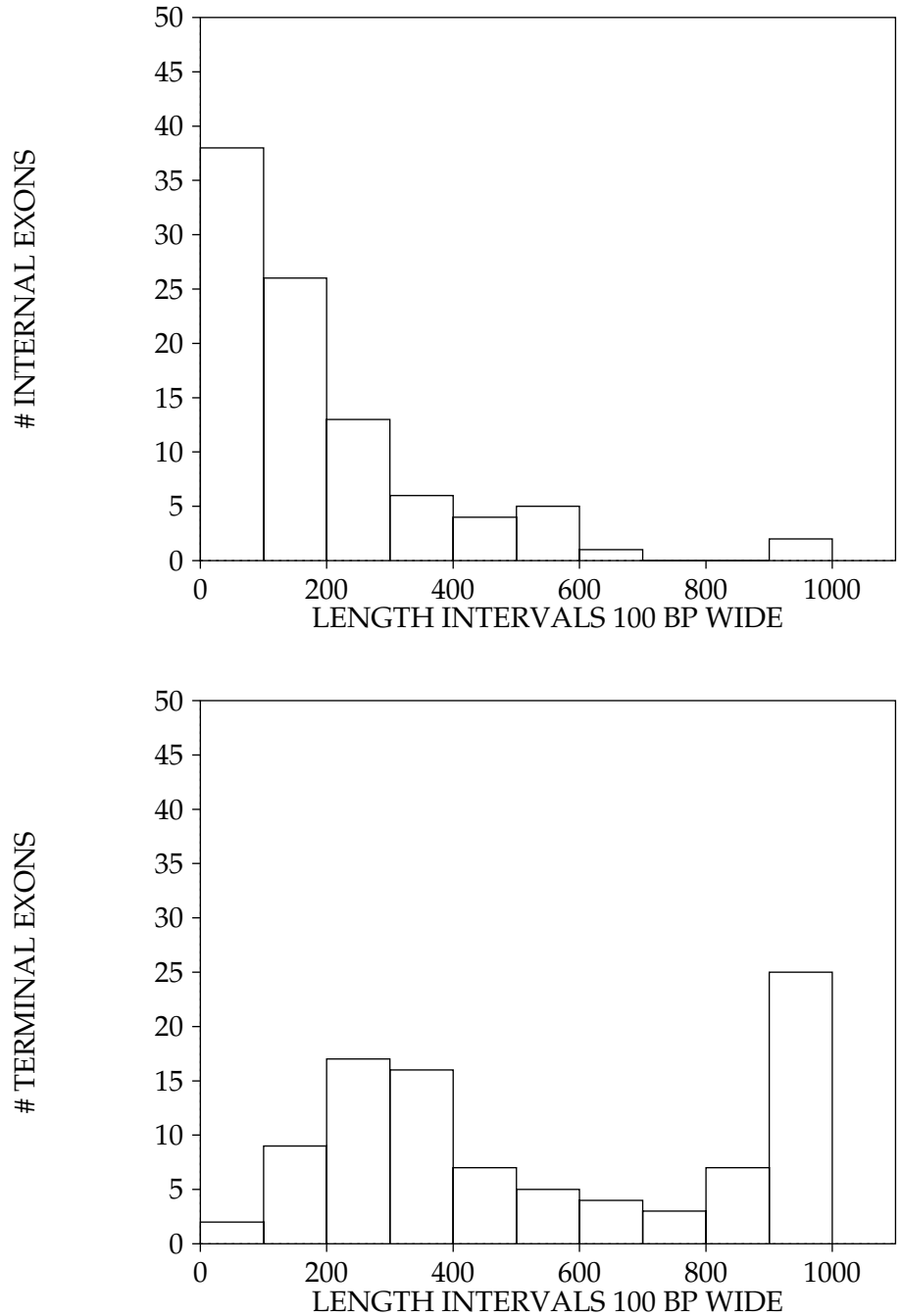

Figure 2: Distribution of lengths of the 95 internal (upper) and 95 terminal (lower) exons. Each channel has a length interval of 100 bp. Terminal exons are with a few exceptions always longer than 100 bp. The distribution for terminal exons peaks in the length interval 201-300 bp. The last interval 901-1000 includes all lengths $\geq$ 901 bp.

2) on a *reduced* data set of sequence segments extracted symmetrically around donor and acceptor sites with a range of 50 nucleotides. Each sequence segment contained 101 nucleotides, and part I of the reduced *donor* and *acceptor* site data sets thus contained 331 segments with a total of $331 \times 101 = 33431$ nucleotides, and part II correspondingly $118 \times 101 = 11918$ nucleotides.

3) on a set of *truncated*, but otherwise complete sequences. This data set version was created in order to take into account the fact that a number of GenBank entries may contain parts of adjacent genes. In the truncated data set version nucleotides outside a distance of 30 bp to the left and to the right of the coding region were removed from each entry. This removal mainly had the effect of reducing the initial part of the sequences, as most entries only include small regions beyond the 3' end of the mature pre–mRNA. The truncation lowered the number of nucleotides to 217436 bp in part I and to 109149 bp in part II, against 385488 and 190987 in the sets of complete sequences. The percentage of nucleotides in the E exon category did increase from 14.2% in the original data set to 18.9% in the truncated version.

2.2.1 Sequence similarity

The alphabetical order was maintained when dividing the data set into training and test in order to avoid strong similarities between genes in the two parts. The reduced donor and acceptor site data sets were searched for sequence similarities with the neural network metric method (Bohr *et al.*, 1988), measuring the similarity between each of the 118 sequence segments in part II and the 331 sequence segments in part I. A comparison between this measure and the measure of the Protein Identification Resource (PIR) (George *et al.*, 1986) can be found in (Bohr *et al.*, 1988; Petersen *et al.*, 1990). With the four–letter nucleotide alphabet, the overlap between any two random sequences is 25%. In the present case, donor site sequence segments shared the central dinucleotide `GT`, acceptor site sequence segments the dinucleotide `AG`, and accordingly the overlap between otherwise uncorrelated strings of length 101 nucleotides was close to 27% instead. We measured the $118 \times 331 = 39058$ overlaps between each of the 118 test set sequence segments in part II and each of the 331 sequence segments in the training set (part I), and only 13 of the 118 maximal overlaps between a donor site sequence segment in part II and part I exceeded 50%, 4 exceeded 75%, in the acceptor site case 10 exceeded 50%, and 3 exceeded 75%, from which we concluded that the alphabetically determined division in training and test was quite effective in avoiding unsatisfactorily large sequence similarities. Another factor which effectively reduced the level of sequence similarity was the fact that the data set contained complete coding sequences only. A large number of the GenBank entries represent small fractions of other entries containing complete sequences. The few large overlaps between segments in part I and II, together with average overlaps and standard deviations, are given in Table 3 (donor site sequence segments) and in Table 4 (acceptor

GenBank Locus (II)	Mean	σ	Max	GenBank Locus (I)
HUMMT2A	23.7	5.1	59.4	HUMMETIF1
HUMMT2A	30.9	6.0	76.2	HUMMETIF1
HUMPLAPL	30.1	7.2	82.2	HUMALPPD
HUMPLAPL	30.5	7.5	92.1	HUMALPHA
HUMPLAPL	28.8	7.9	97.0	HUMALPHA
HUMPPPA	30.0	5.8	50.5	HUMHMG14A
HUMRPS17A	28.0	5.4	52.5	HUMHMG14A
HUMSAACT	31.6	5.9	57.4	HUMACTGA
HUMSPRO	28.5	5.4	52.5	HUMHMG14A
HUMSPRO	29.3	5.9	57.4	HUMHMG14A
HUMTFPB	23.8	6.0	51.5	HUMALBGC
HUMTUBAG	27.3	5.0	51.5	HUMHMG14A
HUMVPNP	27.7	5.8	51.5	HUMHST

Table 3: Sequence similarity exceeding 50% among the 118 donor site sequence segments in the part II and the 331 donor site sequence segments in part I. Listed is the GenBank locus, the mean value for the 331 overlaps, the standard deviation σ, the largest overlap and the GenBank locus for the sequence segment with the largest overlap, for each of the 13 segments. Only 4 of the 118 maximal overlaps exceed 75%.

site sequence segments).

2.2.2 Dinucleotide propensities

If nucleotides in genes were distributed randomly, all pairs of nucleotides would occur with the same probability $\frac{1}{16}$, — *i.e.* in 6.25% of the cases. In Table 5 the actual distribution of `GT` and `AG` dinucleotides in the five functional categories is given for the total data set of 95 genes. `GT` dinucleotides were most abundant in introns, 5.18% against 5.05% on the average, while `AG` dinucleotides were most abundant in translated exons, 7.67%, and in introns, 7.65%, against 7.58% on the average. The total set of sequences contained 29090 `GT` dinucleotides and 43709 `AG` dinucleotides, whereas the test set contained 9620 `GT` and 14560 `AG` dinucleotides. These large numbers are to be compared with the small numbers of true donor/acceptor sites, 449 and 118, in part I+II and II, respectively.

A characteristic of the polypyrimidine tract upstream from the true acceptor sites was that the dinucleotides `TT`, `TC`, `CT`, and `CC`, gave different contributions, with high probabilities of `CT` and `CC` extending much further (in the interval fifteen to thirty nucleotides away from the acceptor site) than those of `TT` and `TC`.

We have computed the Kullback divergence for single nucleotides (Kullback, 1959), quantifying the "contrast" between the surroundings of `GT`/`AG` sites functioning as splice sites and those not functioning as splice sites in an interval of thirty nucleotides to each side of the sites. We found these to be in accordance with the findings of Lapedes *et al.*, 1988, with Kullback divergence in the acceptor site case residing mostly in twenty nucleotides at the 3' end of the introns (the polypyrimidine tract), and in the donor

GenBank Locus (II)	Mean	σ	Max	GenBank Locus (I)
HUMMT2A	28.4	5.6	71.3	HUMMETIF1
HUMMT2A	29.6	5.3	61.4	HUMMETIF1
HUMPLAPL	33.3	8.2	95.1	HUMALPHA
HUMPLAPL	30.5	7.3	100.0	HUMALPHA
HUMPLAPL	28.3	6.8	90.1	HUMALPHA
HUMPPPA	32.1	5.5	51.5	HUMMETIF1
HUMTKRA	31.2	6.0	50.5	HUMGCB1
HUMTKRA	30.2	5.1	50.5	HUMMETIF1
HUMTNFA	30.2	6.6	50.5	HUMERPA
HUMTPA	28.1	5.5	52.5	HUMFABP

Table 4: Sequence similarity exceeding 50% among the 118 acceptor site sequence segments in the part II and the 331 acceptor site sequence segments in part I. Listed is the GenBank locus, the mean value for the 331 overlaps, the standard deviation σ, the largest overlap and the GenBank locus for the sequence segment with the largest overlap, for each of the 10 segments. Only 3 of the 118 maximal overlaps exceed 75%.

Class	GT%	AG%
E	4.26	7.67
I	5.18	7.65
M	4.71	7.19
N	4.95	7.54
S	5.05	7.58

Table 5: The GT/AG dinucleotide distribution in the 95 entries listed in Table 1 given for translated exon (E), intron (I), untranslated exon (M) and non–transcribed (N) DNA separately. S denotes the complete sample.

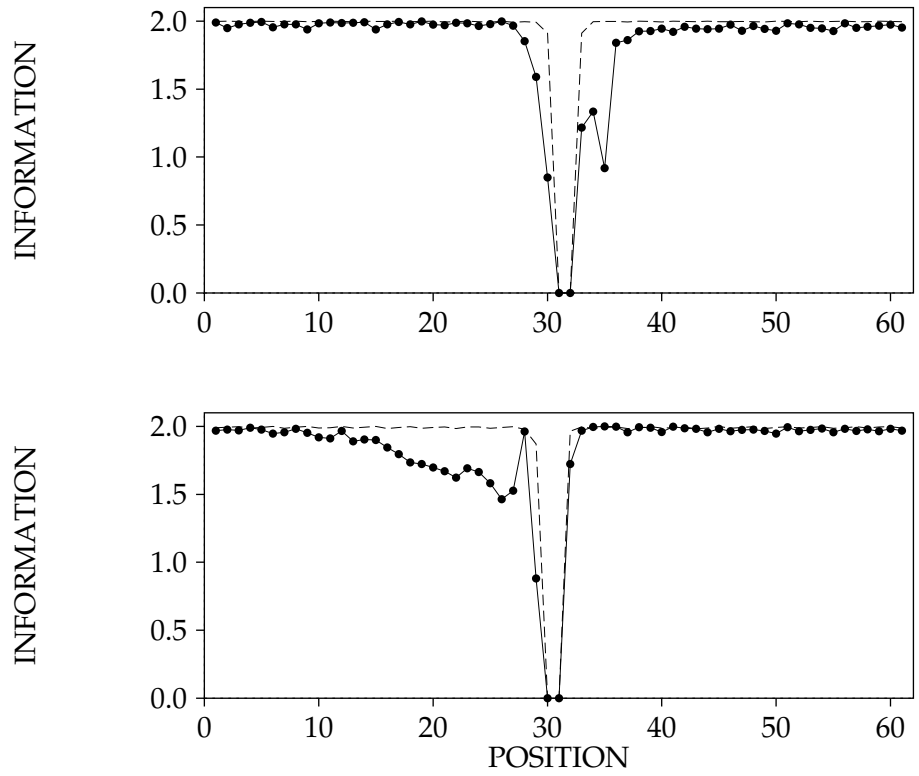

Figure 3: The Shannon information for `GT` sites (upper) and `AG` sites (lower), splice sites (•) and non–splice sites (-), in the 95 GenBank entries listed in Table 1. The central `G` in each site is surrounded by thirty bases in both directions. The unit of information is bits *per* nucleotide.

site case in six bases on the intron side and in three bases on the exon side. We also computed the Kullback divergence for all the sixteen dinucleotides and the sixty–four trinucleotides, and found it to be residing in roughly the same intervals. The variation of the Shannon information (Shannon, 1948) is shown in Figure 3 for both `GT` sites (donor/non–donor) and `AG` sites (acceptor/non–acceptor), where a value of 2.0 bits/nucleotide indicates complete randomness. The context of the `GT/AG` non–splice sites was, apart from small deviations stemming from minor asymmetries in the overall dinucleotide probabilities, completely random, whereas the context of the true splice sites showed a non–randomness exactly matching the one found by inspecting the Kullback divergence.

2.3 Architecture of the neural networks

The neural networks used in this investigation were of the feedforward type. They received input from windows scanning the DNA sequence, each window configuration was represented numerically as a binary string, see Figure 4 and below for details.

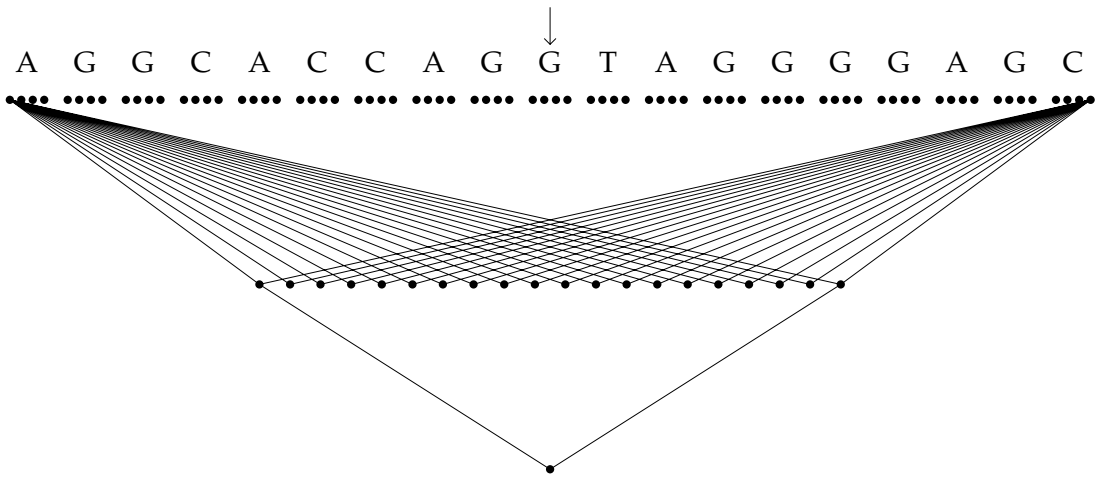

Figure 4: Architecture of the feed-forward neural network used for classifying the middle nucleotide in one of two categories: the dichotomy: splicing donor site or non–splicing donor site, or, splicing acceptor site or non–splicing acceptor site, or coding (E) or non–coding (IMN). The particular architecture shown has nineteen nucleotides visible to it in the input layer, twenty non–linear hidden units and one output unit. The input layer encoded each DNA window configuration as a binary string with $19 \times 4 = 76$ bits, representing nineteen blocks of four elements, an `A` by the binary pattern 1000, `T` by 0100, `G` by 0010 and `C` by 0001. The 76 bits were broadcasted to the twenty non–linear processing units interfacing the input layer and the output layer. Each of the twenty units computed a weighted sum with 76 terms, and subsequently this sum (subtracted an internal threshold value), was made the argument of a non–linear sigmoid function giving a real number, limited downwards by zero and upwards by one. From these twenty reals the single unit in the output layer similarly computed a real number. The activity of the output unit was compared with a cutoff value, separating the two output category assignments from each other. Notice that only connections branching off from the first and the last unit in each layer are shown on the figure. This 19w-20h-1o architecture contained $19 \times 4 \times 20 + 20 \times 1 = 1540$ weights and $20 + 1 = 21$ internal thresholds, making a total of 1561 adjustable parameters.

The networks were internally equipped with processing elements which performed simple computations: summing up input values, I_j, each of which were multiplied by a specific weight factor, W_j; subtracting from this sum an internal threshold value, t, and subsequently making the result the argument of a non–linear function possessing a sigmoidal (*i.e.* S–shaped) form. The value of the latter represented a real–numbered activity, limited downwards by zero and upwards by one

$$O = \sigma(\sum_{j=1}^{N} W_j I_j - t), \tag{1}$$

where $\sigma(x) = 1/(1 + \exp(-x))$, and N the number of inputs. One layer of connections in these neural networks connected the input layer with a hidden layer of non–linear processing elements, and a second layer of connections connected the hidden layer with the output layer. The size in nucleotides S_w of the window into the DNA sequence was in the range from 11 to 401, and the number of hidden units S_h in the range from 0 to 200.

The binary strings encoding the window configuration was composed of S_w blocks of four elements, an `A` being represented by the binary pattern 1000, `T` by 0100, `G` by 0010 and `C` by 0001. The 15 unknown nucleotides were represented by 0000. In the output layer one processing element represented the assignment of the middle nucleotide in the input window by the patterns 1 and 0. (In the actual implementation of the algorithm binary 1 was represented by 0.9 and binary 0 by 0.1, in order to speed up the convergence of the training process.) The binary representation is similar to the one used in the NETtalk network (Sejnowski & Rosenberg, 1987) which assigned phonemes to alphabetic letters of English text and the one used in networks (Qian & Sejnowski, 1988; Bohr *et al.*, 1988; Holley & Karplus, 1989; Kneller *et al.*, 1990) which classifies amino acids into categories of their corresponding protein secondary structure: alpha-helix, beta-sheet and coil. A similar sparse coding scheme was applied in a network predicting distance constraints on the three-dimensional structure of protein backbones (Bohr *et al.*, 1990a).

During training, the network was exposed to input window configurations of size S_w and the corresponding output category assignments, and the strengths of the connections were modified (Rumelhart *et al.*, 1986) so that the numerical value of the output neuron gradually approached the desired value, 0.9 or 0.1, depending on the window configuration in question. Each connection, say between neuron i and j, was modified by a small amount proportional to the negative gradient of the error made by the network

$$\Delta W_{ij} \propto -\frac{\delta E}{\delta W_{ij}}, \tag{2}$$

where the total network error E was the sum of the squared deviations from the desired output values

$$E = \sum_{\alpha} \sum_{k} (O_k^{\alpha} - D_k^{\alpha})^2, \tag{3}$$

α enumerated the input window configurations and k the output neurons. The input window was sequentially shifted, one nucleotide at a time, down the sequence of each gene. The total 65 gene training set (part I) consisted of 385488 input window configurations, the 30 gene test set (part II) of 190987. The reduced sequence segment data sets contained 33431 and 11918 configurations, respectively. For a network with S_w nucleotides in the window, 4^{S_w} possible window configurations exist, a number that grows rapidly with increasing window size, for $S_w = 11$, $4^{11} = 4194304$, for $S_w = 13, 4^{13} = 67108864$, and so on. The goal of the training was, by adjusting the network parameters, to make an internal representation of the systematics in the mapping between the set of window configurations in part I and the two output categories that did generalise well to the corresponding mapping in part II.

During a training session the success of the training was monitored in two different ways, one concerning the training set as a whole, another concerning each window configuration separately. The performance on the training set as a whole was quantified by monitoring the decrease in the total network error E. If E remained constant for a large number of presentations of the training set, it indicated that no improvements would be gained from further training. The network had learnt a single window configuration, if the real–numbered activity of the output neuron was falling on the same side of a cutoff value, as the training target of the window configuration in question. The cutoff value, separating the two output category assignments from each other, was mostly chosen to be 0.5. Thus, at any moment during training, we had an objective criterion for successful learning, a criterion pointing uniquely to those inputs not being classified correctly.

In a previous study (Brunak, *et al.*, 1990a,b) it was shown how the examination of the unlearnable window configurations led to the detection of wrongly assigned donor sites due to insufficient proofreading of databank entries as well as erroneous interpretation of experiments. Information about the degree of regularity of for example donor site window configurations, could be obtained by monitoring the course of the training: For the donor site assignment problem window configurations learned early in the training process showed stronger conformity to the standard 5' consensus sequence $\frac{\mathtt{C}}{\mathtt{A}}\mathtt{AG/GT}\frac{\mathtt{G}}{\mathtt{A}}\mathtt{AGT}$, than those being learned late, see section 3.1 and Table 6 for details.

2.4 Evaluation measures

For each of the three classification tasks dealt with in this paper, the distribution over categories was highly asymmetric. The non–donor sites largely outnumbered the splicing donor sites, the non–acceptor sites largely outnumbered the splicing acceptor sites, and in the coding (E) or non–coding

(IMN) problem, we also had a skewed distribution, although to a smaller extent.

As a consequence of this imbalance, mere percentages of correctly predicted window configurations are bad indicators of the predictive performance. An algorithm which assigns donor sites to all the nucleotides in a gene will correctly predict the presence of donor sites with an accuracy of 100%, but is not of much use. An alternative measure which takes the relation between correctly predicted positives and negatives as well as false positives and negatives into account is the *correlation coefficient* (Mathews, 1975),

$$C(X) = \frac{(P_X N_X) - (N_X^f P_X^f)}{\sqrt{(N_X + N_X^f)(N_X + P_X^f)(P_X + N_X^f)(P_X + P_X^f)}} \qquad (4)$$

where X can be any of the categories E, I, I^D, I^A, M or N. P_X and N_X are the correctly predicted positives and negatives, and P_X^f and N_X^f are similarly the incorrectly predicted positives and negatives. A perfect prediction gives $C(X) = 1$, whereas a truly imperfect one gives $C(X) = -1$. This measure as well as the relation between the number of true and false positives, will be used to quantify the performance of the algorithms in the present paper.

3 Results

3.1 Donor sites

The ability of the network method to predict the function of a nucleotide as being splicing donor or not was assessed by training networks of widely different architectures. Perceptrons with 0, 5, 10, 15, 20, 40 and 50 hidden units were used — all with 11, 13, 15, 17, 21, 31, 41 and 51 nucleotides visible to them in the input layer. For the reduced data set of sequence segments the result of 56 runs with these parameters is shown in Figures 5 and 6.

Figure 5 shows the maximal test correlation coefficient $C_{max}(D)$ obtained during 500 presentations of the 331 sequence segments in the training set. The corresponding test set contained the 118 sequence segments from part II each with a length of 101 nucleotides — 50 nucleotides before and after the donor site. For the networks *with* hidden units $C_{max}(D)$ was in the interval 0.81-0.93, fifteen nucleotides and forty hidden units being superior with $C_{max}(D) = 0.9244$. This network correctly classified 111 of the 118 donor sites (94.1%), and 11789 of the 11800 non–donor sites (99.9%). The probability for false negatives was thus as low as 0.059, and for false positives even lower, 0.001.

The optimal network trained on sequence segments ($S_w = 15$, 40 hidden units) had a $C(D) = 0.41$, when tested on the 30 complete sequences in part II of the data set. The network again correctly classified 111 of the 118 donor sites (94.1%), but incorrectly classified 536 of the 190987 non–donor sites

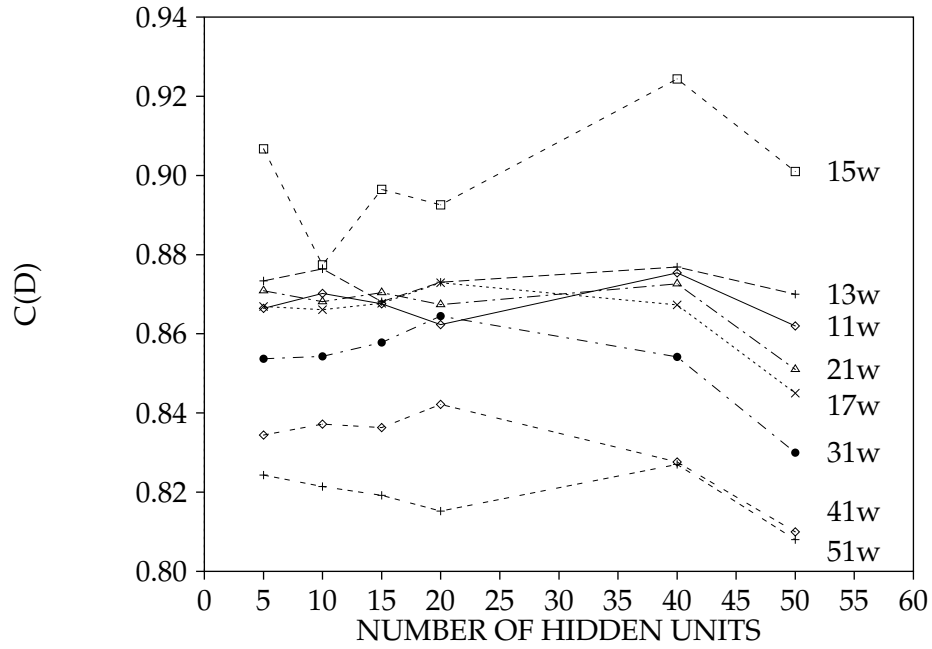

Figure 5: The maximal test correlation coefficient $C_{max}(D)$ as function of the number of perceptron hidden units, for donor site sequence segments. Eight curves show the results for perceptron's with 11, 13, 15, 17, 21, 31, 41 and 51 nucleotides in the window.

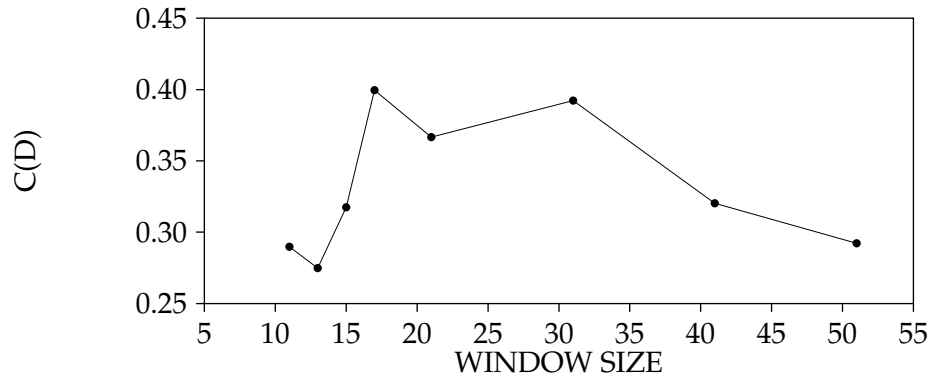

Figure 6: The maximal test correlation coefficient $C_{max}(D)$ as function of window size in nucleotides for perceptron's without hidden units, obtained on donor site sequence segments. For the eight window sizes ($S_w = 11, 13, 15, 17, 21, 31, 41$ and 51) $C_{max}(D)$ is non–zero.

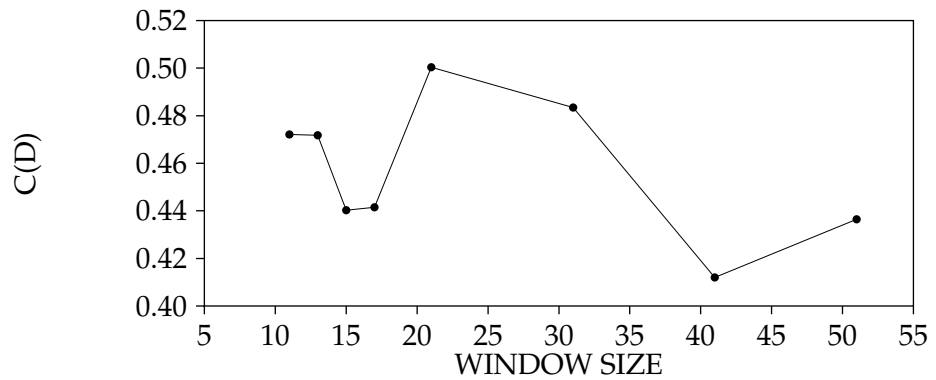

Figure 7: The maximal test correlation coefficient $C_{max}(D)$ on the 30 complete sequences obtained during 500 presentations of the 65 *complete sequences* as function of window size in nucleotides for perceptron's with 40 hidden units.

as being donor sites. The number of GT dinucleotides in part II was 9620, so the network still reduced the number of possible donor site candidates considerably. The false positive donor sites were not distributed at random in the sequences, in section 3.1.1 and 3.2.1 the distributions of false positive donor and acceptor sites are analysed in detail. For the truncated version of the data set of complete sequences the number of false positives was reduced to 423 giving $C(D) = 0.44$.

The results for training and testing on complete sequences are given in Figure 7. The number of hidden units was kept fixed at forty. The best $C_{max}(D)$ was obtained by a window size of 21 nucleotides, $C_{max}(D) = 0.50$. Out of 190987 possible ones the network made only 50 false positive assignments, but 62 of the 118 donor sites was incorrectly classified as being non–donor sites, the recognition percentage being 47.5%.

Information about the degree of regularity of donor site window configurations was obtained by monitoring the order in which the networks learned to assign window configurations correctly. In this way classes of donor sites could be identified, showing a decreasing degree of complementarity to the nucleotides 4-11 of the U1RNA. The largest cluster comprised sequences conforming to the standard 5' consensus sequence $\frac{\mathtt{C}}{\mathtt{A}}AG/GT\frac{\mathtt{G}}{\mathtt{A}}$AGT as well as sequences with substitution at position 6 on the intron side. In primates sequences with substitution at position 6 are close to being as frequent as sequences with complete consensus agreement, (Jacob & Gallinaro, 1989). In Table 6 it is shown how donor site window configurations learned early in the training process conformed more strongly to the standard 5' consensus sequence, than those being learned late. A central substring like GACCCAG/GTGAGTGG, from HUMACCYBA, was learned earlier than for example GGAGTTT/GTAAGCTC from HUMGHN which is further from the consensus sequence. The single donor site from HUMAPRTA which did not conform to the standard GT rule (C substituted for T) was extremely hard to learn. A secondary factor influencing the order in the training was the presence of

non–splicing donor window configurations closely resembling the true splicing donor window configurations. The presence of these almost conflicting window configurations did push the input-output relation of the networks towards increased non–linearity, which was harder to obtain by the infinitesimal adjustment of network parameters (2).

Networks trained on window configurations deprived of either the left or the right context, showed considerable differences in the ability to learn the donor site assignment task. Networks deprived of the left context learned much faster than networks deprived of the right context, which indicated that correlations in the base composition in the intron part is much more important in the donor site recognition than the base composition in the exon part. This observation is in agreement with the magnitudes of the Kullback divergence and the Shannon information given in section 2.2.2. A similar picture emerged in the acceptor site assignment case.

3.1.1 False donor sites

We have examined the densities of false positive donor sites in the functionally different parts of the total data set. For the optimal network trained on donor site sequence segments and tested on the total data set (part I and II in all: 576475 bp) the average density was 0.27 false positives *per* 100 bp, ranging from 0.16 in translated exon (E), 0.22 in untranslated exon (M), to 0.29 in intron (I) and 0.31 in non–transcribed DNA. The density was remarkably low in initial (E/M) exons, 0.09, close to the average exon value in internal exons (E), 0.18, and somewhat higher in terminal (E/M) exons, 0.22.

The robustness of the detection of true donor sites can be assessed from Figure 8, where the effect of changing the value of the cutoff for donor site assignment can be observed. A window configuration was classified as splicing donor only if the real–numbered activity of the output neuron exceeded the cutoff, fixing the confidence level of the assignment. The numbers given in section 3.1 were obtained with a cutoff value of 0.5. Larger correlation coefficients could be obtained by increasing the cutoff value up to around 0.8, at the cost of a considerable decrease in the number of true donor sites detected.

To investigate the possibility of a functional role of "false" donor sites in terminal exons, we have for comparison plotted the probability of false donor sites as a function of the distance from the last acceptor site in internal exons and in terminal exons. Figure 9 shows the probability of the *first* false positive donor site after the acceptor site, for internal exons (upper part) and terminal exons (lower part). If the false positives were distributed at random, we would expect a decreasing probability distribution, like the one found on the plot for the internal exons. The distribution on terminal exons initially shows an increase and then a decreasing probability, signaling a non–random distribution of the false positives. The density of false positive donor sites in terminal exons peaks in the distance 400-450 bp from the

Epoch	GenBank Locus	Sequence
1	HUMA1ATP	TACATCTTCTTTAAAGGTAAGGTTGCTCAACCA
1	HUMA1ATP	CCTGAAGCTCTCCAAGGTGAGATCACCCTGACG
1	HUMACCYBA	CCACACCCGCCGCCAGGTAAGCCCGGCCAGCCG
1	HUMACCYBA	CGAGAAGATGACCCAGGTGAGTGGCCCGCTACC
1	HUMACTGA	GCGCCCCAGACACCAGGTGAGTGGATGGCGCCG
1	HUMACTGA	AGAGAAGATGACTCAGGTGAGGCTCGGCCGACG
1	HUMACTGA	CACCATGAAGATCAAGGTGAGTCGAGGGGTTGG
1	HUMADAG	TCTTATACTATGGCAGGTAAGTCCATACAGAAG
1	HUMALPHA	CGTGGCTCTGTCCAAGGTAAGTGCTGGGCTACC
1	HUMALPI	CCTGGCTCTGTCCAAGGTAAGGGCTGGGCCACC
1	HUMALPPD	TGTGGCTCTGTCCAAGGTAAGTGCTGGGCTACC
1	HUMAPRTA	CCTGGAGTACGGGAAGGTAAGAGGGCTGGGGTG
1	HUMCAPG	GAAGGCTGCCTTCAAGGTAAGGCATGGGCATTG
1	HUMCFVII	GGAGTGTCCATGGCAGGTAAGGCTTCCCCTGGC
1	HUMCP21OH	CACCTTGGGCTGCAAGGTGAGAGGCTGATCTCG
1	HUMCP21OHC	CACCTTGGGCTGCAAGGTGAGAGGCTGATCTCG
1	HUMCS1	GTGGCAATGGCTCCAGGTAAGCGCCCCTAAAAT
1	HUMCSFGMA	AATGTTTGACCTCCAGGTAAGATGCTTCTCTCT
1	HUMCSPB	AAAGACTTCCTTTAAGGTAAGACTATGCACCTG
1	HUMCYC1A	GCTACGGACACCTCAGGTGAGCGCTGGGCCGGG
...	...	...
2	HUMA1ATP	CCTGGGACAGTGAATCGTAAGTATGCCTTTCAC
2	HUMA1ATP	AAAATGAAGACAGAAGGTGATTCCCCAACCTGA
2	HUMA1GLY2	CGCCACCCTGGACCGGGTGAGTGCCTGGGCTAG
2	HUMA1GLY2	GAGAGTACCAGACCCGGTGAGAGCCCCCATTCC
2	HUMA1GLY2	ACCGTCTCCAGATACGGTGAGGGCCAGCCCTCA
2	HUMA1GLY2	GGGCTGTCTTTCTATGGTAGGCATGCTTAGCAG
2	HUMA1GLY2	CACCGACTGGAAAAAGGTAAACGCAAGGGATTG
2	HUMACCYBA	GCGCCCCAGGCACCAGGTAGGGGAGCTGGCTGG
2	HUMACCYBA	CAGCCTTCCTTCCTGGGTGAGTGGAGACTGTCT
2	HUMACCYBA	CACAATGAAGATCAAGGTGGGTGTCTTTCCTGC
2	HUMACTGA	TCGCGTTTCTCTGCCGGTGAGCGCCCCGCCCCG
2	HUMADAG	CTTCGACAAGCCCAAAGTGAGCGCGCGCGGGGG
2	HUMADAG	TGTCCAGGCCTACCAGGTGGGTCCTGTGAGAAG
2	HUMADAG	CGAAGTAGTAAAAGAGGTGAGGGCCTGGGCTGG
...	...	...
11	HUMCS1	AACGCAACAGAAATCCGTGAGTGGATGCCGTCT
11	HUMGHN	AACACAACAGAAATCCGTGAGTGGATGCCTTCT
52	HUMHSP90B	CTCTAATGCTTCTGATGTAGGTGCTCTGGTTTC
80	HUMMETIF1	ACCTCCTGCAAGAAGAGTGAGTGTGAGGCCATC
112	HUMHSP90B	ATACCAGAGTATCTCAGTGAGTATCTCCTTGGC
113	HUMHST	GCGGACACCCGCGACAGTGAGTGGCGCGGCCAG
113	HUMLACTA	GACATCTCCTGTGACAGTGAGTAGCCCCTATAA
151	HUMKAL2	ATCGAACCAGAGGAGTGTACGCCTGGGCCAGAT
157	HUMCS1	CACCTACCAGGAGTTTGTAAGTTCTTGGGGAAT
157	HUMGHN	CACCTACCAGGAGTTTGTAAGCTCTTGGGGAAT
164	HUMALPHA	CAACATGGACATTGATGTGCGACCCCCGGGCCA
622	HUMCFVII	CTGATCGCGGTGCTGGGTGGGTACCACTCTCCC
636	HUMADAG	CCTGGAACCAGGCTGAGTGAGTGATGGGCCTGG
895	HUMAPOCIB	TCCAGCAAGGATTCAGGTTGTTGAGTGCTTGGG
970	HUMALPHA	CGGGCCAAGAAAGCAGGTGGAGCTGGGGCCCGG
2114	HUMAPRTA	ATCGACTACATCGCAGGCGAGTGCCAGTGGCCG

Table 6: Donor site window configurations being learned early and late in the course of training. The applied network was small, (9 nucleotides in the window, 2 hidden units and 1 output unit), and it was trained on small 33 bp segments surrounding the 331 splice sites in part I of the data set. Shown is the training epoch at which a configuration was assigned splicing donor correctly by the network, its GenBank locus, and the nucleotide context surrounding the central G. Segments with large deviations from the standard donor site consensus sequence, $\frac{\mathtt{C}}{\mathtt{A}}\mathtt{AG/GT}\frac{\mathtt{G}}{\mathtt{A}}\mathtt{AGT}$, were learned only after a relatively large number of presentations.

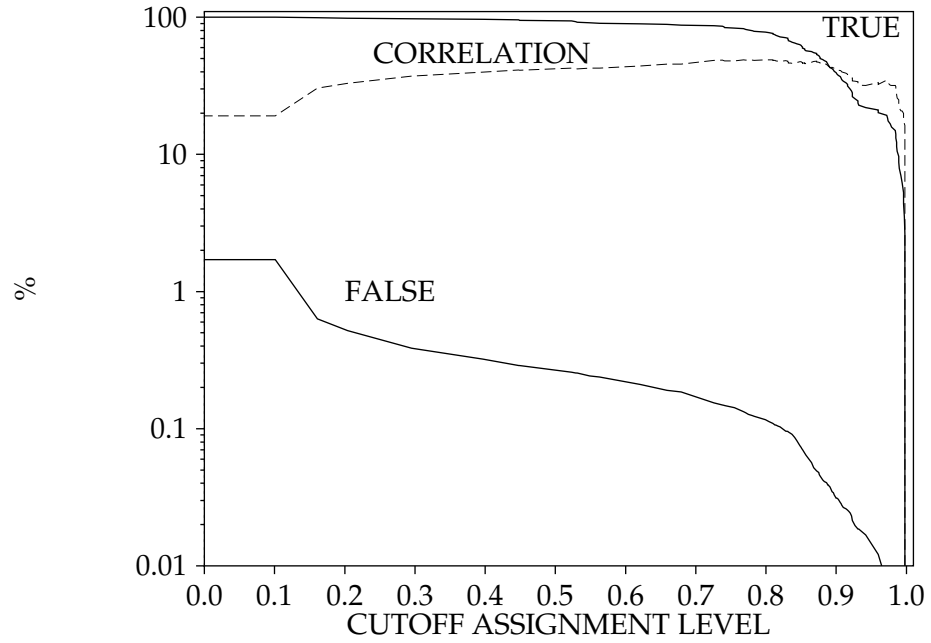

Figure 8: Prediction of donor sites on complete sequences. Shown are percentages of true and false positive donor sites and the correlation coefficient $C(D)$ ($\times 100$) as function of the donor site assignment cutoff value. The transition in the amount of false positives at 0.1 stems from the encoding of binary 0 as 0.1, see section 2.3.

acceptor site. The distribution of *all* the false positives in internal and terminal exons shows a similar behaviour.

3.2 Acceptor sites

In compiling the acceptor site correlation coefficients, $C(A)$, we found these to be smaller than those of the donor sites, $C(D)$, in general, as can be seen from the figures 10, 11, 12, which correspond to figures 5, 6, 7 of the donor sites.

The optimal network architecture for predicting acceptor sites embedded in sequence segments had a window size of 41 nucleotides and 20 hidden units, as opposed to the 15 nucleotides and 40 hidden units in the donor site case. Despite the variation in the number of hidden units in the perceptrons, and hence in the amount of non–linearity available, the decreasing coverage of the window configuration spaces for larger S_w, resulted in decreasing predictive performance as quantified by the correlation coefficients. The optimal window size of 41 nucleotides covers precisely the polypyrimidine tract. The optimal network correctly classified 100 of the 118 acceptor sites (87.4%) and 11778 of the 11800 non–acceptor sites (99.8%), the correlation coefficient being $C(A) = 0.83$.

A characteristic of the 18 false negatives was that the polypyrimidine tract was contaminated with purines. In two cases even with `AG` dinucleo-

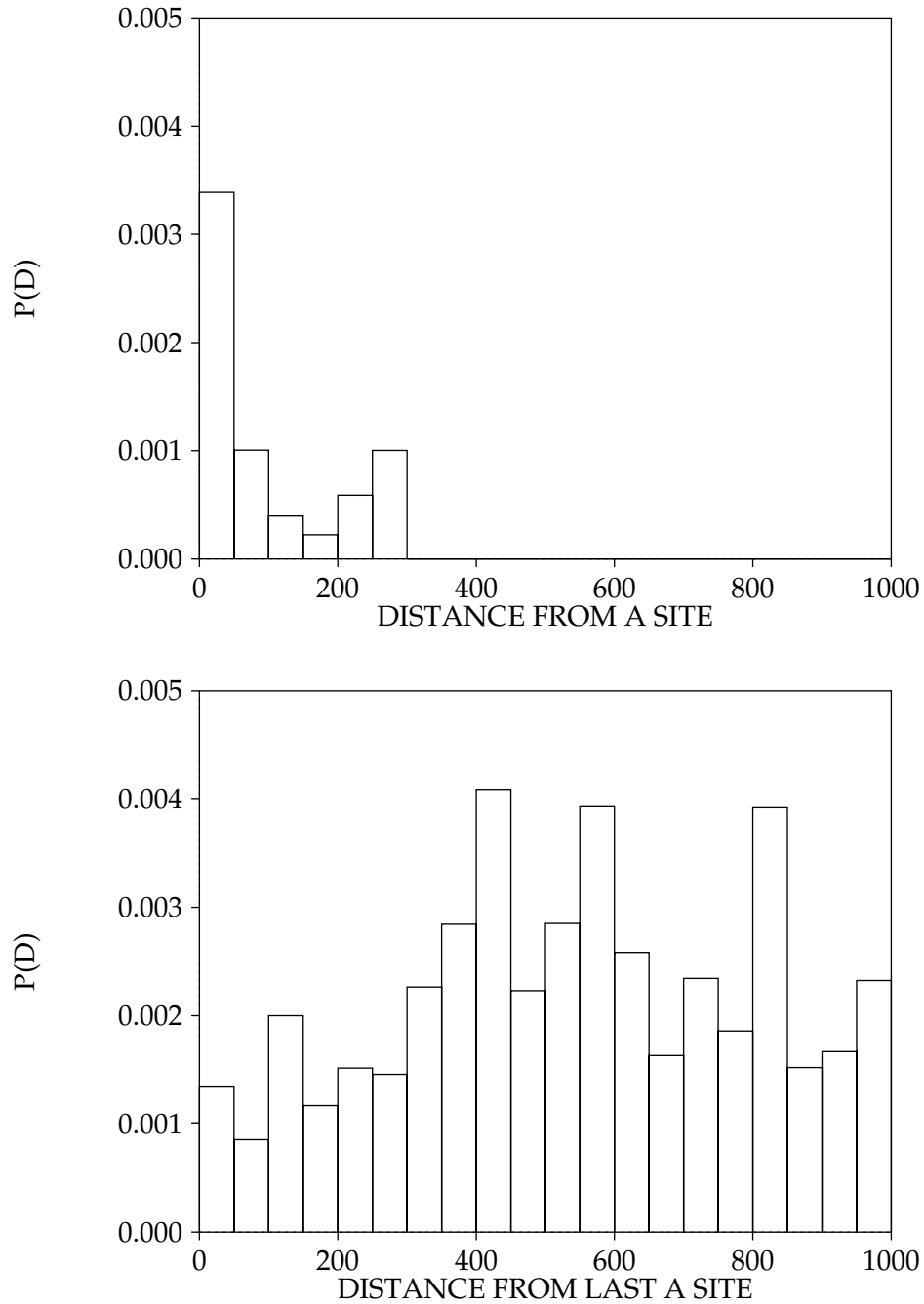

Figure 9: The probability of the *first* false donor site after the acceptor site, in internal exons (upper), and in terminal exons (lower). Probabilities were computed as the number of incorrectly assigned donor sites divided by the number of possible ones. Counts were merged in 50 bp intervals, 1-50, 51-100, ...

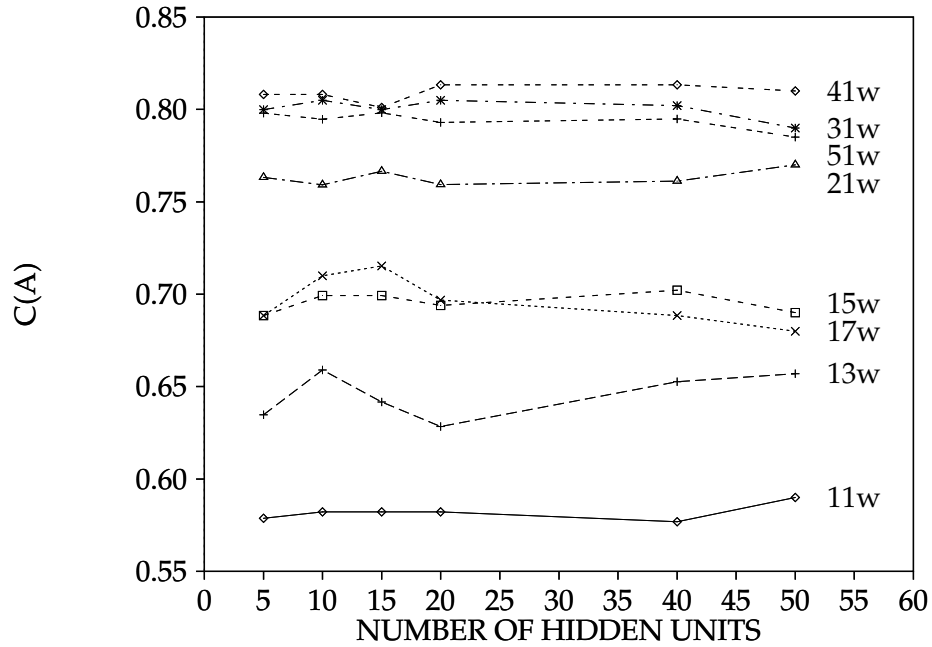

Figure 10: The maximal test correlation coefficient $C_{max}(A)$ as function of the number of perceptron hidden units, for acceptor site sequence segments. Eight curves show the results for perceptron's with 11, 13, 15, 17, 21, 31, 41 and 51 nucleotides in the window.

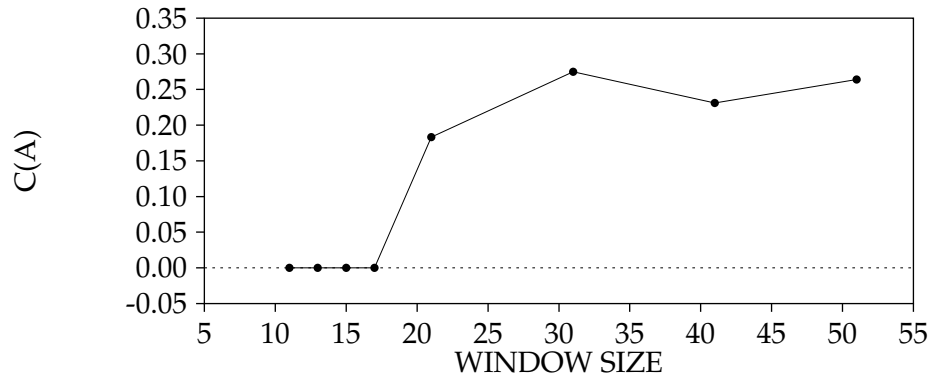

Figure 11: The maximal test correlation coefficient $C_{max}(A)$ as function of window size in nucleotides for perceptron's without hidden units, obtained on acceptor site sequence segments. For the four window sizes 11, 13, 15, and 17 $C_{max}(A)$ is zero and hence the acceptor/non–acceptor assignment problem is not linearly separable.

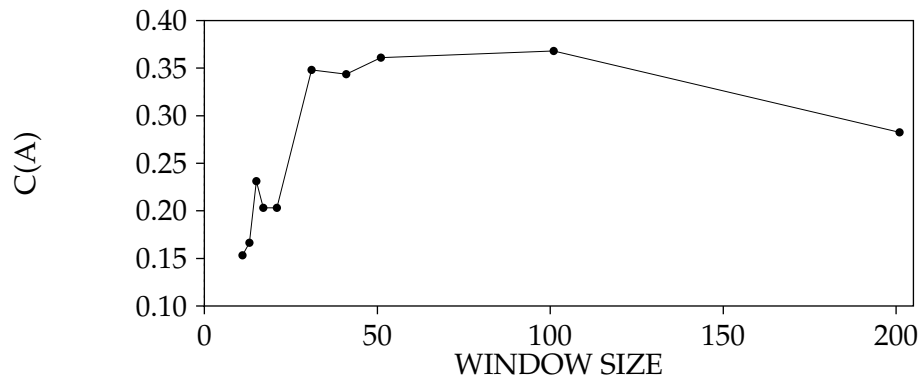

Figure 12: The maximal test correlation coefficient $C_{max}(A)$ on *complete sequences* as function of window size in nucleotides for perceptron's with 40 hidden units. The test set comprised 190987 window configurations taken from the last 30 genes in the data set.

tides being present just to the left of the true acceptor site, in HUMP45C17 as (`TCTGGAGCAG`) and in HUMTKRA as (`GGGAGCACAG`). In HUMMT2A a false positive acceptor site was caused by an additional `AG` dinucleotide just to the left of the true splice site, which itself was classified correctly, (`CTTCTAGCAG`). The presence of these cases (HUMP45C17, HUMTKRA and HUMMT2A) contradicts the proposed model by Smith *et al.*, 1989, of the acceptor site being determined by a scanning process that recognises the first `AG` located 3' of the polypyrimidine tract.

When tested on the 30 complete test sequences from part II of the data set this network again correctly assigned the 100 acceptor sites, but incorrectly assigned 856 of the 190987 non–acceptor sites bp in the test sample as acceptor sites thereby lowering the correlation coefficient $C(A)$ to 0.32. The number of `AG` dinucleotides in the test sample was 14560. For the truncated version of the data set of complete sequences the number of false positives was reduced to 679 giving again a correlation coefficient $C(A) = 0.33$.

The results for training and testing on complete sequences are given in Figure 12. The best $C_{max}(A)$ was obtained by a window size of 51 nucleotides, $C_{max}(A) = 0.36$. Out of 190987 possible ones the network made only 68 false positive assignments, but 77 of the 118 acceptor sites were incorrectly classified as being non–acceptor sites, the recognition percentage being as low as 34.7%. Networks trained on complete sequences learn to be very restrictive in assigning acceptor and donor sites. During training they have been exposed to too many examples inside introns and exons which from a local viewpoint closely resemble true sites, but which are not selected by the biological machinery.

As anticipated, no strong clustering of consensus–like contexts did result from monitoring the course of training on acceptor sites, corresponding to the donor site case as shown in Table 6. However, a weaker pattern was definitely present, in that acceptor sites surrounded by `C` and `G`,

...NNNCAGGNNN... were learned first, together with a smaller number of sites having either one of these present: ...NNNCAGNNNN... or ...NNNNAGGNNN.... It took in general more presentations for the networks to learn the acceptor site assignment when the immediate neighbourhood contained T or A. This picture is in agreement with the 3' consensus sequence $(\frac{\mathtt{T}}{\mathtt{C}})_n\mathtt{N}(\frac{\mathtt{C}}{\mathtt{T}})\mathtt{AG/G}$ described in (Mount, 1982), where 65% of the nucleotides immediately to the left of the the conserved AG nucleotides were found to be C and 52% of the nucleotides immediately to the right were found to be G.

The network had severe difficulties in learning the two non–AG acceptor site assignments in the GenBank entry HUMAK1 (acc. no. J04809). Consultation of the original publication (Matsuura *et al.*, 1989) revealed that the GenBank assignment at position 10517 is in conflict with the assignment at position 10507 in the paper which agrees with the assignment made by the network. Furthermore, the assignment of a splicing acceptor at position 3947 in the paper does not seem to be supported by the chicken cDNA clone (Kishi *et al.*, 1986) used by Matsuura *et al.* to make the assignment. The human and chicken genes have a rather low degree of sequence similarity in this region. The network assigns position 3949 as splicing acceptor. This must however be confirmed by experiment.

3.2.1 False acceptor sites

The densities of false positive acceptor sites in the functionally different parts of the data set were in general higher than in the donor site case. For networks trained on acceptor site sequence segments and tested on the total data set (part I and II in all: 576475 bp) the average density was 0.44 false positives *per* 100 bp, ranging from 0.24 in translated exon (E), 0.46 in intron (I), to 0.47 in untranslated exon (M), and 0.51 in non–transcribed DNA. The density in initial (E/M) exons was 0.36, in internal exons (E), 0.26, and 0.38 in terminal (E/M) exons.

The robustness of the detection of true acceptor sites can be assessed from Figure 13, where the effect of changing the cutoff for acceptor site assignment can be observed. The numbers given in section 3.2 were obtained with a cutoff value of 0.5. Larger correlation coefficients could be obtained by shifting the cutoff value to around 0.6, at the cost of a considerable decrease (down to $\approx 60\%$) in the percentage of true acceptor sites detected.

The distribution of false positive acceptor sites in exons did not have a strong non–random appearance, although the density of false positive acceptor sites in terminal exons was higher than in internal exons, a feature that also was observed in the case of the false positive donor sites as described above. Figure 14 shows the probability of the *first* false positive acceptor site after the true acceptor site, for internal exons (upper part) and terminal exons (lower part). The distribution of *all* the false positives in internal and terminal exons shows a similar behaviour. From inspection of predictions of networks trained to classify nucleotides in translated exon (E) or not (see section 3.3) it seemed clear that the exon–signal from the

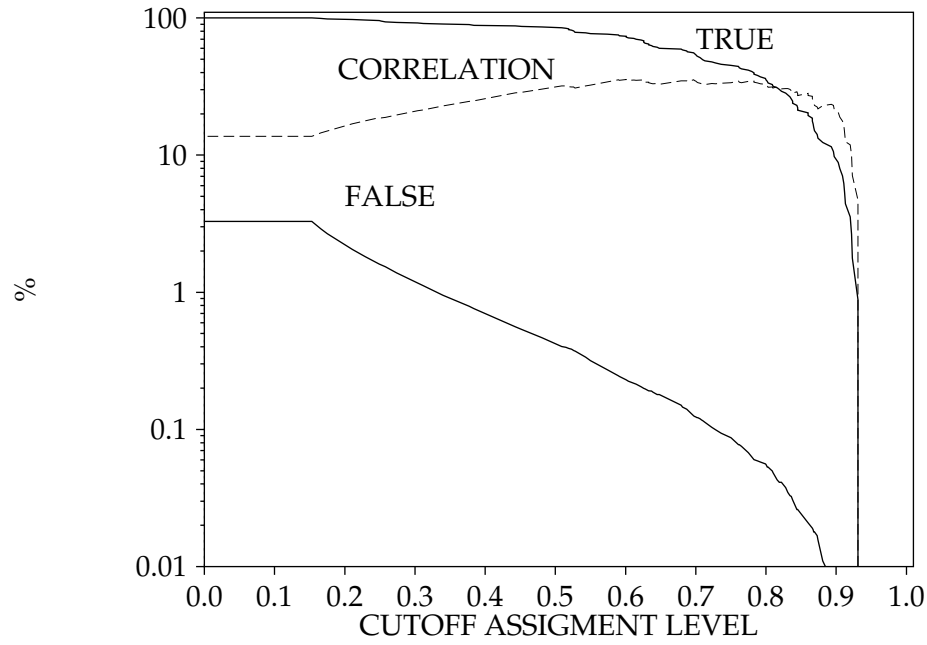

Figure 13: Prediction of acceptor sites on complete sequences. Shown are percentages of true and false positive acceptor sites and the correlation coefficient $C(A)$ ($\times 100$) as function of the acceptor site assignment cutoff value. The transition in the amount of false positives at 0.1 stems from the encoding of binary 0 as 0.1, see section 2.3.

coding part of the terminal exons differed markedly from the exon–signal of internal and initial exons, Figure 15. Large regions of terminal exons seemed to be intron–like, especially in the first third, and in this part densities of false positive acceptor sites approached the average density in introns.

3.3 Coding/non–coding classification

The problem of classifying pieces of DNA sequence according to their function as translated exon, intron, untranslated exon or other alternatives has been addressed by several methods (von Heijne, 1987), primarily either based on the compilation of tables of codon usage (Staden, 1984b), or detection of local nonrandomness in the sequence (Fickett, 1982). This problem is far easier when sequence segments are "pure" (Lapedes, 1988), that is, when they in their entirety stem from one specific class only. Difficulties arise when segments cover regions of transition between the different categories, which in the context of locating true splice sites, are of prime importance.

In order to test the network approach to this problem, networks were trained on the truncated data set with window sizes of 101, 141, 201, 301 and 401 nucleotides, in each case with 200 hidden units in them. The largest network contained 321201 adjustable parameters. Table 7 gives figures for the quality of the prediction for the 30 (truncated) test sequences in part II. Both the percentage of correctly assigned window configurations in the translated exon (E) category and the correlation coefficient C(E) did increase with the size of the window up to $S_w = 301$ nucleotides. When a network with this size was tested on part II of the truncated data set, C(E) became as high as 0.73, with 69.8% of the window configurations in the translated exon (E) category correctly assigned. The false positive non–coding category assignments were most frequent in small exons and in the initial part of terminal exons.

We also tried to train networks to classify window configurations in two categories where the coding exon category (E) and the non–coding exon category (M) were merged in order to get an EM/IN assignment. The result was negative: even networks with 300-500 hidden units had severe difficulties in learning the task, indicating that similarities between window configurations in the E and M categories are very weak or absent.

3.4 Joint coding/non–coding and splice site classification

The general picture concerning the confidence levels on the splice site prediction and the E/IMN classification (as measured by the output neuron activities) was that exons smaller than 75 bp had rather weak (E) exon output neuron levels (0.3-0.6), but relatively strong donor and acceptor site output neuron levels (0.7-1.0). Conversely, longer internal exons in general had rather sharp edges in the output neuron activities, with donor and acceptor site activities being somewhat weaker.

In order to improve the ratio between true and false donor and acceptor sites assigned, we have exploited the complementary relation between

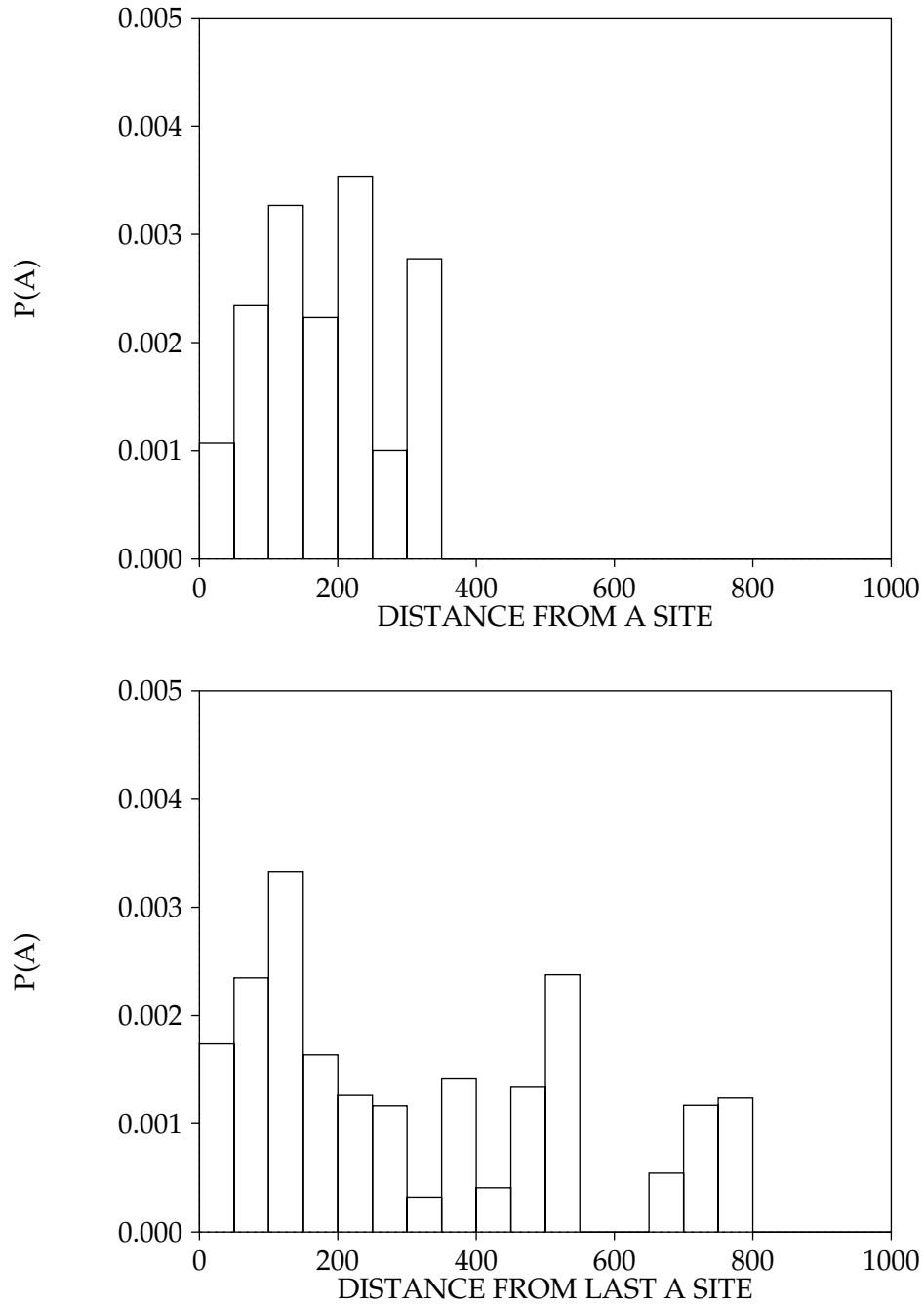

Figure 14: The probability of the *first* false acceptor site after the acceptor site, in internal exons (upper), and in terminal exons (lower). Probabilities were computed as the number of incorrectly assigned acceptor sites divided by the number of possible ones. Counts were merged in 50 bp intervals, 1-50, 51-100, ...

Window size	$C_{max}(E)$	%	E%	IMN%
101	0.54	84.4	66.7	89.2
141	0.63	88.3	68.5	93.3
201	0.67	89.6	67.4	95.3
301	0.73	92.6	69.8	97.5
401	0.64	89.0	67.7	94.6

Table 7: Quality of the coding (E) or non–coding (I/M/N) category prediction. Listed is the maximal test correlation coefficient $C_{max}(E)$, percentages of correctly assigned window configurations for all categories, for the translated exon (E) category, and for the category covering intron (I), untranslated exon (M) and other (N) given for each of the five window sizes.

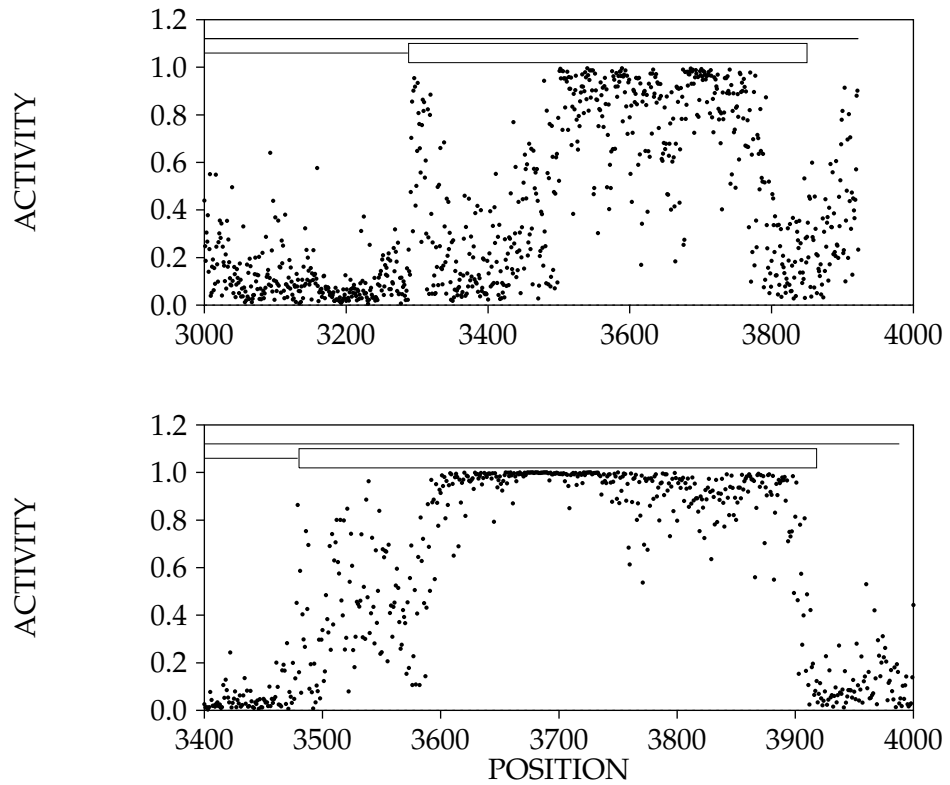

Figure 15: Typical examples of the result of the coding (E) or non–coding (IMN) prediction in terminal (E) exons. The terminal exons shown are from the GenBank entries HUMMYCRT (upper) and HUMPNMTA (lower). Large output neuron activities for the E exon category commence at some distance from the 5' end of many terminal exons in part II of the data set. The boxes correspond to correct exons, the adjoining lines to introns, whereas the top line gives the extension of the transcript. The network had a window size of 301 nucleotides and 200 hidden units.

the exon and the splice site confidence levels. This was done by letting the exon signal control the threshold for assigning splice sites. In regions with abruptly decreasing exon activity, donor sites should be "promoted" and acceptor sites suppressed, and *vice versa* in regions with abruptly increasing exon activity. In regions with only small changes in exon activity, that is, where the level was constantly high (inside exons), and where it was constantly low (inside introns (I), in untranslated exons (M) and in intergenic DNA (N)) a rather high confidence level in the splice site assignment should be demanded in order to suppress false positives.

To detect edges in the E/IMN output neuron levels we computed what was essentially the first derivative of the (E) exon output neuron activity by summing up activities to the right of a given point and subtracting the corresponding sum for the left side, dividing this difference by the number of addends. In order to reduce the number of cases where the individual sums covered both intron or exon regions, we used half the average length of the internal exons in the training set, 75 bp, as the scope when summing, giving derivatives close to $+1$ in the 3' end of introns and close to -1 in the 5' end for perfect E/IMN assignments. In Figures 16 and 17 average quality examples of the E/IMN signal, the derivative Δ, and the forest of donor and acceptor site signals exceeding an activity of 0.25 are given for the GenBank entries HUMOPS and HUMUBILP from part II of the data set. Note that some regions inside introns and in the non–transcribed part of the sequences show exon–like behaviour.

In compiling algorithms which regulates the splice site assignment levels we used the following expressions to assess possible weightings between the strength of the exon signal and the output, O, from the separate splice site networks: if

$$O_{\text{donor}} > e_D \Delta + c_D \tag{5}$$

assign splicing donor site, and if

$$O_{\text{acceptor}} > e_A \Delta + c_A \tag{6}$$

assign splicing acceptor site, where Δ was computed as described above. The two constants c_D and c_A are equivalent to the ordinary cutoff assignment parameters used in figures 8 and 13, whereas e_D and e_A control the impact of the exon signal. Together the four constants controlled the relative strengths between the donor site and the E/IMN network on one side, and between the acceptor site and E the E/IMN network on the other.

The four constants were optimised to give optimal performance in terms of correlation coefficients and in terms of the percentage of true splice sites detected. In the upper part of Figure 18 the percentage of true positive donor sites detected is plotted against the percentage of false positive donor sites assigned for complete sequences. For a given required level of detection of true sites optimal cutoff parameters was found for 1) the separate donor site assignment network, 2) the nucleotide weight matrix method of Staden (Staden, 1984a, 1990), and 3) the joint assignment method. In the upper

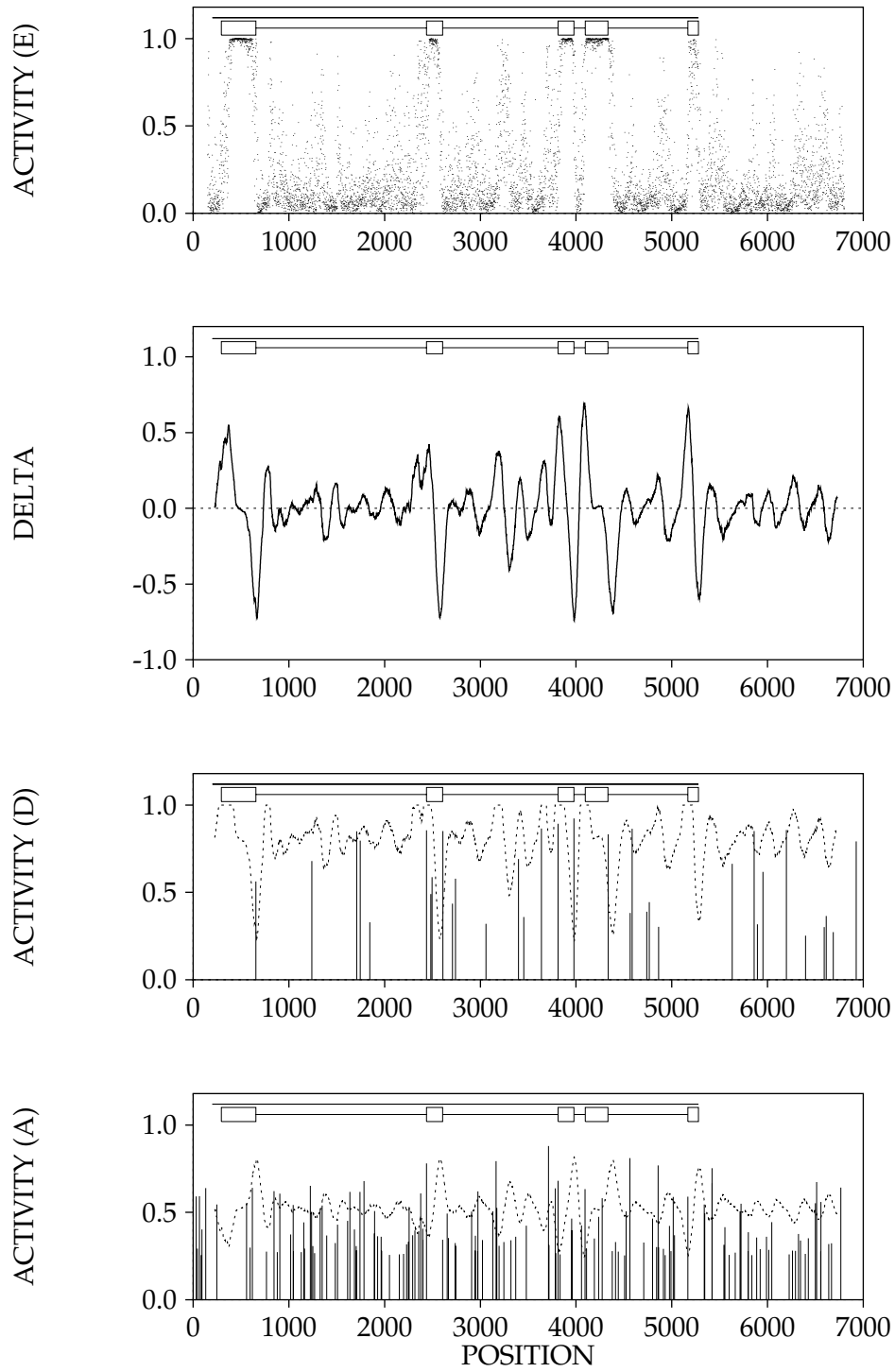

Figure 16: HUMOPS prediction. 1. (Top) Analog output from the output neuron in the E/IMN network showing strong exon signals from the correct exons. Inside the introns and in the terminal non–transcribed part of the sequence a number of regions show exon–like behaviour. The boxes correspond to correct exons, the adjoining lines to introns, whereas the top line gives the extension of the transcript.
2. The derivative of the analog E/IMN output.
3. The donor site activities (≥ 0.25) from the donor site assignment network.
4. (Bottom) The acceptor site activities (≥ 0.25) from the acceptor site assignment network. The variable cutoff assignment level for a 90% detection of true splice sites is shown as a dashed curve on the pin–diagrams 3. and 4.

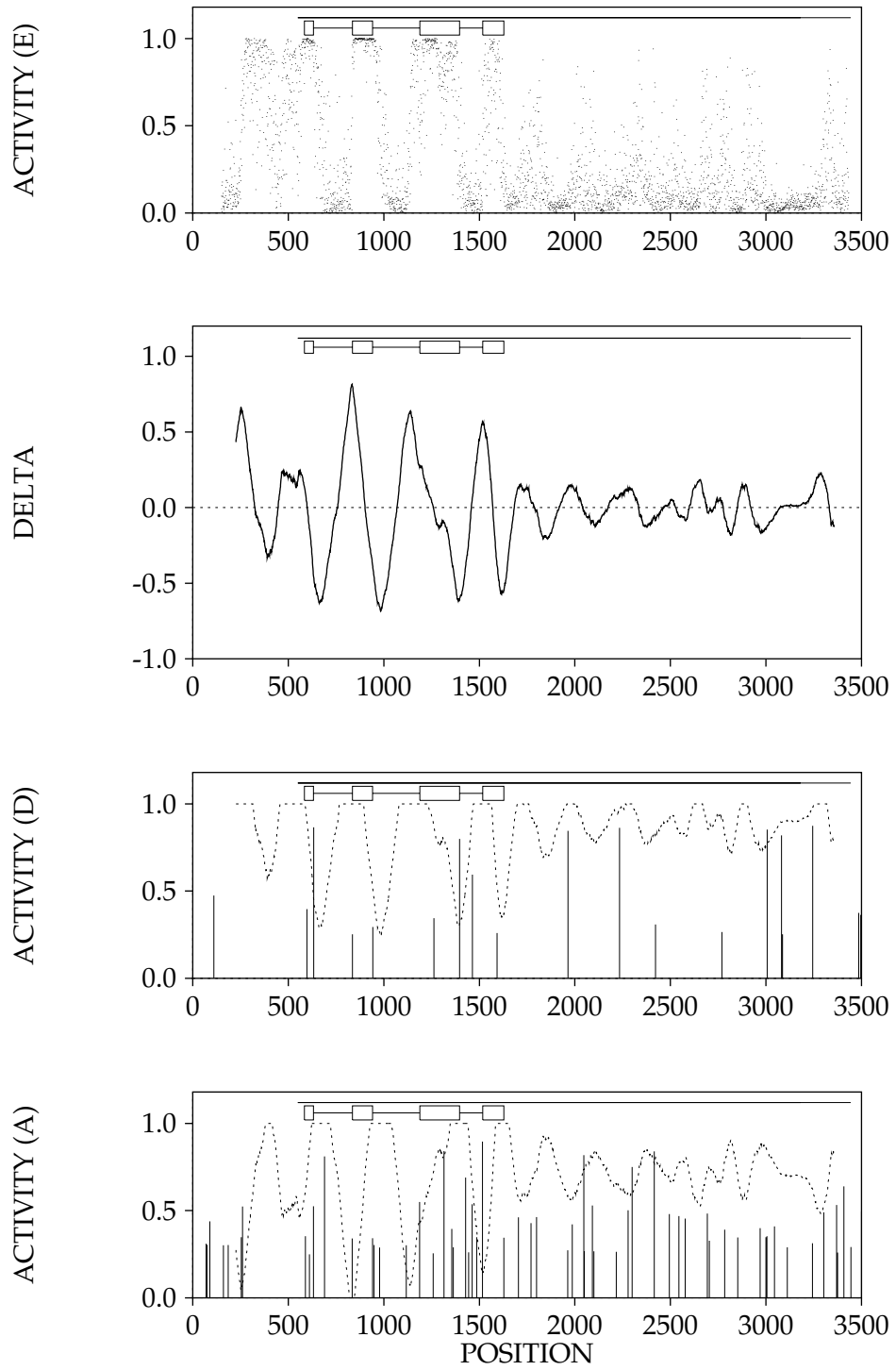

Figure 17: HUMUBILP prediction. 1. (Top) Analog output from the output neuron in the E/IMN network showing strong exon signals from the correct exons. In the initial and the terminal non–transcribed parts of the sequence a number of regions show exon–like behaviour. The boxes correspond to correct exons, the adjoining lines to introns, whereas the top line gives the extension of the transcript.
2. The derivative of the analog E/IMN output.
3. The donor site activities (≥ 0.25) from the donor site assignment network.
4. (Bottom) The acceptor site activities (≥ 0.25) from the acceptor site assignment network. The variable cutoff assignment level for a 90% detection of true splice sites is shown as a dashed curve on the pin–diagrams 3. and 4.

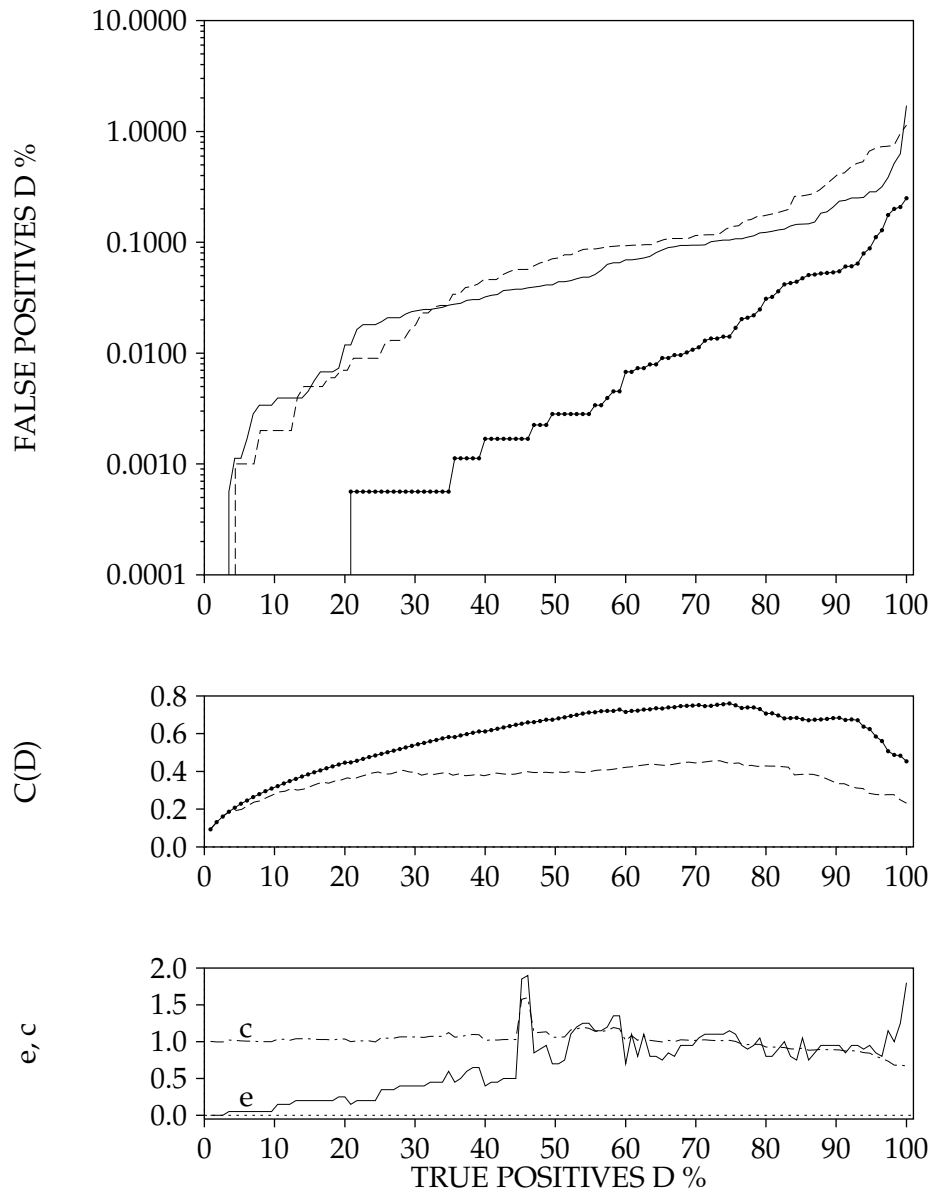

Figure 18: Results for testing donor site assignment on complete sequences. The upper part of the figure shows the percentage of false positive donor sites as function of the percentage of correctly predicted true positive donor sites. The dashed line gives the results for Staden's method obtained with optimal cutoff values for each percentage of true sites detected, the unbroken line the results for the separate donor site network with a simple variation of the cutoff assignment level yielding specific percentages of true donor site detection. The • line gives the result for the joint approach, where the cutoff assignment level was controlled by the exon signal. For 50% detection of true donor sites the three percentages (Staden, separate network, joint assignment) are 0.072, 0.045, 0.003, for 60%, 0.093, 0.071, 0.007, for 70%, 0.116, 0.095, 0.011, for 80%, 0.178, 0.127, 0.031, for 90%, 0.412, 0.238, 0.054, for 95%, 0.703, 0.289, 0.094, for 99%, 0.950, 0.632, 0.207, and for 100%, 1.137, 1.708, 0.250.
The middle part of the figure shows the corresponding correlation coefficients $C(D)$ for the joint assignment (•) and for Staden's method (dashed).
The lower part of the figure shows the values of the constants e_D and c_D giving the smallest number of false positive donor sites for the same succession of points as on the upper and middle parts.

part of Figure 18 the unbroken line represents the situation where the cutoff assignment level was not controlled by the exon signal, but adjusted so that a specified percentage of true splice sites was detected by the separate donor site network. The dashed line refers to the method of Staden, with results obtained by determining optimal cutoff values for each percentage of true sites detected, and the • line gives the results for the joint assignment. Note that the scale for the percentage of false positive donor sites is logarithmic.

The number of false positives assigned was much lower when the cutoff assignment level was controlled by the exon signal, by a factor between two and thirty depending on the required level of detection of true sites.

When a recognition level of 80% for the true donor sites was required from the joint approach, the lowest level of false positives was 0.031%, and $C(D) = 0.71$. The optimal constants for obtaining a recognition level of 80% was $e_D = 0.80$ and $c_D = 0.93$ as can be seen on the lower part of Figure 18, where the values of e_D and c_D correspond to the same succession of points as on the upper curve for the joint assignment. A recognition level of 95% implied a level of false positives close to 0.094% with $C(D) = 0.61$, and optimal constants $e_D = 0.95$ and $c_D = 0.83$. For a recognition level of 21% no false positive donor sites were assigned, meaning that — with the joint approach — close to one fifth of the donor sites could be detected without ambiguity at all.

The effect of the joint approach for prediction of acceptor sites is shown in Figure 19. The number of false positive acceptor sites assigned was again much lower when the cutoff assignment level was controlled by the exon signal. In the interval where the joint method did assign at least one false positive acceptor site, the exon signal was able to reduce the level of false positive assignments by a factor between two and forty. For a recognition level of 27% no false acceptor sites were assigned at all, meaning that — with the joint approach — more than a fourth of the acceptor sites could be detected without ambiguity at all.

When the splice site assignment was based on the coding/non–coding network only, that is by inspection of the Δ values, the quality of the donor site prediction was sligthly worse than the quality of Staden's method for all detection levels of true donor sites. In the interval below 65% detection of true sites the performance for acceptor sites was better than both Staden's method and the separate acceptor site network. Above 65% detection of true sites it was about the same.

For the two examples in figures 16 and 17, HUMOPS and HUMUBILP, the variable cutoff assignment level, with a 90% detection level of true sites, is shown together with the real–numbered output from the separate splice site networks. It can be observed how false splice sites with confidence levels exceeding 0.5 are suppressed, and how the assignment cutoff is lowered for true splice sites with confidence levels below 0.5.

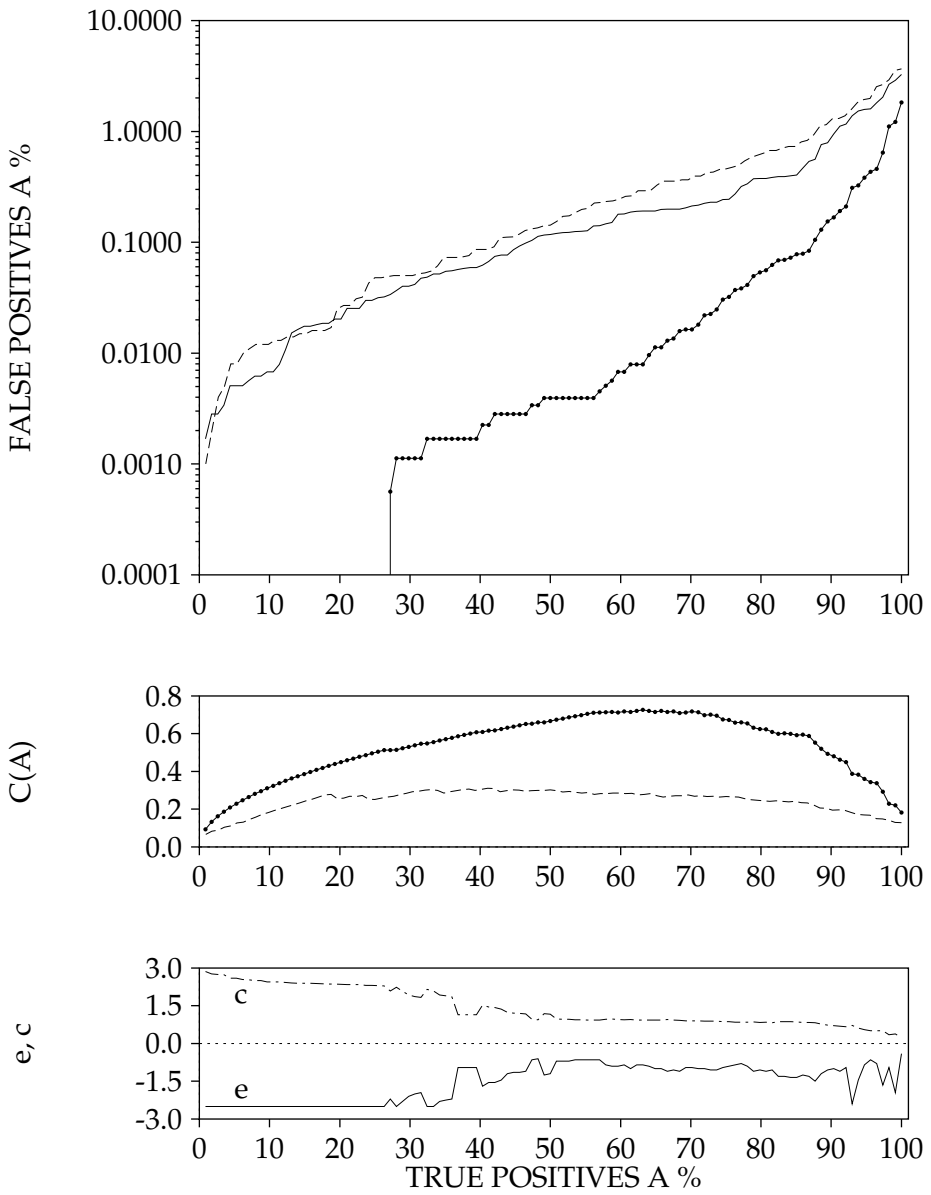

Figure 19: Results for testing acceptor site assignment on complete sequences. The upper part of the figure shows the percentage of false positive acceptor sites as function of the percentage of correctly predicted true positive acceptor sites. The dashed line gives the results for Staden's method obtained with optimal cutoff values for each percentage of true sites detected, the unbroken line the results for the separate acceptor site network with a simple variation of the cutoff assignment level yielding specific percentages of true acceptor site detection. The • line gives the result for the joint approach, where the cutoff assignment level was controlled by the exon signal. For 50% detection of true acceptor sites the three percentages (Staden, separate network, joint assignment) are 0.143, 0.117, 0.004, for 60%, 0.260, 0.179, 0.007, for 70%, 0.396, 0.217, 0.016, for 80%, 0.638, 0.375, 0.054, for 90%, 1.300, 0.950, 0.162, for 95%, 1.986, 1.606, 0.397, for 99%, 3.546, 2.896, 1.202, and for 100%, 3.671, 3.284, 1.828.
The middle part of the figure shows the corresponding correlation coefficients $C(A)$ for the joint assignment (•) and for Staden's method (dashed).
The lower part of the figure shows the values of the constants e_A and c_A giving the smallest number of false positive acceptor sites for the same succession of points as on the upper and middle parts.

4 Conclusion

Artificial neural networks have been trained to predict splice sites in human pre–mRNA. The problem of "sorting" the dinucleotides `GT` and `AG` on the basis of their function is hard — the `GT` dinucleotide constitutes roughly 5% of the dinucleotides in human DNA, but only a very small percentage of these belongs to the splicing donor category, in the order of 1.5%. The dinucleotide `AG` constitutes roughly 7.5% of all the dinucleotides and only around 1% of these function as splicing acceptor sites.

The results obtained in this paper build on a very conservative preparation of the data sets. The test set contained nearly 1/3 of the data extracted from GenBank, both in terms of sequences, specific sites and nucleotides. The training and test sets were searched for mutual sequence similarities, only 4 of 118 maximal overlaps did exceed 75% for the donor sites, and only 3 of 118 maximal overlaps did exceed 75% for the acceptor sites.

If all the splice sites in the data sets used in this study were selected on the basis of the local sequence composition only, we would expect to be able to model the splicing mechanism perfectly, given a training set sufficiently large to match the complexity of the splice site selection process. However, when prediction of splicing donor sites was based on local information only (the immediate neighbourhood — 11-51 bp — of a `GT` dinucleotide), the number of false positive donor site assignments was in the order of five times the number of true positives with correlation coefficients $C(D)$ at the level of 0.4. When the donor site prediction based on local information was joined by a larger network trained to recognise coding regions, building on information of a more global nature, the number of false positive donor site assignments could be reduced by a factor between two and thirty resulting in correlation coefficients at the level of 0.65-0.75 on complete GenBank sequence extracts. Both the separate network and the joint assignment method did perform considerably better than the weight matrix method of Staden. For a 90% detection of true donor sites, Staden's method assigns 0.412% false positives, where the separate network and the joint method assigns 0.238% and 0.054%, respectively. For 95% detection of true donor sites the corresponding percentages are 0.703%, 0.289% and 0.094%.

The network prediction of splicing acceptor sites based only on local information, had a worse performance with a ratio between false and true positive sites close to eight and correlation coefficients close to 0.3. When joined with the coding/non–coding network, the number of false positive acceptor site assignments could be reduced by a factor between two and forty resulting in correlation coefficients at the level of 0.55 on complete GenBank sequence extracts. Also for the acceptor site assignment problem both the separate network and the joint assignment method did perform considerably better than the weight matrix method of Staden. For a 90% detection of true acceptor sites, Staden's method assigns 1.300% false positives, where the separate network and the joint method assigns 0.950% and 0.162%, respectively. For 95% detection of true acceptor sites the corresponding

percentages are 1.986%, 1.606% and 0.397%.

With the joint assignment method more than a fifth of the true donor sites and around one fourth of the true acceptor sites could be detected without accompaniment of any false positive predictions. Highly confident splice sites could not be isolated with either Staden's method or the separate splice site networks. A complementary relation between the confidence levels of the coding/non–coding and the separate splice site networks was observed, with many weak splice sites having sharp transitions in the coding/non–coding signal and many stronger splice sites having a more ill–defined transition between coding and non–coding. We have not utilised the constraint of a continued open reading frame in the selection of compatible splice sites although this may reduce the number of false positive predictions further. In the general case where information on the translation products is not available the combinatorial problem is hard and very vulnerable to frameshift errors in the sequencing.

The joint assignment method was tested on a new independent set of twenty–two sequences extracted from GenBank release 66.0, January 1991. These sequences were not present in release 62.0 and have not been available to us during development of the algorithm. The results were very close to the ones found for the original test set. On the new test set (127919 bp, 127 A/D sites) the results were a little better for the acceptor sites and slightly worse for the donor sites. Both correlation coefficients and percentages of false positive predictions had nearly the same behaviour as function of the originally chosen constants e_D, c_D and e_A, c_A as on Figures 18 and 19. The new test set was also screened for sequences similarities.

The difference in length distributions of initial, internal, and terminal exons suggests that if exons are recognised as a whole, then different mechanisms with different steric constraints may be involved in the recognition of the three types of exons. The non–random distribution of false positive 5' sites (possible cryptic sites) in the terminal exons might be of biological significance. The accumulation of predicted false positive 5' sites around a 400-450 bp distance from the 3' site can be compared with the demonstration by Robberson *et al.* (1990) of inhibition of splicing if the terminal exon was extended beyond 300 basepairs. The inhibition of splicing of the last intron was reversed if a 5' site was included in the extended terminal exon.

In conclusion, we cannot rule out the existence of different classes of splice sites or occasional interaction with cell– or developmental–stage specific factors in the splice site selection process. We also cannot rule out interference from particular secondary structures of the mRNA in specific cases, or models based on local competition between mutually exclusive splice sites or spliceosome assemblies. Consequently, with the network approach we extract general features of all the presented cases, although this leads to a splice site prediction that is not perfect. In the prediction of splice site candidates, therefore, a compromise between finding the maximum number of true splice sites and obtaining the minimal number of false positive splice sites has to be found. The balance of the compromise depends on the

purpose of the prediction. If the purpose is to predict the mRNA (and protein) primary structure of a newly identified gene of which the sequence of a genomic clone exists, the neural network can be used to predict a number of candidate splice sites that can be tested by PCR sequencing directly on the cellular mRNA. In this scenario obtaining the highest prediction rate of true sites is at premium. The false positive predictions can be discarded after testing by the PCR sequencing.

We have therefore provided a set of parameters that allow different compromises for the relation between true and false positives. For large values of the ratio between true sites detected and false sites assigned correlation coefficients in the interval (0.55-0.75) could be obtained, which to our knowledge is better than any other splice site prediction method described in the literature.

We thank Olaf Nielsen for comments on the manuscript. This work was supported in part by the Danish Natural Science Research Council under grants No. J.nr. 11-1173 and 11-8168. The trained NetGene program will be made freely available in versions running under DOS, VMS and UNIX. For instructions, send an e–mail to brunak@nbivax.nbi.dk, or use one of the addresses given above.

References

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Nørskov, L., Olsen, O.H. & Petersen, S.B. (1988). Protein Structure and Homology by Neural Networks. FEBS Letters **241**, 223-228.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Fredholm, H., Lautrup, B., & Petersen, S.B. (1990). A Novel Approach to prediction of the 3-dimensional structures of protein backbones by neural networks. FEBS Letters **261**, 43-46.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1990). Cleaning up gene databases. Nature **343**, 123.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1990). Neural Network Detects Errors in the Assignment of pre–mRNA Splice Sites. Nucl. Acids Res. **18**, 4797-4801.

Chabot, B., Black, D. L., LeMaster, D. M. & Steitz, J. A. (1985). The 3' Splice Site of Pre–Messenger RNA is Recognized by a Small Nuclear Ribonucleoprotein. Science **230**, 1344-1349.

Fickett, J. (1982). Recognition of Protein Coding Regions in DNA Sequences. Nucl. Acids Res. **10**, 5303-5318.

Furdon, P. J. & Kole, R. (1986). Inhibition of Splicing but not Cleavage at the 5' Splice Site by Truncating Human B–Globin Pre–mRNA. Proc. Natl. Acad. Sci. USA. **83**, 927-931.

Furdon, P. J. & Kole, R. (1988). The Length of the Downstream Exon and the Substitution of Specific Sequences Affect Pre–mRNA Splicing In Vitro. Mol. Cell. Biol. **8**, 860-866.

George, D.G., Barker, W.C. & Hunt, L.T. (1986). The Protein Identification Resource (PIR). Nucl. Acids. Res. **14**, 11-15.

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

Harris, N. L. & Senapathy, P. (1990). Distribution and Consensus of Branch Point Signals in Eucaryotic Genes: a Computerized Statistical Analysis. Nucl. Acids Res. **18**, 3015-3019.

Hawkins, J.D., (1988). A Survey on Intron and Exon Lengths. Nucl. Acids Res. **16**, 9893-9908.

Holley, L.H. & Karplus, M. (1989). Protein Secondary Structure Prediction With a Neural Network. Proc. Natl. Acad. Sci. **86**, 152-156.

Jacob, M. & Gallinaro, H. (1989). The 5' splice site: phylogenetic evolution and variable geometry of the association with U1RNA. Nucl. Acids Res. **17**, 2159-2180.

Keller, E. B. & Noon, W. A. (1984). Intron Splicing: A Conserved Internal Signal in Introns of Animal pre-mRNAs. Proc. Natl. Acad. Sci. **81**, 7417-7420.

Kishi, F., Maruyama, M., Tanizawa, Y., & Nakazawa, A. (1986). Isolation and Characterization of cDNA for Chicken Muscle Adenylate Kinase. J. Biol. Chem. **261**, 2942-2945.

Kneller, D. G, Cohen, F. E. & Langridge, R. (1990). Improvements in Protein Secondary Prediction by An Enhanced Neural Network. J. Mol. Biol. **214**, 171-182.

Krainer, A. R., Conway, G. C. & Kozak, D. (1990). The Essential Pre-mRNA Splicing Factor SF2 Influences 5' Splice Site Selection by Activating Proximal Sites. Cell **62**, 35-42.

Krainer, A. R. & Maniatis, T. (1985). Multiple factors including the small nuclear ribonucleoproteins U1 and U2 are necessary for pre-mRNA splicing in vitro. Cell **42**, 725-736.

Kullback, S. (1959). Statistics and Information Theory. New York, J. Wiley and Sons.

Kudo, M., Lida, Y. & Shimbo, M. (1987). Syntactic Pattern Analysis of 5' Splice Site Sequences of mRNA precursors in Higher Eucaryote Genes. CABIOS **3**, 319-324.

Kuhne, T., Wieringa, B., Reiser, J. & Weissmann, C. (1983). Evidence Against a Scanning Model for RNA Splicing. EMBO J. **2**, 727-733.

Lapedes, A., Barnes, C., Burks, C., Farber, R. & Sirotkin, K. (1988) "Application of Neural Networks and Other Machine Learning Algorithms to DNA Sequence Analysis." The Proceedings of the Interface Between Computation Science and Nucleic Acid Sequencing Workshop, Dec. 1988 in Santa Fe, New Mexico. Eds. G.I. Bell and T.G. Marr. Proceedings of the Santa Fe Institute vol. VII, 157-182, Addison-Wesley.

Mathews, B.W., (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. Biochim. Biophys. Acta **405**, 442-451.

Matsuura, S. *et al.* (1989). Human Adenylate Kinase Deficiency Associated with Hemolytic Anemia. J. Biol. Chem. **264**, 10148-10155.

Mengeritsky, G. & Smith, T. F. (1989). New Analytical Tool for Analysis of Splice Site Sequence Determinants. CABIOS **5**, 97-100.

Minsky, M. & Papert, S. (1969, 1988). Perceptrons. MIT Press, Cambridge, Massachusetts.

Mount, S. M., Petterson, I., Hinterberger, M., Karmas, A. & Steitz, J. A. (1983). The U1 Small Nuclear RNA-Protein Complex Selectively Binds a 5' Splice Site in Vitro. Cell **33**, 509-518.

Nakata, K., Kanehisa, M. & DeLisi, C. (1985). Prediction of splice junctions in mRNA sequences. Nucl. Acids Res. **13**, 5327-5340.

Nelson, K. K. & Green, M. R. (1988). Splice Site Selection and Ribonucleoprotein Complex Assembly During In Vitro Pre-mRNA Splicing. Genes & Dev. **2**, 319-329.

Ohshima, Y. & Gotoh, Y. (1987). Signals for the Selection of a Splice Site in Pre-mRNA. J. Mol. Biol. **195**, 247-259.

Orkin, S. H. (1987). The Molecular Basis of Blood Diseases. In Stanatoyannopoulos, G. *et al.* (eds) pp. 106-126, Saunders, Philadelphia.

Petersen, S.B., Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Fredholm, H. & Lautrup, B. (1990). Training neural networks to analyse biological sequences. Trends in Biotechnology, **11**, 304-308.

Reed, R. & Maniatis, T. (1986). A Role for Exon Sequences and Splice-Site Proximity in Splice Site Selection. Cell **46**, 681-690.

Robberson, B. L., Cote, G. J. & Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. **10**, 84-94.

Rosenblatt, F. (1962). Principles of Neurodynamics. Spartan Books. New York.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) In Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, Massachusetts, vol. I, pp. 318-362.

Sejnowski, T.J. & Rosenberg, C.R. (1987). Parallel Networks that Learn to Pronounce English Text. Complex Systems **1**, 145-168.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell System Tech. J. **27**, 379-423, 623-656.

Shapiro, M. B. & Senapathy, P. (1987). RNA Splice Junctions of Different Classes of Eucaryotes: Sequence Statistics and Functional Implications in Gene Expression. Nucl. Acids Res. **15**, 7155-7174.

Smith, C. W. J., Porro, E. B., Patton, J. G. & Nadal-Ginard, B. (1989). Scanning from an Independently Specified Branch Point Defines the 3' Splice Site of Mammalian Introns. Nature **342**, 243-247.

Staden, R. (1984). Computer methods to locate signals in nucleic acids sequences. Nucl. Acids Res. **12**, 505-519.

Staden, R. (1984). Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. Nucl. Acids Res. **12**, 551-567.

Staden, R. (1990). NIP (Nucleotide interpretation program) V1.0 May 1990.

Tazi, J., Alibert, C., Temsamani, J., Reveillaud, I., Cathaia, G., Brunel, C. & Jeanteur, P. (1986). A protein that specifically recognizes the 3' splice site of mammalian pre-mRNA introns is associated with a small nuclear ribonucleoprotein. Cell **47**, 755-766.

Qian, N. & Sejnowski, T.J. (1988). Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. J. Mol. Biol. **202**, 865-884.

von Heijne, G. (1987). Sequence Analysis in Molecular Biology — Treasure Trove or Trivial Pursuit. Academic Press.

Wieringa, B., Hofer, E. & Weissmann, C. (1984). A Minimal Intron Length but no Specific Internal Sequence is Required for Splicing the Large Rabbit B-globin Intron. Cell **37**, 915-925.

Zhuang, Y. & Weiner, A. M. (1986). A Compensatory Base Change in U1 snRNA Suppresses a 5' Splice Site Mutation. Cell **46**, 827-835.