PREDICTION OF HUMAN TRANSLATIONAL INITIATION SITES USING A MULTIPLE NEURAL NETWORK APPROACH

Christian Derst, Martin Reczko, Artemis Hatzigeorgiou

Synaptic, Ltd., Greece

Accepted in the final form: December 01, 2000

Abstract: A method for the identification of human translational initiation sites (hTIS) in genomic DNA sequences has been suggested. The immediate sequence context around start codons show an uneven distribution of nucleotides at a specific position relative to the start codon. Therefore we used a neural network approach to identify translational initiation sites with this sequence context (nucleotide positions -20 to +13) as well as with the prediction of coding regions following the start codons. With the immediate sequence context as the only information we get a correct prediction of about 79 %. To increase the accuracy of the method we split the sequence-pattern into several smaller subpatterns. Using this pattern as the training set in a multiple neural network approach we get an overall prediction improvement of 2 %. The network analysis shows that the nucleotide positions -4 to +5 relative to the start codon dominates the decision in a single network trained with the entire sequence, masking the influence of more distant nucleotide positions. We propose that in a multiple network approach the influence of different parts of the translational initiation sequence can be more accurately fitted.

Keywords: DNA analysis, pattern recognition, translation initiation sites recognition, neural networks.

1 Introduction

With the growing amount of uncharacterized DNA-sequences coming up with large-scale sequencing approaches, such as the human genome project, there is a need for the automated sequence analysis using powerful algorithms. The identification of genomic sequence structures like exon-intron boundaries, translational initiation and termination sites, promotor regions or coding sequences has become a major problem in the computational DNA-sequence analysis

The identification of start codons is of twofold interest. Beside the genomic DNA analysis mentioned above, it is also useful in the analysis of cDNA sequences, as a significant fraction of eucaryotic mRNA do not adhere to the first-AUG rule, with the 5'-nearest AUG-codon in the mRNA not being the site of initiation of translation. Eucaryotic translational initiation sites have been extensively examined by Marilyn Kozak (Kozak 1984, Kozak 1995). In her work M. Kozak showed, that the optimal sequence context for translational initiation involves an especially critical purine site at the -3 position and other nucleotide preferences in the regions around the start-codon, such as a Cytosine-preference at - 4, - 2 and - 1 or a Guanosine-preference at + 4 (with the start codon being nucleotide positions + 1 to + 3). In a suboptimal sequence context the 40 S ribosomal subunit - scanning the 5'-end of the mRNA- is able to bypass the first AUG-codon. Taken together the sequence data of known translational initiation regions of higher eucaryotes, a statistical sequence-matrix is obtained, often used for the prediction of start codons. On the other side, such a compilation-approach could not explain or predict the initiation from start codons with a suboptimal sequence context. Here the interrelation of the individual nucleotide positions in the initiation region may be of particular interest for the translational initiation to occur.

In this paper we describe a neural network approach for the prediction of translational initiation sites, with is probably more accurate for the identification of suboptimal sequence-context start codons, because of the ability of neural networks to find some general rules not obvious in more simple statistical analyses as sequence-matrices or consensus sequences.

Neural networks as powerful learning techniques already have a widespread use in genomic analysis. The prediction of intron slice sites (Brunak et al. 1990, Guigó et al. 1992), promotor regions (O'Neill, 1992, Demler and Zhou, 1991) or the distinction between coding and non-coding regions (Lapedes et al. 1990, Farber et al. 1992) are some examples of the successful use of connectionist methods in the DNA-sequence analysis. Stormo et al. (1982) have used neural network in the prediction of translational initiation sites in E. coli. In fact their use of the perceptron algorithm was the first application of an artificial neural network in genomic analysis.

2 Systems and methods

For the neural network simulation we use the Stuttgarter Neural Network Simulator Version 4.0 and a Silicon Graphics workstation.

2.1 Data

The data for the training and validation of the network was taken from the Translational Termination Database (TransTerm) available from the EMBL file server by anonymous FTP from FTP.EMBL-Heidelberg.de in the directory /pub/databases/transterm. As the data is more a supplement to a collection of translation termination sites, we further improved the dataset by removing duplicate entries and by removing entries without an ATG at the position of the start codon (although there are some non-AUG start codon existing, we found, that some of the entries were not real non-AUG start codons, therefore we removed all entries with non-AUG codons at the appropriate position). In addition 80 entries used as a test set were removed. In the end 2366 database entries were used for training and validation.

The test data was extracted from the Genbank database. As this data was used in other neural network analyses as well, several constrains were used for the extraction, ensuring that the extracted genbank-entry encodes for a entire gene, with genomic and cDNA sequences known (details of the extraction procedure will be published elsewhere). In addition it was shown that none of these 80 entries had a significant homology to any other entry in the test set, using the blast-algorithm of Altschul et al. (1990) and the algorithm of Needlmen and Wunsch (1970) as implemented in the GCG-program package of the Heidelberg Unix Sequence Analysis Resources (HUSAR).

ATG-triplet regions not involved in translational initiation were extracted randomly from human genomic DNA entries of the Genbank database, excluding the test set data as well as an interval of 20 nucleotides on each side of indicated start codons. This ensures that the counter-examples used for training and validation are real ATG-triplets, but no start codons. The relationship of coding to non-coding ATG-triplets in this counter-example set may reflect the ratio of coding versus non-coding regions in genomic DNA entries in the Genbank, with an overrepresentation of the coding sequences.

Unless otherwise stated, the training and the validation set consists of 1183 start codon regions and 1663 non-start ATG-triplet regions. The entire pattern includes 20 nucleotides 5' (numbering - 20 to - 01) and 10 nucleotide 3' (numbering + 04 to + 13) of the ATG-Triplet (positions + 1 to + 3). The frequency-differences of the individual nucleotides in the translational initiation pattern and counter example pattern are shown as diff(ATG)-values in Fig. 1.

Every nucleotide was encoded as a 4 neuron value (T - 1000, G - 0100, C - 0010, A - 0001). Output data was represented by a single output-neuron using binary coding, with 1 coding for translational initiation sites and 0 coding for sequences with non-start ATG-triplets at positions +1 to +3.

2.2 Neural Network Simulation

For the neural network simulation we used the implementations of Standard Backpropagation, Backpropagation Momentum, Quickprop and RProp in the Stuttgarter Neural Network Simulator (SNNS) as learning algorithms (we refer to standard neural network textbooks as Zell (1994) for a detailed description of the learning algorithms). To find the best learning algorithm we varied systematically the respective parameters of each learning algorithm. Although not the best learning algorithm, Standard Backpropagation with a learning rate η of 0.02 (specifying the step width of the gradient descent) and a $d_{\max}$ of 0.1 (as the maximum difference between the teaching value t_j and the output o_j of the output unit which is tolerated), was used as the standard learning algorithm to optimize the network topology, training pattern combination and in the study of partial- and deletion-networks.

The networks usually had 1 input-layer with 132 neurons (when trained with the entire sequence-pattern), 1 hidden-layer (with 1 to 20 neurons) and 1 output-unit. More than 1 hidden-layer and shortcut connections do not improve the prediction rate (data not shown).

pp/np as the ratio of translational initiation sites versus non-start ATG-triplets in the training pattern, has been varied between 0.7 and 1.55. As already mentioned above, in standard training the lowest pp/np-value (0.7) was used, because in the analysis of genomic DNA non-start ATG-triplets are much more frequent than start codons. On the other side with this value one still gets an acceptable amount of start-codons predicted. The training was performed until the standard error for the output-unit in the validation reached a minimum.

To examine the role of particular nucleotide positions around the ATG-codon we trained the network with a shortened input-pattern. In the case of deletion networks the neurons of single nucleotide positions were excluded from the training set, whereas in the partial networks the neurons of whole nucleotide regions were deleted. The later partial networks were the basis for the construction of a multiple neural network. In such a multiple neural network approach up to 6 partial networks were individually trained and the weighted output of all 6 output-units were used for the prediction of human translational initiation regions. The weights of the output-units were systematically varied to reveal the best network combination for the prediction.

2.3 Network analysis

To evaluate some prediction rules of the trained single-pattern networks, we introduced 3 values which may give us some clues about the importance of each nucleotide position, as well as the context-dependencies of different nucleotide positions on the prediction.

First we estimated a ω_i-value for each position as

$$\omega_i = \sum_{h=1}^{H} w_{ih} \cdot w_{ho}$$

where H is the number of hidden-units in a neural network with 1 hidden-layer (usually $H = 6$), w_{ih} and w_{ho} are the weights of the links between input- and hidden-units or between hidden- and output-unit respectively. As a particular input-unit reflects a specific nucleotide (A, C, G or T) at a specific nucleotide position, positive ω_i-values show a positive influence on the prediction of human translational initiation sites, whereas negative values disfavors a positive prediction.

The second value, ζ_i, as

$$\zeta_i = \sum_{h=1}^{H} |w_{ih} \cdot w_{ho}|$$

revealed to us some hints about the importance of each input-neuron on the made prediction. Without regarding the sign of each link, we just sum the absolute weights for each input-neuron. Higher ζ_i-values indicate that the appropriate input-unit has a bigger influence on the prediction than input-neurons with lower ζ_i-values.

In a multi-layered neural network several connections between a particular input-unit and the output-unit(s) exists. As some of these connections may have a positive overall weight $(w_{ih} \cdot w_{ho})$, whereas others have a negative, we introduced a third parameter ψ_i as

$$\psi_i = \xi_i - |\omega_i|$$

If all overall weights of the connections of an input-unit have the same sign, than $\zeta_i = |w_i|$ and $\psi_i = 0$, with all links between input- and output-unit having the same influence on the prediction. In the other case, where some overall weights are positive and some are negative, we get $\zeta_i > |w_i|$ and $\psi_i > 0$. This case can be interpreted in the way that the appropriate input-unit has a context-dependent influence on the prediction, as with the input of all other input-neurons, the influence of a particular link of an input-unit could become bigger than other links. On the other side ψ_i-values > 0 and very low ζ_i-values could also be interpreted as "having no influence at all" (see below). The validity of the estimated ψ_i- and ζ_i-values were tested by a correlation of these values with the diff(ATG)-value of a particular nucleotide positions in the training pattern (see Fig. 6). For each nucleotide position the ζ_i- and ψ_i-values of the 4 input-neurons can be taken together to give us ζ_n- and ψ_n-values. These values are a nucleotide-type independent measure of the importance and the context-dependencies of the different nucleotide positions.

3 Results and discussion

3.1 Single-pattern networks

Various network topologies and learning algorithms were tested, to find the best prediction of hTISs. Network topologies with 4 to 6 hidden units (600 to 800 total trained links) in just one hidden layer were shown to give the best result. More than 6 hidden-units do not give rise to a better prediction, but instead result in a prolonged training. Fewer hidden-units lead to a significantly worse prediction, although the overall prediction-improvement with hidden layers is rather small (about 2 %). This shows, that in a single network the context dependencies are negligeable, because even in the absence of a hidden layer, without any crosstalk between the input-neurons, there is still a correct prediction of 76%. The variation of the pp/np-ratio of the training pattern revealed, that the overall prediction rate is nearly constant in the ratio-range tested (pp/np = 0.6 - 1.4).

In Table 1 the best results with the 4 tested learning algorithms are shown. Although the prediction-differences of the different learning algorithms are rather small, Quickprop significantly performed better than other algorithms. In all cases tested, the best prediction of a single-pattern neural network never exceeds 79.0 %. This may be due to the limitations of the training pattern to a 30 nucleotide context around the start codon. In some cases therefore additional information, like the upstream primer sequences or coding regions downstream, are missing. On the other hand we have to keep in mind the possibility that there are several wrong start-codons indicated in the database, as well as that there may be some unrecognized real start-codons in our counter-example ATG-Triplets. This limitation of the current database entries is still a matter of debate of todays work with genetic algorithms as many (if not most) of the characteristics of the database-entries are not based on experimental results. From the connectionists point of view, there are some limitations in the pattern characteristics, as earlier results as well as the weights of the links in our trained networks revealed that some nucleotide positions are much more important for the prediction than others. This may lead to topological problems during the training of the network, where subtle but important information for the optimal hTIS-prediction may be masked by the dominating influence of the neurons of other nucleotide positions.

learning algorithm	parameter	best prediction
Std. Backprop	$\eta = 0.02; dmax = 0.1$	78.57 %
Backprop Momentum	$\eta = 0.01; \mu = 0.3 \; ; dmax = 0.1$	77.29 %
Quickprop	$\eta = 0.005; \mu = 1.2; \nu = 0.0001; dmax = 0.05$	78.75 %
RProp	$\delta_0 = 0.02; \delta_{\max} = 40; \alpha = 4$	78.29 %

Table 1:

nucleotide pos.	best prediction	best predicted positiv pat.	best wrong pred. negativ pat.
entire pattern	78.57%	78.43%	20.42%
$-20 \to -01$	74.58%	78.17%	26.50%
$-10 \to -01$	74.46%	75.92%	23.08%
$-05 \to -01$	75.00%	78.58%	24.67%
$-03 \to -01$	74.08%	77.58%	29.42%
$-02 \to -01$	67.75%	80.42%	32.68%
-01	61.88%	68.67%	44.92%
$-20 \to -02$	73.46%	76.25%	23.08%
$-20 \to -03$	72.92%	77.50%	27.25%
$-20 \to -04$	63.96%	64.08%	27.58%
$-20 \to -05$	62.96%	62.33%	26.50%
$-20 \to -06$	62.29%	66.00%	32.92%
$-20 \to -11$	60.08%	69.50%	30.25%
$-20 \to -16$	59.38%	65.58%	34.08%
$-20 \to -18$	55.92%	73.33%	30.17%
$+04 \to +13$	63.75%	67.00%	30.92%
$+04 \to +08$	62.79%	65.17%	32.67%
$+04 \to +06$	61.79%	62.67%	36.25%
$+04 \to +05$	61.75%	64.08%	28.25%
$+04$	59.54%	71.00%	51.92%
$+05 \to +13$	61.54%	69.17%	31.00%
$+06 \to +13$	60.17%	64.50%	33.58%
$+08 \to +13$	58.04%	69.92%	29.25%
$-04 \to +05$	76.11 %	69.32 %	14.85 %

Table 2:

nr.	PN1	PN2	PN3	PN4	PN5	PN6	pred.	correct	correct
							(%)	pos.pat.	neg.pat.
01	0.5	–	–	–	–	0.5	78.04	947	926
02	0.6	–	–	–	–	0.4	78.42	955	927
03	0.7	–	–	–	–	0.3	77.42	952	906
04	–	0.2	0.4	–	–	0.4	79.62	955	956
05	–	0.3	0.3	–	–	0.4	79.54	942	967
06	–	0.4	0.2	–	–	0.4	77.62	900	963
07	–	–	–	0.2	0.4	0.4	79.50	948	960
08	–	–	–	0.3	0.3	0.4	79.96	954	965
09	–	–	–	0.4	0.2	0.4	79.17	942	958
10	–	–	–	0.3	0.4	0.3	80.75	956	982
11	–	–	–	0.35	0.35	0.3	80.83	959	981
12	0.2	–	–	0.25	0.25	0.3	79.67	953	959
13	0.2	0.2	0.2	–	–	0.4	79.83	957	959
14	–	0.1	0.1	0.25	0.25	0.3	80.58	951	983
15	0.1	0.1	0.1	0.25	0.25	0.2	79.17	945	955
16	0.1	0.1	0.1	0.2	0.2	0.3	80.17	956	968

Table 3:

nucleotide-type	qw0	m	R
A	$+0.7818$	0.2103	0.95
C	-0.8781	0.1654	0.91
G	-0.3587	0.1611	0.86
T	$+1.0808$	0.2753	0.88
A, C, G, T	-0.1103	0.1332	0.84

Table 4:

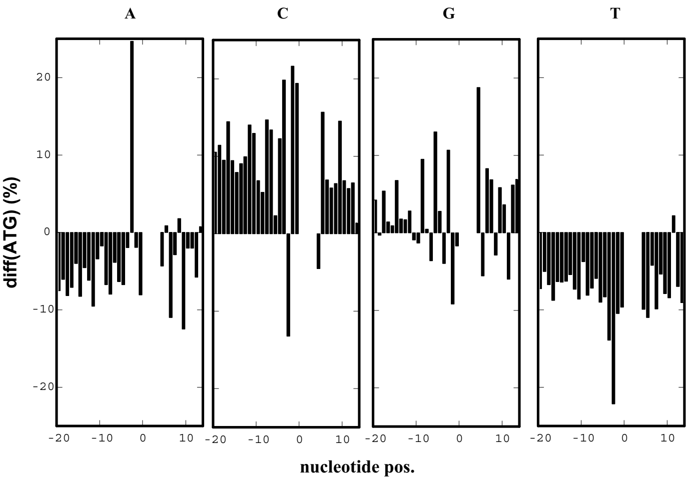

Figure 1: diff(ATG)-values, as the frequency-differences at the individual nucleotide-positions in start-codon regions versus non-start-codon regions, of the training pattern.

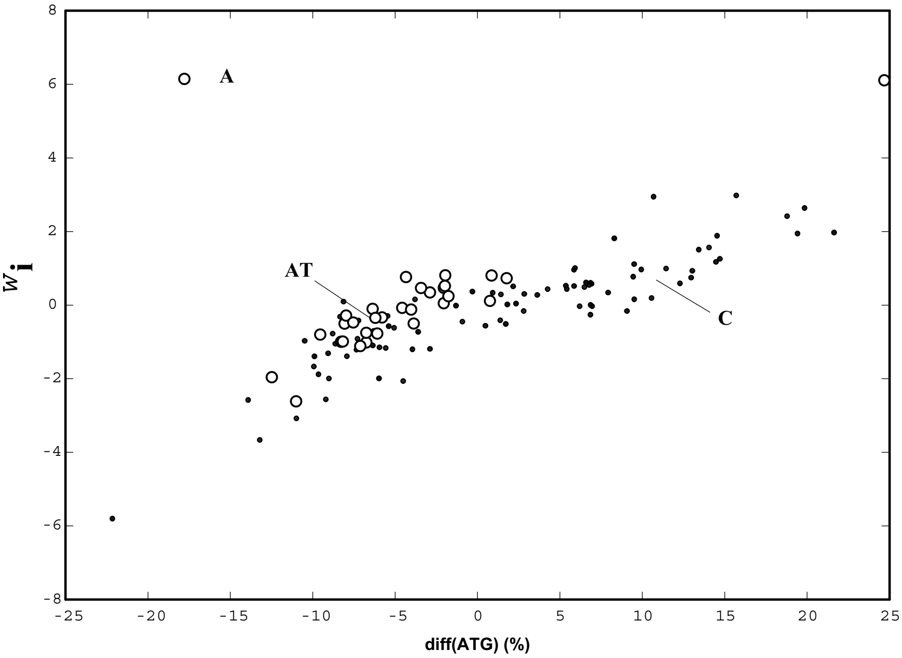

Figure 6: Correlation of ω_i and diff(ATG)-values of the single-pattern neural network of figures 2 and 3. Regions for the datapoints for adenosine and thymidine (AT)-nucleotides and for the cytosine-nucleotides (C) are indicated. In addition adenosine-datapoints are marked by a circle. Notice that the adenosine at position -3 give rise to the right- and upper-most datapoint in the correlation.

3.2 Deletion and Partial Neural Networks

To uncover the influence of special nucleotide positions we constructed deletion networks (exclusion of the neurons of a single nucleotide position) and partial networks (exclusion of the information

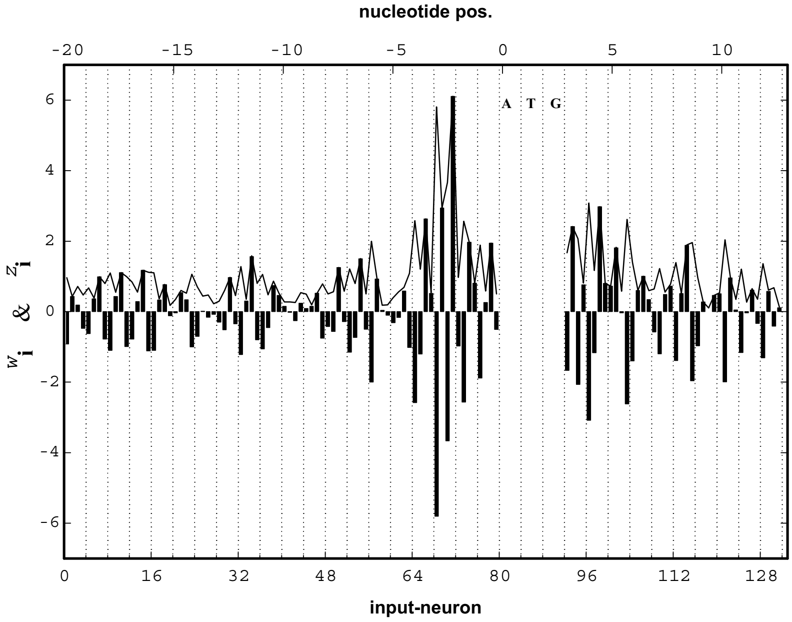

Figure 2: ω_i and z_i- values of a typically trained single-pattern neural network (backpropagation: $\eta = 0.02$; $d_{\max} = 0.1$).

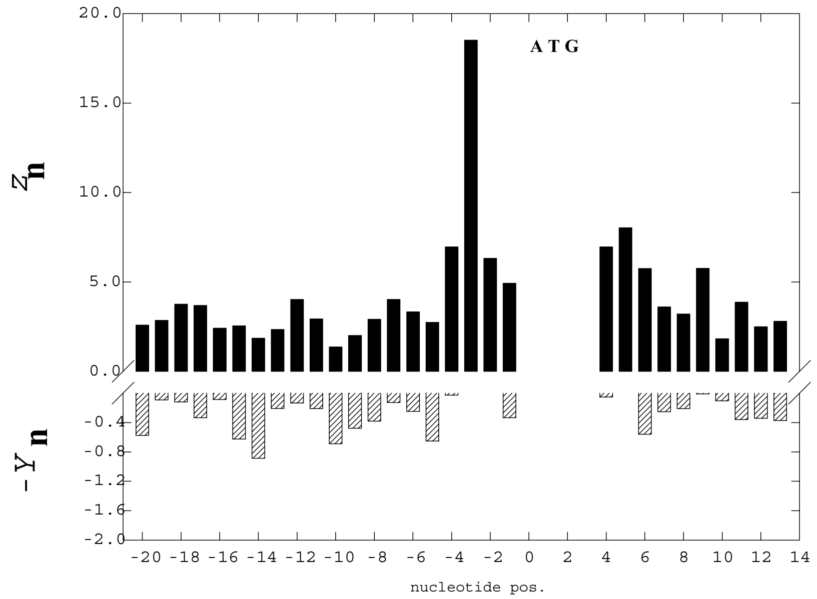

Figure 3: z_n and ψ_n -values of the same single-pattern neural network as in Fig.2.

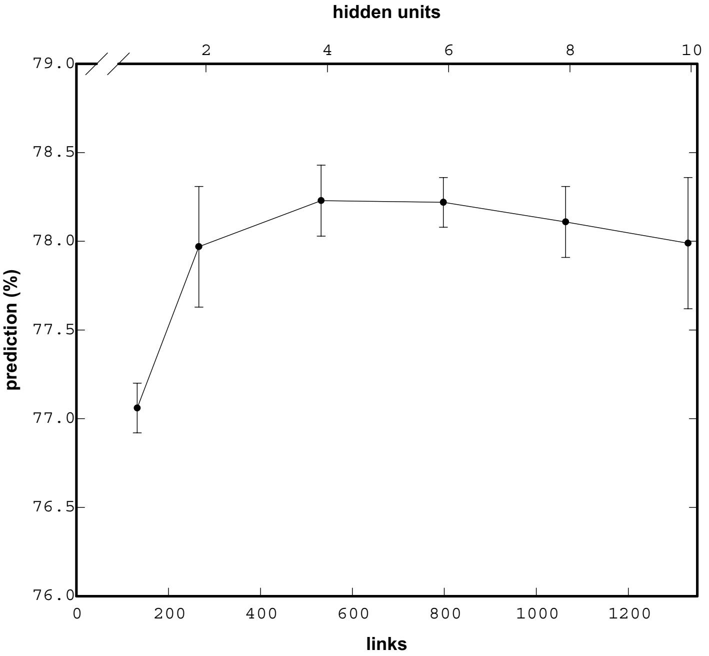

Figure 4: Overall correct prediction in single-pattern neural networks with 1 hidden layer but different numbers of hidden unit/links. For each datapoint more then 5 different neural networks were tested.

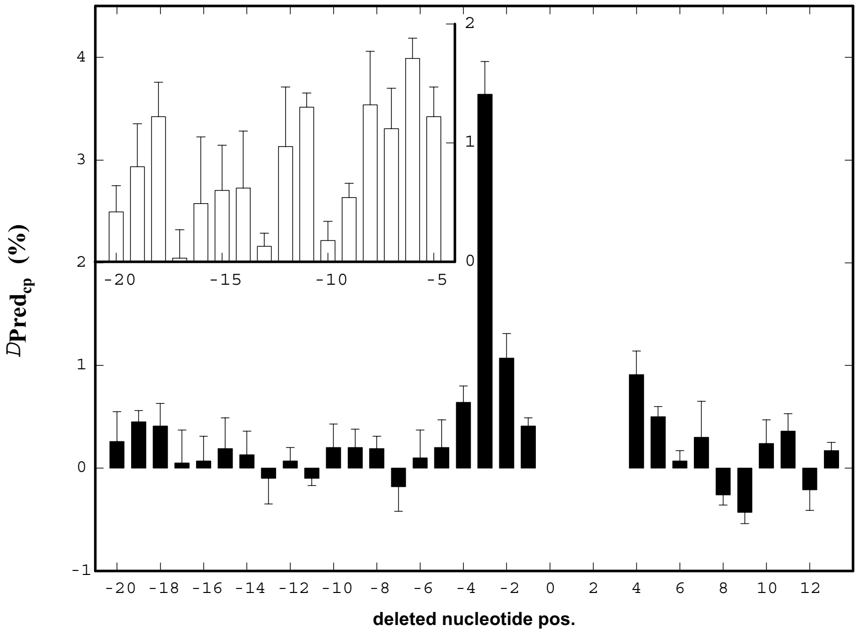

Figure 5: Prediction difference ($\Delta\,\mathrm{Pr}\,ed_{cp}$) between the differnt deletion networks and networks trained with the entire pattern. Small inner figure: The same prediction difference for the -20 to -5 region partial networks and the resulting deletion networks thereof.

from several nucleotide positions) and trained them as described above. The prediction-results of trained deletion- and partial networks are shown in Fig.5 and Table 2. Again in agreement with the Kozak sequence the nucleotide positions in the direct neighborhood of a potential start-codon are most important for the prediction. Very striking is, that only 3 nucleotides on the 5'-side of an ATG-triplet (positions -3 to -1) are enough to give a 72.5 % correct prediction. In addition a network trained with 5 nucleotide positions (-3 to -1, +4 and +5) gives us a prediction only slightly lower than with networks trained with the entire pattern. This data again shows, that the influence of the immediate sequence context around a ATG-Triplet dominates in the hTIS-prediction. On the other hand excluding these dominating regions (for example in a neural network trained with the nucleotide positions -20 to -4) still gives us a prediction significantly better than just by chance (about 62.5 % in this example). The networks trained with pattern not involving the dominating 5 nucleotide positions usually need a prolonged training time with comparable training parameters (more than twice as much training cycles) to reach the prediction optimum.

3.3 Single-Pattern Neural Network Analysis

To reveal some clues about how the decision of the neural networks is made, we analyzed the trained networks as described above. Once more the nucleotide positions in the direct neighborhood to a ATG-triplet are the most important ones defining a hTIS. In Fig.4 the ω_i- and ζ_i-values of a typical trained single-pattern neural network (Backpropagation, $\eta = 0.02$, $d_{\max} = 0.1$) are shown, in Fig. 5 the ζ_n-values of the same network are presented. The purinic site at position -3 seemed to be the major determinant of hTIS, indicated by the very high ζ_i- and ζ_n-values respectively, with a guanosine or a adenosine strongly favoring a sequence to be a hTIS. The positions -4, -2, -1, +4, and +5 also have a strong influence on the decision of the neural network. In the -4 to +5 region all bases at a particular nucleotide position influence the decision as described by M. Kozak. For example a cytosine at position -4, -2 or -1 favors the prediction of a hTIS, whereas at +4 a cytosine may indicate a non-hTIS ATG-triplet. Therefore the weights of the networks connection in the region -4 to +5 are a direct refection of the biological data. Outside the -4 to +5 region the influence of the appropriate nucleotide positions on the decision is significantly smaller, as indicated by the lower ζ_n-values. The prediction may be influenced by the low AT-content in hTIS, with adenosine and thymidine disfavoring a hTIS. Interestingly as seen in Fig. 5 the ω_n-values of nucleotides outside the -4 to +5 region are significantly higher than those for the nucleotides inside this region. As the input of these neurons may influence the prediction of hTIS in a context-dependent manner, these regions seemed to be necessary for predictions above 80 %. In addition the low AT-content is maybe the reason for some different representations of the bases in the neural network. As could be seen in Fig. 6 and Table 4, the correlation of the ω_i- and the diff(ATG)-values is much better for a particular nucleotide-type (for example for all guanosines), than for all nucleotides together. It should also be noted that even at nucleotide positions with a higher AT-content the particular ω_i-value fits the base-type dependent correlation extraordinarily well, as seen with the adenosine at position -3.

3.4 Multiple Neural Networks

The results of the partial networks and the analysis of the single pattern networks leads us to the construction of multiple neural networks where several partial networks are connected by weighted links. Some examples of chosen partial networks, weights and prediction results are shown in Table 3. Surprisingly the prediction of the best weighted multiple network is about 1.5% better than the best single-pattern network. This result clearly shows that the information of several parts of the sequence-context around a start-codon has to be distributed over several networks with independent training to give the best prediction, as distinct parts of the sequence-context may have different demands on the learning algorithms, parameters and the number of training cycles.

References

Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J. (1990) Basic Local Alignment Search Tool. J. Mol. Biol., 215, 403-410

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

Guigó R., Knudsen S., Drake N., Smith T. (1992) Prediction of Gene Structure. J. Mol. Biol., 226:141-157

Kozak M. (1984) Compilation and analysis of sequences upstream from translational start site in eukaryotic mRNAs. Nucleic Acids Res., 12, 857-872

Kozak M. (1995) Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. Proc. Natl. Acad. Sci. USA, 92, 2662-2666

Lapedes A., Barnes C., Burks C., Farber C., Sirotkin K. (1990) Application of neural networks and other machine learning algorithms to DNA sequence analysis. In Computers and DNA: SFI Studies in the Science of Complexity (Bells G., Marr T. Eds.), Addison-Wesley, Reading, MA, 7, pp. 157-182

Needleman S. B., Wunsch C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., 48, 443-453

O'Neill M. C. (1991) Training back-propagation neural networks to define and detect DNA-binding sites. Nucleic Acids Res., 19, 313-318

Stormo G. D., Schneider T. D., Gold L., Ehrenfeucht A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli., Nucleic Acids Res., 10, 2997-3011