Investigations of *Escherichia coli* Promoter Sequences With Artificial Neural Networks: New Signals Discovered Upstream of the Transcriptional Startpoint

Anders Gorm Pedersen* and **Jacob Engelbrecht†**

Center for Biological Sequence Analysis
The Technical University of Denmark, Building 206
DK-2800 Lyngby, Denmark

Abstract

In this paper we present a novel method for using the learning ability of a neural network as a measure of information in local regions of input data. Using the method to analyze *Escherichia coli* promoters, we discover all previously described signals, and furthermore find new signals that are regularly spaced along the promoter region. The spacing of all signals correspond to the helical periodicity of DNA, meaning that the signals are all present on the same face of the DNA helix in the promoter region. This is consistent with a model where the RNA polymerase contacts the promoter on one side of the DNA, and suggests that the regions important for promoter recognition may include more positions on the DNA than usually assumed. We furthermore analyze the *E.coli* promoters by calculating the Kullback Leibler distance, and by constructing sequence logos.

Key words: neural networks, information theory, DNA sequence analysis, *Escherichia coli*, promoters.

Introduction

Initiation of transcription is the first step in gene expression, and constitutes an important point of control in the bacterium *E.coli* (for a review see *e.g.*, (Reznikoff *et al.* 1985)). The initiation event takes place when RNA polymerase—which is the enzyme that catalyzes production of RNA from the DNA template—recognizes and binds to certain DNA sequences termed promoters. The structure that consists of RNA polymerase bound to native, double stranded DNA is known as the closed complex. Subsequent to the formation of the closed complex, a stretch of approximately 10 bp[1] is opened, yielding the open complex, and the polymerase proceeds to initiate RNA chain synthesis by the process of complementary base pairing. It is the sequence of the promoter that determines the precise site and orientation of the 5' end of the mRNA. The sequence is also an important element in determining the frequency of initiation.

In *E.coli* the form of RNA polymerase that is responsible for initiating transcription has the subunit composition $\alpha_2\beta\beta'\sigma$. This so-called holoenzyme can be divided into two functional components: the core enzyme ($\alpha_2\beta\beta'$, also designated E) and the sigma factor (the σ polypeptide). The sigma factor plays an important role in recognizing promoter sequences, and after successful initiation it is released from the holoenzyme (Gross & Lonetto 1992). Several different sigma factors exist, and each type recognizes a specific subset of promoters with distinct nucleotide sequences. In *E.coli* the large majority of promoters are recognized by the holoenzyme Eσ^{70}, whose sigma factor (σ^{70}) is named according to the molecular weight of the protein (70 kDa).

Comparison of *E.coli* promoters has led to the identification of three major conserved features: the "−10 box", the "−35 box", and a purine (A or G) at the initiation site (Rosenberg & Court 1979; Hawley & McClure 1983). The −10 and −35 boxes are conserved hexanucleotide elements that are named according to the approximate position of their central nucleotides relative to the transcriptional start point. The consensus sequences are TTGACA for the −35 box, and TATAAT for the −10 box. Genetical and biochemical studies have demonstrated that there is a good correlation between the proposed consensus sequences and the effect of mutations in promoters, *i.e.*, mutations that result in a promoter that is less identical to the consensus leads to lower levels of transcription, while mutations that cause a promoter to be more like the consensus has the opposite effect. The newer and more extensive compilations of *E.coli* promoters have supported this view of *E.coli* promoters, by suc-

* Phone: (+45) 45 25 24 71, Fax: (+45) 45 93 48 08, E-mail: gorm@cbs.dtu.dk

† Phone: (+45) 45 25 24 78, E-mail: engel@cbs.dtu.dk

[1] Abbreviations used: bp, basepair; RNA, ribonucleic acid; DNA, deoxyribonucleic acid; mRNA, messenger RNA

cessfully aligning newly sequenced promoters against the consensus sequences (Harley & Reynolds 1987; Lisser & Margalit 1993). It should, however, be noted that this procedure is likely to give a picture of *E.coli* promoters that is biased towards the originally discovered consensus patterns.

In this paper we present an analysis of many of the *E.coli* promoter sequences known today. Specifically, we have been interested in finding areas in the promoters that contain information which correlates with the presence of transcriptional start points. For this purpose we have used the database of *E.coli* promoter sequences (with experimentally determined transcriptional start points) that was compiled by Lisser and Margalit (Lisser & Margalit 1993). In addition to determining the information content in the promoters by the Kullback Leibler measure and by making sequence logos, we present a novel method for using the learning capability of an artificial neural network as a measure of information. In order to ensure the unbiasedness of the analysis, we presented only the experimentally determined promoter characteristics (*i.e.*, the location of the transcriptional start point) to the networks. This is unlike previous studies, where the concept of the −35 and −10 boxes have in some form been included in the data—*e.g.*, by training the network on aligned hexanucleotides (Lukashin *et al.* 1989; Demeler & Zhou 1991; O'Neill 1991; 1992; Horton & Kanehisa 1992; Mahadevan & Ghosh 1994). Additionally, our approach differs from those previously used, in that we try to predict whether or not a given nucleotide is a transcriptional start point, while previous studies have focused on predicting whether a given sequence is a promoter or not. It is important to note that it is not our goal to construct a neural network that has a better prediction ability than previous networks. Rather, our goal is to use the learning abilities of neural networks as a measure of the information content in various parts of the promoter sequences.

The soundness of our approach is demonstrated by the fact that we find the previously discovered sequence signals at −10 and −35. However, we also discover hitherto unrecognized signals, that are correlated with the presence of a transcriptional start point. These signals are regularly spaced along the upstream region of the promoter: there is approximately one turn of the DNA helix (10.5 bp) between the central nucleotide in each signal, consistent with a model where the RNA polymerase contacts the promoter sequence mainly on one face of the DNA helix.

Methods

Data

The promoter sequences were taken from the compilation by Lisser and Margalit (Lisser & Margalit 1993). This database, which contains 300 sequences, is superior to most other available *E.coli* promoter databases on two accounts:

- Each sequence has been compared to the original paper, minimizing the chance of database entry errors.
- For each sequence, the assignment of transcriptional start point(s) has been verified with the relevant papers, and the most reliable have been chosen.

We processed the data in the following ways: first, we concatenated the sequences that are partially overlapping (*e.g.*, dnaK-P1 and dnaK-P2). This removed a number of contradictions, since the nucleotide that is marked as a transcriptional start point in one sequence is not labeled as such in the partially overlapping sequence, and vice versa. Concatenation resulted in a subset consisting of 248 sequences. Second, we discarded all the sequences that contain multiple start points. The resulting set, which we use in this study, contains 167 sequences. No division of the data set was performed based on knowledge about which sigma factor transcribes the gene. One reason for this is that for many genes it has not been determined experimentally which sigma factor is responsible for promoter recognition.

In the experiments where a training set and a test set were needed, the subsets were randomly divided into two parts: the training set contained 134 sequences (80%), while the test set contained the remaining 33 sequences. In order to ensure that the specific choice of test set versus training set was not important for the performance of the neural networks, the random divisions were performed in 2 different ways, and the results compared. The results obtained in this way were qualitatively identical.

Negative examples were constructed by choosing sequences from the promoters themselves rather than by constructing random sequence, or by using coding regions of *E.coli* genes as it has been done in previous studies (Lukashin *et al.* 1989; Demeler & Zhou 1991; O'Neill 1991; 1992; Horton & Kanehisa 1992; Mahadevan & Ghosh 1994). We believe that the problem of discriminating between promoter and non-promoter DNA in a promoter-containing region is closer to the biological task encountered in the cell by the RNA polymerase. Furthermore, the possibility exists that a network trained against random sequence or coding regions, actually learns to discriminate between secondary characteristics of the sequences, rather than to

predict transcriptional start points in a natural context. Specifically, training and test examples were constructed by sliding a window over the promoter sequences such that each positive example was preceded and followed by several negative examples from the same promoter. (However, it should be noted that during training, the network is presented to these windows in random order). In order to avoid contradictions due to sequence signals that are not situated at a specific distance from the transcriptional start point, we excluded all negative examples that were shifted from the positive example by between 1 and 5 nucleotides.

Information Measures

The Kullback Leibler distance (or relative entropy) was calculated by the formula:

$$D(i) = \sum_N P_i^N \log \frac{P_i^N}{Q_i^N}$$

where P_i^N and Q_i^N are the probabilities of occurrence for a particular nucleotide N (A, C, G, T) at position i (Kullback & Leibler 1951). The probability distribution P is here taken relative to transcriptional start points, while Q is taken relative to all other positions in the promoter. $D(i)$ has values that range from 0 to ∞. $D(i) = 0$ indicates that the two distributions are identical at position i (*i.e.*, there is totally average occurrence of nucleotides at position i relative to the transcriptional start point), while larger values of $D(i)$ means that the occurrence of a nucleotide at position i is different from the average.

Sequence logos were constructed according to Schneider and Stephens (Schneider & Stephens 1990). Briefly, sequence logos combine the information contained in consensus sequences with a quantitative measure of information, by representing each position in an alignment by a stack of letters. The height of the stack is a measure of the non-randomness at the position (here essentially the Shannon measure (Shannon 1948)), while the height of a letter corresponds to its relative occurrence.

Neural Networks

The neural networks used in this study were of the feed-forward type, and had three layers of neurons (Hertz, Krogh, & Palmer 1991). We implemented the neural network programs in the C programming language, and executed them on UNIX workstations.

Input values were obtained by encoding the DNA sequence into a binary string, using a coding scheme where each nucleotide is represented by 4 binary digits: A=0001, C=0010, G=0100, T=1000 (Brunak, Engelbrecht, & Knudsen 1991). It has been found that this

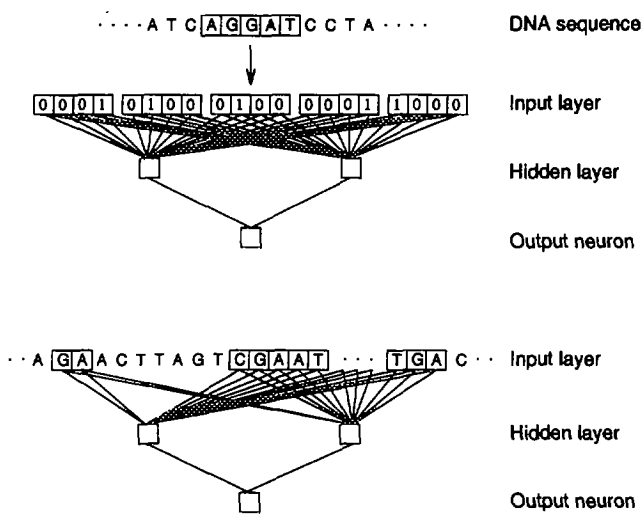

Figure 1: Overview of the two types of neural network architecture used in this study. The upper panel shows a type of network with a single input window, two hidden neurons, and one output neuron. In the example shown the network has an input window of 5 nucleotides, which are encoded as a string of $5 \cdot 4 = 20$ bits. The encoding scheme is: A=0001, C=0010, G=0100, T=1000. In this study, we used networks of this type with input windows of between 1 and 51 nucleotides (between 4 and 204 bits). The lower panel illustrates another type of network, which has an input window covering 65 nucleotides but with a 7 nucleotide hole (although not shown, the nucleotide input is in this case also encoded using the 4-bit scheme). In this study we used networks of this type where the position of the hole was varied.

leads to a significantly better performance than a more compact coding scheme (A=00, T=01, G=10, C=11), presumably due to the identical Hamming distances between the nucleotide encodings (Demeler & Zhou 1991). The output layer consisted of only one neuron, which decided whether the nucleotide at a given position was a transcription initiation site or not. Output values were in the range 0 to 1. During training, transcription initiation sites were encoded as 1.0, while non-initiation sites were encoded as 0.0. When evaluating the output, all values above 0.5 were interpreted as "transcription initiation site" while all values smaller than 0.5 were interpreted as "non-initiation site". Two main types of neural network architectures were used during this study (Figure 1). One type contained a single input window, the size of which was varied during a number of different training sessions. The other type had an input window covering 65 nucleotides except for a 7 nucleotide long hole, the position of which was varied during a number of different training sessions (Figure 1).

Initially the network weights and thresholds were as-

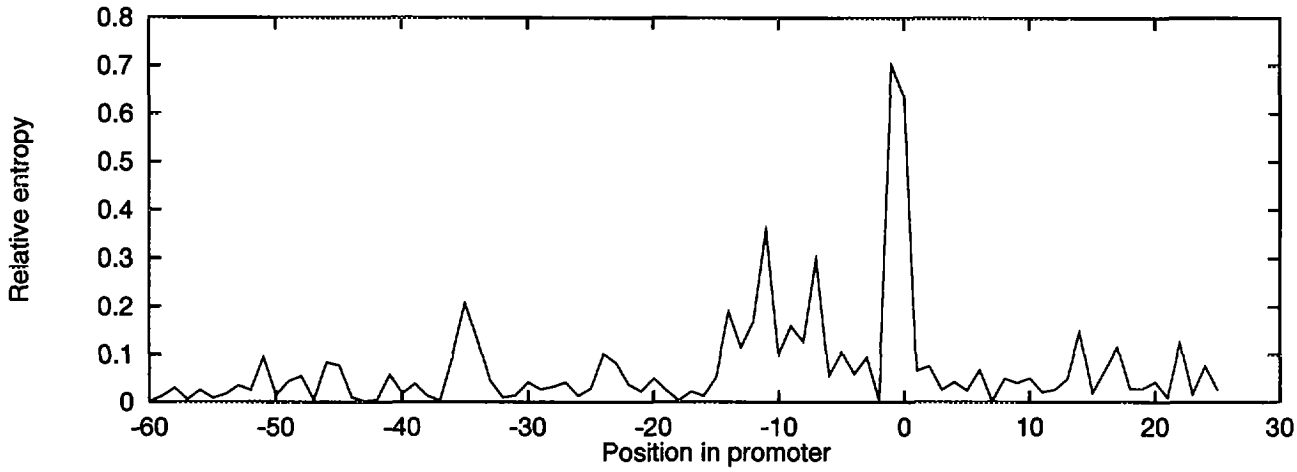

Figure 2: Kullback Leibler distances between the distribution of nucleotides relative to the transcriptional start points, and relative to all other positions in the dataset as a function of the position in the promoter (transcriptional initiation site=0). Notice the peaks around 0, −10, and −35, indicating that the occurrence of nucleotides in these regions is non-average.

signed random values in the range −0.3 to +0.3. Training of the weights and thresholds was performed using the backpropagation algorithm, and was carried out after each training example (online training) rather than after each training epoch (Hertz, Krogh, & Palmer 1991). Specifically, the error measure we used was

$$E = -\log(1-(O-T)^2)$$

where O is the actual output value, and T is the desired target value.

After each training epoch (*i.e.*, after all training examples have been presented to the network), the order of the training set was shuffled randomly. This helps the network avoid being trapped in local minima in the error landscape. The learning rate (η) was fixed at 0.02, which was found to give good performance for all the network architectures assessed in this study. Finally, we found that overtraining (which can be interpreted as memorization of the idiosyncrasies of the training set) could be efficiently reduced by performing backpropagation only when the output and target values were sufficiently different. The criterium we chose after having experimented with a range of thresholds, was that the absolute value of the difference between the output and the target should be above 0.1 for backpropagation to take place.

Evaluation of Results

When different configurations of the network are to be compared it is necessary to have a good measure of the network's performance. Merely comparing the percentage of true positives (*i.e.*, transcription initiation sites predicted as such) is not satisfactory: a network that assigns all nucleotides in a sequence as transcriptional initiation sites, will have a true positive ratio of 100%, but is obviously not of much use. A measure that takes all correctly and falsely predicted initiation sites and non-initiation sites into account, is the correlation coefficient

$$C = \frac{PN - P'N'}{\sqrt{(N+N')(N+P')(P+N')(P+P')}}$$

where P and N are the correctly predicted positives and negatives, and P' and N' are the falsely predicted positives and negatives (Mathews 1975). The correlation coefficient C is +1.0 for a perfect prediction, −1.0 for a completely imperfect prediction, and 0.0 when uncorrelated. In the special case when $N+N'=0$ or $P+P'=0$ (*i.e.*, all nucleotides are predicted to belong to the same category) C was taken to be 0.0.

Results and Discussion

Information Measures

The Kullback Leibler distance $D(i)$ was calculated for the 167 sequences in our data set (Figure 2). It is obvious that non-average areas are present around positions 0, −10, and −35. This is consistent with the positions of the conserved sequence elements described above. When $D(i)$ was calculated for the distributions of dinucleotides, the result was qualitatively identical.

Signals at 0, −10, and −35 can also be seen on the sequence logos (Figure 3). In agreement with previous studies, the signal at the initiation site was mainly caused by the dinucleotide CA (or more precisely: a pyrimidine followed by a purine). The signal at −10 can be seen to be a stretch of 6–8 A's or T's, in good

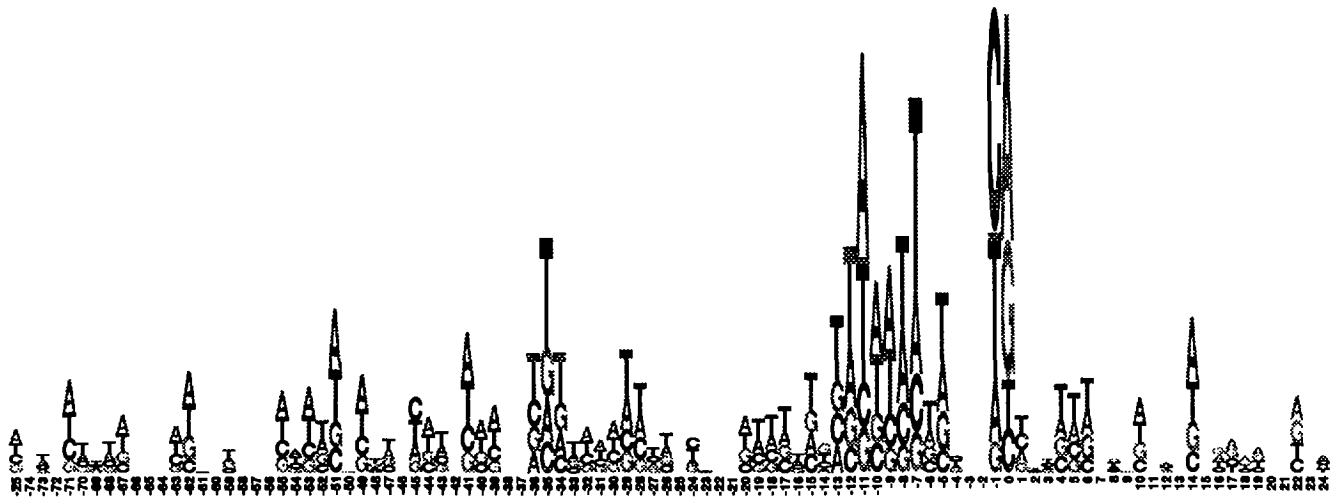

Figure 3: Sequence logos of the sequences in the data set. The sequences are aligned by their transcriptional initiation site (position 0). Notice that non-random areas are present around 0, −10 and −35, and that the sequences seen in this logo are in good agreement with previous results.

agreement with the previously determined consensus: TATAAT. No clear consensus emerges for the signal at −35 from this analysis. This can probably be explained in part by the fact that the position of the −35 box, relative to the transcriptional start point, is somewhat flexible. Consequently, the sequence will not be clearly recognized without alignment. However, another part of the reason for the unclear consensus might be that the −35 box is less well conserved than the −10 box (Galas, Eggert, & Waterman 1985; Harley & Reynolds 1987). It has been noted that in *E.coli* promoters the most frequently occurring three-letter word (found within a 7 bp window) is TTG, which is present in the −35 region (Galas, Eggert, & Waterman 1985). This can be seen to be in reasonable agreement with our data (Figure 3).

In conclusion, we find signals that correlate with the presence of a transcriptional start point using methods from information theory. The position, of these signals are consistent with the position of conserved sequence elements that have previously been described. Additionally, the predominant nucleotides, as determined by sequence logos, are in good agreement with the previously described consensus sequences.

Neural Networks

It is well known that neural networks are able to learn complex correlations when they are given suitable training examples (Brunak, Engelbrecht, & Knudsen 1990; 1991; Hertz, Krogh, & Palmer 1991). Hence, we reasoned that if the ability of a neural network to learn correlations from different input data is compared, it should be possible to determine the relative importance of the input data in the learning process. *I.e.*, the ability of a network to learn can be used as a relative measure of the information content in the input data. Thus, it might be expected that it is practically impossible for a neural network to learn to predict transcriptional start points if the input presented to the network is five nucleotides present at positions −537 to −533, whereas a network whose input is sequence around the initiation site is more likely to succeed.

The method we used to monitor the learning state of the network was to calculate the correlation coefficient C mentioned above. Specifically, we trained the networks until no improvement could be observed in the prediction of the training set. After each training epoch the test and training correlation coefficients were calculated, and after training was stopped the maximum test correlation coefficient obtained (C_{max}) was noted. For each input configuration, the network was trained five times, and the average of C_{max} obtained in this way was used as a measure of information in the input data. The C_{max} values obtained in different trainings on the same data were always very similar. After having investigated architectures with various numbers of neurons in the hidden layer, we decided to use the networks with 2 or 3 hidden neurons.

The first method we used was to make series of network trainings where the size of the input window was varied in one direction from a fixed point. Thus, we started with a window of 1 nucleotide at position +5 relative to the initiation site, and proceeded to make successive runs where the input window was increased

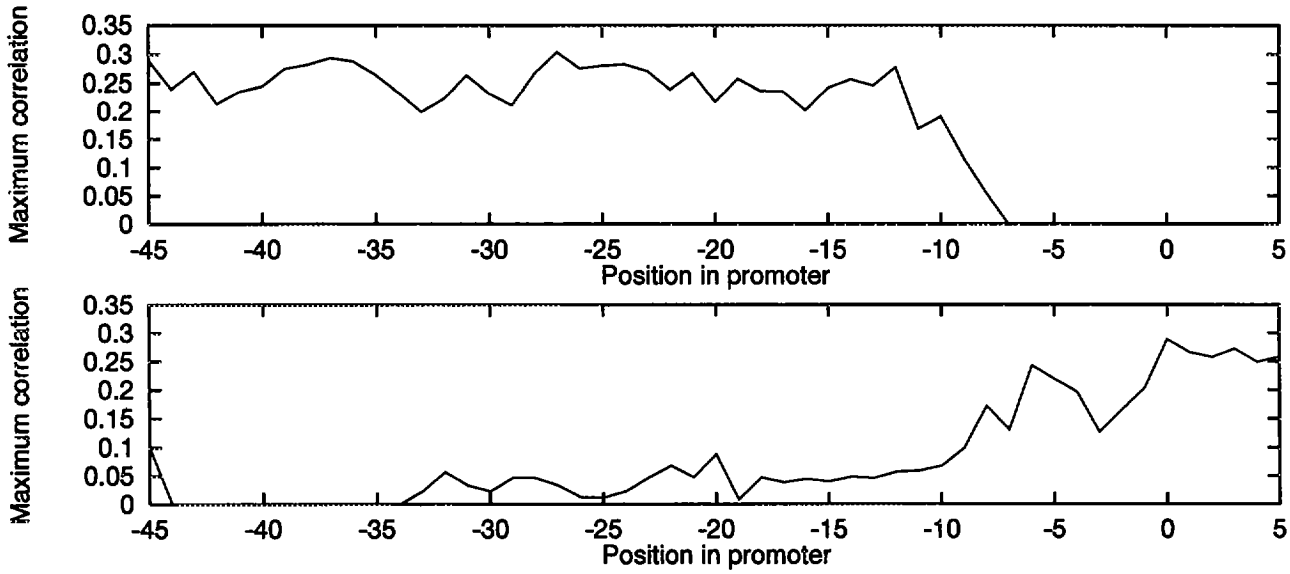

Figure 4: Varying the border of the neural network's input window in one direction. The maximum correlation coefficient C_{max} is shown as a function of the position of the moving border of the input window (transcriptional initiation site=0). The upper curve shows the results obtained by having the right border of the input window fixed at +5 relative to the transcriptional start point, and varying the left border from +5 to −45 in steps of one nucleotide. The lower curve shows the result when the left border is fixed at −45 and the right border is varied in the opposite direction. All C_{max} values are the average obtained after five runs with different (and random) start values for the network weights and thresholds.

in size by moving the left border in an upstream direction (Figure 4 upper panel). Following runs had windows covering positions +4 to +5, +3 to +5, and so on until the last run was performed with an input window covering from −45 to +5. We also performed network trainings where scanning was performed from the opposite direction (*i.e.*, the fixed border was −45, and the moving border was varied from −45 to +5, Figure 4 lower panel). We have previously used this method to identify a signal positioned downstream of transcriptional initiation sites in mammalian genes (Larsen, Engelbrecht, & Brunak 1995).

As it can be seen, the network is completely unable to generalize when the input window only covers sequence downstream of −7 (*i.e.*, between −7 and +5, Figure 4 upper panel). After the moving border has passed this position a steep increase in the learning ability can be observed. This indicates that important information is present around position −10, which is in perfect agreement with what was found above, and what has been demonstrated previously. However, no further increase is seen in C_{max} after the border has passed the −10 area (Figure 4 upper panel). When the scanning border is varied from the opposite direction, small peaks appear on the C_{max}-plot between −35 and −15 (Figure 4 lower panel). However, only after the moving border has passed the −10 area a significant increase in C_{max} can be observed. This again indicates that important information is present around −10. Furthermore, the small peaks suggest that regions further upstream are also important. The reason that these less important peaks are not visible when scanning from the opposite side, is presumably that the additional information relative to the −10 region is too small to be noticed.

In addition to the "border-scanning" technique, we developed a new method for finding information-containing areas: neural networks were presented with input windows that covered positions −55 to +9, but which also contained a 7 bp hole. In a series of runs, the position of the hole was shifted along the input window, and the maximum C values for the test set were determined. In this manner it should be possible to detect local regions with important information by looking for positions of the hole that causes the learning ability to be partly destroyed. Specifically, these positions can be seen as local minima in the plot of C_{max} versus the position of the hole. Compared to our other network method, this technique has the advantage that a constant number of weights and thresholds are used in the different networks. Therefore, differences in learning abilities should mainly reflect the difference between the information content of the inputs.

When scanning the input window with a hole, we found the same areas of high information content that was identified above. This is visible in the C_{max} plots

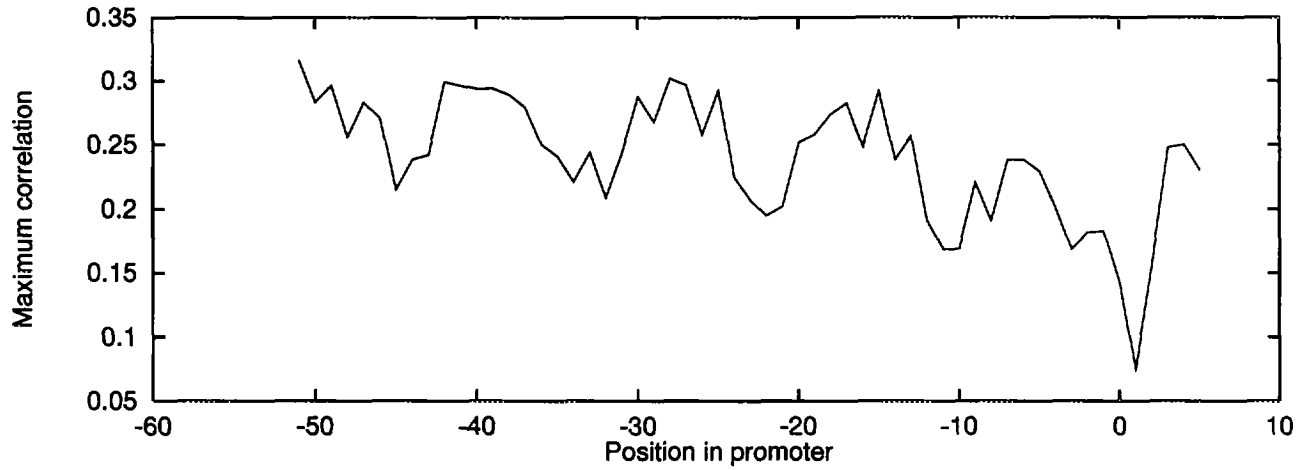

Figure 5: Scanning the input window of neural networks with a 7 nucleotide hole. The maximum correlation coefficient C_{max} is shown as a function of the position of the middle nucleotide in the hole (transcriptional initiation site=0). Notice the local minima around position 0, −10, −22, −33, and −44, which indicate that important information is present in these areas. All C_{max} values are the average obtained after five runs with different (and random) start values for the network weights and thresholds.

as clear local minima around 0, −10, and −35 (Figure 5). Interestingly, two additional minima can be seen around positions −22 and −44. The reason that these were not identified in the Kullback Leibler plots or sequence logos, could be because the problem of predicting transcription initiation sites is nonlinear. *I.e.*, the signals do not consist of conserved nucleotides at any single position, but rather of correlations between nucleotides at different positions. A signal at −44 has also been noticed in a previous study where *E.coli* promoter regions were searched for conserved words of different sizes (Galas, Eggert, & Waterman 1985), and in multiple alignment studies (Harley & Reynolds 1987), but to our knowledge, a signal at −22 has never been reported.

We find it very interesting that all the signals observed in the above analysis are spaced regularly along the promoter region with a period of 10–11 bp (positions of local minima: +1, −10, −22, −33, −44). This spacing corresponds to the helical periodicity of DNA, and the signals we have observed are therefore all present on the same face of the DNA helix. This is consistent with a model in which the RNA polymerase holoenzyme contacts the promoter on one face of the DNA, as experimental studies indicate that it does (*e.g.*, see (Siebenlist, Simpson, & Gilbert 1980)).

Concluding Remarks

In this paper we have described a novel method for using the learning ability of a neural network as a measure of information: neural networks are presented with different windows on the input data, and the maximum obtainable test set correlation coefficients (C_{max}) are compared. When local regions that contain important information are excluded from the input data, C_{max} will be lower than when the region is not excluded. Using a version of this method where the data is scanned with a hole in the input window, we have discovered what appears to be regularly spaced signals in the promoter region of *E.coli* genes. Specifically, the spacing we observe corresponds to the helical periodicity of B-form DNA (approx. 10.5 bp/turn), and therefore the signals are all present on the same face of the DNA helix. These results are in agreement with previous theoretical and experimental results, which show that RNA polymerase makes important contacts with DNA at the so-called −10 and −35 boxes, and that these are present on the same side of the DNA (Siebenlist, Simpson, & Gilbert 1980; Harley & Reynolds 1987). However, our results furthermore suggest that the regions important for promoter recognition may include more positions on the DNA than usually assumed. The credibility of our findings is supported by the fact that these additional regions are all situated on the same side of the DNA helix as the previously described −10 and −35 boxes. It will, however, be necessary to await experimental confirmation before any firm conclusions can be made.

Acknowledgements

We thank Dr. Lise Hoffmann for comments on the manuscript. Both authors were supported by a grant from the Danish National Research Foundation.

References

Brunak, S.; Engelbrecht, J.; and Knudsen, S. 1990. Neural network detects errors in the assignment of mRNA splice sites. *Nucleic Acids Res.* 18:4797–4801.

Brunak, S.; Engelbrecht, J.; and Knudsen, S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220:49–65.

Demeler, B., and Zhou, G. W. 1991. Neural network optimization for E. coli promoter prediction. *Nucleic Acids Res.* 19:1593–9.

Galas, D. J.; Eggert, M.; and Waterman, M. S. 1985. Rigorous pattern-recognition methods for DNA sequences. analysis of promoter sequences from Escherichia coli. *J. Mol. Biol.* 186:117–28.

Gross, C. A., and Lonetto, M. 1992. Bacterial sigma factors. In *Transcriptional regulation.* Cold Spring Harbor Laboratory Press.

Harley, C. B., and Reynolds, R. P. 1987. Analysis of E. coli promoter sequences. *Nucleic Acids Res.* 15:2343–61.

Hawley, D. K., and McClure, W. R. 1983. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res.* 11:2237–55.

Hertz, J.; Krogh, A.; and Palmer, R. G. 1991. *Introduction to the Theory of Neural Computation.* Studies in the Sciences of Complexity. Addison–Wesley Publishing Company.

Horton, P. B., and Kanehisa, M. 1992. An assessment of neural network and statistical approaches for prediction of E. coli promoter sites. *Nucleic Acids Res.* 20:4331–8.

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *Ann. Math. Stat.* 22:79–86.

Larsen, N. I.; Engelbrecht, J.; and Brunak, S. 1995. Analysis of eukaryotic promoter sequences reveals a sytematically occurring CT-signal. *Nucleic Acids Res.* In press.

Lisser, S., and Margalit, H. 1993. Compilation of E. coli mRNA promoter sequences. *Nucleic Acids Res.* 21:1507–16.

Lukashin, A. V.; Anshelevich, V. V.; Amirikyan, B. R.; Gragerov, A. I.; and Frank-Kamenetskii, M. D. 1989. Neural network models for promoter recognition. *J Biomol. Struct. Dyn.* 6:1123–33.

Mahadevan, I., and Ghosh, I. 1994. Analysis of E.coli promoter structures using neural networks. *Nucleic Acids Res.* 22:2158–2165.

Mathews, B. W. 1975. Comparison of predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta* 405:442–451.

O'Neill, M. C. 1991. Training back-propagation neural networks to define and detect DNA- binding sites. *Nucleic Acids Res.* 19:313–8.

O'Neill, M. C. 1992. Escherichia coli promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. *Nucleic Acids Res.* 20:3471–7.

Reznikoff, W. S.; Siegele, D. A.; Cowing, D. W.; and Gross, C. A. 1985. The regulation of transcription initiation in bacteria. *Annu. Rev. Genet.* 19:355–87.

Rosenberg, M., and Court, D. 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. *Annu. Rev. Genet.* 13:319–53.

Schneider, T. D., and Stephens, R. M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* 18:6097–6100.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Tech. J.* 27:379–423, 623–656.

Siebenlist, U.; Simpson, R. B.; and Gilbert, W. 1980. E. coli RNA polymerase interacts homologously with two different promoters. *Cell* 20:269–81.