A Neural Network System for Prediction of RNA Polymerase II Promoters

Sherri Matis, Manesh Shah, Richard Mural¹, and Edward Uberbacher

Informatics Group
Computer Science and Mathematics Division and ¹Biology Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6364

"The submitted manuscript has been authorized by a contractor of the U.S. Government under contract No. DE-AC05-84OR21400. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes."

To be published as full article Proceedings of the First World Congress on Computational Medicine, Public Health and Biotechnology, 1995

*Research was supported by the Office of Health and Environmental Research, U.S. Department of Energy under contract No. DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED
Abstract

One of the most difficult problems in the analysis of eucaryotic genes is the detection of RNA polymerase II promoter regions. Although promoter regions vary in the primary DNA sequence, a basic group of core promoter elements has been suggested in the literature. Many human promoter sequences contain a TATAA sequence element at approximately 30 bases upstream of the cap site (transcription start site). Other elements are the GC box which binds SP-1 and upregulates transcription, the CAAT box, and the ATG initiator codon. To characterize promoters, we constructed frequency matrices for each element using experimentally mapped human promoter regions. Additionally, we constructed histograms for the distances separating the various elements. We then used a neural network to combine these informational elements. The output of the neural network is then processed using a set of expert rules which depend on GRAIL's ability to find exons in anonymous DNA. This improves the selectivity of promoter detection and reduces the false positive rate.

Keywords: RNA Polymerase II promoters, neural network, Computational Biology, Human Genome Project, Informatics.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
Introduction

Detailed characterization of the genes within genomic sequence is a complex endeavor requiring a combination of a number of complementary technologies (10,12,13,29,30,31). Previous to this study we developed GRAIL, a modular expert system designed to recognize, characterize and model genes from DNA sequence data (29,30,31). GRAIL recognizes coding regions based on statistical parameters such as hexamer bias and other gene features such as splice junctions or polyA sites (19). Computational models of a single gene are constructed by dynamic programming using neural network predictions for starting, middle, and terminal exons. The best translation of an individual exon or gene model can then be used as a substrate for database queries via the ORNL genQuest server. The accuracy of GRAIL gene model prediction, when tested on known genes is greater than 90% (31).

As longer and longer DNA sequences are determined, the need for computational recognition of multiple gene structures contained in a single stretch of DNA will increase greatly. The major difficulty in multiple gene modeling comes from the lack of effective ways to locate the boundaries of genes. Regulation of gene expression is facilitated through the interaction of a number of factors with control regions. Transcription begins following binding by protein factors to signals within the promoter region and ends in recognition of a transcription terminator region. The recognition of these signals within the DNA can be used to delimit gene boundaries and guide multiple gene modeling using computational tools.

Many signal sequences are known to act in cis to control gene expression. Promoter and terminator regions, found at the 5' and 3' boundaries of the gene, control transcription and are thus not present in the nascent RNA transcript (3,8,18,22,23,27). Other signals such as splice junctions and polyadenylation sites are present in the primary transcript and function to direct post-transcriptional RNA processing. All of the previously mentioned signals are known to be constitutively controlled but there are examples of tissue specific, inducible, or developmentally regulated control elements. Recognition of these types of regions can be influential in constructing accurate gene models, particularly where multiple genes exist in one stretch of test DNA. Definition of signals known to induce gene expression as a response to external stimuli or as part of a developmental or tissue specific pattern also provides valuable information about the biology of gene expression.

Although a number of signals have been identified in experimentally determined promoters, there is limited detail available about the relevant protein-nucleic acid interactions. While footprinting experiments can show regions of protein-nucleic acid interaction, specific contact points between individual nucleotides and residues in the regulatory protein often remain speculative (2). Furthermore, many known control sequences are not highly conserved, allowing gaps and sequence substitutions while still retaining function and are therefore not adequately represented by a consensus. Only a few regulatory sequences, such as the TATA box (3,8,18,22,23) and the AATAAA polyadenylation hexamer (27), are well conserved. The underlying cause for the lack of sequence conservation may reflect promoter or polyadenylation site strength or perhaps the ability of the cellular machinery to recognize conserved structural signals which are not well conserved at the sequence level.

The position of control elements relative to other parts of the gene and to each other is also subject to considerable variability in most cases. A major class of RNA pol II promoters contain a TATAAA sequence at approximately 30 bases upstream from the site of transcription
initiation (cap site). However, additional sequence signals maybe present at varying distances relative to the cap site. A small group of promoters lack a consensus TATAA and function by means of a pyrimidine rich initiator sequence. Also, control of tissue specific expression of many genes is mediated by interactions with tissue specific sequence elements (8,11,21,23). Induction of promoters containing inducible elements require the binding of the hormone and cognate receptor complex to a specific sequence element in the promoter (21).

Cell processes accurately tolerate many sequence and distance variations in promoters without loss of function. Developing methods for finding complex promoter regions, in which the individual sequence elements are not conserved and are variably spaced, represents a fundamental computational challenge. Additionally, much of the structural, physical and chemical information used by the cell is largely hidden to current sequence-based analysis methods (2).

A number of methods for finding control regions have been used previously with limited success. Early methods predicted single elements from a preset consensus (9,25). These have been only marginally successful due to the sequence variability, the presence of gaps in many control sequences and differences in the combinations of control elements within a control region (18,26). Methods which detect control elements by comparison to an a priori defined frequency matrix represent an improvement in methodology but still most often result in the recognition of numerous false positive signals (3,22). Single and multiple alignments, while very powerful for the analysis of proteins, are of less use at the DNA sequence level because of numerous false signals with coincidental similarity (28). Several methods have been developed which combine simple elements to recognize possible complex or compound types of features. For example, EMV (6) and ConsIndex (10) can detect multiple consensus at variable distances. While these are a step in the right direction for dealing with control regions such as polyadenylation sites and RNA pol II promoters, so far, the success of these methods is unclear for the analysis of eucaryotic sequences. Neural network methods involving training on primary sequence data have also been tried for prediction of some gene features, such as splice sites (13). However, for long target sequences such as the Pol II promoter, the number of inputs and network nodes would be very large requiring an impractically large number of training examples. Although systems to define procaryote promoters have been developed (reviewed in 13), the structure of eucaryotic promoters is more complex and so far no reliable system has been implemented for their detection.

What is needed for Pol II promoters is a system capable of encompassing the significant variability and flexibility of these compound features, which apparently contain few hard constraints, but potentially many soft constraints. The design of our system, described in detail in this work, accomplishes this goal by using an array of syntactic and statistical information in a weighted way within a hybrid neural network-based system. This neural network uses scores from the individual sequence elements found in the promoter regions. Distance relationships between the various promoter elements are also potentially important so we included score information derived from spatial relationships between promoter elements. In many cases multiple CAAT and GC elements maybe present and we include this information in the system. In order to further reduce the false positive rate of the neural network system, a set of heuristic rules were applied to its predictions. The rules incorporate the results of exon prediction by GRAIL II (30,31) in combination with the network score. This combined system demonstrates significant performance improvement compared to previous methods for the prediction of human RNA polymerase II promoters.
Methods

Data sets

A group of 206 human promotor sequences were extracted from the eucaryotic promotor database, EPD(4). Release 36 was used in this work and is available via the computational biology Gopher server at the Johns Hopkins University. All data sets and matrices used in this paper can be obtained via e-mail from the authors at "grailmail@ornl.gov". This version of EPD contains experimentally determined promoter regions of 600 bases. All sequences in the database begin at approximately -499 with respect to the cap site (site of transcription initiation) and end 100 bases downstream of the cap site.

Classification of promoter regions based on the presence of consensus elements

A schematic of a general RNA Pol II promoter is shown in Figure 1. The first step our analysis was to search suitable promotor examples for the presence of putative elements and construct frequency matrices. This was done using shortened versions of the consensus sequences for these elements. To find initial examples for the TATA element we searched promotor examples with the ATA subsequence. Examples of two other promotor elements were also located by consensus analysis. CAAT was used to detect the CAAT element, and GGGCGG for the GC element. The cap site is approximately defined positionally within the data files (position 500). It should be noted that in many cases the functionality of specific sites has not been experimentally verified or annotated in EPD, however, the assumption of functionality is necessary to construct an example set.

We have defined 4 classes of functional human promotor regions based on the occurrence of the TATA element and other elements which have been shown to function in transcription initiation. Detection of the largest class of sequences, class I, are those which contain a TATA near -30 with respect to the cap. The detection of class I promoters will be addressed in this work. The remaining three classes of promoter regions were defined based on our analysis of promoter regions for other promotor elements and may or may not be mechanistically distinct. Detection of these three classes of promoters will be handled separately for each group.

In class I, TATA containing promoters, a basal level of transcription is supported by recognition of a TATA or TATA-like core promotor element. Consensus analysis shows that 169 out of 206 human EPD sequences in the dataset have a TATA or TATA-like element at or near -30 bases.

A second class of promotor regions was found which contain a GC box and lack a TATA element. The GC box upregulates the basal level of transcription of pol II promoters through binding of the factor SP-1 (23). It can be found in promoters containing a TATA element (class I or II) and maybe be present in multiple copies. In addition, the GC element is required for the activation of some TATA-less promoters and been proposed to enable TATA-less promoters to function via interaction with a proposed but still undefined ‘tethering factor’ which may exist as part of the TFIID complex (23). The mechanism of RNA Pol II transcription from class II promoters may vary from that of TATA containing promoters.

The third promotor class we define contains neither TATA or GC elements. There is
experimental evidence for the binding of TBP at the -30 region even in the absence of a canonical TATA box in DHFR, TdT, SV40 major late promoters (1,20,24). This class of promoters have been shown to be under the control of a pyrimidine rich initiator element and indeed for some of the promoters in this group the control region has been mapped close to the cap site (1,20,24). There are few examples of this type of promoter.

The final class of promoters (class IV) which could be defined were those which have a TATA which is found not at the usual -30 position, but the 17 members have a TATA box at a distance which is a multiple of 30 ± 1 bases. This class of promoters has not been previously recognized and may bind multiple copies of protein factors. The functional details of this group are unclear and it may be mechanistically distinct from the other classes of promoters.

The CAAT element could be detected in many of the test sequences. In some cases, multiple CAAT sequences could be found in promoter regions. Experimental evidence shows that the CAAT element functions as an adjunct element to upregulate transcription (8). Although we will use information about the CAAT element later in network training, we do not classify promoters based on its presence in the absence of the TATA or GC elements.

Construction of frequency matrices and distance histograms.

We constructed frequency matrices for the TATA, CAAT, cap, GC and ATG elements from 206 human promoter sequences. This provides somewhat more examples than previous work done by Penotti who used 44 human TATA sequences and 66 cap sites to construct frequency matrices (22). To construct a frequency matrix for the TATA, the sequence upstream of the cap site were searched for matches to ATA. A frequency matrix was constructed from a group of 127 out of 169 class I promoters (figure 2a) using the ATA region near position -30. These signals (counted from the first “T” in “TATA”) had an average distance of 32 ± 2 nucleotides upstream of the designated cap site. An additional 42 sequences displayed a TATA-like element in the -30 region but without the ATA consensus. These examples were not used in the matrix. The nucleotide frequency was calculated for the core (TATA) and for 10 nucleotides flanking both ends of the sequence.

The region upstream of the cap was searched for matches to the short CAAT and GC subsequences (CAAT and GGGCGG). The 26 base GC frequency matrix, shown in figure 2b, was constructed from sequence information flanking 132 GC elements. A 24 base frequency matrix was constructed for the CAAT element using the data from the 232 matches (figure 2c). There were 232 CAATs found in 140 sequences with up to 6 per sequence and 132 GC elements in 59 sequences with a maximum of 7 per sequence.

The consensus sequence for the cap has been reported as CA (3). However, searches for CA are likely to reveal many false signals. In addition, many cap sites do not contain the CA dinucleotide. The EPD dataset has the sequences aligned with the cap mapped approximately to position 500 and a 20 base sequence cap matrix was constructed from 206 sequences. The cap sites are expected to be accurate to within 5 bases or less. However, in the 7% of the sequences multiple initiation sites are annotated. In these cases, the major initiation site is located near position 500 and none of the minor sites were used in the matrix. Examination of the matrix is shown in figure 2d demonstrates there is a predominance CA dinucleotide beginning at position 7 in the matrix, despite the fact that no realignment of these sites has been attempted. This finding suggests that typically cap sites in the EPD dataset are misaligned by about -3
nucleotides. Based on this assumption, the average distance between the TATA and the cap in this dataset is actually about 29 bases.

Over 500 Genbank sequences were analyzed for the position of the translation initiation site and a matrix was constructed from the initiator ATG and flanking nucleotides. Our matrix (figure 2e) is in better agreement with a matrix constructed by Cavner (7) who used significantly more examples than these previous studies (16).

We collected data on the distance relationships between pairs of individual elements found within the promoter region. Histograms of the TATA-CAAT, GC-CAAT, TATA-GC, TATA-cap and cap-ATG distances were generated and scaled to make the maximum histogram height equal to 1.0. The results are shown in figure 3a-e.

The TATA-cap distance histogram is shown in figure 3a. All sequences selected for the current study had a TATA-cap distance of 25-35. A large proportion of the distances were between 32-34 nucleotides and we speculate that this preferred distance may reflect a systematic bias in many of the cap sites (see figure 2e). Considering the apparent discrepancy in the position of the cap site, the preferred TATA-cap distance is closer to 29-31 bases.

The CAAT-TATA distances span a wide range. The preferred distances are 50 and 70 nucleotides but distances across the entire range of 465 bases are found. The histogram is shown in figure 3b.

The GC-TATA distances are much more conserved than the TATA-CAAT distance with the predominate distance being less than 100 bases. The distance histogram is shown in figure 3c.

Analysis of the cap-ATG distance of an independent set of 169 human genes shows that only 14% had 5' untranslated regions of greater than 240 nucleotides, while 57% of the 5' untranslated regions were less than 100 nucleotides. The histogram for the cap-ATG distance is shown in figure 3d.

Within a promoter region, our analysis shows that the GC element usually precedes the CAAT box. In a few cases, the CAAT element may be upstream of the GC element. The GC-CAAT distance is plotted in the histogram shown in figure 3e. The case where the CAAT element upstream of the GC element is considered as negative distance. If multiple CAAT or GC elements were present, distances of each possible combination were included in the generation of the histogram.

**Construction of neural network training and test sets**

We trained standard back-propagation neural networks on parameters used to characterize RNA Pol II promoters. Each training example was made up of a vector containing the 13 separate scores calculated for a 600 base region containing an "ATA". The 13 different parameters used in network training were the TATAA, cap, CAAT, GC, and ATG scores; the number of CAAT and GC elements; the TATA-cap, TATA-CAAT, TATA-GC, Cap-ATG and GC-CAAT distances; and the GC content for a region of 600 basepairs beginning approximately 465 bases 5' of the TATA and ending at approximately 135 bases 3' of the TATA. If more than one CAAT or GC element was present, the highest scoring element was used for the CAAT or GC score. The highest scoring cap candidate within a region of 25-35 bases from the TATA was used. In the case where an element is absent from the promoter region, the appropriate element score and distance score were set to 0. The score for the promoter elements were scaled to
between 0 and 1.0 by determining the maximum and minimum values possible for summation of the matrix, subtracting the minimum score from the raw score and then dividing by the range. The number of CAATs or GCs was scaled by dividing the number of elements in each region by the maximum number of CAAT [6] or GC [5] elements per sequence in the EPD dataset and in the event that a test region had more than the measured maximum, a value of 1.0 was used. Out of the 127 Class I promoters, 107 promoters which were larger than 200 nucleotides were used as true training examples.

The false promoter examples were constructed from 62 human sequence files from Genbank, with false promoter candidates constructed around ATA subsequences found within internal exons and introns. This strategy was used to eliminate unannotated true examples which might be present at the 3' or 5' ends of published sequences. False promoter candidates were 600 nucleotide regions with the TATA at position 465. 570 false promoter candidates were used in the training set.

In the test set, 10 true examples from EPD were included as well as 49 true promoters annotated in Genbank files. There were 1503 false test examples taken from Genbank sequences for a total number of 1562 test promoter examples. The high ratio of false to true examples in the training and test sets mimics the actual density of false examples in genomic sequence, where on average there are approximately 33 promoter candidates (ATA subsequences) found in a 6000 base gene.

**Configuration of the neural network**

We trained a feed-forward, back propagation neural network consisting of 13 input nodes, 2 layers of 3 and 2 hidden nodes, and a single output node and a schematic of the network is shown in figure 4. Training of the network was monitored by testing the network's ability to discriminate true and false promoter examples in the test set. Several neural network configurations were evaluated for this study, including networks with both one and two hidden layers. Networks with two hidden layers outperformed those with a single hidden layer. During network training, testing was done periodically on the set of test examples and performance of the network stopped improving after 70,000 training examples.

**Rules for classification of eucaryotic promoters by GRAIL**

The context of a putative promoter within other gene features such as exons can provide additional information useful in eliminating false candidates. We derived several simple rules which we applied to the network output data to eliminate some false positive promoters detected by the neural network. These rules are based on the prediction of exons by GRAIL II (Gene Recognition and Analysis Internet Link (29,30,31). GRAIL II finds about 94% of coding exons in anonymous DNA sequence. The exon edges predicted by this system are quite accurate (the average error is 10 nucleotides). The exon prediction information is used in conjunction with the output of the promoter neural network to recognize impossible or very unlikely promoter candidates. The combined neural network and rule-based promoter recognition system has been implemented in the X-GRAIL client-server system. The client version of X-GRAIL 1.2 can be obtained via anonymous ftp from "arthur.epm.ornl.gov".

The strategy used for the rules employs two promoter scoring thresholds. Any putative
promoter with a network score below 0.34 was classified as a false example. Promoter regions with very high network scores, above 0.75, were classified as true promoters. Regions scoring between these two thresholds were evaluated in light of GRAIL II's exon prediction. For these, any promoter candidate which occurs within a GRAIL II exon is classified as false. Promoter candidates which occur within 110 nucleotides upstream of a predicted first exon and are more than 125 basepairs downstream of any exon are classified as true. Many times when two candidates promoters occur within close proximity of one another, the higher scoring one is the true candidate, so when two promoters occur within 1900 bases of one another, the higher scoring one is classified as true.

Results

Detection of RNA polymerase II promoters

The results of the promoter recognition system on a set of 93 previously unseen GenBank sequences is summarized in table I. This is an independent test set, separate from the set used to evaluate the network during training and consisting of whole gene regions in which exon prediction is possible. The performance of the system was measured on both strands of the sequence because GRAIL is designed to predict exons in anonymous double stranded DNA where the coding strand is either unknown or both strands may contain coding sequences. GRAIL II correctly predicted most first exons. Out of the entire group, 52 correct promoters were found, as well as 61 false promoters. 11122 out of 11183 negative promoter candidates were properly classified, leaving less than 1% false positives. The sensitivity was 56% and the specificity of the system was 46%. The correlation coefficient value of 0.99 was calculated for the entire set of genes. There was less than one false positive per 7 kb of DNA. The sensitivity, specificity and correlation coefficient for the combined neural network and rule based system are shown in table I.

Discussion

The promoter recognition system described here has practical utility for uncharacterized DNA sequence. It is capable of recognizing most human class I (TATA containing promoters, defined in this manuscript) Pol II promoters, and recognizing approximately equal numbers of real and false positives in sizable stretches of genomic DNA sequence. The efficacy of promoter recognition is the result of several factors; [1] The construction of better matrices for individual elements based on larger numbers of examples than in previous studies (22), [2] the use of only human promoter examples (to remove heterogeneity for calculation of the matrices), [3] the combination of many weak information elements, including distance relationships, which strengthens the promoter recognition by the network, and [4] the use of gene context information provided by GRAIL II.

We compared this system with a system for promoter detection derived by Bucher (3). Bucher developed a set of frequency matrices for the TATA, GC, CAAT, and cap elements found in a collection of higher and lower eucaryote promoters. We tested the matrices on a set
of human genes from Genbank and found that both Bucher’s matrices and ours find many false positive TATAs within genes while also finding most of the annotated promoters. Figure 5 compares the results of promoter detection using Bucher’s TATA matrix (figure 5a) with our matrix (figure 5c), on a 13,865 basepair test gene, HUMTFPB.gb_pr. This gene has 192 TATA candidates in the entire gene with 63 in the annotated promoter region. In this case and in general, the correct TATA scores higher relative to other potential candidates using our matrix. In the promoter region of the test example shown in figure 5b, Bucher’s threshold of -8.16 eliminated all but 6 candidates on the forward strand when used in conjunction with his matrix. When the neural network was implemented with our matrices, it eliminated all but 2 candidates, found in the same region on opposite strands of the DNA. The TATA candidate on the predicted coding strand is 27 bases upstream of the annotated transcription start and fits the criteria for a true TATA.

The primary advantage of our system is in a more effective combination of the various information elements, and additional insight about gene context. Additionally the Bucher TATA matrix was constructed from examples which are taxonomically heterogeneous, containing examples from both higher and lower eucaryotes. While Bucher’s strategy is designed to find different types of promoters, our’s is specific for human sequences.

Several methods have been reported which can generate a consensus element or elements from a collection of sequences. The method reported by Frech et al. appears to accurately recognize multiple protein binding sites by consensus in DNA sequence. While their method can define consensus binding elements, the overall success of the method in detecting promoter regions in multiple test sequences is not reported (10). Kel et al. reported the development of a method for promoter detection which has an error of about 15%. Unfortunately, the explanation of error calculation is unclear and evaluation of their method is difficult (14,15).

One of the limitations of multi-layer neural network systems for pattern recognition is that the method used by the network to discriminate features is obscure. However, insight into the importance of the various input parameters used in the decision making process can be obtained by examining the strength of connections made between the input layer and nodes in the first hidden layer. Careful study of the connection weights allows one to determine which parameters have the most influence on the network’s decision, and can provide clues about the biology of promoter function. As might be expected, the TATA score was most influential in the network’s prediction. The cap-ATG distance is well conserved and again is very important to the network’s decision. The GC content of the promoter region was surprisingly important in the network’s decision; we have observed that true TATA elements tend to be flanked by regions that have higher GC content than false TATAs ((3), see figure 2a). The number of GC and CAAT boxes was more important than the actual scores of the elements so regions with many GC or CAAT boxes were more likely to be classified as true promoters. Of the remaining distance relationships, all had some value, with the more conserved distance relationships more useful in the network’s decision.

A number of imperfections have been identified which limit the performance of the current system: (1) TATA-like elements which do not contain the ATA string will be missed. If the two additional subsequences (TATT and TAA) are used to identify potential candidates, the number of matches but also the number of false positives would increase. Thus modification of the system to detect the remaining TATA-like elements is not feasible at the present time. (2) GRAIL failed to accurately predict some starting exons and the system relies on GRAIL’s ability to detect these. As a result some promoter candidates with scores between 0.34 and 0.75,
promoters were missed by the combined system. (3) The rule which accepts very high scoring candidates is designed to alleviate some of the false negatives due to inaccurate first exon prediction, and also to consider high-scoring promoters which are at a large distance from the predicted first coding exon. This creates some false positives. (4) Some true promoters received very low network scores primarily because of poor agreement between individual elements and the frequency matrices. These missed examples may simply be very weak promoters. Many well studied genes are strongly expressed and thus there is probably a higher proportion of such sequences present in the databases. This may tend to bias the element matrices. These limitations pose challenges for further study.
References


Table 1. The results of neural network and expert system for a test set of 93 sequences. The results shown are calculated for both strands of the DNA sequence.
Results of Neural Network/Expert System

<table>
<thead>
<tr>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>+ Correlation Coefficient(31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>61</td>
<td>11183</td>
<td>41</td>
<td>56%</td>
<td>46%</td>
<td>.99</td>
</tr>
</tbody>
</table>

The system test set consisted of 93 Genbank Sequences and was used to evaluate the neural network/expert system. The sequences are humchymb, humdef5a, humelafin, humembpa, humigfbp1a, humil4a, humil8a, humjuncna, humkertra, humkrtlx, humlori, hummchemp, hummetiii, hummgpa, hummyc31, humpp14b, humpreelas, humprot1b, humproti2, humpsaa, humregb, humsemi, humsfipla, humsprpc, humtm, humtrpylb, humactaf, humadhlx, humatpl3a3s, humcd33a, humchat1, humcrekin, humdp01, humelaif, humelamb, humep2p01, humg0s2pe, humg0s3pp, humgphaa, humgr2, humilla, humlae, humld78a, humlplstn1, humaobbb, hummyhec08, humnak21, humntri, humpgdf, humsodb, humth01, humtnp2ss, hum019a, humactga, humadag, humakt1, humaldb1, humanfa, humant1, humapoac2, humapoc2, humapoe4, humbnop, humbsf3, humcrygb, humcs3, humef1a, humng019a, humg019b, humgasta, humghn, humhsp90b, humi11b, humi2, humi5, humkerep, hummetia, hummyca, humorahbbe, humotnpi, humpalc, humpgamng, humpla, humplsp, humpsa, humpsaa, hump,sap, humsaact, humtftp, humtga. These sequences were not present in either the training or network test set.
Figure Legends

Figure 1. A general schematic of a class I RNA polymerase II promoter. For our analysis, we analyzed 600 nucleotide regions, surrounding a TATA element at position 465, from EPD or Genbank files. The GC and CAAT elements are found 5' of the TATAA element. The TATAA element is found at about -30 nucleotides with respect to the cap. In many cases an ATG is found within 100 nucleotides 3' of the cap site. Certain element pairs are used to derive distance histograms (shown linked by lines). Note that certain elements such as CAAT and GC occur in multiplicity and the number of these elements is important score information.

Figure 2. Frequency matrices generated by analysis of 206 EPD sequences. The frequency at each position is shown as a percentage. The score for each element is calculated by summing the corresponding matrix value for each position. (a) TATA matrix, (b) GC matrix, (c) CAAT matrix, (d) Cap matrix, (e) ATG matrix

Figure 3. Histograms for the distances separating the various elements in the RNA pol II promoter. These distances were scaled from 0 to 1.0 for network training. EPD sequence data was used to construct histograms a-c and e. The cap-ATG distance histogram shown in d was constructed from 169 sequence files from Genbank. (a) TATA-cap distance, (b) CAAT-TATA distance, (c) TATA-GC distance, (d) Cap-ATG distance, (e) GC-CAAT distance.

Figure 4. The architecture of the neural network. Each input node corresponds to one sequence score parameter. The network has 2 hidden layers of 3 and 2 input nodes and 1 output node.

Figure 5. We used the TATA matrix of Bucher (3) and his threshold to score TATA elements and compared the results to those obtained using our matrices and neural network. (a) Results obtained using Bucher’s TATA matrix to score elements found within a promoter region of the sequence HUMTFPB.gb_pr. A total of 192 TATAs were found in the entire sequence with 63 candidates in the region from 1-2300 basepairs. The annotated TATA is at position 722 and is marked with an arrow. (b) Use of Bucher’s stated threshold eliminates all but 6 potential TATAA elements in this region including the real TATA at 772. (c) The TATA scores calculated with our matrix. (d) Resulting predictions after neural network processing and application of the expert rules. The only candidate promoter on the forward strand includes the real TATA element found at 772.
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td><strong>C</strong></td>
<td><strong>G</strong></td>
<td><strong>T</strong></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>27</td>
<td>42</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>28</td>
<td>44</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>32</td>
<td>36</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>24</td>
<td>25</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>32</td>
<td>20</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>31</td>
<td>32</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>52</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>24</td>
<td>22</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>33</td>
<td>52</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>21</td>
<td>22</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>30</td>
<td>44</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>28</td>
<td>21</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>23</td>
<td>21</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>25</td>
<td>48</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>26</td>
<td>17</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>27</td>
<td>47</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>19</td>
<td>22</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>35</td>
<td>28</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>15</td>
<td>23</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>29</td>
<td>30</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>35</td>
<td>28</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>37</td>
<td>20</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>26</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>40</td>
<td>15</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>29</td>
<td>30</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>58</td>
<td>19</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>6</td>
<td>43</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>46</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>15</td>
<td>23</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>29</td>
<td>30</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>46</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>a. TATA Matrix</td>
<td>b. GC Matrix</td>
<td>c. CAAT Matrix</td>
<td>d. Cap Matrix</td>
<td>e. ATG Matrix</td>
</tr>
</tbody>
</table>
a. TATA-Cap Distance

b. CAAT-TATA Distance
c. GC-TATA Distance
d. Cap-ATG Distance
e. GC-CAAT Distance
RNA Polymerase II Promoter Network

- #CAATS
- CAAT Score
- #GC
- GC Score
- TATA Score
- cap Score
- ATG Score
- CAAT-TATA DIS
- CAAT-GC DIS
- TATA-ATG DIS
- GC-TATA DIS
- cap-TATA DIS
- GC Content

Promoter Score