Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences

S. Burden1,*, Y.-X. Lin1 and R. Zhang2

1Department of Mathematics and Applied Statistics and 2Department of Biological Sciences, University of Wollongong, Wollongong, NSW 2522, Australia

Received on July 22, 2004; revised on September 23, 2004, accepted on September 24, 2004

Advance Access publication September 28, 2004

ABSTRACT

Motivation: Although a great deal of research has been undertaken in the area of promoter prediction, prediction techniques are still not fully developed. Many algorithms tend to exhibit poor specificity, generating many false positives, or poor sensitivity. The neural network prediction program NNPP2.2 is one such example.

Results: To improve the NNPP2.2 prediction technique, the distance between the transcription start site (TSS) associated with the promoter and the translation start site (TLS) of the subsequent gene coding region has been established. This information is combined with the coding region has been studied for promoters and the translation start site (TLS) of the subsequent gene. By proxy, the promoters are also associated with at least one gene coding region and after transcription, the relevant coding regions in the RNA are translated into a protein. Hence immediately downstream of the promoter region and TSS there exists at least one gene coding region. In this paper, the first nucleotide downstream of the promoter region to be translated (i.e. the first nucleotide in the immediately subsequent gene coding region) will be denoted by the relevant translation start site (TLS) for the promoter.

Availability: The data used in this project and the prediction results for the tested sequences can be obtained from: http://www.uow.edu.au/~yanxia/E_Coli_paper/SBurden_Results.xls

Contact: alh98@uow.edu.au

1 INTRODUCTION

Owing to the availability of vast amounts of genomic data, there is a need for prediction techniques that can rapidly and accurately evaluate sequences for the presence of promoters. Although a great deal of research has been undertaken in the area of promoter prediction, the problem has not yet been resolved. Algorithms tend to exhibit either poor specificity, generating many false positives, or poor sensitivity. The problems inherent in promoter prediction occur because the transcription process is initiated by specific interactions between proteins and the DNA sequence in the promoter region. The interactions are complex and the various elements involved are highly degenerate, so recognition of characteristic sequences within promoters is difficult. Consequently, the most distinguishing feature of many promoters is often the presence of one or more characteristic motifs upstream of the +1 position and the spacing and distance between the conserved elements. Various other signals, motifs and/or clusters of elements in the surrounding region have also been reported (Chasov et al., 2002; Ozoline et al., 1997; Pedersen et al., 2000; Ioshikhes et al., 1999).

Additional complication arises because the detectable motifs within the promoters also occur randomly throughout the genome. As such, the promoter elements do not possess specific identifying characteristics, and prediction algorithms must incorporate multiple features that may or may not be present in any given promoter. Even though the promoter elements are degenerate in nature, they contain the nucleotide sequences which indicate the starting point for RNA synthesis. Hence, they are always located immediately upstream of the first nucleotide (or bp) that is transcribed during gene expression. This nucleotide is often called the +1 position or transcription start site (TSS).

By proxy, the promoters are also associated with at least one gene coding region and after transcription, the relevant coding regions in the RNA are translated into a protein. Hence immediately downstream of the promoter region and TSS there exists at least one gene coding region. In this paper, the first nucleotide downstream of the promoter region to be translated (i.e. the first nucleotide in the immediately subsequent gene coding region) will be denoted by the relevant translation start site (TLS) for the promoter.

Currently, features that are commonly utilized in promoter prediction algorithms include homology with known promoters, the presence of particular motifs within the sequence, DNA structural characteristics and the relative signatures of different regions in the sequence.

Algorithms that use the presence of particular motifs within the sequence include those based on position weight matrices (PWMs). The usefulness of simple examples such as TATA (Bucher, 1990) is limited because they assume independence between adjacent bases and do not allow for the presence of multiple promoter elements, insertions, deletions or variable spacing between elements. However, more complex algorithms such as Eponine (Down and Hubbard, 2002) incorporate multiple promoter elements using probability distributions relative to the TSS.

Markov models in several forms have also been used for promoter prediction (Ohler et al., 1999, 2000; Audic and Claverie, 1997). They take advantage of the different characteristics of particular signatures in the sequence. Unlike PWM, they do not rely on the presence of a motif and allow for a departure from independence between adjacent nucleotides.

Word frequency techniques are also based on over-represented patterns of nucleotides of different lengths. Some algorithms such as PromoterInspector (Scherf et al., 2000) utilize classifiers to match sequences against a database of known promoter elements. However, as the probability of a word decreases exponentially with
its length, balancing sensitivity and specificity using single words is problematic. Others, such as PromFD (Chen et al., 1997) find over-represented patterns of 5–10 bp length and create information matrices to predict promoter location. Multiple promoter elements are incorporated into the Mitra Algorithm (Eskin et al., 2003) using disjoint groups of words with a given separation to find frequent signals in several bacterial genomes. Finally, the evolutionary algorithm has also been applied to promoter prediction (Corne et al., 2001).

Neural networks (NNs), another pattern recognition technique, have also been applied to promoter prediction. Despite the simple architecture of earlier algorithms (Demeler and Zhou, 1991), prediction accuracy was high but there was a correspondingly high incidence of false positives. More complex architectures including algorithms incorporating several NNs in series (Knudsen, 1999; multiple hidden layers and time delay neural networks (TDNNs) (Reese and Eckman, 1995) have also been applied to the problem. Interactive optimization of the number of nodes during training has been used (Yang et al., 1999) as has preprocessing of promoter sequences to extract features based on their information content (Ma et al., 2001).

Recent algorithms often incorporate several analysis techniques. PromH (Solovyev and Shahmuradov, 2003) uses linear discriminant functions to evaluate the structural characteristics of DNA in the promoter regions and also detects conservation features and nucleotide sequences of promoters from pairs of orthologous genes. Similarly, CONPRO (Liu and States, 2001) compares predictions from five algorithms, including the NNPP program, for sequences upstream from the TLS. Finally, Dragon gene finder (Bajic and Seah, 2003) compares sequences to the oligonucleotide positional distributions for a particular functional region of a gene. The output from several modules is then passed to an NN which integrates the information and provides a prediction.

Although all of these techniques attempt to maximize the number of promoters detected while minimizing the number of false predictions (incorrect predictions), there is always a trade-off between sensitivity (the percentage of true positives) and specificity (the ratio of true to false predictions).

To improve currently available algorithms, it is therefore necessary to introduce other measures to help in reducing the number of false predictions. The use of structural information has a lot of potential for assisting with promoter prediction and work is continuing in this area. For example, methods to predict areas of helix destabilization are being developed to help in reducing false positives (Bentham, 1996) and distance correlations between large sets of elements that have been used to identify over-represented correlations without the need for training (Quandt et al., 1996).

The distance between the TSS and TLS has not been explicitly utilized in promoter prediction algorithms. Some algorithms informally incorporate this information by restricting their search areas using the TLS as a reference (Liu and States, 2001), while others incorporate probability distributions. However, none incorporate distance information into their algorithms. Consequently, another measure to assist with promoter prediction can be defined: the distance between a promoter (defined by its associated TSS) and the TLS of the subsequent coding region. Using the calculated distance from known promoters to the start of the subsequent gene coding region, it is possible to estimate the probability of a promoter occurring at a given distance from the TLS and use this information to improve prediction specificity for existing promoter prediction tools.

The structure of this paper is as follows. In the next section the application of NNs to promoter prediction is described. The subsequent two sections first describe the distribution of distances between the TSS and TLS of known promoter sequences and then assess the sensitivity of this measure to the number of known promoters. The usefulness of the distance distribution for improving promoter prediction from NN techniques is tested in the following section. Finally the methods utilized are discussed.

2 APPROACH

2.1 Neural network prediction

NNs attempt to model the learning process in the brain. NNs have been used extensively for promoter prediction and a comprehensive summary of their application can be obtained from Wu (1997). In particular, they have been applied successfully to the *Escherichia coli* genome (Demeler and Zhou, 1991; Ma and Wang, 1999) and other bacteria (Kalate et al., 2003). NNs have the advantage that they can learn to recognize the degenerate patterns that characterize promoter motifs although they are often unable to distinguish novel promoters.

Although the results were not published, the program NNPP2.2 has also been applied to the *E. coli* genome (M. Reese, personal communication, http://www.fruitfly.org/seq_tools/promoter.html). The program uses a series of TDNNs to incorporate multiple promoter elements with variable spacing. To train the TDNN for prokaryotes, a carefully cross-validated test set of 272 *E. coli* promoters was used.

To use the program, a DNA sequence of length \( \geq 51 \) bp is inputted into the program and the output of NNPP2.2 is a list of predictions with scores greater than the user-defined threshold. Each prediction includes the prediction score and the associated 46 bp sequence corresponding to the positions \(-41\) to \(+5\), where \(+1\) is the TSS. All scores lie between 0 and 1 and represent the probability of the instance being a promoter. Because promoter elements may appear at different relative positions to one another and the TSS, the positional accuracy of promoter prediction using NNPP2.2 is \( \pm 3 \) bp. In this paper, a threshold score of 0.1 is used, to provide an initial limit to the number of predictions obtained.

The results obtained for eukaryotic organisms from this technique have been cited by several references. Most recently, Liu and States (2001) compared several available techniques during the development of their own technique, showing that NNPP2.2 is competitive with several other freely available techniques. The results show that reasonable results are available, although the technique suffers from a high level of false positives.

2.2 TSS–TLS distance

The TSS–TLS distance can be used to improve the promoter prediction for several reasons. First, the TSS and TLS can be defined for every gene and both locations can be experimentally defined (or verified) and second the location of the TSS is closely related to the promoter region.

The TLS in a gene is readily identifiable due to the structure of the coding regions in DNA. In many genomes, the first annotation of a sequence defines open reading frames that are essentially the predicted coding regions in the genome. As the TLS corresponds to the first nucleotide in the gene coding region, its position is easily defined.
Improving promoter prediction

![Graphical representation of TSS-TLS distance for a single DNA strand.](https://example.com/image)

Promoter regions are not well defined. They generally consist of several characteristic elements, but the exact number and the location of individual promoter elements varies. Since they are always located immediately upstream of the TSS, it is expedient to define a promoter region in terms of its TSS. By definition, the TSS is also located upstream of the associated TLS. Hence, all promoters are related to at least one gene coding region downstream of the promoter sequence. In *E. coli*, the distance between the TSS and TLS is relatively small, ranging from 0 to 1000 bp at the known extremes. As large samples of experimentally defined promoters exist for the *E. coli* genome, the distribution of the distance between the TSS and TLS can be estimated if all the TLS are given.

Formally, the distance between the TSS and the TLS is defined in terms of the number of nucleotides between these positions. As shown in Figure 1, the +1 position in the promoter (the TSS) is the first nucleotide counted and the nucleotide before the TLS is the last to be counted. This measure will be referred to as the TSS–TLS distance.

For the *E. coli* K12 bacteria, we found 771 experimentally defined promoters from the EcoCyc database V7.1 (Karp et al., 2002, http://biocyc.org/ecocyc/) associated with 587 genes. For each promoter we calculated the TSS–TLS distance and evaluated the distribution of the distances. The resulting distribution has a range from 0 to 1000 bp at the known extremes. As large samples of experimentally defined promoters exist for the *E. coli* genome, the distribution of the distance between the TSS and TLS can be estimated if all the TLS are given.

To check whether all promoters share the same distribution, we estimated the shape of the final distribution. However when less promoters were included, the peak density was reduced and more weight was observed at the tail of the distribution.

The distribution appears to be robust when the number of known promoters exceeds 200. As such the technique is still applicable when significantly fewer promoters have been experimentally defined.

### 3 ALGORITHM

The NN technique NNPP2.2 is based on the nucleotide sequence alone. It recognizes only the presence and relative location of patterns and motifs within a promoter, rather than the location of promoter motifs relative to the TLS. It predicts the probability that a tested sequence +3 bp (denoted by s) belongs to the class of true promoters (S).

Given a potential promoter sequence s, the relevant TSS and TLS can be determined. The TSS–TLS distance is denoted by d, and measures the distance between the TSS and TLS for sequence s. Since the position of sequence s is usually randomly located, d is a random variable.

Here, we propose a new approach for improving the predictive ability of NNPP2.2 using additional information provided by d. Instead of just predicting the probability that sequence s belongs to the class of true promoters $P(s \in S)$, we predict the probability that sequence s belongs to the class of true promoters and that $d \in (x - a, x + a)$. This probability can be written as $P[s \in S \land d \in (x - a, x + a)]$, where x is an integer and a is a predefined non-zero integer.

As it is relatively easy to identify the locations of TLS in the sequence, incorporating this information into the prediction algorithm improves specificity. That is, instead of predicting whether a sequence is a promoter by just using the nucleotide content of the sequence, we are also finding out whether the sequence is located in proximity to a gene coding region. As such we can theoretically guarantee that $P[s \in S \land d \in (x - a, x + a)]$ predicts promoter

---

**4.2 Sensitivity of TSS–TLS distance**

In the next section, a new measure for predicting promoter locations based on the empirical probability distribution of TSS–TLS distance is developed. The empirical probability distribution used in this paper is based on the information from the 771 *E. coli* promoters. Theoretically, to accurately predict promoters by using the measure developed in this paper, this empirical distribution should be updated frequently whenever any new promoter is identified. However, it is not practical to update the empirical distribution so often. Thus, it is interesting to know whether the empirical probability distribution is very sensitive to the number of promoters used when the number is already reasonably high.

To carry out this study, a subset of promoters was randomly selected from the set of 771 promoters and the smoothed density function for the TSS–TLS distance was recreated. Subsets of 400, 200 and 100 promoter sequences were tested.

As shown in Figure 4, the resulting distributions closely approximated the shape of the final distribution. However when less promoters were included, the peak density was reduced and more weight was observed at the tail of the distribution.

The distribution appears to be robust when the number of known promoters exceeds 200. As such the technique is still applicable when significantly fewer promoters have been experimentally defined.
sequences more accurately. This probability can be calculated as follows:

\[
P[s \in S \text{ and } d_i \in (x - a, x + a)] = P(s \in S)P[d_i \in (x - a, x + a)|s \in S].
\]

(1)

Therefore, \(P[s \in S \text{ and } d_i \in (x - a, x + a)]\) is estimated by

\[
\hat{P}[s \in S \text{ and } d_i \in (x - a, x + a)] = \hat{P}(s \in S)\hat{P}[d_i \in (x - a, x + a)|s \in S],
\]

(2)

where \(\hat{P}(s \in S)\) is provided by NNPP2.2 and \(\hat{P}[d_i \in (x - a, x + a)|s \in S]\) is given by the empirical distribution produced in this paper, we chose \(a = 3\) as this corresponds to the positional accuracy of NNPP2.2. Hereafter, this prediction technique is referred to as the TLS–NNPP technique.

4 IMPLEMENTATION

In general, implementing the TLS–NNPP technique for an organism involves several steps. First, the empirical distribution of TSS–TLS distances for known promoters must be generated. Then, DNA sequences must be passed through the NNPP2.2 prediction algorithm to obtain the locations of predicted promoters. These predictions must be associated with the closest subsequent TLS in the sequence and the TSS–TLS distance for each prediction determined. Finally, for each prediction, Equation (2) is calculated and the final set of predictions created using an appropriate cut-off probability.
For the *E. coli* genome, implementation involved each of these steps. First, as described, the empirical TSS–TLS distance distribution was generated. Then, using the NNPP2.2 module for prokaryotes, 510 sequences from the *E. coli* genome containing a total of 671 known promoters were tested. Tested sequences were randomly chosen from the set of known promoter sequences. Each sequence was 500 bp long, starting 501 bp upstream from the TLS and finishing at the −1 position before the TLS. Sequences contained between 1 and 7 known promoters, with an average of 1.32 promoters per sequence. Promoters were considered to be correctly predicted when the actual TSS of the promoter fell within ±3 bp of a predicted TSS. Only one true positive prediction was allowed for each tested promoter. The results are shown in Table 1.

In Table 1, TP denotes the number of true predictions with an NNPP2.2 score greater than the chosen threshold of 0.1. No. Pred. denotes the total number of tested sequences with a score above the threshold. \( \hat{P}(\text{Prom}) \) denotes the estimated probability from the data that a tested sequence is a true promoter and % Recog denotes the percentage of correct predictions obtained in total (% Recog = TP/671), which is a measure of the coverage of the NNPP2.2 program. Using a threshold of 0.1, a total of 7584 predictions resulted and of these, only 470 corresponded to true (or known) promoters. The number of false predictions generally exceeded 85% of the total number of predictions, even at very high thresholds.

To incorporate the new prediction measure proposed in this paper, for each of the 7584 sequences obtained from NNPP2.2 with a cut-off >0.1, the TSS–TLS distance, denoted by \( x \), was calculated. Using this distance, \( P(d_i \in x \pm 3 | s \in S) \) was calculated from the TSS–TLS distance empirical distribution. \( P[s \in S \text{ and } d_i \in (x - 3, x + 3)] \) was then calculated by multiplying the score from NNPP2.2 by \( P(d_i \in x \pm 3 | s \in S) \). Hence the probability of each prediction obtained from NNPP2.2 was modified to give an adjusted probability of the sequence being a promoter. It was found that after adjustment, the maximum prediction probability was 0.0672. New cut-off values were created as a percentage of this value and a modified set of predictions was created.

### Table 1. NNPP2.2 prediction frequency for 510 *E. coli* sequences of length 500 bp containing 671 known promoters

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>TP</th>
<th>No. pred.</th>
<th>% Recog</th>
<th>( \hat{P}(\text{Prom}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0.98</td>
<td>42</td>
<td>303</td>
<td>6.3</td>
<td>0.139</td>
</tr>
<tr>
<td>&gt;0.9</td>
<td>135</td>
<td>1055</td>
<td>20.1</td>
<td>0.128</td>
</tr>
<tr>
<td>&gt;0.8</td>
<td>208</td>
<td>1774</td>
<td>31.0</td>
<td>0.117</td>
</tr>
<tr>
<td>&gt;0.7</td>
<td>255</td>
<td>2339</td>
<td>38.0</td>
<td>0.109</td>
</tr>
<tr>
<td>&gt;0.6</td>
<td>297</td>
<td>2896</td>
<td>44.3</td>
<td>0.103</td>
</tr>
<tr>
<td>&gt;0.5</td>
<td>333</td>
<td>3502</td>
<td>49.6</td>
<td>0.095</td>
</tr>
<tr>
<td>&gt;0.4</td>
<td>356</td>
<td>4116</td>
<td>53.1</td>
<td>0.086</td>
</tr>
<tr>
<td>&gt;0.3</td>
<td>401</td>
<td>4917</td>
<td>59.8</td>
<td>0.082</td>
</tr>
<tr>
<td>&gt;0.2</td>
<td>446</td>
<td>6124</td>
<td>66.5</td>
<td>0.073</td>
</tr>
<tr>
<td>&gt;0.1</td>
<td>470</td>
<td>7584</td>
<td>70.0</td>
<td>0.062</td>
</tr>
</tbody>
</table>

### Table 2. TLS–NNPP prediction frequency for the 510 *E. coli* sequences of length 500 bp containing 671 known promoters

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>TP</th>
<th>No. pred.</th>
<th>% Recog</th>
<th>( \hat{P}(\text{Prom}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0.0605</td>
<td>27</td>
<td>56</td>
<td>4.0</td>
<td>0.482</td>
</tr>
<tr>
<td>&gt;0.0538</td>
<td>69</td>
<td>150</td>
<td>10.3</td>
<td>0.460</td>
</tr>
<tr>
<td>&gt;0.0471</td>
<td>97</td>
<td>245</td>
<td>14.5</td>
<td>0.396</td>
</tr>
<tr>
<td>&gt;0.0403</td>
<td>122</td>
<td>380</td>
<td>18.2</td>
<td>0.321</td>
</tr>
<tr>
<td>&gt;0.0336</td>
<td>159</td>
<td>529</td>
<td>23.7</td>
<td>0.301</td>
</tr>
<tr>
<td>&gt;0.0269</td>
<td>202</td>
<td>766</td>
<td>30.1</td>
<td>0.264</td>
</tr>
<tr>
<td>&gt;0.0202</td>
<td>245</td>
<td>1097</td>
<td>36.5</td>
<td>0.223</td>
</tr>
<tr>
<td>&gt;0.0134</td>
<td>303</td>
<td>1616</td>
<td>45.2</td>
<td>0.188</td>
</tr>
<tr>
<td>&gt;0.0067</td>
<td>383</td>
<td>2638</td>
<td>57.1</td>
<td>0.145</td>
</tr>
<tr>
<td>&gt;0.0034</td>
<td>430</td>
<td>3939</td>
<td>64.1</td>
<td>0.109</td>
</tr>
</tbody>
</table>

Table 2 shows the results of the TLS–NNPP technique. In general, compared with the results from NNPP2.2 alone, the total number of true predictions has declined, but the number of false predictions has also declined to a great extent. Hence, the percentage...
of false predictions is now always <85%, and reaches as low as 52%. The benefit of the TLS–NNPP technique can also be seen by comparing the total number of predictions for a given number of true predictions. From Table 1, at a threshold of 0.9, NNPP2.2 provides 135 true predictions from a total of 1055 predictions \[ \hat{P}(\text{Prom}) = 0.128 \]. Using the TLS–NNPP technique to obtain 135 true predictions, a threshold of 0.0383 is required (not included in Table 2). At this threshold, a total of only 423 predictions are obtained \[ \hat{P}(\text{Prom}) = 0.319 \], which is a 60% improvement on the unmodified algorithm.

Figure 5 shows the reduction in false positives obtained when using the TLS–NNPP technique compared with NNPP2.2 alone. Both techniques suffer from overall low recognition rates for the promoters because the prediction program NNPP2.2 was trained only on 293 E.coli promoters, and therefore many of the newer promoters are not recognized by the program.

5 DISCUSSION
This study has shown that the TSS–TLS distance distribution can be used to reduce the incidence of false predictions from an existing prediction technique. By combining the distance information with the pre-existing NNPP2.2 prediction algorithm, we have demonstrated that the prediction of promoters can be significantly improved. In addition, the technique ensures that each promoter that is predicted is associated with a gene coding region.

For the E.coli genome, there is one limitation for using the empirical TSS–TLS distance distribution to improve promoter prediction by the NNPP2.2 program. Promoters that are located further upstream than 500 bp will, by the very nature of the empirical distance distribution for E.coli, score very low and will not be located using this technique. Hence, while it is useful for most E.coli promoters, prediction for those few located further upstream will not be improved.

Several sources of errors have been identified in this study. The size of the dataset used in this study does not necessarily make it representative of the E.coli genome. The fact that these promoters have been experimentally defined may mean that they are all the more homogeneous than the set of unknown promoters within the genome. As such, the presence of bias in these results is possible.

Conversely, within each sequence there could exist other promoter sequences that have not yet been identified. As such some of the false positives may in fact be unknown or as yet unreported promoters, suggesting that the results obtained are conservative.

The additional benefit of this technique is that there is a potential to use it more generally. That is, the distance information could be used to post-process results from other traditional prediction algorithms. Potentially, it could also be incorporated into a new promoter prediction technique for E.coli bacteria. Hence, while this paper has focused on the NNPP2.2 prediction algorithm, the technique could also be applied to other prediction algorithms and there is a scope for further research in this area.

Finally, although the results cannot be directly generalized to other organisms, there is a potential for this technique to be applied more widely. Even higher organisms, whose promoter structure and TSS–TLS distance distribution may be significantly different from those found in E.coli promoters, may have a similarly stable TSS–TLS distance distribution that could be used to improve the prediction specificity. The only requirement is access to a sufficiently large number of experimentally defined promoter sequences from similar organisms.

REFERENCES