Identification of ribosome binding sites in
Escherichia coli using neural network models

David Bisant* and Jacob Maizel

Neuroscience Program (151 B), Stanford University, Stanford, CA 94305, USA and 1National Cancer Institute, FCRF, Building 469, Room 151, PO Box B, Frederick, MD 21701, USA

Received August 26, 1994; Revised and Accepted February 8, 1995

ABSTRACT
This study investigated the use of neural networks in the identification of Escherichia coli ribosome binding sites. The recognition of these sites based on primary sequence data is difficult due to the multiple determinants that define them. Additionally, secondary structure plays a significant role in the determination of the site and this information is difficult to include in the models. Efforts to solve this problem have so far yielded poor results. A new compilation of E.coli ribosome binding sites was generated for this study. Feedforward backpropagation networks were applied to their identification. Perceptrons were also applied, since they have been the previous best method since 1982. Evaluation of performance for all the neural networks and perceptrons was determined by ROC analysis. The neural network provided significant improvement in the recognition of these sites when compared with the previous best method, finding less than half the number of false positives when both models were adjusted to find an equal number of actual sites. The best neural network used an input window of 101 nucleotides and a single hidden layer of 9 units. Both the neural network and the perceptron trained on the new compilation performed better than the original perceptron published by Stormo et al. in 1982.

INTRODUCTION
Neural networks and molecular sequence analysis
Several researchers have exploited the pattern recognition capabilities of neural networks in analyses of protein sequences. An example of such a study is the work of Qian and Sejnowski (1) in their rigorous application of neural networks to predicting protein secondary structure. The evaluation of their network was comprehensive and the combined success level they reported on both a homologous and a non-homologous test set (~68%) was quite good. Since their study, several researchers have improved upon their results by incorporating non-local information into the network input (2–4). Other applications of neural networks to protein analysis have included systems to predict β-turns (5), to identify cysteines which participate in disulfide bonding (6) and to predict the surface exposure of amino acids (7).

There have also been many applications of neural networks to nucleic acid sequence analysis. A popular problem has been the recognition of Escherichia coli promoters. Two early attempts (8,9) used primitive types of neural networks and achieved decent results. A more rigorous study by O’Neill (10) was more informative. Using feedforward backpropagation networks, he selected only the 17 base promoter spacing class. He trained his network on 39 of these promoters. Using an input window of 58 nt, his trained network was ~77% successful at identifying new promoters, with a false positive rate of ~0.1%. This performance was ~8% better than the previous best method, which was a hand-crafted rule-based system also developed by O’Neill and his assistant (11,12). The rule-based system took them a year to develop, whereas the neural network approach took 2 months. Another area where neural networks have been applied to nucleic acid sequence analysis is the identification of eukaryotic coding regions (13,14).

In each of the studies mentioned above, neural networks significantly improved upon the previous best methods. The improvement from using neural networks was especially noticeable when compared with ad hoc or hand-crafted systems.

Early and current research on ribosome binding sites
Several reviews (15–17) have described the early research on translation in E.coli and other prokaryotes. The 1981 review by Gold and co-workers (15) and the 1982 review by Stormo and co-workers (18) were especially comprehensive and came at a time when much certainty could be placed in the models which were formulated to explain the collected data. This study will not seek to duplicate their work, but will summarize the research up to that point and will discuss the work which has been performed since then. Emphasis will be placed on the areas of what defines a ribosome binding site and on the computer methods which have been developed to locate them.

Briefly, ribosome binding sites were first defined by several groups performing nuclease protection experiments (19–21). In these experiments, complexes formed of phage mRNA, ribosomes and fMet-tRNAfMet were exposed to ribonuclease. The 20–40 base long segments of mRNA which remained were called ribosome binding sites. The AUG start codon was positioned about
12 bases from the 3'-end of the protected segment. Since then, hundreds of sequences corresponding to gene starts have been determined and catalogued. At the time of their 1981 review (15), Gold and co-workers had collected 120 ribosome binding sites. Currently, there are well over 1000 ribosome binding sites known.

In 1974, Shine and Dalgarno (22) sequenced the 3' terminal sequence of the _E.coli_ 16S ribosomal RNA (rRNA). They hypothesized that the terminal sequence they found, ACCUCUCUUA-3', recognized a conserved sequence, UGGAGG-5', found in most ribosome binding sites. The next year, Steitz and Jakes (23) confirmed this hypothesis by disassembling ribosomal mRNA complexes which had been clipped with colicin. A complex was found which consisted of the mRNA ribosome binding site and the 3'-end of 16S rRNA, apparently maintained by a 7 base complementarity.

The base pairing model of Shine and Dalgarno did not explain everything. One of the sequences they used to construct their model showed a high degree of complementarity to 16S rRNA, yet was translated very poorly. This led them to conclude that other determinants played a role in translation initiation, such as the degree to which the secondary structure of the mRNA makes the site available to the ribosome. To this day, there is a strong belief that the structure of the mRNA plays a significant role in translation initiation. This process has yet to be characterized precisely, though progress (24) is being made.

Work by Steege in 1977 (25) helped to further understand and define ribosome binding sites and the role of the initiator codon. She identified a GUG as an initiation codon in the _E.coli_ lacZ repressor mRNA. Both GUG and UUG initiation codons are found in a small percentage of ribosome binding sites. Steege further found that if multiple in-frame initiator codons were present in the sequence, only the one which followed an appropriately positioned Shine-Dalgarno sequence was utilized for initiation. She further determined that after translation termination, the ribosome simply did not initiate at the next available initiator codon. A genetic study by Dunn et al. (26) further solidified the role of the Shine-Dalgarno element and the initiation codon.

More examples were found of sequences containing very complementary Shine-Dalgarno and initiator elements yet which did not initiate translation efficiently. Work with phages λ (27,28) and T7 (26) revealed that some proteins were translated 1000-fold more efficiently, even though the levels of mRNA corresponding to each of the proteins were the same. Work of a different nature also pointed to the existence of other determinants. A series of protection experiments (29,30) with the phage MS2 RNA replicase cistron yielded a fragment from the 5'-end of that gene containing the Shine-Dalgarno sequence, the first several codons and some 5' regulatory regions. This fragment was then digested further from the 5'-end to create a series of fragments of different lengths, each of which had the Shine-Dalgarno element and the initiator codon. Fragments which had fewer than 35 upstream bases, from the AUG, were unable to bind ribosomes, whereas longer fragments were able to bind. One of the most confusing studies to come out of this era was a 1979 paper by Roberts et al. (31). They looked at translation of a lacZ-cro fusion protein using a series of 5' deletion mutants. The interesting mutants were those with deletions in the −59 to −31 area. Though differing by only a few bases in this region, some mutants were 12-20 times more efficient in synthesis of the product. It was apparent that the other determinants which had long been suspected to play a role could have significant effects on translation initiation. Most researchers felt the secondary structure of the mRNA was responsible for the observations which had pointed to the existence of other determinants.

Though computer models of secondary structure were primitive at the time, some researchers used them to examine initiation regions. Selker and Yanofsky (32) examined the ribosome binding sites of _Salmonella typhimurium_. They found these regions were capable of forming stable stem-loop structures with most of the Shine-Dalgarno element and the initiator codon in the single-stranded portions. Iser and associates (33) constructed secondary structure models which explained the results of the Roberts et al. study mentioned above, by placing these elements in the single-stranded portions. Unfortunately these studies did not test their hypotheses against experimental data.

Incremental improvements in our understanding and definition of ribosome binding sites have been made in the last 12 years, but precise characterization is far from being achieved. Though some binding sites are more efficient than others, it is not possible to tell that by examining the message and its primary sequence. Little improvement has been made in the ability to precisely identify ribosome binding sites merely by examining primary sequence data. Below is listed a consensus ribosome binding site as determined by Scherer and co-workers in 1980 (17), followed by several of the actual sites analyzed in this study, along with the locus name. The Shine-Dalgarno element and the initiator codon are in lower case.

**Consensus:**

```
UUAUUCUUUU AAAAAUuaug gaggauAUUUU angAAAAAAA UUAAAAC
```

**ECODKSA:**

```
UGAAACGAGU CAGAAACACU uCCggaWuG augAcGCCUG AGCGCUAG
```

**ECOMICA:**

```
UUGCCCCCAA CAAACAGGau UUCggaGACC VugCAAGCGU CCAUUUUU
```

**ECOPUP:**

```
AAU AUUCCUCCCU UuaGcGcaCga augCUUUAGU CUAUACAGA
```

**ECOPURCA:**

```
GCUUCAUAGC AGCGUUACAG CsCggcAuG gugACGUAUA UUUCCGGC
```

**ECOPOLBDA:**

```
CUG UUUUUGuGauG gauUUUUCACG gcgCgCGAG CAGGUUUUA
```

**ECOPURCA:**

```
AACGUAAAUU CACACCgag uggCuAUAAAG augCAAAAA CAGCGUAGU
```

**ECORPSOP:**

```
UACGUUCUAA GAAAGAGAaa gauUUUACAA uagCUUAUCG CGAUCGUC
```

Currently, the following factors have been identified in the definition and function of a ribosome binding site:

(i) the Shine-Dalgarno element (34);
(ii) the start codon and the bases immediately surrounding it (35);
(iii) the presence of multiple competing initiator codons (36);
(iv) the distance between the start codon and the Shine-Dalgarno element (37);
(v) the sequence of the bases in between the start codon and the Shine-Dalgarno element (36,38);
(vi) the composition of the bases far downstream (39);
(vii) the composition of the bases far upstream (40);
(viii) the secondary structure (35,41);
(ix) the presence of translational enhancers (42,43).

These multiple factors which contribute to translation initiation make it a difficult process to model.
Identification of ribosome binding sites

The pattern recognition problems found in the area of genetic sequence analysis are typically very difficult. Few methods are successful. As a result, little research has been published in the area. The work on identification of ribosome binding sites is no exception. There were primarily two papers published in the early 1980s. They were, and still are, the only primary reports which have been published on the subject.

The first paper, by Stormo et al. (18), is perhaps the most comprehensive study done to date on the identification problem. This paper, along with another of their papers published in the same volume, explored properties of ribosome binding sites in E. coli and what methods helped in exploiting these properties to distinguish these sites from the surrounding RNA.

From a database of 78 612 bases of transcribed RNA, Stormo and co-workers identified 124 ribosome binding sites. For each of their 124 ribosome binding sites, instead of directly counting the complementarity with the Shine–Dalgarno element, they calculated the free energy of each ribosome binding site coupled with 16S rRNA. They found a wide range in the free energies, with a peak at −10. They also found in this analysis many base paired structures which yielded a low free energy but which would not have had a high degree of complementarity with 16S rRNA.

Using a string search procedure, they also developed a set of rules to identify ribosome binding sites. Their rule-based procedure also identified many false positives or non-sites. To determine if there were any differences between the ribosome binding sites and the non-sites, they looked at the distribution of mononucleotides in the region −60 to +40. The distribution differed markedly between the two sets, especially in the region −30 to +20. For the ribosome binding sites, G residues especially were absent outside the Shine–Dalgarno element and the initiator.

Given the histogram data, they then tested the ribosome binding sites to determine where in the area they deviated from the expected. Six distinct peaks were found. Besides the expected peaks in the Shine–Dalgarno element and the initiator, peaks were also found at positions −20, −3, +5 and +12. This data was in agreement with similar results from Scherer et al. (17). From this additional information, Stormo and associates added more rules to their procedure to help distinguish the ribosome binding sites. Their analysis did not utilize secondary structure information, since these structures are difficult to determine for lengthy mRNAs with high reliability. This, of course, is a weakness which affects all models used for ribosome binding site identification. Another weakness of this method, however, is the qualitative nature of the prediction. The system of rules simply produces a prediction of whether it is a ribosome binding site or whether it is not. It gives no information about the certainty of the prediction or the strength of the site.

Even more significant than their initial attempts using a hand-crafted, rule-based system, was their attempt at using the perceptron algorithm (44). Using this procedure, Stormo and associates trained three different perceptrons with window sizes of 51, 71 and 101 nucleotides. For their training data they used all the ribosome binding sites selected from their dataset for the positive examples (S+), and slightly more non-sites (S−) which their previous rule-based method mistakenly identified as ribosome binding sites. Training typically took about 100 rounds before a plateau was reached and training was terminated. To evaluate the performance of their trained perceptrons, they used several sequences which were not in their training set. This step was not very rigorous, utilizing only eight sequences with a total of 10 ribosome binding sites.

The researchers found that a window of 101 worked best, identifying seven of the 10 genes, with five false positives. This was a bit of a surprise, since the ribosome is known to interact with many fewer nucleotides than 101. The researchers surmised that the context surrounding the ribosome binding site is important, perhaps decreasing the secondary structure in the area so the ribosome is free to bind. In an analysis of the variability at each position in their weight matrix, they found peaks at positions corresponding to the Shine–Dalgarno sequence and the start codon, as expected.

These studies by Stormo and co-workers contributed significantly towards advancing this area of research. Their perceptrons performed much better than a consensus search and much better than their hand-crafted, rule-based system mentioned earlier. It also allowed quantitative evaluation at each site, so that the sites could be ordered. At the time, the authors thought it was possible that the value given to each site could correlate with the translational efficiency of that site. Research (22, 45) from early on has indicated that some sites with a consensus ribosome binding site are poor at initiating translation and hence the correlation is not true for every site. It is still probable that some correlation will be found, however.

Though the performance of the perceptron was significantly better than competing approaches at that time, it was far from ideal. The researchers attributed this poor performance to noisy data, whereby there were other ribosome binding sites in the training data, but they were unknown and hence labeled incorrectly. Secondly, they attributed a great deal to secondary structure effects and felt that better performance would not be obtained until these effects could be modeled properly.

Summary of results

The present study sought to improve upon the work of Stormo and associates by applying neural networks to the same problem. A new and larger database of E. coli sequences was compiled specifically for this study. The sequences were aligned with the published perceptron of Stormo and co-workers, with a perceptron trained on the new data and with a variety of neural networks trained on the new data. A new method was also introduced to compare the performances of the systems across the entire range of thresholds used for classification. The results indicate that the neural network performed significantly better than the perceptron, finding less than half the number of false positives when the two models were adjusted to locate an equal number of sites. As an example, when each method was adjusted to identify 75% of the ribosome binding sites in the database, the perceptron of Stormo and co-workers (44) had 0.79% (1660) false positives, the perceptron method applied to the new compilation had 0.20% (415) false positives and the neural network had 0.08% (172). The improvement was seen over the entire range of thresholds.

METHOD

The database of ribosome binding sites was constructed by first extracting all non-redundant E. coli sequences from GenBank release 73 (46). Each of these GenBank entries was then examined to determine the presence of transcribed regions. To
Additionally, the convergence of a network can be much faster than learning the ribosome binding sites would be more evenly represented. A subset of the training data was used for this investigation, so that any potential bias itself to the over-represented training examples. Hence, only positive examples presented to the network are under-represented when training the neural networks minimize squared error. If the input could be presented to the network. Much like the scheme used to perform a perceptron was published in their 1982 study (44). The activation function for each unit was the logistic sigmoid  

$$A(x) = \frac{1}{1 + e^{-x}}$$

co-workers in their previous study of the problem. Their best performing perceptron was published in their 1982 study (44). This data was typed into a computer file and implemented with a neural network simulator developed at the University of Maryland and partially designed by one of the authors was used in the analysis of the data. This simulator was based on sequences in the general area and on the similarities seen in all E.coli mRNA sequences.

After all the transcribed regions were identified, they were extracted from the entries and 75% of the regions (328) were randomly selected for the first file, which was used as the training data, while 25% (122) were selected for the second file, which was used as the testing or evaluation data. There were ~590,000 bases of transcribed RNA in the training file and ~210,000 bases in the testing file. The reason that two separate files were created is so that one file could be used in the generation of the analysis or search algorithms, while the other file could be used to evaluate how well the generated algorithm performs on data it has not seen before. This cross-validation method is often used to determine the generalization ability of a trained algorithm.

With the data collected, corrected and split into a testing and training set, an alignment was made of all ribosome binding sites so they could be visually examined for errors and cross-checked against their corresponding referenced article if necessary. Once that was completed, the first step in the analysis was to analyze the testing data with the best method produced by Stormo and co-workers in their previous study of the problem. Their best performing perceptron was published in their 1982 study (44). This data was typed into a computer file and implemented with the same algorithm used to execute the other perceptrons produced by this study.

The next step in the analysis looked at the perceptron training method of Stormo and co-workers, which was trained in this study instead of the limited data which was available for their study. Input windows of 51, 71 and 101 nucleotides were tried.

The final step in this study was to analyze the data using neural network models. A neural network simulator developed at the University of Maryland and partially designed by one of the authors was used in the analysis of the data. This simulator was chosen for its data handling capabilities and for its familiarity. For training the neural network only a subset of the training data was used. This subset consisted of ~21,000 different patterns and was used as the testing data, 651 (91.7%) were AUG, 49 (6.9%) were GUG and 10 (1.4%) were UUG. From the aligned training data, a consensus binding site was also determined. It is presented below, along with a consensus developed by Scherer and co-workers in 1980 (17) on the second line:

\[
\text{COCOA AATUCAUUUU OAAUCUMAA AAUAAAAAO UTOAAAAAOC •ugAADAAOA AAAUUAIJUAU CCDCQ}
\]

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>AUG</th>
<th>GUG</th>
<th>UUG</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>U</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

For the perceptron, for each nucleotide position four input units were used. Only one of these four was set to 1.0 to represent the specific base. All others were set to 0.0. As an example, a network which examined 51 nt at a time would have 204 total input units, with only 51 of those set to 1.0 for each presentation. At the beginning or end of a sequence there was a need to represent a blank or indicate that there was no nucleotide present at a specific position. To encode this at a specific nucleotide location, all of the input units were set to 0.0.

The weights of each network were initialized with pseudo-random numbers between 0.1 and −0.1. The initial learning rate and momentum were 0.2 and 0.6 respectively. These were decreased during training to 0.1 and 0.4. No weight decay was used. The activation function for each unit was the logistic sigmoid  

$$A(x) = \frac{1}{1 + e^{-x}}$$

Various window sizes were explored, including sizes of 51, 71 and 101. For each of these sizes the networks were trained to predict whether the initiator codon began at position 31, 41 or 61 respectively in the input window. Networks were trained with one and two hidden layers. Those with a single hidden layer were tried with the number of hidden units adjusted to be 7, 9, 11 and 13. Four training runs were made with each different network architecture examined. During each training run, the network was presented with 60,000–100,000 pattern presentations. The weights of the network were saved 20 times during the training run at evenly spaced intervals. The best performers of these networks on the testing set are reported in the results section.

**RESULTS**

After the data was prepared and split into two sets, all ribosome binding sites were extracted and aligned for visual examination. Any that deviated significantly in appearance from the norm were checked against the referenced publication for the GenBank entry they appeared in. A total of 41 GenBank-annotated sites disagreed with the referenced publication. The main source of error was found on complementary sequences, where the start codon was annotated to begin one codon before or after the actual start codon. Other mistakes were variable, ranging from annotated sites which actually did not exist to sites where there was little evidence to indicate that a gene start was actually present. The latter case especially is a problem, since some authors annotate the start of every open reading frame on their sequence as being a gene start, even though there is no experimental evidence to support the presence of a translated region.

While the sites were aligned, a count was taken of the different initiator codons. Out of 710 total ribosome binding sites in both the testing and training data, 651 (91.7%) were AUG, 49 (6.9%) were GUG and 10 (1.4%) were UUG. From the aligned training data, a consensus binding site was also determined. It is presented below, along with a consensus developed by Scherer and co-workers in 1980 (17) on the second line:

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>AUG</th>
<th>GUG</th>
<th>UUG</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>U</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

It is clear from both consensus determinations that the ribosome binding sites are heavily biased towards A and U nucleotides in the region −35 to +20. Both determinations are in strong agreement on this property, though, as might be expected, there are many differences at the individual base positions. It should
also be noted here that Scherer and co-workers used a procedure for determining their consensus that preserved the Shine–Dalgarno sequence.

Each model tested in this study produces a quantitative value when a specific input sequence is evaluated. Some threshold is then selected by the modeler to determine actual sites from non-sites. If the value produced by the model is above the threshold, the input sequence is predicted to be a site, otherwise it is predicted by the model to be a non-site. If performance is reported for such a model using just one threshold, it is difficult to compare against another model, which may use a different threshold. Additionally, some models may perform better than others on weaker sites, but worse on stronger sites. Unfortunately, this has been the method of comparison with previous studies on molecular sequence analysis. To solve this problem, the performance of each model tested in this study is reported in the form of a ‘receiver-operating-characteristic’ (ROC) graph (49,50). It displays performance over a continuous range of thresholds by plotting the percentage of sites correctly identified versus the false positives. This method allows comparison over the entire range of thresholds. Better performing systems are easily identified, since they will have a larger area under the curve. It also allows comparison with future systems by other investigators.

The first step in the analysis was to analyze the testing data with the best method produced by Stormo and co-workers in their previous study of the problem. Their best perceptron used an input window of 101 nt with 60 of those nucleotides being upstream of the recognition site. Their data is reported in Figure 4, along with the results from the other procedures. The ROC curve for the perceptron is clearly shifted towards poorer performance than the other methods. When the threshold was adjusted to yield 75% correct identifications (135 out of 180), 0.79% false positives (1660 out of 210 000) were also identified.

When examined across a range of thresholds, such as is found in the ROC curve, a much clearer picture is given of where each method of Stormo and co-workers, but trained the perceptron on stronger sites. Unfortunately, this has been the method of comparison with previous studies on molecular sequence analysis. To solve this problem, the performance of each model tested in this study is reported in the form of a ‘receiver-operating-characteristic’ (ROC) graph (49,50). It displays performance over a continuous range of thresholds by plotting the percentage of sites correctly identified versus the false positives. This method allows comparison over the entire range of thresholds. Better performing systems are easily identified, since they will have a larger area under the curve. It also allows comparison with future systems by other investigators.

The next step in the analysis looked at the perceptron training method of Stormo and co-workers, but trained the perceptron on the data collected for this present study instead of the limited data which was available for their study. Input windows of 51, 71 and 101 nt were used. For each window size the perceptron was trained until the number of changes made on each pass through the data plateaued. At various points after training had reached a plateau the perceptron was tested on the test data and the results of the best performance were saved. Several perceptrons were tried for each window size and the best performer was selected for final comparison. The ROC curves for each window size are given in Figure 1. The performance difference based on window size is so small that it is hardly worth mentioning. The better performing perceptrons come from window sizes of 51 and 101 nt. These two were very close, with the larger window size performing better on sites that are easier to recognize, whereas the smaller window size is better for sites that are more difficult to recognize. In their 1982 study, Stormo and co-workers (44) felt that a window size of 101 was significantly better. They made this conclusion, however, based on performance at a single threshold. When examined across a range of thresholds, such as is found in the ROC curve, a much clearer picture is given of where each window size performs best. In the case of the perceptrons, there is no clear winner between sizes of 101 or 51. The performance of the perceptron with a window size of 101 would probably be of more use to molecular biologists, due to its better performance with sites in the middle percentages of recognition. For this reason alone, it was selected for comparison with the other methods, as can be seen in Figure 4. It performed better than the perceptron of Stormo et al. In order to recognize 75% of the ribosome binding sites (135 out of 180), about 0.20% false positives (415 out of 210 000) are also identified.

The final step in the analysis examined the use of neural networks. Input windows of 51, 71 and 101 nt were used. A plot of the performance improvement observed as training progresses is seen in Figure 2 for a network with an input window of 51 nt and 13 hidden units. This network reached convergence after being trained on 50 000 training patterns. Little if any improvement from further training is observed after this point. Training plots for all the networks in this study were similar. For each window size the number of units in the hidden layer was varied between 7 and 13. The differences in performance for hidden units of 7, 9, 11 and 13, with the input window size held constant, were very close. Generally, networks with 9 and 11 hidden units appeared to perform better. Close performances were also observed when the input window size was varied. The best performing networks with input window sizes of 51, 71 and 101 are given in Figure 3. For the networks that used an input window of 51 nt, the best performing network had a single hidden layer with 9 hidden units. For an input window of 71 nt the best performing network had a single hidden layer with 11 hidden units. The best performing neural networks used an input window of 101 nt and a single hidden layer of 9 or 11 units. These networks required 100 000 training iterations to achieve their performance. The performance data for a neural network with an input window of 101 and a hidden layer of 9 units is plotted in Figure 4, along with the best of the other methods tried. It performed well, scoring 75% correct identifications with 0.08% false positives (172 out of 210 000).

Figure 5 graphs the scores produced by the best neural network across one cistron of a polycistronic mRNA taken from the test set. The particular locus illustrated is Ecokdpaec, an operon with 4933 bases and three cistrons. Graphs of the other two cistrons were very similar. It is apparent that for this particular sequence.
the ribosome binding sites have a very high score, whereas the other portions of the sequence are very low.

Some interesting information was found by examining some of the false positives produced by the neural network in its analysis of the testing set. In Figure 6 are some of the false positives which were sampled from the larger set of 170 that was produced by using a threshold of 0.09 with the best performing neural network. The closer that each predicted score is to 1.0, the more the neural network predicted the sequence to be a binding site. Each of the sampled false positives was examined closely to determine if it was actually a false positive and, if so, why it was mistakenly identified.

A significant portion of the false positives were actually competing initiator codons, both in-frame and out-of-frame. Initiation actually does take place at these codons, but for consistency only the primary initiator codon was accepted as a correct prediction. Competitive initiator codons accounted for ~18% of the false positives. The scores produced by the neural network for these sequences were very high.

Another significant portion of false positives came from portions of the RNA sequence which were near the start or the end of the mRNA. Apparently, the neural network learned that the ends, and especially the 5'-end, are likely to have a ribosome binding site in the vicinity. In those few cases where this is not the case, it can lead to a false positive.

One of the false positives in Figure 6 appeared to be an actual ribosome binding site which had yet to be identified. It had a short open reading frame (ORF) which was translationally coupled to a much larger ORF which made up the last half of the α-glucan phosphorylase gene. The score for this site on the Ecoglpa sequence was very high, ~0.60. The higher scores appeared to come from competing initiators or from previously undiscovered sites. As a result, 14 of the highest scoring false positives were examined. Some of these are included in Figure 7. One of these 14 was a competitive initiator. Another five appeared to be genes which had not yet been identified. The lengths of the reading
frames, codon bias and the locations of the genes were carefully examined in drawing this conclusion.

DISCUSSION

The data preparation portion of this study turned out to be very time consuming. The number of entries which had to be hand checked was large. The number of errors found in the annotations of the sequences was significant. Judging from the number of these errors, which were relatively easy to detect, there are probably many more errors which are present, but which are not so easy to detect. This type of error could be something like a ribosome binding site which is present, but which is unannotated and unknown. A check of the false positives identified by the best neural network revealed that >20% of them are actually misannotated genes or what appear to be genes that have not previously been discovered. Conversely, some of the annotated ribosome binding sites are probably not binding sites at all. Though many of the algorithms used in this study deal well with noisy data, it still degrades performance and will remain a significant obstacle to progress on this problem and others in the field.

An important contribution of this study was the introduction of the ROC curve into this area of research. When used to compare performances of identifiers against each other, this method of visualizing the performance data allows comparison over the range of thresholds which identify from 0 to 100% of the ribosome binding sites. It not only provides a clear picture of which system performs better, but it also indicates where a specific system performs best. This can help in actual usage of the system, since thresholds can be selected that maximize the performance of the best system. The ROC curve allows other researchers to compare performance of their systems against the one developed for this study. This has been a problem in the sequence analysis area. An example of this problem is seen with the promoter recognition problem, where O'Neill (10) claimed a recognition rate of 77% with fewer than 0.1% false positives and that this rate was superior to the results of Lukashin and co-workers (9), since their system produced many more false positives, even though they had selected their threshold to identify 90% of the actual promoters. Comparison of these two systems is impossible, since O'Neill selected his threshold to minimize false positives, while Lukashin and co-workers selected their threshold to maximize promoter identification. Hopefully, this problem will be avoided in the future by using the ROC curve to report performance.

The most important contribution of this study is the improvement that neural networks provide in identifying ribosome binding sites. The performance graphs clearly indicate that the neural network finds more ribosome binding sites with fewer false positives, regardless of where the threshold is set. Selecting thresholds so that both the neural network and perceptron identify 75% of the testing sites produces 415 false positives for the perceptron, but only 172 for the neural network. Additionally, the best performing perceptron, which used an input window of 100 nt, was significantly outperformed by all the neural networks applied in this study, regardless of the input window size or size of the hidden layer. Perhaps neural networks with a larger window size may actually perform better. Further study will determine this.

There is a limit to the success which can be expected from neural networks or any other network which takes as input only primary sequence information. Secondary structure plays a major role in ribosome scanning and recognition. Improvements in automated recognition algorithms will be limited until this secondary structure information is incorporated into the model. Hopefully, further advances in the ability to successfully determine the secondary structure of long mRNAs will make this possible in the future.

The improvement found with a neural network should not detract from the insight of Stormo and co-workers in their application of perceptrons. The backpropagation algorithm was
not widely known until the mid eighties and computational power was not sufficient for a problem of this magnitude at that time. This type of study would not have been possible back then, or for many years after.

The proper application of neural networks to molecular sequence analysis has been and will continue to be a fruitful avenue of research in expanding our understanding of this area. They will also provide improved tools for molecular biologists. The improved results of this study added to ORF searches and codon bias information could significantly improve upon present methods of locating genes in E.coli.

ACKNOWLEDGEMENTS

The authors would like to thank David Rumelhart, Stanford University, and the National Cancer Institute’s Division of Mathematical Biology for computer and technical support of this project. A portion of this research was completed as part of David Bisan’s graduate training at The George Washington University. An executable copy of the identification program (Sun Sparc implementation) is available from the first author for $100.00. It will be made available for other systems at a later date.

REFERENCES