To appear in: Nucleic Acids Research, 1995.

Identification of Ribosome Binding Sites in *Escherichia coli* Using Neural Network Models

David Bisant
Neuroscience Program (151 B)
Stanford University
Stanford, CA 94305
bisant@decatur.stanford.edu

Jacob Maizel
National Cancer Institute, FCRF
Bldg 469 Rm 151, PO Box B
Frederick, MD 21701
jmaizel@ncifcrf.gov

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 *Escherichia 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 to 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 nonhomologous test set, about 68%, was quite good.

Since their study, several researchers have improved upon their results by incorporating nonlocal information into the network input [2, 3, 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 *E.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 nucleotides, his trained network was about 77% successful at identifying new promoters with a false positive rate of about 0.1%. This performance was about 8% better than the previous best method which was a hand-crafted rule-based system also developed by O'Neill and 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 to ad hoc or hand-crafted systems.

EARLY AND CURRENT RESEARCH ON RIBOSOME BINDING SITES

Several reviews [15, 16, 17] have described the early research on translation in *Escherichia coli* and other prokaryotes. The 1981 review by Gold and coworkers [15] and the 1982 review by Stormo and coworkers [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 (RBS), 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, 20, 21]. In these experiments, complexes formed of phage mRNA, ribosomes, and fMet-tRNA$_{fMet}$ were exposed to ribonuclease. The 20–40 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 coworkers 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 *Escherichia coli* 16S ribosomal RNA (rRNA). They hypothesized that the terminal sequence they found, ACCUCCUUA-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* lactose 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, Buzash-Pollert, and Studier [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 a thousand fold more efficiently even through 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, 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, Kacich, and Ptashne[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 to 20 times more efficient in the 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, Kacich, and Ptashne study mentioned above, by placing these elements in the single-stranded portions. Unfortunately these studies did not test their hypotheses with 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 coworkers 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   UUAAUCUUUU AAAAAUuaag gagguAUAUU augAAAAAAA UUAAAAAAC
ECODKSA     UGAACGAGGU CAGAAACACU uCGggagUuG augAGCGUAG CGGCUGAAG
ECOMICA     UUGCCCCCAA CAACAGUGaa UUCgguGACC augCAAGCGU CGCAAUUUU
ECOPUV             AAU AUUCCUCCCU UuGaCgaCgA augCUUAUGU CUAUACAGA
ECOPURCA    GCUUCCAUGC AGCGUUACAG CaCggagAuG augAACGUUA UUUCCGCCG
ECOPOLBDA         CUGG UUUUUUGaUg gaUUUUCAGC gugGCGCAGG CAGGUUUUA
ECOPURCA    AACGUUAAUU CACACCCagg aguGAUAAAG augCAAAAGC AAGCUGAGU
ECORPSOP    UACUGUCUAA GAAAGAGAaa ggaUAUUACA uugCUUAAUC CGAUCGUUC
```

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

1. the Shine-Dalgarno element [34]
2. the start codon and the bases immediately surrounding it [35]
3. the presence of multiple competing initiator codons [36]
4. the distance between the start codon and the Shine-Dalgarno [37]
5. the sequence of the bases in between the start codon and the Shine-Dalgarno [36, 38]
6. the composition of the bases far downstream [39]
7. the composition of the bases far upstream [40]
8. the secondary structure [41, 35]
9. 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 1980's. 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 his coworkers identified 124 ribosome binding sites. For each of their 124 ribosome binding sites, instead of directly counting the complementarity with the Shine-Dalgarno, they calculated the free energy of each ribosome binding site coupled with 16S. 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.

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, this region was especially absent of G residues outside of the Shine-Dalgarno 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 at the Shine-Dalgarno 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 mRNA's 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 3 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 8 sequences with a total of 10 ribosome binding sites.

The researchers found that a window of 101 worked best, identifying 7 of the 10 genes with 5 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 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 his coworkers 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 the 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 would 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 labelled 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 modelled 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 analyzed with the published perceptron of Stormo and coworkers, 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 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 coworkers [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 assist in this process, a map of Ecoseq 7.0 [47, 48] was used. Ecoseq is a database of *E. coli* sequences represented in global coordinates with many of the ambiguities and redundancies resolved when compared to the Genbank entries. When consulting the Ecoseq map and the sequence entry was not enough to determine the transcribed areas, the actual journal article referenced in the Genbank entry was examined. A list of the transcribed regions for each locus was compiled. In many cases, it was impossible to precisely tell where a transcribed region started or ended. In such cases, a conservative guess was made 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 approximately 590,000 bases of transcribed RNA in the training file and approximately 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 a 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 his coworkers 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 his coworkers, 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 nucleotides were tried.

The final step in this study was to analyze the data using neural network models. A neural network simulator developed at University of Maryland and partially designed by the author, 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 about 21,000 different patterns and was made up of ribosome binding sites, non-sites which were very similar to the ribosome binding sites, and random sequences selected from the training data. Unlike the other algorithms such as the perceptron, the backpropagation algorithm used for training the neural networks minimizes the squared error. If the positive examples presented to the network are underrepresented, as in the case of ribosome binding sites, the network will bias itself to the overrepresented training examples. Hence, only a subset of the training data was used for this investigation so that the ribosome binding sites would be more evenly represented. Additionally, the convergence of a network can be much faster with a subset of the data.

A local encoding scheme was used to encode the sequence so it could be presented to the network. Much like the scheme used for the perceptron, for each nucleotide position, 4 input units were used. Only 1 of these 4 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 nucleotides 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 pseudorandom 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 variety of network structures were tried. 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 the 31st, 41st, or 61st position, respectively, in the input window. Networks were tried with 1 and 2 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 to 100 thousand 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 is especially 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 coworkers in 1980 [17] on the second line:

```
CUGGA AACUCAUUUU UAAUCUAAAA AAUAAAAAAG UUUAAAAAUC augAAUAAUA AAAUUAUUAU CCUCG
                            UU UU AAAAAUUAAG GAGGUAUAUU augAAAAAAA UUAAAAAACU GAA
```

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 coworkers 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 modeller 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 percent 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 his coworkers in their previous study of the problem. Their best perceptron used an input window of 101 nucleotides 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 (1,660 out of 210,000) were also identified.

The next step in the analysis looked at the perceptron training method of Stormo and his coworkers, 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 nucleotides 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

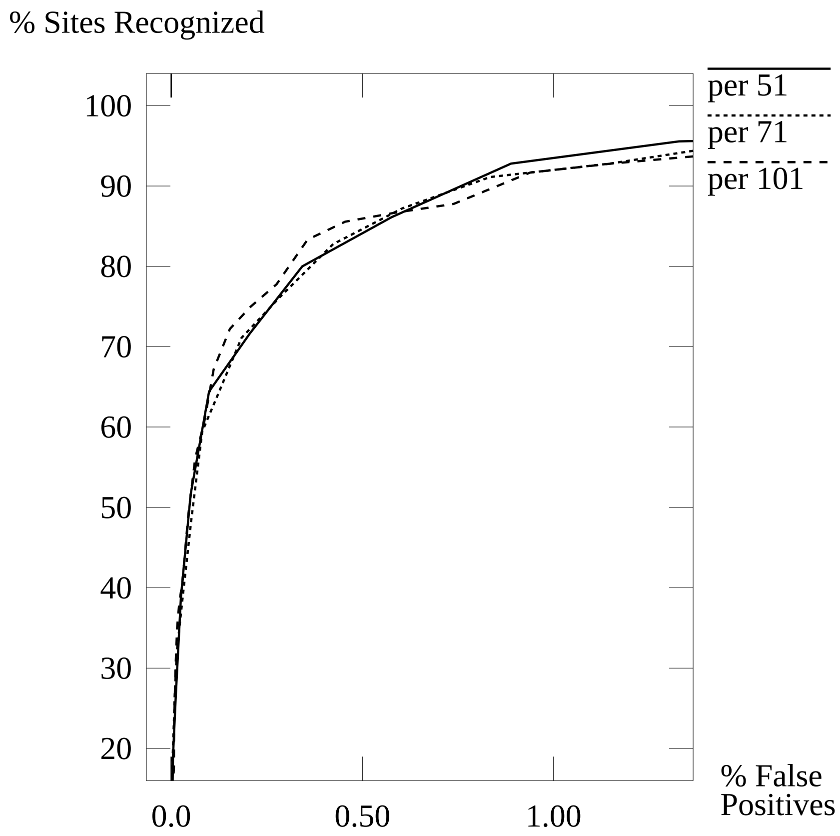

Figure 1: Test data ROC curves for perceptrons with window size of 51, 71, and 101.

worth mentioning. The better performing perceptrons come from window sizes of 51 and

101 nucleotides. 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 coworkers [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 nucleotides 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 nucleotides and 13 hidden units. This network reached convergence after being trained on 50,000 training pat-

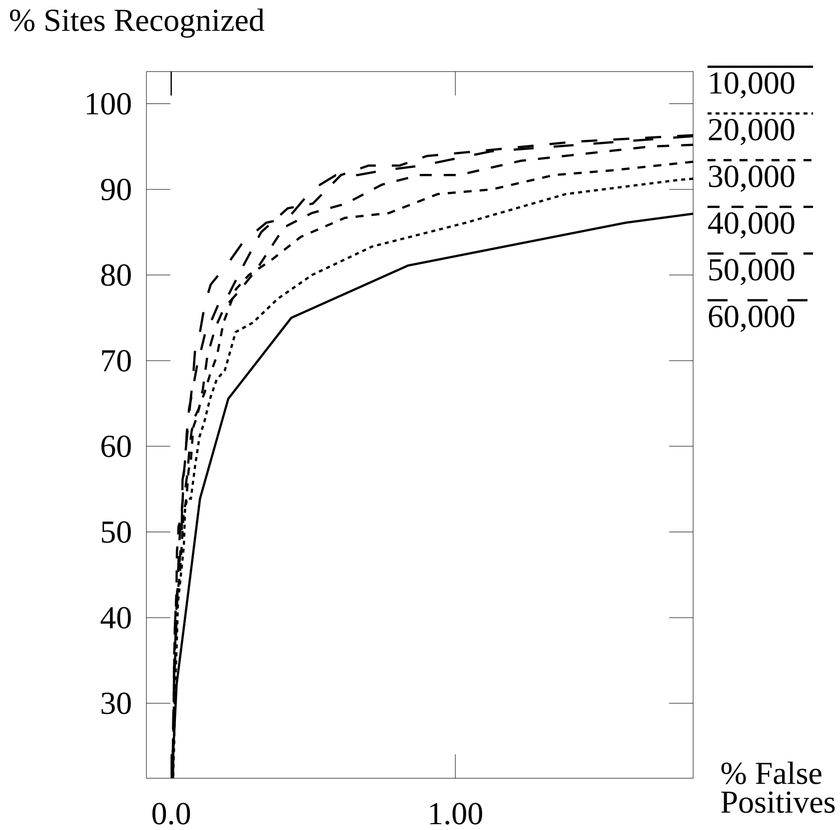

Figure 2: Test data ROC curves plotted at different points during training for a typical neural network. This network had an input window of 51 with 13 hidden units.

terns. 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 difference in performance for hidden units of 7, 9, 11, and 13, with the input window size held constant, was 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 nucleotides, the best performing network had a single hidden layer with 9 hidden units. For an input window of 71 nucleotides, the best performing network had a single hidden layer with 11 hidden units. The best performing neural networks used an input window of 101 nucleotides and a single

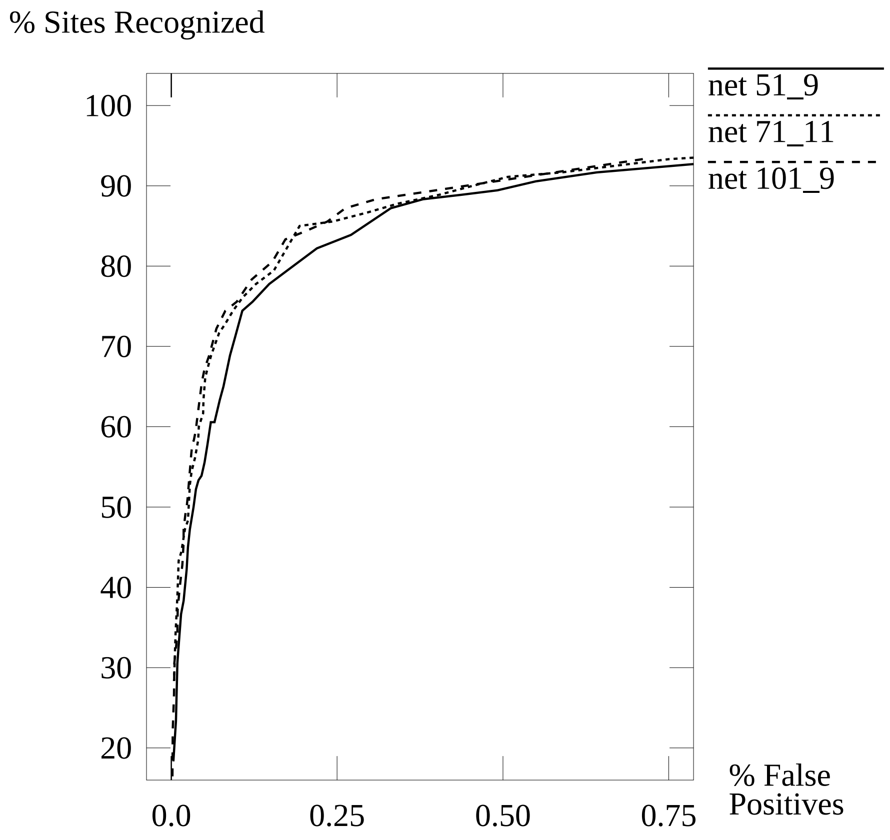

Figure 3: Test data ROC curves for the best neural network from each input window size.

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 Ecokdpabc, an operon with 4933 bases and 3 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.

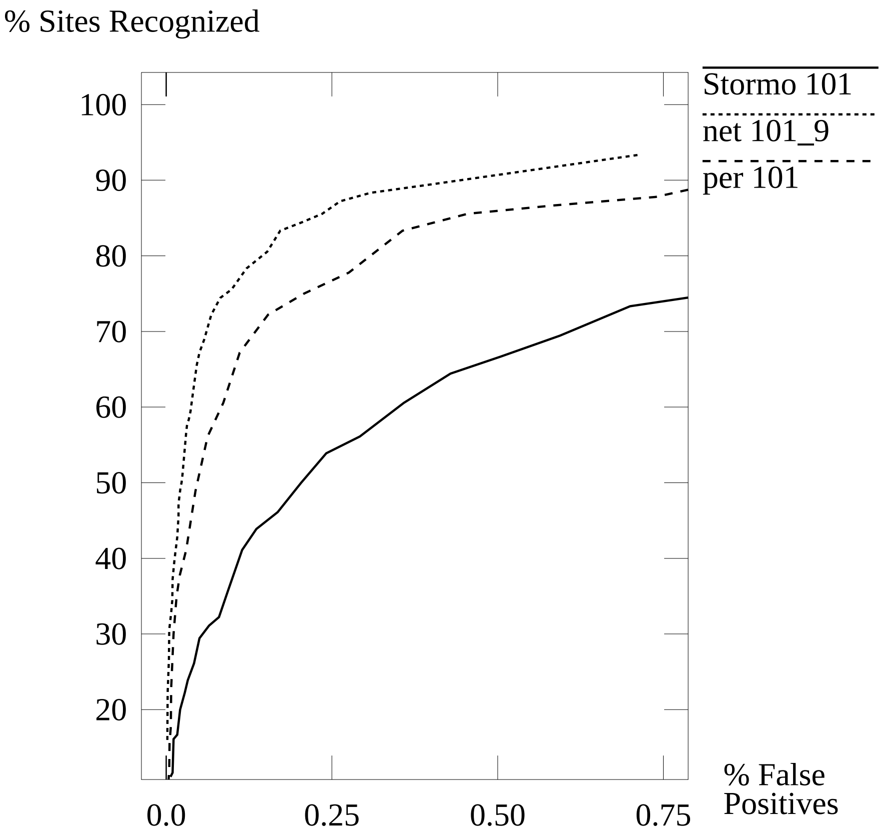

Figure 4: Test data ROC curves for the original perceptron of Stormo et al. (Stormo 101), the best perceptron trained on our new data (per 101), and the best neural network (net 101_9). Each used a window size of 101.

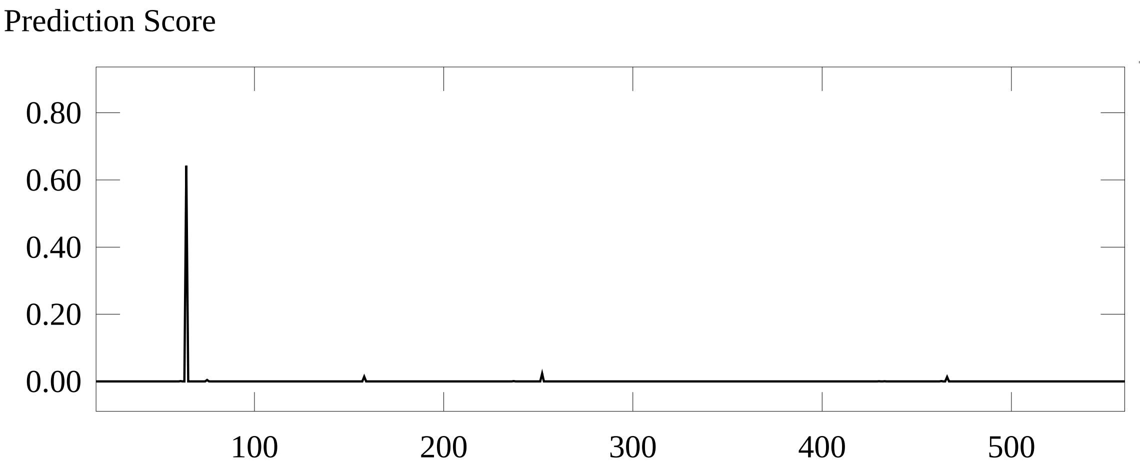

Figure 5: The neural network predictions for the first cistron of the Ecokdpabc locus. The ribosome binding site is located at position 65.

```
J01581  ECOALAS   0.1554      (not an RBS, short ORF))
AACUUUGUUGUUUACCAACGCCGGGAUGAACCAGUUCAAGGAUGUGUUCC UUGGGCUCGACAAGCGUAAUUAUUCCC

K02498  ECOALKA   0.1631      (not an RBS, start)
                                         GGAUGAAUA AGGAGAUGCGaugUAUACCCUGAACUG

M73326  ECOASCBFG_c   0.7901  (competing initiator)
                         GAUUGCGUGUAAAGGGAUGAAAaug AUGACGACGAUGCUGGAAGUGGCGAAG

M27058  ECOFUMB   0.7457      (competing out-of-frame initiator)
UCCACUUUGCCAAUAACGCCGUCCAUAAUCAGCCACGAGGUGCGCGaugA AUGCGACUGCAG

M55161  ECOKATE   0.2444      (not an RBS, A & U bias, note invalid initiator)
ACUCACUGCGUGCCGGUAGCCGUGGUCCAACGCUGCUGGAAGAUUUUAUU CUGCGCGAGAAAAUCACCCACUUUGAC

J03966  ECOGLGPA   0.5994     (a second start, translational coupling)
GUAUAUCGCCCAGCAGCUGAAUGUGGUGGUGAAUCCAAAGGCGUUGUUCG AUGUACAAAUCAAACGUAUUCACGAAU
```

Figure 6: A representative sample of false positives produced by the best neural network on the testing data. The accession and locus are given along with the score produced by the neural network and an explanation or description in parenthesis. The mistaken initiator has a space before it, and any true initiators present are in lower case.

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 approximately 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, then 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 alpha-glucan phosphorylase gene. The score for this site on the Ecoglgpa sequence was very high, around 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

```
M32253  ECOILVGMED  0.7952   (probable leader peptide, translational coupling)
                                            AGGAGCAGACA AUGAAUAACAGCACAAAAUUCUGUUUC

 K02670  ECOKDPABC   0.6098   (new protein, ORF > 160)
 CCAGCCGGUUGUCAACAUUGUUGAACUCAAUCUGGCGCUGGAUAAACUUG AUGAAUAA

 M33928  ECOPURCA    0.776    (new protein, ORF=133)
                GCAUGCCGGUUUGCUGUAAAGUUUAGGGAGAUUUG AUGGCUUACUCUGUUCAAAAGUCGCGC

 M19501  ECOPURLA   0.8946    (not an RBS, short ORF)
 CGUACUGUCAGCAACUCCUGGCAUCCGGAAAACUGGGGCGAGGAUGGCCC AUGGAUGCGCAUUUUCCGCAAUGCGCG

 K00032  ECOTNAA   0.9352     (out-of-frame competitor, translational coupling)
 GCAGAUGUAAUAUUCACAGGGAUCACUGUAAUUAAAAUAAAUGAAGGAUU AUGUAaugGAAAACUUUAAACAUCUCC

 M16487  ECOBGLO   0.8113     (not an RBS, short ORF)
 GAGCAUUACGCCCGCACGGUCUUUACUCGCUACCAACAUAAAGUGGCGUU AUGGCUGACGUUUAAUGAAAUCAACAU
```

Figure 7: Some of the highest scoring false positives produced by the neural network on the testing data.

in Figure 7. One of these 14 was a competitive initiator. Another 5 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 over 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 number of sites identified while minimizing the number of false positives.

This study also found that each system has a unique range of thresholds. If a threshold is selected for one system, that same threshold would not work the same on another system. In such a case, it is difficult to compare performance of the two systems. The ROC curve is not affected by this problem since it is independent of the actual threshold.

An example of what happens when ROC data is not available is seen with the perceptron study of Stormo and coworkers [44]. When they compared the performances of each of their perceptrons, they selected a single threshold and evaluated each system at that threshold. Since the perceptron with an input window of 101 performed best, they concluded that there is primary sequence information outside of the ribosome-protected segment which the perceptron is using and which must influence translation initiation. Other researchers have also used this conclusion in designing their experiments. The analysis of perceptrons in this study could very well have come to the same conclusion if the different perceptrons had been evaluated at a single threshold. The ROC curves, however, indicated that the perceptrons with input windows of 51 and 101 are essentially equal in performance. This could mean that no primary sequence information outside of the ribosome-protected area is used in translation initiation. A critical reexamination of this issue could be in order. Lastly, an important additional benefit of the ROC curve is that it 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 coworkers [9] since their system produced many more false positives even though they had selected their threshold to identify 99% of the actual promoters. Comparison of these two systems is impossible, since O'Neill selected his threshold to minimize false positives while Lukashin and coworkers 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 at 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 that used an input window of 100 nucleotides 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. Improvement 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 mRNA's will make this possible in the future.

The improvement found with a neural network should not detract from the insight of Stormo and his coworkers 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 Bisant'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. It will be made available for other systems at a later date.

References

[1] Qian, N. and Sejnowski, T. J. (1988) *Journal of Molecular Biology* **202**, 865–884.

[2] Kneller, D. G., Cohen, F. E., and Langridge, R. (1990) *Journal of Molecular Biology* **214**, 171–181.

[3] Vieth, M., Kolinski, A., Skolnick, J., and Sikorski, A. (1992) *Acta Biochimica, Polonica* **39(4)**, 369–392.

[4] Rost, B. and Sander, C. (1993) *Proceedings of the National Academy of Science* **90**, 7558–7562.

[5] McGregor, M. J., Flores, T. P., and Sternberg, M. J. E. (1989) *Protein Engineering* **2(7)**, 521–526.

[6] Muskal, S. M., Holbrook, S. R., and Kim, S. (1990) *Protein Engineering* **3(8)**, 667–672.

[7] Holbrook, S. R., Muskal, S. M., and Kim, S. (1990) *Protein Engineering* **3(8)**, 659–665.

[8] Ezhov, A. A., Kalambet, Y. A., and Cherny, D. I. (1989) *studia biophysica* **129(2-3)**, 183–192.

[9] Lukashin, A. V., Anshelevich, V. V., Amirikyan, B. R., Gragerov, A. I., and Frank-Kamenetskii, M. D. (1989) *Journal of Biomolecular Structure & Dynamics* **6(6)**, 1123–1133.

[10] O'Neill, M. C. (1991) *Nucleic Acids Research* **19(2)**, 313–318.

[11] O'Neill, M. C. (1989) *Journal of Molecular Biology* **207**, 301–310.

[12] O, Neill, M. C. and Chiafari, F. (1989) *The Journal of Biological Chemistry* **264(10)**, 5531–5534.

[13] Brunak, S. and Engelbrecht, J. (1991) *Journal of Molecular Biology* **220**, 49–65.

[14] Snyder, E. E. and Stormo, G. D. (1992) *Nucleic Acids Research* **21(3)**, 607–613.

[15] Gold, L., Pribnow, D., Schneider, T., Shinedline, S., Singer, B. S., and Stormo, G. (1981) *Annual Review of Microbiology* **35**, 365–403.

[16] Lodish, H. F. (1976) *Annual Review of Biochemistry* **45**, 39–72.

[17] Scherer, G. F. E., Walkinshaw, M. D., Arnott, S., and Morre, D. J. (1980) *Nucleic Acids Research* **8(17)**, 3895–3907.

[18] Stormo, G. D., Schneider, T. D., and Gold, L. M. (1982) *Nucleic Acids Research* **10(9)**, 2971–2996.

[19] Gupta, S. L., Chen, J., Schaefer, L., Lengyel, P., and Weissman, S. M. (1970) *Biochemical and Biophysical Research Communications* **39(5)**, 883–888.

[20] Hindley, J. and Staples, D. H. (1969) *Nature* **224**, 964–967.

[21] Steitz, J. A. (1969) *Nature* **224**, 957–964.

[22] Shine, J. and Dalgarno, L. (1974) *Proceedings of the National Academy of Science* **71(4)**, 1342–1346.

[23] Steitz, J. A. and Jakes, K. (1975) *Biochemistry* **72(12)**, 4734–4738.

[24] deSmit, M. H. and vanDuin, J. (1994) *Journal of Molecular Biology* **235**, 173–184.

[25] Steege, D. A. (1977) *Proceedings of the National Academy of Science* **75(10)**, 4163–4167.

[26] Dunn, J. J., Buzash-Pollert, E., and Studier, F. W. (1978) *Proceedings of the National Academy of Science* **75(6)**, 2741–2745.

[27] Ray, P. N. and Pearson, M. L. (1974) *Journal of Molecular Biology* **85**, 163–175.

[28] Ray, P. N. and Pearson, M. L. (1975) *Nature* **253**, 647–650.

[29] Jansone, I., Berzin, V., Gribanov, V., and Gren, E. J. (1979) *Nucleic Acids Research* **6(5)**, 1747–1760.

[30] Borisova, G. P., Volkova, T. M., Berzin, V., Rosenthal, G., and Gren, E. J. (1979) *Nucleic Acids Research* **6(5)**, 1761–1774.

[31] Roberts, T. M., Kacich, R., and Ptashne, M. (1979) *Proceedings of the National Academy of Science* **76(2)**, 760–764.

[32] Selker, E. and Yanofsky, C. (1979) *Journal of Molecular Biology* **130**, 135–143.

[33] Iserentant, D. and Fiers, W. (1980) *Gene* **9**, 1–12.

[34] Gillam, S., Astell, C. R., Jahnke, P., Hutchison, C. A., and Smith, M. (1984) *Journal of Virology* **52(3)**, 892–896.

[35] Munson, L. M., Stormo, G. D., Niece, R. L., and Reznikoff, W. S. (1984) *Journal of Molecular Biology* **177**, 663–683.

[36] Matteucci, M. D. and Heyneker, H. L. (1983) *Nucleic Acids Research* **11(10)**, 3113–3121.

[37] Hui, A., Hayflick, J., Kindelspiel, K., and deBoer, H. A. (1984) *The EMBO Journal* **3(3)**, 623–629.

[38] Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. D., and Gold, L. (1992) *Molecular Microbiology* **6(9)**, 1219–1229.

[39] Petersen, C. (1987) *Molecular and General Genetics* **209**, 179–187.

[40] Coleman, J., Inouye, M., and Nakamura, K. (1985) *Journal of Molecular Biology* **181**, 139–143.

[41] Davis, M. A., Simons, R. W., and Kleckner, N. (1985) *Cell* **43**, 379–387.

[42] Thanaraj, T. A. and Pandit, M. W. (1989) *Nucleic Acids Research* **17(8)**, 2973–2985.

[43] Ivanov, I. G., Alexandrova, R., Dragulev, B., Leclerc, D., Saraffova, A., Maximova, V., and Abouhaidar, M. G. (1992) *European Journal of Biochemistry* **209**, 151–156.

[44] Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982) *Nucleic Acids Research* **10**, 2997–3011.

[45] Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W. (1972) *Nature* **237**, 82–88.

[46] Bilofsky, H. S. and Burks, C. (1988) *Nucleic Acids Research* **16**, 1861–1864.

[47] Rudd, K. E., Miller, W., Werner, C., Ostell, J. Tolstoshev, C., and Satterfield, S. G. (1991) *Nucleic Acids Research* **19**, 637–647.

[48] Rudd, K. E. (1993) *ASM News* **59**, 335–341.

[49] Green, D. M. and Swets, J. A. (1988) Signal Detection Theory and Psychophysics, Peninsula Publishing, Los Altos, California.

[50] Meistrell, M. L. (1989) *Computer Methods and Programs in Biomedicine* **32**, 73–80.