Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 15, Issue Number 3, (1997)

Calliper Randomization: An Artificial Neural Network Based Analysis of *E. coli* Ribosome Binding Sites

http://www.albany.edu/chemistry/sarma/jbsd.html

T. Murlidharan Nair*
Department of Medicinal Chemistry,
308 Skaggs Hall,
University of Utah,
Salt Lake City,
Utah 84112
USA

Abstract

An artificial neural network based approach has been used in analyzing the translation initiation region of *E. coli*. The approach is based on using a trained network capable of recognizing a particular region and presenting the network with randomized calliper inputs of the true sequence. The network responds with an error when the regions which have been the main source of knowledge are randomized. Analysis of the *E. coli* ribosome binding sites using this approach reveal that the initiation codon and the Shine/Dalgarno sequence which are known to be important for translation initiation are also important in imparting knowledge to the network. Further, selectively changing the usually occurring initiation codon AUG, to GUG, UUG and AUU, which occur less frequently, decreases the network performance in accordance with the frequency of their occurrence. This approach can be used as a general method to derive consensus.

Introduction

Macromolecular binding to specific sites of DNA/RNA involves the recognition of a specific sequence pattern. This sequence pattern recognition is seen in almost all macromolecular binding processes (repressors, polymerases, ribosomes, etc.). In some cases the sequence pattern may be very distinct while in others the sequence pattern may be diffuse. In studying molecular binding sites in DNA or RNA it is a conventional practice to derive a consensus by aligning an ensemble of sequences recognized by a common macromolecule. It is most often found that the sequence pattern is never completely conserved. In the case of *E. coli* translation initiation codons, the first position has 95% A, 5%G, 1%U and 0%C. (1). Initiation of protein biosynthesis plays a major role in the process of gene expression. The initiation process requires the formation of a complex between a ribosome, an mRNA and an aminoacylated initiator tRNA. The site of formation of this complex on mRNA is protected against nuclease attack and is known as the ribosome binding site (RBS). This region which usually extends about 35-40 nucleotides has the initiation codon located at about two-thirds of this region. Most of them contain the AUG triplet as the initiation codon. Others contain GUG or UUG, with a very few having AUA and AUU as their initiation codons (2-5). Gold *et al.*, (4) and Stormo *et al.*, (6) have compressed the histogram representing the frequency of occurrence of each base at each position into a single curve using a X^2 function. Schneider *et al.*, (1) have evaluated the information content of sites recognized by a particular macromolecule. Sequence logos have also been used in representing consensus for these regions (7). Perceptron algorithms have been employed to distinguish translation initiation sites in *E.coli* (8). Bisant and Maizal have developed a neural network model using a new set of ribosome binding sites (9).

Another important feature of the RBSs is a polypurine sequence called the Shine-Dalgarno (SD) sequence, occurring 5′ to the initiation codon and is complementary to nucleotides at the 3′ end of 16S rRNA (10). A few RBSs that do not seem to har-

*For author correspondence. Phone: 801-581-5301; Fax: 801-581-7087; E-mail: nair@alanine.pharm.utah.edu

bor SD sequence have also been reported (11,12). While rigorous statistical analysis of ribosome binding sites has not resulted in an exact consensus, many important features in the process of recognition have come forth. Recognition of binding sites based on primary sequence data is difficult since many positions in a consensus sequence of the site are degenerate and there are multiple determinants that define them (13).

The present study uses a neural network based approach in deriving the sequences that are important in the process of recognition by a trained network by employing a calliper randomization approach. Neural networks which are mathematical approximations of a biological synapse were initially developed to simulate the brain's learning process (14, 15). This massively parallel computational device has since been exploited in recognizing patterns rather than in understanding brain function *per se*. The application of neural networks in solving computational problems in biology and in other fields exceeds its biological significance (16-19). The calliper randomization approach detailed here is based on the assumption that the sequences important in the process of recognition by a neural net are also biologically important. This approach has been used in the analysis of the RBSs and the results are suggestive of the fact that the approach can be used as a general method in deriving consensus.

Materials and Methods

Data

The data for training the network were taken from the compilation by Rudd and Schneider (20). Out of the total of 1055 translation initiation sequences (of length 40), 500 sequences were used for training the network. The remaining 555 sequences were used as a test data set. A pseudo-random number generator was used in constructing random sequences with equal composition of all the four nucleotides. The random sequences were combined with the translation initiation sequences in a ratio of 1:4. Thus the total learning space comprised of 2000 different patterns. These sequences were presented to the network by coding them in binary similar to that used by Demeler and Zhou (21), called the CODE-4 representation. (C=0001; T=1000; A=0100; G=0010). The target to each translation initiation sequence was coded as 1 and 0 for a random sequence.

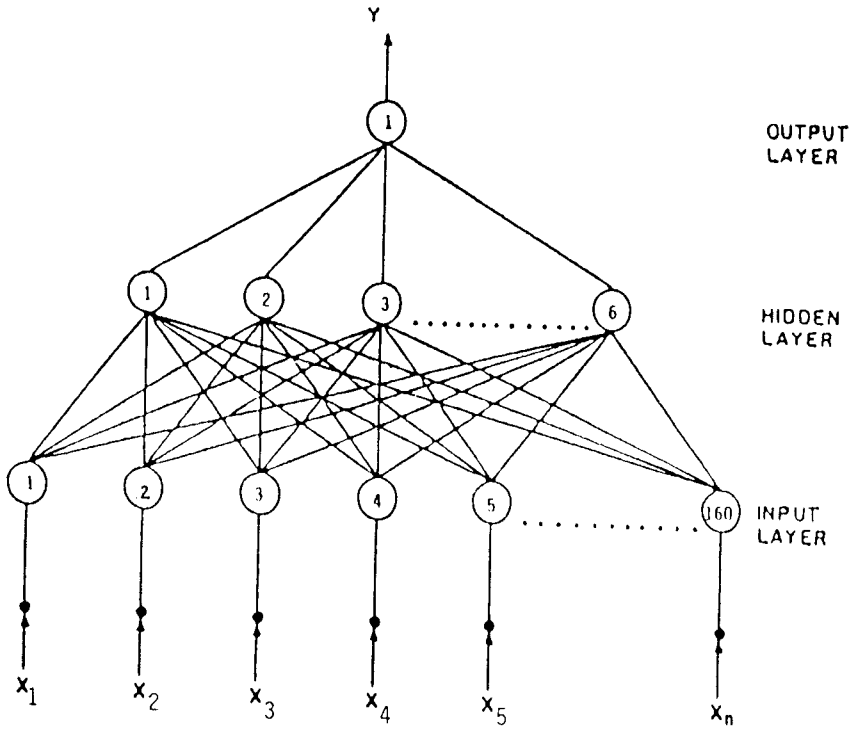


Figure 1: Architecture of a three layered feed forward neural network used in the simulation. The circles represent the artificial neurons which integrate input from the preceding layer and propagate the signal to the next layer.

The neural network simulations were done on a Silicon Graphics Indigo^2 workstation and programs used were written in FORTRAN. A multi-layered feed-forward type neural network was used for the simulation. The architecture of the network used is shown in Figure 1. The network consists of three layers of artificial neurons: input, hidden and output. Each neuron in one layer is connected to every other neuron in the following layer. These connections are represented by weights and thresholds. Training a feed-forward network involves presenting the network with an input pattern, calculating the network output by propagating the pattern through the network architecture, comparing the network output to the desired output or target, and using this difference to alter the weights in the direction which minimizes the difference between the actual output (network calculated) and the desired output (target). The algorithm used for training is the method of error back-propagation (EBP). This approach involves two passes through the network, a forward pass and a reverse pass. The forward pass is the least computation intensive and generates the network output. The reverse pass on the other hand is computationally very time consuming, since it not only involves propagating the error through the network but also assigns errors to each neuron that contributed to the initial error. The objective function that the EBP algorithm attempts to minimize, namely the summed squared error is defined as follows

$$E = \Sigma \Sigma (t_{ij} - out_{ij})^2 \qquad [1]$$

where the index i ranges over the set of input patterns and j ranges over the set of output neurons. t_{ij} is the target and out_{ij} is the network calculated output of the j^{th} neuron in the output layer when the i^{th} pattern is presented. A detailed description of the algorithm can be found in several articles (*e.g.* (14-16, 18)).

The calliper randomization approach involves randomizing a fixed window of sequence and presenting this sequence to the trained network. The calliper window is moved from one end of the sequence to the other, and the calliper randomized sequences are presented to the network for prediction. The error values for each window position is computed for the entire set of sequences.

Results and Discussion

A neural network has been trained to capture the internal representations of the translation initiation region. The network architecture consisted of 160 neurons

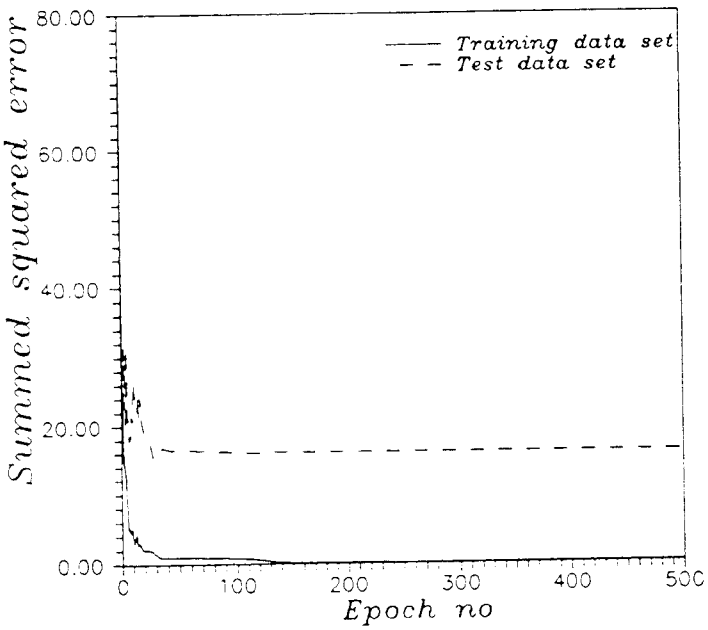


Figure 2: Error profile of the training and test data set.

(sequence length x 4) in the input layer, a hidden layer with 6 neurons and an output layer with 1 neuron. The number of neurons in the input and the output layers are fixed by the problem under consideration, while the neurons in the hidden layer are varied so as to obtain a generalized network with maximum prediction capability. Optimal predictions were obtained by using 6 neurons in the hidden layer. Increasing the number of neurons in the hidden layer, however did not increase the prediction capability of the network. Using too many neurons in the hidden layer would result in over parametrization, and would result in a network that memorizes the patterns, having poor generalization capabilities. Decreasing the number of neurons in the hidden layer hampered the prediction capability of the network. The momentum term was optimized to 0.9 and was fixed throughout the training process while the learning rate was decreased during the process of training to obtain optimal prediction capability and to prevent the network from getting stuck in any local minima. During the process of training the network was presented with a total of 2000 different patterns consisting of 500 translation initiation sites and 1500 random sequences. After every training epoch, which corresponds to propagating all the 2000 patterns once through the network architecture, the weights were extracted and used for determining the performance of the network. This was achieved by using these weights to predict the outcomes of the network when it was presented with a test data set, consisting of 555 translation initiation sites combined with 1110 random sequences. These sequences were not presented to the network during the process of training. A network output between 0.5 and 1 for a pattern suggested that the pattern is a translation initiation site, while an output less than 0.5 corresponded to a random sequence. Cross validating the network during the process of training, helps in capturing the weights of the network with maximum generalization capability. The error profiles of the training and the test data set are shown in Figure 2. The weights corresponding to the minimum error for the test data set were taken as optimal and used for the calliper randomization approach. The network predicted the test patterns with a very good degree of accuracy (96%) and there were no false positives. Further, another set of 4000 random sequences presented to the net were correctly predicted as random.

A trained network so obtained is a model-free tool with the capability of distinguishing a random sequence from a ribosome binding site sequence. While the trained network is able to distinguish a completely random sequence from a non-random sequences, would it be able to recognize partially randomized sequences? If the network acquires the knowledge to distinguish a random sequence from a non random sequence from a few sequences at particular positions, then randomizing these sequences should transform them into a random sequence. Further, if the

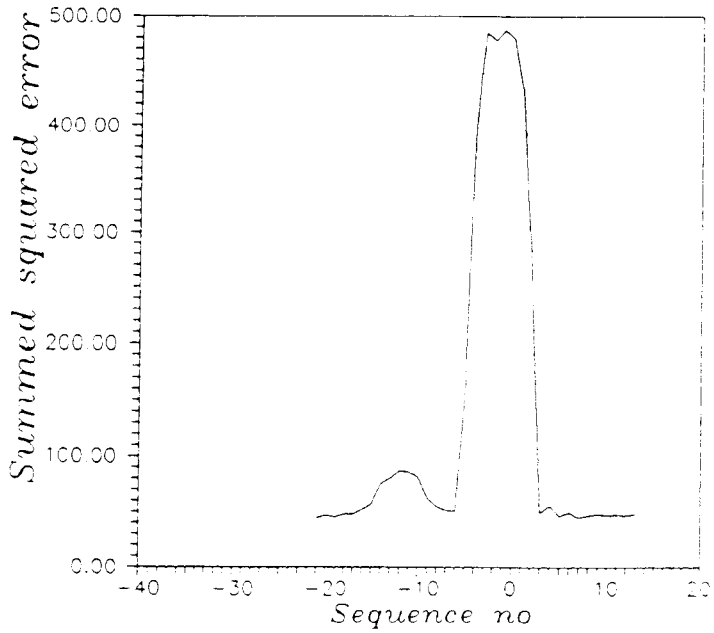
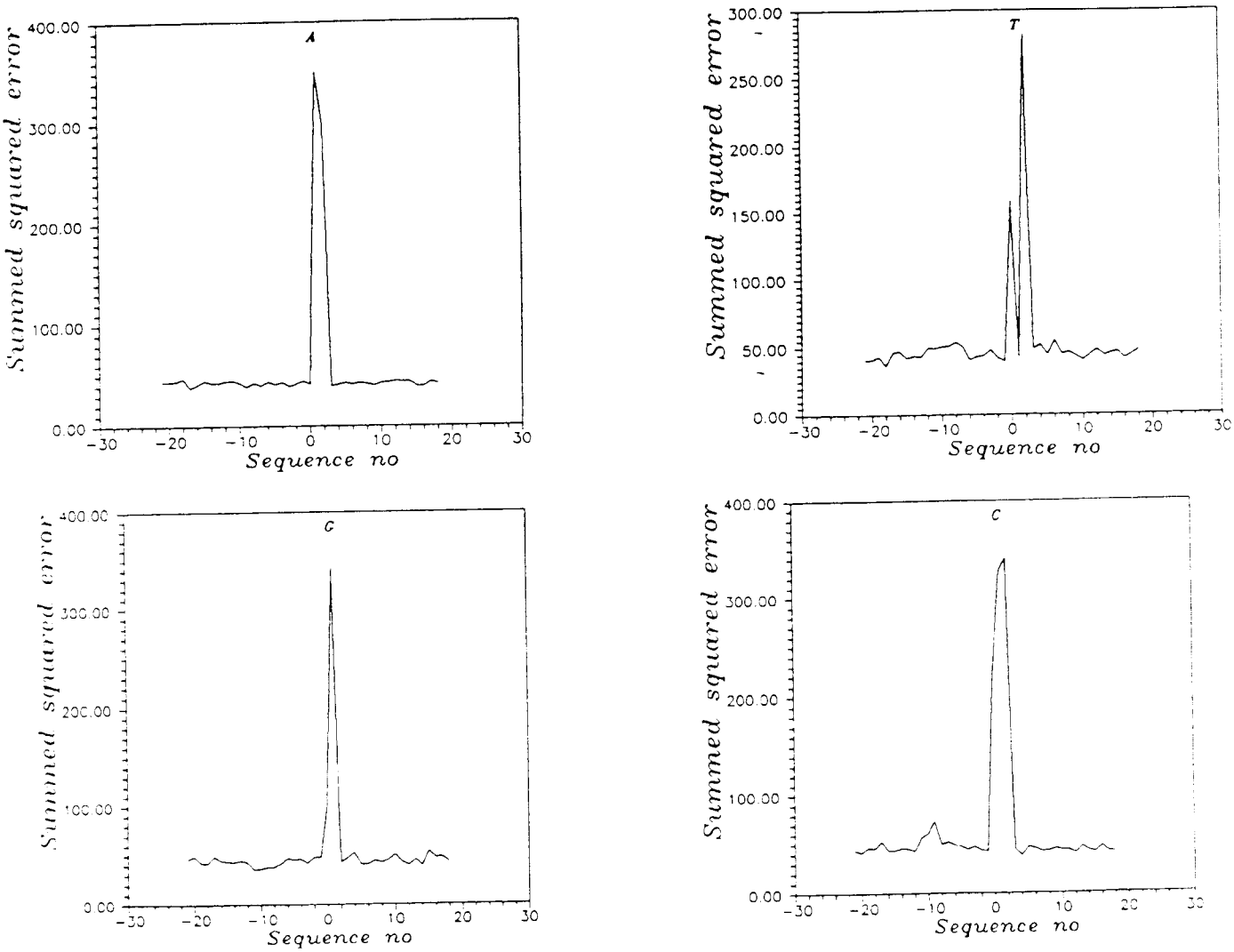


Figure 3: Performance of the network when a fixed calliper of sequence is randomized. The peaks correspond to the regions encompassing the initiation codon and the Shine/Dalgarno sequence.

sequences that impart knowledge to the network are ones that are biologically important, then randomizing these sequences would result in an approach to derive functionally and biologically important sequences. This is the basis of the calliper randomization approach. To test this assumption the trained network was presented with the ribosome binding site sequences that were randomized at fixed calliper lengths. The results of the network predictions for these partially randomized sequences are shown in Figure 3. The results are indicative of the fact that the network looses its prediction capability, when regions around the initiation codons are randomized. The calliper also encompasses the region preceding the initiation codon, which is a short polypurine stretch known as the "Shine-Dalgarno" (SD) sequence. The SD sequence is known to base pairs with the 3′ end of 16-S rRNA during initiation site selection. Further, the network was also presented with RBS sequences wherein each of the positions was substituted with all four nucleotides. The results of the network prediction are shown in Figure 4. In all the cases the network looses its prediction capability when the nucleotides of the initiation codon are changed. Thus making the initiation codon the most important in the process of recognition. Earlier reports on mutations that destroy the initiation codon, as seen in the case of bacteriophage T7 0.3 gene (22), the *E. coli trp* leader peptide (23), the bacteriophage T4 rIIB gene (24, 25), the lambda *xis* gene (26, 27) and the *Salmonella typhimurium his* leader peptide (28) have been shown to greatly reduce or abolish translation from the affected cistron. It is noteworthy to mention that randomizing a large portion of the region preceding the initiation codon did in fact cause a loss in the prediction capability of the network. However, mono-nucleotide substitution at these positions did not cause a significant loss in the recognition signal. These results point to the fact that the signals harbored in the Shine and

Figure 4: Performance of the network as a result of mono-nucleotide substitution at each position



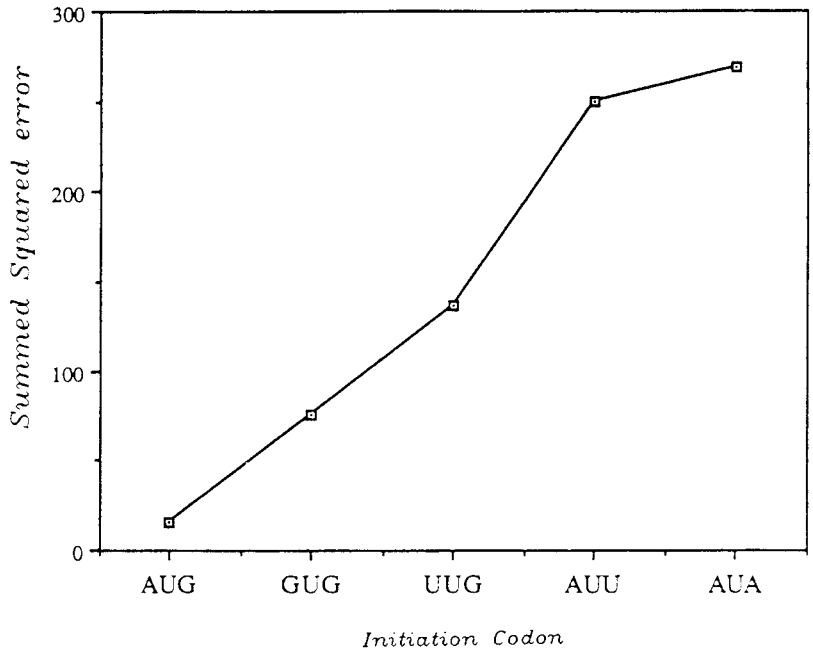


Figure 5: Performance of the network when different initiation codons were substituted in place of the naturally occurring ones.

Dalgarno region are not completely lost by the substitution of a single nucleotide. It is further interesting to note that when the network was presented with sequences wherein the initiation codon was changed to the less frequently used ones, the network prediction decreased in accordance with the frequency of codon used. The network prediction for the different codon substitutions are given in Figure 5. The prediction capability of the network further increased when all the initiation codons were changed AUG (97 %).

These results verify the assumption that the network acquires knowledge from regions that are biologically and functionally important. This approach has an advantage over the other approaches in deriving consensus. The network is not biased towards any specific region of the sequence presented to it. It learns from the weighted sums of all the features to distinguish between the more important ones and less important ones. Further, this method can be exploited in deriving consensus for other biologically important regions for which weak conservation in sequence is observed.

Acknowledgments

The author thanks Prof. Thomas Schneider for useful discussions.

References and Footnotes

1. Schneider, T.D., Stormo, G.D., Gold, L., and Ehrenfeucht, A., *J. Mol Biol. 188*, 415-431(1986).
2. Gold, L. *Ann. Rev. Biochem. 57*, 199-233 (1988).
3. De Smit, M.H. and van Duin, J., *Progress in Nucl. Acid Res and Mol Biol. 38*, 1-35 (1990).
4. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S. and Stormo, G. *Ann. Rev. Microbiol. 35*, 365-403 (1981).
5. *"The Ribosome: Structure, function and evolution"* (ed: Hill, W.E., Dahlberg, A., Garrett, R.A., Moore, P.B., Schlessinger, D., and Warner, J.R.)
6. Stormo, G.D., Schneider, T. D. and Gold, L.M., *Nucl. Acids Res. 10*, 2971-2996 (1982).
7. Schneider, T.D. and Stephens, R.M., *Nucl. Acids Res. 18*, 6097-6100 (1994).
8. Stormo, G.D., Schneider, T.D., Gold, L.M. and Ehrenfeucht, A., *Nucl. Acids Res. 10*, 2997 3011(1982).
9. Bisant, D. and Maizel, J. *Nucl. Acids Res., 23*, 1632-1639 (1995).
10. Shine, J., Dalgarno, L., *Proc. Natl. Acad. Sci USA 71*, 1344-1346 (1974).
11. Waltz, A., Pirotta, V., and Ineichen, K., *Nature 262*, 665-669(1976).
12. Van Gemen, B., Koetes, H.J., Plooy, C. A. M., Bodlaenda, J. and Van Knippenberg, P.H., *Biochimie 69*, 841-848 (1987).
13. Trifonov, E.N., *CABIOS 12*, 423-429 (1995).
14. Rumelhart, D.E., Hinton, G.E. and Williams, R.J.,*Nature 323*, 533-536 (1986)

15. Rumelhart, D.E. and McClelland, J.L. in : *Parallel and Distributed Processing :Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA (1986).
16. Zupan, J. and Gasterger, J. *An. Chem. Acta 248*, 1-30 (1991).
17. Nair, T.M., Tambe, S.S. and Kulkarni, B.D. , *FEBS. Lett 346*, 273-277 (1994).
18. Zupan, J. and Gasterger, J., *Angew. Chem. Int. Ed. Engl. 32*, 503-527(1993).
19. Nair, T.M., Tambe, S.S. and Kulkarni, B.D., *CABIOS 11*, 293-300 (1995).
20. Rudd, K.E. and Schneider, T.D., in : *A Short course in Bacterial Genetics : A Laboratory Manual and Handbook for Escherichia coli and related bacteria*. (Miller, J., Ed.) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. Pp. 17.19-17.45 (1992).
21. Demeler, B. and Zhou, G. *Nucl. Acids Res. 18*, 1593-1599 (1991).
22. Dunn, J. J., Buzash-Pollert, E., Studier, F.W. *Proc. Natl. Acad. Sci USA 75*, 2741-2745 (1978).
23. Zurawski, G., Elseviers, D., Stauffer, G.V., Yanofsky, C., *Proc. Natl. Acad. Sci. USA 75*, 5988-5992 (1978).
24. Belin, D., Epistein, R. H., *Virology 78*, 537-553 (1977).
25. Belin, D., Hedgpeth, J., Selzer, G. B., Epistein, R.H., *Proc. Natl. Acad. Sci USA 76*, 700-704 (1979).
26. Abraham, J., Mascarenhas, D., Fischer, R., Benedik, M., Campbell, A., Echols, H., *Proc. Natl. Acad. Sci. USA 77*, 2477-2481 (1980).
27. Hoess, R.H., Foeller, C., Bidwell, K., Landy, A., *Proc. Natl. Acad. Sci. USA 77*, 2124-2128 (1980).
28. Johnston, H. M. and Roth, J. R., *J. Mol. Biol. 145*, 735-756 (1981).

Date Received: August 5, 1997

Communicated by the Editor Rick Ornstein