# Ribosome Binding Site Recognition Using Neural Networks[+]

Marcio Ferreira da Silva Oliveira[(1)*]

Daniele Quintella Mendes[(2)]

Ana Tereza Ribeiro Vasconcelos[(1)]

[(1)] Laboratório de Bioinformática - Laboratório Nacional de Computação Científica – LNCC/MCT – RJ, Brasil

[(2)]Coordenação de Pós-graduação e Pesquisa de Engenharia (COPPE) – Programa de Engenharia de Sistemas e Computação (PESC) – Universidade Federal do Rio de Janeiro (UFRJ) – RJ, Brasil

## ABSTRACT

Pattern recognition is an important process to localization of genes and other functional sequences in genomes. The prokaryotic mRNA ribosome binding site (RBS) usually contains a domain UAAGGAGGU known as the Shine-Dalgarno (SD) sequence found just upstream to the translation initiation codon. The RBS is an important pattern to identification of genes in the DNA sequence, since every bacterial mRNA usually has an RBS, to every polipeptidical product to be produced.

The RBS sequence is highly degenerated, and may have a great variation in its base composition and localization in the genome [1][3][4][8]. Due to this great flexibility, the conventional methods mostly used to recognize RBS's may have a very high error rate in its predictions.

In the present work, we will deal with the problem using a Machine Learning method [2][6][9] known as Multi Layer Neural Netwoks (MLNN), with the learning algorithm Backpropagation [5][10]. These MLNN are able to deal with problems where few information is available or where the data to be analyzed is very inaccurate. The organism we chose to apply the models of Neural Networks is the bacteria *E. coli* K-12 [7].

The MLNN models used in this work allow an economic representation of this biological concept highly degenerated, taking advantage of the learning process properties to drastically reduce the size of the learning set. Instead of using a training set that would include explicitly all the possibilities, we used the intrinsic power of generalization of the Neural Network, combined with (extreme) sequences, to allow a simple and biologically appropriate representation of the problem.

## REFERENCES

[1]Baris E. Suzek; Maria D. Ermolaeva, Mark Schreiber & Steven L. Salzberg; "A Probabilistic Method for identifying Start Codons in Bacterial Genomes"; Bioinformatics, vol. 17 n° 12, p. 1123-1130, 2001.

[2] Carvalho, L. A. V.; "Data Mining", São Paulo, Érica Press, 2001.

[3] Lodish, Harvey F. & Freeman, W. H. "Molecular Cell Biology". 4th edition, New York, W.H. Freeman, 2000.

[4] Lewin, Benjamin. "GENES VII". Oxford University Press, 1997.

[5] Haykin, Simon., "Redes Neurais - Princípios e prática". 2nd edition; Bookman Press, 2001.

[6] Mitchell, Tom M.; "Machine Learning"; McGraw-Hill Press, 1997.

[7] NCBI - National Center for Biotechnology Information. Available in http://www.ncbi.nlm.nih.gov. Access in 2002.

[8] Shultzberger, Rayn K., Bucheimer R. Elaine, Rudd, Kennth E. & Schneider, Thomas D. "Anatomy of Escherichia coli Ribossome Binding Sites". J. Molecular Biology (2001) 313, pp. 215-228, 2001.

[9] Salvini, Rogério L. "Uma Nova Abordagem para Análise de Agrupamentos Baseada no Algoritmo Elástico ". Master thesis – COPPE / UFRJ, Programa de Engenharia de Sistemas e Computação, 2000.

[10] Wu, Cathy H. & McLarty, Jerry W. "Neural Networks and Genome Informatics". 1st edition. Methods in Computacional Biology and Biochemistry vol. 1, Elsevier Press, 2000

[+] We are submitting this paper to be presented as a poster

[*] Corresponding author e-mail address: marciofs@lncc.br.