Curvature, bendability, stability and nucleotide composition applied to *E. coli* σ^{28} promoter prediction and recognition

Tahila Andrighetti[1], Priscila Portela[1], Ivaine Taís Sauthier Sartor[3], Günther J. L. Gerhardt[2], Sergio Echeverrigaray[1], Scheila de Avila e Silva[1]

[1]Universidade de Caxias do Sul, Instituto de Biotecnologia
[2]Universidade de Caxias do Sul, Departamento de Física e Química
Rua Francisco Getúlio Vargas, 1130 - CEP 95070-560, Caxias do Sul - RS – Brasil
Phone: 55 54 3218 21 00 extension 2075 - Fax number 55 54 3218 21 49

Universidade Federal do Rio Grande do Sul – UFRGS, Instituto de Ciências Básicas da Saúde
Centro de Estudos em Estresse Oxidativo - Laboratório 32, Departamento de Bioquímica
Rua Ramiro Barcelos, 2600 – anexo – CEP 90035 003, Porto Alegre/RS Brasil
Phone: + 55 51 3308 5577

{Tahila Andrighetti tahila.a@hotmail.com, Priscila Portela pri-portela@hotmail.com; Ivaine T. S. Sartor ivaine.sauthier@yahoo.com.br, Günther Gerhardt gunther_lew@yahoo.com.br, Sergio Echeverrigaray selaguna@yahoo.com, Scheila de Avila e Silva sasilva6@ucs.br}

The first step in gene expression process is the promoter recognition by RNA polymerase enzyme (RNAP). The RNAP is formed by five subunits and an additional sigma (σ) subunit factor. The σ factor is essential to lead RNAP binding on specific promoter regions and activate given genes for the best environmental changes response. Several studies have reported that prokaryotic σ^{70}-dependent promoter sequences have lower stability, higher curvature and lesser bendability than coding sequences. The melting behavior of DNA double strand is related with the DNA stability. Bendability is twists and short bends of approximately 3 base-pairs, while curvature are loops and arcs involving around 9 base-pairs. In this paper, σ^{28}-dependent promoter sequences (which control the expression of flagellar genes during the normal growth of bacteria) were predicted, recognized and characterized based on curvature, bendability, stability and nucleotide composition by using Multilayer Perceptron Neural Networks (NN). The simulations were carried out in the R environment. The back-propagation algorithm with a 2-fold-cross validation was chosen, in order to obtain statistically valid results. The 20 available positive examples were obtained from RegulonDB database. The negative examples were randomly chosen from *E.coli* non-promoter intergenic regions in the same number and length of positive examples. Four simulations were carried out, each one with different input data: orthogonal codification, stability, curvature and bendability. The algorithms for data generating were based on literature. The data sets (excluding orthogonal codification) were smoothed with a moving average using the software Low121. Among all architectures tested, was possible to select the simplest architecture which can better classify the sequences. For each simulation, the results achieved for accuracy, sensitivity and specificity were, respectively: *(i)* curvature: 55%, 56%, 54%; *(ii)* bendability: 79%, 78% 79%; *(iii)* stability: 80%, 77%, 83%; *(iv)* orthogonal codification 70%, 73%, 66%. The best results were showed by the DNA properties which do not present wide correlations between bases (stability, bendability). The orthogonal codification is directly related with the nucleotide position. As the motifs do not present high sequence conservation, this reduced the NN accuracy. The lowest accuracy was presented by curvature. This can be explained by the fact that this measure is obtained from wide correlation between nucleotides, and the promoter is a short sequence. In spite of the performance values, these results are an indicative that the combination of different promoter features can provide relevant biological information and improve the promoter prediction and recognition.