Curvature and Flexibility as Promoter Regions Classifiers in Gram-Negative Bacteria

I. T. S. Sartor¹, T. Andrighetti¹, G. J. L. Gerhardt¹, S. Echeverrigaray¹, S. de Avila e Silva¹.

¹ Universidade de Caxias do Sul, Instituto de Biotecnologia Rua Francisco Getúlio Vargas, 1130 - CEP 95070-560 Caxias do Sul - RS - Brasil Phone: 55 54 3218 21 00

1. Introduction

The gene expression control is a fundamental process in cellular activities, performed through the interaction of multiple regulatory mechanisms. The proper regulation of transcription is crucial for a single-cell prokaryote since its environment can change dramatically and instantly. The promoters are recognized as one of the transcription regulatory regions, as recruit the transcriptional machinery through the binding of regulatory proteins in their DNA sequences. The characterizing promoter regions in silico has difficulties, since these elements are short and degenerated, providing a high probability of finding similar sequences in other parts of the genome. Therefore, the embedding of structural characteristics can increase the accuracy of prediction methods [1-2].

In bacteria, RNAP holoenzyme is responsible for promoter recognition and the gene expression starts. This enzyme consists of five subunits (2α,β,β',ω) and an additional sigma (σ) subunit factor. A collection of different σ subunits act as key regulators of bacterial gene expression. The substitution of one σ factor by another can initiate the transcription of different groups of genes.

A promoter sequence is characterized by the presence of two conserved DNA elements called -10 and -35 (upstream). These elements are defined according to the distance which have in relation to the transcriptional start site (position 1) and are represented by TATAAT-TTGACA nucleotides [3]. The upstream region (promoter) has distinct sequence properties compared to the downstream region (non-promoter), such as differences in the structural characteristics of flexibility, stability and curvature [4].

Artificial neural networks (ANNs) have been widely used in nucleic acid sequences analysis, since they present ability to recognize and classify quantitative and qualitative patterns in data analysis [5]. This work aims to predict, recognize and characterize promoter regions recognized by sigma factor 28 (σ28) employing an approach of artificial neural networks using as input parameter curvature and flexibility data of the sequence.

2. Methods

The 20 sequences of σ28 promoter regions of E. coli were obtained from biological database RegulonDB [6]. The random sequences were generated with 0.28 frequency for A and T nucleotides and 0.22 for C and G nucleotides - the same probability found in E. coli promoter sequences. The promoters were defined as positive examples of training, whereas the random sequences were the negative examples.

2.1. Data preparation: The flexibility and curvature values of promoter and non-promoter sequences were generated using Bend algorithm, applied with different sizes of sliding window.

2.2. Coding for training: After generation of curvature and flexibility values, these were smoothed by the software Tisean, at different degrees, using a low-pass filter.

2.3. Neural Networks: The model of network architecture employed in these simulations was a feedforward multilayer perceptron, which is composed of interconnected artificial neurons, grouped in non-recursive layers, that is, the outputs of a given layer are connected to the input of the next layer.

2.4. Network training: Simulations were performed at R ambient using the back-propagation algorithm and the methodology used was 2-fold-cross-validation.

2.5. Classification Analysis: The average accuracy of classification obtained by the NN was calculated from the number of true positives (TP), true negative (TN), false positives (FP) and false negative (FN). The sensitivity (SN) and specificity (S) also calculated, representing the TP and TN correctly classified, respectively.
3. Results

In total, this approach performed 6400 simulations using curvature and flexibility values for promoter recognition and prediction. Among all architectures tested, it was possible to select the simplest architectures which can classify the sequences.

3.1. Curvature results: The best architecture was composed by an input layer with 76 neurons, 3 hidden neurons and a output layer with 1 neuron. These architectures achieved 55% of accuracy value. The sensitivity and specificity values presented were 56% and 54%, respectively.

3.2. Flexibility results: The best architecture was composed by 74 neurons in the input layer, 6 hidden neurons, 1 neuron on output layer. The results presented for accuracy, sensitivity and specificity were 79%, 78.66% and 79.33%, respectively. In the simulations with the both input data, a greater number of neurons in the hidden layer did not increase accuracy, specificity or sensitivity values

4. Discussion and Conclusions

In this paper, we have presented an approach for prediction and classification of E. coli promoters based on NN trained with the structural features of the promoter sequences. As an efficient classification tool, it is expected to recognize promoter as well non-promoters. The similar specificity and sensitivity values obtained for both input data values are indicative of the consistence of the NN learning process.

The results of these analyses showed that the curvature of these sequences has not been consolidated as a great feature for classification, since the best networks exhibited low rates of accuracy. However, for flexibility, the best network showed, in addition to the high rates, approximate values of accuracy, sensitivity and specificity, which means that the NN equivalently recognized promoter and non-promoter sequences. These results indicate that the flexibility is an important feature using as a parameter to classifying promoter regions for σ28 factor. In contrast to the most of tools previously reported in the literature, this paper is devoted in the analysis of other kind of promoter sequence, not only the σ70-dependent promoters.

Acknowledgements

This work is supported by UCS.

5. Bibliography