Promoter Recognition by Non-Homogeneous VOMT Models

Ben-Gal I.[1], Arviv S.[1], Shmilovici A.[2] and Grosse I.[3]
[1] Department of Industrial Engineering, Tel-Aviv University
[2] Department of Industrial Engineering, Ben-Gurion University
[3] Cold Spring Harbor Laboratory, NY, USA

We suggest a new class of learning models for patterns classification of DNA sequences. The models are based on the context tree that was originally proposed in [7] for data-compression purpose and later modified in [1][2][3].

The suggested models can be described as varying-order Markov Trees (VOMT). Unlike the fixed-order Markov models, the order of various contexts in the VOMT do not have to be equal and, therefore, is not necessarily fixed to account for the maximum dependence found in the data. As a result, the VOMT obtains a great reduction in the number of parameters that need to be estimated, and a smaller probability for over-fitting. Such reduction in the estimation effort is especially important for database of a limited size.

In the present case study we consider the E. coli supervised promoter-recognition (see, [4], [5], [6]). Based on a given dataset that contains 238 E. coli promoters, the VOMT are trained to distinguish between promoters and non-promoters. In particular a non-homogenous VOMT is constructed for the promoters data, such that a context-tree is built for each position in the promoter sequence. **It is shown that the proposed VOMT achieve superior results in comparison to all previously published results.** In particularly, for a cross-validation experiment with a 99.9% true-negatives (TN) value, one obtains a 48.76% level of true-positive (TP) – an improvement of almost 10% to that of the PWM model. Alternatively, for a threshold which is set to zero, one obtains a 92.32% accuracy level, where TP=89.92% and TN=94.73%. Further examples will be given in the talk.

[1] Ben-Gal I., Shmilovici A., Morag G., (2000) Design of Control and Monitoring Rules for State Dependent Processes, *The International Journal for Manufacturing Science and Production*, 3, NOS. 2-4, pp. 85-93.

[2] Ben-Gal I., Shmilovici A., (2001) "Promoters Recognition by Varying-Length Markov Models", Artificial Intelligence and Heuristic Methods for Bioinformatics, 30 Sept. – 12 Oct., San-Miniato, Italy.

[3] Ben-Gal I., Shmilovici A., Morag G., Singer G (2001) US Provisional Patent Application No. 60/269,344 filed February 20th 2001

[4] Fickett J. W., Hatzigeorgiou A.G., Eukaryotic Promoter Recognition, *Genome Research* 7:861-878,1997.

[5] Holste D., Grosse I., Buldyrev, S. V., Stanley H. E., and Herzel H. Optimization of Protein Coding Measures Using Positional Dependence of Nucleotide Frequencies. J. *of Theoretical Biology*, 206, 525--537 (2000)

[6] Ohler U., Harbeck S., Neimann H., Noth E., Reese M. G., Interpolated Markov Chains for Eukaryotic Promoter Recognition, *Bioinformatics*, 15,5, 362-369.

[7] Rissanen, J., 1983, A Universal Data Compression System, *IEEE Transactions on Information Theory,* 29(5), 656-664.