BMC Systems Biology

Poster presentation

A probabilistic context-free grammar for the detection of binding sites from a protein sequence

Witold Dyrka*[1,2] and Jean-Christophe Nebel[1]

Address: [1]Faculty of Computing, Information Systems and Mathematics, Kingston University, Kingston upon Thames, KT1 2EE, UK and [2]Faculty of Fundamental Problems of Technology, Wroclaw University of Technology, 50-370 Wroclaw, Poland

Email: Witold Dyrka* - k0543192@kingston.ac.uk

* Corresponding author

from BioSysBio 2007: Systems Biology, Bioinformatics and Synthetic Biology
Manchester, UK. 11–13 January 2007

Published: 8 May 2007

BMC Systems Biology 2007, **1**(Suppl 1):P78 doi:10.1186/1752-0509-1-S1-P78

This abstract is available from: http://www.biomedcentral.com/1752-0509/1?issue=S1

Introduction

The analysis of a protein, through the evaluation of interactions between the amino acid composing its sequence, is a very challenging problem where pattern recognition techniques based on Hidden Markov Model (HMM) have proved to be the most efficient [1]. Although HMM is a powerful technique, it has limitations. According to formal language theory, its expressive power is similar to probabilistic regular grammars. A more powerful grammar, Context-Free Grammar (CFG), has been applied successfully for the recognition and prediction of RNA structure [1,2]. However, its utilisation in the field of protein pattern recognition is a more challenging task due to the larger set of terminals and less straightforward relations between residues. In this piece of work, we propose a Probabilistic Context-Free Grammar (PCFG) to represent features of protein structures. In order to deal with the size of the protein alphabet, we use quantitative properties of amino acids to reduce the number of symbols. Based on that grammar we designed a tool allowing the detection of protein regions which are involved in binding sites. The PCFG is evolved using a genetic algorithm (GA) to describe a pattern shared by a set of proteins.

Methodology

The method is described schematically in Figure 1a. The general idea is to use quantitative properties of amino acids to limit the number of symbols present in the PCFGs describing the binding sites of interest. First, we select the amino acid properties which are relevant to the binding site of interest and we create terminal rules which express those properties in a probabilistic manner. This process is detailed in the next section. Then non terminal rules are generated and their associated probabilities are induced using a genetic algorithm (GA) from a positive training set. Obtained grammar could then be pruned from rules of low probability. Finally, protein sequences of unknown function are scanned using a Cocke-Kasami-Younger style parser. Binding sites are detected at a given position if probability at that position is above a threshold set automatically during the learning stage. An example is given in Figure 1b. In order to achieve more robust results, grammars based upon different properties can be combined.

Terminal rules based on amino acid properties

Our method relies on the selection of amino acid properties in order to deal with the size of the protein alphabet. We use quantitative properties taken from AAindex database to reduce the number of symbols [3]. This database consists of values of the 20 amino acids for over 500 properties which can be clustered into 6 categories: beta propensity, alpha and turn propensities, composition, physiochemical properties, hydrophobicity and other properties. An appropriate small set of properties can be chosen by either expert knowledge or PCA analysis of preselected properties reflecting the learning set composition (Weighted PCA). For each property, 3 non-terminals – low, medium and high level – are created. Then terminal

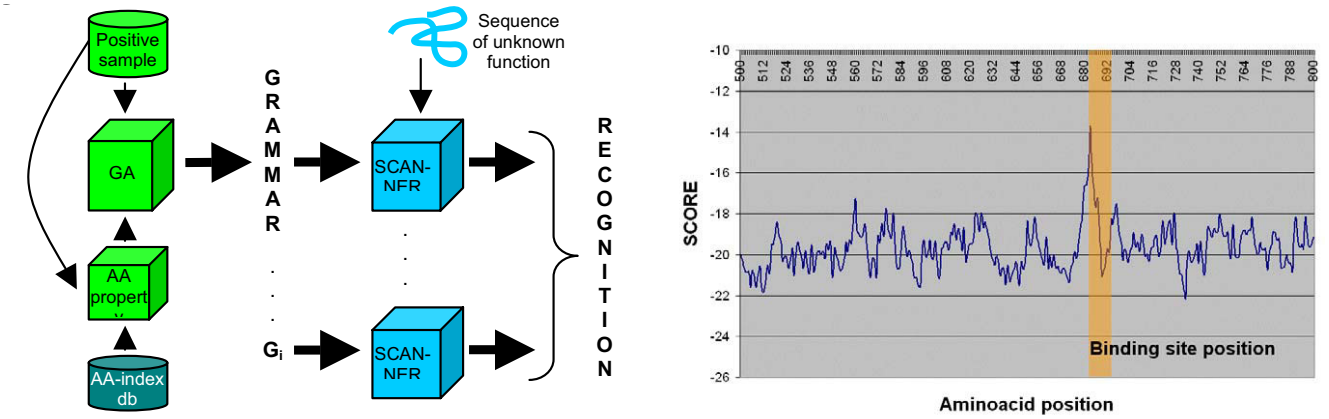

Figure 1
Left: General scheme of the method; Right: Typical output showing the peak corresponding to the pattern.

rules are produced to associate a set of 3 probabilities to each amino acid.

Results

Our technique was successfully tested on a PROSITE pattern (PS00219) which has a high false negative rate. As expected, results show the choice of the amino acid property is key to prediction accuracy. In this case, good prediction rate was achieved using grammars based on either charge or van der Waals volume of amino acids. Results for other properties like beta sheet, alpha helix and relative frequency are poor as they seem to be weakly related to the binding site function. Grammar based on accessibility performs slightly better. In order to automate the property selection, we processed a set of 5 amino acid properties representing the following 5 clusters: beta propensity, alpha and turn propensities, composition, physiochemical properties and hydrophobicity. Very good results were achieved for the first wPCA vectors (standard PCA did not perform as well). However, due to the poor results obtained with the second wPCA vector further investigations are necessary to conclude regarding the validity of this approach. We also observed that window size did not have a major impact on detection accuracy. Finally, our experiments showed that the best performances were achieved by combining grammars. These grammars proved to be more accurate than the PROSITE pattern.

Conclusion

PCFG based on quantitative representation of amino acid properties proved to be successful for PS00219. Also our process of automated property selection based on weighted PCA is encouraging as it contributed to results of the best accuracy. For the future, we plan to improve the automated property selection process and refine our procedure of grammar combination. We will also introduce a scoring scheme independent from window size and speed up convergence of evolution process. Finally, tests will be performed on large variety of binding sites.

References

1. Sakakibara Y: **Grammatical Inference in Bioinformatics.** *IEEE Trans on PAMI* 2005, **27:**1051-1062.
2. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjolander K, Underwood RC, Haussler D: **Stochastic context-free grammars for tRNA modelling.** *Nucleic Acids Res* 1994, **22:**5112-5120.
3. Kawashima S, Kanehisa M: **AAindex: amino acid index database.** *Nucleic Acids Res* 2000, **28:**374.