

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

Witold Dyrka and Jean-Christophe Nebel

Faculty of Computing, Information Systems & Mathematics, Kingston University,
Kingston-Upon-Thames, KT1 2EE, UK

k0543192@kingston.ac.uk

Abstract

Analysis of protein sequences to predict their functions is a very challenging problem where pattern recognition techniques based on Hidden Markov models (HMMs) have proved to be the most efficient. However HMMs have limitations. According to formal language theory, their expressive power is similar to probabilistic regular grammars. Here, we propose a pattern recognition method based on a more powerful grammar. We developed a Probabilistic Context-Free Grammar to detect protein regions that are involved in binding sites. In order to deal with the size of the protein alphabet, we use quantitative properties of amino acids to reduce the number of symbols. Our technique was successfully tested on a PROSITE pattern which has a high false negative rate. Results show the potential of our method for detecting patterns in proteins.

Rational

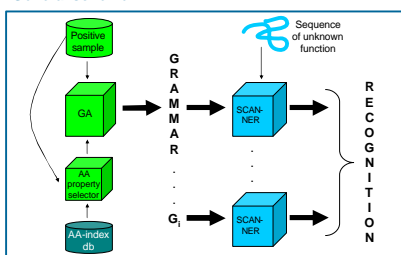
- Detection of binding sites in a protein sequence is key to the problem of protein annotation
- State of the art Hidden Markov Models (HMMs) based methods [1] cannot take into account long distance relationships between amino acids (AA)
- Probabilistic Context-Free Grammar (PCFG) is a more powerful language which has been successfully applied to the prediction of RNA secondary structures [1,2]

Method

Principle

Use quantitative properties of AAs to limit the number of symbols present in the PCFGs describing the binding sites of interest. Grammars are generated using a Genetic Algorithm (GA).

General scheme



- Selection of AA property relevant to binding site
- Extraction of grammar rules using GA from a positive training set
- Scanning of sequence of interest using grammar
- Detection of binding site if probability of a position is above an automatically generated threshold
- Grammars based upon different properties can be combined to achieve more robust results

Amino acid properties

- Definition: AAindex database [3] of quantitative values of the 20 AAs for over 500 properties.

They cluster in 6 categories:

- beta propensity
- alpha & turn propensities
- composition
- physicochemical properties
- hydrophobicity
- other properties

- Selection by either expert knowledge or PCA analysis of preselected properties reflecting learning set composition (Weighted PCA)

- Usage: 3 non-terminals are created for low, medium and high level of property

Implementation

PCFG parser

- Based on Cocke-Kasami-Younger algorithm

Genetic algorithm

- M. Wall's GAlib 2.4.6 customised for evolution of grammar rule probabilities

Parameters

- Number of non-terminals, including property non-terminals
- Diversity pressure during evolution
- Cut-off level of rule parsing for higher robustness

Automated property selection

- Weighted PCA using several property categories

Experiment

Binding site of interest:

ANION_EXCHANGER_1 (PS00219)

- Training set (random 80% used for training)

8 different instances of PROSITE pattern:

F-G-G-[LIVM](2)-[KR]-D-[LIVM]-[RK]-R-R-Y length: 12

1 pattern missed by PROSITE pattern:

F-G-G-L-L-L-D-I-K-R-K length: 11

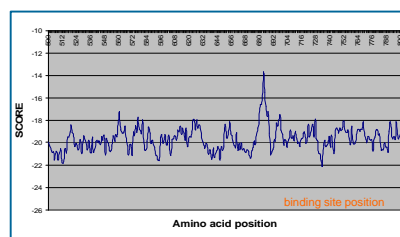
Positive set*

76 Swiss-Prot/TrEMBL entries matching PS00219

Negative set**

829 single chain sequences of 300-500 residues returned by PDB query for similarity below 30%

Typical output



- SCORE indicates log of probability that pattern at a certain position is generated by the grammar

- For best grammars, the binding site is always at the position with the highest value

Results

TP & TN rates for single and combined grammars

- For positive set, result is counted as TP when the highest peak is at the position of PS00219 pattern. Their lowest value is set as the detection threshold.
- For negative set, result is counted as TN when the highest peak is below the detection threshold
- Score of combined grammars is the average score of grammars involved

Property	Window size	TP rate	TN rate
Change	11	79%	96%
+wPCA+vol	11	100%	100%
+wPCA	11	100%	100%
Volume	12	100%	99%
+wPCA	11	100%	100%
Change	12	100%	100%
+wPCA	11	100%	99%
+volume	12	96%	99%
Change	11	100%	99%
wPCA	12	100%	95%
PCA	11	83%	83%
Beta sheet propensity	12	75%	88%
Var det	11	62%	42%
Markov volume	12	100%	97%

Analysis

- Good prediction using charge or volume
- Very good results for 1st wPCA vector, but poor for 2nd wPCA vector (not shown)
- Poor prediction for beta sheet, alpha helix and relative frequency properties (not shown). Accessibility (not shown) performs slightly better.
- Window size does not have a major impact
- Best performance achieved by combined grammars (better accuracy than PROSITE pattern)

Conclusion

- PCFG based on quantitative representation of AA properties proved to be successful for PS00219
- Automated property selection is encouraging

Future work

- Investigate further automated property selection and combination of grammars
- Introduce a scoring scheme independent from window size
- Speed up convergence of evolution process
- Tests on variety of binding sites

References

- [1] Y. Sakakibara, IEEE TPAMI, 7:1051-1062, 2005
- [2] Y. Sakakibara *et al.*, Nucleic Acids Res., 22:5112-20, 1994
- [3] S. Kawashima *et al.*, Nucleic Acids Res., 28:374, 2000

* Positive sample based on the UniProt (9th December 2006)

** Negative sample based on the PDB (12th December 2006)