Frequent Subgraph Summarization by Substitution Matrices

Wajdi DHIFLI[1,2], Rabie SAIDI[1,2], Engelbert MEPHU NGUIFO[1,2]

[1] Clermont University, Blaise Pascal University , LIMOS, BP 10448, 63000 Clermont-Ferrand, France
[2] CNRS, UMR 6158, LIMOS, 63173 Aubière, France

Background and aims

Motivation And Background

Proteins have been recently seen as graphs of amino acids and studied based on graph theory concepts. Indeed, algorithms of frequent subgraph discovery [1] have been applied on protein structures to find motifs that could be interesting in any further analysis. However, when the support threshold is low, the number of frequent subgraphs is expected to be very large which may hinder rather than help.

Contribution

We claim that in the set discovered subgraph-motifs, there exist a subset of representative subgraph-motifs that can substitute several others and hence can summarize the whole set. We propose a novel approach that selects these motifs based on the amino acids mutation quantified in the substitution matrices. We term them the **unsubstituted patterns**. These selected motifs can be used instead of the whole set, in order to enhance and facilitate any motif-based-analysis such as classification, clustering, visual inspection, drug molecule prediction, etc.

Methods

Unsubstituted pattern selection

During the evolution, amino acids that compose the protein mutate. A mutation is a substitution that exchanges one amino acid to another. This phenomenon was quantified in literature in the form of substitution matrices [2].

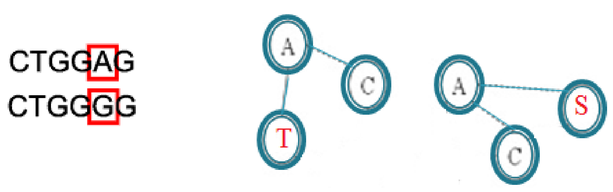

Fig.1 An example of amino acids mutation during the evolution of proteins (left: protein sequence, right: protein tertiary structure)

We explore this information to select representative motifs from the set of discovered frequent spatial motifs. Each one of the selected motifs is a representative of all the spatial motifs it substitutes.

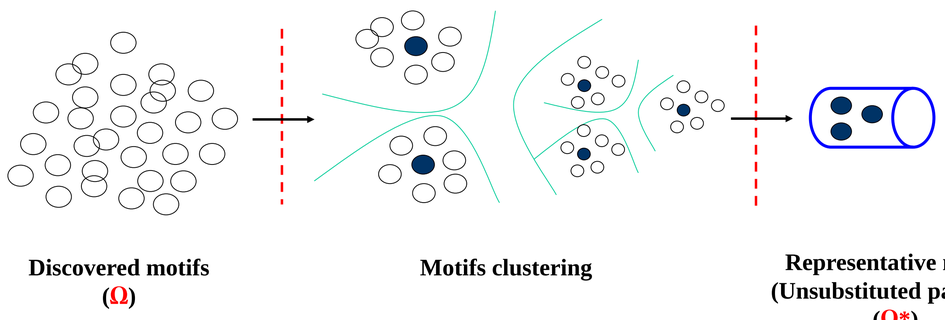

Fig.2 The general process of unsubstituted patterns selection.

As shown in the figure above, the general process of the selection is as follow:

1. We divide Ω into subsets of patterns having the same size.
2. Each subset is sorted in a descending order by the mutation ability of the patterns (computed based on the used substitution matrix).
3. Each subset is browsed starting from the pattern having the highest mutation ability.
4. For each pattern in the subset, we remove all the patterns it substitutes.
5. The remaining patterns represent the unsubstituted patterns set Ω*.

The remaining set Ω* can not be summarized by a subset of it but itself.

Experimental settings

Dataset	SCOP ID	Family name	Pos	Neg	# motifs
DS1	52592	G proteins	33	33	799 094
DS2	48942	C1 set domains	38	38	258 371
DS3	56437	C-type lectin domains	38	38	114 792
DS4	88854	Kinases, catalytic subunit	41	41	1 073 393

Tab.1 Experimental data from [3]. SCOP ID: identifier of protein family in SCOP [4], Pos: positive proteins sampled from a selected protein family, Neg: negative proteins randomly sampled from the Protein Data Bank [5].

- Proteins are parsed into graphs of amino acids using Cα method as in [3].
- We use gSpan [1] to extract frequent-subgraphs (spatial motifs) (freq >= 30%).
- We use UnSubPatt to select unubstituted patterns.
- We compare the number and the interestingness of the selected patterns with the original set.
- We perform a 10-CV classification on the datasets then we compare the performances using frequent-subgraph motifs then unsubstituted patterns.

Results

Experimental results

Dataset	|Ω|	|Ω*|	Selection rate (%)
DS1	799094	7291	0.91
DS2	258371	15898	6.15
DS3	114792	14713	12.82
DS4	1073393	9958	0.93

Tab.2 Number of frequent spatial motifs (30%), unsubstituted spatial motifs (30%) and the selection rate.

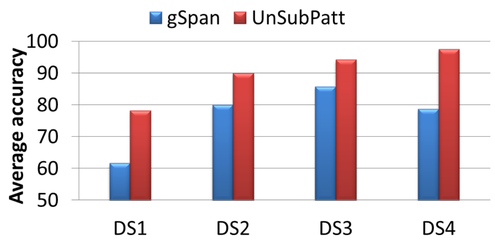

Fig.3 Classifcation accuracy by NB using frequent spatial motifs (gSpan)(30%) and unsubstituted spatial motifs (UnSubPatt)(30%).

The selection rate shows that our approach descreases dramatically the number of spatial motifs. This reduction comes with a significant enhancement in the classification accuracy with the four datasets.

Impact of Substitution Threshold

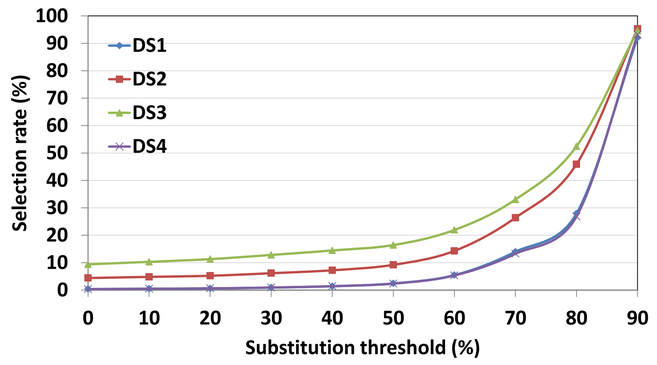

Fig.4 Rate of unsubstituted patterns from the initial set of spatial motifs (Ω) depending on the substitution threshold.

We notice that UnSubPatt reduces considerably the number of frequent spatial motifs especially with lower substitution thresholds.

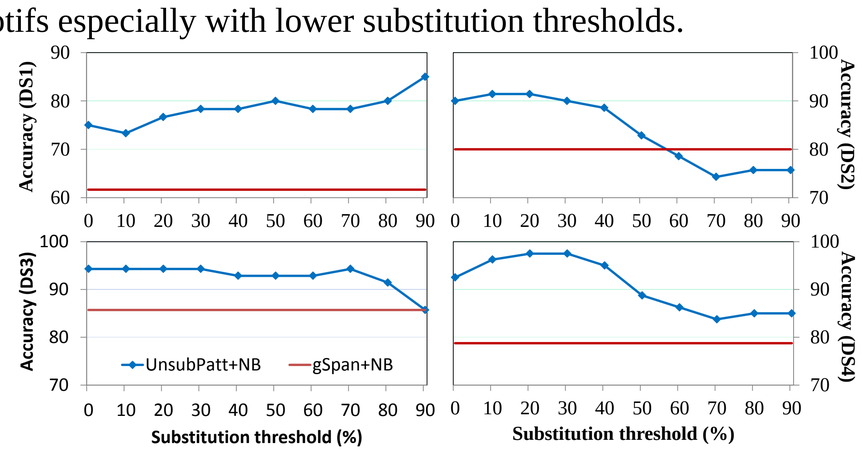

Fig.5 Classification accuracy by NB.

With all the datasets, unsubstituted patterns allow a significant enhancement of the classification accuracy compared to the original set of spatial motifs.

As future goals, we plan to test our approach using other substitution matrices (BLOSUM80, PAM250, …). Moreover, we intend to test our approach in other classification contexts as well as in other different applications.

References

[1]: V. Krishna and al. A comparative survey of algorithms for frequent subgraph discovery. Current Science, 100(2):190, 2011.
[2]: S. R. Eddy. Where did the blosum62 alignment score matrix come from? Nature Biotechnology, pages 1035–1036.
[3]: H. Fei and J. Huan. Boosting with structure information in the functional space: an application to graph classification. In ACM KDD conference, pages 643–652, 2010.
[4]: A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the scop database: new developments. Nucleic Acids Research, 36(1):D419–D425, 2008.
[5]: H. M. Berman, J. D. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. Nucleic Acids Research, 28(1):235–242, 2000.

Acknowledgements

This work is supported by a PhD grant from the French Ministry of Higher Education to the first author.