Proteins

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](Image)

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](Image)

As shown in the figure above, the general process of the selection is as follows:
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* cannot be summarized by a subset of it but itself.

Experimental settings

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SCOP ID</th>
<th>Family name</th>
<th>Pos</th>
<th>Neg</th>
<th># motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>52902</td>
<td>G proteins</td>
<td>33</td>
<td>33</td>
<td>709 004</td>
</tr>
<tr>
<td>DS2</td>
<td>48942</td>
<td>C1 r t domains</td>
<td>38</td>
<td>38</td>
<td>253 171</td>
</tr>
<tr>
<td>DS3</td>
<td>56437</td>
<td>Cmp. lactn domains</td>
<td>38</td>
<td>38</td>
<td>181 942</td>
</tr>
<tr>
<td>DS4</td>
<td>88554</td>
<td>Kinases, catalytic subunit</td>
<td>41</td>
<td>41</td>
<td>1 073 939</td>
</tr>
</tbody>
</table>

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 Cu method as in [3].
- We use gSpan [1] to extract frequent subgraphs (spatial motifs) (freq > 30%).
- We use UnSubPatt to select unsubstituted 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

The selection rate shows that our approach decreases 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 motif (Ω) depending on the substitution threshold.](Image)

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

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