A web interface to extract protein motifs by constrained co-clustering

Francesca Cordero[1,2], Alessia Visconti[2], Marco Botta[2].

[1]*Molecular Biotechnology Center, Department of Clinical and Biological Science, University of Turin, Via Nizza 52, 10126 Torino, Italy and* [2]*Department Computer Science, University of Turin, Corso Svizzera 185, 10149 Torino, Italy*

Pattern discovery in biological sequences is a fundamental problem in both computer science and molecular biology.

We propose a framework based on a constrained co-clustering technique that simultaneously groups protein sequences and their associated patterns. The main goal of our approach is to split a set of (possibly unknown function) protein sequences in groups characterized by some common motifs.

The novelty of our approach is the protein motif discovery methodology, which relies on the following two main ideas: exhaustive search, based on prefix tree, and automatic association between motifs and protein sub-families using a constrained co-clustering algorithm.

Motif discovery is performed on a complete *ab-initio* technique where any biological knowledge is considered *a-priori*.
We use a prefix tree as data structure, since it allows to store all possible motifs of a given length extracted from the protein sequences. In this way, we maintain *all* those patterns present there along with both the protein name and their frequencies.
A constrained co-clustering technique finds both protein motif classes and protein groups: in such a way, it is possible to correlate every protein group with one or more motif classes. This technique is applied to the frequency matrix built on those values extracted from prefix tree. Therefore, the co-clustering association is given by a statistical measure based on cluster cardinality.

In test on experimentally determined protein datasets, the presented framework is able to identify the correct pairs of protein family and motifs that are very similar to PROSITE's patterns.

We have built a web interface in which the best identified motifs are displayed in association with a list of related protein names. In particular, the patterns in each motif cluster have been multiple aligned with ClustaW. In such way the information content is graphically pictured with sequence logos.

Our approach is able to group together protein sequences belonging to the same families and, at the same time, provides a set of characterizing motifs.