A Metaheuristics Approach to Protein Active Site Detection

Elisa Cilia and Mauro Brunato
Department of Computer Science, University of Trento
Via Sommarive 14, 38050, Povo (Trento), ITALY
{cilia,brunato}@dit.unitn.it

One of the aims of modern Bioinformatics is to discover the molecular mechanisms that rule the protein operation. This would allow us to understand the complex processes involved in living systems and possibly correct dysfunctions. Given the evident interest of the pharmaceutical industry, an active research has been conducted in the classification of proteins and their interactions. The first step in this direction is the identification of the functional sites of proteins. As there is a huge amount of variables and data involved in such task, intelligent approaches for the automatic detection of active sites are needed.

There may be many ways to deal with the problem of automatic protein active site identification. We have defined it as a binary classification task and we have applied efficient linear maximal margin classifiers as SVMs extended with the use of kernel methods.

We modeled the objects we want to predict to be active sites or not, which are basically topological regions of a protein, as spheres in the three-dimensional space, centered on an amino acid (or residue) of the protein. This representation brings us to have for each protein a number of spheres which is equal to the number of amino acids of the protein.

Each of the previously mentioned objects (spheres in the three-dimensional space) can be described by both linear and structural features. We have devised innovative attribute-value and structured computational representations derived from adequately preprocessed biological and spatial information of protein crystals retrieved from the Protein Data Bank.

The portion of the protein contained in a spherical three-dimensional region can be represented with a completely connected graph $G = (R, D)$. R is the set of vertices of the graph which represent residues (side-chain centroids), while D is the set of edges which represent the distances in the three-dimensional space between the pairs of vertices.

In order to design the computational model, we provided two possibilities: (1) feature vectors able to capture the most important properties of the graph and (2) graph based kernels in kernel-based machines such as SVMs. For example a graph-based *ad hoc* solution can evaluate similarities between two graphs G_1 and G_2 using a naive kernel function or a more complex one in a more relevant way from a biological point of view, that is, basing the similarity measure computation on the scores of the similarity matrices PAM or BLOSUM.

Point (2) often leads to high computational complexity. We approached such problem by

extracting a tree forest from the target graph and applying efficient tree kernels [1]. Each of the feature vectors and of the trees of a forest has an assigned and proper semantics. For example we can include in each training example a feature vector associated to each of the residues contained in the topological region of the protein. Different kernels can be used with different features subsets, i.e. it can be evaluated the similarity among different tree representations of an example preserving their semantics during the evaluation. Formally, given two objects we want to classify $o_1 = \{\vec{t_1}, ..., \vec{t_m}, \vec{u_1}, ..., \vec{u_n}\}$ and $o_2 = \{\vec{t'_1}, ..., \vec{t'_{m'}}, \vec{v_1}, ..., \vec{v_{n'}}\}$, their similarity is evaluated with the following kernel function:

$$K(o_1, o_2) = \tau \sum_{i=1}^{min(m,m')} K_t(\vec{t_i}, \vec{t'_i}) + \sum_{i=1}^{min(n,n')} K_p(\vec{u_i}, \vec{v_i}),$$

where τ is a parameter which rules the contributions of tree kernels $K_t()$ with respect to $K_p()$ which is a polynomial kernel.

In addition, labeling the catalytic residues in each individuated functional region allows us not only to state if a region of a protein is an active site or not but also to indicate with high probability which are the catalytic residues. The model to pass from the pointwise local information to global one may produce an exponential computational complexity, e.g. if it is considered the labeling of a region as the labeling task of all possible binary configurations of residues (each taken as a different instance). Moreover, the number of negative examples in opposite to the number of positives would highly increase. So, the already existent imbalance in the dataset might become more evident making the separation of the examples in the hyperplane still more difficult. To alleviate such complexity, ranking algorithms already used in other classification tasks can be applied to reduce the proliferation of negative examples. These ranking algorithms allow us to consider in the dataset only the most probable labeling configurations.

In [2], the authors have shown that the best performing algorithm for the prediction of active site amino acids (or residues), among the set of WEKA software package classifiers, is a SVM.

Moreover our previous work has shown that structural kernels used in combination with polynomial kernels can be effectively applied to discriminate an active site from other regions of the protein for a specific class of enzymes, the hydrolases [3].

Thanks to the proposed improvements our new approach is at the same time more general and accurate, possibly leading to higher performance.

References

[1] Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees. In: Proceedings of The 17th European Conference on Machine Learning, Berlin, Germany, 2006, Berlin, Germany (2006)

[2] Petrova, N.V., Wu, C.H.: Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. BMC Bionformatics 2006 (7) (2006) 312–324

[3] Cilia, E., Moschitti, A., Ammendola, S., Basili, R.: Structured kernels for the automatic detection of protein active sites. In: Proceedings of the International Workshop on Mining and Learning with Graphs. (2006) 117–124