ECCB 2010 Poster
Predicting protein function with the relative backbone position kernel
Leander Schietgat, Sunil Aryal and Jan Ramon

1 Teaser
We propose a kernel method for predicting the function of proteins that makes use of 3D structural information. Using the kernel in support vector machines, we obtain a state-of-the-art accuracy on two datasets of protein structures.

2 Motivation
Proteins are macromolecules that play crucial roles in many biological processes. As more data about proteins become available, the automatic classification of their function is an important challenge in bioinformatics. A lot of techniques have been proposed that predict function based on the primary and secondary structure of the proteins. However, little attention has gone to 3D structures of proteins, while these carry a lot of additional information. Recently, it was shown that kernel methods provide good results on protein classification tasks. However, none of them use 3D information.

3 Materials and Methods
We propose a new kernel for proteins called the relative backbone position kernel (RBPK). It makes use of 3D information by comparing Euclidean distances between the residue atoms and the backbone atoms of the protein. In this way, the kernel can select spatial features that are important for interactions with ligands or other proteins and will influence protein function. We evaluate our kernel through two datasets: one contains protein structures that have to be classified into enzymes and non-enzymes, while for the second one, the task is to predict the resistance of HIV protease structures.
4 Results

We compare the performance of RBPK with the Fast Subtree Kernel (FSTK), which is a state-of-the-art kernel for protein function classification. FSTK uses a graph with amino acids as vertices and distances as edges for the representation of the proteins. Although FSTK is more efficient to compute than RBPK, the latter obtains a higher predictive accuracy, resulting in state-of-the-art results for the two datasets.

5 Discussion

Our experimental results show that RBPK, which exploits 3D information of the protein, leads to more accurate predictions over a recently proposed graph-based kernel. The accuracy of above 85% for the first dataset (D&D) is an encouraging result. However, the computational efficiency of the kernel still remains an important issue, especially for large proteins. There are several ideas to improve this, by limiting the amount of distances that are computed by the kernel.

6 Acknowledgements

This research is funded by the ERC Starting Grant 240186: Mining Graphs and Networks: a Theory-based approach.