Predicting Protein Function and Protein-Ligand Interaction with the 3D Neighborhood Kernel (Extended Abstract)[1]

Leander Schietgat Thomas Fannes Jan Ramon

Dept. of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium

Introduction Kernels for structured data have gained a lot of attention in a world where increasingly complex data are continuously generated. For example, biological databases contain thousands of 3D structures of proteins, sometimes with small molecules, called ligands, bound to them. However, learning methods exploiting these kinds of data are scarce and have several limitations. They usually convert the 3D structures into graphs and use existing graph mining methods such as kernels or distance measures to compute structural similarities [2]. By transforming the 3D structures into graphs, information about angles and exact distances is lost. Moreover, these methods are dealing with severe complexity issues while handling proteins, which are an order of magnitude larger than small molecules. We propose a new kernel for 3D data, called the 3D Neighborhood Kernel (3DNK), which takes spatial distances directly into account, focusing on geometry rather than relationships in a graph. We evaluate 3DNK on two biological tasks: predicting the function of proteins and predicting interactions between proteins and ligands. While we apply this kernel to proteins and ligands, it is applicable to any kind of 3D data where objects follow a common schema, such as RNA, cars, or faces.

The 3DNK kernel For an introduction to kernel methods, we refer to [3]. We represent the 3D structures as a point set, each point p having 3D coordinates and a label $\lambda(p)$. The idea of the 3DNK kernel is to compare point sets based on their 3D structure: *(i)* for each of both point sets, a subset of points is selected (called the *selected* points) according to a user-specified criterion Δ; *(ii)* for each selected point, its neighborhood is retrieved according to a user-specified neighborhood function Φ; and *(iii)* for each point in the set of selected points, d_Φ returns a feature vector describing the local spatial conformation of that point in its neighborhood, i.e. the distances to the other points in that neighborhood. The kernel or similarity between two point sets X and Y is then calculated by comparing the feature vectors of all pairs of identically labeled, selected points:

$$K_{\Delta,\Phi}(X,Y) = \sum_{a \in \Delta(X)} \sum_{b \in \Delta(Y)} K_G\left(d_\Phi(X,a), d_\Phi(Y,b)\right) \cdot I\left(\lambda(a) = \lambda(b)\right),$$

where K_G is a Gaussian-based distance kernel, and $I(x) = 1$ if x is true, 0 otherwise.

In order to solve the two biological tasks, we create several instantiations of 3DNK by using an appropriate selection function Δ and neighborhood function Φ. To predict protein function, the selected points are the side chain atoms of the protein, while the neighborhood consists of either the nearest n backbone atoms (3DNK$_{nn}$) or the n backbone atoms in the window defined by the protein sequence around the nearest backbone atom (3DNK$_{sw}$). To predict protein-ligand interaction, the selected points are the ligand atoms and the neighborhood consists of the nearest atoms from the protein binding pocket, either a set containing all atom type labels (3DNK$_{nn}$) or clustered by atom type label (3DNK$_{at}$).

[1]The full paper has been published in *Proceedings of the International Conference on Discovery Science*, 2015.

Table 1: AUROC of 3DNK and the state-of-the-art methods for the benchmark classification datasets. The best scoring method per dataset is indicated in bold. For EC and GO, averaged results are reported.

DATASET	3DNK$_{sw}$	3DNK$_{nn}$	FSTK	NSPDK	MAMMOTH
HIV	0.848 ± 0.008	0.853 ± 0.008	0.717 ± 0.010	$\mathbf{0.896 \pm 0.007}$	0.863 ± 0.008
EC	0.575 ± 0.021	$\mathbf{0.600 \pm 0.021}$	0.573 ± 0.021	0.535 ± 0.021	0.536 ± 0.021
GO	0.744 ± 0.033	0.710 ± 0.035	0.687 ± 0.035	0.660 ± 0.036	$\mathbf{0.859 \pm 0.026}$

Experimental evaluation We evaluate the predictive performance of the 3DNK variants on four datasets and compare it with four state-of-the-art methods.

Datasets We use three classification datasets (with the number of examples ranging from 998 to 2048) in the context of protein function prediction: predicting HIV resistance (HIV), predicting enzyme class (EC) and predicting Gene Ontology term (GO) [5]. For the prediction of protein-ligand interactions, we use the regression dataset PDBbind (1300 examples) [1]. The goal is to predict the logarithm of the binding affinity between a ligand and a protein, which is a real number.

State-of-the-art methods We compare with two well-known graph kernels and two methods designed specifically to solve the aformentioned biological tasks. The Fast Subtree Kernel (FSTK) is a graph kernel based on the Weisfeiler-Lehman test for graph isomorphism [4]. The Fast Neighborhood Subgraph Pairwise Distance Kernel (NSPDK) is a graph kernel based on pairwise distances of neighborhood subgraphs [2]. The Mammoth kernel is based on the 3D structural alignment between two proteins [5]. RF-Score uses random forests using features representing occurrence counts of certain atom-type pairs between proteins and ligands [1].

Experimental methodology We evaluate the performance of the kernel methods by running support vector machines on their kernel matrices. We used 10-fold cross-validation on HIV, EC and GO and reported AUROC. We optimized the parameters of the different methods using an internal 5-fold cross-validation. For PDBbind, we used the same training and test split as in [1], tuned parameters with a 10-fold cross-validation on the training set and reported the Pearson's correlation coefficient (R) in order to compare with the published results of RF-Score [1].

Results The results in Table 1 and on the PDBbind dataset show that the different instantiations of 3DNK perform competitively when compared to the state-of-the-art methods. For PDBbind, 3DNK$_{at}$ ($R = 0.730$) and 3DNK$_{nn}$ ($R = 0.652$) were ranked 2nd and 4th, respectively, out of 22 methods, while 3DNK$_{at}$ was not significantly different than the top-scoring method RF-Score ($R = 0.776$). FSTK and Mammoth were not able to produce results on this type of dataset.

Conclusions We introduced the 3DNK kernel, which acts on 3D structures, applied it to two biological tasks and compared it to four state-of-the-art methods. The 3DNK is more broadly applicable than these methods and can solve both tasks equally well. In future work, we will explore various aspects (such as the parameter space and runtimes) of the 3DNK family further and consider other application domains.

Acknowledgements This research was supported by ERC-StG 240186 MiGraNT and IWT-SBO Nemoa.

References

[1] P.J. Ballester, J.B.O. Mitchell. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

[2] F. Costa and K. De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 255-262, 2010.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel Based Methods*. Cambridge University Press, UK, 2000.

[4] N. Shervashidze and K. Borgwardt. Fast subtree kernels on graphs. In: *Advances in Neural Information Processing Systems 22*, pp. 1660–1668, 2009.

[5] J. Qiu, M. Hue, A. Ben-Hur, J.-P. Vert, and W. Stafford Noble. A structural alignment kernel for protein structures. *Bioinformatics*, 23(9):1090–1098, 2007.