Predicting HIV Resistance with the 3D Neighborhood Kernel

Leander Schietgat LEANDER.SCHIETGAT@CS.KULEUVEN.BE
Ward Lanssens WARD.LANSSENS@STUDENT.KULEUVEN.BE
Thomas Fannes THOMAS.FANNES@CS.KULEUVEN.BE
Department of Computer Science, Celestijnenlaan 200A, 3001 Leuven, Belgium

Jan Ramon JAN.RAMON@INRIA.FR
INRIA Lille Nord Europe, 40 avenue Halley, 59650 Villeneuve d'Ascq, France

Keywords: 3D data, kernel methods, classification, regression, HIV, proteins, homology modelling

Abstract

Recently, we developed the 3D Neighborhood Kernel (3DNK), which acts on 3D structures of small molecules and proteins. We showed its state-of-the-art performance on several biological datasets. However, 3D data are in many cases difficult to obtain. For this reason, we adopt a different strategy: instead of requiring actual 3D structures, we use as input protein sequences, of which we approximate the 3D structure through homology modelling. Then, we apply 3DNK on the approximated 3D protein structures and show that, on the task of predicting HIV resistance, we obtain better results than when using a kernel function based on the protein sequences alone.

1. Introduction

An increasing number of machine learning techniques are able to exploit structure in data, with many applications, not in the least in the biological domain, where sequences and structures of proteins, interaction networks, or RNA structures have been used to predict a variety of interesting properties (King et al., 1996; Deforche, 2008; Shervashidze & Borgwardt, 2009). Recently, we introduced the 3D Neighborhood Kernel (3DNK), which can exploit geometrical relationships within 3D structures of proteins and small molecules (Schietgat et al., 2015).

In this work, we will address a specific biological task: predicting resistance of HIV proteins to drugs. Drug resistance is a recurrent problem in modern medicine, not only relevant for HIV treatment but also for treatment with antibiotics, for example. Being able to accurately predict whether a virus or bacteria is resistant against a particular drug can improve the treatment significantly. A number of algorithms tackling this task have been introduced (Vercauteren & Vandamme, 2006), but they are limited to sequence data. Such methods are not able to exploit the geometrical relationships between the protein's atoms, which are important when considering the binding to a drug for example. Therefore, we use a 3D method (3DNK), which takes into account distances between atoms from the backbone (or skeleton) of the protein and atoms from the side chains (which form the protein's exterior).

However, because it is still much more expensive to obtain a 3D structure of a protein than obtaining a sequence, 3D data are scarce. In order to circumvent this limitation, we will convert protein sequences into structures using homology modelling and then use the predicted structures to train a model with 3DNK. Interestingly, while the predicted protein structures are not perfectly accurate, the results show that the 3DNK model based on the predicted structures works better than a kernel method based on the sequences (also called 1D structure) of the same proteins.

2. Methods

Homology modelling Given an amino acid sequence of a protein, homology modelling computes an approximation of its 3D structure (Šali & Overington, 1994). The underlying assumption is that proteins with a similar sequence have a similar structure. Therefore, the input sequence is first aligned with other sequences

Appearing in *Proceedings of Benelearn 2016*.

for which the structure is known (called *templates*), and then the structure of the most similar ones are used to model the structure for the input sequence. A number of possible solutions are generated and scored according to an energy-based criterion. The resulting protein structure is then the solution with the lowest score. The method works best when there is high similarity between the input sequence and the templates, which is especially the case for HIV data.

3D Neighborhood Kernel (3DNK) For an introduction to kernel methods, we refer to (Cristianini & Shawe-Taylor, 2000). 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 (Schietgat et al., 2015). To predict HIV resistance, the selected points are the side chain atoms of the protein, while the neighborhood consists of the nearest n backbone atoms (with n a parameter specified by Φ).

3. Experimental Evaluation

3.1. Datasets

Classification Sequences of HIV-1 protease proteins were retrieved from genotype-treatment data of the Stanford HIV resistance database (Shafer & Rhee, 2016). These sequences were extracted from patients who were untreated (labeled not resistant) or treated with the inhibitors IDV and NFV (labeled resistant). This leads to two classification datasets: IDV_{Cl} (2159 sequences) and NFV_{Cl} (2192 sequences).

Regression We also extracted two regression datasets from genotype-phenotype data of the same database, which contains activity data of in-vitro experiments (a real number between 0 and 1000): IDV_{Re} (276 sequences) and NFV_{Re} (326 sequences).

Table 1. AUROC and MSE of the different methods for the classification and regression datasets.

Dataset	1D	3DNK	Rega
Classification (AUROC)			
IDV_{Cl}	0.81	**0.99**	0.73
NFV_{Cl}	0.83	**0.99**	0.76
Regression (MSE)			
IDV_{Re}	2.01	**1.46**	-
NFV_{Re}	2.05	**1.73**	-

3.2. Methodology

We compare 3DNK to a kernel which acts on protein sequences. $K_{1D}(p,q) = \exp\frac{-d(p,q)}{\sigma^2}$, with $d(p,q)$ the Hamming distance between sequence p and q and σ a parameter of the Gaussian distance kernel. We also compare 3DNK to the Rega algorithm, which was developed by domain experts. It consists of a number of expert rules relating certain mutations to resistance (REGA institute, KU Leuven, 2013). To model the protein structures, we used Modeller v9.15 (Webb & Sali, 2014) with standard parameters and HIV-1 templates from PDB (Berman et al., 2000).

We evaluate the performance of the kernel methods by running support vector machines (using SVM^{light} (Joachims, 2002)) on their kernel matrices. We used 10-fold cross-validation and reported AUROC for the classification datasets and mean squared error (MSE) for the regression datasets. We optimized the parameters of the different methods using an internal 5-fold cross-validation. For the regression datasets, we used the log values of the original targets.

3.3. Results

For both the classification and regression datasets, the 3D kernel outperforms the 1D kernel. Moreover, both kernel methods outperformed the REGA algorithm (Table 1). This suggests that 3DNK can extract relevant information from the 3D structures.

4. Conclusions and Further Work

In this work, we showed that, when looking at the prediction of HIV resistance, even when there are no 3D data available, it is worth approximating the 3D structure and running a 3D method. In further work, we will check whether the same conclusions hold for other datasets and tasks, and for other machine learning techniques.

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

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*, 235–242.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel based methods*. Cambridge, UK: Cambridge University Press.

Deforche, K. (2008). *Modelling HIV resistance evolution under drug selective pressure*. Doctoral dissertation, Katholieke Universiteit Leuven.

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory, and algorithms*. Springer.

King, R., Muggleton, S., Srinivasan, A., & Sternberg, M. (1996). Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences*, *93*, 438–442.

REGA institute, KU Leuven (2013). Rega algorithm. URL: `https://rega.kuleuven.be/cev/avd/software/rega-algorithm`.

Šali, A., & Overington, J. P. (1994). Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Science*, *3*, 1582–1596.

Schietgat, L., Fannes, T., & Ramon, J. (2015). Predicting protein function and protein-ligand interaction with the 3D neighborhood kernel. *Discovery Science* (pp. 221–235).

Shafer, R., & Rhee, S. (2016). Hiv drug resistance database. URL: `http://hivdb.stanford.edu/index.html`, data retrieved on 09-04-2016.

Shervashidze, N., & Borgwardt, K. M. (2009). Fast subtree kernels on graphs. *Advances in Neural Information Processing Systems* (pp. 1660–1668).

Vercauteren, J., & Vandamme, A.-M. (2006). Algorithms for the interpretation of HIV-1 genotypic drug resistance information. *Antiviral research*, *71*, 335–342.

Webb, B., & Sali, A. (2014). Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 5–6.