# Kernel Methods for Protein Function Prediction

Karsten M. Borgwardt*, Hans-Peter Kriegel [I]
Institute for Informatics, LMU Munich, Oettingenstr. 67, 80538 Munich, Germany

*To whom correspondence should be addressed: kb@dbs.ifi.lmu.de

## 1. INTRODUCTION

From a computer scientist's point of view, protein function prediction can be regarded as a classification problem, namely to correctly classify a newly discovered protein into its functional class. Besides, this computational function prediction often means dealing with different data formats, data types and data sets with missing entries. Kernel methods are a popular method from Machine Learning which can handle all these problems of classification, data compatibility, data integration, and data completion. Numerous applications of kernel methods to protein function prediction tasks have been published over recent years. In the following, we will review applications, potential and limits of kernel methods in this field.

## 2. KERNEL METHODS AND THEIR POTENTIAL

Kernels methods are based on measures of similarity called *kernel functions* that allow us to perform classification, regression and related tasks (for a complete introduction please refer to (7)). They work implicitly by mapping input data, such as proteins, into a (usually) higher-dimensional feature space and by finding a suitable hypothesis in this feature space.

In the case of classification, this hypothesis is a hyperplane in feature space which separates two classes of input data; new data points can then be classified into one of these two classes, depending on the half-space they are located in. This so-called *Support Vector Machine (SVM)* classifier maximizes the *margin*, i.e. the minimum distance between the hyperplane and data points from both classes.

The transformation from input space into feature space is connected with three huge advantages: First, two classes of data points that are not linearly separable in input space can become linearly separable if mapped into an adequate feature space. Second, all calculations in feature space can be performed implicitly via evaluating a kernel function on data points in input space; this kernel function is a dot-product in feature space and represents a measure of similarity between data points. Third, any type of data can be classified, as long as a kernel function can be defined on it. Kernels have been developed for data types such as vectors, graphs, trees and strings. Several kernel functions can be combined into one joint kernel which integrates several sources of information. Furthermore, kernel methods allow dealing with missing information in large data sets and to approximate these unknown features.

## 3. APPLICATIONS TO FUNCTION PREDICTION

The enormous potential of kernel methods in protein function prediction is reflected by a large number of publications over recent years.

One common approach is to define a kernel on proteins to quantify their similarity based on certain characteristics, e.g. their sequences, structures, chemical features, special amino acid motifs or phylogenetic profiles. Cai et al. (3) represent proteins as feature vectors comprising approximate chemical characteristics of their sequences. Dobson and Doig (4) describe protein structures as feature vectors including information about molecular surface, secondary structure, ligands, bonds and surface clefts. Borgwardt et al. (2) integrate both sequence and structure information into one graph model of proteins that is further enriched by approximate chemical properties. In all three studies, representations of proteins are then classified into functional classes using SVMs. Vert (9) proposes a tree kernel to analyze phylogenetic profiles by incorporating knowledge about the phylogenetic relationship among species. Via SVMs and other kernel methods, Vert then detects functional relationships based on these profiles. Ben-Hur and Brutlag (1) define kernels based on discrete functional motifs, i.e. proteins are deemed similar if they share sequence patterns that have been found to be associated with certain functions.

[I] This work was supported in part by the German Ministry for Education, Science, Research and Technology (BMBF) under grant no. 031U112F within the BFAM (Bioinformatics for the Functional Analysis of Mammalian Genomes) project which is part of the German Genome Analysis Network (NGFN).

As an advantage, kernel methods can also be employed for data integration. Lanckriet et al. (6) were the first to show a principled way of integrating various types of genomic information into one kernel for function prediction that performs better than any kernel on a single data source.

Dealing with missing information is another problem that kernel methods can help to tackle. Kin et al. (5) estimate structural similarity between proteins without known structure by approximating unknown structure kernel matrix entries via a sequence kernel matrix. Tsuda and Noble (8) define a kernel on interaction networks that allows to predict functions of unannotated proteins by maximizing entropy.

## 4. LIMITS

Despite their wide range of applicability in function prediction, kernel methods are not devoid of problems, both from a computational and a biological point-of-view.

Adequate kernel design, class representation and multi-class classification can bear difficulties for the practitioner. First, kernel methods will perform well if the kernel function employed is a good measure of similarity in the area of application. Kernel functions must therefore be designed by an expert with domain knowledge, as automatic choice of good kernel functions is not possible yet. Second, classification via kernel methods and SVMs provides support vectors which represent “borders” of classes in feature space, not the classes themselves. Kernel methods are less adequate if one is interested in a direct representation of classes which helps to define characteristics of these classes. Third, SVMs are originally defined as binary classifiers for two classes of data. Standard multi-class SVMs are either based on binary “one class versus rest” classifiers or binary “one class versus one class” classifiers for all pairs of classes. In the first case, the number of classifiers to be trained grows linearly with the number of classes, in the latter case, it grows quadratically with the number of classes, which can result in runtime problems.

From a biological perspective, a major disadvantage of current kernel methods for function prediction is the fact that they are based on simplified models of proteins. While this simplification keeps the calculation of the kernel matrix computationally feasible, it may lead to a loss of information that is important for exact function determination. Due to these simplifications, kernel methods are not able to distinguish functions of proteins correctly that are closely related, yet functionally different. In short, current kernel methods are more adequate for *function class prediction* than for *specific function determination*. Consequently, as other computational approaches to function prediction, kernel methods may reduce the number of lab experiments required to determine exact protein function. However, they cannot make experiments completely unnecessary.

To conclude, we feel that protein function prediction will further benefit from kernel method applications in future, while creating new challenges to improve kernel design, kernel evaluation performance, data integration and data approximation.

## 5. REFERENCES

1. Ben-Hur, A. and Brutlag, D. 2003. *Remote homology detection: a motif based approach*. Bioinformatics, 19 Suppl 1:i26-33.
2. Borgwardt, K. M., Ong, C. S., Schönauer, S., S.V.N. Vishwanathan, Smola, A.J., and Kriegel, H.-P. 2005. *Protein Function Prediction via Graph Kernels*. ISMB 2005, in press.
3. Cai, C. Z., Han, L. Y., Ji, Z. L., and Chen, Y. Z. 2004. *Enzyme family classification by support vector machines*. Proteins, 55(1):66–76.
4. Dobson, P. D. and Doig, A. J. 2003. *Distinguishing enzyme structures from non-enzymes without alignments*. J Mol Biol, 330(4):771–783.
5. Kin, T., Kato, T., and Tsuda, K. 2004. *Protein Classification via Kernel Matrix Completion*. In: *Kernel Methods in Computational Biolog*y, 261-274. (Eds.) Schölkopf, B., Tsuda, K. and Vert, J.P., MIT Press (2004).
6. Lanckriet, G. R., Deng, M., Cristianini, N., Jordan, M. I., and Noble, W. S. 2004. *Kernel-based data fusion and its application to protein function prediction in yeast.* PSB 2004, pages 300–311.
7. Schölkopf, B., and Smola, A. J. 2002. *Learning with Kernels*. MIT Press.
8. Tsuda, K., Noble, W.S. 2004. *Learning kernels from biological networks by maximizing entropy*. Bioinformatics, 20 Suppl 1:I326-I333.
9. Vert, J. P. 2002. *A tree kernel to analyse phylogenetic profiles*. Bioinformatics, 18 Suppl 1:S276-84.