

Similar sequence or structure: A kernel function that brings together the best of two worlds

Isye Arieshanti¹
i.arieshanti@imb.uq.edu.au

Mikael Bodén¹
m.boden@uq.edu.au

Stefan Maetschke¹
s.maetschke@imb.uq.edu.au

Fabian A. Buske¹
f.buske@imb.uq.edu.au

¹ The University of Queensland, Institute for Molecular Bioscience, Qld 4072, Australia

Abstract

Sequence and structure homology can be combined to infer protein function. A kernel function is proposed that combines sequence and structural features. When utilized with a support vector machine (distinguishing between enzymes and non-enzymes) this kernel shows higher accuracy than a kernel that exploits either sequence or structural features exclusively. We show that the proposed kernel makes both sequence homology and structure homology more accessible compared to the best kernels that specialize in either. The homogeneous organization of the feature space explains the superior classification accuracy of the combined kernel on the enzyme challenge problem.

Keywords: kernel function, homology

1 Introduction

Inferring function of novel proteins by homology to experimental data is regarded very reliable at high levels of similarity. Problems such as enzyme classification demand methods that exploit both sequence and structural similarity: functionally similar enzymes can be diverse both in sequence and structure. In specific studies, the *ad hoc* combination of sequence and structural features has indeed attained superior prediction accuracy [1,2]. Such evidence of partial success begs explanations and continued exploration of methods that are able to perform such transfer of annotation between weakly homologous proteins.

In this study, we propose and evaluate a new *kernel function* that combines sequence and structural features. Acting like an interface between real biological data and the arsenal of kernel machine learning methods, our kernel simply embeds one sequence and one structure kernel. The ability of this *hybrid* kernel to leverage both sequence and structural features of proteins is evaluated in two ways: First, we benchmark a support-vector machine equipped with alternative kernels for classifying novel proteins as enzymes. Second, we analyze the kernel-specific feature spaces to illustrate how *accessible* biologically meaningful features are. Indeed, as described below, classification accuracy is at its peak when both aspects of protein similarity are utilized. Our analysis highlights the complementarities of sequence and structure captured by our hybrid kernel function to detect and leverage weak homology.

2 Method and Results

Recent years have seen increasing interest in applying kernel methods (e.g. support-vector machines) using domain-tailored kernel functions. Based on the contribution of all possible local alignments between sequences, the local alignment kernel [3] shows superior ability to detect amino acid sequence similarity. A profile-based adaptation of the local alignment kernel (incorporating the scoring matrix from a PSI-BLAST search) is particularly capable of finding remote homologues [4]. Based on atomic distances between structural elements of a protein structure, the graph kernel function establishes a purely structural view of a protein [5]. Inside a support-vector machine, this kernel was shown to outperform other kernels to distinguish between known enzyme classes. Our hybrid kernel is the simplest convolution of the profile local alignment and the graph kernels: by summing the two individual kernel results, the hybrid kernel feature space represents the concatenation of the sequence and structure feature spaces.

About one thousand proteins with limited sequence similarity are used in this study. Only 53 proteins have more than 30% sequence identity, ensuring that most pairs are beyond the level at which a standard alignment algorithm can safely generate meaningful results. Half of the proteins are known enzymes, each belonging to one of six enzyme classes. The data has been used in past studies to challenge classification models by the absence of sequence homology [6]. Since our aim is to develop a component that is able to operate under conditions with weak sequence similarity, we re-use this set. It should be noted that, by treating all enzyme classes as one group, we further challenge a classifier. It essentially needs to construct a

decision boundary that takes a diverse set of features into account. For purposes of analysis, we also tag each protein with its SCOP class (there are 737 different classes in our set), and a sequence identity cluster identifier generated by BlastClust (in total we found 961 such groups at 30% identity).

Table 1 summarizes the classification tests (enzyme vs. non-enzyme) of a support-vector machine, equipped with the different kernels studied. The table presents the prediction accuracies and the corresponding standard deviations. Prediction accuracy was measured as the average AUC (area under ROC curve), over 15 repeats of five-fold cross-validation.

Table 1: Classification performance comparison

Kernel	Average AUC
Profile Local Alignment	0.705 \pm 0.005
Graph	0.718 \pm 0.003
Hybrid (Graph + PLA)	0.730 \pm 0.003

To understand what biologically meaningful information that is accessible, we performed a cluster analysis of all the feature spaces. In brief, each group of proteins belonging to a single cluster was evaluated by looking at the protein tags. In turn, we computed an information theoretic entropy score for (a) SCOP classification, (b) sequence cluster identifier, and (c) enzyme class, indicating if proteins are grouped in accordance with them (see Table 2). Low entropy indicates a homogeneous feature space (that members sharing the tag are also appearing close). High entropy indicates that organization is random.

Table 2: Cluster entropy for different kernels and protein categories

Kernel	SCOP classification (a)	Sequence-cluster (b)	Enzyme (c)
Profile Local Alignment	3.162 \pm 0.015	3.190 \pm 0.013	0.916 \pm 0.014
Graph	2.427 \pm 0.075	2.466 \pm 0.073	0.645 \pm 0.048
Hybrid (Graph + PLA)	2.360 \pm 0.079	2.402 \pm 0.081	0.332 \pm 0.029

3 Discussion

The results for the support-vector machine with the hybrid kernel show a small, but statistically significant increase in classification performance compared to those of the other kernels ($p < 0.01$). Embedding sequence features (as mapped by the profile local alignment kernel) and structural features (as mapped by the graph kernel) thus improves the ability of the kernel machine learning method to distinguish between the two broad functional (but sequentially and structurally diverse) classes of proteins. Previous studies are in agreement with this insight but suggest little in terms of explanation of any improvement.

The kernel function establishes a feature space to which all proteins are mapped. It is in this feature space that the kernel method (the support-vector machine) operates and any similarities can be leveraged. The hybrid feature space embeds the profile local alignment and the graph kernel feature spaces. The feature space analysis in Table 2 shows that the feature space of the hybrid kernel is more homogeneous for all analyzed protein categories (a), (b) and (c) compared to the feature spaces of both the profile local alignment and the graph kernel. It is suggested that the hybrid kernel makes both sequence and structure features more accessible to the kernel method. The greater homogeneity of the hybrid kernel feature space explains the superior classification accuracy on the enzyme vs. non-enzyme problem.

References

- [1] Armon A, Graur D, Ben-Tal N: ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, 307(1):447-463, 2001.
- [2] Dobson PD, Doig AJ: Distinguishing enzyme structures from non-enzymes without alignments. *J Mol Biol*, 330(4):771-783, 2003.
- [3] Vert J-P, Saigo H, Akutsu T: Local Alignment kernels for Biological Sequences. *In: Kernel Methods in Computational Biology*. Edited by Scholkopf B, Tsuda K, Vert J-P: The MIT Press: 131-154, 2004.
- [4] Rangwala H, Karypis G: Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, 21(23):4239-4247, 2005.
- [5] Borgwardt KM, Ong CS, Schonauer S, Vishwanathan SVN, Smola AJ, Kriegel H-P: Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1):i47-56, 2005.
- [6] Qiu J, Hue M, Ben-Hur A, Vert J-P, Noble WS: A structural alignment kernel for protein structures. *Bioinformatics*, 23(9):1090-1098, 2007.