

Detecting Sequence and Structure Homology via an Integrative Kernel: A Case-Study in Recognizing Enzymes

Isye Arieshanti, Mikael Bodén, Stefan Maetschke and Fabian A. Buske

Abstract—Sequence and structure are complementary pieces of information that can be used to infer protein function. We study and compare sequence, structure and sequence-structure integrative kernels to recognize proteins with enzymatic function. Using a support-vector machine, we show that kernels that combine sequence and structure information typically perform better (AUC 0.73) at this task than kernels that exploit either type of information exclusively. We find that the feature space of structure kernels complements that of sequence kernels, making both sources of similarity more accessible to kernel methods

I. INTRODUCTION

HOMOLOGY refers to similarity due to descent from a common ancestor. Inferring function of novel proteins by homology with existing experimental data is regarded very reliable at high levels of similarity. Homology-based methods thus hold promise to efficiently annotate the increasing amounts of genomic data amassed by current sequencing efforts. To advance methods for functional annotation, this study recognizes and leverages the complementarities of two forms of homology: by sequence and by structure. Specifically, we explore a *kernel*-based integration of sequence and structural similarity for transferal of enzyme classification to novel proteins. Often based on catalytic mechanisms manifested structurally, enzymes can have identical function with limited sequence similarity (see Figure 1). Conversely, proteins with similar sequences may have disparate function [1] (see Figure 2) making them particularly challenging to annotate.

Several studies use approximately 40% sequence identity as a threshold for transferring Enzyme Commission (EC) numbers with partial success [2], [3], [4], [5]. Such sequence similarity can be derived by multiple means, some directly based on sequence alignment. Sequence similarity becomes a poor indicator of homology at lower similarity levels - with 15% to 30% being the "twilight zone", and below 15%, the "midnight zone" [6]. To compensate for poorly supported sequence homology, other studies use fold similarity to transfer functional annotation to a query protein. Fold similarity can be measured using methods based on structural alignment [7], [8], graph theory [9], [10], [11], and statistics of local structural features [12]. It is generally believed that structure is more conserved than sequence [13] - attesting to the importance of considering structural features. However, it has been shown that a single fold can be responsible for diverse biological functions [14], [15], complicating matters further.

This work was supported in part by the ARC Centre of Excellence in Bioinformatics. All authors are with The University of Queensland, Institute for Molecular Bioscience (e-mail: i.arieshanti@imb.uq.edu.au).

Problems like enzyme classification thus demand methods that rely on sequence and structural similarity combined. Several studies have explored *ad hoc* features of both kinds [16], [17]. The common theme amongst these studies is to increase prediction accuracy by combining features that complement each other. We ask what is the underlying basis of such prediction improvements on poorly aligned protein sequences?

In this paper, we study three different approaches to integrate sequence and structural features. Our focus is hereby on kernel methods, due to their capability to process structured data directly and the availability of highly efficient training algorithms. Kernel methods are based on data structure specific kernel functions that define a feature space in which decisions are made for any input. We employ a Support Vector Machine (SVM) equipped with a range of different kernels to combine sequence and structure features at what Pavlidis *et al.* call the "early", the "intermediate" or the "late" stage [18]. The quality of the integration is evaluated by comparing the performance of the SVM on classifying novel proteins as enzymes or non-enzymatic proteins. We contrast the organization of kernel-specific feature spaces, utilizing kernel *K*-Means [19]. The analysis of a hybrid kernel feature space highlights the importance of using complementary sources to detect and leverage homology beyond the twilight zone.

II. METHODS AND MATERIALS

A. Data set

All our experiments are based on a dataset by Qiu *et al.* [8] that itself is derived from a dataset constructed by Dobson and Doig [17]. The dataset is balanced and comprises 498 enzymes and 498 non-enzymatic proteins. Removing of all sequences that contain an unknown residue ('X'), which could not be encoded by the profile local alignment kernel, leads to a final dataset with 493 enzymes and 492 non-enzymes.

The data has been used in past studies to challenge classification models by the absence of sequence homology [8]. We use this set since our aim is to develop a component that is able to operate under conditions with weak sequence similarity. 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.

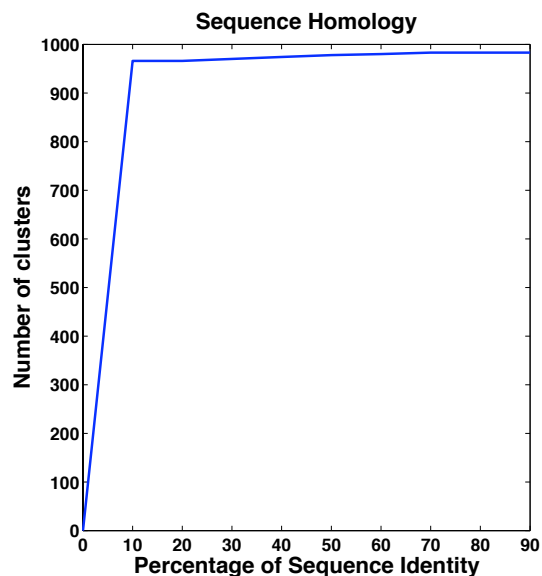


Fig. 1. Number of cluster groups against percentage of sequence identity. Clustering is performed using the NCBI BLASTCLUST program. The plot shows that proteins in the dataset have very low sequence identity.

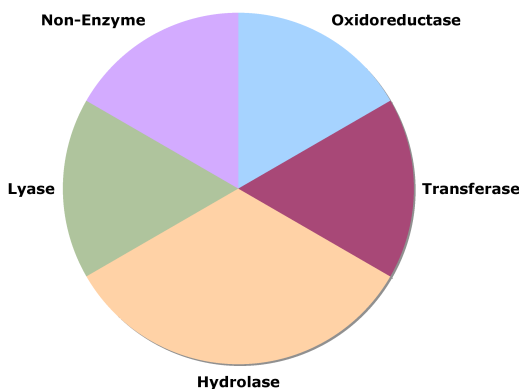


Fig. 2. An example of how a single superfamily (alpha/beta-Hydrolases) in the dataset has members with several functions: transferase, oxidoreductase, lyase, hydrolase and non-enzyme.

B. Kernel Methods and Kernels

The core of our prediction model is a Support Vector Machine (SVM) that utilizes Platt’s Sequential Minimal Optimization method [20] for training. The SVM finds a decision boundary in the kernel feature space defined in terms of a subset of the training samples known as the support-vectors. To analyze the homogeneity of the feature space, we employ kernel K -Means [19]. This method identifies groups of samples, deemed similar by distances in the kernel-defined feature space.

We equip our prediction models with different kernels that exploit specific features of a protein’s sequence or structure. The *spectrum* kernel calculates the similarity between two protein sequences by calculating the dot product between

vectors that hold the frequencies of l -mers for any pairs of sequences [21]. In this study we use $l = 3$, a setting that has been successful in several protein feature predictions studies. The *mismatch* kernel measures the similarity between two protein sequences in a similar way to the spectrum kernel, but it allows up to m mismatches. We use $l = 4$ and $m = 1$. The *local alignment* kernel computes the similarity between two protein sequences by exploring all possible alignments [22]. We employ the BLOSUM62 substitution matrix with the default setting [22] to evaluate each alignment. The *profile local alignment* kernel operates just like the local alignment kernel but evaluates the profile generated by PSI-BLAST in place of the BLOSUM62 substitution matrix [23]. Profiles disclose important evolutionary information tied to each residue in sequence. PSI-BLAST profiles are created from a repeated multiple alignment of the highest scoring hits of a BLAST search. We used the following typical parameter settings of 0.001 for the E -value and three iterations, with the *uniref90* non-redundant dataset as database.

The *graph* kernel represents a protein as a graph, with secondary structure elements (SSE) as nodes, and SSE distances and/or spatial distances as edges. The similarity between two proteins is determined by counting common sub-graphs between two full graphs [9]. Features for the graph kernel include secondary structure elements [24], physical-chemical properties of amino acids [25] contained in secondary structure elements and spatial distance of the C-alpha atoms. The weighting factor was set to 1.

The *contact* kernel represents a protein as a set of c -body residue interactions among the amino acids in its structure. A set of c residues is considered to be in mutual contact if, for each pair of residues from the set, the distance between any pair of atoms is below a predefined distance threshold [8]. Two proteins are considered similar if they have similar contact spectra. For both the graph kernel and the contact kernel, we utilize structure data from PDB [26]. In the experiment we report on here $c = \{2,2,2,3,3,4\}$, distance threshold = $\{6.5, 8.0, 9.5, 8.0, 9.5, 9.5\}$ with respect to the C-alpha atom and the sequence separation between residue pairs $s = 3$.

C. Integration models

We explore three basic approaches of integrating sequence and structure features: via “early”, “intermediate” and “late” integration (see Figure 3). The late integration model consists of two SVMs, each equipped with a separate kernel (one for sequence and one for structural information). The output of the model is the sum of the outputs of the two SVMs. (Note that the SVM output is not subject to the usual 0-threshold.)

For intermediate integration, a sequence and a structure kernel are first computed separately and then element-wise added to create a hybrid kernel, which is used by a single SVM.

Early integration refers to a model that directly exploits sequence and structure information within a single kernel. In our experiments we employ the contact kernel for this purpose as described below.

III. RESULTS

We have three aims in this study. The first aim is simply to construct a predictive model capable of integrating sequence and structural features. We embed sequence features by choosing the best performing sequence kernel from the selection of sequence kernels, including the spectrum, mismatch [21], local alignment [22] and profile local alignment kernel [23]. The structural features in our model are evaluated by the graph kernel [9].

The second aim is to assess the integration of sequence and structure features on basis of enzyme vs. non-enzymatic protein classification. The classification performance of the SVM is measured by the AUC (Area Under ROC curve) using 5-fold cross-validation, run fifteen times (from different dataset divisions). For this, the decision value of the SVM is the ROC converted to a binary output (positive or negative) by the use of a variable threshold. Features are integrated at an early, intermediate or late stage [18] (see Figure 3). We compare the performance of the three different integration models and also discuss the achieved prediction accuracies in relation to previous work on enzyme classification.

Intermediate integration involves superimposing sequence and structural features. Thus, the third and final aim is to understand single-kernel compared to mixed-kernel feature spaces. Specifically, we ask how the complementarities of individual kernels practically manifest themselves in the mixed setting. We do so by probing the organization of single and mixed feature spaces in terms of biologically meaningful sequence and structural features.

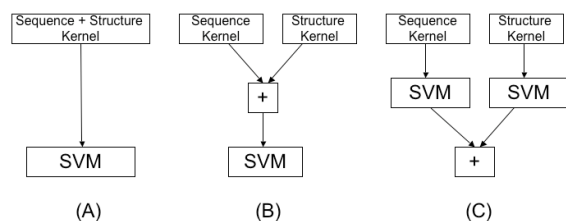


Fig. 3. (A) Early integration model. (B) Intermediate integration model. (C) Late integration model.

A. Sequence Features

To make an informed decision of which sequence kernel to incorporate in the intermediate and late integration models, we compare the performance of the spectrum, mismatch, local alignment and profile local alignment kernel when used to classify proteins as enzymes or non-enzymatic proteins. The profile local alignment kernel produces the best ROC with the smallest standard error compared to the other three sequence kernels (see Figure 4).

The profile local alignment kernel also yields the top AUC value of 0.705 (see Table I), which is significantly higher ($p < 0.01$) than any of the alternatives according to a *Kruskal-Wallis* rank sum test. This result is not surprising, since alignment-based string kernels and those that leverage the evolutionary information reflected by “profiles” have

performed strongly for many other protein classification problems reported in the literature, e.g. remote homology detection [23]. The PSI-BLAST derived profile encoding used in the profile local alignment kernel clearly contributes to higher classification accuracy. We thus choose the profile local alignment kernel to represent sequence features.

B. Structure Features

Both the graph and the contact kernel exploit structural information of the protein to classify. The average AUC values for the graph kernel and the contact kernel are 0.718 and 0.730 respectively (see Table I). The ROC curves with standard error bars are shown in Figure 5. Note however, that the contact kernel is not a pure structure kernel but incorporates sequence information as well. Similar to the spectrum kernel, it evaluates spectra built from the amino acids that take place in all the c -body contacts of a protein. The contact kernel’s usage of such rich information of the protein structure appears to contribute positively to the classifier’s ability to distinguish between enzymes and non-enzymatic proteins. We employ the graph kernel to represent structure features in the intermediate and late integration models, while the contact kernel is used as an example of early integration.

C. Integration of Sequence and Structure Features

Dobson and Doig [17] perform early integration by concatenating a vector representing sequence features and a vector representing structural features of the query protein. The concatenated vector is then processed via a standard kernel by a SVM. However, proteins are intrinsically variable and do not naturally map to fixed-length vectors as required by early integration. Instead, we leverage the power of data structure tailored kernels. Specifically, we perform early integration by employing the contact kernel (Contact) and utilize the profile local alignment kernel (PLA) and the graph kernel (Graph) for intermediate (Graph+PLA) and late integration (Graph/PLA).

It should be noted that early and intermediate integration allows one SVM to detect and exploit dependencies across the sequence and structure features spaces. The two SVMs used in the late integration model optimize their decision boundaries in isolation, thus unable to access inter-feature space relations.

Table I compares the prediction performances of the different sequence and structure kernels and integration models. Based on the *Kruskal-Wallis* rank sum test, it is apparent that models that take structural information into account (the last four entries in the table) achieve significantly ($p < 0.01$) higher AUC values than kernels that rely on sequence information only. Furthermore, combining sequence and structure information as performed by the integration models generally results in increased prediction performance relative to pure sequence or structure kernels. However, in contrast to Pavlidis *et al.* [18] we find very little difference in prediction accuracy between early, intermediate and late integration models for the enzyme vs. non-enzyme classification task.

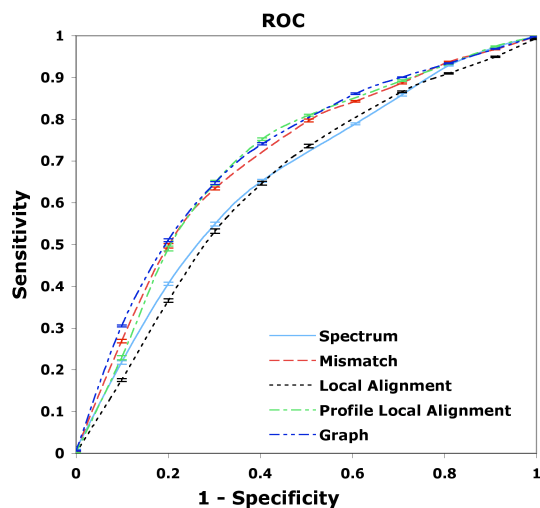


Fig. 4. ROC curves with standard error bars representing the average accuracies of a classifier utilizing the individual kernels (sequence kernels and structure kernel) listed in the legend.

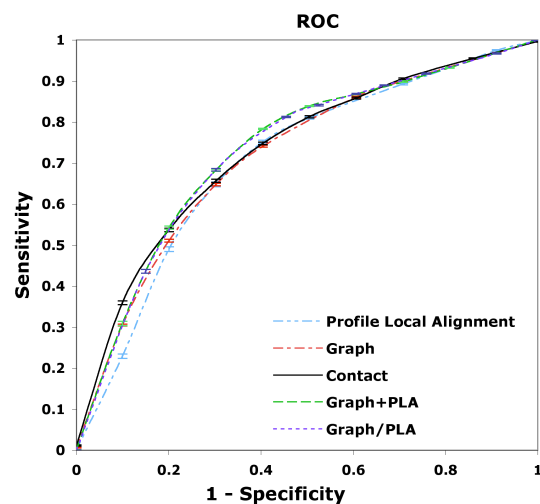


Fig. 5. ROC curves with standard error bars representing the average accuracies of integration models (in contrast to one sequence-only and one structure-only model).

IV. ANALYSIS

To gain a better understanding of the differences in performance of the evaluated kernel methods, we analyze the corresponding kernel feature spaces using K -Means clustering. It is intuitive that the more similar two samples of the same class appear in a given feature space, the more likely a predictor is able to classify them equally. In the following, a biologically meaningful label is attached to each sample of the data set. We subsequently test if samples that are close in a given feature space share the same label (a "homogeneous cluster") or if they have different labels (a "heterogeneous cluster"). The degree of homogeneity in each cluster is determined by calculating its information entropy. The average homogeneity calculated over all clusters thus illustrates how well the feature correlates with the label space.

TABLE I
KERNEL AND INTEGRATION MODEL COMPARISON

Kernel	Feature	Average AUC	Standard Deviation
Spectrum	Sequence	0.657	0.005
Mismatch	Sequence	0.700	0.004
Local Alignment	Sequence	0.644	0.006
Profile Local Alignment	Sequence (profiles)	0.705	0.005
Graph	Structure -SSE (turn,helix,sheet) -spatial distance (C α distance) -SSE length	0.718	0.003
Contact	Hybrid -sequence -C α distance	0.730	0.004
Graph+PLA	Hybrid -sequence (profiles) -structure features (the graph kernel features)	0.730	0.003
Graph/PLA	Hybrid -sequence (profiles) -structure features (the graph kernel features)	0.729	0.004

In more detail, a set of K clusters is first identified using kernel K -Means. A cluster $k \in \{1, \dots, K\}$ has n_k members. For each label $l \in \{1, \dots, L\}$ (where L is the number of labels), $cnt_l(k)$ is the total count of cluster k members that are labeled l .

The entropy for the k^{th} cluster is denoted as E_k

$$E_k = \sum_{l=1}^L -((cnt_l/n_k) \log_2(cnt_l/n_k)) \quad (1)$$

The average entropy over all clusters is denoted as \hat{E}

$$\hat{E} = \frac{1}{K} \sum_{k=1}^K E_k \quad (2)$$

We label the same dataset (as used for the kernel evaluation above) in nine different ways. First, proteins receive labels according to sequence homology as determined by BLASTCLUST (with similarity threshold 90%). Proteins appearing in the same BLASTCLUST group are assigned the same label (961 labels). Second, proteins are labeled with their SCOP super-family [27] category. This assignment thus indicates structural homology (737 different labels). It should be noted that both of these two label spaces overlap and do not represent two independent ways of looking at the data. Furthermore, since the data set has low sequence identity (see Figure 1), the sequence similarity is very limited and we find few labels that have more than one member [8]. Thirdly, proteins were labeled according to their enzymatic category (EC) including "Oxidoreductase" (EC 1) vs. "non-Oxidoreductase", "Transferase" (EC 2) vs. "non-Transferase", "Hydrolase" (EC 3) vs. "non-Hydrolase", "Lyase" (EC 4) vs. "non-Lyase", "Isomerase" (EC 5) vs. "non-Isomerase", "Ligase" (EC 6) vs. "non-Ligase". Finally, proteins were labeled as "enzyme" or "non-enzyme".

To identify any dependencies with suitable number of clusters for the K -Means analysis of the label spaces, we plot the averaged information entropy over an increasing number of clusters. Figure 6 shows the entropy for different kernels and the enzyme non-enzyme labeling (the graphs for other label spaces display a similar behavior - data not shown). We find that for 100 or more clusters the relative performance of the different kernels stabilizes. We thus choose 100 clusters for our subsequent analysis of label spaces (see Table II).

Table II compares the information entropies for the selected sequence, structure and hybrid kernels and the nine different label spaces described above. Lower entropy indicates higher cluster homogeneity. Note that the number of proteins that belong to a specific enzymatic category (EC) varies greatly (annotated in brackets) and therefore the information entropy varies considerably. For instance, there are only 19 ligases that can easily be grouped together and the corresponding cluster entropy is therefore small. However, the entropy values for a specific label category are comparable across the different kernels evaluated.

For the label spaces "BLASTCLUST sequence homology" and "SCOP super-family" we find comparatively high entropy values due to the large number of different labels for these spaces (961 and 737, respectively). The absolute higher values for these spaces do not indicate more heterogeneous feature spaces per se but are related to the more difficult clustering task than presented by the EC labels. As expected, structure or hybrid kernels generally show relatively higher homogeneity for the organization of the various label spaces (see mean values) than the profile local alignment kernel, which relies on sequence information exclusively.

The Graph+PLA model achieves the lowest entropy values over all label spaces. We find the Graph+PLA model to create the most homogeneous feature spaces but this seems to be specific to this kernel. It cannot be argued that the integration of sequence and structure information consistently leads to better-organized feature spaces, since the graph kernel shows higher homogeneity than the contact model for instance.

Furthermore, cluster entropy appears not to be closely correlated with the prediction accuracies of the different kernels as the results for the enzyme non-enzyme problem show. For instance, information entropy of the contact kernel for this task is considerably higher than that of the graph kernel but the former achieves a significantly better AUC value. Similarly, the hybrid kernels show higher homogeneity but do not perform better in terms of prediction accuracy than the contact kernel.

This discrepancy between intuition and prediction results can be explained by the fact that cluster entropy is a simplified estimator of the separability of the feature space. Or in other words, high cluster entropy does not ensure that the clusters themselves can easily be separated by a hyper plane.

V. DISCUSSION

Our results on the enzyme non-enzyme classification task show that models that take sequence and structure infor-

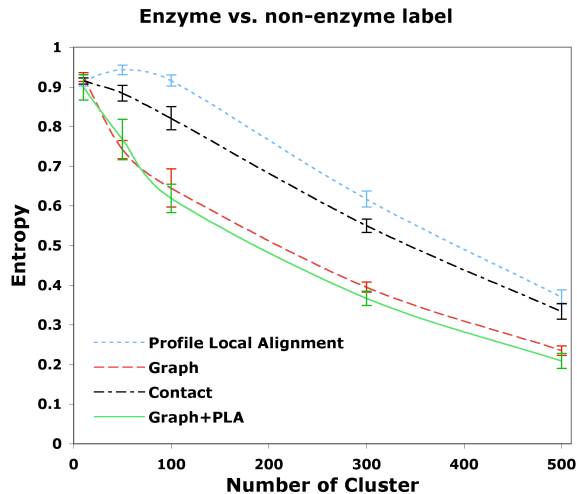


Fig. 6. Average entropy over different numbers of clusters for the profile local alignment kernel (red), the contact kernel (black), the graph kernel (blue), and the Graph+PLA kernel (green). Proteins are labeled with "enzyme" and "non-enzyme". Error bars show the standard error of the average entropy over ten repeats.

TABLE II
CLUSTER HOMOGENEITY MEASURED AS AVERAGED INFORMATION ENTROPY OVER 100 CLUSTERS FOR NINE DIFFERENT LABEL SPACES. THE LAST ROW SHOWS THE MEAN VALUES OVER TABLE COLUMNS.

	Profile Local Alignment	Graph	Contact	Graph+PLA
Sequence homology	3.19	2.47	2.93	2.40
SCOP	3.16	2.42	2.90	2.36
Oxidoreductase (79)	0.32	0.21	0.26	0.20
Transferase (127)	0.47	0.34	0.38	0.33
Hydrolase(157)	0.54	0.41	0.49	0.39
Lyase (60)	0.25	0.18	0.19	0.18
Isomerase (51)	0.22	0.16	0.18	0.15
Ligase (19)	0.09	0.06	0.07	0.06
Enzyme (493)	0.91	0.65	0.82	0.62
Mean	1.02	0.77	0.91	0.74

mation into account achieve higher prediction accuracies than models that exploit sequence or structure information exclusively. In contrast to Pavlidis *et al.* [18] we find only minor differences in the prediction performance of the early, the late and the intermediate integration model. It is, however, to note that Pavlidis *et al.* [18] applied their models to different classification tasks and employed feature selection.

In comparison to Qiu *et al.* [8], the accuracy of the hybrid kernels presented herein is similar to the performance of their aggregate kernel (the sum of all six different kernels they explored achieved an AUC of 0.73). Noteworthy is that Qiu *et al.* report an AUC of 0.75 for the contact kernel by itself, while our implementation of the contact kernel only reached an AUC of 0.73.

Compared to the work of Dobson and Doig [17] (Qiu and colleagues' replication) the accuracy of our hybrid kernels is a slight improvement to their early integration model (AUC

of about 0.72).

According to the cluster analysis, kernels that exploit structural information (graph and contact kernel) show greater homogeneity than the sequence kernel (profile local alignment kernel) for all label spaces (see Table II). However, the highest homogeneity over all label spaces is achieved by the hybrid kernel (Graph+PLA) that combines sequence and structure information.

VI. CONCLUSIONS

Enzymes challenge current protein feature prediction methods. There are several examples where an identical catalytic mechanism has evolved from genes with no sequence homology (e.g. serine proteases). We set out to explore the complementarities of sequence and structural homology for the purpose of accurately classifying enzymes. By evaluating a range of standard sequence and structure kernels, and comparing their prediction performance to models that integrate sequence and structure information, we establish that combining complementary information leads to higher prediction accuracy.

The integration models outperform all homogenous kernels when used in the context of a SVM trained and tested to distinguish between enzymes and non-enzymatic proteins. An average AUC of 0.73 is better than a SVM fitted with alternative kernels ($p < 0.01$).

We note that the models that include the graph kernel organize their feature space differently compared to the other models. Clusters of proteins when viewed in the feature space of the intermediate integration model (Graph+PLA) or the graph kernel appear to share biologically meaningful features to a greater degree than when viewed using a sequence-based kernel. With the intermediate integration model, it should be easier for a kernel method to distinguish between sequence specific properties of proteins, structure specific properties of proteins, and combinations thereof than using any of the other kernels that were studied.

We conclude that the combination of features increases the prediction accuracy for a domain that challenges homology-based inference. The presentation of combined features in terms of kernels allows a support-vector machine to balance the impact of sequence and structural information. Our analysis highlights the complementarities of sequence and structure captured by a combined kernel function to detect and leverage weak homology.

REFERENCES

- [1] A. E. Todd, C. A. Orengo, and J. M. Thornton, "Sequence and Structural Differences between Enzyme and Nonenzyme Homologs," *Structure*, vol. 10, pp. 1435-1451, 2002.
- [2] D. Devos and A. Valencia, "Practical limits of function prediction," *Proteins: Structure, Function, and Genetics*, vol. 41, pp. 98-107, 2000.
- [3] W. Tian and J. Skolnick, "How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity?," *Journal of Molecular Biology*, vol. 333, pp. 863-882, 2003.
- [4] A. E. Todd, C. A. Orengo, and J. M. Thornton, "Evolution of function in protein superfamilies, from a structural perspective," *Journal of Molecular Biology*, vol. 307, pp. 1113-1143, 2001.
- [5] C. A. Wilson, J. Kreychman, and M. Gerstein, "Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores," *Journal of Molecular Biology*, vol. 297, pp. 233-249, 2000.
- [6] B. Rost, "Twilight zone of protein sequence alignments," *Protein Eng.*, vol. 12, pp. 85-94, February 1, 1999 1999.
- [7] L. Holm and C. Sander, "Protein Structure Comparison by Alignment of Distance Matrices," *Journal of Molecular Biology*, vol. 233, pp. 123-138, 1993.
- [8] J. Qiu, M. Hue, A. Ben-Hur, J.-P. Vert, and W. S. Noble, "A structural alignment kernel for protein structures," *Bioinformatics*, vol. 23, pp. 1090-1098, May 1, 2007 2007.
- [9] K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. N. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, pp. 147-56, 2005.
- [10] M. Jambon, A. Imbert, G. Delage, and C. Geourjon, "A new bioinformatic approach to detect common 3D sites in protein structures," *Proteins: Structure, Function, and Genetics*, vol. 52, pp. 137-145, 2003.
- [11] P. P. Wangikar, A. V. Tendulkar, S. Ramya, D. N. Mali, and S. Sarawagi, "Functional Sites in Protein Families Uncovered via an Objective and Automated Graph Theoretic Approach," *Journal of Molecular Biology*, vol. 326, pp. 955-978, 2003.
- [12] A. Stark, S. Sunyaev, and R. B. Russell, "A Model for Statistical Significance of Local Similarities in Structure," *Journal of Molecular Biology*, vol. 326, pp. 1307-1316, 2003.
- [13] E. V. Koonin and M. Y. Galperin, "Evolutionary Concept in Genetics and Genomics," in *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics*: Kluwer Academic Publishers, 2002.
- [14] J. M. Thornton, A. E. Todd, D. Milburn, N. Borkakoti, and C. A. Orengo, "From structure to function: Approaches and limitations," *Nat Struct Mol Biol*, 2003.
- [15] A. E. Todd, C. A. Orengo, and J. M. Thornton, "Evolution of protein function, from a structural perspective," *Current Opinion in Chemical Biology*, vol. 3, pp. 548-556, 1999.
- [16] A. Armon, D. Graur, and N. Ben-Tal, "ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information," *Journal of Molecular Biology*, vol. 307, pp. 447-463, 2001.
- [17] P. D. Dobson and A. J. Doig, "Distinguishing enzyme structures from non-enzymes without alignments," *Journal of Molecular Biology*, vol. 330, pp. 771-783, 2003.
- [18] P. Pavlidis, J. Weston, J. Cai, and W. S. Noble, "Learning Gene Functional Classifications from Multiple Data Types," *Journal of Computational Biology*, vol. 9, pp. 401-411, 04/91 2002.
- [19] S. D. Inderjit, G. Yuqiang, and K. Brian, "Kernel k-means: spectral clustering and normalized cuts," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining Seattle, WA, USA*: ACM, 2004.
- [20] C. P. John, "Fast training of support vector machines using sequential minimal optimization," in *Advances in kernel methods: support vector learning*: MIT Press, 1999, pp. 185-208.
- [21] C. Leslie, R. Kuang, and E. Eskin, "Inexact Matching String Kernels for Protein Classification," in *Kernel Methods in Computational Biology*, B. Scholkopf, K. Tsuda, and J.-P. Vert, Eds.: The MIT Press, 2004, pp. 95-112.
- [22] J.-P. Vert, H. Saigo, and T. Akutsu, "Local Alignment kernels for Biological Sequences," in *Kernel Methods in Computational Biology*, B. Scholkopf, K. Tsuda, and J.-P. Vert, Eds.: The MIT Press, 2004, pp. 131-154.
- [23] H. Rangwala and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition," *Bioinformatics*, vol. 21, pp. 4239-4247, December 1, 2005 2005.
- [24] R. W. W. Hooft, C. Sander, M. Scharf, and G. Vriend, "The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value," *Comput. Appl. Biosci.*, vol. 12, pp. 525-529, December 1, 1996 1996.
- [25] H. Cid, M. Bunster, M. Canales, and F. Gazitua, "Hydrophobicity and structural classes in proteins," *Protein Eng.*, vol. 5, pp. 373-375, July 1, 1992 1992.
- [26] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, 2000.
- [27] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden Markov

models that represent all proteins of known structure," *J Mol Biol*, vol. 313, pp. 903-919, 2001.