

# PARTIAL PROFILE ALIGNMENT KERNELS FOR PROTEIN CLASSIFICATION

*Thanh Ngo and Rui Kuang*

Department of Computer Science and Engineering, University of Minnesota, Twin Cities,  
200 Union Street SE, Minneapolis, MN 55455, USA,  
thango@cs.umn.edu, kuang@cs.umn.edu (Correspondence)

## ABSTRACT

Remote homology detection and fold recognition are the central problems in protein classification. In real applications, kernel algorithms that are both accurate and efficient are required for classification of large databases. We explore a class of partial profile alignment kernels to be used with support vector machines (SVMs) for remote homology detection and fold recognition. While existing profile-based kernels use the whole profiles to determine the similarity between pairs of proteins, the partial profile alignment kernels are derived from part of the position specific scoring matrices (PSSMs) in the profiles for alignment. Specifically, at each position in the PSSM, only amino acids in the mutation neighborhood of the corresponding amino acid in the original protein sequence are considered for alignment to remove noise and improve computing efficiency. Our experiments on SCOP bench datasets show that the partial profile alignment kernels achieved overall better classification results for both fold recognition and remote homology detection than profile kernels and profile-alignment kernels. In addition, our algorithm using only a fraction of the profiles saves the cost of computing the kernels significantly, compared to the full-profile alignment methods.

## 1. INTRODUCTION

In the post-genomic era, one important task is to annotate new genes/proteins encoded by the genome of newly sequenced species. The most widely applied large-scale approach is to classify the proteins into their corresponding protein families, superfamilies or folds defined by a taxonomy of protein structural classification. In protein classification, homologous proteins within the same family can be easily detected with less ambiguity using sequence alignment. However, the problem of detecting subtle sequence similarity between proteins sharing remote evolutionary relation or only similar folding patterns, is much more challenging. Such problems are called remote homology detection and fold recognition respectively.

Discriminative classification approaches using SVMs have shown superior performance for remote homology detec-

tion and fold recognition [1, 2, 3, 4, 5, 6, 7]. At the heart of SVM-based methods is the kernel, which is designed to capture the subtle similarity between protein sequences. For example, [1] derived a kernel from Fisher scores of HMM models; and, in [2] proteins are represented by pairwise sequence alignment scores against a protein database, while [3, 4, 6] built kernels from representations of proteins based on  $k$ -mers, short length- $k$  subsequences of amino acids. Saigo et al. [5] proposed a convolution kernel to summarize the alignment scores of all possible local sequence alignments between two protein sequences. Rangwala and Karypis [7] further explored the approach using profile-profile local alignment as a similarity function and then the kernel matrix is made positive semi-definite by adding a small constant on the diagonal.

## 2. PROFILE-BASED ALIGNMENT KERNELS

The PSI-BLAST profiles [8] of a protein sequence  $X$  of length  $n$  are a  $n \times 20$  matrix, either in form of a position-specific scoring matrix (PSSM) or a position-specific frequency matrix (PSFM). The columns of PSSM and PSFM are indexed by the alphabet of amino acids and each row corresponds to a position in the protein sequence. PSSM and PSFM contain the substitution scores and the frequencies, respectively, of the amino acids at different positions in the protein sequence. Each row of PSFM is normalized to sum to 1. The PSI-BLAST profiles of a sequence  $X$  are computed by aligning  $X$  with multiple sequences in the database that have statistically significant sequence similarities with  $X$ . Therefore, it contains more general evolutionary and structural information of the protein family that protein  $X$  belongs to, and thus, provides valuable information for remote homology detection and fold recognition.

Rangwala and Karypis [7] used the profile-profile local alignment derived from the scoring scheme proposed by [9] to achieve improved classification results. Specifically, the similarity score between the  $i$ -th position of protein  $X$ 's profile and  $j$ -th position of protein  $Y$ 's profile is

$$S_{X,Y}(i,j) = \sum_{k=1}^{20} PSFM_X(i,k) \times PSSM_Y(j,k) + \sum_{k=1}^{20} PSFM_Y(j,k) \times PSSM_X(i,k).$$

This similarity scoring scheme exploits the profile information to capture the evolutionary sequence conservation between proteins that are remotely related for better remote homology detection. Kernels based on sequence alignment can hardly compete with profile-based methods since remote homologous protein sequences share too little similarity to generate a good alignment. Clearly, profile-profile alignment overcomes this problem by the fact that the profiles can be used to evaluate the evolutionary conservation between two sequence positions.

### 3. PARTIAL PROFILE ALIGNMENT KERNELS

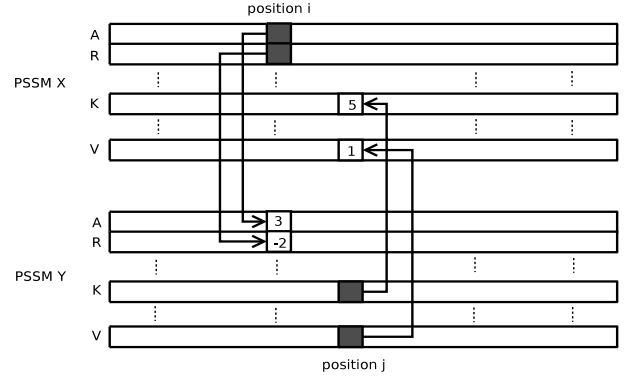
One observation of the profile-based alignment kernels is that most amino acids have a small frequency close to 0 or a negative substitution score at each position, which often represents noises that should not be used in the evaluation of alignments. Based on this observation, we assume that keeping the amino acids in a certain mutation neighborhood of the original one in the protein sequence are sufficient to represent the evolutionary information of the class of proteins that it belongs to. Specifically, we propose partial-profile alignment kernels in which, at each position in the PSSMs, only a mutation neighborhood of the amino acid at that position in the protein sequence is used to compute the similarity between proteins. We consider two different scoring schemes for our partial-profile alignment algorithms: unweighted and weighted.

In the unweighted case, the similarity score between the  $i$ -th position of protein  $X$ 's profiles and the  $j$ -th position of protein  $Y$ 's profiles is given by

$$S_{X,Y}^{(k)}(i,j) = \sum_{l \in N(X,i)} PSSM_Y(j,l) + \sum_{l \in N(Y,j)} PSSM_X(i,l)$$

where  $N(X,i)$  and  $N(Y,j)$  are mutation neighborhoods of the amino acids at position  $i$  in  $X$  and at position  $j$  in  $Y$ , respectively;  $PSSM_X(i,l)$  and  $PSSM_Y(j,l)$  are the values corresponding to the  $l$ -th amino acids at the  $i$ -th position of PSSM of  $X$  and  $j$ -th position of PSSM of  $Y$ , respectively. In the weighted case, the scoring function is

$$Sw_{X,Y}^{(k)}(i,j) = \sum_{l \in N(X,i)} PSFM_X(i,l) \cdot PSSM_Y(j,l) + \sum_{l \in N(Y,j)} PSFM_Y(j,l) \cdot PSSM_X(i,l).$$



**Fig. 1.** Top-2 partial profile alignment (PSSM matrices are transposed for viewable reason)

The weighted score function above is the mutation neighborhood truncated version of the score function used in SW-PSSM. The mutation neighborhood at each position are defined in three different ways, (i) the original amino acid in the protein sequence; (ii) the top-k scored amino acids in the PSSM; or (iii) the amino acids with scores in the PSSM higher than some threshold. Figure 1 shows an example of how the similarity score between position  $i$  of profile  $X$  and position  $j$  of profile  $Y$  is computed with *top-2* neighborhood. The top-2 scored amino acids at position  $i$  in  $PSSM_X$  are A and R, while the top-2 scored amino acids in  $PSSM_Y$  at position  $j$  are K and V. The top-2 unweighted similarity score between the two positions is  $3 + (-2) + 5 + 1 = 7$ . For weighted alignment,  $PSFM_X(i,A)$ ,  $PSFM_X(i,R)$ ,  $PSFM_Y(j,K)$  and  $PSFM_Y(j,V)$  are used to weight corresponding PSSM values. When the mutation neighborhood is the amino acid in the protein sequence, the alignment is a combination of two sequence-to-profile alignments.

The similarity matrix between protein profiles is symmetric but not necessarily positive semi-definite. Therefore it might not be a valid kernel according to Mercer's conditions. We employed the technique described in [5] and [7] to subtract from the diagonal of the similarity matrix its smallest negative eigenvalue. The resulting matrix is a valid kernel and differs from the original matrix only on the diagonal.

## 4. EXPERIMENTS

### 4.1. Dataset and evaluation

Partial-profile alignment kernels are tested on a benchmark dataset of 7329 domains from SCOP 1.59, in which no pair of sequences share more than 95% identity. Sequence profiles are obtained by sequence alignment using PSI-BLAST against nr database with e-value of  $10^{-3}$ . In case of remote

**Table 1. Remote homology detection results.**

Kernels	ROC	ROC-50	mRFP
SPK(3.0, 1.0)	0.985	0.864	<b>0.0037</b>
Top1-U-SW(3.0, 1.0)	0.979	0.873	0.0232
Top2-U-SW(3.0, 0.5)	0.985	0.863	0.0227
Top3-U-SW(3.0, 0.75)	0.985	0.851	0.0222
Top4-U-SW(3.0, 0.75)	0.966	0.779	0.0288
Top1-W-SW(3.0, 0.75)	0.978	0.871	0.0232
Top2-W-SW(3.0, 1.0)	0.983	<b>0.889</b>	0.0227
Top2-U-Conv(3.0, 1.0)	0.976	0.831	0.0268
Top2-W-Conv(3.0, 1.0)	0.980	0.875	0.0241
Thres1-U-SW(3.0, 1.0)	<b>0.987</b>	0.871	0.0241
Thres1-U-Conv(3.0, 0.75)	0.985	0.880	0.0244
profile-kernel(4, 4)	0.880	0.595	0.0793
profile-kernel(4, 6.0)	0.974	0.837	0.0288
profile-kernel(4, 8.0)	0.979	0.827	0.0253
profile-kernel(5.0, 7.5)	0.984	0.874	0.0230
profile-kernel(5, 10.0)	0.981	0.852	0.0244
profile-kernel(6, 9.0)	<b>0.987</b>	0.866	0.0228
SW-PSSM(3.0, 0.125, 0.0)	0.969	0.784	0.0279
SW-PSSM(3.0, 0.25, 0.0)	0.972	0.810	0.0259
SW-PSSM(3.0, 0.75, 1.5)	0.980	0.872	0.0041
SW-PSSM(3.0, 0.75, 2.0)	0.980	0.874	0.0225
SW-PSSM(3.0, 1.0, 2.0)	0.980	0.876	0.0040

SPK: Sequence-Profile Alignment Kernel; U:Unweighted; W:Weighted; Thres1:Score threshold = 1;

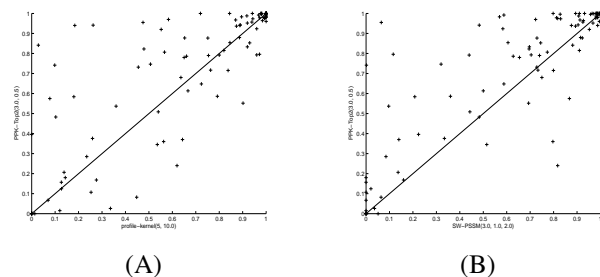
SW: Smith-Waterman; Conv: Convolution Kernel

homology detection, for each super-family, a SVM classifier is trained to detect proteins belongs to the same superfamily but not the same family. The same setting as in [6] is applied to 54 target families. A similar setting is applied for fold recognition. We constructed the fold recognition data by holding out a positive test superfamily with more than 5 sequences and the corresponding positive training set (sequences in other superfamilies under the same fold) with at least 10 sequences which leaves us 102 superfamilies.

We tested our kernels against two other best performing profile-based kernels: profile kernels [6] and SW-PSSM [7]. SVM training and classification are done with SPIDER 1.71 with SVM<sup>light</sup> engine. To evaluate classification results, we use receiver operating characteristic (ROC) score, ROC-50 score, which is the ROC score computed upto the first 50 false positives [10], and median rate of false positives (mRFP) [1]. ROC-50 is considered to be more important than ROC, since in practice, only some top results are viable to users. We also implemented SW-PSSM within our alignment framework based on available source code on the authors' website to test the running time of our kernels against SW-PSSM kernels.

## 4.2. Results

Partial profile kernels are defined by the type of mutation neighborhood, the scoring scheme (weight/unweighted) and



**Fig. 2.** Fold recognition results. (A) Top2-U-SW(3.0,0.5) v.s. profile-kernel(5,10.0). (B) Top2-U-SW(3.0,0.5) v.s. SW-PSSM(3.0,10.0,2.0)

the alignment algorithm. We tested both Smith-Waterman (SW) local alignment algorithm and the convolution kernel algorithm proposed in [5] with a scaling factor  $\beta=0.5$ . Our naming convention of partial profile kernels is *neighborhood type-scoring scheme-alignment algorithm*( $go, ge$ ). For example, Top2-W-SW(3.0, 0.75) means top-2 neighborhood, weighted scoring and SW algorithm with  $go = 3.0$  and  $ge = 0.75$ ; Thres1-U-Conv(0.5, 3.0, 0.75) means mutation neighborhood with a score threshold 1, unweighted scoring and convolution kernel with  $\beta = 0.5$ ,  $go = 3.0$  and  $ge = 0.75$ . When the mutation neighborhood is the amino acid in the protein sequence, we only use unweighted scoring scheme and SW algorithm, the kernels are denoted as SPK, as the alignment is a combination of two sequence-profile alignments. The gap opening cost ( $go$ ) used in all partial profile alignments is 3.0. Gap extension cost ( $ge$ ) values tested are 0.5, 0.75 and 1.

Table 1 shows the average ROC scores and ROC-50 scores over the 54 families achieved by the kernels for remote homology detection. In this set of experiments, partial profile kernels, profile kernels and SW-PSSM kernels with the best choice of parameters give similar classification performance with slight difference. Top2-W-SW kernels achieve the best ROC-50 scores; profile-kernel(6,9.0) gives the best ROC score, but Top2-W-SW and Top2-U-SW give similar results. It is interesting that SPK kernels and Top-1 kernels, while running significantly faster, achieve comparable results against SW-PSSM.

Table 2 lists average ROC and ROC30 scores for fold recognition over 102 superfamilies. Top2-U-SW kernels show significant improvement on ROC-50 with 5.4% improvement over the best of profile-kernels and about 8% better than the best of SW-PSSM kernels. Moreover, the improvement is consistent over different Top- $k$  kernels. This result suggests that top-ranked amino acids in the profiles are most informative for fold recognition.

The scatter plots in figure 2 visualize experiment-wise comparisons between Top2-U-SW and the best of profile-kernels and SW-PSSMs respectively. Top2-U-SW beats the

**Table 2. Fold recognition results.**

Kernels	ROC	ROC-50	mRFP
SPK(3.0, 1.0)	0.945	0.648	0.0339
Top1-U-SW(3.0, 1.0)	0.958	0.677	0.0257
Top2-U-SW(3.0, 0.5)	0.961	<b>0.717</b>	0.0238
Top2-U-SW(3.0, 1.0)	0.960	0.712	<b>0.0222</b>
Top3-U-SW(3.0, 1.0)	0.958	0.688	0.0238
Top4-U-SW(3.0, 0.75)	0.948	0.647	0.0317
Top1-W-SW(3.0, 0.75)	0.957	0.671	0.0252
Top2-W-SW(3.0, 0.75)	0.959	0.685	0.0279
Top2-U-Conv(3.0, 1.0)	0.950	0.648	0.0351
Top2-W-Conv(3.0, 1.0)	0.953	0.652	0.0311
Thres1-U-SW(3.0, 1.0)	0.949	0.650	0.0384
Thres1-U-Conv(3.0, 0.75)	0.950	0.628	0.0348
profile-kernel(4, 4)	0.885	0.333	0.0824
profile-kernel(4, 6.0)	0.935	0.591	0.0375
profile-kernel(4, 8.0)	0.934	0.626	0.0366
profile-kernel(5.0, 7.5)	0.959	0.614	0.0284
profile-kernel(5, 10.0)	0.953	0.664	0.0230
profile-kernel(6, 9.0)	<b>0.964</b>	0.592	0.0275
SW-PSSM(3.0, 0.125, 0.0)	0.920	0.516	0.0614
SW-PSSM(3.0, 0.25, 0.0)	0.923	0.535	0.0585
SW-PSSM(3.0, 0.75, 1.5)	0.943	0.635	0.0431
SW-PSSM(3.0, 0.75, 2.0)	0.944	0.628	0.0421
SW-PSSM(3.0, 1.0, 2.0)	0.945	0.636	0.0400

SPK: Sequence-Profile Alignment Kernel; U:Unweighted; W:Weighted; Thres1:Score threshold = 1;

SW: Smith-Waterman; Conv: Convolution Kernel

other two in majority of the experiments and give comparable results for most of the remaining.

We also measured and compared the running time of partial profile alignment using SW algorithm to the full-profile alignment used in SW-PSSM. Table 3 shows the average running time of the methods. The running time is measure by actual CPU usage using Unix’s ps command. At each step of Smith-Waterman algorithm, to compute the similarity between the two positions, sequence-profile alignment is roughly 20 times faster than full-profile alignment. However, due to the overhead of traversing through the dynamic programming table and other comparisons, the overall improvement is reduced to about 5.5 times faster but is still significant.

**Table 3. Average running time**

Kernels	Running time (minutes)
SPK	448
Top1-U-SW	834
Top2-U-SW	972
Top3-U-SW	1118
SW-PSSM	2552

## 5. CONCLUSION AND FUTURE WORK

We propose partial profile alignments for computing similarity scores between proteins and derived kernels for SVM-based remote homology detection and fold recognition. At each position. Only a mutation neighborhood of the amino acid in the protein sequence is considered for alignment. This helps remove noises in PSI-BLAST profiles as well as improve computing efficiency. The proposed kernels have gained improvements especially in fold recognition. In the future, we will use the proposed kernels to attack the harder problem, multi-class protein classification.

## 6. REFERENCES

- [1] T. Jaakkola, M. Diekhans, and D. Haussler, “A discriminative framework for detecting remote protein homologies,” *J. Comput. Biol.*, vol. 7, pp. 95 – 114, 2000.
- [2] L. Liao and W. S. Noble, “Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships,” *J. Comput. Biol.*, vol. 10, pp. 857 – 868, 2003.
- [3] C. Leslie, E. Eskin, and W. S. Noble, “The spectrum kernel: a string kernel for svm protein classification,” in *Pac. Symp. Biocomput.*, 2002, pp. 566 – 575.
- [4] C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, “Mismatch string kernels for svm protein classification,” in *Adv. Neural Inf. Process. Syst.*, 2003, pp. 1441–1448.
- [5] H. Saigo, J. P. Vert, N. Ueda, and T. Akutsu, “Protein homology detection using string alignment kernels,” *Bioinformatics*, vol. 20, pp. 1682 – 1689, Feb. 2004.
- [6] R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie, “Profile-based string kernels for remote homology detection and motif extraction,” *J. Comput. Biol.*, pp. 527 – 550, Jun. 2005.
- [7] H. Rangwala and G. Karypis, “Prole-based direct kernels for remote homology detection and fold recognition,” *Bioinformatics*, vol. 21, pp. 4239 – 4247, Sep. 2005.
- [8] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped blast and psi-blast: a new generation of protein database search programs,” *Nucleic Acids Res.*, pp. 3389 – 3402, 1997.
- [9] D. Mittelman, R. Sadreyev, and N. Grishin, “Probabilistic scoring measures for proleprole comparison yield more accurate short seed alignments,” *Bioinformatics*, vol. 19, pp. 1531 – 1539, Feb. 2003.
- [10] M. Gribskov and N. L. Robinson, “Use of receiver operating characteristic (roc) analysis to evaluate sequence matching,” *Comput. Chem.*, vol. 20, pp. 25 – 33, 1996.