Spectral Clustering on Neighborhood Kernels with Modified Symmetry for Remote Homology Detection

Anasua Sarkar, Macha Nikolski, Ujjwal Maulik

Spectral clustering on neighborhood kernels with modified symmetry for remote homology detection

Anasua Sarkar, Macha Nikolski
LaBRI
Laboratoire Bordelais de Recherche en Informatique, UMR CNRS 5800
France
Email: anasua.sarkar@labri.fr, macha.nikolski@labri.fr

Ujjawl Maulik
Computer Science and Enginnering
Jadavpur University
Kolkata, India
Email: umaulik@cse.jdvu.ac.in
German Cancer Research Center
Heidelberg, Germany

***Abstract*—Remote homology detction among proteins in an unsupervised approach from sequences is an important problem in computational biology. The existing neighborhood cluster kernel methods and Markov clustering algorithms are most efficient for homolog detection. Yet they deviate from random walks with inflation or similarity depending on hard thresholds. Our spectral clustering approach with new combined local alignment kernels more effectively exploits state-of-the-art neighborhood vectors globally. This appoarch combined with Markov clustering similarity after modified symmetry based corrections outperforms other six cluster kernels for unsupervised remote homolog detection even in multi-domain and promiscuous proteins from Genolevures database with better biological relevance. Source code available upon request.**

***Keywords*-Spectral clustering,; kernel matrix; modified symmetry distance measure; Remote homology detection;**

I. Introduction

The remote homology detection with very subtle sequence similarity from unlabeled protein sequences is one of the challenging problems in computational biology. Historically, besides the pairwise sequence similarity aproaches like BLAST or PSI-BLAST [1], the discriminative kernel methods with SVMs (Support Vector Machines) [2] or MAMMOTH score based structure kernel [3] proved to detect remote homologs with extensive training data.

The most successful methods for remote homology detection based on MCL algorithms are OrthoMCL [4] and TribeMCL [5], which utilizes random walks on the Markov transition matrix to analyze the emergence of clusters in the graph that encodes this matrix. These methods attempt to solve multi-domain and promiscuous domain problems, although they bias the random walks with 'inflation' parameter to promote the cluster emergence from graph. Earlier non-kernel aproach of Paccanaro et. al [6] utilized spectral clustering on protein sequences, while Weston et. al [2] introduced the neighborhood vector in cluster kernels for semi-supervised protein clustering. Symmetry is an inherent feature that enhances recognition and reconstruction of shapes and objects even in kernel space. Chou et. al [7] proposed a symmetry based distance measure d_s, which fails to detect clusters, that have inherent symmetry relative to some intermediate point. Therefore, they corrected the distance norm in [8] to handle overlapping symmetrical clusters.

In this work, at first we develop three positive semi-definitive local alignment kernels based on the singular value decompositions of similarities explicitly in local alignment methods such as BLAST, PSI-BLAST and MCL clustering. We enhance the Markov cluster similarity kernel further with neighborhood feature vectors to reduce the diagonal dominance issue problem. To reduce the problem of promiscuous domains further, we incorporate the spectral clustering approach over kernel matrices implicitly selecting the leading eigenvectors from 'global' distances. Finally, we incorporate the modified symmetry based correction in Hilbert space reducing number of singletons and classifying multi-domain proteins into more biologically significant clusters with closest nearest neighbor.

We experiment all our seven kernels over the multi-domain proteins from Genolevures yeast database [9]. When comparing the performance of our modified symmetry corrected combined spectral kernels, the experiments demonstrate the superiority of introducing modified symmetry corrections with combined neighborhood spectral kernels to detect remote homologs more accurately even for multi-domains and promiscuous domain proteins.

II. Background

In this section, we briefly describe existing state-of-the-art cluster kernels for protein classification and the modified symmetry based distance measure for clustering.

A. Spectral clustering

In semisupervised learning, the spectral clustering kernel boils down to be the spectral graph partitioning into the sub-space of the k largest eigenvectors of a normalized affinity/kernel matrix [10]. Let us assume an undirected graph $G = (V, E)$ with vertices $v_i \in V$, for $i = 1, \cdots, n$ and edges $e_{i,j} \in E$ with non-negative weights $s_{i,j}$ expressing the

similarity between vertices v_i and v_j. Then the eigenvectors $(v_1, \cdots, v_k)$ are computed as $D^{-1/2}KD^{-1/2}$, where D is a diagonal matrix computed as $D_{ii} = \sum_n K_{in}$, where K is the RBF kernel on the graph. The spectral clustering approach produced over protein sequences in [6] simultaneously analyzed k eigenvectors.

B. Neighborhood mismatch kernel

In [2] Weston et al defined a neighborhood kernel over the feature representation $\phi_{nbd}(x) = \frac{1}{|Nbd(x)|}\sum_{x' \in Nbd(x)} \phi_{orig(x')}$ as:

$$K_{nbd}(x,y) = \frac{1}{|Nbd(x)||Nbd(y)|} \sum_{x' \in Nbd(x), y' \in Nbd(y)} K_{orig}(x', y') \tag{1}$$

,where $Nbd(x)$ denotes a neigborhood average vector for sequence x over a set of sequences x' with E-value less than a fixed threshold in PSI-BLAST/BLASTP search. The use of this kernel boosts up the classification performance of protein sequences.

C. Modified symmetry based distance measure

As existing distance measures such as Euclidean or Pearson correlation can detect symmetry in clusters, Su and Chou [7] proposed a symmetry based distance norm d_s between a pattern x and a reference centroid c as follows: $d_s(x,c) = \frac{(d_1)}{d_e(x,c)+d_e(x_1,c)}$, where x_1 is the symmetrical point of x with respect to c and $d_e(x,c)$ and $d_e(x_1,c)$ are Euclidean distances between c and respectively between x and x_1. If x' represents first nearest neighbor of x and is computed as $x' = (2 * c - x)$, then d_1 is the Euclidean distance of x_1 and x'. To improve this symmetry based distance norm even for inter-symmetrical clusters, Chou et. al [8] proposed a modified measure d_c as defined below:

$$d_c(x,c) = d_s(x,c)d_e(x,c) \tag{2}$$

III. Methods

A. Data

The Genolevures Release 3 candidate 3 database (2008-09-24) [11] explores nine complete genomes from the class of Hemiascomycete yeasts. We use 323 sequences as unlabeled data from 23 Multiple choice families $GL3M.*$ which have a very variable composition dependent on statistical parameters and are complicated families such as polyproteins and repeat domains. We use Genolevures [11] familiy structure as the true clusters for ROC analysis.

B. Local alignment-based kernels

The non-symmetric probablistic profiles generated by PSI-BLAST are recently used in kernels instead of sequence encoding in [6]. The maximal or average interpretation of logarithmic E-values produces a symmetric kernel solving this problem [5]. However utilizing the HSP(high-scoring segment pair) score of BLASTP results directly resembles the functionality of mismatch string kernel [12] to some extent. In stead of E-values, we utilize the BLASTP HSP (high-scoring segment pair) score within the threshold in kernel formation(**I**)which also satisfies the biological relavance of searching homologs.

1) Position specific scoring kernel: We treat the position-specific PSI-BLAST [1] score directly for kernel formation, as it represents the similarity of homologos in descending order more accurately than BLASTP [13]. The PSI-BLAST similarity matrix P is not positive semidefinitive, as all-vs-all PSI-BLAST scores are not symmetric. However P is symetric with singular value decomposition $P = U^T DV$,where D is the diagonal matrix $diag(\lambda_1, \cdots, \lambda_n)$ with singular value entries $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. Therefore we define the PSI-BLAST kernel(**II**) by $K = U^T \psi(D) V$ where $\psi(D) = diag(\psi(\lambda_1), \cdots, \psi(\lambda_n))$ and $\psi(\lambda) = 1+\lambda$ if $\lambda > 0$, and 0 otherwise. We normalize the kernel with unit sphere projection via, $K_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$. A related protein structure kernel, based on MAMMOTH score [3] previously yielded good performane in classifying proteins.

2) Markov cluster similarity kernel: The Markov Cluster alogorithm simulates random walks on a graph to detect the transition probabilities among its edges using Markov matrices. TribeMCL [5] and OrthoMCL [4] apply the MCL algorithm to detect protein clusters which consists of multi-species orthologs or recent paralogs. The scoring matrix in OrthoMCL [4] algorithm is intially computed as the average $-log10(P-value)$ from pairwise WU-BLASTP similarities with an average edge weight normalization to minimize impact of recent paralogs in cross-species ortholog clusters. We generate a kernel from this score to solve the diagonal dominance issue for $K(x,x)$ to be orders of magnitudes larger than $K(x,y)$, by assigning arbitray values to $K(x,x)$.

Neighborhood similarity kernel: We incorporate the neighborhood feature vector of Eq. 1 of each sequence over the MCL similarity scores, following earlier neighborhood mismatch kernel [2]. To satisfy the positive semidefinitive property of our kernel(**III**), we compute the singular value decomposition of this matrix with a normalization to [0,1] interval.

C. Combined spectral kernel clustering

For unsupervised classification, we utilize the spectral clustering method [3] directly to the local alignment kernel matrices without using a transductive setting like in [2]. We combine all modified local alignment kernels using normal product(**IV,V**). However this random walk based graph partitioning method cuts the inter-cluster edges, which explicitly removes the promiscuous domains. This algorithm also constructs the Markov transition matrix following a Markovian relaxation process [3] to utilize the eigenvectors corresponding to the leading eigenvalues. As this method does not need to modify the random walks with a relaxation

parameter called 'inflation' in OrthoMCL [4] and TribeMCL [5], it outperforms those methods in the accuracy of the result clusters with respect to the true classifications.

D. Modified symmetry in kernel space

The modified-symmetry based distance norm d_c [8] in Eq. 2 considers the nearest neighbor of symmetrical points among clusters to compute distances. The distance of a point and its nearest neighbor in the Hilbert space produces significant higher values distinguishing outliers. Correcting clusters with lower modified symmetry norm (d_c) less than the pre-defined threshold $\theta = 0.18$ [7] value, imposes compact clusters reducing outliers over kernel space(**VI,VII**). With respect to the original "true" clusters, this yields good overlapping symmetrical clusters as discussed in Section V.

IV. Results

In this section, we describe the experimental framework and comparative results of all implemented spectral alignment kernels with modified symmetry based corrections.

A. Evaluation framework

To experiment all our seven kernels, we utilize several different frameworks to be implemented. The $PSI-BLAST$ [1] iterations with composition based statistics are performed on a Cluster with 62 Opteron nodes [2.60 GHz, 322.4 GFLOPs] using $MPIBlast$. We use OrthoMCL version 2.0 [4] for our experiments. All the kernels are generated in Matlab v7.10 (R2010a) 64-bit. We use the normalized spectrum kernel with sub-sequence/string length = 4 settings in the Kernel-based Machine Learning Lab $(kernlab)$ package [14] in R from CRAN for spectral clustering. For the ROC analysis of the kernel matrices, we use the $ROCR$ packages [15]. We implement the modified symmetry based clustering approach using MPICH.

B. Performance of local alignment-based specral kernels

Table I summarizes the performance achieved by BLASTP, PSI-BLAST and OrthoMCL Neighborhood Mismatch $(OMCL\ NM)$ kernels (**I,II,III**) for familiy-level classification of multi-domain proteins from our dataset with mean ROC and mean ROC50 scores. $OMCL\ NM$ kernel performs best over other simple kernel methods indicating the influence of neighborhood in homolog detection. As an illustartion, the number of families whose ROC50 scores are greater than a given threshold in the range [0,1] are shown in Figure 1. The $OMCL\ NM$ kernel retrieves approximately two times more ROC50 scores than the simple BLASTP and PSI-BLAST kernels for a similar number of families.

While investigating our combined spectral kernels, combined $PSI-BLAST\ OMCL\ NM$ kernel(**V**) provides ROC50 values 0.757, which is superior to the values 0.738 obtained by the combined $BLASTP\ OMCL\ NM$ kernel(**IV**). $PSI-BLAST$ with $OMCL\ NM$ kernel produces a highest ROC score of 0.945. In Figure 1, $OMCL\ NM$ kernel combined respectively with $PSI-BLAST$ and

Table I
ROC, ROC50 averaged over 23 families for different simple and combined local alignment based spectral kernels with modified symmetry based corrections

ID	Kernel	Mean ROC-50	Mean ROC
I	BLASTP kernel	0.481	0.836
II	PSI-BLAST kernel	0.495	0.939
III	OMCL NM kernel	0.741	0.949
IV	BLASTP + OMCL NM kernel	0.738	0.942
V	PSI-BLAST + OMCL NM kernel	0.757	0.945
VI	BLASTP + OMCL NM kernel + Modsym	0.742	0.946
VII	PSI-BLAST + OMCL NM kernel + Modsym	0.798	0.962

OMCL NM=OrthoMCL Neighborhood Mismatch kernel

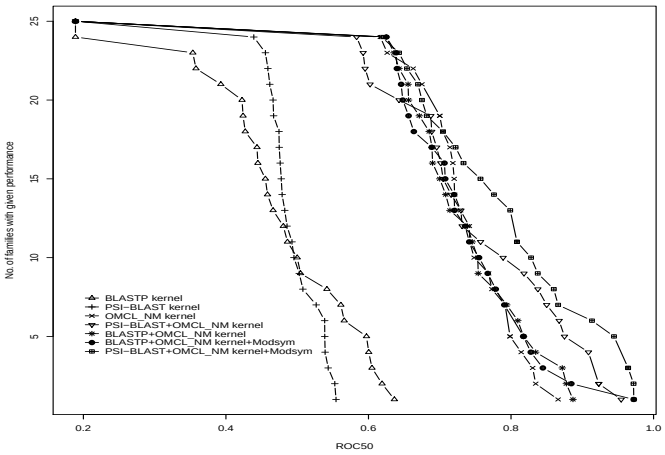

Figure 1. Comparison of ROC50 score distribution for all spectral local alignment kernels

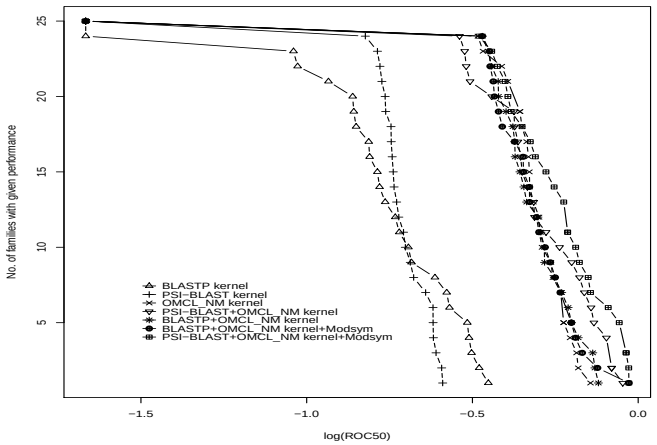

Figure 2. Comparison of log ROC50 score distribution for all spectral local alignment kernels

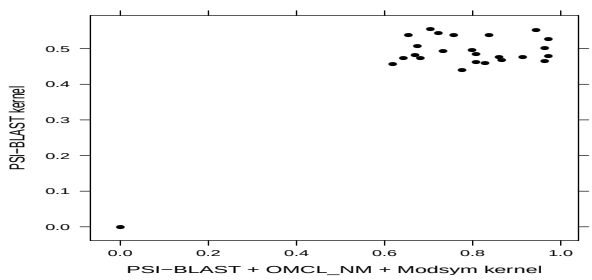

Figure 3. Family-by-family comparison of PSI-BLAST kernel and PSI-BLAST OMCL NM kernel after modified symmetry based enhancement.

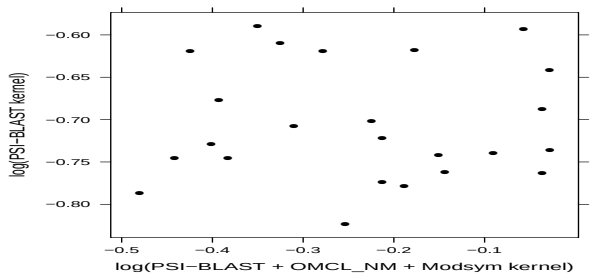

Figure 4. Family-by-family comparison of PSI-BLAST kernel and PSI-BLAST OMCL NM kernel (log transformed) after modified symmetry based enhancement.

$BLASTP$ kernels consistently shows compartive superior performance in ROC50 distributions with each other.

All combined spectral kernels show better performances after modified symmetry-based enhancement in Table I. The major impact of modified proximity norm d_c can be observed in ROC50 score of 0.798 for combined $PSI-BLAST\ OMCL-NM$ spectral kernel(**VII**). In Figure 1, our $PSI-BLAST$ kernel combined with $OMCL\ NM$ kernel after modified symmetry based redistribution, consistently outperforms other combined kernels. Figure 3 shows the familiy distribution for ROC50 scores between $PSI-BLAST$ kernel and its improvement after combination with our $OMCL\ NM$ kernel including modified symmetry based enhancements. For most of the families, the $PSI-BLAST\ OMCL\ NM$ kernel with modified symmetry provides higher ROC50 scores than simple $PSI-BLAST$ kernel. Therefore all experiments demonstrate the power of combined spectral kernels with modified symmetry corrections in the remote homolog detection on unlabeled data.

V. Discussion

We have experimentally evaluated seven spectral kernels for remote homology detection in proteins with the explicit evaluation of modified symmetry based proximity norm. In this unsupervised kernel setting, we show that our spectral clusetring approach with combined local alignment kernels perform competitively with respect to the state-of-the-art neighborhood [2] mismatch kernel method. Finally our experiments with introducing modified symmetry in kernel space outperform other cluster kernels in this framework.

Four major observations can be made by analyzing the performances achieved by different experiments we represent in this article. First, BLASTP and PSI-BLAST local-alignment kernels with singular value decomposition prove to be a valid kernel in terms of kernel definition for homology detection. Second, incorporation of OrthoMCL scores to reduce the "recent" paralog effects gains significance in creating Markov cluster similarity kernel. The neighborhod kernel over OrthoMCL scores also proves to be significant, reducing the diagonal dominance issue with arbitrary lower magnitude distribution in diagonals.

Third, without making diagonal matrix of all labeled and unlabebled data [2], in our spectral kernel methods we compute global distances from all-vs-all local alignment scores without using any hard cut-off threshold. Implicit reduction of inter-cluster edges in computing leading eignevectors over kernel in spectral clustering also reduces promiscuous domain problem by restricting to a one-to-one allocation between proteins and families, which TribeMCL [5] did with 'inflation' parameter as a relaxation over random walks.

Four, the modified symmetry based reallocation in kernel space imposed to be biologically significant to exclude outliers as shown in Section III-D. Smaller distance with the nearest neighbor in intra-symmetrical clusters therefore signifies more compactness in kernel space. Therefore detecting modified symmetry among multi-domain proteins classifies a protein to a cluster selecting more accurate domain with closer homologs increasing biological significance.

The fact that both the widely used cluster kernels [2] and OrthoMCL [4] are able to detect homologous clusters even in multi-domain protein families reinforces the validity of our approaches. In the context of remote homolog detection our approach produces more statistically significant protein clusters with biological relevance of modified symmetry correction.

VI. Conclusion

The homologous protein family detection tools indicating conservation of function are an important tool in comparative genomics. Errors in protein assignemnt to families is frequent especially as homology becomes distant. Therefore, we propose a systematic approach for computing local alignment based combined spectral kernels from unlabeled protein sequences for remote homlogy detection. We experiment the corrections by the modified symmetry based proximity norm producing improved clusters with reduced outliers/singletons and selecting more biologically significant domains for multi-domain proteins. Our position specific scoring kernel combinned with modified symmetry,

achieves state-of-the-art prediction performance on unsupervised homology detection. This approach outperforms other cluster kernels, combined with Markov cluster similarity kernels in well-known neighborhood feature space. Therefore to detect homologs among multi-domain proteins, our spectral clustering approach with combined local alignment kernels leads to achieve biological significance with modified symmetry based corrections in neighborhood kernel space.

REFERENCES

[1] S. F. Altschul, T. L. Madden, A. A. Schffer, R. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psiblast: a new generation of protein database search programs," *NUCLEIC ACIDS RES*, vol. 25, pp. 3389–3402, 1997.

[2] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble, "Semi-supervised protein classification using cluster kernels," *Bioinformatics*, vol. 21, no. 15, pp. 3241–3247, 2005.

[3] M. Hue, M. Riffle, J.-P. Vert, and W. S. Noble, "Large-scale prediction of protein-protein interactions from structures," *BMC Bioinformatics*, vol. 11, p. 144, 2010.

[4] L. Li, C. J. Stoeckert, and D. S. Roos, "Orthomcl: identification of ortholog groups for eukaryotic genomes." *Genome Res*, vol. 13, no. 9, pp. 2178–89, 2003. [Online]. Available: http://v2.orthomcl.org

[5] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucl. Acids Res.*, vol. 30, no. 7, pp. 1575–1584, 2002.

[6] A. Paccanaro, J. A. Casbon, and M. A. S. Saqi, "Spectral clustering of protein sequences," *Nucleic Acids Research*, vol. 34, no. 5, pp. 1571–1580, 2006.

[7] M. C. Su and C. H. Chou, "A modified version of the k-means algorithm with a distance based on cluster symmetry," *IEEE Trans Pattern Anal Mach Intell*, vol. 23, no. 6, pp. 674–680, 2001.

[8] C. H. Chou, M. C. Su, and E. Lai, "Symmetry as a new measure for cluster validity," in *Second WSEAS International Conference on Scientific Compuation and Soft Computing*, 2002, pp. 209–213.

[9] D. J. Sherman, T. Martin, M. Nikolski, C. Cayla, J.-L. Souciet, and P. Durrens, "Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes," *Nucleic Acids Research*, vol. 37, no. Database-Issue, pp. 550–554, 2009.

[10] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 849–856.

[11] M. Nikolski and D. J. Sherman, "Family relationships: should consensus reign? - consensus clustering for protein families," *Bioinformatics*, vol. 23, no. 2, pp. 71–76, 2007.

[12] C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, vol. 20, no. 4, pp. 467–476, 2004.

[13] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "A basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, October 1990.

[14] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab – an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004. [Online]. Available: http://www.jstatsoft.org/v11/i09/

[15] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "Rocr: visualizing classifier performance in r," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005. [Online]. Available: http://rocr.bioinf.mpi-sb.mpg.de/