Application of Geometric Kernel Data Fusion in Protein Fold Recognition and Protein Sub-nuclear Localization

Pooya Zakeri[1], Ben Jeuris[3] , Raf Vandebril[3] , Yves Moreau[1],[2]

[1]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, [2]iMinds Medical IT and [3]Department of Computer Science, KU Leuven, 3001 Leuven, Belgium

Abstract: Finding an efficient technique for integrating heterogeneous biological data sources has received growing attention. In particular, kernel methods are an interesting class of techniques for data fusion. We look into the possibility of using the geometric mean of matrices instead of the arithmetic mean for multiple kernel learning (MKL). While computing geometric means of matrices is challenging, it hints at an intriguing research direction in data fusion. Geometric kernel fusion (GKF) is used for protein fold recognition and sub-nuclear localizations. It offers a significant improvement over MKL approaches. The experimental results demonstrate that GKF can effectively improve the accuracy of the state-of-the-art kernel fusion model for protein fold recognition and predicting protein sub-nuclear prediction.

Moreover, the limitation of convex linear combinations in coping with integration of different protein features that carry complementary information is investigated. Our proposed fusion frameworks, by contrast, can be used to detect these features with complementary information, which provides an insightful contrast for combining different features of other problems in bioinformatics.

***Keywords*:** Genomic Data Fusion, Protein Fold Recognition, Predicting Protein Sub-nuclear localization, Geometric Kernel Fusion, Geometric Mean, Multiple Kernel Learning

GENERAL GUIDELINES

Various sequence-based protein features and often machine learning methods have been developed to solve Bioinformatics tasks. More attention needs to be paid to finding an efficient and cost-effective method for fusing these different discriminatory data sources for these problems.

This study presents kernel-based computational frameworks for fusing heterogeneous biological data sources by taking more involved geometry means of their corresponding kernel matrices instead of convex linear combinations. It has been observed that geometric data fusion is less sensitive in dealing with complementary and noisy kernel matrices compared to typical multiple kernel learning approaches [1]. Since biological kernels often encode the complementary characteristics of biological data, this motivates us to see the application of geometric data fusion in bioinformatics tasks.

We address two challenging problems in bioinformatics: (1) Protein fold recognition, and (2) Predicting protein sub-nuclear localization. Both problems are among the most essential objectives in cell biology, molecular biology and proteomics. In particular, protein fold and protein sub-nuclear information can help to understand better cellular process and molecular function in a cell.

We consider various sequence-based protein features including pseudo amino acid composition, predicted secondary structure, as well as information extracted directly from position-specific scoring matrices and local alignment kernels. GKF frameworks are employed to integrate these complementary biological data sources. Moreover, our computational model has been developed by incorporating the functional domain composition of proteins through a kernel-based hybridization model.

Our GKF frameworks offer a significant improvement for both protein fold recognition and protein sub-nuclear localization prediction. It is observed that using GKF, we achieve an overall protein fold accuracy of 86.86% on SCOP PDB-40D benchmark dataset [2] and 96.86 on newDD benchmark dataset [3], which represent 20.8% and 7.6% improvement over the state-of-the-art protein fold predictor respectively. Experimental results on the SNL9 benchmark dataset [4] also show an improvement of 4.1% over the state of the art in protein sub-nuclear localization.

INTRODUCTION

Early and late integration are typical approaches to integrate various protein data sources. In addition, the heterogeneous biological data sources can be integrated intelligently and efficiently using partial integration, such as kernel-based data fusion approaches. Using kernel methods is an interesting strategy because it decouples the original data from the machine-learning algorithms by using a representation of the data as a kernel matrix. Symmetric positive definite (SPD) kernel matrices are in fact the nonlinear extension of covariance/correlation matrices and encode the similarity between examples in their respective input space. This demonstrates that the heterogeneous data can all be replaced by appropriately scaled kernel matrices with the same size, and consequently the data heterogeneity vanishes. Then other classification, clustering and prioritization algorithms can access the same data, which is currently not

possible. Indeed, constructing the same representation for all data sets and integrating these representations systematically is the main insight behind kernel fusion methods.
There are several methods for obtaining a valid and fitting kernel by tuning the kernel matrices weights [5]. Finding such weights from training data and replacing the single kernel by a convex linear combination of weighted base kernels is usually referred to as MKL. These weights reflect the relative potential importance of the different data sources in the combined kernel. Several MKL approaches have been proposed in the literature [6-9], during the last decade. Several complicated convex optimization-based approaches have been proposed [6-9], to improve the efficiency of kernel data fusion. They try to optimize the kernel weights based on different optimization criteria. However, on the one hand, the optimized weights of base kernels convex combination of kernel matrices often leads to mixed results and causes an improvement in performance only when dealing with redundant or noisy kernel matrices [9]. Indeed, this type of averaging has usually sensitive behavior in coping with kernel matrices containing complementary and non-redundant information and fails to completely capture all the information for these kernels. Since genomic kernels often encode the complementary characteristics of biological data, applying convex linear combination of base kernels is not very appropriate for biological application.
On the other hand, it has been showed that even the results obtained by employing uniformly weighted kernel fusion are comparable to the results of the best existing MKL approaches in general applications. This is also supported by the equal weights theorem [10], which states when all optimized weights are uniformly distributed on the interval [0.25; 0.75], the performance is hardly changed using equal weights.
Using the Euclidean distance on a convex cone whose interior contains all SPD matrices $P(n)$, we can obtain the arithmetic mean which is uniformly weighted average of the base kernels. By contrast, since it has been illustrated that this type of averaging fails to completely capture all the information for kernels containing complementary, non-redundant information, Euclidean distance on SPD matrices might not be appropriate. Moreover, SPD matrices form a convex cone and not a vector space. This has an effect on the "natural" geometry of SPD matrices, which may not be Euclidean, but rather should rely on concepts from Riemannian geometry. This inspires us to consider other means between SPD matrices that are not relative to the Euclidean distance on $P(n)$ and necessarily a linear of SPD matrices [1]. For example, the mean corresponding to Riemannian distance on $P(n)$ is the geometric mean.

METHODS

GEOMETRIC KERNEL FUSION

In this study, we design, and develop several methods to integrate kernel matrices by taking more involved, geometry inspired means of these matrices instead of convex linear combinations [1]. Such averaging of the base kernels can be interpreted as a kind of integration that expresses the nonlinear relationship between the individual kernels. In particular, we focus on taking the matrix geometric mean of base kernels. For a general number of matrices, the fused kernel is obtained by taking the geometric mean

$$K = F(K_1, K_2, \ldots, Kn) = G(K_1, K_2, \ldots, Kn)$$

Similar to the arithmetic mean (AM), other types of means of SPD matrices such as the harmonic mean (HM), Log-Euclidean mean (LogEM) [11], and geometric mean(GM) result in SPD kernels. However, computing the geometric mean of a general number of SPD matrices is a challenge. In fact, for a general number of SPD matrices, a proper definition of a geometric mean with some natural properties has only recently been developed [12]. We present two methods for computing the geometric mean [1]. The first approach is focused on computing the actual geometric mean using the definition of the Karcher mean (Karcher-KF) [13]. The second, however, only computes a rough approximation of the actual geometric mean using a proposed, heuristic method based on Arithmetic-Geometric-Harmonic (AGH) mean (AFH-KF). We show in the second section that it is a computationally scalable method for computing an approximate geometric mean.
However, computing the geometric mean for a general number of SDP matrices is hard and computationally expensive, which is why we also discuss the Log-Euclidean mean (LogE) [11] . It can be considered as a consensus between the arithmetic and geometric mean. The Log-Euclidean mean of n SPD matrices can be obtained explicitly as [11]:

$$K_{LE}(K_1, \ldots, K_n) = \exp\left(\frac{1}{n}\left(\sum_{i=1}^{n} \log(K_i)\right)\right)$$

Protein Fold Recognition and Predicting Protein Sub-nuclear Localization

Protein fold recognition is a challenging problem in computational biology. Tertiary structural information of proteins can provide new knowledge on their function. In addition, understanding the three-dimensional structure of proteins can be facilitated through the knowledge of protein folds; hence, determining this structure is among the most essential

objectives in molecular biology, cell biology, and proteomics. Structural information is also potentially useful for drug design study.

Predicting protein sub-nuclear localization is another challenging problem in bioinformatics. Knowledge on functions of proteins can be provided by information of subcellular locations. Another area which has received little attention in the literature [4, 14, 15], is the prediction of protein localization in the organelles of the cell, such as nucleus, chloroplast and mitochondria. The nucleus is a membrane-enclosed organelle in eukaryotic cells, and contains the DNA organized into chromosome. It is the principal location of DNA and RNA synthesis. Information about the sub-nuclear location for a nuclear protein can also provide much better understanding about its function. Furthermore, using the reliable automatic sub-nuclear localizer the design of appropriate drugs could be sped up for many kind of complex diseases linked to human genome and cancers.

In the recent years, various sequence-based protein features and often machine learning methods have been proposed for protein fold recognition [1, 2, 3, 17], and predicting protein sub-nuclear localizations [4, 14]. In order to evaluate the efficiency of the GKF in bioinformatics tasks, we address these two problems.

Feature Vectors

For protein fold recognition, we consider 26 sequence-based protein features including 6 types of structural information (Amino Acid composition (C), Predicted Secondary Structure (S), and 4 pseudo amino acid composition (PseAAC) [4] with four different *landa*, 4 kinds of physicochemical properties of constituent amino acids (Hydrophobicity (H), Polarity (P), van der Waals volume (V) and Polarizability), and 2 local pairwise sequence alignment-based feature spaces, as well as sequence evolution information extracted directly from position specific scoring matrices in 14 different ways.

For predicting protein sub-nuclear localization, we also consider various sequence-based protein features including 2 types of information extracted directly from position specific scoring matrices, 3 types of structural information and local sequence alignment.

RESULTS & DISCUSSION

Gaussian RBF kernel function is employed for different protein features. To see the advantage of fusing heterogeneous data sources for protein fold classification (protein sub-nuclear localization prediction) through intermediate-based data integration, we focus on combining 26 (10) RBF kernel matrices derived from each view on protein domains (nuclear proteins).

The kernel matrices are combined through various types of means like Karcher-KF, AGH-KF, AM, and LogE. Subsequently, the combined kernel is used to determine the performance. Then the classification is performed using a one-versus-rest (OVR) support vector machine (SVM).

We evaluate our method for classification on the SCOP PDB-40D benchmark dataset [2] which consists of 27 SCOP fold classes. Moreover, to compare the performance of our proposed approaches, we also consider three types of MKL approaches [8, 16, 18]. Figure 1 provides the success rates of our proposed kernel fusion approaches based on averaging of the kernel matrices, these MKL approaches, as well as the best existing methods for classification of protein folds in DD data set [1, 2, 3, 16, 17] .

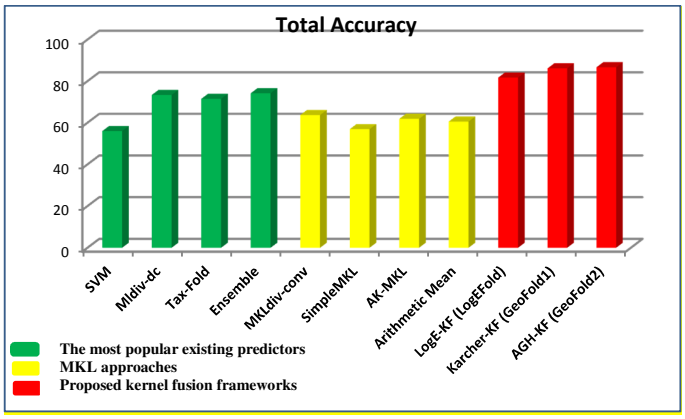

Figure 1. Comparison of the total accuracies of our proposed kernel fusion frameworks with existing predictor and Meta-predictors for classification of protein folds in the DD dataset[2] .

According to Figure 1, classification results of the combined kernels using Karcher-KF, AGH- KF, and LogE-KF show a clear improvement compared to the state of the art.

Furthermore, to incorporate the available functional domain information (FunD) of proteins, we consider the FunD composition of protein sequences using integrated FunD databases. For this purpose, we use the InterPro database [19], and Conserved Domain Database (CDD) [20]. Next, our computational model has been developed by incorporating the functional domain composition of proteins through a hybridization model 9 (Figure 1). It is observed that by using our proposed hybridization model the protein fold recognition accuracy is further improved to 89.30%.

The performance is estimated on an independent test set. The general architecture of the proposed approaches for classifying protein folds and predicting protein sub-nuclear localization are shown in Figure 1.
Furthermore, we investigate the performance of our approach on the newer SCOP database (version 1.75) [3] using 10-fold cross validation. Our results suggest that only by combining the evolutionary and secondary structural kernels through GKF, we obtain competitive results compared with the best ensemble approaches proposed for this problem [17]. In addition, it is observed that by incorporating the available functional domain information through our proposed hybridization model, we are able to tackle the protein fold recognition problem for 27-folds.

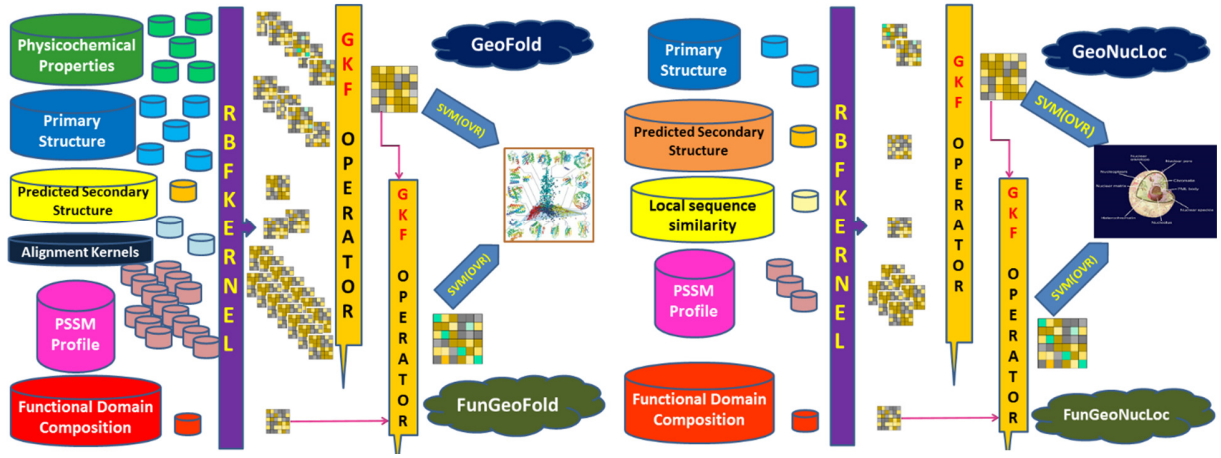

FIGURE 2. (a) The architecture of our fusion model for protein fold recognition (b) and predicting protein sub-nuclear localization.

We also investigate the performance of our geometric kernel fusion approach for predicting protein sub-nuclear localization. We evaluate our models by applying them to the SNL9 dataset [4] data sets using leave-one-out cross-validation. This data set contains 714 nuclear protein classified into 9 sub-nuclear locations. We achieve competitive results compared with the best ensemble approaches proposed for predicting protein sub-nuclear localization.
Moreover, our computational model has been developed by incorporating the functional domain information through the hybridization model. Experimental results on the SNL9 benchmark data set [4] demonstrate that our kernel-based integration approach can effectively improve the accuracy of the state-of-the-art protein sub-nuclear localization predictor.

CONCLUSIONS

In this study, our experimental results demonstrate that the geometric mean of base kernels can effectively improve the accuracy of the state-of-the-art kernel fusion model for protein fold recognition. Moreover, we enhance protein sub-nuclear localization prediction results on the SNL9 data set through our proposed kernel data fusion framework based on the geometric mean of kernel matrices. In addition, incorporating the available knowledge on functions of protein domains into our kernel data fusion framework offers a clear improvement in empirical performance.
Furthermore, the limitation of convex linear combinations in coping with combining different protein features which carry complementary information is considered [1]. By contrast, our proposed fusion frameworks can be used to detect these features with complementary information, which provides an intuitive strategy for integrating different features of other problems in bioinformatics. In particular, our results suggest that integrating the evolutionary and secondary structure information could be crucial to understand the relationship between primary and tertiary structure in proteins.

REFERENCES

1.Zakeri, P., *et al*., *J. THEOR. BIOL*, 269(1), 208 – 216 (2011).
2.Ding, C. H. and Dubchak I., *Bioinformatics,* 17(4), 349-358 (2001).
3.Yang, J.-Y. and Chen, X. *PROTEINS,* 79(7), 2053–2064. (2011)..
4.Shen HB, and Chou KC, *Protein Eng Des Sel*, 20: 561–56712, 2211–2268. (2007).
5.G¨onen, M. and Alpaydin, E. . *J. Mach. Learn. Res*, 12, 2211–2268. (2011).
6.Lanckriet, G. R. G., *et al*., *J. Mach. Learn. Res*, 5, 27–72. (2004).
7.Sonnenburg, S., *et al., J. Mach. Learn. Res*, 328-347 (2006).
8.Rakotomamonjy A., *J. Mach. Learn. Res,* **9**, 2491–2521, (2008).
9.Lanckriet, G. R. G., *et al. Bioinformatics,* 20(16), 2626–2635.(2004).
10.Wainer, H. *Psychological Bulletin*, 83, 213–217. (1976).
11.Arsigny, V. *et al., SIAM J Matrix Anal A,*29(1), 328-347 (2007
12.Bhatia, R., "*Positive Definite Matrices*", Princeton University Press, Princeton Series in Applied Mathematics. (2007).
13.Jeuris, B., *et al., ELECTRON T NUMER ANA,* 39, 379–402 (2012).
14.Lei ZD, and Dai Y., *BMC Bioinformatics*, 6: 291 (2005)
15.Zakeri, P., *et al*., *J. THEOR. BIOL*, 269(1), 208 – 216 (2011).
16.Ying, Y., *et al. BMC Bioinformatics*, 10(1), 267. (2009).
17.Lin, C., *et al, PLoS ONE*, 8(2), e56499. (2013).
18.Qiu, S. and Lane, T. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6(2), 190–199. (2009).
19.Hunter S *et al. Nucleic Acids Res,* 40(D1), 306-312 (2012).
20.Marchler-Bauer, A., *et al*., *Nucleic Acids Research*, 41(D1), D348–D352.. (2013).