PROTEIN FOLD RECOGNITION AND PROTEIN SUB-NUCLEAR LOCALIZATION USING GEOMETRIC KERNEL DATA FUSION

Pooya Zakeri (1,2), Yves Moreau (1,2)

1) Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium
2) iMinds Medical, Leuven, Belgium

BACKGROUND: Various sequence-based protein features and often machine learning methods developed to solve Bioinformatics tasks such as protein fold recognition and predicting protein sub-nuclear localization. More attention needs to be paid to finding an efficient method for fusing these different discriminatory data sources for these problems. In particular, kernel methods are an interesting class of techniques for data fusion. 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. While computing geometric means of matrices is challenging, it hints at an intriguing research direction in data fusion. 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 Protein fold recognition and predicting protein sub-nuclear localization.

METHOD: 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 design and develop several methods to integrate kernel matrices by taking the matrix geometric mean of base kernels. 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. It can be considered as a consensus between the arithmetic and geometric mean.

We consider various sequence-based protein features including structural information such as pseudo amino acid composition, predicted secondary structure, physicochemical properties of constituent amino acids, as well as sequence evolution information extracted directly from position-specific scoring matrices and local alignment kernels. Geometric kernel fusion (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.

RESULTS: Our GKF frameworks offer a significant improvement for both protein fold recognition and predicting protein nuclear localization. It is observed that using GKF, we achieve an overall protein fold accuracy of 86.86% on SCOP PDB-40D benchmark dataset [2]. It is observed that by using our proposed hybridization model the protein fold recognition accuracy is further improved to 89.30%. 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 GFK, we obtain competitive results compared with the best ensemble approaches proposed for this problem [4]. 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.

Moreover, experimental results on the SNL6 [5] and Nuc-PLoc [6] benchmark data set demonstrate that our integration technique can effectively improve the accuracy of the state-of-the-art protein sub-nuclear localization predictors.

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) Lin, C., et al, PLoS ONE, 8(2), e56499. (2013).
5) Lei ZD, and Dai Y., BMC Bioinformatics, 6: 291 (2005)
6) Shen HB, and Chou KC, Protein Eng Des Sel, 20: 561–56712, 2211–2268. (2007).