Protein Classification via Kernel Matrix Completion

Taishin Kin[1]
taishin@cbrc.jp

Tsuyoshi Kato[1]
kato-tsuyoshi@aist.go.jp

Koji Tsuda[1,2]
koji.tsuda@tuebingen.mpg.de

Kiyoshi Asai[1,3]
asai@cbrc.jp

[1] Computational Biology Research Center, AIST, Aomi-Frontier 17F, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany
[3] Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-8562, Japan

Keywords: information geometry, kernel matrix, incomplete matrix, protein structure

1 Introduction

Although 3D structure of a protein is valuable to predict its function, it is still far more difficult and costly to measure coordinates of atoms in a protein than sequencing its amino acids. We often do not know the 3D structures of all the proteins at hand. Let us consider a kernel matrix that consists of kernel values representing protein similarities in terms of their 3D structures where some of the entries are *missing* because structure information of some proteins are unavailable whereas their amino acid sequences are readily available. We proposes to estimate the missing entries by means of another kernel matrix derived from amino acid sequences. Basically a parametric model is created from the sequence kernel matrix, and the missing entries of the structure kernel matrix are estimated by fitting this model to existing entries. For model fitting, we adopt two algorithms: **e-projection** and **em algorithm** based on the information geometry of kernel matrices. We performed protein classification experiments by using support vector machines. Our results show that these algorithms can effectively estimate the missing entries.

2 Method and Results

Let us consider a kernel function as the similarity measure between two proteins. We consider two types of such functions: $k_{st}(\cdot,\cdot)$ for structure similarity and $k_{sq}(\cdot,\cdot)$ for sequence similarity. We define the following two matrices for ℓ proteins: *structure kernel matrix* D: $[D]_{ij} = k_{st}(x_i, x_j)$ and *sequence kernel matrix* M: $[M]_{ij} = k_{sq}(x_i, x_j)$ $(i, j = 1, \cdots, \ell)$, where $[M]_{ij}$ is (i, j)th element of a matrix M and x_i denotes the i th protein. We deal with a condition where D contains some missing entries. Let us rewrite D as an incomplete kernel matrix as: $D = \Big(K_I,\ D_{vh};\ D_{vh}^\top,\ D_{hh}\Big)$, where K_I is an $n \times n$ matrix corresponds to available entries, D_{vh} $(n \times m)$ and D_{hh} $(m \times m$ symmetric) correspond to missing entries. We propose several algorithms to estimate the missing entries by utilizing M as an auxiliary information source. By treating D_{vh} and D_{hh} as parameters, we define parametric model $\mathcal{D}$ to represent all of the admissible estimations for D: $\mathcal{D} = \{D \mid D_{vh} \in \Re^{n\times m},\ D_{hh} \in \Re^{m\times m},\ D_{hh} = D_{hh}^\top\}$. Let us partition M^{-1} so that its sub-matrices correspond to ones of D: $M^{-1} = \Big(S_{vv},\ S_{vh};\ S_{vh}^\top,\ S_{hh}\Big)$. One of our algorithms, named **e-projection** provides estimation of D_{vh} and D_{hh} with a closed form solution [3]: $D_{vh} = -K_I S_{vh} S_{hh}^{-1}$ and $D_{hh} = S_{hh}^{-1} + S_{hh}^{-1} S_{vh}^\top K_I S_{vh} S_{hh}^{-1}$. Another algorithm we propose,

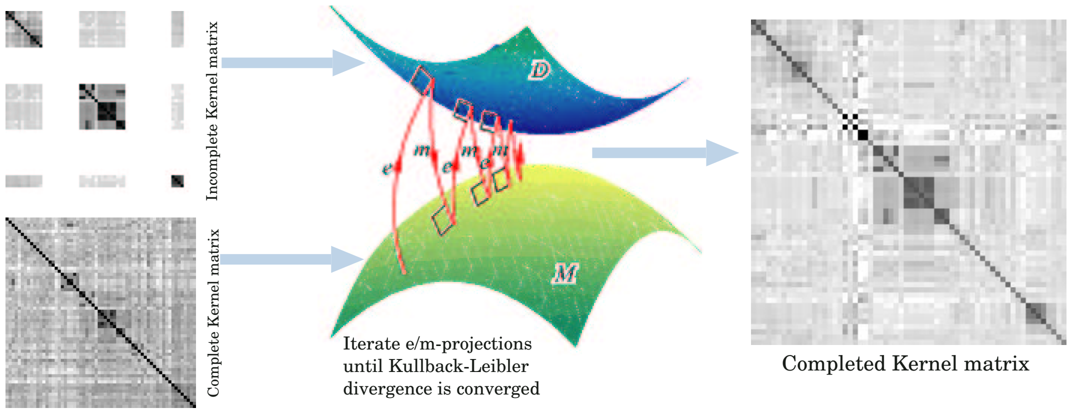

Figure 1: An information geometric view of kernel matrix completion with e/m-projections.

which is called **the em algorithm**, involves *m-projection* which does reverse of *e-projection.* According to our previous paper [3], we define the parametric model $\mathcal{M}$ as all *spectral variants* of M: $\mathcal{M} = \{M \mid M = \sum_{j=1}^{\ell} b_j \mathbf{v}_j \mathbf{v}_j^\top,\ \beta \in \Re^\ell\}$, where v_j is j th eigenvector of M. *m-projection* is computed as: $b_i = 1/tr(\mathbf{v}_i \mathbf{v}_i^\top D)$ [3]. However, we should note that the iteration of *e-projection* and *m-projection* usually does not map same points of each side back and forth. *The em algorithm* exploits this feature of the two projections. It alternatively iterates each projection until it reaches to a certain criterion. We use Kullback-Leibler (KL) divergence for this. In terms of information geometry, $\mathcal{M}$ and $\mathcal{D}$ can be viewed as two different probability distributions [3]. Therefore, KL divergence is defined among these two parametric models: $L_e \equiv KL(M, D) = tr(D\ M^{-1}) - \log\det D$ and $L_m \equiv KL(D, M) = \sum_{j=1}^{\ell} b_j tr(M_j\ D) - \log\det(\sum_{j=1}^{\ell} b_j M_j)$. These are expected to be converged to minimal through the iteration.

We perform protein classification experiments by using our kernel completion algorithms. We classified the proteins in a SCOP [2] superfamily (or fold) (glycosidases, NAD(P)-binding and TIM beta/alpha-barrels(fold)) into its families (or superfamilies) by using our algorithms. Given a complete kernel matrix of structure similarities (D), the fraction of removed rows/columns is changed from 10% to 90% by 10 point step. After completing the missing entries with one of the completion methods, the whole set of samples is randomly shuffled and divided into 50% training and 50% test set. The accuracies of SVM classifications are computed on these datasets. The result is that **em algorithm** performed the best at lower fraction of missing values, which indicates the algorithm can be a powerful tool for real world problems.

References

[1] Amari, S. and Nagaoka, H., *Methods of Information Geometry, Translations of Mathematical Monographs volume 191*, American Mathematical Society, 2001.

[2] Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C., *SCOP*: A structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, 247:536–540, 1995.

[3] Tsuda, K., Akaho, S., and Asai, K., The *em* algorithm for kernel matrix completion with auxiliary data, *Journal of Machine Learning Research*, 4:67–81, 2003.