Bioinformatics Center
– Bio-knowledge Engineering –

http://www.bic.kyoto-u.ac.jp/pathway/

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
TAKIGAWA, Ichigaku
(D Eng)

Assist Prof
HANCOCK, Timothy Peter
(Ph D)

PD
NGUYEN, Hao Canh
(Ph D)

Proj Res
NATSUME, Yayoi
(D Agr)

Students

du VERLE, David (D3)
TAKAHASHI, Keiichiro (D1)
CHEN, Zhuoxin (RS)

Visitors

Assoc Prof ZHU, Shanfeng	Fudan University, China, P.R., 23 June–1 July
Assoc Prof KIHARA, Daisuke	Purdue University, U.S.A., 29 August
Assoc Prof WANG, Fang	Southwest University of China, China, P.R., 19 September 2011–31 March 2012
Prof FRISHMAN, Dmitrij	Technische Universitat Munich, Germany, 11 October
Prof VARNEK, Alexandre	University of Strasbourg, France, 18 November

Scope of Research

We are interested in graphs and networks in biology and chemistry, which include metabolic networks, protein-protein interactions and chemical compounds. We have developed original techniques in machine learning and data mining for analyzing these graphs and networks, sometimes combining with table-format datasets, such as gene expression. We applied our techniques to real data to find new scientific insights. You can see our Research Topics regarding more detailed contents of our achieved or ongoing research.

KEYWORDS

Bioinformatics
Computational Genomics
Data Mining
Machine Learning
Systems Biology

Selected Publications

Nguyen, C. H.; Mamitsuka, H., Discriminative Graph Embedding for Label Propagation, *IEEE Transactions on Neural Networks*, **22(9),** 1395-1405 (2011).

Kayano, M.; Takigawa, I.; Shiga, M.; Tsuda, K.; Mamitsuka, H., ROS-DET: Robust Detector of Switching Mechanisms in Gene Expression, *Nucleic Acids Research*, **39(11),** e74 (2011).

Takigawa, I.; Tsuda, K.; Mamitsuka, H., Mining Significant Substructure Pairs for Interpreting Polypharmacology in Drug-target Network, *PLoS One*, **6(2),** e16999 (2011).

Shiga, M.; Takigawa, I.; Mamitsuka, H., A Spectral Approach to Clustering Numerical Vectors as Nodes in a Network, *Pattern Recognition*, **44(2),** 236-251 (2011).

Takigawa, I.; Mamitsuka, H., Efficiently Mining d-Tolerance Closed Frequent Subgraphs, *Machine Learning*, **82(2),** 95-121 (2011).

Latent Feature Models for Biological Networks

It is common in Systems Biology to represent biological systems as networks. Examples include the networks of protein-protein interactions (PPI) and gene regulatory networks (GRN). Biological networks, like many other types of networks, are known not to be random. Instead, network structures have models and patterns. By understanding the patterns and models of network structures, we would have insights into the biological mechanisms that generate the networks, leading to understanding of the generating processes of the networks. In the end, we would be able to infer knowledge of the biological systems, making judgments on what is on what is missing, what is erroneous. The understanding is to speed up experiment process by suggesting only relevant experiments to validate the understanding.

In this work, we aim to construct statistical models that describe network structures of PPI and GRN networks in order to *predict new links* (edges) in the networks. The motivation is that, for the case of physical protein- protein interactions, the binding site of each protein has to complement that of the other protein in shape. We hypothesize that the network structures follow *latent feature models*. The proof is that a link in the network is generated by a certain features of the nodes. These features alone determine the network structure around the nodes. By knowing all the features for the whole network, we would be able to generate the complete network. What we need is just to generate these latent features to fit the currently known networks obtained from high-throughput experiments. However, the methods for generating the features to describe the networks are usually computationally intractable.

We propose to use nonparametric models to describe network structures that follow latent feature models. Instead of generating latent feature explicitly, we encode them implicitly by providing a similarity function (kernel) on the nodes of the networks. The method is described as in Figure 1. The input of the method is a network as in (1), we use its adjacency matrix (2) to compute the similarity function. With the assumption of latent feature model (4), we embed pairs of network nodes (either known links or not links (5)) into a space (3). In the space, we use usual classification techniques to classify the link class versus the rest. The half-space inferred from classifiers for the link class is used to predict all the pairs of nodes to make a new adjacency matrix (6). The new adjacency matrix is used for the predicted network (7), showing the new links as well as the erroneous ones.

We apply the method to predict new links in the networks of PPI (yeast and fruit fly) and GRN (*E. Coli*). Our method was able to run on the whole networks within minutes (while other methods do not stop for days). Our method gives high prediction scores on these networks, making our method a reliable and scalable one to predict new links on these biological networks.

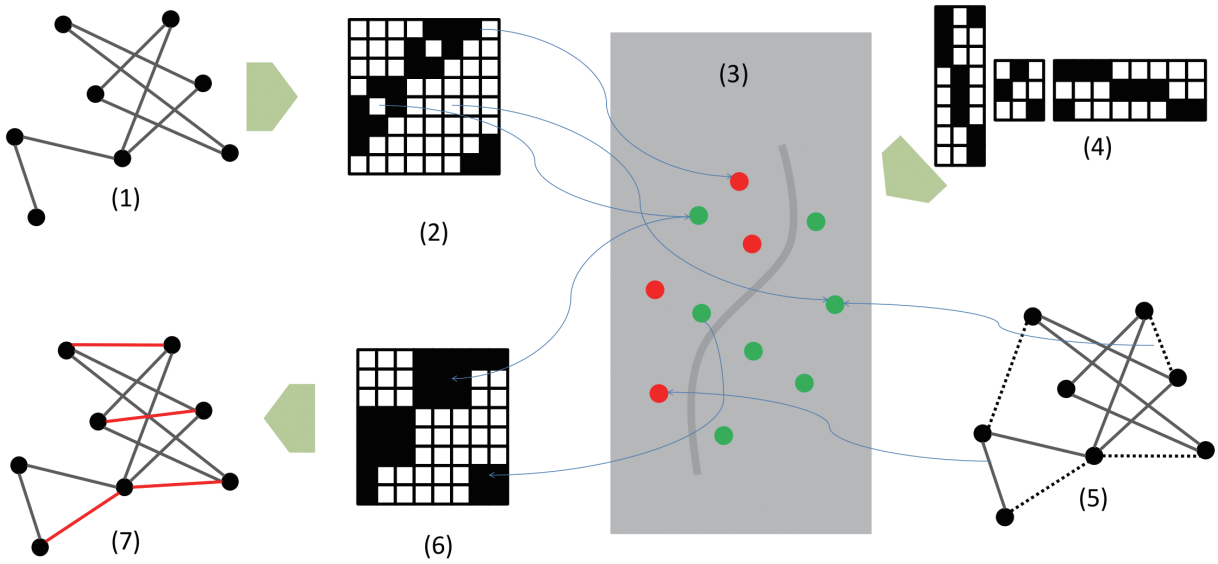

Figure 1.