Bioinformatics Center
– Bio-knowledge Engineering –

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

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
NGUYEN, Canh Hao
(D Knowledge Science)

Assist Prof
YAMADA, Makoto
(D Statistical Science)

Students

MOHAMED, Ahmed (D3) YOTSUKURA, Sohiya (D3) TOHZAKI, Yudai (UG)

Guest Scholar

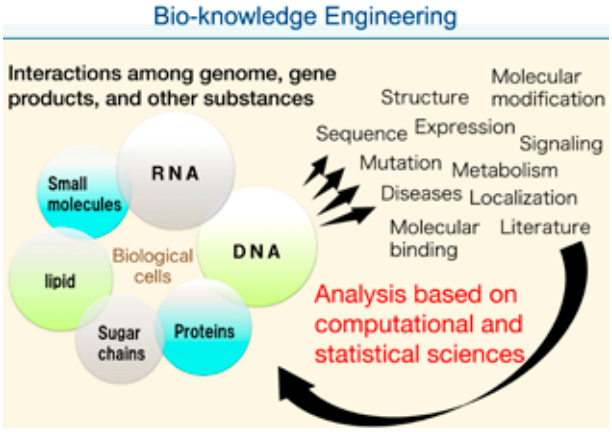
KASKI, Samuel Aalto University, Finland, 13 October-22 December

Guest Res Assoc

GAO, Junning Fudan University, China, P.R., 14 October 2015-8 January 2016
KANGASRÄÄSIÖ, Antti Aalto University, Finland, 26 October-23 December

Scope of Research

We are interested in graphs and networks in biology, chemistry, and medical sciences, including 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, occasionally combining with table-format datasets, such as gene expression and chemical properties. We have applied the techniques developed to real data to demonstrate the performance of the methods and find new scientific insights.



KEYWORDS

Bioinformatics Computational Genomics Data Mining
Machine Learning Systems Biology

Selected Publications

Zheng, X.; Zhu, S.; Gao, J.; Mamitsuka, H., Instance-wise Weighted Nonnegative Matrix Factorization for Aggregating Partitions with Locally Reliable Clusters, *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 4091-4097 (2015).

Liu, K.; Peng, S.; Wu, J.; Zhai, C.; Mamitsuka, H.; Zhu S., MeSHLabeler: Improving the Accuracy of Large-scale MeSH indexing by Integrating Diverse Evidence, *Bioinformatics 31 (12) (Proceedings of the 23rd International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2015))*, i339-i347 (2015).

Yotsukura, S.; Mamitsuka, H., Evaluation of Serum-based Cancer Biomarkers: A Brief Review from a Clinical and Computational Viewpoint, *Critical Reviews in Oncology/Hematology*, **93 (2),** 103-115 (2015).

Shiga, M.; Mamitsuka, H., Non-negative Matrix Factorization with Auxiliary Information on Overlapping Groups, *IEEE Transactions on Knowledge and Data Engineering*, **27 (6),** 1615-1628 (2015).

Mohamed, A.; Nguyen, C. H.; Mamitsuka, H., Current Status and Prospects of Computational Resources for Natural Product Dereplication, (in press).

Selecting Graph Cut Solutions Using Global Graph Similarity

A graph is a general way to represent complex data for analysis. It is a natural way to represent networks, especially in biology, such as metabolic, protein–protein interactions or regulatory networks. Graphs are particularly useful for analyzing high-dimensional data with complicated distributions encountered in various situations in high-throughput biological experiments. Graphs are an integrated framework to analyze expression data of many genes at various time points under many different conditions.

A common data analysis task is *clustering*, which groups similar data points into the same cluster. In the context of graph, similar data points are well-connected clusters with many short paths within the clusters. By grouping these nodes, it is equivalent to *cutting* the edges between the clusters to retain well-connected clusters. Hence, clustering is usually considered a graph cut method.

Clustering on graphs is usually formulated as an optimization problem, of which the objective functions are usually the cut's quality on graphs involving small cut's value (the number of edges being cut) and the balance of a clustering solution in terms of sizes of clusters. However, easily computable objective functions are usually not expressive enough to capture many different scenarios of data. A serious problem is that optimizing the objective functions does not always lead to well-connected clusters. In fact, optimal clustering solutions may even contain disconnected clusters, which defeats the purpose of clustering (Figure 1). Furthermore, detecting disconnected clusters is computationally too expensive to be used in objective functions of clustering algorithms.

Our idea to solve this problem is to use a *global graph similarity* measure named *ged*. It is a measure that, even though computationally too demanding to be used as a clustering objective function, is more expressive for distinguishing undesired clustering solutions that usual objective functions cannot recognize. The motivation is that undesired solutions with disconnected clusters are topologically more different from ones with well-connected clusters. Therefore, global graph similarity of the original graph with clustering solutions could show undesired clustering solutions, even though they could be as optimal as the desired ones.

The measure is formulated with eigenvectors and eigenvalues of the graph Laplacians. It was proved that the measure is equivalent to embedding graphs into a space, then comparing the embedded node sets on the space. In fact, it is equivalent to the standard method for independence measure of HSIC. The global graph similarity measure was used in simulated and real networks and shows that it could detect undesired clustering solutions.

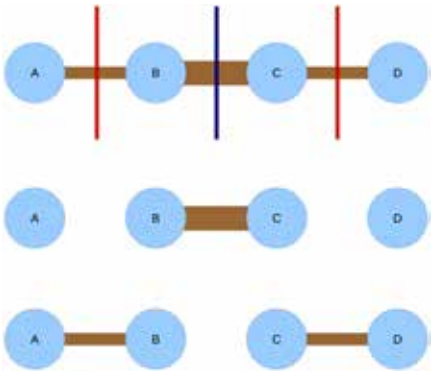


Figure 1. Equally optimal solutions of graph cut clustering on the original graph (top), resulting in an undesired clustering solution (middle), as well as a desired solution (bottom).

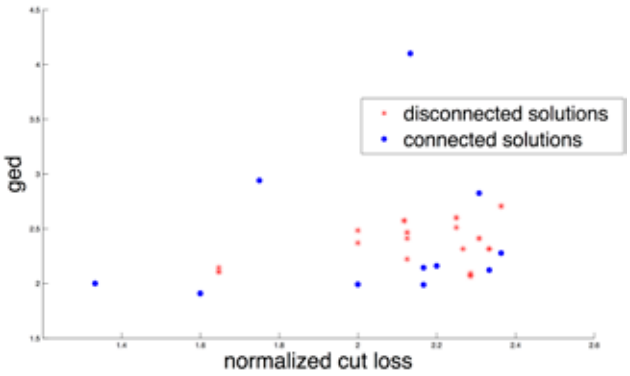


Figure 2. In karate data, *ged* can show that desired solutions with connected clusters tend to have small *ged* compared with the original graph, and can be a candidate for selecting clustering solutions.