Bioinformatics Center
– Bio-knowledge Engineering –

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

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
NGUYEN, Canh Hao
(D Knowledge Science)

Program-Specific Res
WIMALAWARNE, Kishan
(D Eng)

Program-Specific Res
SUN, Lu
(D Eng)

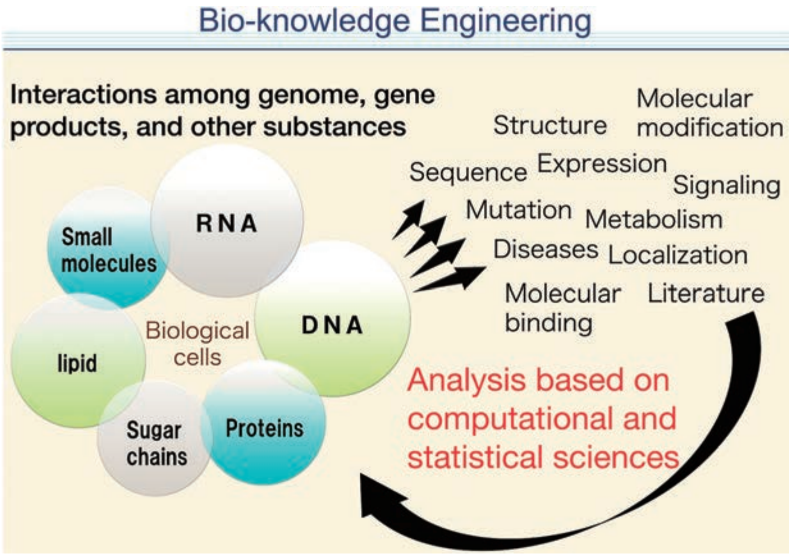
Students

NGUYEN, Dai Hai (D2) NGUYEN, Duc Anh (D1) TOHZAKI, Yudai (M2)

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

Wimalawarne, K.; Mamitsuka, H., Efficient Convex Completion of Coupled Tensors Using Coupled Nuclear Norms, *Proceedings of the Thirty-second Conference on Neural Information Processing Systems (NIPS 2018)*, 6902-6910 (2018).

Wimalawarne, K.; Yamada, M.; Mamitsuka, H., Convex Coupled Matrix and Tensor Completion, *Neural Computation*, **30(11),** 3095-3127 (2018).

Nguyen, D. H.; Nguyen, C. H.; Mamitsuka, H., SIMPLE: Sparse Interaction Model over Peaks of MoLEcules for Fast, Interpretable Metabolite Identification from Tandem Mass Spectra, *Bioinformatics*, **34(13),** i323-i332 (2018).

Mamitsuka, H., Data Mining for Systems Biology: Methods and Protocols (2nd Edition), *Methods Mol. Biol.*, **1807,** (2018).

Karasuyama, M.; Mamitsuka, H., Factor Analysis on a Graph, *Proceedings of Machine Learning Research (Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018))*, **84,** 1117-1126 (2018).

Advanced Machine Learning for Mass Spectrometry

Metabolites are small molecules, which are used in, or created by, the chemical reactions occurring in every cell of living organisms. They play lots of important roles including signaling, energy transport, building blocks of cells, etc. Identifying metabolites or interpreting their biochemical characteristics is an essential part of the metabolomics to enlarge the knowledge of biological systems. It is also a key to development of many applications and areas such as biotechnology, biomedicine or pharmaceutical sciences. Mass spectrometry is a common technique in analytical chemistry for metabolite identification. A mass spectrometer analyzes a chemical sample by fragmenting it and measuring the mass-to-charge ratios (m/z) of its fragments to obtain a mass spectrum (MS). A MS can be represented by a list of peaks, each of which corresponds to a fragment captured by the device. MS can provide more structural information about the sample and be helpful to deal with small-sized molecules such as metabolites.

There have been a number of computational methods proposed for metabolite identification from MS data. In general, they can be divided into three main categories: i) searching in spectral libraries; ii) in silico fragmentation and iii) machine learning. We focus on the machine learning approach, where the common scheme is, given a set of mass spectra, to learn a mapping from a spectrum to a molecule (Figure 1). This has two steps: 1) fingerprint prediction: predict a fingerprint with supervised ML and 2) candidate retrieval: use the predicted fingerprint to query the database. Kernel learning methods have been shown to be powerful tools in fingerprint prediction. For example, FingerID and CSI:FingerID are notable ones, which used support vector machine with kernels for pairs of MS and for pairs of corresponding fragmentation trees. However, existing methods are mainly based on individual peaks in the spectra, without explicitly considering the co-occurrence of peaks, which we call peak interactions. Also, these are computationally heavy and not desirable for the interpretation purposes.

The aim of our research is to propose and develop statistical learning models for identifying metabolites with the following main criterions: 1) High accuracy: given a query MS of a unknown metabolite, the proposed models are expected to produce a highly accurate list of candidate metabolites with most similar MS spectra; 2) Fast prediction: in order to be able to process large-scale datasets of metabolites in reality, it is desirable for the proposed model to produce good lists of candidates with fast prediction as well. Based on these, we developed a sparse interaction model, which we call SIMPLE, allowing to incorporate peak interactions for fingerprint prediction and is computationally lighter than existing kernel-based methods. As shown in Figure 2, the proposed methods achieved comparative prediction accuracy with much faster prediction (around 100 times). Furthermore, thanks to the interpretability, SIMPLE clearly revealed individual peaks and peak interactions, which contribute to enhancing the performance of fingerprint prediction.

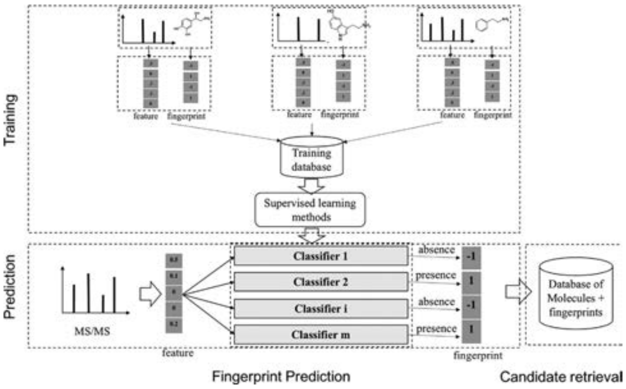
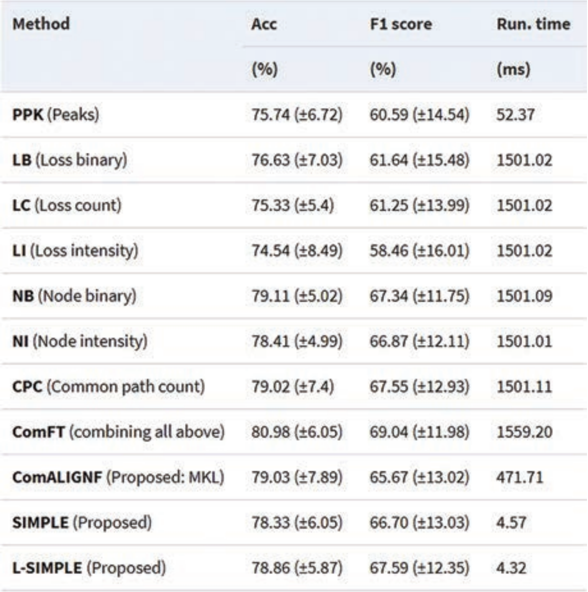


Figure 1. A general scheme to identify unknown metabolites based on molecular fingerprint vectors. There are two main steps: 1) fingerprint prediction; 2) Candidate retrieval.



Method	Acc	F1 score	Run. time
	(%)	(%)	(ms)
PPK (Peaks)	75.74 (±6.72)	60.59 (±14.54)	52.37
LB (Loss binary)	76.63 (±7.03)	61.64 (±15.48)	1501.02
LC (Loss count)	75.33 (±5.4)	61.25 (±13.99)	1501.02
LI (Loss intensity)	74.54 (±8.49)	58.46 (±16.01)	1501.02
NB (Node binary)	79.11 (±5.02)	67.34 (±11.75)	1501.09
NI (Node intensity)	78.41 (±4.99)	66.87 (±12.11)	1501.01
CPC (Common path count)	79.02 (±7.4)	67.55 (±12.93)	1501.11
ComFT (combining all above)	80.98 (±6.05)	69.04 (±11.98)	1559.20
ComALIGNF (Proposed: MKL)	79.03 (±7.89)	65.67 (±13.02)	471.71
SIMPLE (Proposed)	78.33 (±6.05)	66.70 (±13.03)	4.57
L-SIMPLE (Proposed)	78.86 (±5.87)	67.59 (±12.35)	4.32

Figure 2. Micro-average performance and prediction time of kernel-based methods and proposed methods.