Bioinformatics Center
– Bio-knowledge Engineering –

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

Prof
MAMITSUKA, Hiroshi
(D Sc)

Senior Lect
NGUYEN, Hao Canh
(D Knowledge Science)

Program-Specific Res
WIMALAWARNE, Kishan
(D Eng)

Program-Specific Res
SUN, Lu
(D Eng)

Students

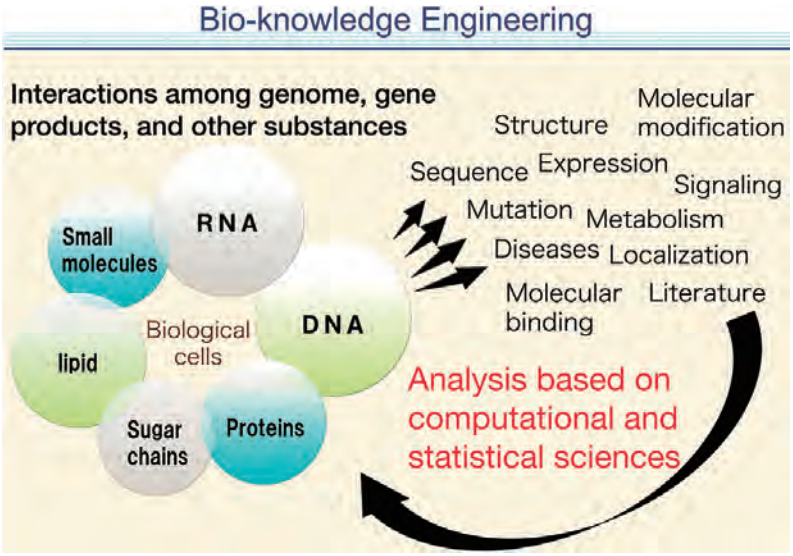
NGUYEN, Dai Hai (D3) NGUYEN, Duc Anh (D2)

Guest Scholar

KASKI, Samuel (Ph D) Aalto University, Finland, 2 April–31 May

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

Sun, L.; Nguyen, C. H.; Mamitsuka, H., Multiplicative Sparse Feature Decomposition for Efficient Multi-View Multi-Task Learning, *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 3506-3512 (2019).

Sun, L.; Nguyen, C. H.; Mamitsuka, H., Fast and Robust Multi-View Multi-Task Learning via Group Sparsity, *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 3499-3505 (2019).

You, R.; Yao, S.; Xiong, Y.; Huang, X.; Sun, F.; Mamitsuka, H.; Zhu, S., NetGO: Improving Large-scale Protein Function Prediction with Massive Network Information, *Nucleic Acids Res.*, **47,** W379-W387 (2019).

Nguyen, D. H.; Nguyen, C. H.; Mamitsuka, H., ADAPTIVE: leArning DAta-dePendenT, ConcIse Molecular VEctors for Fast, Accurate Metabolite Identification from Tandem Mass Spectra, *Bioinformatics (Proceedings of the 27th International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2019))*, **35 (14),** i164-i172 (2019).

Gillberg, J.; Marttinen, P.; Mamitsuka, H.; Kaski, S., Modelling G×E with Historical Weather Information Improves Genomic Prediction in New Environments, *Bioinformatics*, **35(20),** 4045-4052 (2019).

Advanced Machine Learning for Metabolite Identification from Mass Spectrometry

Metabolites are small molecules and play important functions in living cells such as energy transport, signaling, building blocks of cells and so on. Identifying their biochemical characteristics or so-called metabolite identification is an essential task in metabolomics to increase the knowledge of biological systems. However, it is still a challenging task due to the size or coverage of spectra libraries. Mass spectrometry is a widely used technique in analytical chemistry for dealing with metabolite identification task. In detail, a chemical compound is decomposed into fragments, of which mass-to-charge ratios (m/z) are measured to obtain a mass spectrum. The spectrum can also be represented by a list of peaks, each of which corresponds to a fragment captured by MS. The MS spectra provide structural information about the measured compound, which makes MS more useful for tackling the task of metabolite identification.

Computational methods proposed for identifying metabolites from MS data can be categorized into three main groups: (i) spectra library search; (ii) *in silico* fragmentation; and (iii) machine learning [1]. Our research focuses on machine learning based approach, where the common scheme is to predict a chemical structure of a given spectrum through an intermediate representation called molecular fingerprints. It consists of two steps: (i) predicting molecular fingerprints from spectra; (ii) searching molecular structures in database corresponding to the predicted fingerprints. Molecular fingerprints are often binary feature vectors, which should be large to cover all possible substructures and chemical properties, and therefore heavily redundant, in the sense of having many substructures irrelevant to the task, causing limited predictive performance and computational efficiency.

We propose a machine learning framework for metabolite identification task, named ADAPTIVE [2], which allows to learn representation for molecular structures, which we call molecular vectors, instead of using molecular fingerprints to characterize or represent molecules. It has two subtasks in learning step: (i) learning a mapping from structures to molecular representation vectors and (ii) learning another mapping from spectra to molecular vectors as illustrated in Figure 1. In Subtask 1, ADAPTIVE learns a mapping to generate molecular vectors for metabolites using their chemical structures and these vectors are specific to both data and task, and therefore less redundant. The mapping is parameterized by a model, namely message passing neural network (MPNN), and its parameters are trained so that the correlation (measured by Hilbert-Schmidt Independence Criterion, HSIC) between spectra and molecular vectors are maximized. For Subtask 2, ADAPTIVE use IOKR [3], standing for Input Output Kernel Regression, to learn another mapping from spectra to molecular vectors generated by the Subtask 1.

We conducted experiments using a benchmark data to evaluate the proposed method against existing ones in terms of predictive performance and computational efficiency. As shown in Figure 2, ADAPTIVE achieved the best predictive performance, outperforming the second best method, IOKR, with the difference of around 3–5% of *top*-20 under the same conditions. Furthermore, ADAPTIVE was significantly faster than IOKR (4–7 times) because molecular vectors by ADAPTIVE are much more concise and adaptive to given data and task than molecular fingerprints used in existing methods.

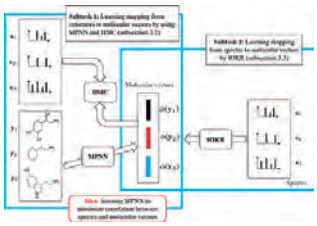


Figure 1. Overview of ADAPTIVE. It has two subtasks: (i) Subtask 1: estimate parameters of a mapping from structures to molecular vectors, given a set of spec-structure pairs; (ii) Subtask 2: learn another mapping from spectra to molecular vectors, generated by subtask 1.

Method	Vec. size	MKL	Accuracies (mean/SD %)		
			Top 1	Top 10	Top 20
FingerID	2765	None	17.74	49.59	58.17
CSI:FingerID unit	2765	ALIGNF	24.82	60.47	68.20
CSI:FingerID mod	2765	ALIGNF	28.84	66.07	73.07
Platt					
IOKR linear	2765	UNIMKL	30.58/2.23	65.99/2.46	73.53/2.47
		ALIGNF	28.34/2.54	65.77/2.39	73.19/3.11
ADAPTIVE linear	100	UNIMKL	29.42/2.83	70.01/2.79	77.48/2.98
		ALIGNF	29.19/3.21	69.52/2.89	77.64/3.23
	200	UNIMKL	29.57/3.96	69.38/3.05	76.95/2.98
		ALIGNF	29.11/3.45	69.53/2.52	77.56/2.43
	300	UNIMKL	30.22/3.47	70.48/2.72	78.18/2.67
		ALIGNF	**30.61/3.23**	**70.51/2.52**	**78.23/2.79**
IOKR Gaussian	2765	UNIMKL	30.66/2.34	66.51/2.87	73.94/2.54
		ALIGNF	**29.59/2.58**	**66.13/2.09**	**73.62/1.85**
ADAPTIVE Gaussian	100	UNIMKL	29.47/3.21	70.01/2.83	77.31/2.11
		ALIGNF	29.37/3.21	69.91/2.64	77.48/2.33
	200	UNIMKL	29.44/3.86	69.84/2.78	77.08/2.95
		ALIGNF	28.98/3.32	69.65/2.71	77.15/2.74
	300	UNIMKL	30.31/3.48	71.10/2.73	78.51/2.65
		ALIGNF	31.03/3.40	70.89/2.74	78.52/2.52

Method	Mol. vec. size	prediction time (ms/example)	
		Linear	Gaussian
IOKR	2765	140.22	3332.4
ADAPTIVE	100	20.32	802.6
	200	39.88	844.33
	300	54.14	1071.8

Figure 2. Evaluation of ADAPTIVE against existing methods in terms of *top-k* (k = 1,10 and 20) accuracies and compution time.

[1] Nguyen, D. H. *et al.*, Recent Advances and Prospects of Computational Methods for Metabolite Identification: a Review with Emphasis on Machine Learning Approaches, *Brief. Bioinf.*, doi:10.1093/bib/bby066 (2018).

[2] Nguyen, D. H. *et al.*, Adaptive: Learning Data-dependent, Concise Molecular Vectors for Fast, Accurate Metabolite Identification from MS/MS, *Bioinformatics*, **35,** i164-i172 (2019).

[3] Brouard, C. *et al.*, Fast Metabolite Identification with Input Output Kernel Regression, *Bioinformatics*, **32,** i28-i36 (2016).