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 NGUYEN, Anh Duc (D Pharm Sc)	Distinguished Visiting Senior Lect PETSCHNER, Peter (Ph D)

Students

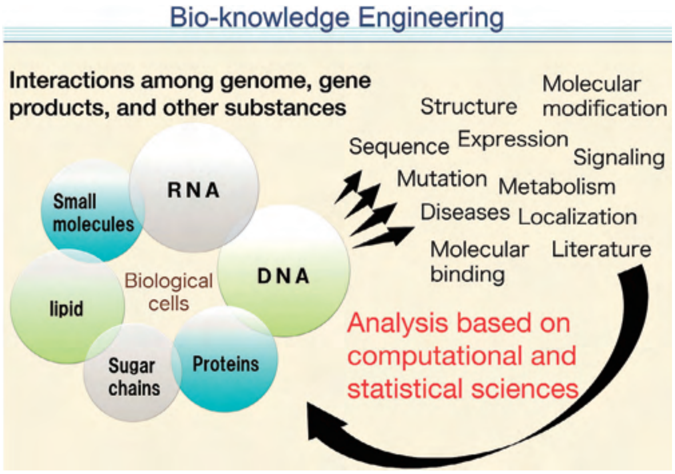
JIANG, Zhiqian (RS)
OTAGIRI, Yuan (UG)

Guest Res Assoc

LI, Yufei Xi'an Jiaotong University, China, 2 May 2022–23 April 2023

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 Machine Learning
Data Mining Artificial Intelligence Systems Biology

Recent Selected Publications

Wang, X.; Sun, L.; Nguyen, C. H.; Mamitsuka. H., Multiplicative Sparse Tensor Factorization for Multi-View Multi-Task Learning, *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*, 2560-2567 (2023).

Nguyen, D. A.; Nguyen, C. H.; Petschner, P.; Mamitsuka, H., SPARSE: A Sparse Hypergraph Neural Network for Learning Multiple Types of Latent Combinations to Accurately Predict Drug-drug Interactions, *Bioinformatics (Proceedings of the 30th International Conference on Intelligent Systems for Molecular Biology (ISMB 2022))*, **38** (Supplement 1), i333-i341 (2022).

You, R.; Qu, W.; Mamitsuka, H.; Zhu, S., DeepMHCII: A Novel Binding Core-Aware Deep Interaction Model for Accurate MHC II-peptide Binding Affinity Prediction, *Bioinformatics (Proceedings of the 30th International Conference on Intelligent Systems for Molecular Biology (ISMB 2022))*, **38** (Supplement 1), i220-i228 (2022).

Nguyen, C. H.; Mamitsuka, H., Learning on Hypergraphs with Sparsity, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43(8),** 2710-2722 (2021).

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).

Data Integrative Machine Learning: DIVERSE, An Example Approach to Personalized Medicine.

Multiple datasets can be found in any applications. For example, the main E-commerce data is a matrix of individuals (users) and items. Additionally, matrices on user demographic data and item contents can be obtained. In this case, the three matrices can be given, sharing the two dimensions, i.e. those of users and of items. Our focus is personalized medicine, where the main data is a matrix of individuals (patients, eventually cell lines) and their drug responses. The problem to be addressed is the drug response prediction, i.e. to predict unknown effective drugs for patients (cell lines). For this purpose, additional datasets can be used, such as a drug similarity matrix, drug-target interactions (a target is a protein, which is equivalent to the corresponding gene). These relevant data sources are called omics data, particularly in biology. Fig. 1 shows a schematic picture of omics data in drug response prediction, consisting of five matrices, including the main matrix *R* of drugs vs. cell lines.

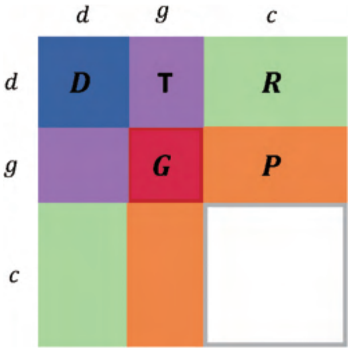


Figure 1. Conceptual integration configuration of the multiple data from three types of entities; *d*, *g* and *c* denote drugs, genes and cell lines, respectively.

Although large-scale omics data have been generated for drug response prediction, many machine learning methods have failed to achieve good performance for multiple heterogeneous data sources, because these methods have been designed for only a single type of data. Thus, a challenging task is to build precise prediction models on diverse data, coming from different sources, which are difficult to compare. In fact, data integration has to overcome several obvious problems, such as different data sizes, complexity, and noisiness. However, more importantly, data-integrative machine learning methods need to decide which information is useful to be incorporated and how significant the information is for the prediction task. This is the most critical problem to be addressed for machine learning models with diverse multi-omics data. For this problem, we propose DIVERSE, a framework to efficiently integrate scientifically diverse data, i.e. genomic, chemical and molecular interaction information. DIVERSE has two unique features: 1) It is methodologically flexible. Most existing studies ignore uncertainty, and hence cannot accept missing values.to predict missing drug responses of cancer cell lines. 2) It allows to compute *importance weights* over given multiple matrices, showing the contribution of the given matrices to prediction. DIVERSE solves these two practically important problems by using a Bayesian setting of matrix factorization. Fig. 2 shows the systematic framework of DIVERSE for the given matrix combination, shown in Fig. 1. In this framework, each dataset is incorporated into the matrix factorization framework of DIVERSE sequentially,

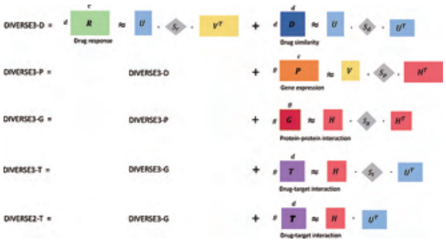
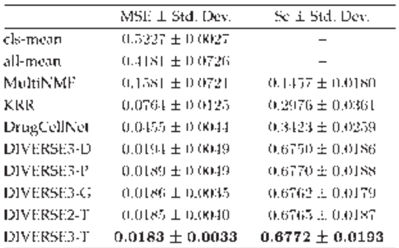


Figure 2. Overview of our systematic framework, DIVERSE, of integrating multiple data sets: importance weight tri-(or bi-)matrix factorization. We start with adding *D* to *R* (first row: DIVERSE3-D). We then add *P* to DIVERSE3-D (second row: DIVERSE3-P). Similarly, we add *G* to DIVERSE3-P (third row: DIVERSE3-G) and *T* to DIVERSE3-G (fourth row: DIVERSE3-T). Another option of the last addition is bi-matrix factorization, and this is the last row: DIVERSE2-T.

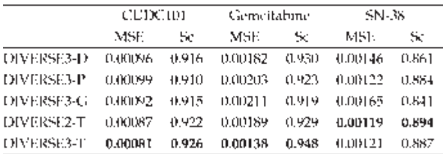
We empirically validated the performance of DIVERSE, comparing with five other methods, including three state-of-the-art methods, under 5x5-fold cross-validation. Experimental results indicate that DIVERSE significantly outperformed all compared methods in both mean-squared error (MSE) and Spearman correlation coefficient (Sc), particularly for out-of-matrix prediction, which is a real-world setting and much harder than in-matrix prediction. Results clearly show the performance advantage of DIVERSE over the current methods for predicting drug responses. Table 1 shows one typical example of the results, where ten methods are compared.



	MSE ± Std. Dev.	Sc ± Std. Dev.
cls-mean	0.5227 ± 0.0027	–
all-mean	0.4181 ± 0.0726	–
MultiNMF	0.1581 ± 0.0721	0.1457 ± 0.0180
KRR	0.0764 ± 0.0125	0.2976 ± 0.0361
DrugCellNet	0.0455 ± 0.0044	0.3423 ± 0.0259
DIVERSE3-D	0.0194 ± 0.0049	0.6750 ± 0.0186
DIVERSE3-P	0.0189 ± 0.0049	0.6770 ± 0.0188
DIVERSE3-G	0.0186 ± 0.0035	0.6762 ± 0.0179
DIVERSE2-T	0.0185 ± 0.0040	0.6765 ± 0.0187
DIVERSE3-T	**0.0183 ± 0.0033**	**0.6772 ± 0.0193**

Table 1. MSE and Sc (average scores of 5x5 cross-validation) of ten compared methods in out-of-matrix prediction.

Furthermore, the results indicate that the MSE and Sc of DIVERSE were smoothly improved by the step-wise addition of each data set. Table 2 shows the performances of different data integration types of DIVERSE for three different cancer cell line datasets.



	CUDC101		Gemcitabine		SN-38	
	MSE	Sc	MSE	Sc	MSE	Sc
DIVERSE3-D	0.00096	0.916	0.00182	0.930	0.00146	0.861
DIVERSE3-P	0.00099	0.910	0.00203	0.923	0.00122	0.884
DIVERSE3-G	0.00092	0.915	0.00211	0.919	0.00165	0.841
DIVERSE2-T	0.00087	0.922	0.00189	0.929	**0.00119**	**0.894**
DIVERSE3-T	**0.00081**	**0.926**	**0.00138**	**0.948**	0.00121	0.887

Table 2. Average MSE and Spearman correlation scores over 5x5-fold cross-validation for three different types of cancer cell lines.

Finally, these advantages of DIVERSE were confirmed by several case studies. Overall, DIVERSE is useful for performing integrative machine learning for given multiple omics data sources, which has not been handled by a regular machine learning algorithm.