様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １５ 日現在

機関番号：１４６０３
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００４０６

研究課題名（和文）A big data approach to function prediction of metabolites by clustering of structural similarity networks

研究課題名（英文）A big data approach to function prediction of metabolites by clustering of structural similarity networks

研究代表者
ＡＭＩＮ　ＭＤ．ＡＬＴＡＦＵＬ（AMIN, MD. ALTAFUL）

奈良先端科学技術大学院大学・先端科学技術研究科・准教授

研究者番号：３０３７９５３１

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：私たちは代謝物の活性を予測する手法を開発し、この手法を用いて1340種類の未知の代謝物の機能を予測することができた。またプロジェクトの一環として、単純グラフと2部グラフのクラスタリングのためのDPClusSBOというツールを開発した。このプロジェクトを中心に論文を多数発表していて、マレーシア、ブラジル、インドネシアの共同研究者もこのツールを使用している。将来的にはさらに広くこのツールが使用されることを期待していて、またツール自体も伝統医学の抗生物質の探索など応用範囲を広げていきたいと考えている。

研究成果の学術的意義や社会的意義
By utilizing our method we predicted the functions of 1340 unknown metabolites. As part of this project we developed a tool for clustering of simple and bipartite graphs which we have utilized in several other research works. This tool is a significant academic and social achievement.

研究成果の概要（英文）：We have developed a method to predict the activity of metabolites and have been able to predict the function of 1340 unknown metabolites using this method. As part of the project, we have also developed a tool called DPClusSBO for clustering simple graphs and bipartite graphs. We have published many papers centered around this project and the tool we developed is being used by our collaborators in Malaysia, Brazil and Indonesia. We hope that this tool will be used more widely in the future, and we would like to expand the application of the tool ourselves such as for searching for antibiotics based on traditional medicines.

研究分野：Systems Biology

キーワード：Metabolomics　Metabolic Network　Network Clustering　Metabolite Activity　Big Data Biology　Algorithm

１．研究開始当初の背景

Background

Numerous chemical reactions of various types called metabolic reactions are continuously occurring in living organisms. It is to gain energy, and to obtain molecules that make up cell, and to relate to the environment to increase the survival probability.

Molecules biosynthesized by this metabolic reaction are called metabolites. Metabolites are divided into primary and secondary metabolites according to their role. Primary metabolites are metabolites essential for maintaining normal physiological processes such as cell growth, development, and reproduction. The primary metabolites are amino acids, sugars, nucleic acids, etc. Many organisms biosynthesize primary metabolite by primary metabolic pathway.

Secondary metabolites are metabolites that are not classified as primary metabolites. Secondary metabolites are mainly biosynthesized by bacteria, fungi, and plants, and many are endemic to the species. Deficiencies in secondary metabolites do not result in immediate death, but affect the survival and health of the organism over an extended period of time. In some case, no major change is observed. Secondary metabolites are important to adapt to the environment and increase the probability of survival.

Through evolution process, organisms acquired pathways that biosynthesize diverse metabolites. Evolutionary pressures have led production of various secondary metabolites and constructed pathways to improve fitness for survival of organisms. Metabolites that are involved in growth, development, and reproduction of an organism are called “primary metabolites” and their production is referred as the “central metabolic pathway” which are key components in maintaining normal physiological processes. On the other hand, “secondary metabolites” are remaining compounds, that, although important, are not essential for the survival of an organism under suitable environment.

In species-metabolite relational database KNApSAcK, currently there are 111199 records involving 50899 metabolites and 22350 species. However, the activities of only 3,210 metabolites are recorded in KNApSAcK metabolite activity database. Given the importance of metabolites in agriculture, ecology and healthcare it would be of great significance to predict the functions of the other metabolites included in the KNApSAcK database. In this regard a computational method would be less expensive and fast approach.

In the current work we propose to develop a computational method to predict the functions of a huge set of metabolites whose functions are unknown, utilizing structural similarity networks. Such networks contain relations between metabolites concerning chemical similarity of metabolites.

２．研究の目的

Objectives

To understand the interaction of plants with environment, the interdependence of various organisms and ecology in general, it is first necessary to know the functions of metabolites, specially the secondary metabolites. Knowing the functions of secondary metabolites will also help to analyze the evolution of metabolic pathways and diversity of enzymes in plants.

Conventional agricultural industry relies on a wide use of chemical pesticides and fertilizers. However, increased demand for organic products shows that consumers prefer reduced chemical use. Volatile organic compounds (VOCs) are special type of metabolites. VOCs emitted by bacteria and fungi might have the potential to be alternatives to the use of chemical pesticides to protect plants from pests and pathogens. microbial VOCs are seen as biocontrol agents to control various phytopathogens and as biofertilizers for plant growth promotion. Furthermore, metabolites are deeply involved in human health care. The use of

VOCs as biomarkers to detect human diseases is rapidly increasing. Plant metabolites are a major sources of drugs and they are widely used in pharmaceutical industries. The above discussion implies the importance of function prediction of huge set of metabolites.

The secondary metabolites synthesized by plants, fungi and microorganisms are diverse, e.g. there are at least 30,000 terpenoids, 9,000 flavonoids, 1,600 isoflavonoids and 12,000 alkaloids. These metabolites have been deposited in the species-metabolite relational database(DB) known as the KNApSAcK Core DB, which contains 109,976 species-metabolite relationships involving altogether 22,399 species and 50,897 metabolites. Out of 50,897 metabolites of the KNApSAcK database the activity of only 3,161 metabolites are recorded. Given the importance of metabolites in agriculture, ecology and healthcare it would be of great help to various sectors to predict the activities of the unknown metabolites by using some computational means. However, a computational method would be less expensive and fast approach. Therefore, as a first step it is wise to develop and apply a computational approach which is the focus of our work.

Previously it was reported that structural similarity between metabolites imply similar activities. Several study used primary, secondary and tertiary structure similarities between proteins for predicting protein functions (*e.g. Nature Reviews Molecular Cell Biology 8.12 (2007): 995-1005*). At metabolite level also there have been a number of researches showing activity structure relationships (*e.g. Free radical biology and medicine 20.7 (1996): 933-956*). It has been reported that structural similarity between metabolites implies similar functions.

Network clustering algorithms have been successfully applied in different fields in recent years. In systems biology, network clustering has been applied in protein-protein interaction networks, gene expression networks and metabolic pathways. Many more such applications of network clustering can be mentioned. In the present work, we applied network clustering for activity prediction of metabolites.

In summary, the purpose of the present proposal is to develop a computational method for function prediction of metabolites by clustering the structural similarity based network utilizing the concepts of network clustering and False Discovery Rate (FDR).

３．研究の方法

Data Set

We collected molecular structure data of 50,037 metabolites recorded in KNApSAcK Core database. We obtained molecular structures from KNApSAcK database as MOL formats and input to COMPLIG algorithm for calculating structural similarity between metabolites. From KNApSAcK Metabolite Activity database, we downloaded 9,809 secondary metabolite-biological activity relationships involving in total 155 types of activity categories.

Flow of the proposed method

Figure 1 shows prediction method. First, we determined Highly Structurally Similar (HSS)

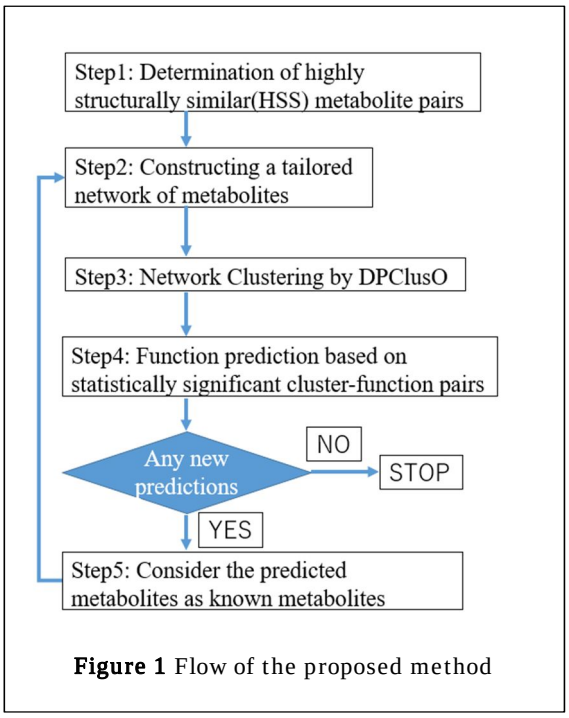

Figure 1 Flow of the proposed method

metabolite pairs based on COMPLIG algorithm and then constructed a restricted network of secondary metabolites. Next, we extracted clusters from the network by overlapping graph clustering algorithm DPClusO, and validated the relationship between cluster and bioactivity by calculating *p*-values using a hypergeometric distribution. Finally, we select statistically significant cluster-activity pairs based on false discovery rate (FDR) analysis and predict activities of some activity unknown metabolites.

４．研究成果

We developed a method for prediction of metabolite activities. Also as part of this project we developed a tool called DPClusSBO for clustering of simple and bipartite graphs. We published a number of papers keeping this project at the core in reputed journals including *BMC Medical Genomics, IEEE/ACM transactions on computational biology and bioinformatics, BMC Bioinformatics, Reproductive biomedicine online, Applied Network Science, Journal of Computer Aided Chemistry, Academia Journal of Medicinal Plants.*By utilizing our method, we predicted the functions of 1340 unknown metabolites. As part of this project we developed a tool for clustering of simple and bipartite graphs which we have utilized in several other research works. Our collaborators in Malaysia, Brazil and Indonesia are also using this tool and we hope this tool will be utilized by many other researchers in future. Also, we will use it for finding antibiotic drugs based on traditional medicine formulas under the next KAKENHI project we got.

Based on this project and other related works we achieved the following publications:

Book Chapters

1. **Altaf-Ul-Amin, M**., & Kanaya, S. (2020). Applications of Network Clustering in Natural Product Research. In book: Reference Module in Chemistry, Molecular Sciences and Chemical Engineering https://doi.org/10.1016/B978-0-12-409547-2.14785-7
2. Kanaya, S., **Altaf-Ul-Amin, M**., Aki, M. H., Huang, M., & Ono, N. (2020). Databases for Natural Product Research. In book: Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. https://doi.org/10.1016/B978-0-12-409547-2.14744-4
3. Azian Azamimi Abdullah**, Md. Altaf-Ul-Amin**, and Shigehiko Kanaya, Insight into KNApSAcK Metabolite Ecology Database: A Comprehensive Source of Species-VOC-Biological Activity Relationships (chapter 9), In the book: Volatile Organic Compound Analysis in Biomedical Diagnosis Applications, edited by Raquel Cumeras, PhD and Xavier Correig, PhD (http://www.appleacademicpress.com/volatile-organic-compounds-analysis in-biomedical-diagnosis-applications)
4. **Altaf-Ul-Amin, M.**, Mohamed-Hussein, Z.-A., & Kanaya, S. (2018). Investigating Metabolic Pathways and Networks, In the book: Reference Module in Life Sciences published by Elsevier. (doi:10.1016/b978-0-12-809633-8.20140-4)

Journal Papers

1. **Altaf-Ul-Amin, M**., Karim, M. B., Hu, P., Naoaki, O. N. O., & Kanaya, S. (2020). Discovery of inflammatory bowel disease-associated miRNAs using a novel bipartite clustering approach. BMC Medical Genomics, 13(3), 1-10.
2. Afiqah-Aleng, N., **Altaf-Ul-Amin, M**., Kanaya, S., & Mohamed-Hussein, Z. A. (2020). Graph cluster approach in identifying novel proteins and significant pathways involved in polycystic ovary syndrome. Reproductive biomedicine online, 40(2), 319.
3. Wakamatsu, N., Huang, M., Ono, N., **Altaf-Ul-Amin, M.**, & Kanaya, S. (2019). Prediction of Metabolite Activities by Repetitive Clustering of the Structural Similarity Based Networks. Journal of

Computer Aided Chemistry, 20, 76-83

4. Mohammad Bozlul Karim, Shigehiko Kanaya, **Md. Altaf-Ul-Amin**, "Implementation of BiClusO and its comparison with other biclustering algorithms", Applied Network Science, 4(1), 79.
5. Mohammad Bozlul Karim, Ming Huang, Naoaki ONO, Shigehiko Kanaya, **Md. Altaf-Ul Amin** ; "BiClusO: A novel biclustering approach and its application to species-VOC relationship data", IEEE/ACM transactions on computational biology and bioinformatics, 2019 May 14. doi: 10.1109/TCBB.2019.2914901
6. Hossain M.Mofazzal, M. S. A. Gazi, M. Mahbub, A. A. P. Sayed, S. Kanaya and **M. Altaf-Ul-Amin**, "On the determination of important plants for ayurvedic formulas in Bangladesh using unsupervised machine learning approach", Academia Journal of Medicinal Plants 7(2): 036-041, February 2019 DOI: 10.15413/ajmp.2019.0106 ISSN: 2315-7720
7. Eguchi, R., Karim, M. B., Hu, P., Sato, T., Ono, N., Kanaya, S., & **Altaf-Ul-Amin, M**. (2018). An integrative network-based approach to identify novel disease genes and pathways: a case study in the context of inflammatory bowel disease. BMC bioinformatics, 19(1), 264.
8. Liu, K., Abdullah, A. A., Huang, M., Nishioka, T., **Altaf-Ul-Amin, M**., & Kanaya, S. (2017). Novel Approach to Classify Plants Based on Metabolite-Content Similarity. *BioMed research international*, *2017*.
9. Liu, K., Morita, A. H., Kanaya, S., & **Atlaf-Ul-Amin, M**. (2018). *Metabolite-Content-Guided Prediction of Medicinal/Edible Prop-erties in Plants for Bioprospecting. Curr Res Complement Altern Med: CRCAM-130. DOI: 10.29011*. CRCAM-130/100030

Conference Proceedings/Posters

1. **Altaf-Ul-Amin, M.,** Karim, M. B., Hu, P., Naoaki, O. N. O., & Kanaya, S. (2020). Discovery of inflammatory bowel disease-associated miRNAs using a novel bipartite clustering approach. International Conference on Genome Informatics (GIW 2019) held in Sydney, Australia from 9th to 11th December, 2019.
2. Shaikh Farhad Hossain, **Md. Altaf-Ul-Amin**, Shigehiko Kanaya, Ming Huang and Naoaki Ono. Inter Disease Relations Based on Human Biomarkers by Network Analysis. Accepted for presentation in The 19th annual IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2019) held in Athens, Greece during October 28-30, 2019.
3. **Md. Altaf-Ul-Amin**, Nobutaka Wakamatsu and Shigehiko Kanaya. "An Approach to Function Prediction of Metabolites by Clustering the 3D-Chemical Structural Similarity Based Network", Translational Bioinformatics Conference 2016 (TBC 2016)', Oct. 15-17, 2016 in Jeju Island, Korea.
4. Nobutaka Wakamatsu, Tsuyoshi Shirai, Ryohei Eguchi, Mohammad Bozlul Karim, Naoaki Ono, Shigehiko Kanaya and **Md. Altaf-Ul-Amin**, "A Network based Approach to Predict Functions of Metabolites", International Conference on Metabolomics, 2018, Bangkok.
5. Shaikh Farhad Hossain, Sony Hartono Wijaya, Ming Huang, Irmanida Batubara, Shigehiko Kanaya and **Md. Altaf-Ul-Amin**, Prediction of Plant-Disease Relations Based on Unani Formulas by Network Analysis, The 18th IEEE International Conference on BioInformatics and BioEngineering (BIBE 2018), Taichung, Taiwan.

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Eguchi, R., Karim, M. B., Hu, P., Sato, T., Ono, N., Kanaya, S., & Altaf-Ul-Amin, M.	４．巻 19(1)
２．論文標題 An integrative network-based approach to identify novel disease genes and pathways: a case study in the context of inflammatory bowel disease	５．発行年 2018年
３．雑誌名 BMC bioinformatics	６．最初と最後の頁 264
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12859-018-2251-x	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Liu, K., Morita, A. H., Kanaya, S., & Atlaf-Ul-Amin, M	４．巻 CRCAM-130
２．論文標題 Metabolite-Content-Guided Prediction of Medicinal/Edible Prop-erties in Plants for Bioprospecting	５．発行年 2018年
３．雑誌名 Current Research in Complementary & Alternative Medicine	６．最初と最後の頁 2577-2201
掲載論文のDOI（デジタルオブジェクト識別子） 10.29011/2577-2201/100030	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Watanabe, Satoshi, K. Y. O. Hoko, K. A. N. G. Liu, Ryohei Eguchi, Md Altaf-Ul-Amin, Aki MORITA, Minako OHASHI et al.	４．巻 15(1)
２．論文標題 Data Intensive Study of Accessibility of Edible Species and Healthcare Across the Globe	５．発行年 2018年
３．雑誌名 Japanese Journal of Complementary and Alternative Medicine	６．最初と最後の頁 37-60
掲載論文のDOI（デジタルオブジェクト識別子） 10.1625/jcam.15.37	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 3.Antonio, V. A. A., Ono, N., Saito, A., Sato, T., Altaf-Ul-Amin, M., & Kanaya, S.	４．巻 13(12)
２．論文標題 Classification of lung adenocarcinoma transcriptome subtypes from pathological images using deep convolutional networks	５．発行年 2018年
３．雑誌名 International journal of computer assisted radiology and surgery	６．最初と最後の頁 1905-1913
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11548-018-1835-2	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Hossain M. Mofazzal, M. S. A. Gazi, M. Mahbub, A. A. P. Sayed, S. Kanaya and M. Altaf-Ul-Amin	4．巻 7(2)
2．論文標題 On the determination of important plants for ayurvedic formulas in Bangladesh using unsupervised machine learning approach	5．発行年 2019年
3．雑誌名 Academia Journal of Medicinal Plants	6．最初と最後の頁 036-041
掲載論文のDOI（デジタルオブジェクト識別子） 10.15413/ajmp.2019.0106	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Sony Hartono Wijaya, Irmanida Batubara, Takaaki Nishioka, Md. Altaf-Ul-Amin, and Shigehiko Kanaya	4．巻 36
2．論文標題 Metabolomic Studies of Indonesian Jamu Medicines: Prediction of Jamu Efficacy and Identification of Important Metabolites	5．発行年 2017年
3．雑誌名 Molecular Informatics	6．最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/minf.201700050	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Eguchi, Ryohei, Naoaki Ono, Aki Hirai Morita, Tetsuo Katsuragi, Satoshi Nakamura, Ming Huang, Md Altaf-Ul-Amin, and Shigehiko Kanaya	4．巻 18
2．論文標題 Classification of Alkaloid Compounds Based on Subring Skeleton (SRS) Profiling: On Finding Relationship of Compounds with Metabolic Pathways	5．発行年 2017年
3．雑誌名 Journal of Computer Aided Chemistry	6．最初と最後の頁 58-75
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.2751/jcac.18.58	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Mohammad Bozlul Karim, Nobutaka Wakamatsu and Md. Altaf-Ul-Amin	4．巻 18
2．論文標題 DPClusOST: A Software Tool for General Purpose Graph Clustering	5．発行年 2017年
3．雑誌名 Journal of Computer Aided Chemistry	6．最初と最後の頁 76-93
掲載論文のDOI（デジタルオブジェクト識別子） 10.2751/jcac.18.76	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Liu, K., Morita, A. H., Kanaya, S., & Atlaf-Ul-Amin, M.	4．巻 130
2．論文標題 Metabolite-Content-Guided Prediction of Medicinal/Edible Prop-erties in Plants for Bioprospecting	5．発行年 2018年
3．雑誌名 Curr Res Complement Altern Med	6．最初と最後の頁 1-15
掲載論文のDOI（デジタルオブジェクト識別子） 10.29011. CRCAM-130/100030	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計7件（うち招待講演　3件／うち国際学会　2件）

1．発表者名 Md. Altaf-Ul-Amin
2．発表標題 Applications of KNApSAcK Database and DPClus Algorithm: Plants to Metabolites to Target Proteins in the Context of Jamu Medicines and IBD Gene Prediction
3．学会等名 ICBET 2019, Meiji University, Tokyo（招待講演）
4．発表年 2019年

1．発表者名 Md. Altaf-Ul-Amin
2．発表標題 Introduction to Big Data Science Focusing Health Care and Ecology: A Case Study based on Indonesian Traditional Medicines
3．学会等名 Big data: Linking Ecosystem, health and wellbeing, BUET, Dhaka, Bangladesh（招待講演）
4．発表年 2019年

1．発表者名 Md. Altaf-Ul-Amin
2．発表標題 Data-Intensive Science of Jamu Medicines
3．学会等名 Universidade Federal de São Paulo, Brazil（招待講演）
4．発表年 2018年

１．発表者名
Mohammand Bozlul Karim

２．発表標題
Comparison of BiClusO with five different biclustering algorithms using biological and Synthetic data

３．学会等名
7th International Conference on Complex Networks and Their Applications, Cambridge, UK

４．発表年
2018年

１．発表者名
Shaikh Farhad Hossain

２．発表標題
Prediction of Plant-Disease Relations Based on Unani Formulas by Network Analysis

３．学会等名
18th IEEE International Conference on BioInformatics and BioEngineering, Taichung, Taiwan

４．発表年
2018年

１．発表者名
Md. Altaf-Ul-Amin

２．発表標題
A Network based Approach to Predict Functions of Metabolites

３．学会等名
International Conference on Metabolomics（国際学会）

４．発表年
2018年

１．発表者名
Yasue Aoyama

２．発表標題
Multifaceted analysis of sequence similarity based network of allergens

３．学会等名
International Conference on Computational Mathematics , Physics and It's Applications（国際学会）

４．発表年
2018年

〔図書〕　計2件

１．著者名 Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, and Shigehiko Kanaya	４．発行年 2018年
２．出版社 Apple Academic Press	５．総ページ数 20
３．書名 Volatile Organic Compound Analysis in Biomedical Diagnosis Applications(Chapter 9)	

１．著者名 M Altaf-Ul-Amin, S Kanaya, ZA Mohamed-Hussein	４．発行年 2018年
２．出版社 Elsevier	５．総ページ数 16
３．書名 Investigating Metabolic Pathways and Networks, In the book: Reference Module in Life Sciences	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考