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　　元 年　　６ 月 ２５ 日現在

機関番号：１４６０３
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ０７２２３

研究課題名（和文）生物活性にもとづいた二次代謝パスウェイにおけるモジュール構造の解明

研究課題名（英文）Relationship between activity and groups of metabolic pathways

研究代表者
金谷　重彦（Shigehiko, Kanaya）

奈良先端科学技術大学院大学・先端科学技術研究科・教授

研究者番号：９０２２４５８４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：生命維持に関わる代謝物である一次代謝物に対して、即時に死に至らないが、太陽光防御、ストレス対応、対生物など生存にかかわる代謝物質を二次代謝物という。地球全体で生物が生合成する二次代謝物は、顕花植物に限っても102万種と推定されており、非常に多様な生物活性を有する。本研究では、このような要請にバイオインフォマティクスより応え、生物活性による二次代謝パスウェイの機能単位（代謝モジュール）を抽出する目的で二次代謝パスウエイデータベースを構築し、バイオインフォマティクス技術、特に、深層学習を活用し生物の進化の過程で創出された代謝パスウェイと生物活性の関係における法則性を見出した。

研究成果の学術的意義や社会的意義
文献調査をもとに、640種のアルカロイドを中心に３０枚に及ぶ二次代謝経路マップを整理しウエブにより公開した。本代謝経路で、それぞれのアルカロイドの生合成開始物質を明確に確認できる。また、現在までに、12000種のアルカロイド化合物のうち深層学習により、10,051種について、生合成開始物質が予測でき、二次代謝経路予測における基盤となるデータを整備し、生物活性の関係も把握できる公開データベースを構築した。アルカロイド化合物の生合成開始物質の推定が可能になり、二次代謝研究における学術的意義が非常に高いことを示す。また、データを一般公開しており、社会的にも十分な貢献を果たしている。

研究成果の概要（英文）：In this study, we constructed a model to predict their precursors based on a novel kind of neural network called the molecular graph convolutional neural network and examined the relationships between activities and metabolites based on metabolic pathways. In order to investigate alkaloid biosynthesis, we trained the network to distinguish the precursors of 566 alkaloids, which are almost all of the alkaloids whose biosynthesis pathways are known, and showed that the model could predict starting substances with an averaged accuracy of 97.5%. The prediction of pathways contributes to understanding of the alkaloids and the application of graph based neural network models to similar problems in bioinformatics would therefore be beneficial. We applied our model to evaluate the precursors of biosynthesis of 18000 alkaloids found in natural organisms and found some rules of relationships between chemical structure and activity.

研究分野：バイオインフォマティクス

キーワード：代謝マップ　グラフコンボリューションネットワーク　深層学習

１．研究開始当初の背景
生命を維持するためのエネルギーならびに栄養素を生合成するために必要とされる一次代謝物に対して、即時に死に至らないが、太陽からの光防御、ストレス対応、対生物、生育戦略など生物の生存にかかわる代謝物質が二次代謝物である。地球全体で生物が生合成するユニークな二次代謝物の総数は、顕花植物に限っても 102 万種と推定されており、さまざまな生物活性を有している。

２．研究の目的
本研究では、二次代謝物に注目し、生物活性による代謝パスウェイの機能単位と、生物の進化の過程で創出された代謝パスウェイと生物活性の関係における法則性をバイオインフォマティクス技術により解明する。

３．研究の方法
(1)文献調査：本研究では、アルカロイド化合物に注目した。アルカロイド化合物の生合成の開始代謝物は、タンパク質合成に関わるアミノ酸類、テルペン、プリン、ピリミジンなどの核酸関連化合物と比較的限定されている。また、それぞれの代謝経路で生合成されるアルカロイド化合物には、類似の生物活性があることが知られている。そこで、文献調査にもとづいて、アルカロイド類の代謝経路を構築を行った。２年間にわたり約 850 報の論文を精査し、566 種のアルカロイド化合物についての代謝経路情報を獲得した。

(2)データベース構築：文献調査をもとに、アルカロイド生合成経路を３２枚の代謝マップに整理した。すべてのマップは、CobWeb データベースとして、http://www.knapsackfamily.com/CobWeb/top.jsp より公開した。

(3)アルカロイド化合物の生合成開始代謝物質の予測：現在までに、18000 種のアルカロイド化合物が収集され、KNApSAcK Core データベースに格納されている(http://www.knapsackfamily.com/knapsack_jsp/top.html)。一方で、生合成経路が既知のアルカロイド化合物は、千に満たず、本研究での調査では 566 種類にとどまった。そこで、これら 18000 種のアルカロイド化合物の生合成開始物質を推定する目的で、深層学習の一つである分子グラフ・コンボリューション・ニューラル・ネットワーク(MGCNN)を活用した。MGCNN において、原子を頂点、結合を辺とみなしたグラフにより化合物を取り扱うことができる(**図 1a**)。また、通常の深層学習における畳込み処理とプーリング処理を、グラフにおける原子の隣接性を考慮してコンボリューション・ニューラルネット・ネットワーク・モデルを構築できる。
　本研究では 566 種の化合物により、MGCNN モデルを構築し、KNApSAcK Core データベースに登録されている 18000 種のアルカロイド化合物に対して、生合成開始化合物の予測を行った。なお、データ数の制限から 15 種の生合成開始物質（L-Ala, L-Arg, L-Asp, L-His, L-Lys, L-Phe, L-Pro, L-Trp, L-Tyr, アントラニル酸、セコロガニン、IPP, GGPP, コレステロール、IGP）についての MGCNN モデルを構築した。

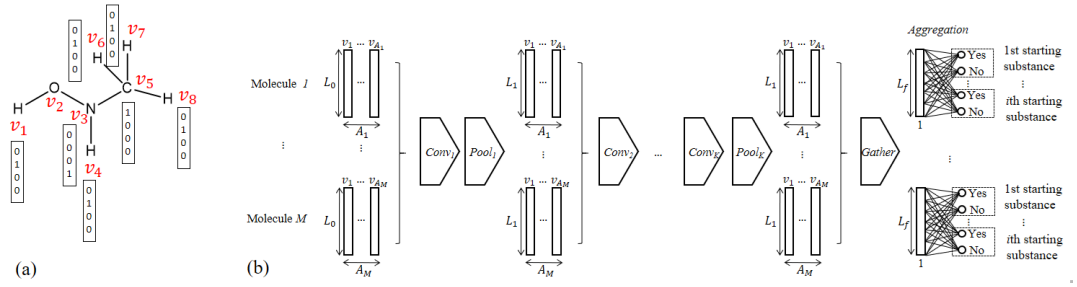

図 1: MGCNN の概念図. (a) 分子のグラフ表現, (b) MGNN

(4) 活性情報：　Metabolite Activity データベース（ http://www.knapsackfamily.com/MetaboliteActivity/top.jsp）を活用し、アルカロイド化合物の活性情報を獲得した。本データベースは、文献調査に基づく、2,356 種の代謝物からなり、9,584 対の代謝物と活性情報の関係がデータベース化されており、活性の種類としては、オントロジー解析から、140 種の活性グループに分類されている。これら 140 種の活性グループを２０種の包括的記述子に分類されている(**図 2**)。この活性情報に焦点をあてて代謝経路の関係を検討した。

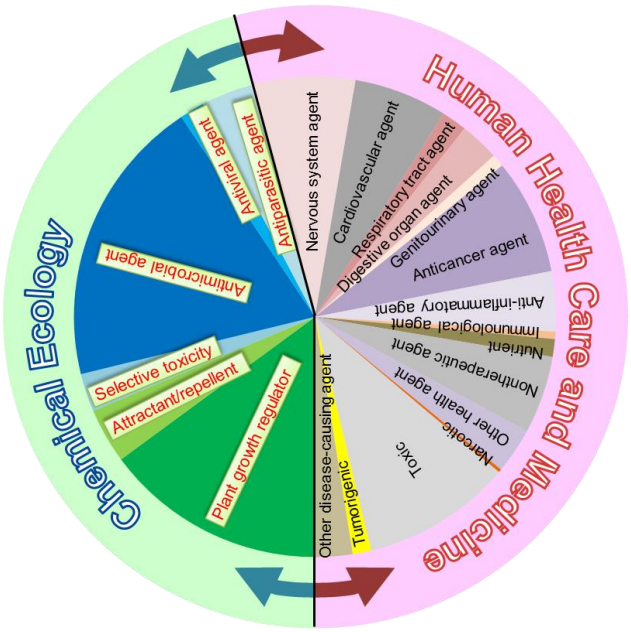

図 2: Metabolite Activity データベースにおける 20 種の包括的活性記述子の出現頻度

４．研究成果

(1)アルカロイド２次代謝経路データベース

アルカロイドにおける文献調査に基づいて収集した代謝反応情報をもとに代謝マップを構築した。その概略を**図 3** に示す。現状で、33 種の開始物質からアルカロイド化合物への代謝経路のデータベース構築が完了した。

左側の Group of SS をクリックすると、詳細な代謝経路情報を得ることができる(**図 4**)。図４は、文献情報にもとづいて構築されたアルカロイド代謝経路の一例である。この代謝マップ(アルカロイド代謝マップ ID=12)は、一つのアミノ酸を開始物質(L-Tyr)として、生合成されるアルカロイド化合物である。非常にさまざまなアルカロイド化合物が L-Tyr から生合成されることがわかる。本データへのアクセスは、ウエブを介して全て無償でアクセスできるので、様々な研究者が現在活用しているところである。このように、生合成開始物質に着目して、32 種のアルカロイド代謝マップを構築した。さらに、化合物名をクリックすると、生物種の情報、ならびに生物活性情報へとリンクづけがされているので、生物種固有、あるいは活性固有の代謝情報を得ることができる。

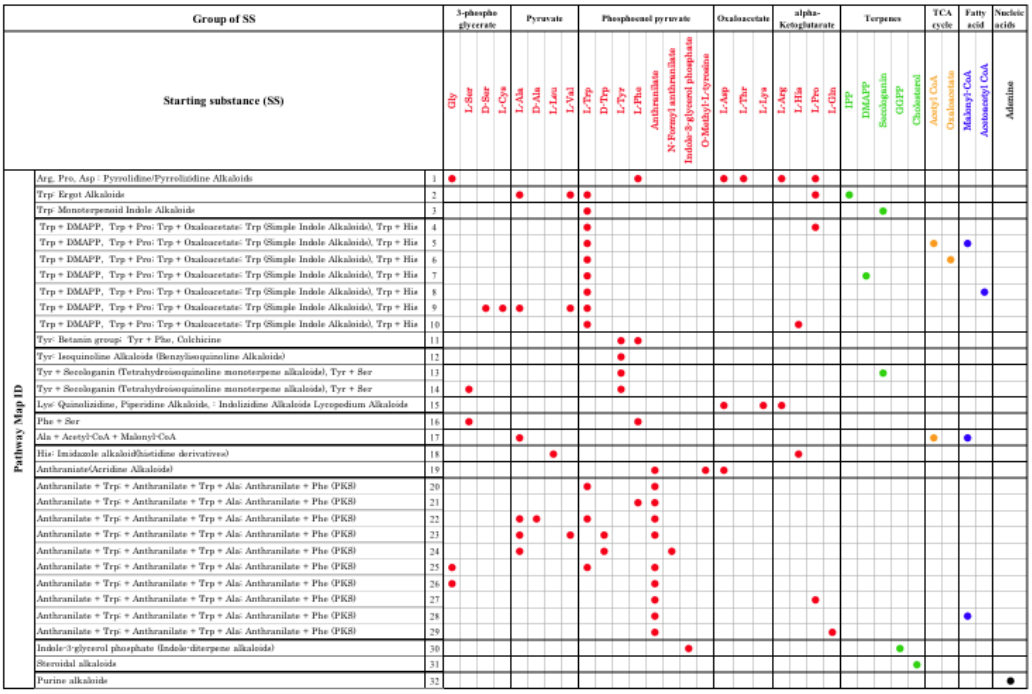

図 3: 32 種のアルカロイド化合物の 33 種の開始物質からの生合成経路

(2)MGCNN モデルによる代謝開始物質予測
化学構造をケミカルフィンガープリントにより構造表現し、通常のニューラルネットならびにランダムフォレストにおける認識率はそれぞれ、75%および 83%であるのに比べて、MGCNN モデルにおける認識率は 97.5%であり、その分類性能の高さが確認できた。MGCNN により、いまのところ生合成が未知の 18000 個の化合物について、その生合成開始物質を予測しウエブにより公開を進めている。現状で 17050 化合物については 15 種の生合成開始物質を割り振ることができた。さらに、アルカロイドに関わる代謝マップを整備したことにより、例えば sedative (M05)、analgesic (M02)、antispasmodic (M06)、nervous system agent, antitussive (M30) narcotic (M76)、antidiarrheic (M34)といった神経系に作用する活性における植物代謝モデュールを抽出することにも成功した。このように、代謝データの構築、深層学習による分類、活性情報という３つの基盤を活用することによる、生合成代謝経路と生物活性の関係を解明する生物間相互作用への新たな研究パラダイムへとさらに研究を進めるべきであると確信した。

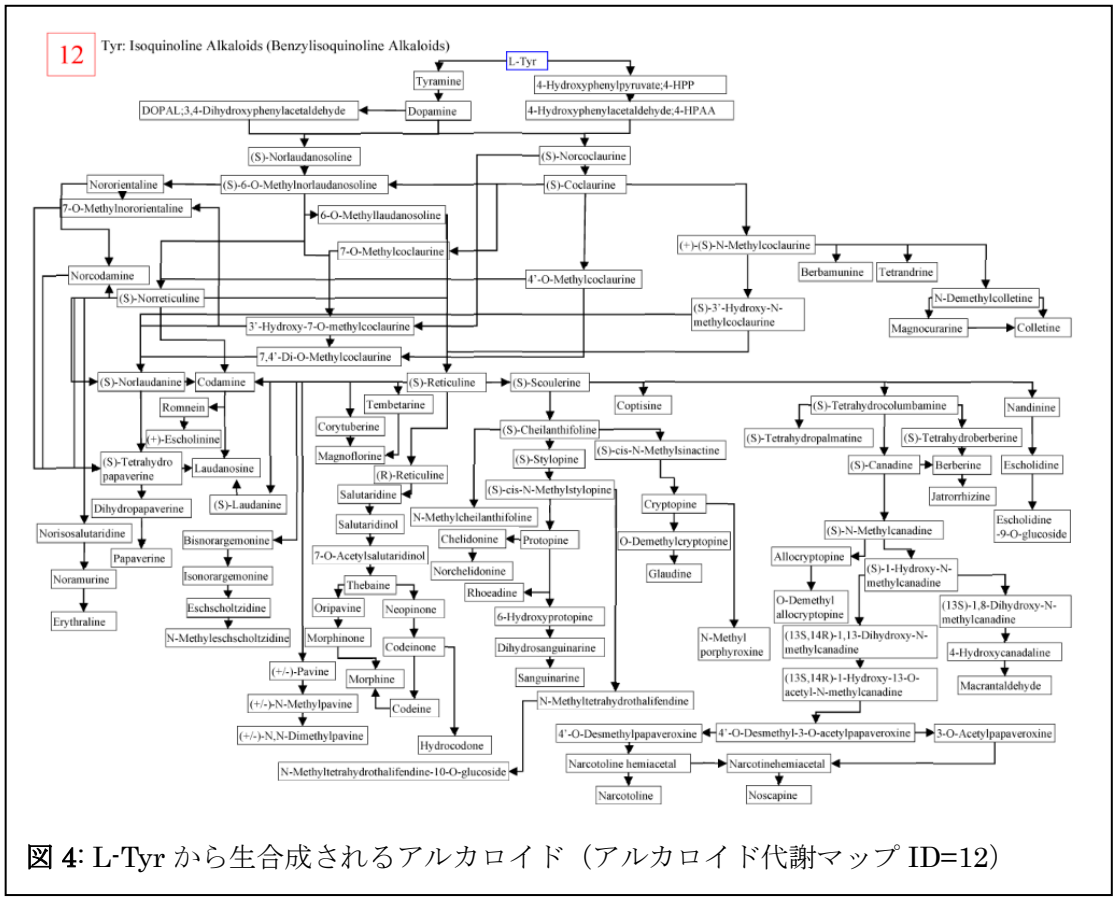

図 4: L-Tyr から生合成されるアルカロイド（アルカロイド代謝マップ ID=12）

５．主な発表論文等
〔雑誌論文〕（計 5 件）
① R. Eguchi, N. Ono, A. Hirai (Morita), T. Katsuragi, S. Nakamura, M. Huang, Md. Altaf-Ul-Amin, S. Kanaya, Classification of alkaloids according to the starting substances of their biosynthetic pathways using graph convolution neural networks, BMC Bioinformatics, in press, (2019)（査読あり）
②M. Karim, M. Huang, N. Ono, S. Kanaya, Md. Altaf-Ul-Amin, BiClusO: A novel biclustering approach and its application to species-VOC relational data, IEEE TRANSACTIONS on COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, MANUSCRIPT, (2019) doi: 10.1109/TCBB.2019.2914901（査読あり）
③R. Eguchi, M. B. Karim, P. Hu, T. Sato, N. Ono, S. Kanaya, Md. Altaf-Ul-Amin, An integrative network-based approach to identify novel disease genes and pathways:a case study in the context of inflammatory bowel disease, BMCBioinformatics, 19(1), 264, (2018) doi: 10.1186/s12859-018-2251-x.(査読あり)
④M. Ohnishi, A. Anegawa, Y. Sugiyama, K. Harada, A. Oikawa, Y. Nakayama, F. Matsuda, Y. Nakamura, R. Sasaki, C. Shichijo, P. G. Hatcher, H. Fukaki, S. Kanaya, K. Aoki, M. Yamazaki, E. Fukusaki, K. Saito, T. Mimura, Molecular components of arabidopsis intact

vacuoles clarified with metabolomic and proteomic analyses, Plant Cell Physiol., 59(7), 1353-1362, (2018) doi: 10.1093/pcp/pcy069. （査読あり）
⑤K. Liu, A. H. Morita, S. Kanaya and Md. Altaf-Ul-Amin, Metabolite-content-guided prediction of medicinal/edible properties in plants for bioprospecting, Current Research in Complementary & Alternative Medicine, ppCRCAM-130, (2018), doi: 10.29011/CRCAM-130/100030. （査読あり）

〔学会発表〕（計２件）
①金谷 重彦、二次代謝物の生合成経路の悉皆的解析を目指したデータサイエンス」第 64 回日本放線菌学会学術講演会、2019.3.14(東京)
②金谷 重彦、データサイエンス全般から化学の話題まで、Future Trend in Polymer Science 2018, 2019.3.2（東京）

〔図書〕（計１件）
① Md. Altaf-Ul-Amin, S. Kanaya, M.A. Hirai, M. Huang, N. Ono, Comprehensive Natural Products III: Chemistry and Biology, Databases for natural product research (Elisevia, Edited by Kazuki Saito)(2019)

〔産業財産権〕

○出願状況（計０件）

〔その他〕
○プレス発表：
①データ整理の計算手順を開発、先端大　ソフト無償公開へ、毎日新聞朝刊 2019.6.14.第 23 面
②生物・代謝物の関係視覚化、奈良先端科技大、分析ソフトの無償公開、日刊工業新聞 2019.6.11 第 21 面掲載
③画期的な分析方法考案、関係性の密な要素グループ化：世界で高い評価、多様分野で応用化、情報科学領域、先端大研究グループ、奈良新聞 2019.6.6 朝刊第３面掲載

○ホームページ：
アルカロイド代謝マップ：http://www.knapsackfamily.com/CobWeb/top.jsp
代謝物活性データベース：http://www.knapsackfamily.com/MetaboliteActivity/top.jsp

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。