

修士論文

ネットワークに基づく三次元分子構造類似性の
クラスタリングによる二次代謝物の機能推定

若松 信孝

2017 年 3 月 15 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である。

若松 信孝

審査委員：

金谷 重彦 教授	（主指導教員）
笠原 正治 教授	（副指導教員）
Md. Altaf-UL-Amin 准教授	（副指導教員）
小野 直亮 助教	（副指導教員）

ネットワークに基づく三次元分子構造類似性の クラスタリングによる二次代謝物の機能推定

若松 信孝

内容梗概

二次代謝物は主に植物や微生物などによって生産される低分子であり、天敵の忌避物質や紫外線防御などの形で生産者の生存戦略に用いられている。また、その生物活性の多様さは、医薬品や香料、嗜好品などの形で古くから人類に利用されてきた。そして近年、解析手法の進歩のために数多くの代謝物の分子構造が決定され、その情報がデータベースに蓄積されている。しかし、分子構造が決定された代謝物の数に比して、代謝物の持つ生物活性が決定された例は少ない。本研究では代謝物の3次元分子構造情報からネットワークを構築し、そのクラスタリングを行うことで分子構造と生物活性とを関連付け、その関係性に基づいて生物活性未知の代謝物の活性を予測した。KNApSAcK データベースの Core System に収録されている 50,037 件の代謝物の分子構造情報と、KNApSAcK Metabolite Activity データベースに収録されている 3,210 件の代謝物と 155 種類の生物活性との関連付け情報からなるデータセットに本研究の手法を適用した。本研究手法の予測精度は約 73% だった。

キーワード

クラスタリング、ネットワーククラスタリング、代謝物、生物活性、ケモインフォマティクス

* 奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文,
NAIST-ISMT1551120, 2017 年 3 月 15 日.

An approach to function prediction of metabolites by clustering the 3D-chemical structural similarity based network

Nobutaka Wakamatsu

Abstract

Secondary metabolites are small molecules generated mainly by plants and microorganisms. They are used for survival strategies such as ultra violet protection, insect repellent and pollinator attractant. Since these metabolites have diverse biological activities, humans have used them for various purposes such as medicines, flavor and recreational drugs etc. Recently, a large number of metabolites' molecular structures are determined by advanced molecular analysis methods. These structural data are accumulated in the databases. However, there are few metabolites whose biological activity was determined. In this thesis, we constructed a metabolite network from three dimensional structure information of metabolites. Next, clustering of metabolite network was performed to construct the relationship between molecular structure and biological activity. Finally, we predicted unknown biological activities of metabolites using data set containing 50,037 molecular structure information of metabolites and relationship between 3,210 metabolites and 155 biological activities. The accuracy of our method was about 73%.

Keywords:

clustering, network clustering, metabolite, biological activity, cheminformatics

* Master's Thesis, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1551120, March 15, 2017.

目次

1. はじめに	1
1-1. 背景	1
1-2. 二次代謝物について	2
1-3. 目的	2
2. 方法	3
2-1. データセット	5
2-2. 代謝物の分子構造類似度に基づくネットワークの構築 (Step1)	5
2-3. ネットワークからのクラスタ抽出 (Step2)	7
2-4. クラスタと生物活性の関係性の統計的な評価 (Step3)	14
2-5. 生物活性が未知の代謝物の活性予測 (Step4)	14
2-6. 予測手法の精度の評価	15
2-7. 生物活性間の関係性の調査	15
3. 結果と考察	16
3-1. 代謝物の分子構造類似度に基づくネットワークの構築	16
3-2. ネットワークからのクラスタ抽出	18
3-3. クラスタと生物活性間の関係の統計的な評価	18
3-4. 生物活性が未知の代謝物の活性予測	20
3-5. 予測手法の精度の評価	23
3-6. 生物活性間の関係性について	24
4. 結論	40
謝辞	41
参考文献	42

図目次

図 1. 研究手法の概要	4
図 2. DPCLUSO のアルゴリズムの概要	10
図 3. 各閾値以上の構造類似度を持つ代謝物ペア数の遷移	16
図 4. 構築された代謝物ネットワークの概観	17
図 5. p-value の区間ごとのクラスター-生物活性ペアの出現頻度	19
図 6. KnapSack データベース内の生物活性の出現頻度	21
図 7. 本研究の予測結果における生物活性の出現頻度	22
図 8. 10 回以上機能未知として扱われた生物活性と各生物活性との 同時出現頻度のヒートマップ	39

表目次

表 1. 予測精度評価の際の各生物活性の予測回数	25
表 2. 各生物活性の出現回数と予測結果における出現回数の比	31

1. はじめに

1-1. 背景

二次代謝物は植物や微生物によって生産される低分子化合物の総称であり、人間によって様々な目的のために使用されている。また、生物種間の生態学的な関係を形成するために重要な役割を果たしている。二次代謝物は生物に対する様々な作用(生物活性)を持つ。しかし、多くの二次代謝物の生物活性は解明されていない。このような生物活性が未知の代謝物の活性予測を行うことで、予測結果を用いた効率的な生物活性の探索および決定が可能になり、利用可能な代謝物-生物活性間の対応情報というリソースが蓄積されていくことで新しい薬剤や機能性食品の開発につなげることができると考えられる。また、利用可能なデータの増加によりメタボロミクス分野の研究を促進すると考えられる。

KNApSAcK Metabolite Activity データベースは生物の代謝産物とそれらの生物学的活性およびそれらの対象となる生物種間の三つ組みの関係性を提供するデータベースである。現在、3,161 種類の代謝物、155 種類の生物活性カテゴリ、912 種類の対象となる生物種の関連データが 9,809 件蓄積されている。

KNApSAcK Metabolite Activity データベースを用いて生物活性と代謝物の分子構造間の関係性が調査されている(Nakamura *et al.*, 2013)。この研究では、よく似た生物活性を持つ代謝物がグループにまとめられ、各グループにおける 11 種類の分子構造の出現頻度が調べられた。その結果、各グループに特徴的な構造が発見され、一部の分子構造と生物活性の関係性が示された。

代謝物の 3 次元化学構造と生物活性の関係性についても調べられている(Ohtana *et al.*, 2014)。この研究では、KNApSAcK Metabolite Activity データベース内に存在する活性が既知の代謝物について包括的な解析が行われた。この研究では、データベース内の代謝物を分子構造 3 次元的に類似している代謝物グループに分類された。これらの代謝物グループのうち 73%が少なくとも単一の生物活性と重要な関係があることが示された。

本研究では上記研究の手法を参考にして、代謝物の 3 次元分子構造と生物活性間の関係性に基づいた生物活性の予測を行った。

1-2. 代謝物について

本研究で扱う代謝物について述べる。代謝物は生物が行う代謝反応により生成される低分子である。代謝物は一次代謝物と二次代謝物に分類される。

一次代謝物は生物の生命維持や成長、生殖に直接かかわるものである。例えば、人間において肉体を構成するタンパク質や遺伝子を形成する核酸、エネルギーとして使用される糖や脂質は一次代謝物に分類される。多くの生物種に共通して存在することも一次代謝物の特徴である。

二次代謝物は生物の成長、発達および生殖に直接関与しない有機化合物を指す。一次代謝物とは異なり、二次代謝物の欠乏によって生物は即時の死には至らない。また、生物の系統群内の狭い範囲の種に特異的なものが数多く存在することも二次代謝物の特徴の一つである。人間によって一次代謝物は栄養として利用されるのに対して、二次代謝物はその多様な生物活性から香料、医薬品、嗜好品など様々な用途に利用されている。

二次代謝物の薬剤としての利用の例として、ベラドンナなどのナス科の植物に含まれるアルカロイドである scopolamine (Hyoscine) は薬剤として利用されている。この二次代謝物には乗り物酔いを抑える作用、胃酸の分泌を抑える作用、内臓周りの筋肉の痙攣を抑える作用などがある。そのため、消化管の痙攣による疼痛の抑制や胃炎による痛みの抑制、手術後の吐き気を抑えるために使用される。この scopolamine はそのまま用いた場合には人体に対する副作用があるので、構造を変えてブチルスコポラミン臭化物という形で副作用を大幅に軽減したものが薬剤として使用されている。

1-3. 目的

本研究の目的は、KNApSAcK Metabolite Activity データベースに蓄積された代謝物の分子構造と既知の活性情報を用いて、代謝物の 3 次元構造と生物活性間の関係性にに基づき、未知の生物活性を予測する事である。そのために、各代謝物を分子構造の類似度に基づいて接続した代謝物ネットワークを構築することで活性が未知の代謝物と活性情報のある代謝物を結び付ける。ネットワーククラスタリングを用いることで分子構造の類似性に基づく活性予測を行う。

2. 方法

本研究では生物活性情報が未知の代謝物の活性情報の予測を行うために、4つの Step からなる手法を用いた。Step 1 では代謝物間の分子構造に基づいたネットワークを構築した。Step 2 では構築したネットワークから密度の高い部分をクラスタとして抽出するためにネットワーククラスタリングアルゴリズムを適用した。このクラスタリングによって、分子構造の似た代謝物がクラスタにまとめられる。Step3 ではネットワーククラスタリングアルゴリズムの適用により得られた各クラスタと生物活性の間の関係性を統計的に検証した。このとき、クラスタ数×生物活性数の回数の検定を行うことになり、偽陽性の割合が高くなってしまう多重検定の問題が発生する可能性がある。この問題の補正を行うために False Discovery Rate(FDR)に基づいて有意なクラスタ-生物活性のペアを選出した。Step4 では、有意なクラスタ-生物活性ペアの情報から各クラスタに含まれる機能未知の代謝物の活性の推定を行った。

本研究で用いた手法の概要図を図 1 に示す。

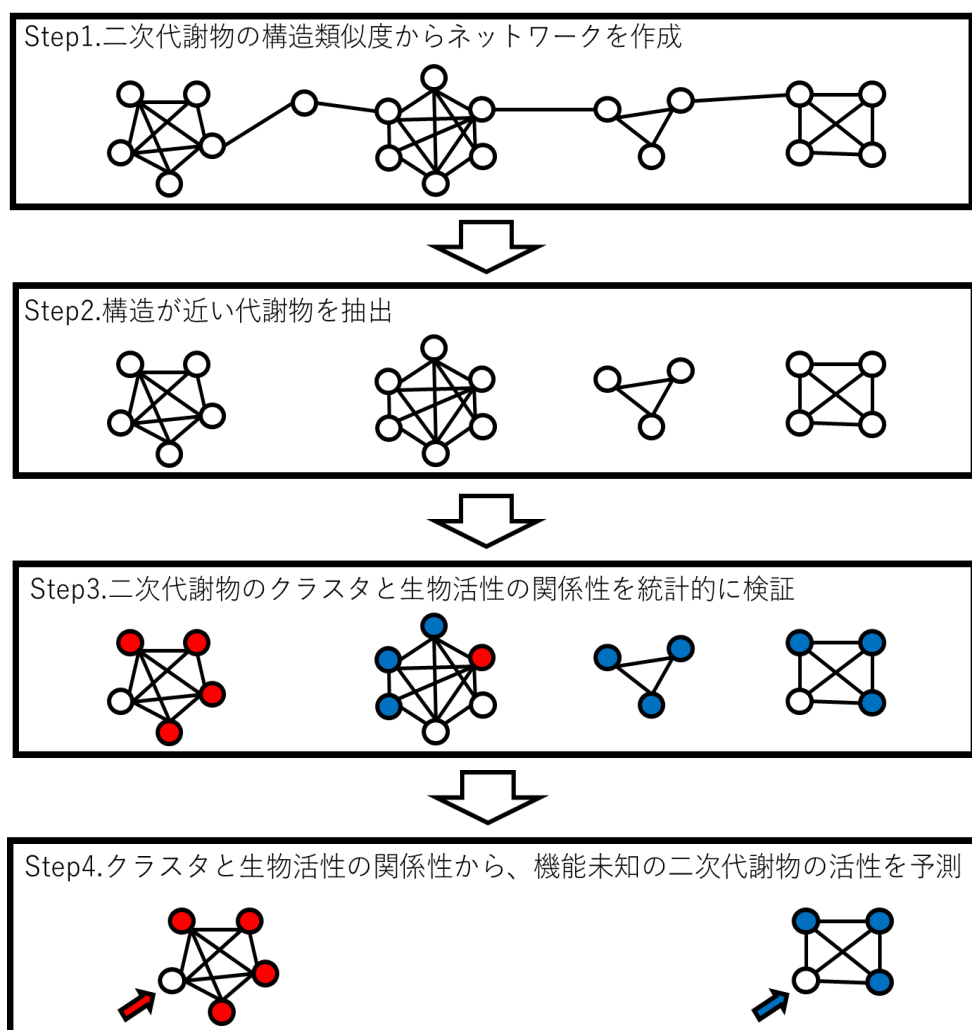


図 1. 研究手法の概要

2-1 データセット

本研究で利用したデータは、代謝物データベースである KNApSAcK Family データベース (http://kanaya.naist.jp/KNApSAcK_Family/) から取得した。KNApSAcK データベースの Core System から 50,037 件の代謝物の 3 次元分子構造情報を取得した。加えて、KNApSAcK Metabolite Activity データベースから 3,210 件の代謝物の情報と、これらの代謝物が持つ生物活性の情報を取得した。これらの生物活性情報は、KNApSAcK Metabolite Activity データベース内で "Activity categories" として扱われている 155 種類の生物活性情報を取得した。3,210 件の代謝物と 155 件の生物活性の間の対応付け情報は 6,649 件得られた。

2-2 代謝物の分子構造類似度に基づくネットワークの構築(Step 1)

代謝物をノードとして表現し、2 代謝物間の分子構造が 95% 以上類似している代謝物ノードの間にエッジを引くことにより代謝物ネットワークを構築した。

代謝物間の分子構造の類似度の計算には、高速グラフマッチアルゴリズムである COMPLIG (Saito *et al.*, 2012) を用いた。2 つの分子間の類似性の計算のために広く用いられる方法としては、各分子の分子フィンガープリントを求め、そのフィンガープリントの類似性を比較することで分子間の類似度を求める方法が広く使われている。ここで分子フィンガープリントとは、各分子に対して複数種類の特定の分子構造を持つか否かを調べ、分子構造が見つかった場合に 1、見つからなかった場合に 0 を記述することで 0 と 1 の羅列という形で分子を表現する方法である。また、分子フィンガープリントの類似性の尺度には tanimoto 係数や Jaccard 係数などが用いられる。分子フィンガープリントを求め、その類似性を比較する方法では分子構造の直接的な比較はできないが、COMPLIG を用いることで 2 分子間の原子配置や原子間の結合情報を入力とした、直接的かつ 3 次元的な分子比較が可能である。また、COMPLIG はグラフマッチアルゴリズムなので、比較する分子のサイズが大きい場合には計算時間が大きくなってしまいうために適さないが、本研究で扱う代謝物は小分子であるため、この問題

は発生しない。上記の理由により、本研究では分子構造の比較のために COMPLIG を用いた。

COMPLIG は Protein Data Bank(PDB)に登録されている低分子リガンドに対して体系的な構造分類を行う際に使用されたアルゴリズムである。COMPLIG は入力として分子を構成する各原子の 3 次元座標上の位置と、どの原子とどの原子が結合している、といった結合情報を使用する。そして 3 つの要素(結合距離、結合次数、回転可能結合か否か)に基づく対応付けを行うことにより、2 分子間の構造の類似度を算出する。この類似度は、式(1)において $M(A,B)$ に対応する。これは 2 つの分子を重ね合わせたときに、3 次元座標上の同じ位置に配置される原子と結合の総数を表す。このように 2 つの分子間の 3 次元的な比較が可能である。

2 つの代謝物間(代謝物 A と代謝物 B)の構造類似度スコア $S(A,B)$ は式(1)により求めた。

$$S(A,B) = \frac{M(A,B)}{\max\{N(A),N(B)\}} \quad (1)$$

ここで、 $M(A,B)$ は代謝物 A と代謝物 B を重ね合わせた際に同じ位置に存在し、重なり合う原子と結合の総数を示す。 $N(A)$ は代謝物 A を構成する原子数と結合数の総数を示す。 $N(B)$ は代謝物 B を構成する原子数と結合数の総数を示す。 $N(A), N(B)$ のより大きい方で $M(A,B)$ を割ったものが構造類似度スコア $S(A,B)$ となる。

ノード間にエッジを引く際に基準となる類似度スコア S の閾値を決定するために、データセットの全代謝物の全てのペアに COMPLIG を適用して得られた累積度数分布を参考にした。本研究では類似度スコア 95% を閾値とした。

構造類似度が 95% 以上の代謝物ペアの中から、生物活性情報が未知の代謝物同士の代謝物ペアを取り除いた。選択された構造類似度 95% 以上の代謝物ペアの中には生物活性情報が未知の代謝物同士のペアが多く含まれている。これにより後のネットワーククラスタリングの際に生物活性が未知の代謝物のみのクラスタが多数作り出される。このことにより活性予測が困難になる可能性があるためにこのような処理を行った。

代謝物をノードとして、構造類似度 95% 以上で類似している代謝物間にエッ

ジを引くことで代謝物ネットワークを構築した。このときノードとエッジに重みを定義しないので、ネットワークは重みなしグラフとなる。

2-3 ネットワークからのクラスタ抽出(Step 2)

Step1 で構築した代謝物ネットワークから高密度の部分、つまり分子構造が似ている代謝物が集まっている部分をクラスタとして抽出するために、ネットワーククラスタリングアルゴリズムである DPCLUSO(Md. Altaf-Ul-Amin *et al.*, 2012)を用いた。DPCLUSO は DPCLUS アルゴリズム(Md. Altaf-Ul-Amin *et al.*, 2006)に似た概念を用いてオーバーラッピングクラスタリングを行うために開発された。このアルゴリズムは酵母のタンパク質間相互作用ネットワークのクラスタリングを行う際に用いられている。

DPCLUSO の特徴は 3 つある。

1. オーバーラップは、複数のクラスタに同一のノードが出現することを許すということを意味する。
2. DPCLUSO の適用によって同一の内容のクラスタが複数出力されることはない。
3. ネットワークに含まれる各ノードが少なくとも 1 つのクラスタに分類される。

DPCLUSO によるクラスタリングのアルゴリズムについて概要を述べる。

DPCLUSO は設定された密度(density)とクラスタの周辺状況(cluster property)に基づいてクラスタリングを行う。本研究ではネットワークから高密度の部分のクラスタとして抽出するためにパラメータに density = 0.9, cluster property = 0.5 を設定した。

DPCLUSO の入力は無向グラフの隣接リストであり、この入力はネットワーク G と G' として保存される。

1. シードノードを求める際に、一度だけネットワーク内の全てのエッジの重みを計算する。ここでエッジの重みは、エッジの両端の 2 つのノードに共通する隣接ノードの数である。
2. 続いてネットワーク内の全てのノードの重みを計算する。ここでノード

の重みは、各ノードに隣接するエッジの重みの合計値である。

3. シードノードには最も大きい重みをもつノードが選択される。ただし、ネットワーク内の全てのノードの重みが0の場合は、各ノードの結合次数によりシードが決定される。
4. シードと結合している隣接ノードの中から優先ノードの選択を行う。優先ノードは、次の i)、ii)をもとに選択される。
 - i) 優先ノードには、隣接ノードの中からシード(またはクラスタ内の各ノード)とのエッジ重みの合計値が最も大きいノードが選択される。
 - ii) ii 複数のノードが最大かつ同じ重みをもつ場合は、シード(クラスタ)との結合数が最も多い隣接ノードが優先ノードとなる。このとき、結合次数が最大のノードが複数存在する場合、その中からランダムに優先ノードを選択する。
5. シードの決定後は、次の2つの条件を満たす限り、隣接ノードの中から優先ノードを再帰的に追加していく。
 - (1)優先ノードをクラスタに追加しても、クラスタ k の密度 d_k は d_{in} より小さくならない。
 - (2)優先ノードを追加しても、クラスタ・プロパティ cp_k が cp_{in} 以上である。ここで d_{in} と cp_{in} は DPCLUSO の起動時にユーザーが設定するパラメータであり、それぞれ密度とクラスタ・プロパティを示す。 d_k と cp_k はそれぞれ式(2)と式(3)により求められる。

$$d_k = \frac{|E_k|}{|E_k|_{max}} = \frac{2 \times |E_k| |N_k|}{|E_k| (|N_k| - 1)} \quad (2)$$

$$cp_k = \frac{|E_{nk}|}{d_k \times |N_k|} \quad (3)$$

6. ここで $|E_k|$ はクラスタ k 内に存在するエッジの数を示し、 $|E_k|_{max}$ はクラスタ k 内に引くことができる最大のエッジ数を示し、 $|N_k|$ はクラスタ k 内に存在するノードの数を示し、 $|E_{nk}|$ はノード n とクラスタ k の各ノードの間のエッジの総数を示す。DPCLUSO は、初めに読み込んだネットワークを変数 G と G' に格納する。クラスタへの優先ノードの追加は、初め

にネットワーク G' について探索される。 G' 内に追加できる優先ノードがなくなった場合、次にネットワーク G についての探索が実行される。ネットワーク G' と G の両方に追加できるノードがなくなったらクラスタを保存し、両方のノードがクラスタに属するエッジをネットワークから削除したものをネットワーク G' に保存する。このようにシードの選択、シードに隣接するノードから優先ノードの選択、クラスタへの優先ノードの再帰的な追加、クラスタの保存とネットワーク G' の更新の手順をネットワーク G' 内にエッジがなくなるまで繰り返す。

図 2 に DPCLUSO アルゴリズムによるクラスタリングの例を示す。

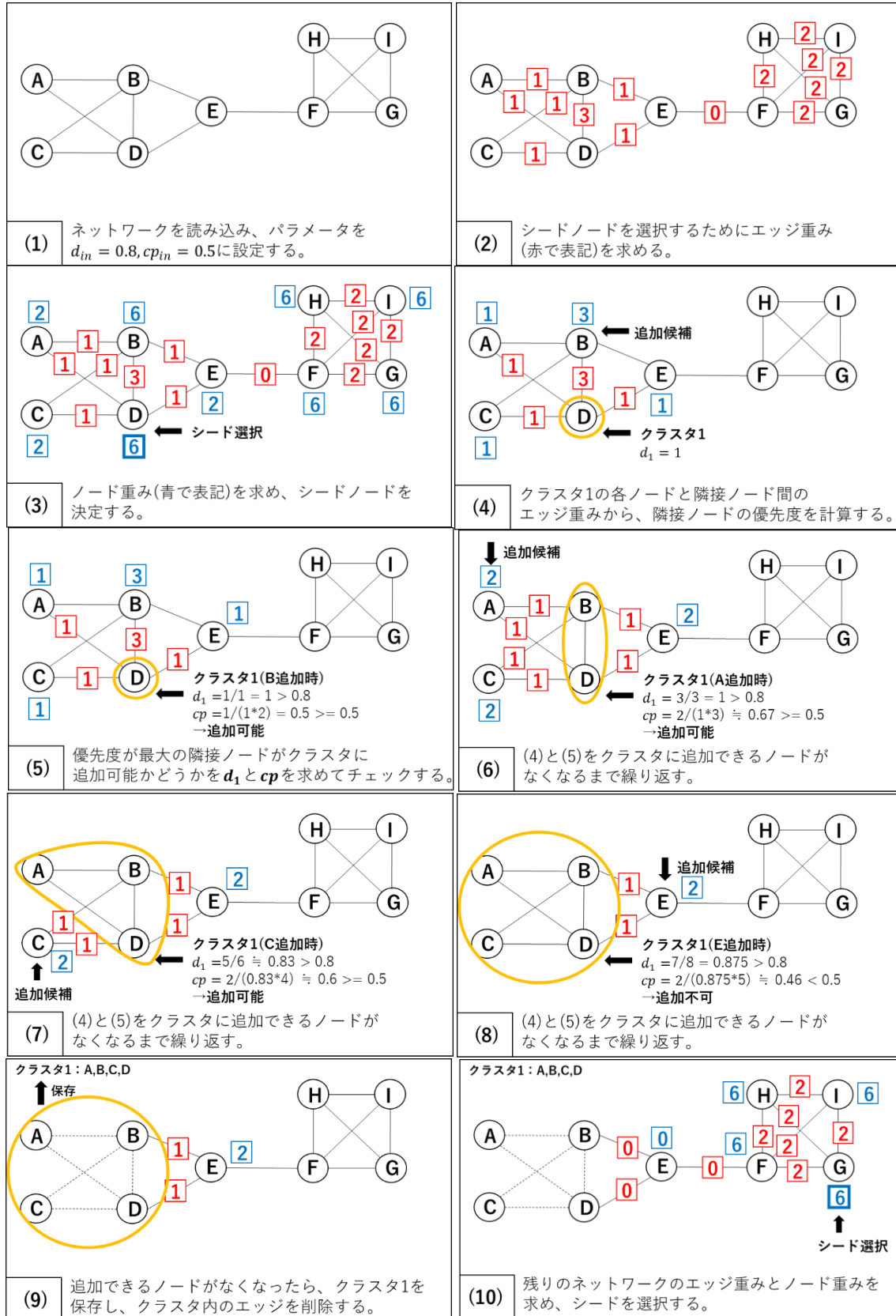


図 2. DPCLUSO のアルゴリズムの概要

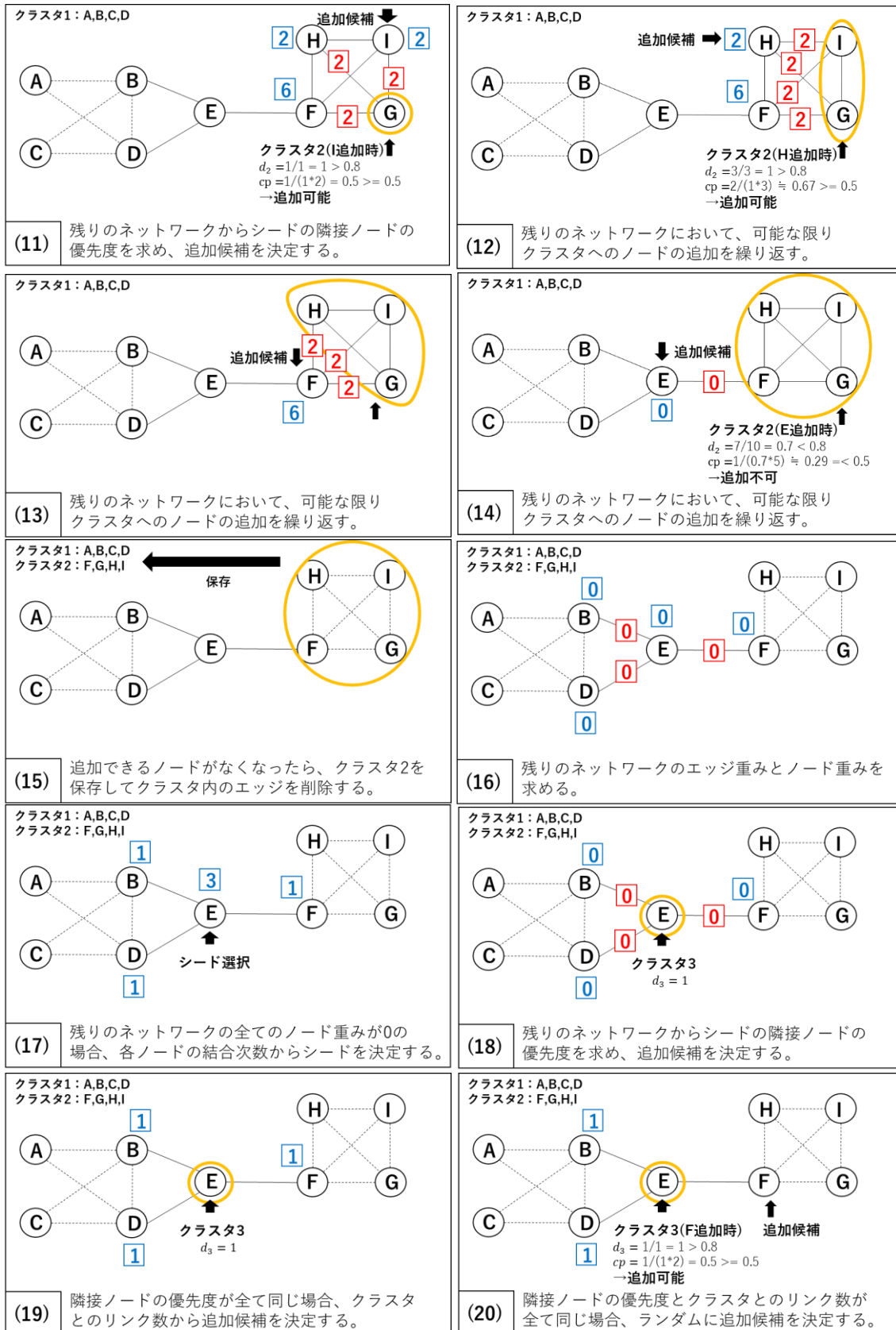


図2. DPCLUSO のアルゴリズムの概要(続き)

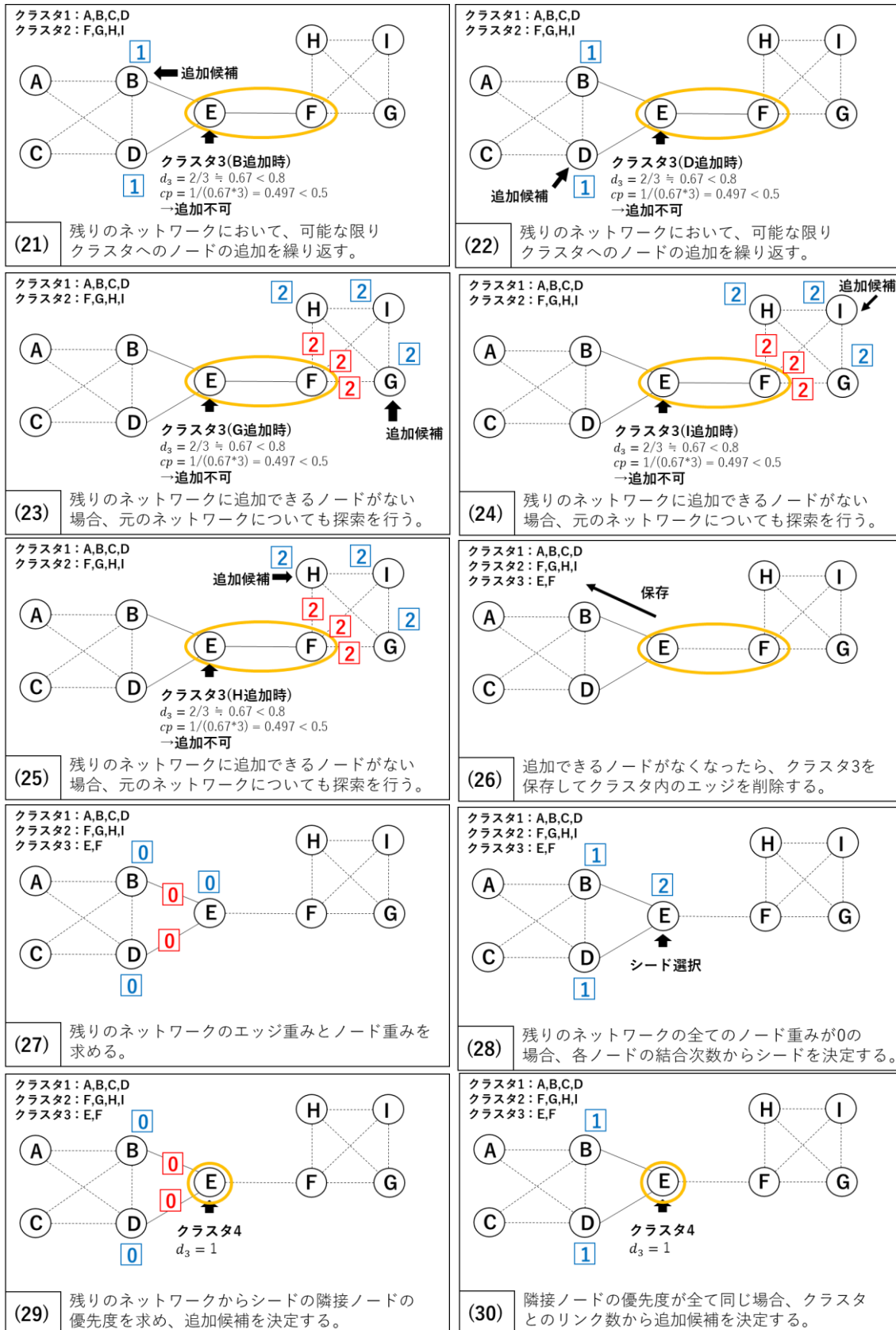


図2. DPCLUSO のアルゴリズムの概要(続き)

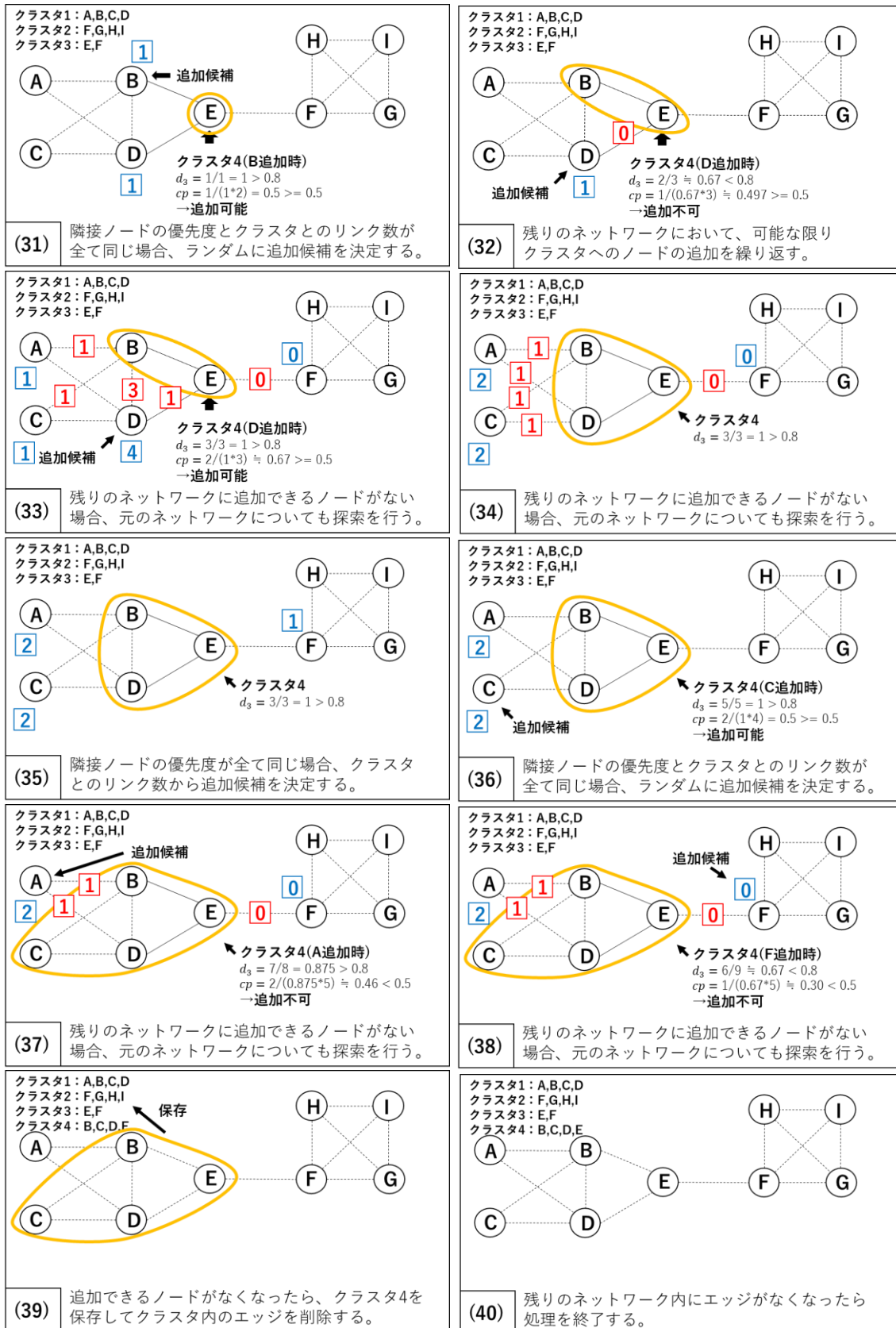


図 2. DPCLUSO のアルゴリズムの概要(続き)

2-4 クラスタと生物活性間の関係の統計的な評価(Step 3)

代謝物ネットワークから得られた各クラスタに対して 155 種類の生物活性との関係について統計的な評価を行った。特定の生物活性を持つ代謝物が有意に集まっているクラスタを選択するために、帰無仮説を「クラスタ 1 と生物活性 1 の間には関連性がない」、対立仮説を「クラスタ 1 と生物活性 1 は有意に関連している」という検定を全てのクラスタ、全ての生物活性に対して行う。DPCLUSO により得られた各クラスタとデータセットに含まれる 155 種の生物活性の間の p 値を求めた。 p 値の計算には超幾何分布を用いた。超幾何分布を用いた p 値の計算式を式(4)に示す。

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{F}{i} \binom{N-F}{C-i}}{\binom{N}{C}} \quad (4)$$

ここで、 F はネットワーク内で特定の生物活性を持つ代謝物の数を示し、 N はネットワーク内の代謝物の総数を示し、 C はクラスタ内の代謝物の数を示し、 k はクラスタ内で特定の生物活性を持つ代謝物の数を示す。

このとき、各クラスタに対してネットワーク内に現れる各生物活性との検定を行うことにより False Positive(偽陽性)が大きくなる多重比較の問題が生じる可能性がある。そのため、 p 値に関するいくつかの閾値に対して False Discovery Rate(FDR)を計算し、この FDR を基準として最適な p 値の閾値を選択した。

2-5 生物活性が未知の代謝物の活性予測(Step 4)

有意な関係性を持つクラスタ-生物活性のペアの中から、クラスタ内の代謝物の 40%以上が対応する生物活性を持っているクラスタ-生物活性ペアを選択し、そのクラスタ内に存在する機能未知の代謝物とクラスタに対応する生物活性と対応付けることにより、生物活性の予測を行う。

2-6 予測手法の精度の評価

本研究における 4 つの Step からなる活性予測手法の予測精度の確認を行った。Step2 の DPCLUSO を実行した後、Step3 の p 値を求める前にネットワーク内に存在する活性が既知の代謝物のうち 10% をランダムに選択し、活性が未知のものとして扱う。この操作により p 値の計算時に、ネットワーク内で特定の生物活性を持つ代謝物数 F とクラスタ内で特定の生物活性を持つ代謝物数 k が小さくなる。このように、ネットワーク内の活性情報を減らした状態で活性予測を行い、ランダムに選択した各代謝物に対して、生物活性を 1 つでも正しく予測出来たなら 1、1 つも正しく予測できなかったなら 0 をスコアとして割り当てた。1 つの代謝物が複数の生物活性を持つ場合、そのうちの 1 つでも正しく予測できたなら 1 をスコアとして割り当てる。ランダムに選択した活性が未知のものとして扱っている代謝物のうち、生物活性と有意な関係性を持つクラスタ内に含まれているものの総数でスコアの合計値を割ったものを予測精度とした。このようなランダム選択、活性予測、予測精度の確認を 10 回繰り返し、各予測精度の平均を最終的な予測精度とした。

2-7 生物活性間の関係性の調査

生物活性が既知の代謝物が本来持つ活性と、これらの代謝物に対する活性予測の結果を比較することで生物活性間の関係性を調べる。データセットの生物活性既知の代謝物から 10% をランダムに選択して活性未知として扱い、本研究の方法によりこれらの代謝物の活性予測を行う。これにより活性既知の代謝物に対する活性予測の結果を得て、代謝物の本来の生物活性と予測された生物活性との比較を行う。

3. 結果と考察

3-1. 代謝物の分子構造類似度に基づくネットワークの構築

代謝物ネットワークを構築するために、KNApSAcK Core データベースから得られた 50,037 件の代謝物間の分子構造の比較を行い、分子構造の類似度スコアを算出した。この分子構造の類似度スコアの算出には COMPLIG を用いた。その結果、1,251,739,702 件の構造類似度スコアが得られた。この中から構造類似度スコア 95%以上を閾値として代謝物ペアを 177,179 対選択した。閾値となる構造類似度(%)以上の構造類似度を持つ代謝物ペア数の遷移を図 3 に示す。

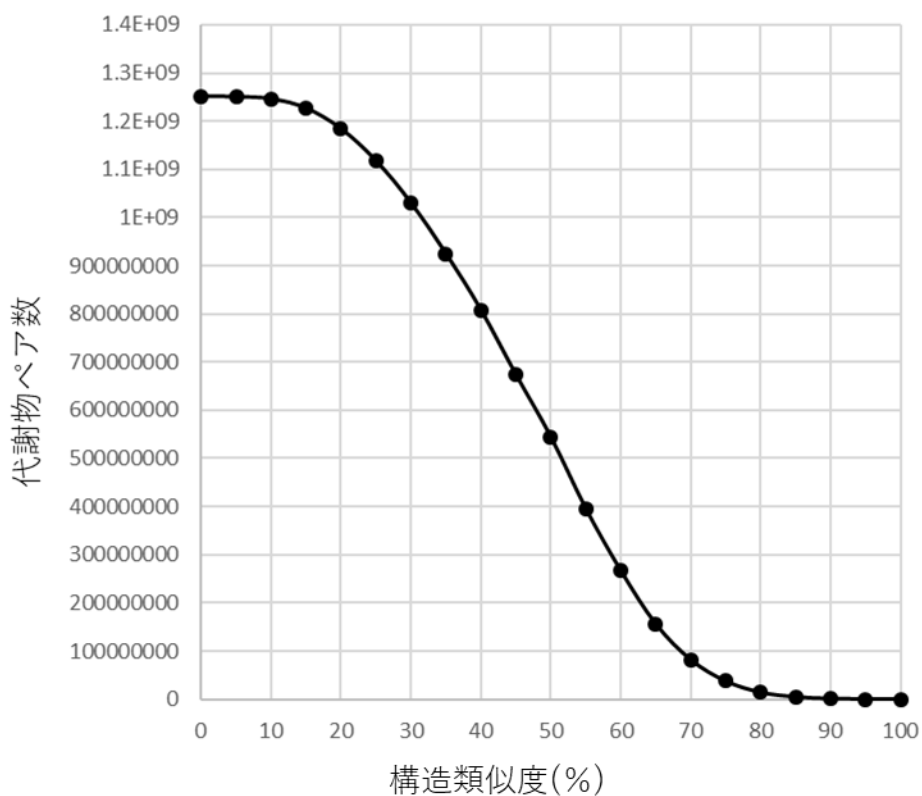


図 3. 各閾値以上の構造類似度を持つ代謝物ペア数の遷移

各構造類似度(%)を閾値としたときの、閾値以上の構造類似度を持つ代謝物ペアの数をプロットした。70%以上の構造類似度を持つ代謝物ペアは 81,984,992 件、80%以上の構造類似度を持つ代謝物ペアは 14,374,757 件、90%以上の構造

類似度を持つ代謝物ペアは 1,379,238 件存在した。本研究では分子構造の類似度 95%を閾値として COMPLIG により得られた比較結果全体の 0.00014%である 177,179 組の代謝物ペアを選択した。

構造類似度が 95%以上の代謝物ペアの中から、生物活性情報が未知の代謝物同士の代謝物ペアを取り除いた。その結果、10,773 件の代謝物を含む代謝物ペアが 22,482 組選択された。

これらの代謝物ペアから代謝物ネットワークを構築した。このネットワークでは代謝物はノードとして表現され、構造類似度が 95%以上の代謝物ノード同士がエッジでつながれる。得られた代謝物ネットワークの概観を図 4 に示す。

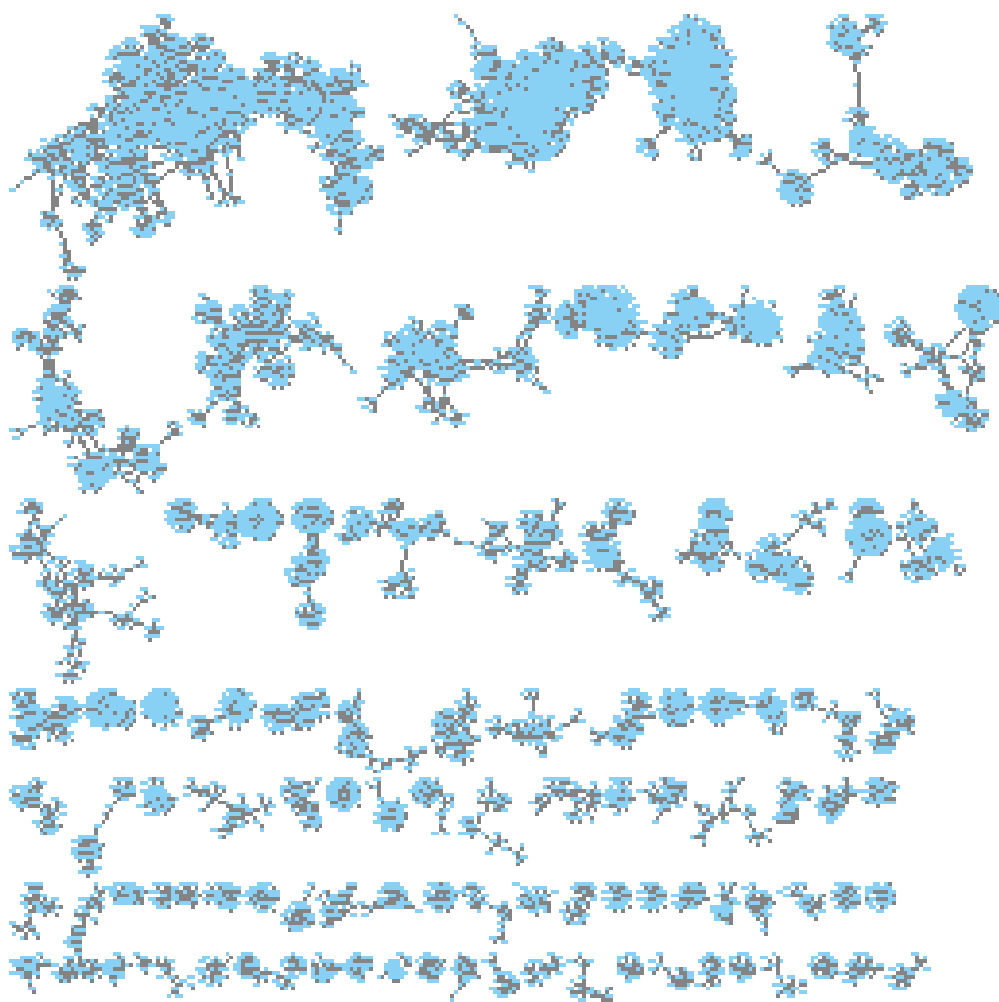


図 4. 構築された代謝物ネットワークの概観

代謝物ネットワークの可視化は Cytoscape (P. Shannon *et al.*, 2003)により行った。得られたネットワークはひとつつながりの巨大なものではなく、複数の島のように分かれたネットワークだった。

3-2. ネットワークからのクラスタ抽出

代謝物ネットワークからクラスタの抽出を行った。このクラスタリングには、ネットワーククラスタリングアルゴリズム DPCLUSO を用いた。DPCLUSO はネットワーク内の高密度の部分クラスタとして抽出するアルゴリズムである。この代謝物ネットワークにおいては分子構造が似ている代謝物がクラスタとしてまとめられる。DPCLUSO を適用する際のパラメータは $d_{in} = 0.9$, $cp_{in} = 0.5$ に設定した。その結果、代謝物ネットワーク内で高度に関連する代謝物をまとめたクラスタが 8,047 件得られた。この時点でネットワーク内に存在する生物活性は 153 種類、代謝物は 9740 種類だった。

3-3. クラスタと生物活性間の関係の統計的な評価

得られた各クラスタに対して、同一の生物活性を持つ代謝物が有意に集まったクラスタであるかどうかを確認するために超幾何分布により各クラスタ-生物活性間の p-value を求めた。このとき、得られた各クラスタ 8,047 件のそれぞれに対して生物活性 153 種類との関連性を評価することになる。これにより多重比較の問題が発生する可能性があるため、FDR(累積 q-value)を検討した。FDR の算出には、クラスタ-生物活性ペア 24,296 対の p-value を用いた。p-value の 0 から 1 の範囲を 0.001 区切りで 1,000 の区間に分割した。このときの p-value ごとのクラスタ-生物活性ペアの頻度の一部を図 5 に示す。それぞれの区間においてランダムにクラスタ-生物活性ペアをカウントした場合、24,296 件 ($24,296/1000$)のクラスタ-生物活性ペアが現れることになる。このことから、p-value が 0 から 0.001 の区間に現れる偽陽性の数は 24,296 件 ($24,296 \times 0.001$)と推定される。同様にして p-value が 0 から 0.002 の区間に現れる偽陽性の数は

48.592($24,296 \times 0.002$)である。p-value < 0.001 のときの FDR は Benjamini-Hochberg 法(Benjamini *et al.*, 1995)により 7.21×10^{-3} と推定された。p-value < 0.002 のときは $FDR = 0.10 \times 10^{-2}$ であり、p-value < 0.003 のときは $FDR = 0.11 \times 10^{-2}$ となった。本研究では FDR が 0.05 未満となるクラスタ-生物活性ペア、すなわち 0.046 未満の p-value を持つクラスタ-生物活性ペアを 22,428 対選択した。

クラスタ-
生物活性ペア数

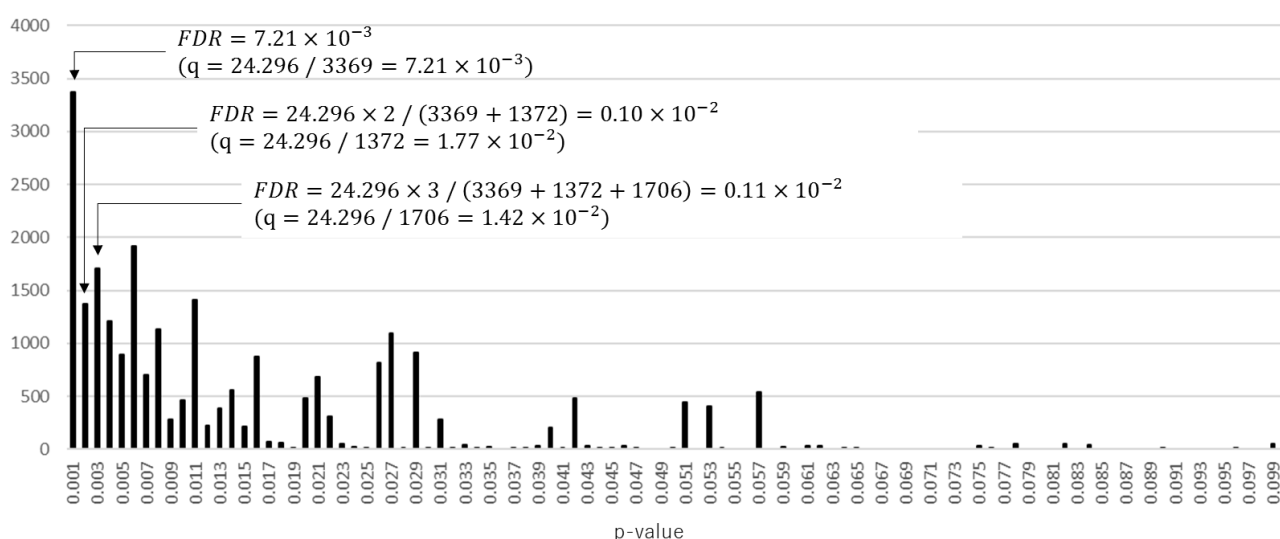


図 5. p-value の区間ごとのクラスタ-生物活性ペアの出現頻度

p-value を 0 から 0.001 ごとに区切り、0 から 0.1 までの p-value を持つクラスタ-生物活性の出現頻度を示す。

より豊富な情報に基づいて活性予測を行うために、FDR により選択されたクラスタ-生物活性ペアの中からさらにデータを選択を行った。FDR に基づき選択されたクラスタ-生物活性ペアのうち、クラスタに対応する生物活性を持つ代謝物がクラスタ内の 40%以上を占めるデータを選択した。例えば、クラスタ 1 が生物活性”Flavor”と有意に関係しており、かつクラスタ 1 に 12 の代謝物が含まれている場合を考える。この場合、クラスタ 1 に生物活性”Flavor”を持つ代謝物が 5 種類以上含まれていれば選択され、4 種類以下ならクラスタ 1 と生物活性”Flavor”とは有意な関係性として選択されない。このようなデータ選択を行っ

た結果、13,855 件のクラスタ-生物活性ペアが予測結果として選択された。

3-4. 生物活性が未知の代謝物の活性予測

得られた 13,855 件のクラスタ-生物活性データから活性予測を行った。そのために、クラスタ内の活性情報が未知の代謝物とクラスタに結び付けられた生物活性情報の対応付けを行った。対応付けの例を挙げると、活性が既知の代謝物 8 種類と活性が未知の代謝物 4 種類を含む、クラスタサイズが 12 のクラスタ 1 と生物活性”Flavor”が有意に関係していた。この場合、クラスタ 1 に含まれる活性が未知の代謝物 4 種類が生物活性”Flavor”を持つと予測される。このようにして活性予測を行った結果、6,480 件の活性が未知の代謝物の生物活性を予測することができた。これらの代謝物は、149 種類の生物活性と結びつけられた。

KNAPSAcK データベースから取得したデータ内の生物活性の出現頻度と予測結果における生物活性の出現頻度を比較した。KNAPSAcK データベース内の生物活性の出現頻度を図 6 に示す。また、予測結果における生物活性の出現頻度は図 7 に示す。

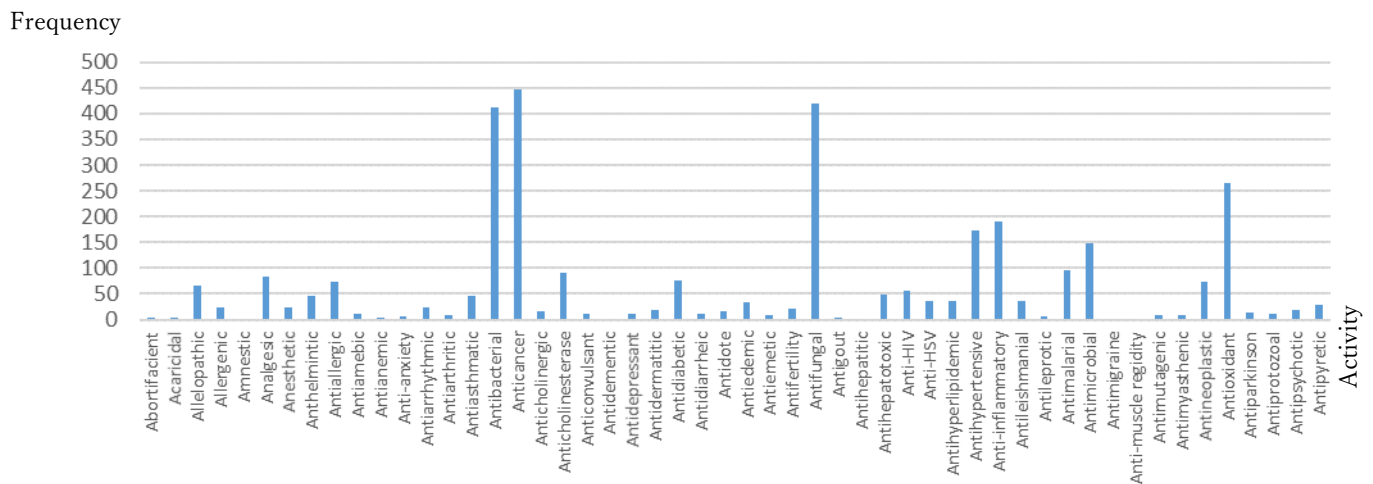


図 6. KNApSacK データベース内の生物活性の出現頻度

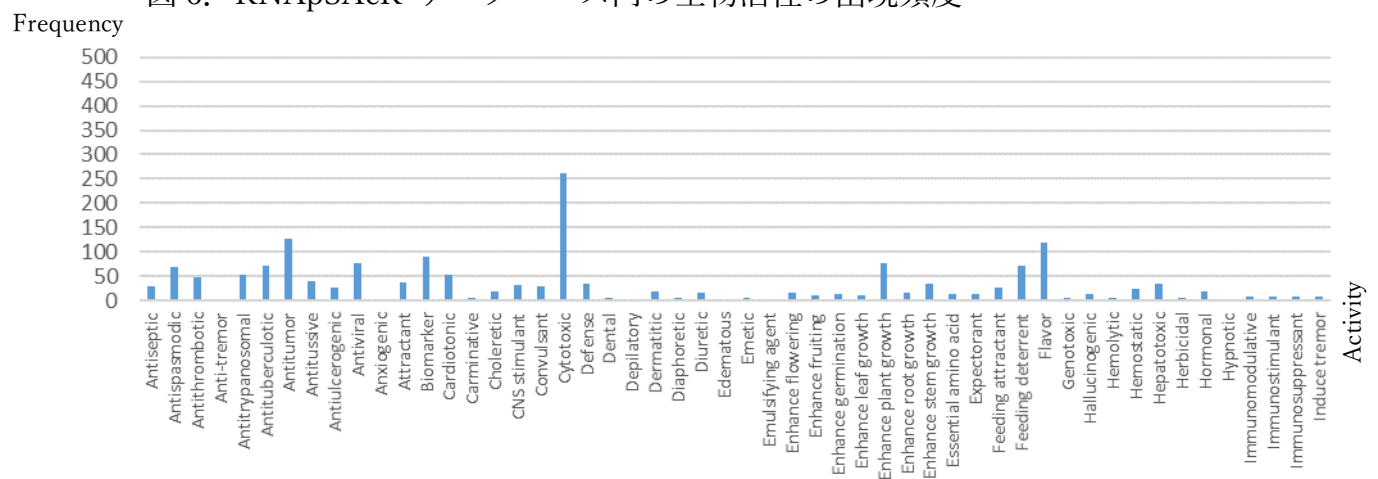


図 6. KNApSacK データベース内の生物活性の出現頻度(続き)

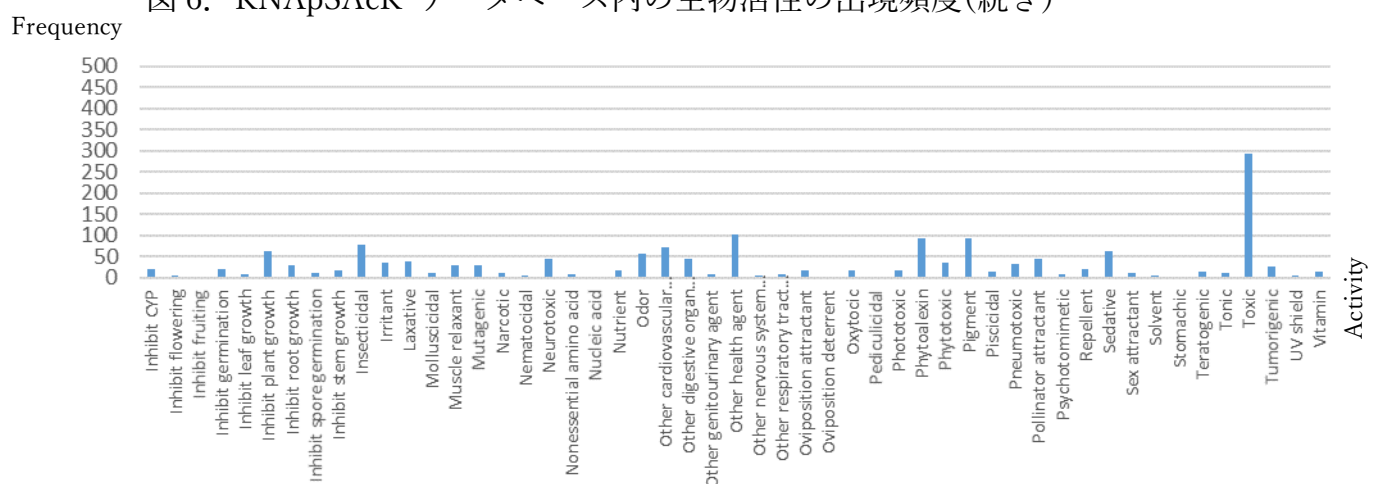
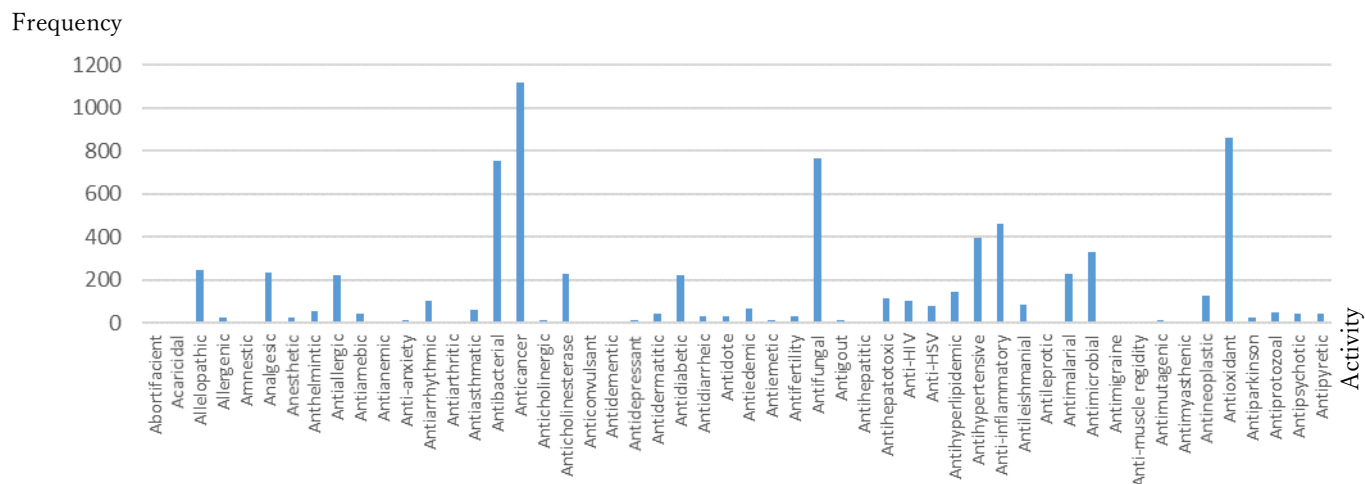
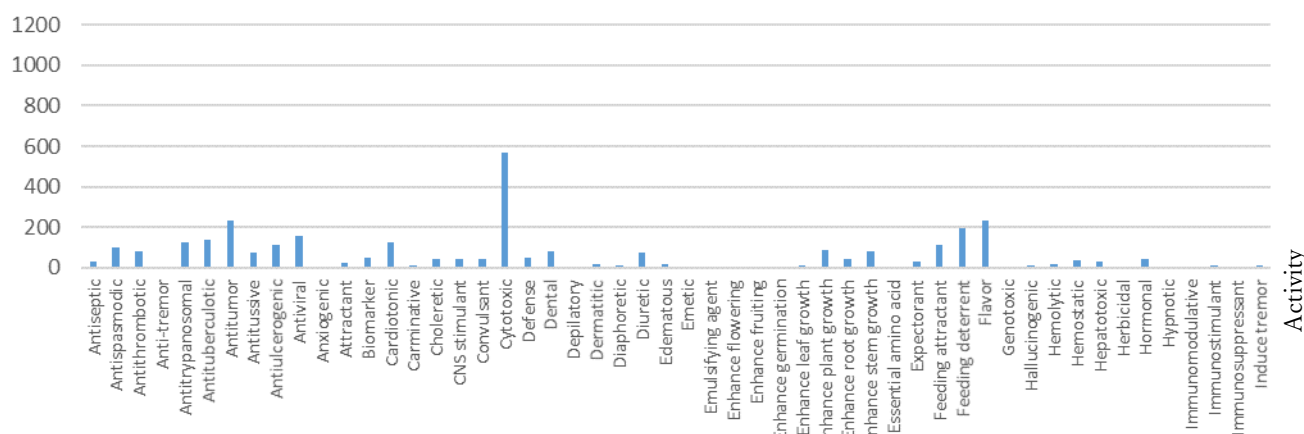


図 6. KNApSacK データベース内の生物活性の出現頻度(続き)



Frequency 図 7. 本研究の予測結果における生物活性の出現頻度



Frequency 図 7. 本研究の予測結果における生物活性の出現頻度(続き)

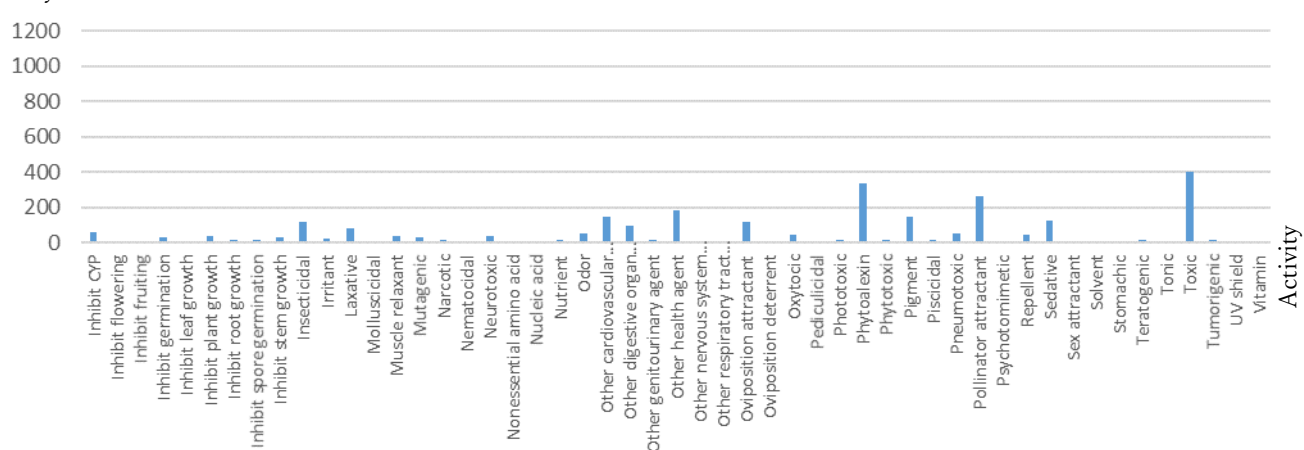


図 7. 本研究の予測結果における生物活性の出現頻度

KNApSAcK データベース内に高い頻度で出現する生物活性には Anticancer(447 件), Antifungal(420 件), Antibacterial(447 件), Toxic(293 件), Antioxidant(264 件), Cytotoxic(262 件)が見られた。また、予測結果において特に高い頻度で出現する生物活性には Anticancer(1,115 件), Antioxidant(860 件), Antifungal(766 件), Antibacterial(756 件), Cytotoxic(570 件)が見られた。KNApSAcK データベース内のデータと本研究の予測結果に共通の生物活性が頻繁に現れていることから、本手法による活性予測はある程度データベース内の生物活性の頻度の影響を受けていると考えられる。

3-5. 予測手法の精度の評価

本研究における活性予測手法の予測精度の確認を行った。Step2 の DPCLUSO を実行した後、Step3 の hyper geometric p-value を求める前にネットワーク内に存在する活性が既知の代謝物 2,130 件のうち 213 件をランダムに選択し、活性が未知のものとして扱った。このようにしてネットワーク内の活性情報を減らした状態で活性予測を行った。ランダムに選択した各代謝物に対して、生物活性を 1 つでも正しく予測出来たなら 1、1 つも正しく予測できなかったなら 0 をスコアとして割り当てた。ランダムに選択した活性が未知のものとして扱っている代謝物のうち、生物活性と有意な関係性を持つクラスタ内に含まれているものの総数でスコアの合計値を割ったものを予測精度とした。このランダム選択から予測精度の算出を 10 回繰り返した結果、平均予測精度は 71.934%だった。

3-6. 生物活性間の関係性について

代謝物が持つ既知の生物活性と、その代謝物に対して予測される生物活性の関係について調べた。生物活性が既知の代謝物のうち 10%を選択して活性未知として扱い、これらの代謝物の活性予測を行った。表 1 は活性既知の代謝物が持つ生物活性と本研究の方法による予測状況を示す。表 1 の縦軸は、活性既知の代謝物の本来の生物活性を示す。横軸は、縦軸に示されている生物活性を持つ代謝物に対して予測された生物活性を示す。ただし横軸の”Real”の列は縦軸に示される各生物活性を本来持っている代謝物の数を示す。表内の赤い数字は縦軸に示されている生物活性を予測することができた回数を示す。”Real”列以外の表内の数値は、縦軸に示される各生物活性を持つ代謝物に対して、横軸の各生物活性が予測された回数を示す。一例を挙げると、1 行目の生物活性 Allelopathic を持つ代謝物は 5 件だった(”Real”列に示されている)。活性予測の結果、生物活性 Allelopathic を本来持つ代謝物 5 件に対して、生物活性 Allelopathic が予測された代謝物数は 5 件だった(赤い数字)。このとき生物活性 Allelopathic を持つと予測された 5 件の代謝物のうち、同時に他の生物活性である Allergenic, Anticancer, Anticholinesterase などを併せ持つと 1 回予測された。Antidiabetic, Antifungal, Antihepatotoxic などを併せ持つと 2 回予測された。Antibacterial, Anti-inflammatory, Other cardiovascular agent などを併せ持つと 3 回予測されたことが示されている。各生物活性ごとの予測の割合を表 2 に示す。この表の縦軸と横軸は表 1 と同じものを示す。ただし表 2 の数値は、縦軸に示される各生物活性を持つ代謝物数で横軸に示されている予測結果における生物活性の出現回数を割った割合(%)として示されている。

[illegible]

表 1. 予測精度評価の際の各生物活性の予測回数(続き)

	Real	Allelopathic	Allergenic	Analgesic	Anesthetic	Anthelmintic	Antiallergic	Anti-anxiety	Antiarrhythmic	Antiarthritic	Antisthmatic	Antibacterial	Anticancer	Anticholinergic	Anticholinesterase	Anticonvulsant	Antidepressant	Antidiabetic	Antidote	Antiedemic	Antifertility	Antifungal	Antigout	Antihepatotoxic	Anti-HIV	Anti-HSV	Antihyperlipidemic	Antihypertensive	Anti-inflammatory	Antileishmanial	Antileprotic	Antimalarial	Antimicrobial	Antimutagenic	Antineoplastic	Antioxidant	Antiparkinson	Antiprotzoal	Antipsychotic	Antipyretic	Antiseptic	Antispasmodic	Antithrombotic	Anti-tremor	Antitypanosomal	Antituberculous	Antitumor	Antitussive	Antulcerogenic	Antiviral	Attractant	Biomarker	Cardiotonic	Carminative	Choleretic	CNS stimulant		
Allelopathic	5	100	20	-	-	-	20	-	-	-	-	60	20	-	20	-	-	40	-	20	20	40	-	40	20	-	40	-	60	-	-	-	20	-	40	20	-	-	-	-	40	-	-	-	20	20	20	40	-	-	40	-	-	-				
Allergenic	2	50	50	-	-	-	100	-	-	-	-	100	50	-	50	-	-	100	-	50	-	50	-	50	100	-	50	-	100	-	-	-	50	-	50	-	-	-	-	-	50	-	-	-	50	50	-	50	-	-	50	-	-	-				
Analgesic	6	17	-	33	-	-	17	-	17	17	-	67	17	-	17	-	-	33	-	-	-	17	-	33	-	-	17	50	17	-	17	17	17	-	33	-	17	-	17	-	-	-	-	17	-	-	17	17	-	17	17	-	-	-				
Anesthetic	2	-	-	-	50	-	-	-	-	-	-	-	-	-	-	50	-	-	50	-	-	-	-	50	-	-	-	-	-	-	-	-	-	-	-	-	50	50	-	50	-	-	-	-	-	-	-	-	-	-	50	-	-					
Anthelmintic	2	-	-	50	-	-	-	-	50	-	-	50	-	-	50	-	-	-	-	-	-	100	-	-	-	-	-	50	-	-	-	-	-	-	-	-	50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-					
Antiallergic	5	-	-	20	-	-	40	-	-	-	-	20	20	-	20	-	40	20	-	-	-	-	-	40	-	-	-	-	-	-	-	20	-	20	-	-	-	-	-	20	20	20	20	-	20	-	20	20	20	-	20	20	-	-				
Anti-anxiety	1	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-					
Antiarrhythmic	3	-	-	-	-	-	33	-	-	-	-	-	33	-	-	-	-	33	-	-	-	-	-	-	33	-	33	33	-	33	33	-	33	33	-	-	-	-	33	-	-	33	33	-	33	33	-	33	-	-	-	-	-					
Antiarthritic	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	-	100	-	-	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
Antisthmatic	2	50	50	50	-	-	100	-	-	-	50	50	-	-	50	-	-	100	-	50	-	50	-	100	50	50	50	50	100	50	-	50	-	-	-	100	-	-	-	-	50	-	50	50	-	50	50	-	100	-	-	50	-	-				
Antibacterial	27	26	3.7	11	-	-	11	-	3.7	-	3.7	44	15	-	11	7.4	-	19	3.7	3.7	3.7	26	-	19	3.7	3.7	11	11	22	11	-	11	3.7	7.4	-	22	3.7	3.7	3.7	7.4	3.7	15	3.7	3.7	7.4	-	3.7	7.4	19	-	3.7	11	7.4	-	-			
Anticancer	28	14	3.6	14	-	-	18	-	3.6	-	3.6	36	50	-	7.1	7.1	-	14	3.6	7.1	-	25	-	18	11	3.6	3.6	11	25	3.6	-	7.1	14	7.1	3.6	21	-	3.6	3.6	7.1	3.6	11	3.6	-	3.6	7.1	11	7.1	11	14	-	3.6	7.1	3.6	-	-		
Anticholinergic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-				
Anticholinesterase	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-					
Anticonvulsant	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	-	100	-	-	-	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	100	100	-	100	-	-	-	-	-	-	-	-	-	-	100	-	-				
Antidepressant	1	-	-	-	-	-	-	-	-	-	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-				
Antidiabetic	6	83	17	17	-	-	33	-	-	-	83	17	17	-	-	83	17	33	-	17	-	17	-	83	17	-	17	17	67	33	-	33	-	67	17	50	-	-	-	17	-	17	17	-	33	-	17	17	67	83	-	-	33	-	17	-		
Antidote	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
Antiedemic	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	-	100	-	-	-	-	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	100	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
Antifertility	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
Antifungal	26	12	-	3.8	-	3.8	3.8	-	-	-	-	27	19	-	7.7	3.8	-	-	7.7	-	3.8	58	-	3.8	-	-	7.7	15	7.7	7.7	-	3.8	3.8	-	-	15	3.8	-	-	3.8	7.7	7.7	7.7	-	3.8	-	3.8	3.8	-	-	3.8	7.7	3.8	-	-			
Antigout	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
Antihepatotoxic	5	40	20	-	-	-	20	-	-	-	40	-	20	-	20	-	40	-	20	-	20	-	20	-	60	20	-	20	-	40	-	-	-	20	-	40	-	-	-	-	-	20	-	20	40	40	-	20	20	-	20	-	-					
Anti-HIV	2	-	-	50	-	-	-	-	50	-	50	-	50	-	50	-	-	-	-	-	-	50	-	-	50	-	-	50	-	-	-	-	-	-	-	-	50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
Anti-HSV	2	-	-	-	-	-	-	-	-	-	50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	50	-	50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
Antihyperlipidemic	3	33	-	-	-	-	-	-	-	-	33	33	-	-	-	-	-	-	-	-	33	33	-	-	-	33	33	-	33	-	33	-	-	-	33	100	33	-	-	-	33	-	-	-	-	-	-	33	-	-	33	-	-	-	-			
Antihypertensive	17	18	5.9	5.9	-	-	18	-	5.9	-	53	12	-	12	-	-	18	5.9	5.9	-	18	-	18	5.9	5.9	5.9	47	24	12	5.9	18	12	12	5.9	29	-	5.9	-	5.9	-	18	-	-	12	5.9	18	18	12	24	-	-	24	-	-	-	-	-	
Anti-inflammatory	19	32	11	21	-	-	32	-	5.3	5.3	5.3	53	11	5.3	16	5.3	-	47	11	11	-	16	-	42	11	11	11	21	58	16	-	16	11	21	11	37	-	5.3	-	16	5.3	16	11	5.3	16	5.3	21	16	21	47	-	5.3	11	5.3	5.3	-		
Antileishmanial	4	25	-	25	-	-	25	-	-	-	75	50	25	-	-	25	25	25	-	25	-	25	-	25	-	-	-	25	25	50	-	25	-	25	25	50	-	-	25	-	25	-	25	-	25	-	25	-	25	-	25	-	25	-	-	-	-	
Antileprotic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
Antimalarial	7	43	-	29	-	-	14	-	14	-	57	57	14	14	-	-	14	14	14	14	43	-	14	-	-	29	29	29	14	-	14	-	14	14	43	14	14	-	14	-	14	-	14	-	14	14	14	29	-	-	29	-	14	-	-			
Antimicrobial	5	-	-	40	-	-	40	-	20	-	40	40	-	20	-	-	20	20	-	20	-	20	-	20	-	-	40	40	-	20	20	40	-	20	-	20	-	20	-	20	-	20	-	20	-	20	20	20	-	-	-	20	-	-	-	-		
Antimutagenic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Antineoplastic	5	-	-	-	-	-	40	-	-	-	-	60	-	-	-	-	-	20	20	-	20	-	-	20	20	-	40	40	-	20	40	20	-	20	-	-	-	-	20	-	20	-	20	-	20	40	40	20	-	-	-	-	-	-	-	-	-	
Antioxidant	26	27	3.8	15	-	-	31	-	-	-	3.8	46	35	3.8	3.8	3.8	-	35	7.7	12	3.8	23	-	35	3.8	12	12	7.7	42	12	-	12	7.7	12	7.7	58	3.8	-	-	7.7	3.8	12	7.7	3.8	12	-	7.7	12	12	35	-	3.8	7.7	3.8	3.8	-	-	
Antiparkinson	2	-	-	50	-	-	-	-	-	-	-	-	-	-	50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	50	50	-	-	-	-	50	-	-	-	-	-	-	-	-	-	-	-	-	-

表 2. 各生物活性の出現回数と予測結果における出現回数の比

[illegible]

表2. 各生物活性の出現回数と予測結果における出現回数の比(続き)

[illegible]

表2. 各生物活性の出現回数と予測結果における出現回数の比(続き)

[illegible]

表2. 各生物活性の出現回数と予測結果における出現回数の比(続き)

[illegible]

未知の生物活性として扱われた回数の多い生物活性である Antibacterial, Anticancer, Antihypertensive, Cytotoxic, Toxic, Antifungal, Anti-inflammatory, Antioxidant と、これら 8 件の生物活性を持つ代謝物が同時に持っているとして予測された各生物活性と対応を図 8 に示す。図 8 はマップ上の色が白に近いほど予測結果への出現回数が少なく、赤に近いほど予測結果への出現回数が多いことを示すヒートマップである。

図 8 より、Antibacterial は縦軸に示されている 8 種類全ての生物活性と頻繁に同時に予測される。すなわち、上記の 8 種類の生物活性を持つ代謝物は生物活性 Antibacterial も同時に持っているとして頻繁に予測されている。詳しく見ていくと生物活性 Antioxidant を持つと予測された代謝物 15 件のうち、12 件の代謝物が生物活性 Antibacterial を同時に持っているとして予測されている。また、生物活性 Anti-inflammatory を持つと予測された 11 件の代謝物のうち、10 件が生物活性 Antibacterial を持つと予測されている。このことから、上記の 8 種類の生物活性と Antibacterial の間には何らかの関係性が存在する可能性が考えられる。

生物活性 Anticancer は縦軸に示されている代謝物のうち、生物活性 Cytotoxic と頻繁に同時に予測されており、生物活性 Anticancer を持つ代謝物は同時に生物活性 Cytotoxic を持つと頻繁に予測された。同様に生物活性 Antihypertensive は同時に Antibacterial, Antioxidant, Hemostatic, Pigment, Antiviral を持つ代謝物が多いと予測された。生物活性 Antifungal を持つ代謝物は同時に Antibacterial, Phytoalexin, Toxic を持つと頻繁に予測された。生物活性 Toxic を持つ代謝物は同時に Antibacterial, Anti-inflammatory, Neurotoxic を持つと頻繁に予測された。生物活性 Antioxidant を持つ代謝物は、同時に Antibacterial, Anticancer, Cytotoxic, Anti-inflammatory, Phytoalexin, Allelopathic, Antihepatotoxic, Antiviral, Antidiabetic, Antiallergic, Pigment, Pollinator attractant, Feeding attractanttwo を持つと頻繁に予測された。生物活性 Anti-inflammatory を持つ代謝物は同時に Antibacterial, Cytotoxic, Antioxidant, Phytoalexin, Allelopathic, Antihepatotoxic, Antiviral, Antidiabetic, Antiallergic, Pigment, Pollinator attractant, Feeding attractant を持つと頻繁に予測された。図 8 において、縦軸に示される生物活性と頻繁に同時出現する生物活性が少なく、生物活

性 Attractant, Antihematitic, Enhance germination などのように縦軸に示される生物活性とほとんど同時に予測されないものも多く存在する。以上のことから生物活性の間には、生物活性 Toxic を持つと予測された代謝物は生物活性 Neurotoxic を持つ事が多いというように特定の生物活性を持つ代謝物は同時に別の特定の生物活性を持つ、という関係性が存在する可能性があると考えられる。

4. 結論

KNApSAcK Core データベースに収録されている 50,037 件の代謝物の 3 次元構造情報および KNApSAcK Metabolite Activity データベースに収録されている 6,649 件の代謝物-生物活性間の関連付け情報を使い、代謝物の分子構造の類似度に基づいて生物活性情報が未知の代謝物 6,480 件の生物活性を予測することができた。生物活性情報が存在する 3,210 件の代謝物データのうち 10%をランダムに選択し、機能未知の代謝物として扱い、その生物活性を予測することにより本研究の予測手法の精度の確認を 10 回繰り返したとき、平均予測精度は 71.934%であった。また、特定の生物活性を持つと予測された代謝物は別の特定の生物活性を持つといった生物活性間の関係性がみられた。

謝辞

本研究に際して、熱心かつ的確な御指導、御教授を賜りました奈良先端科学技術大学院大学 情報科学研究科 金谷重彦 教授に深く感謝申し上げます。

研究発表の際、貴重な御意見を下さいました奈良先端科学技術大学院大学 情報科学研究科 笠原 正治 教授に厚く御礼申し上げます。

本研究において、様々な御指導、御指摘を賜りました奈良先端科学技術大学院大学 情報科学研究科 Md. Altaf-Ul-Amin 准教授に心より感謝申し上げます。

研究発表の際、的確な御意見や御指導を賜りました奈良先端科学技術大学院大学 情報科学研究科 西岡 孝明 教授に深く感謝申し上げます。

研究発表の際、非常に貴重な御意見、ご指導を賜りました奈良先端科学技術大学院大学 情報科学研究科 小野 直亮 助教に深く感謝申し上げます。

研究発表の際、貴重なご意見をくださいました奈良先端科学技術大学院大学 情報科学研究科 佐藤 哲大 助教に心より感謝申し上げます。

本研究において、COMPLIG アルゴリズムについてのご指導だけでなく、研究全体の御助言、御指導を賜りました長浜バイオ大学 バイオサイエンス学部 コンピュータバイオサイエンス学科 白井 剛 教授に深く御礼申し上げます。

研究生生活において多くの知識や示唆を頂きました桂樹 哲雄 様、博士研究員の森田 晶 様、長浜バイオ大学 バイオサイエンス学部 コンピュータバイオサイエンス学科 和田 健之助 様、和田 佳子 様、研究室の博士課程の江口 遼平 様、白石 磨貴男 様、Azian Azamimi binti Abdullah 様に心より感謝申し上げます。

日頃よりともに研究生生活を送り、いつも暖かく支えて下さった計算システムズ生物学研究室の皆様、心より感謝致します。

参考文献

K. Kai, H. Takahashi, H. Saga, T. Ogawa, S. Kanaya, D. Ohta, *Plant Biotechnology*, 2011, 28, 379-385.

K. Saito, F. Matsuda, *Annual Review of Plant Biology*, 2010, 61, 463-489.

M. Saito, N. Takemura, T. Shirai, *Journal of Molecular Biology*, 2012, 424, 379-390.

Md. Altaf-Ul-Amin, M. Wada, S. Kanaya, *ISRN Biomathematics*, 2012, ID726429, 1-12.

Md. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, S. Kanaya, *BMC Bioinformatics*, 2006, 7, 207, 1-13.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, *Genome Research*, 2013, 13, 2498-2504.

R. Croteau, T. M. Kutchan, N. G. Lewis, *Biochemistry and Molecular Biology of Plants*, 2000, 24, 1250-1318.

V. D. Luca, B. S. Pierre, *Trends in Plant Science*, 2000, 5, 4, 168-173.

Y. Benjamini, Y. Hochberg, *Journal of the Royal Statistical Society. Series B Methodological*, 1995, 57, 289-300.

Y. Nakamura, F. M. Afendi, A. K. Parvin, N. Ono, K. Tanaka, A. H. Morita, T. Sato, T. Sugiura, Md. Altaf-Ul-Amin, S. Kanaya, *Plant Cell Physiology*, 2014, 55, e7.1-9.

Y. Ohtana, A. A. Abdullar, Md. Altaf-Ul-Amin, M. Huang, N. Ono, T. Sato, T. Sugiura, H. Horai, Y. Nakamura, A. Morita(Hirai), K. W. Lange, N. K. Kibinge, T. Katsuragi, T. Shirai, S. Kanaya, *Molecular Informatics*, 2014, 33, 11-12, 790-801.