

NAIST-IS-DD1661003

博士論文

グラフ構造に基づいたアルカロイド化合物の分類および
代謝経路予測

江口 遼平

2018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
博士（工学）授与の要件として提出した博士論文である。

江口 遼平

審査委員：

金谷 重彦 教授	（主指導教官）
笠原 正治 教授	（副指導教官）
MD.ALTAf-UL-AMIN 准教授	（副指導教官）
小野 直亮 准教授	（副指導教官）

グラフ構造に基づいたアルカロイド化合物の分類および代謝経路予測*

江口 遼平

内容梗概

二次代謝産物の生物の生体内における機能は、害虫および病原体に対する防御反応、太陽光などの刺激に対する防衛反応などを含む進化過程において獲得された生体分子活性に関連する。そのため二次代謝物と化学構造および代謝経路との関連性を明らかにすることは、進化過程における二次代謝産物の機能獲得メカニズムの解釈につながる可能性がある。

本研究では構造多様性が非常に高いアルカロイド化合物を研究対象としており、これらの情報を最新の学術文献から収集することで経路情報の拡充を行った。そして、環構造に基づくアルカロイド化合物の分類および、グラフ畳み込みネットワークを用いた学習モデルによる出発物質の予測を通じて、二次代謝物構造-代謝経路間の関係性の網羅的な理解を試みた。本研究においては 478 の化合物から構成される 32 のアルカロイド経路を文献情報から再構築した。そして環構造に基づくアルカロイド化合物の分類を行った結果、29 のクラスタと 25 のサブクラスタ、187 種の環骨格を抽出してアルカロイド化合物の化学構造分類との妥当性を検証した。また、同様にアルカロイド代謝経路の情報に対して、グラフ畳み込みネットワークを用いることで各化合物の出発物質を予測したところ、平均して約 98% と非常に高い精度で予測を行うことができた。

キーワード

ケモインフォマティクス, バイオインフォマティクス, アルカロイド化合物, 環骨格, 深層学習, グラフ畳み込みネットワーク

*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 博士論文, NAIST-IS-DD1661003, 2018 年 3 月 15 日.

Classification of alkaloid compounds and prediction of metabolic pathway based on graph structure*

Ryohei Eguchi

Abstract

Secondary metabolites are defined as a group of natural compounds that are not directly involved in growth, development or reproduction of organisms. The term 'secondary' in the context of metabolic pathways, was introduced in 1891. The functions of secondary metabolites are related to their prevalent potent biomolecular activity acquired in evolution process involving pest and pathogen defense, UV-B-sunscreens etc. The classification of secondary metabolites taking chemical structure and metabolic pathway into consideration could provide important clues to activities of metabolites which lead to interpretation of function acquisition mechanisms of secondary metabolites in evolutionary process. In the present study, we examined whether or not classification of alkaloids by ring skeletons can be related to metabolic pathways.

Keywords:

Chemoinformatics, Bioinformatics, Alkaloid compounds, Ring skeleton, Deeplearning, Graph Convolutional Network

*Doctor's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1661003, March 15, 2018.

目次

図目次		v
第 1 章	はじめに	1
1.1	はじめに	1
1.1.1	代謝とは	2
1.2	関連研究-代謝物の 3 次元構造と生物活性の関連性	2
1.3	本研究の目的	3
1.4	本論文の構成	4
第 2 章	アルカロイド代謝経路の構築	5
2.1	概要	5
2.1.1	アルカロイド化合物	5
2.1.2	既存の代謝経路情報	6
2.1.3	目的	6
2.2	手法	6
2.3	文献調査結果	7
2.4	まとめ	22
第 3 章	部分環構造に着目した化合物のアルカロイド代謝経路上へのマッ ピング	23
3.1	概要	23
3.2	手法	23
3.2.1	計算機上における化合物の構造表現	24
3.2.2	化合物の類似性評価	26
	フィンガープリント法	26
3.2.3	提案手法-部分環構造に基づく分類	26
3.2.4	二次代謝物の分類	30
3.3	結果及び考察	31
3.3.1	部分環構造に基づくアルカロイド化合物の分類	31

3.3.2	環構造に基づくアルカロイド代謝経路の代謝モジュール抽出	46
3.3.3	KNAPSAcK Database 中のアルカロイド化合物における 分類	64
3.3.4	化合物の部分環構造の抽出に関する問題点と解決策	64
3.4	まとめ	68
第 4 章	グラフ畳み込みニューラルネットワークを用いた二次代謝物の化学構造による出発物質予測	70
4.1	概要	70
4.2	手法	70
4.2.1	ディープラーニング	71
4.2.2	畳み込みニューラルネットワーク	73
	畳み込み層	74
	プーリング層	74
	全結合層	74
4.2.3	グラフ畳み込みネットワーク	75
4.2.4	入力層	75
	畳み込み層	76
	プーリング層	79
	集約層	81
4.3	グラフ畳み込みネットワークモデルにおける重みの学習	83
4.4	データセット	84
4.5	結果および考察	85
4.6	まとめ	107
第 5 章	総括	108
	謝辞	109
	研究業績	110
	参考文献	111

目次

1.1	KNApSAcK Core Database	3
2.1	各アルカロイド代謝経路と出発物質との対応関係	8
2.2	アルカロイド代謝経路 1	9
2.3	アルカロイド代謝経路 2	10
2.4	アルカロイド代謝経路 3	11
2.5	アルカロイド代謝経路 4~10	12
2.6	アルカロイド代謝経路 11	13
2.7	アルカロイド代謝経路 12	14
2.8	アルカロイド代謝経路 13~14	15
2.9	アルカロイド代謝経路 15~16	16
2.10	アルカロイド代謝経路 17~19	17
2.11	アルカロイド代謝経路 20~29	18
2.12	アルカロイド代謝経路 30	19
2.13	アルカロイド代謝経路 31~32	20
3.1	本研究における環構造に基づく化合物分類の流れ	24
3.2	MDL Molfile における化学構造の記述例	25
3.3	フィンガープリント法による化合物のベクトル表現例	27
3.4	PubChem フィンガープリントの各ビットが表す特徴の一部例	28
3.5	化合物のグラフ表現および Bemis-Marco らによる構造の定義	29
3.6	フラボノイドから部分環構造を抽出した例	30
3.7	化合物の構造情報から部分環構造群を抽出するための疑似アルゴリズム	31
3.8	各化合物を部分環構造に基づき表現したバイナリ行列表	32
3.9	アルカロイド代謝経路上に存在する化合物における部分環構造の出現頻度	33
3.10	アルカロイド化合物の階層的クラスタリング	34

3.11	各クラスタ (1~9) とそれらを構成する環骨格	36
3.12	各クラスタ (10~16) とそれらを構成する環骨格	37
3.13	各クラスタ (17~21) とそれらを構成する環骨格	38
3.14	各クラスタと (22~24) それらを構成する環骨格	39
3.15	各クラスタ (25~29) とそれらを構成する環骨格	40
3.16	各クラスタ (27A~27J) とそれらを構成する環骨格	42
3.17	各クラスタ (27K~27P) とそれらを構成する環骨格	43
3.18	各クラスタ (27Q~27T) とそれらを構成する環骨格	44
3.19	各クラスタ (27U~27Y) とそれらを構成する環骨格	45
3.20	各アルカロイド代謝経路とクラスタの関係	47
3.21	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 1)	48
3.22	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 2)	50
3.23	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 3)	51
3.24	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 4~10)	52
3.25	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 11)	54
3.26	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 12)	55
3.27	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 13~14)	56
3.28	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 15~16)	58
3.29	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 17~19)	59
3.30	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 20~30)	61
3.31	アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 31~32)	63
3.32	ZDD の構造例 (a. ある化合物のグラフ表現、b.a の化合物から部 分環構造を抽出する際の二分木表現、c.b の二分木の ZDD)	66
3.33	二分木から ZDD に圧縮する際の規則 (冗長節点の削除)	67

3.34	二分木から ZDD に圧縮する際の規則 (等価節点の共有)	68
4.1	ニューロンモデル	71
4.2	順伝播型ニューラルネットワークの構造例 (4 層)	72
4.3	畳み込みニューラルネットワークの概念図	73
4.4	畳み込みニューラルネットワークにおけるプーリングの一例	75
4.5	グラフ畳み込みネットワークへの入力として用いる各原子に特徴ベクトルが割り当てられた化合物	76
4.6	畳み込み層における処理の概念図	78
4.7	プーリング層における処理の概念図	80
4.8	グラフ畳み込みネットワークによる処理の全体像. 特徴ベクトルを更新後、集約層においてベクトルを 1 つにまとめる.	82
4.9	学習の全体像. 集約層で出力された特徴ベクトルは重み W_f によって学習の予測値に変換し、訓練データとの誤差を損失関数 L によって評価する.	83
4.10	本研究におけるグラフ畳み込みネットワークの構成	85
4.11	L-Tryptophan と誤って予測された化合物	88
4.12	L-Proline と誤って予測された化合物	89
4.13	Anthranilate と誤って予測された化合物	90
4.14	L-Lysine と誤って予測された化合物	91
4.15	L-Arginine と誤って予測された化合物	92
4.16	L-Alanine と誤って予測された化合物	93
4.17	Malonyl-CoA と誤って予測された化合物	94
4.18	L-Phenylalanine と誤って予測された化合物	95
4.19	L-Phenylalanine と誤って予測された化合物	96
4.20	IPP と誤って予測された化合物	97
4.21	誤って予測された化合物の経路上へのマッピング (1)	98
4.22	誤って予測された化合物の経路上へのマッピング (2)	99
4.23	誤って予測された化合物の経路上へのマッピング (3)	100
4.24	誤って予測された化合物の経路上へのマッピング (4)	101
4.25	誤って予測された化合物の経路上へのマッピング (5)	102

4.26	誤って予測された化合物の経路上へのマッピング (6)	103
4.27	誤って予測された化合物の経路上へのマッピング (7)	104
4.28	誤って予測された化合物の経路上へのマッピング (8)	105

第 1 章 はじめに

1.1 はじめに

英米の自然史博物館図書館、植物園図書館、研究機関が協同で構築した生物多様性に関する文献をデジタル化した Biodiversity Heritage Library や次世代シーケンサー技術、マイクロアレイ技術をはじめとする計測機器の高速化によって、近年、生物のエコシステムを理解するために活用できるデータは増え続けている。そのため、生物学もまたデータ集約型のビックデータサイエンスに移行している。生物の多様性をそれらが産出する生体分子に基づいて理解するためのデータベースシステムを開発することは生物学分野のビックデータサイエンスにおいて最も重要なトピックスであるといえる [1]。この目的を達成するために、これまで数多くのデータベースが開発されている。KNApSAcK Family Database もそのうちの一つであり、これは特にメタボロミクスに関する研究を遂行することを目的として開発された [2-4]。科学研究における KNApSAcK Database の利用報告は池田らによって示されており [5]、古典的な知識に基づいた植物の医薬的な利用のためのデータのほか、近代的な知識に基づいた健康に関する料理のレシピデータ、代謝物相互関係のデータのみならず、メタボロミクスを根幹として生物に関連した事象の包括的な理解を可能にするために KNApSAcK Family Database では数多くのデータを収集し、現在もなおデータベースの拡充を進めている [6]。

また、二次代謝物とそれらを生合成あるいは蓄積する生物種の相互関係を整理したデータベースが KNApSAcK Core Database(http://kanaya.naist.jp/knapsack_jsp/top.html) である。現在に至るまで、約 2 万生物種と約 5 万種の二次代謝物を含んだ、おおよそ 10 万対の生物種—二次代謝物間の相互関係のデータを KNApSAcK Core Database として公開している。また、二次代謝物とそれが有する生物活性・生理活性の関係性を体系化し、それらの相互関係の包括的な理解を可能にするため、KNSpSAcK Metabolite Database が公開されている (<http://kanaya.naist.jp/MetaboliteEcology/top.jsp>)。これには 2963 種の生物活性・薬理活性とそれらが属する 140 種の活性カテゴリおよび 778 種の生物種を含む、おおよそ 9000 の代謝物-生物活性-対象種の関係が整理され、公開されている。

1.1.1 代謝とは

生物の体内では、外部から取り込まれた物質が新たな別の物質へと変化する化学変化が連鎖的に起こっている。このような化学変化は代謝反応と呼ばれ、代謝反応にかかわる物質を代謝物質と呼ぶ。代謝の主な役割は外部から吸収した食物からエネルギーを得ることおよび外部から取り込まれた物質から生物が必要とする物質を合成することである。生物はその進化の過程において、潜在的に酵素遺伝子を保存しながら共に機能分化させることで選択的な機能を得る。このような選択圧は異なった二次代謝物やそれに関連する代謝経路の出現を引き起こした。

代謝物は生物の生命活動に直接関与する一次代謝物と、それが無くとも即座に致死に至ることはないが生物の生存戦略に関与する二次代謝物に分けられる。二次代謝物は生物の生体内で病原体防御や太陽光防御をはじめとする生存戦略に用いられている、またこれらは多種多様な生物活性を有しているため、医薬品、燃料、香料などを主な用途として人類に活用されてきた。そのため生物にとって有用な二次代謝物を安定的に得ることは重要な課題であり、これを実現するために二次代謝物と生物活性の関係を包括的に理解する必要がある。

1.2 関連研究-代謝物の 3 次元構造と生物活性の関連性

二次代謝物と生物活性の関係については、KNApSAcK Metabolite Activity Database(<http://kanaya.naist.jp/MetaboliteActivity/top.jsp>) [7] として設計を進め、約 5000 の代謝物と生物活性の関係を整理し、拡充を進めている。このデータベースでは、科学論文において報告がなされている生物活性が 140 種の活性グループ(活性オントロジー)に分類され、体系的に整理されている。このうち半数の活性オントロジーは、誘引作用、忌避作用などの生物-生物間相互作用を関係付ける生態学に関連するものであり、もう残りの半数は医薬品等に活用される生理活性を有するヘルスケアに関連するものである。

このように二次代謝物の化学構造、それが有する生理活性、代謝経路に関する情報を画一的に集積し、データサイエンスにおいて活用する体制の確立は進められており、そこで、生物活性が報告されている約 2000 種の二次代謝物を対象に、三次元立体構造の類似性により分類し、分類結果により得られた各々のグループに含ま

れる二次代謝物の生物活性の共通性を KNApSAcK Metabolite Activity における活性オントロジーにより検討したところ、各々のグループを共通の生物活性により特徴付けられた [8]。

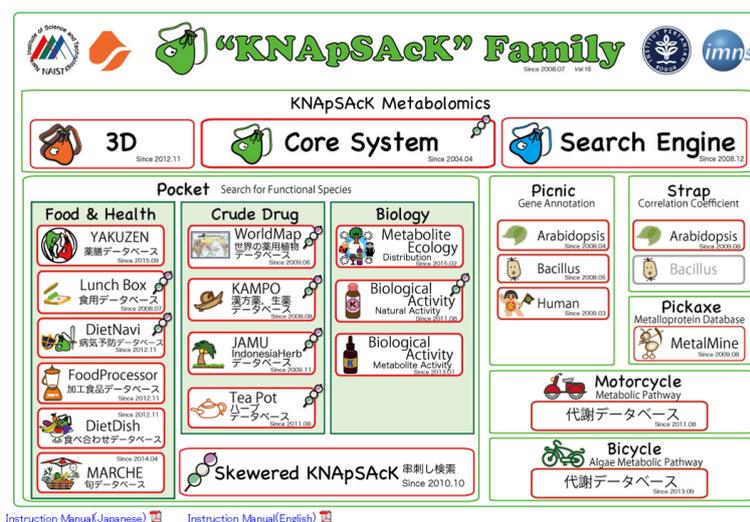


図 1.1 KNApSAcK Core Database

1.3 本研究の目的

本研究の目的はデータサイエンスを通じて、代謝物とそれを創出する代謝経路の相互関係性を網羅的に理解することである。この目的を達成するために、本研究においては二つの観点からそれに対するアプローチを行う。

第一の観点として二次代謝物の構造情報から本研究において提案する環構造に基づく分類を行い、その分類を代謝経路上に投影することで代謝モジュールとして抽出することを試みる。第二の観点としてグラフ畳み込みネットワークを用いた機械学習モデルを適用することで、二次代謝物の構造情報からそれらがどの出発物質から誘導されるのか、即ちいずれの代謝経路に由来する二次代謝物であるのかの予測を試みる。言うまでもなく前述した二つの観点からのアプローチはいずれにしても、二次代謝物-代謝経路あるいは二次代謝物-生物活性間の関係性を網羅的に理解するという点で一致しており、第一の観点においては任意の代謝経路のモジュール

と代謝物の構造を結び付ける点で局所的な理解を、第二の観点においては代謝物からその属する代謝経路を予測するという点で広域的な理解を試みた。

1.4 本論文の構成

本論文は5章から構成されており、まず第2章において本研究における検証の対象としたアルカロイド化合物情報を学術文献から収集、蓄積することでより大規模なアルカロイド代謝経路の構築を行った。第3章においてはアルカロイド代謝経路に属する二次代謝物に対して構造上に基づく分類を行い、それらの結果と代謝経路の関係性についての考察を行った。また、第4章においてはグラフ畳み込みネットワークを用いることで、アルカロイド化合物の構造情報から代謝経路の予測を行い、その妥当性について議論を行った。そして、第5章において本研究での総括を記述する。

第 2 章 アルカロイド代謝経路の構築

2.1 概要

第 1 章で述べたように本研究の目的は、二次代謝物の代謝経路から化学構造における部分環構造の類似性に基づいて代謝モジュールを抽出すること、また、二次代謝物の構造情報からその出発物質を予測することである。これらを検証する対象として、本研究ではアルカロイド化合物およびアルカロイド代謝経路に着目した。本章では研究の対象であるアルカロイド化合物情報を最新の学術文献から収集し、それらに基づきアルカロイド代謝経路の構築を行う。

2.1.1 アルカロイド化合物

アルカロイド化合物は、特徴的な毒性及び薬理活性を有する低分子量の窒素含有化合物である。植物種のおおよそ 20 %がアルカロイドを生体内に蓄積し、草食動物および病原体に対する防御反応において重要な役割を果たしている。また、これらは主に Phenylalanine、Tyrosine、Tryptophan、Lysine などのアミノ酸、核酸をはじめとする生体含窒素代謝物を出発物質として生合成される [9,10]。現在では 12000 種類を超えるアルカロイド化合物が発見されており、他の二次代謝産物と比較しても構造的な多様性が非常に高い。

天然資源から得られるアルカロイド化合物には、人にとって有用な性質を持ったものも多数報告されており、これまでに医療用途をはじめとして様々な分野において活用されている [11]。例えば麻薬性鎮痛作用を有するモルヒネ、鎮痛および鎮咳性作用を有するコデイン、また、抗がん剤治療に活用されるビンクリスチンおよびビンブラスチンなどがこれに当たる。アルカロイド化合物は生物から抽出したものを直接活用するだけでなく、創薬のための新たな発見をもたらすという点においても非常に重要である。アルカロイド化合物の生合成の解明に向けた研究は現在も進められているものの、アルカロイド化合物のうちその代謝経路が解明されているものは非常に限られている [12,13]。

2.1.2 既存の代謝経路情報

自然界で生産される二次代謝物の総数は 100 万種類以上あるとの推測がなされている。しかしながら、既に発見されている二次代謝物についても、それらがどの代謝経路によって生産されているかが不明なものは非常に多く、研究者らによって様々な生合成経路の推定が行われ報告がなされている。アルカロイド化合物においてもこの例に漏れず、生合成経路が不明な化合物は多く存在している。バイオインフォマティクス研究用の著名なデータベースである Kyoto Encyclopedia of Genes and Genomes(KEGG) [14] において、公開しているアルカロイド代謝経路 (Biosynthesis of alkaloids derived from shikimate pathway, Biosynthesis of alkaloids derived from ornithine, lysine and nicotinic acid, Biosynthesis of alkaloids derived from histidine and purine, Biosynthesis of alkaloids derived from terpenoid and polyketide) は現時点で約 300 種のアルカロイド化合物により構成されている。

2.1.3 目的

本研究では二次代謝物の代謝経路からのモジュールの抽出、また、二次代謝物の構造情報からの出発物質予測することを試み、その対象として構造に多様性があるアルカロイド化合物を対象として選択をした。しかしながら前述した通り、構造および性質が既知である二次代謝物であってもそれらが由来する代謝経路の情報は未知であることが多い。そこで本章においては最新の文献情報に基づいてこれらの検証を行うため、対象であるアルカロイド化合物の代謝経路情報の拡充および代謝マップの構築を行う。

2.2 手法

本研究では文献情報に基づいた知識ベースでのアルカロイド化合物代謝経路の体系化を行う。生物、生化学分野における最新の文献情報から二次代謝物間の二項関係を抽出し再構築することで経路情報の拡充を行う。各文献にて報告されている代謝反応は非常に断片的なものが多いが、それらを多数収集することで異なる代謝反

応を繋ぎ合わせアルカロイド代謝経路を構築する。

2.3 文献調査結果

アルカロイド化合物およびそれらに関連する代謝経路情報を 68 件の文献情報 [15–82] から収集した (収集した全アルカロイドのリストは付録を参照)。その結果、478 のアルカロイド化合物情報と化合物間を関連付ける経路情報を得ることが出来た。ここで構築したアルカロイド代謝経路と各出発物質の関係を図 2.1 に示す。また、これらの情報に基づき構築した 32 のアルカロイド代謝経路を図 2.2~図 2.13 に示す。構築した 32 のアルカロイド代謝経路は、それぞれの化合物が由来する出発物質に基づき分割をしている。本研究におけるアルカロイド代謝経路上の出発物質として次の 33 種類を定義した: Glycine、L-Serine、D-Serine、L-Cysteine、L-Alanine、D-Alanine、L-Leucine、L-Valine、L-Tryptophan、D-Tryptophan、L-Lysine、L-Phenylalanine、Anthranilate、Fornylanthranilate、Indole-3-glycerol phosphate、O-Methyltyrosine、L-Asparatic acid、L-Threonine、L-Lysine、L-Arginine、L-Histidine、L-Glutamine、Isopentenyl diphosphate、Dimethylallyl diphosphate、Secologanin、Geranylgeranyl diphosphate、Cholesterol、Acetyl-CoA、Oxaloacetate、Malonyl-CoA、Acetoacetyl-CoA、Adenine。これらはいずれもアミノ酸、アミノ酸生合成に関連する化合物、TCA サイクル、脂肪酸、核酸に関連する化合物である。

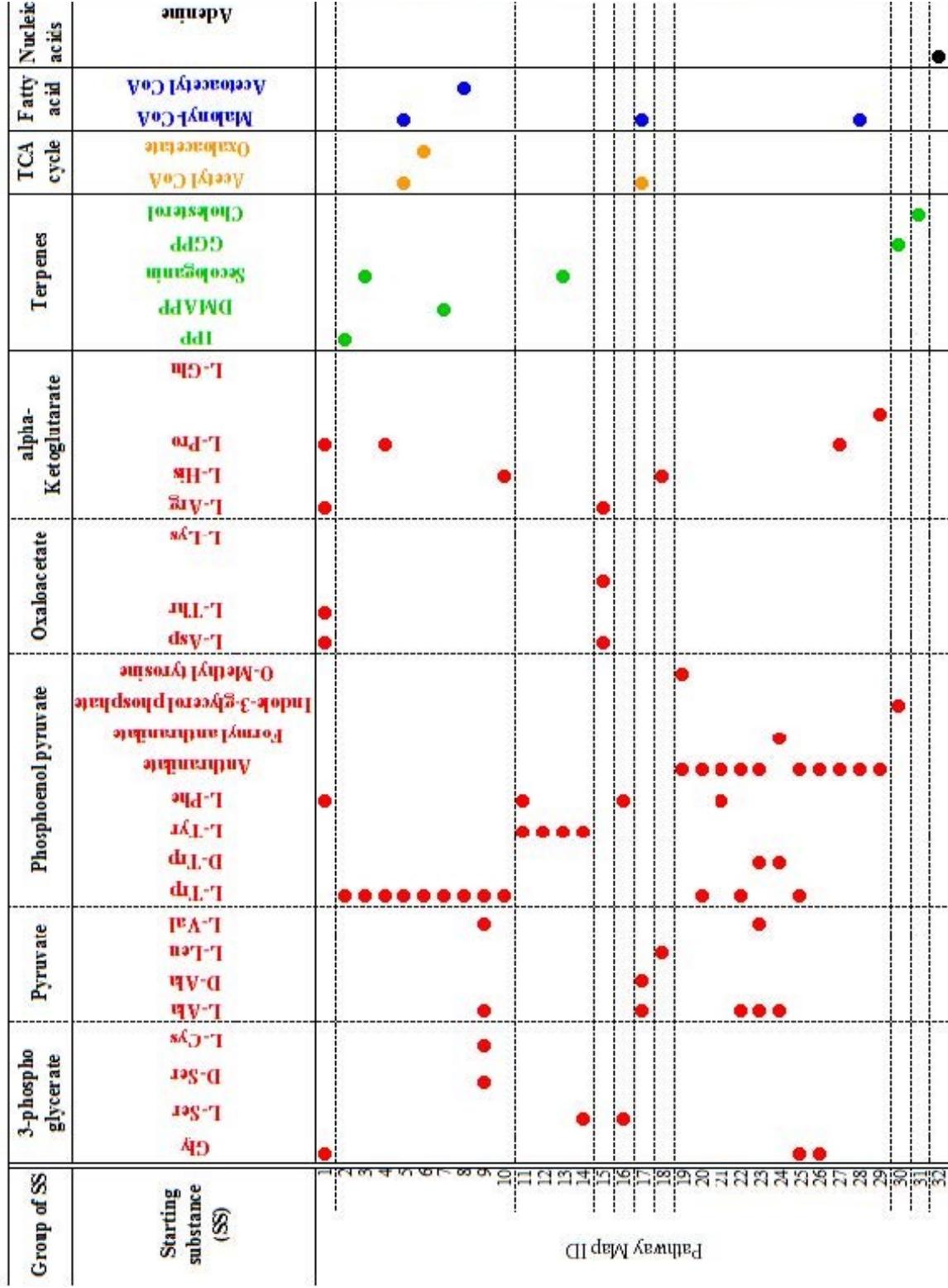


図 2.1 各アラルカロイド代謝経路と出発物質との対応関係

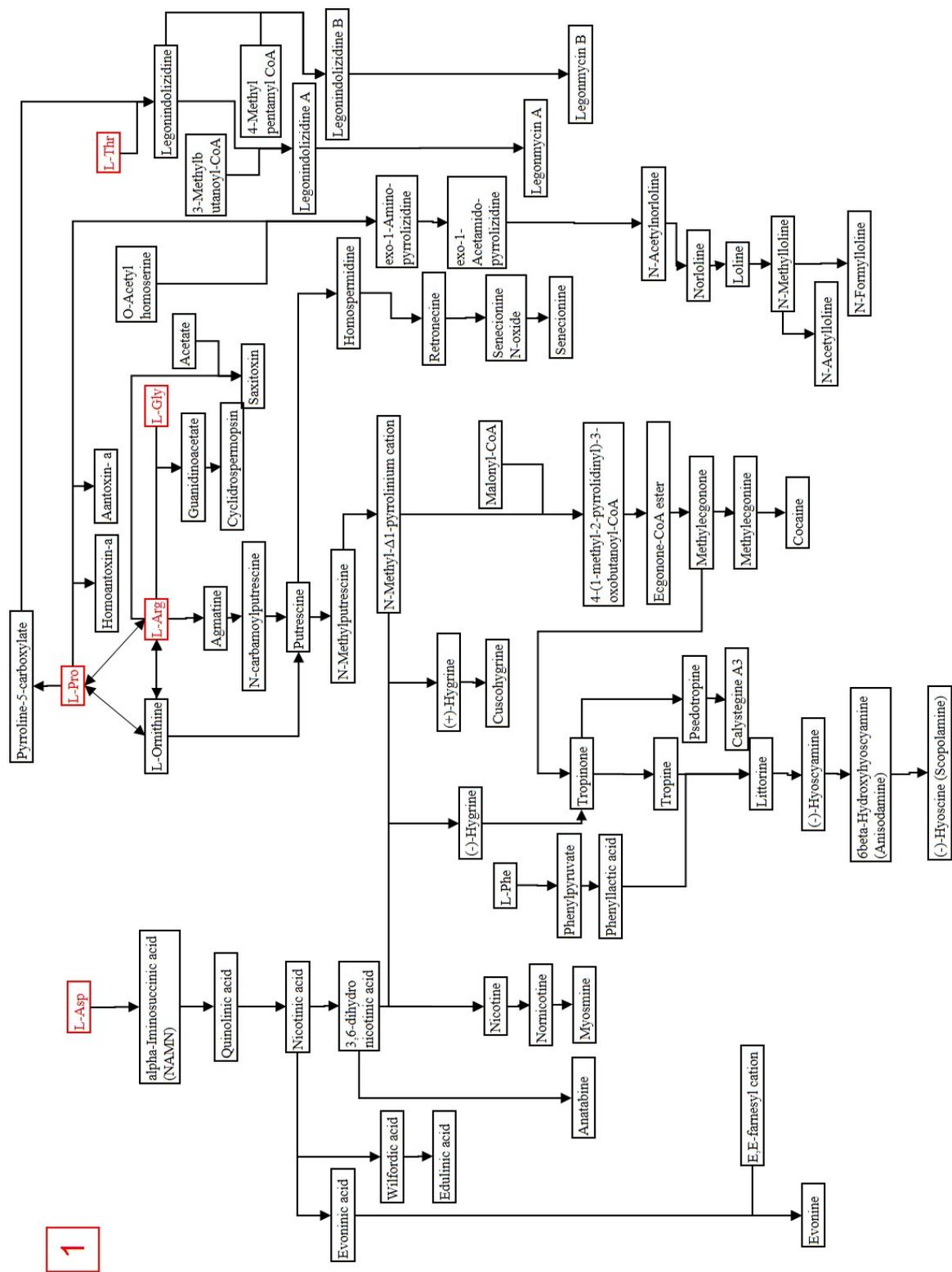


図 2.2 アルカロイド代謝経路 1

2

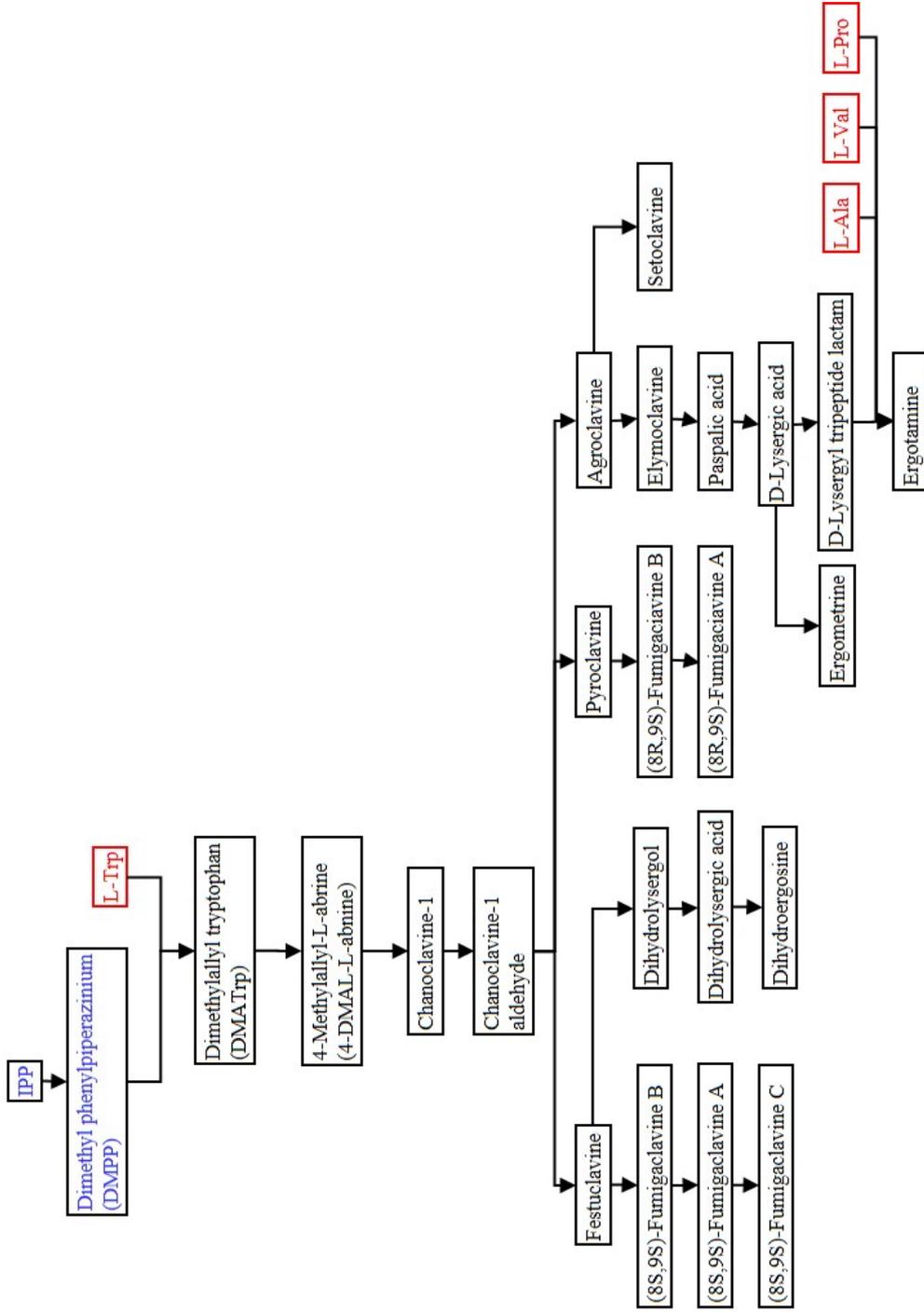


図 2.3 アルカロイド代謝経路 2

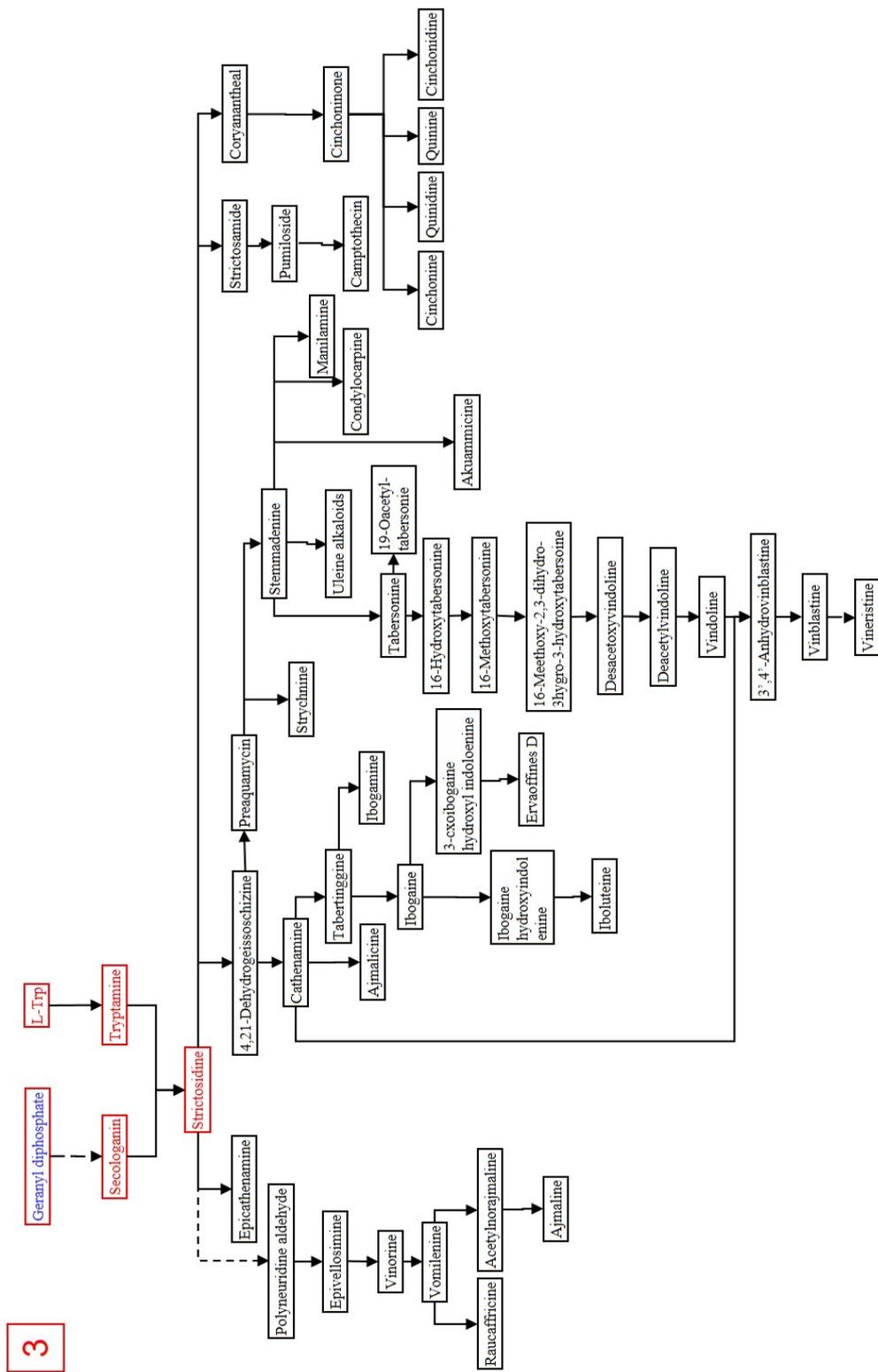


図 2.4 アルカロイド代謝経路 3

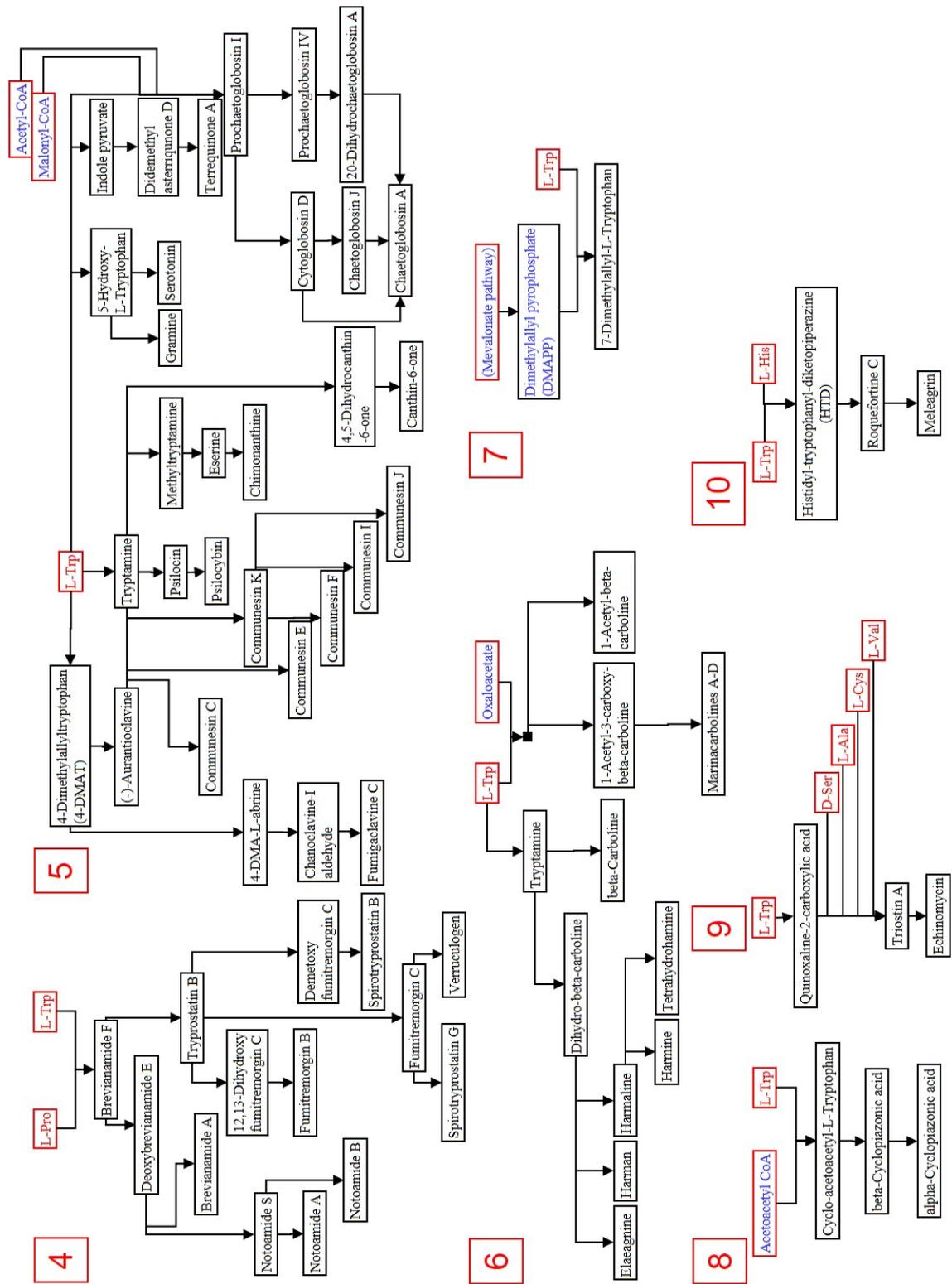


図 2.5 アルカロイド代謝経路 4-10

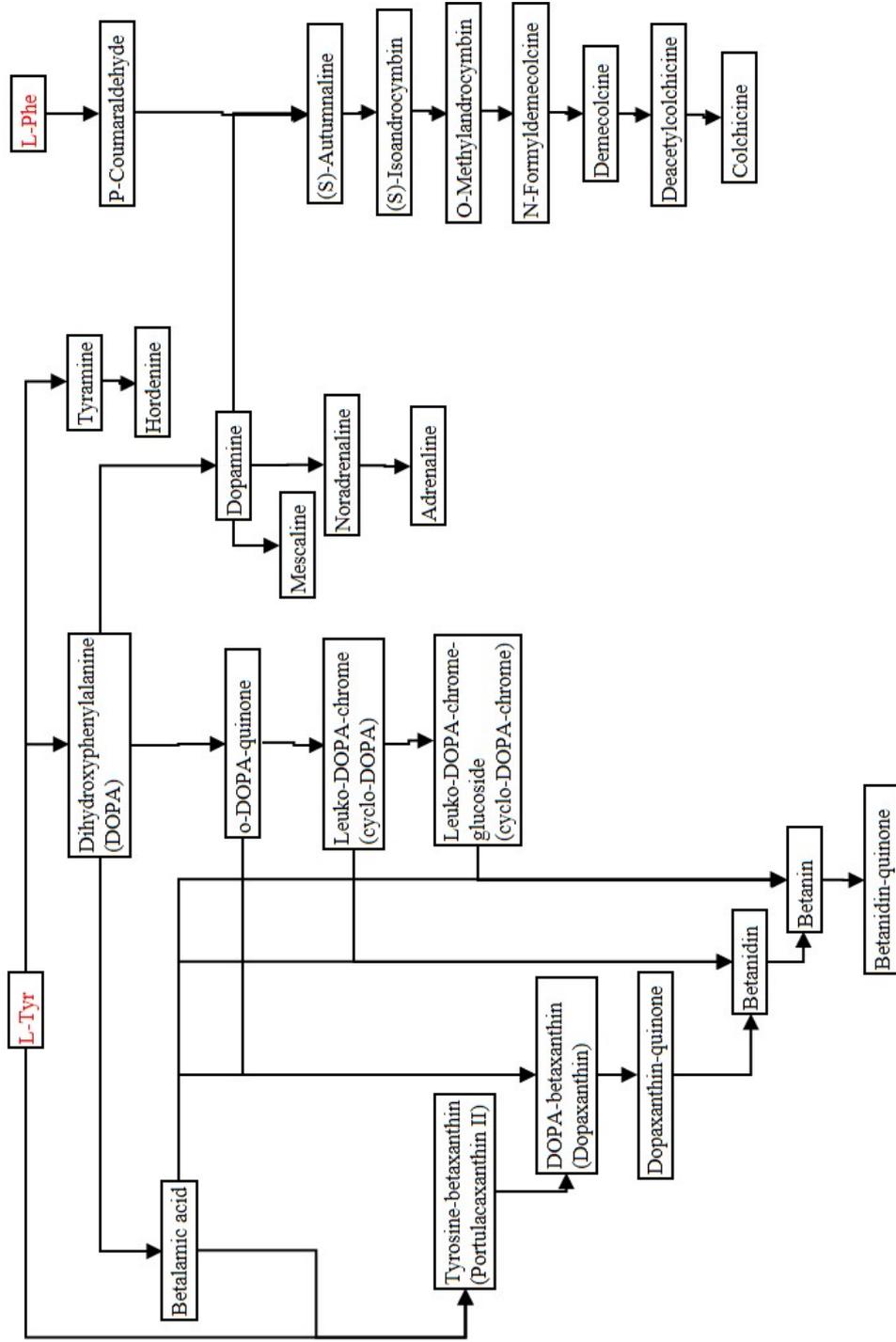


図 2.6 アルカロイド代謝経路 11

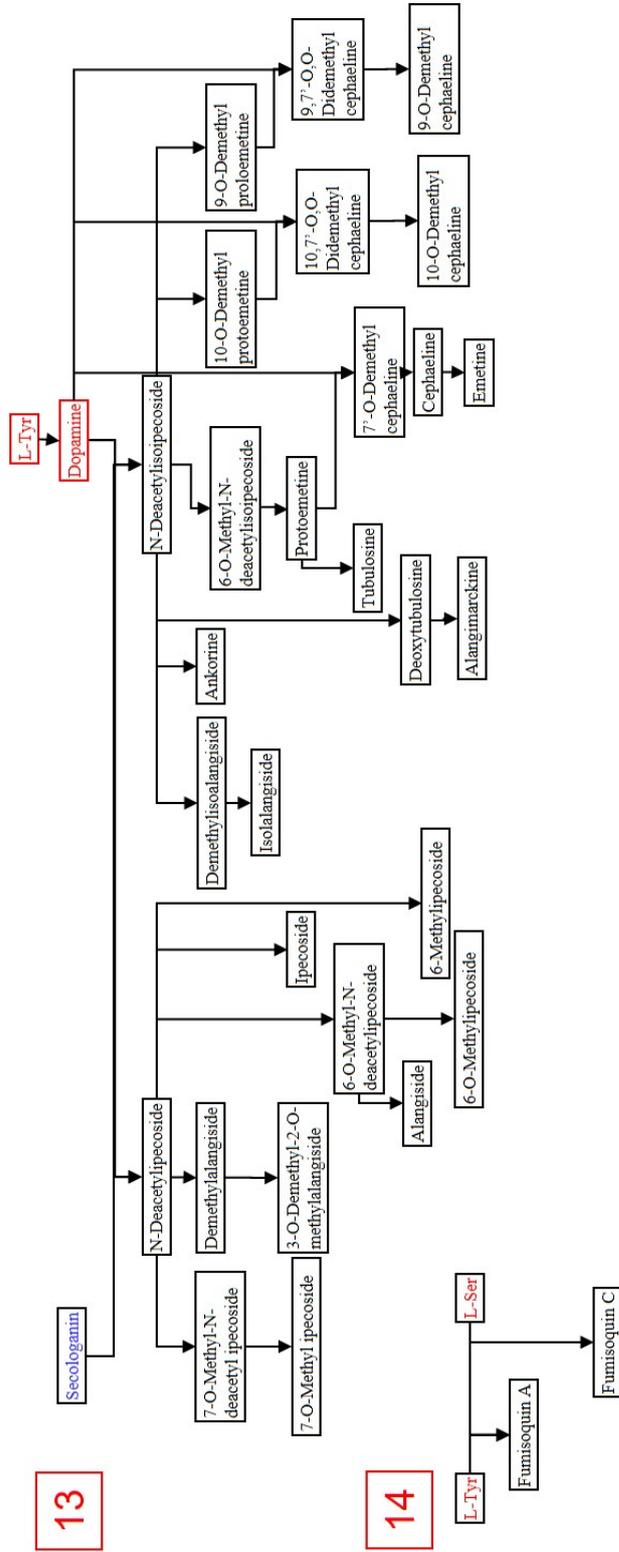
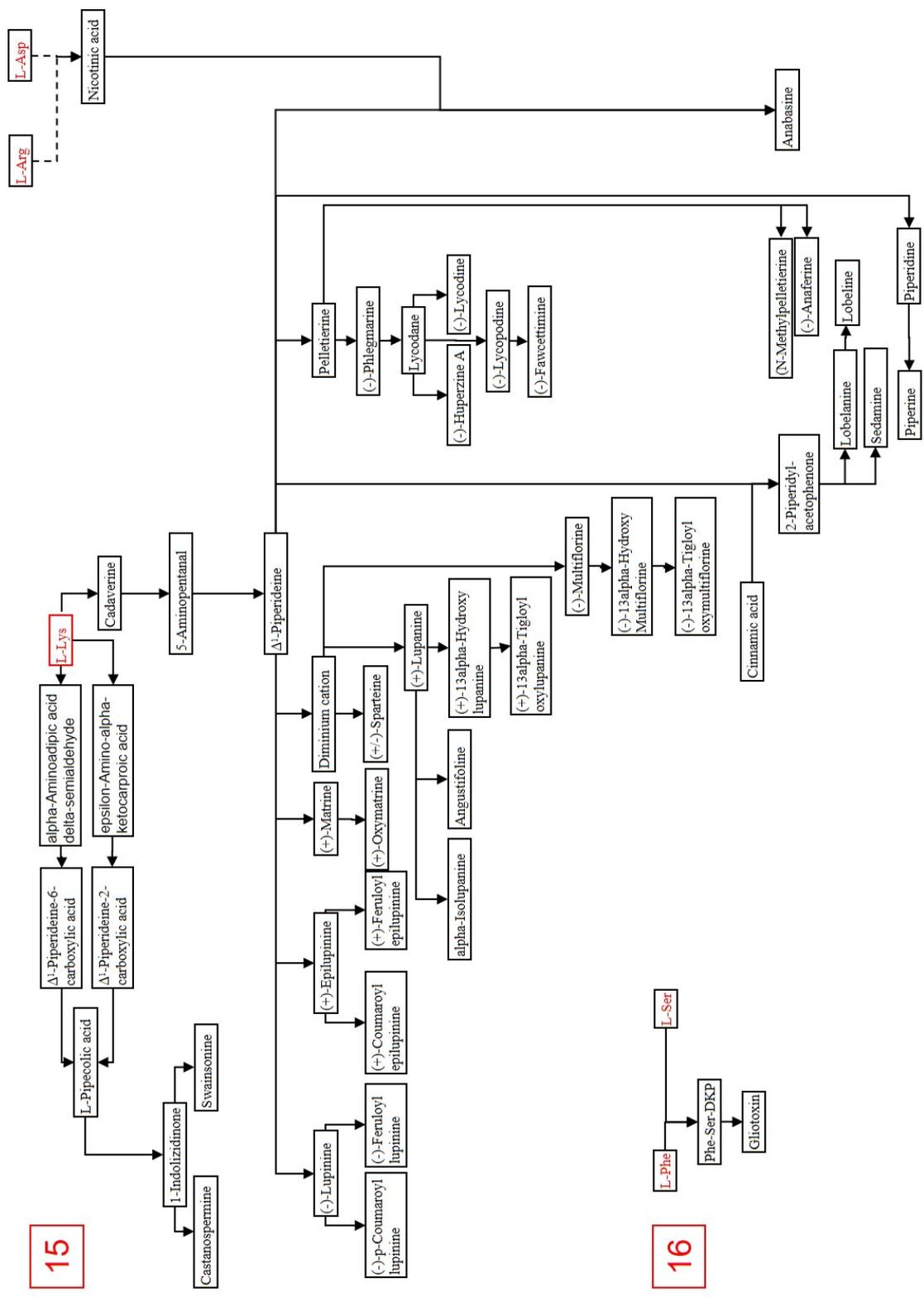


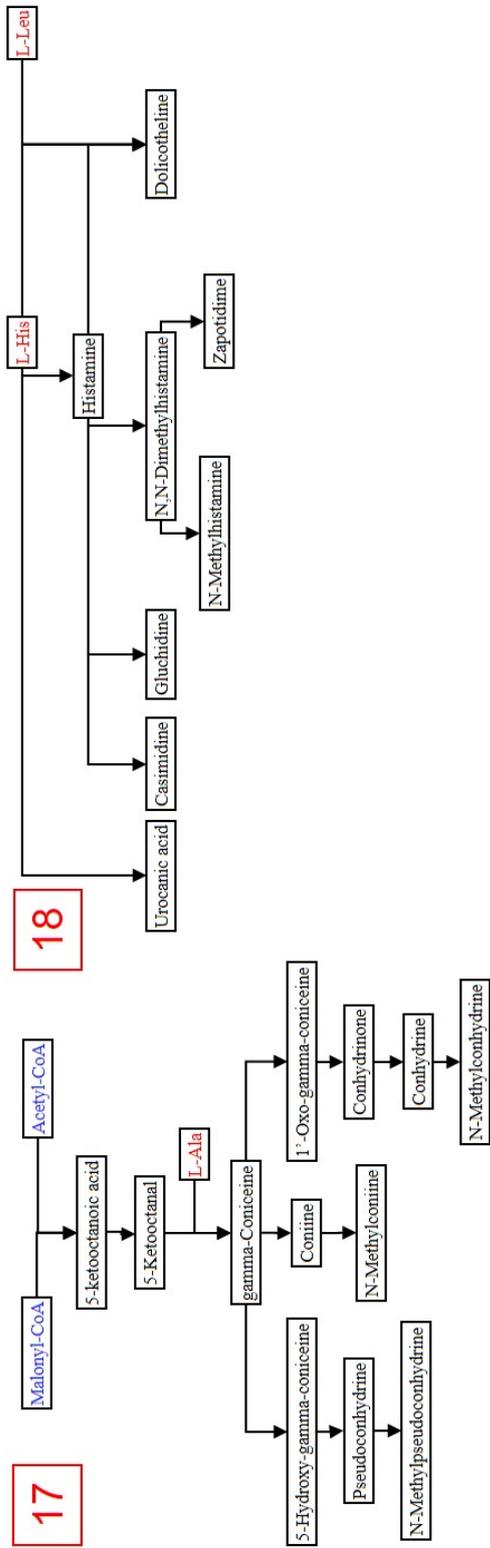
図 2.8 アルカロイド代謝経路 13~14



15

16

図 2.9 アルカロイド代謝経路 15~16



17

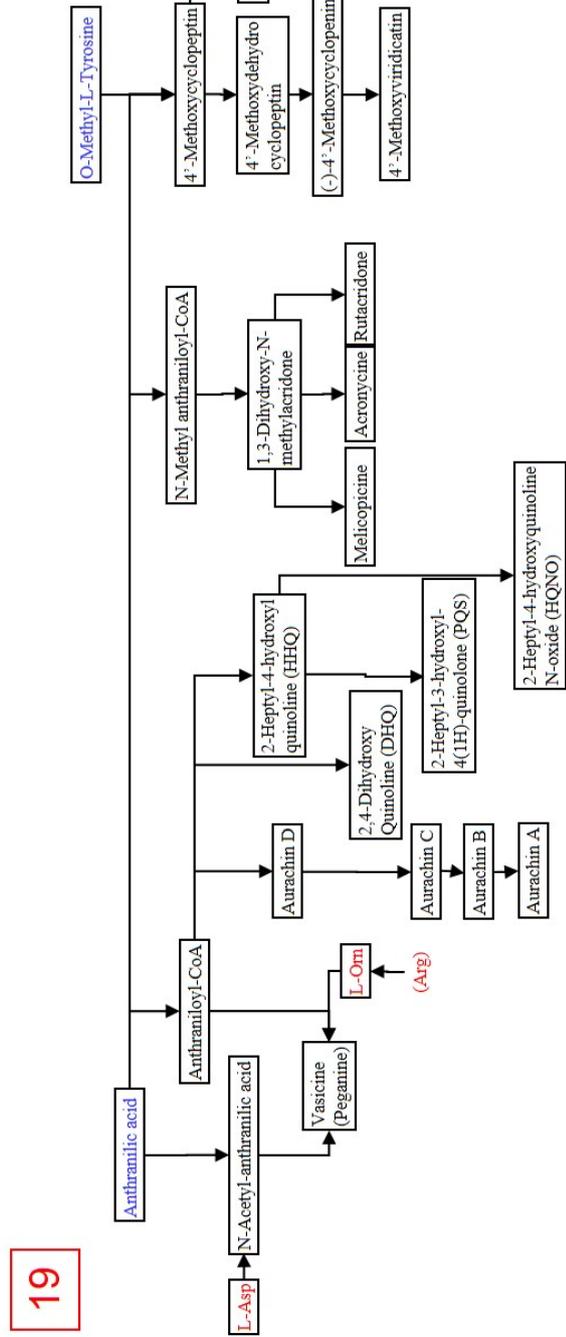


図 2.10 アルカロイド代謝経路 17~19

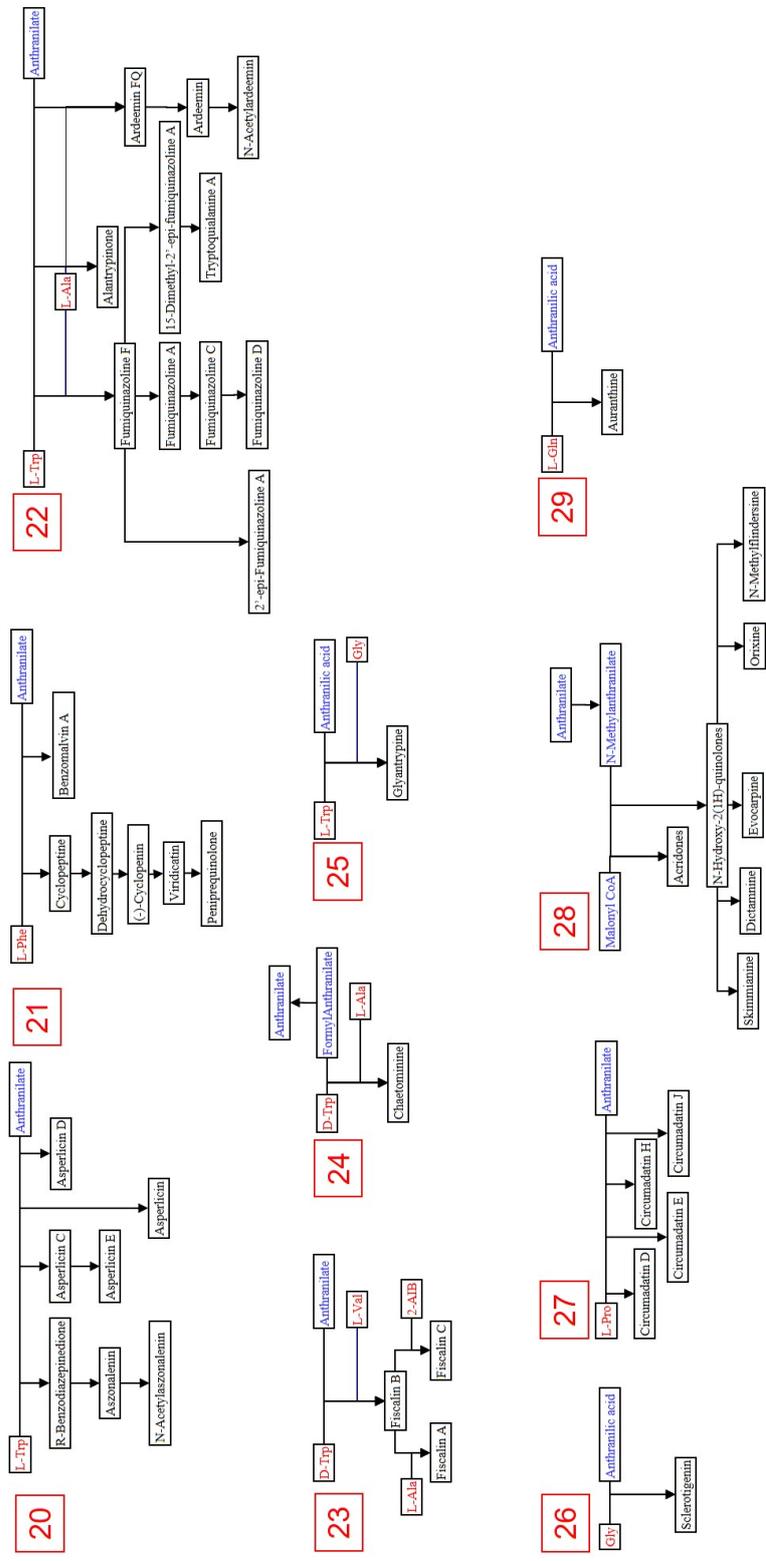


図 2.11 アルカロイド代謝経路 20~29

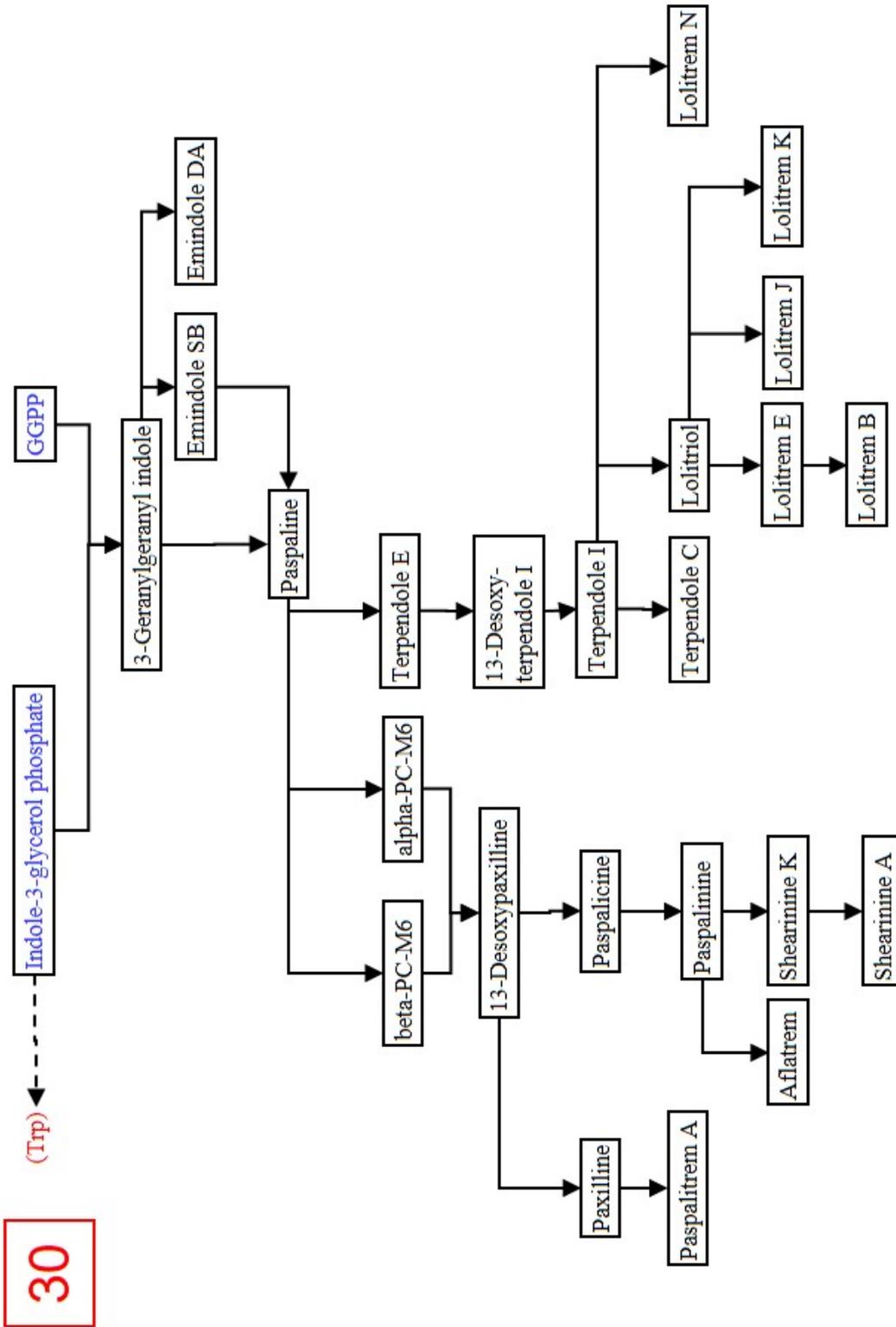


図 2.12 アルカロイド代謝経路 30

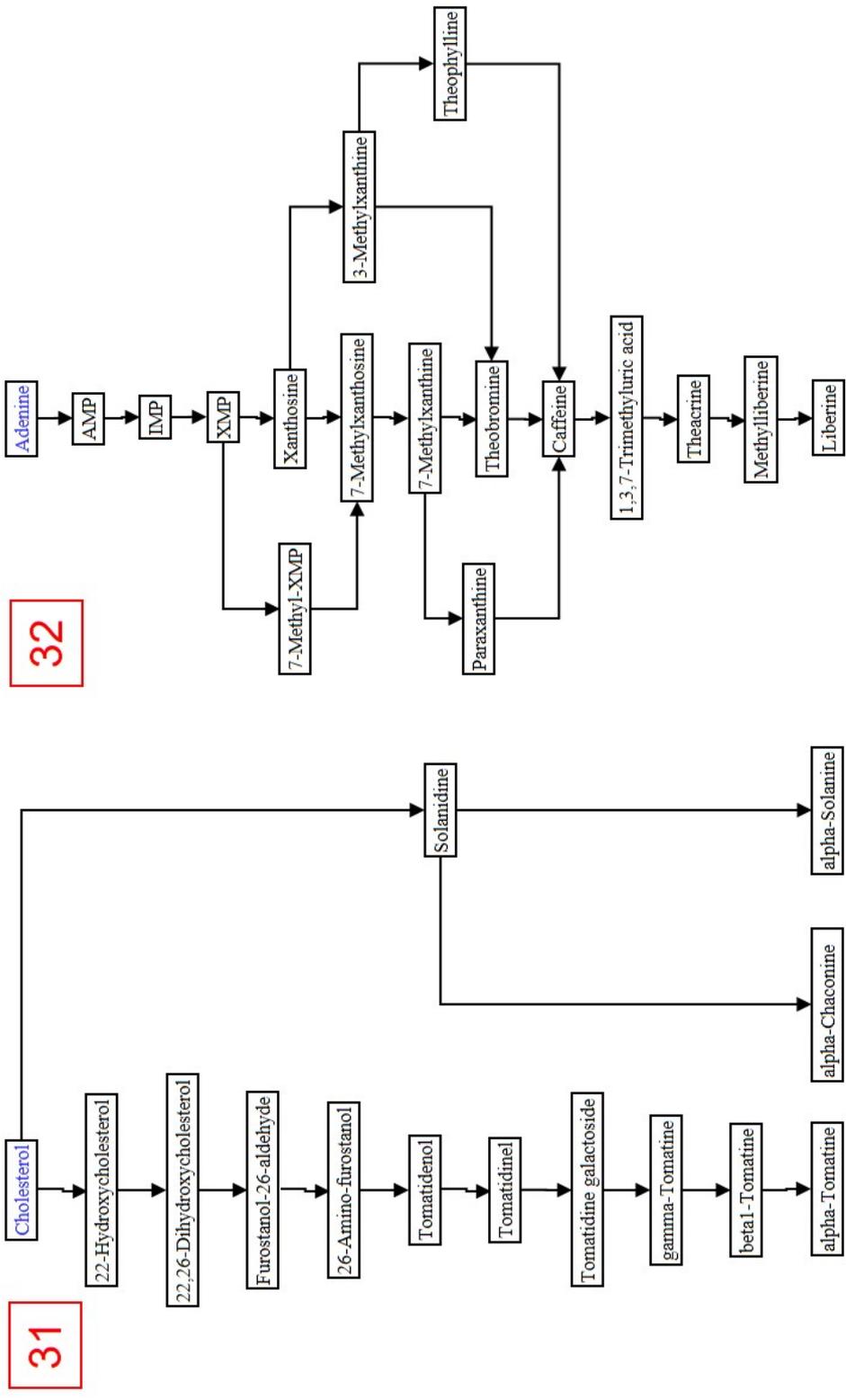


図 2.13 アルカロイド代謝経路 31~32

アルカロイド代謝経路 1 は L-Aspartic acid、L-Proline および L-Arginine を主な出発物質とし、それらに誘導された物質と L-Phenylalanine、L-Glycine、Malonyl-CoA、L-Threonine によって多種多様な化合物を生合成している。

アルカロイド代謝経路 2~10 は L-Tryptophan を主な出発物質とし、アルカロイド代謝経路 2 では Isopentenyl pyrophosphate、アルカロイド代謝経路 3 では Secologanin、アルカロイド代謝経路 4 では L-Proline、アルカロイド代謝経路 5 では Acetyl-CoA および Malonyl-CoA、アルカロイド代謝経路 6 では Oxaloacetate、アルカロイド代謝経路 7 では Dimethylallyl pyrophosphate、アルカロイド代謝経路 8 では Acetoacetyl-CoA、アルカロイド代謝経路 10 では L-Histidine がそれぞれの経路を特徴付けている。また、アルカロイド代謝経路 9 においては、最終生成物が D-serine、L-Alanine、L-Cysteine、L-Valine の 4 種のアミノ酸から誘導されるという特徴がある。

アルカロイド代謝経路 11~14 は L-Tyrosine を主な出発物質としている。アルカロイド代謝経路 11 では L-Phenylalanine、アルカロイド代謝経路 13 では Secologanin、アルカロイド代謝経路 14 は L-Serine がそれぞれ生合成に関与しているが、アルカロイド代謝経路 12 は L-Tyrosine を唯一の出発物質としているのにも関わらず、多くの化合物を生合成する経路である。また、アルカロイド代謝経路 13 では、L-Tyrosine から誘導される Dopamine が生合成に深く関わっている。

アルカロイド生合成経路 15 では L-Lysine を出発物質としている。この経路では Anabasine のみが L-lysine、L-Arginine、L-Aspartic acid によって誘導される化合物であり、その他の物質は全て L-Lysine によって誘導される物質である。

アルカロイド代謝経路 16 は L-Phenylalanine および L-Serine、アルカロイド代謝経路 17 は Malonyl-CoA、Acetyl-CoA および L-Alanine、アルカロイド代謝経路 18 は L-Histidine および L-Leucine を出発物質とした経路である。

アルカロイド代謝経路 19~29 は主に Anthranilate を出発物質としている。これらは比較的小規模な経路ではあるが、Anthranilate が様々な物質との組み合わせにより、多種多様な構造を有する化合物を誘導していることが見て取れる。また、これらの経路を特長付ける出発物質はそれぞれ、O-Methyltyrosine(アルカロイド代謝経路 19)、L-Tryptophan(アルカロイド代謝経路 20)、L-Phenylalanine(アルカロイド代謝経路 21)、L-Tryptophan および L-Alanine(アルカロイド代謝経

路 22)、D-Tryptophan、L-Valine および L-Alanine(アルカロイド代謝経路 23)、D-Tryptophan および L-Alanine(アルカロイド代謝経路 24)、L-Tryptophan および L-Glycine(アルカロイド代謝経路 25)、L-Glycine(アルカロイド代謝経路 26)、L-Proline(アルカロイド代謝経路 27)、Malonyl-CoA(アルカロイド代謝経路 28)、L-Glutamine(アルカロイド代謝経路 29) である。

アルカロイド代謝経路 30 は、Tryptophan を誘導する Indole-3-glycerol phosphate およびテルペノイドの重要な中間生成物でもある Geranylgeranyl diphosphate を出発物質としている。

アルカロイド代謝経路 31 および 32 は Cholesterol、Adenine をそれぞれの唯一の出発物質としている。また、Cholesterol は脂肪酸、Adenine は核酸と関与しており、これらは生物の生体内において、広く分布する重要な物質である。

2.4 まとめ

本章では、68 件の文献情報に基づき知識ベースによる 32 のアルカロイド代謝経路の構築を行った。構築されたアルカロイド代謝経路上の化合物数は 478 であり、これは KEGG で公開されている既存の代謝経路情報と比較すると約 200 以上のアルカロイドが追加された形となる。本研究におけるアルカロイド代謝経路上の化合物は Glycine、L-Serine、D-Serine、L-Cysteine、L-Alanine、D-Alanine、L-Leucine、L-Valine、L-Tryptophan、D-Tryptophan、L-Tyrosine、L-Phenylalanine、Anthranilate、Fornylanthranilate、Indole-3-glycerol phosphate、O-Methyltyrosine、L-Asparatic acid、L-Threonine、L-Lysine、L-Arginine、L-Histidine、L-Glutamine、Isopentenyl diphosphate、Dimethylallyl diphosphate、Secologanin、Geranylgeranyl diphosphate、Cholesterol、Acetyl-CoA、Oxaloacetate、Malonyl-CoA、Acetoacetyl-CoA、Adenine の 33 種類の出発物質から誘導される。この後の第 3 章では、構築したアルカロイド代謝経路の情報を用いることで、化合物の環構造に焦点を当てた分類を行うことでモジュールの抽出を試みる。また第 4 章では、グラフ畳み込みネットワークを用いてアルカロイド化合物の構造情報からそれらの経路上の出発物質の予測を試みる。

第 3 章 部分環構造に着目した化合物のアルカロイド代謝経路上へのマッピング

3.1 概要

生物が生産する二次代謝物にはその構造や由来に基づき、様々な分類が過去の研究からなされている [11]。例えば、テルペノイドは五炭素化合物であるイソフレンユニットを構成単位とする有機化合物群、フラボノイドはクマル酸 CoA およびマロニル CoA が重合してできるカルコン派生する有機化合物群である。また、これらの分類を構造あるいはその化学的性質に基づき、さらに細分化した分類が定義されている場合がある。しかしながら、これらの細分化された分類と化学構造および代謝経路との関係性についての体系化は未だに進んでいないのが現状である。そのためシステムズ代謝工学の観点から二次代謝物に着目し、二次代謝物の構造情報による代謝パスウェイの構造単位を抽出する必要がある。この化学構造および代謝経路を考慮した二次代謝物の分類は、二次代謝産物の機能獲得メカニズムの解釈につながる代謝物の活動に重要な手掛かりを与える。

本章では、第 2 章で構築した図 2.2~図 2.13 のアルカロイド代謝経路から、二次代謝物中に存在する環構造に焦点を当てた代謝モジュールの抽出を行い、文献情報に基づく既知の分類との比較をすることで両者の関係性について考察を行う。またこれを実現するため、既存の化学物質の類似性評価手法が代謝経路を考慮した分類に適していないことに着目し、化合物が有する環構造に基づいた分類手法を提案する。

3.2 手法

第 2 章で構築したアルカロイド代謝経路上に存在する 478 のアルカロイド化合物に対して化学構造及び代謝経路を考慮した分類を行う。このとき、構築されたアルカロイド代謝経路上の代謝物についてこれらを原子をノード、原子間の結合をエッジとして表現したグラフとして扱い、二次代謝物が有する環構造に着目して各々の分類を行う。これらの一連の流れを図 3.1 に示す。

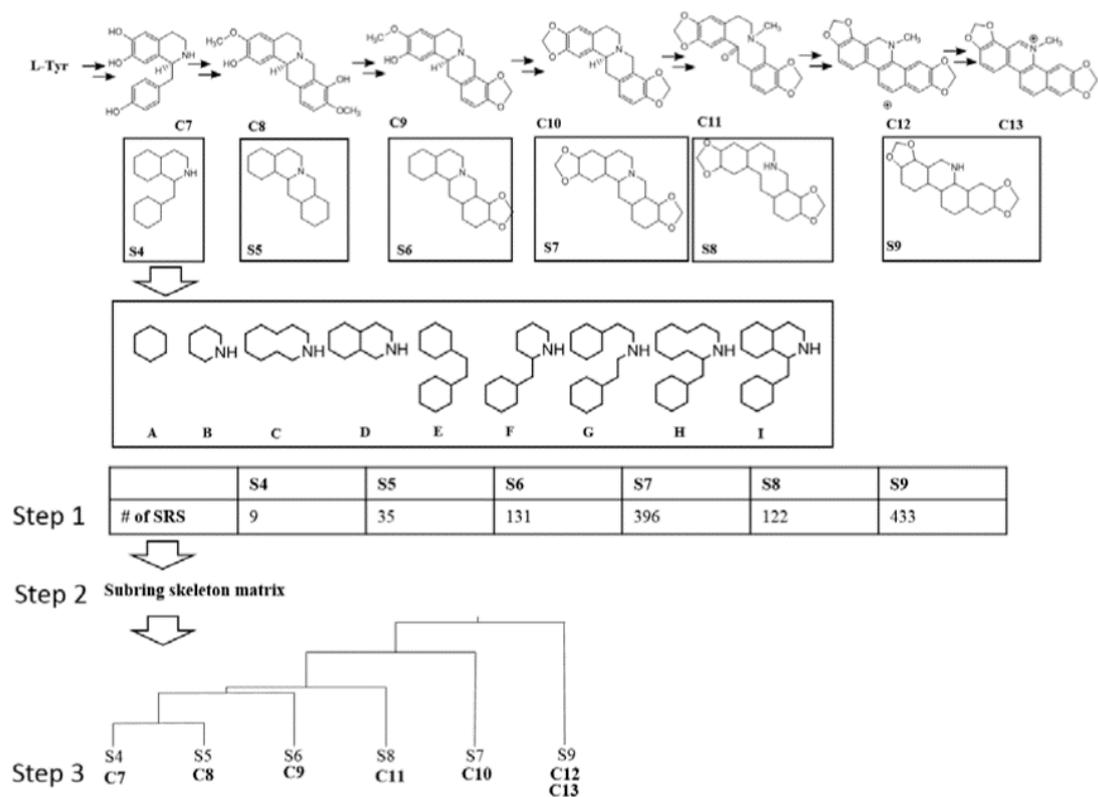


図 3.1 本研究における環構造に基づく化合物分類の流れ

Step1. 各化合物から環構造を抽出、Step2. 環構造を特徴として化合物を表現しバイナリ行列を作成、Step3. 階層的クラスタリングにより化合物を分類

3.2.1 計算機上における化合物の構造表現

化学構造を計算機上でどのように表記するかは重要な問題であり、現在に至るまで数多くの表記法が開発されてきた。それらの多くは、化学構造は原子をノード、結合をエッジとしたグラフとして表現されており、化学構造とそのグラフ表現は相互に変換が可能であるという利点がある。

化学構造の表記方法の一つである MDL Molfile フォーマットは、化合物の構造情報を扱う際のデファクトスタンダードであり、前述した化学構造のグラフ表現に加えて原子の座標情報も保持している (図 3.2)。しかしながら、化合物の構造類似性や同一性を評価する場合においては、原子の座標情報が保持されていない InChI

コードや後述するフィンガープリント法によって構造を記述することがほとんどである。これは立体構造の類似性を評価することの困難さに加えて、化合物中の単結合まわりで起こる回転などによって原子の空間的配置が変化し、立体構造が変化するためである。

立体異性体を考慮する際に原子の座標情報は重要であるが、InChI コードなどは構成する原子の座標情報は保持していないものの立体異性体を区別可能な表記方法である。立体異性体とは、構成する原子の種類および原子間の繋がり方は同一であるものの原子の空間配置が異なる化合物である。

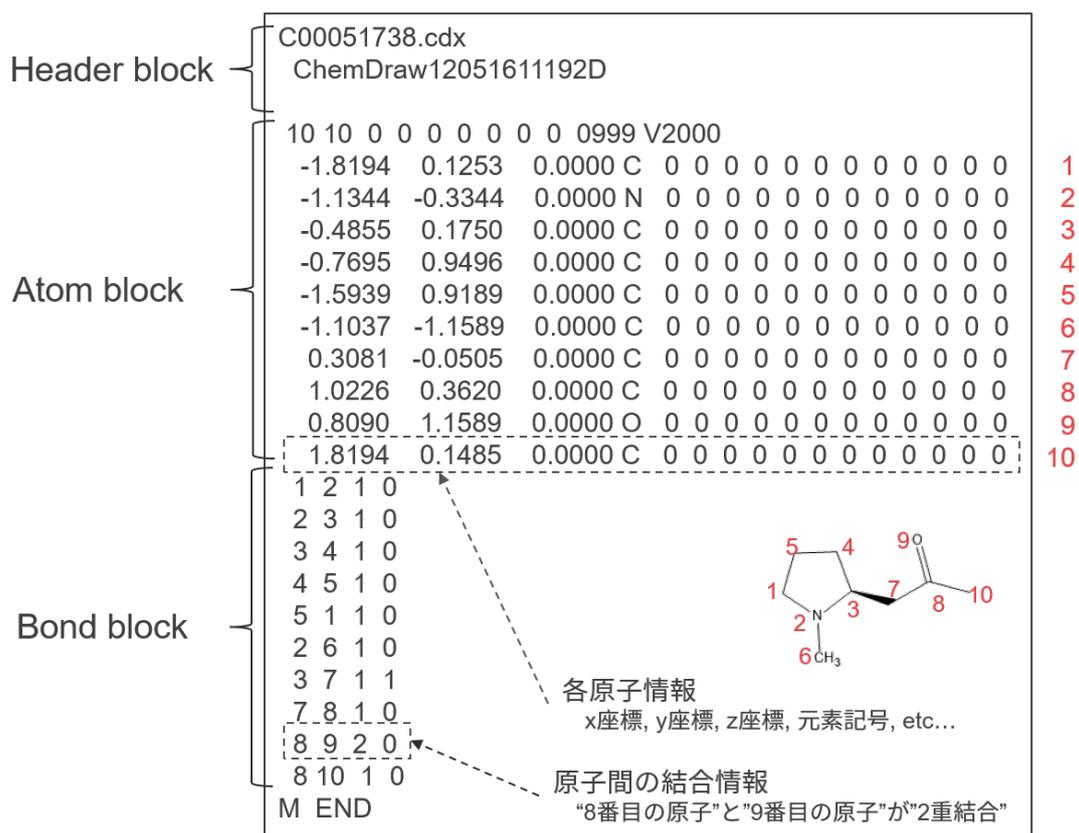


図 3.2 MDL Molfile における化学構造の記述例

3.2.2 化合物の類似性評価

化合物間の類似性評価手法は主に定量的構造活性相関の分野において発展してきた。化合物の類似性評価の複雑さは二つの化合物の構造が同一か否かを判定する問題を考えたとき、それがグラフの同型判定問題と等価になることからもいかに困難な問題であるかが分かる。そのため化学の分野では化合物間の類似性を評価するための様々な手法が開発されてきた。現在、化合物間の類似性評価において多くの手法がフィンガープリント法を用いている。本研究における提案手法とも関連するため、まずフィンガープリント法についての説明を記述する。

フィンガープリント法

フィンガープリント法では、事前に Fragment library と呼ばれる固定長の特徴セットを定義し、各化合物がどの特徴を保有しているかを 0-1 特徴ベクトルで表現する (図 3.3)。Fragment library として良く用いられるものは PubChem フィンガープリント (881bits) [83]、CDK フィンガープリント (1024bits)、MACCS フィンガープリント (166bits) [84]、Atom pairs フィンガープリント (780bits) [85] 等が挙げられる。また、PubChem フィンガープリントの Fragment library に含まれる特徴の一部を図 3.4 に示す。

Fragment library が事前に設計されているため、フィンガープリントに基づく評価手法は二つの化合物間の類似性を高速に評価することが可能である。しかしながら、Fragment 間に連続性がないために評価の曖昧さが高まる点が過去の研究で指摘されている。また、同様の理由で代謝経路を考慮した分類には適していない。

3.2.3 提案手法-部分環構造に基づく分類

前述した通り、既存の類似性評価手法として高頻度で用いられるフィンガープリント法では各 Fragment 間に連続性がない。そのため、化学構造および代謝経路を考慮した二次代謝物の分類という目的にはそぐわない。そこで本研究では代謝経路を考慮した分類を行うために化合物中の部分環構造に基づく分類手法を提案する。

ここでの部分環構造とは、化合物中に含まれる全ての部分構造のうち、Bemis-Marco らによって定義された Molecular Framework [86] の条件を満たす部分構造

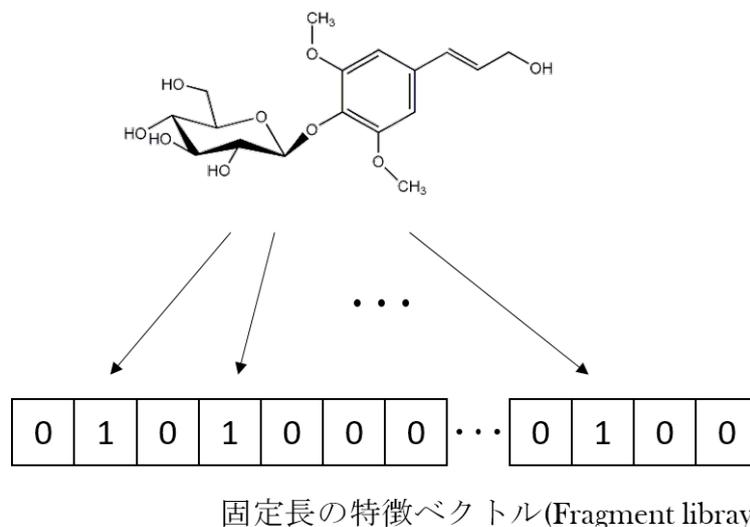


図 3.3 フィンガープリント法による化合物のベクトル表現例

と定義する。Bemis-Marco らは図 3.5 に示すようなベンゼン環やピリジン環などの環構造 (Ring) 間を鎖状構造 (Linker) で繋ぎ合わせたものを Molecular Framework と定義している。一例としてフラボノイドから部分環構造を抽出した結果を図 3.6 に示す。この例では、フラボノイドから 9 つの部分環構造が抽出され、これらは重複しておらず、いずれも Bemis-Marco らが定義した Molecular Framework の条件を満たしている。

本研究においては、各化合物から部分環構造を抽出する際に二重結合および三重結合 (不飽和結合) は考慮していない。この理由は、二重結合および三重結合を考慮した際に部分環構造の多様性が非常に大きくなってしまい、計算量的に膨大なコストがかかるためである。また、多くの化合物において頻出する環構造であるベンゼン環は炭素原子間の結合は全て等価であるが、この化学構造を表現するときには通常、炭素間を二重結合と三重結合が繰り返されているような構造として表現する。ベンゼン環をはじめとして、本来等価であるはずの構造を区別することが困難になることも不飽和結合を考慮していない理由である。

次に化合物から部分環構造を抽出するための疑似アルゴリズムを図 3.7 に示す。ここで化合物中の各原子をノード x 、各原子間をつなぐ結合をエッジ e としたグラ

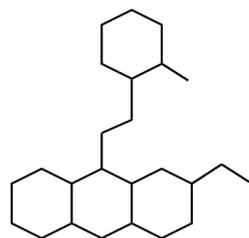
Bit Position	Bit Substructure
0	>= 4 H
1	>= 8 H
2	>= 16 H
3	>= 32 H
4	>= 1 Li
5	>= 2 Li
⋮	⋮
115	>= 1 any ring size 3
116	>= 1 saturated or aromatic carbon-only ring size 3
117	>= 1 saturated or aromatic nitrogen-containing ring size 3
118	>= 1 saturated or aromatic heteroatom-containing ring size 3
119	>= 1 unsaturated non-aromatic carbon-only ring size 3
120	>= 1 unsaturated non-aromatic nitrogen-containing ring size 3
121	>= 1 unsaturated non-aromatic heteroatom-containing ring size 3
122	>= 2 any ring size 3
⋮	⋮
874	SC1C(Br)CCC1
875	NC1C(N)CCC1
876	NC1C(Cl)CC1
877	NC1C(Br)CCC1
878	ClC1C(Cl)CCC1
879	ClC1C(Br)CCC1
880	BrC1C(Br)CCC1

図 3.4 PubChem フィンガープリントの各ビットが表す特徴の一部例

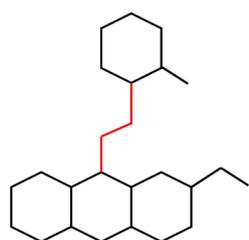
フ G として表現をしている。

このアルゴリズムでは、まず化合物中の側鎖を除去したグラフを生成する (DelChain)。その後グラフ中の任意のエッジを削除し、その結果に対して側鎖の除去を繰り返すことにより化合物中の全ての部分環構造をグラフセットの形式で生成している (DetRing)。

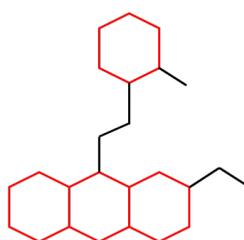
また、この状態では抽出した部分環構造に重複した構造が含まれる可能性が



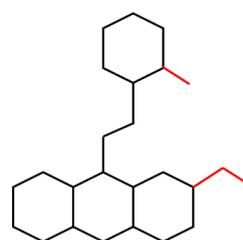
化合物のグラフ表現



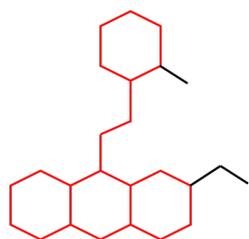
鎖状構造(Linker)



環構造(Ring System)



側鎖(Side chain)



Molecular Framework

図 3.5 化合物のグラフ表現および Bemis-Marco らによる構造の定義

あるため、InChI(International Chemical Identifier) コードを生成することで異同判定を行い構造の重複を回避している。InChI は IUPAC(International Union of Pure and Applied Chemistry) によって開発された化合物構造情報の表現手法であり [87]、構造情報の表現方法としてよく用いられている SMILES(Simplified molecular input line entry system) [88] とは異なり、IUPAC の命名法を独自に用いることで化合物構造を一意に表現することが可能である。

部分環構造に基づく類似性評価手法では、以上の手順によって抽出された部分環

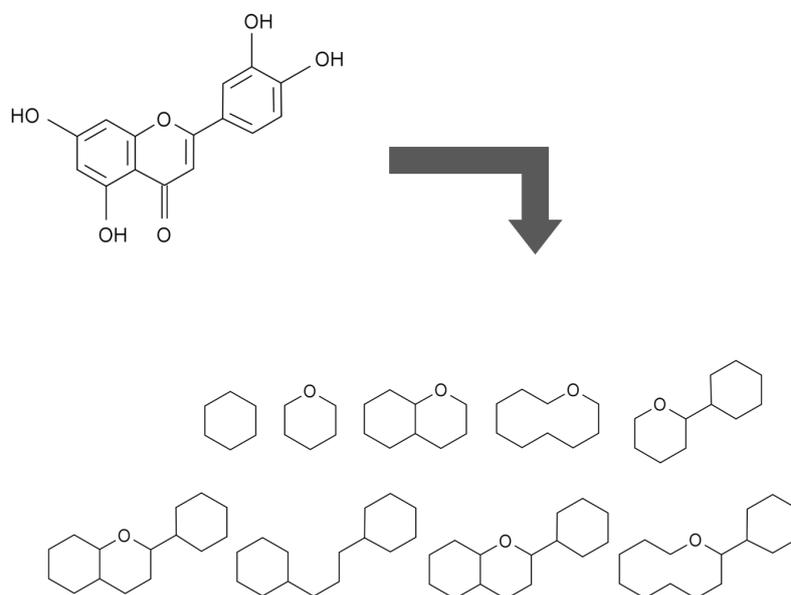


図 3.6 フラボノイドから部分環構造を抽出した例

構造を特徴セットとしたベクトルとして各化合物を表現する。図 3.8 においては、列が化合物、行が環構造を表現している。このとき、化合物が任意の環構造を有している場合は 1 の値を、有していない場合は 0 の値を取る。二つの化合物間の類似性を評価する際はフィンガープリント法の場合と同様に、各々の化合物の構造情報を任意の部分構造を選択してベクトルとして特徴を表現し、それらのデータ間距離を算出する。

3.2.4 二次代謝物の分類

各化合物を抽出した部分環構造に基づく特徴ベクトルとして表現し、これら进行分类するために階層的クラスタリングを行う。各ベクトル間の距離はユークリッド距離を用い、また、階層的クラスタリングにはワード法を用いる。

```

DetRing(G)
X := DelChain(G)
If Xに辺がない then return  $\phi$ 
  R := { X }; D := { X }
  while D  $\neq \phi$ 
    S := Dから1個取り出す
    for all Sの任意の辺e
      G' := Sからeを除去したグラフ
      G' := DelChain(G')
      if E(G')  $\neq \phi$  & G'は連結グラフ & G'  $\notin$  R then
        R := R  $\cup$  { G' }; D := D  $\cup$  { G' }
  return R

DelChain(G)
G' := G
for all 点x in G'
  if xの次数 == 1 then
    G'から点xとxに接続する辺eを除く
  else if xの次数 == 0 then
    G'から点xを除く
If 点が除かれた then 先頭に戻る
return G'

```

図 3.7 化合物の構造情報から部分環構造群を抽出するための疑似アルゴリズム

3.3 結果及び考察

3.3.1 部分環構造に基づくアルカロイド化合物の分類

まずはじめに各化合物に対する予備解析として、それらが有する部分環構造の種類を集計を行う。例えば図 3.6 で示したように、フラボノイドに関して環構造の抽出を適用した場合、9 種類の部分環構造が得られる。

第 2 章で構築したアルカロイド代謝経路上に存在する 478 の化合物に対して部分

二次代謝物	環構造1	環構造2	...	環構造N
C1	1	0	...	1
C2	1	1	...	0
C3	0	1	...	1
:	:	:	...	
C _N	1	0	...	0

図 3.8 各化合物を部分環構造に基づき表現したバイナリー行列

環構造の抽出を行ったところ、2546 のユニークな環構造が得られた。図 3.9 は抽出された部分環構造の出現頻度を示している。ここで、横軸が化合物数、縦軸が部分環構造数であり、化合物数を x 、部分環構造数を y とした際、 x の化合物に共通して含まれる部分環構造は y ということを表示している。また、これらの部分環構造それぞれの二次代謝物における多様性をみると化合物数が少ないほどそれが有する部分環構造は多様性があり、化合物数が多いほどそれが有する部分環構造はより多くの二次代謝物に共通して含まれるものである。縦軸上に存在する点は単一の代謝物のみが有する環構造の種類を表しているが、この時縦軸上に点は存在しないため、抽出された 2546 の部分環構造はいずれも複数の化合物に含まれていることがわかる。

次にアルカロイドパスウェイマップ上に存在する 478 の化合物を抽出された 2546 の部分環構造に基づいた特徴ベクトルで表現し、これらを 29 のクラスタ (クラスタ ID= 1~29) に分類することで階層的クラスタリングとして可視化した。その結果を図 3.10 に示す。

このとき全てのアルカロイド化合物は 187 種の構造によって構成されており、これを環骨格として定義した。187 種類の化合物を構成する環骨格と得られたクラスタとの対応関係を図 3.11 ~図 3.15 に示す。クラスタ 27 に関しては属する化合物の構造多様性が高く、さらにクラスタを細分化して議論するため後述する。

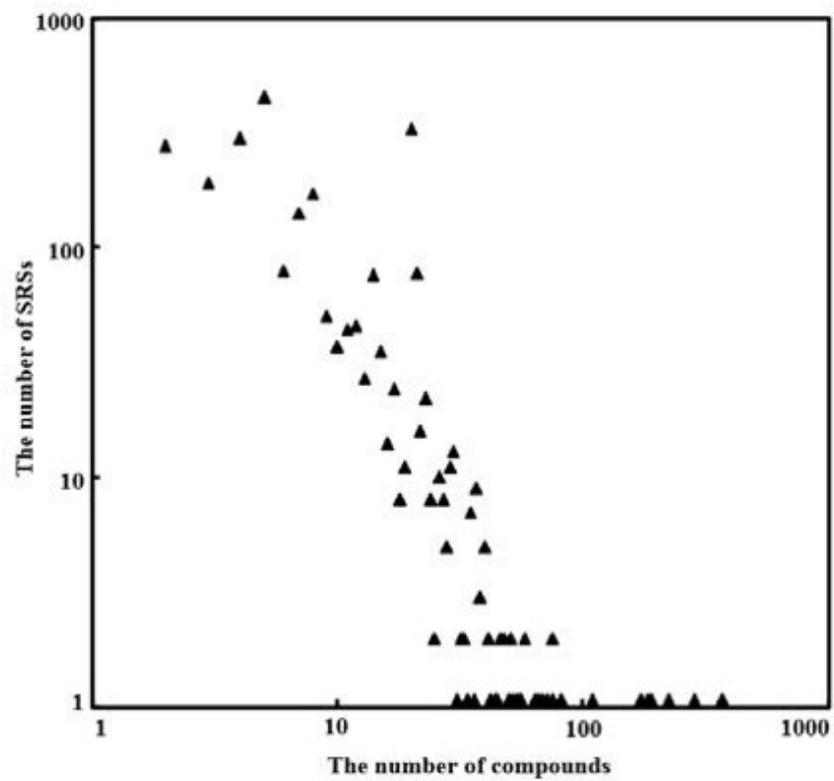


図 3.9 アルカロイド代謝経路上に存在する化合物における部分環構造の出現頻度

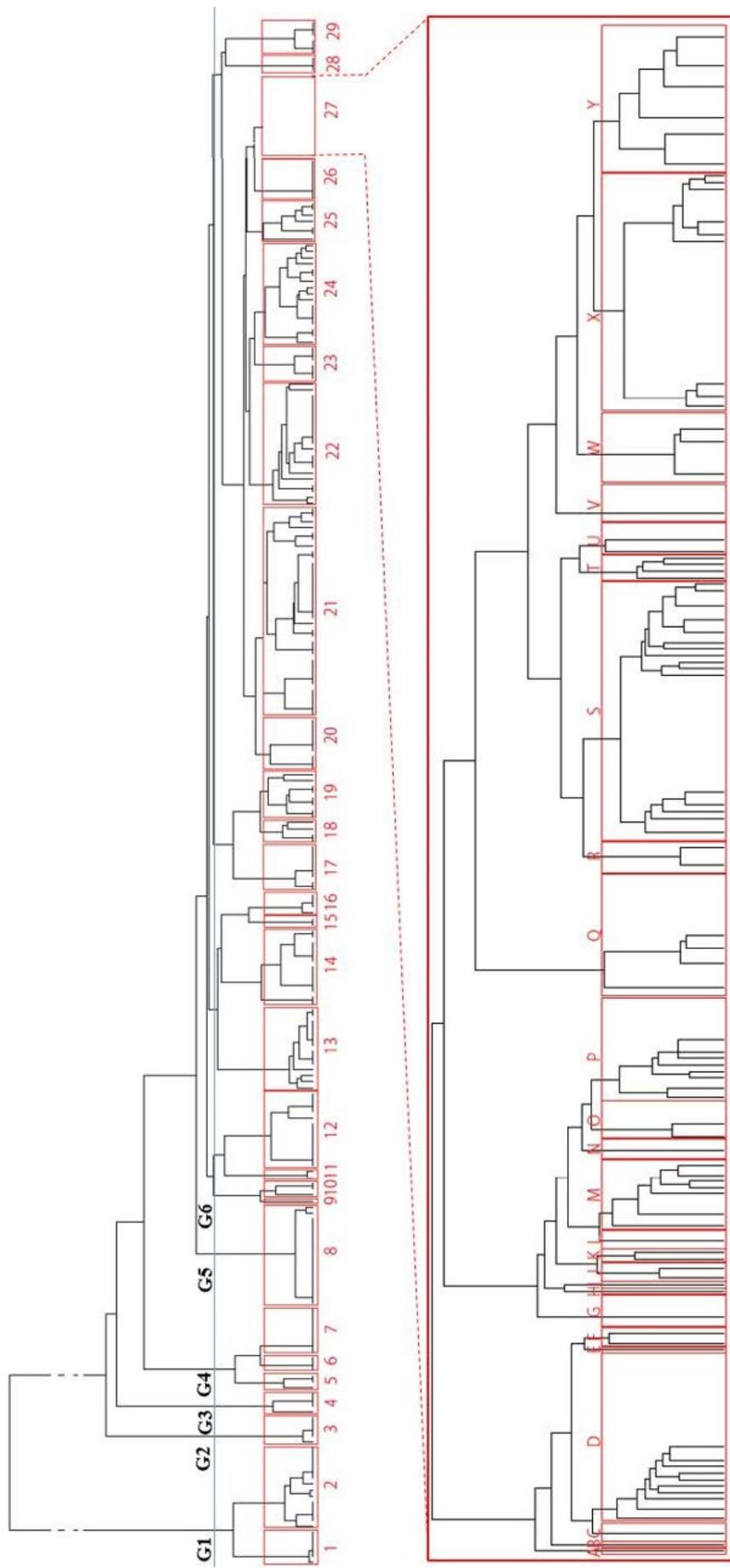


図 3.10 アルカロロイド化合物の階層的クラスタリング

図 3.10 で示した通り、クラスタリングの結果を大別すると G1(クラスタ 1 および 2)、G2(クラスタ 2)、G3(クラスタ 4)、G4(クラスタ 5 クラスタ 7)、G5(クラスタ 8)、G6(クラスタ 9~29) の 6 グループに分類される。

G1 および G2 はともに Indole-diterpenes に関連しており、そのうち G1 は Paxilline、Terpendoles および lolitrems に、G2 は Strictosides と関与している。これらは両者ともに 6~9 の五員環または六員環で構成される多環式化合物であり、一つの窒素原子を有しているという特徴を持つ。

G3(クラスタ 4) に分類される化合物は Alpha-solanine、Alpha-chaconine および Alpha-tomatine を含む Glucosides と関連しており、これらは Steroidal アルカロイドとして割り当てられる。G4(クラスタ 5~7) に含まれる化合物は窒素原子を 2 つ以上有する Vindoline 誘導体として特徴づけられ、また、G5(クラスタ 8) は Ergot アルカロイドと対応付けることが出来る。

その他のアルカロイド化合物が属する G6(クラスタ 9~29) は各々が多様な環骨格を有している。クラスタ 9 から 20 は次のように特徴づけることができる：Aglycones(クラスタ 9)、Glucosides(クラスタ 10)、Bis-isoquinolines(クラスタ 11)、Morphinans(クラスタ 12)、Quinazoline 誘導体 (クラスタ 13)、Indolizidine(クラスタ 14)、Polyneuridines(クラスタ 15)、Ajmalicines(クラスタ 16)、Berberine および Relatives(クラスタ 17~19)、Ipecac アルカロイド (クラスタ 20)、Isoquinolines(クラスタ 21)、Beta-carbolines(クラスタ 22)、Chateoglobosins(クラスタ 23)、Roquefortines および Acetyl-Aszonaleins(クラスタ 24)、Emindoles(クラスタ 25)、Quinolizidines(クラスタ 26)、Iboganes(クラスタ 28)、Communesins(クラスタ 29)。

クラスタ 27 はアルカロイドパスウェイマップ上に存在する化合物の内、おおよそ半数の化合物が属するクラスタである。従って、このクラスタを部分環構造と特徴づけるためにさらに 25 のサブクラスタに分割した。187 種類の化合物を構成する環骨格と得られた 25 のサブクラスタとの対応関係を図 3.16~図 3.19 に示す。

この 25 のサブクラスタ中に含まれるアルカロイド化合物は、その他のクラスタに属する化合物と対比させると、比較的単純な複素環式化合物であることが確認できる。A、B、C、D、E、F、H の 6 つのサブクラスタはインドール (C₈H₇N) の構造を環骨格として有しており、特に A は Brevianamides、B は Cyclopiazonate

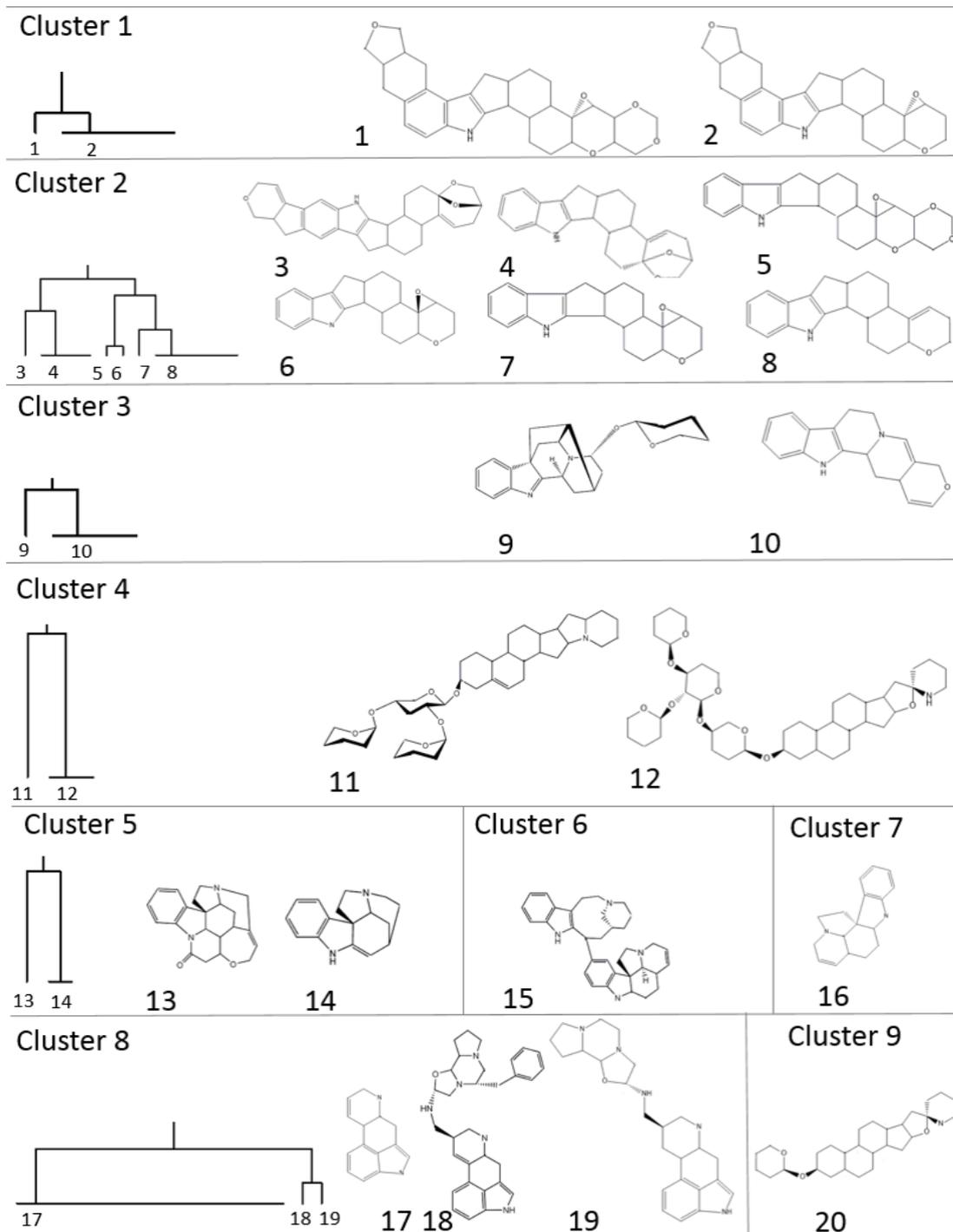


図 3.11 各クラスタ (1~9) とそれらを構成する環骨格

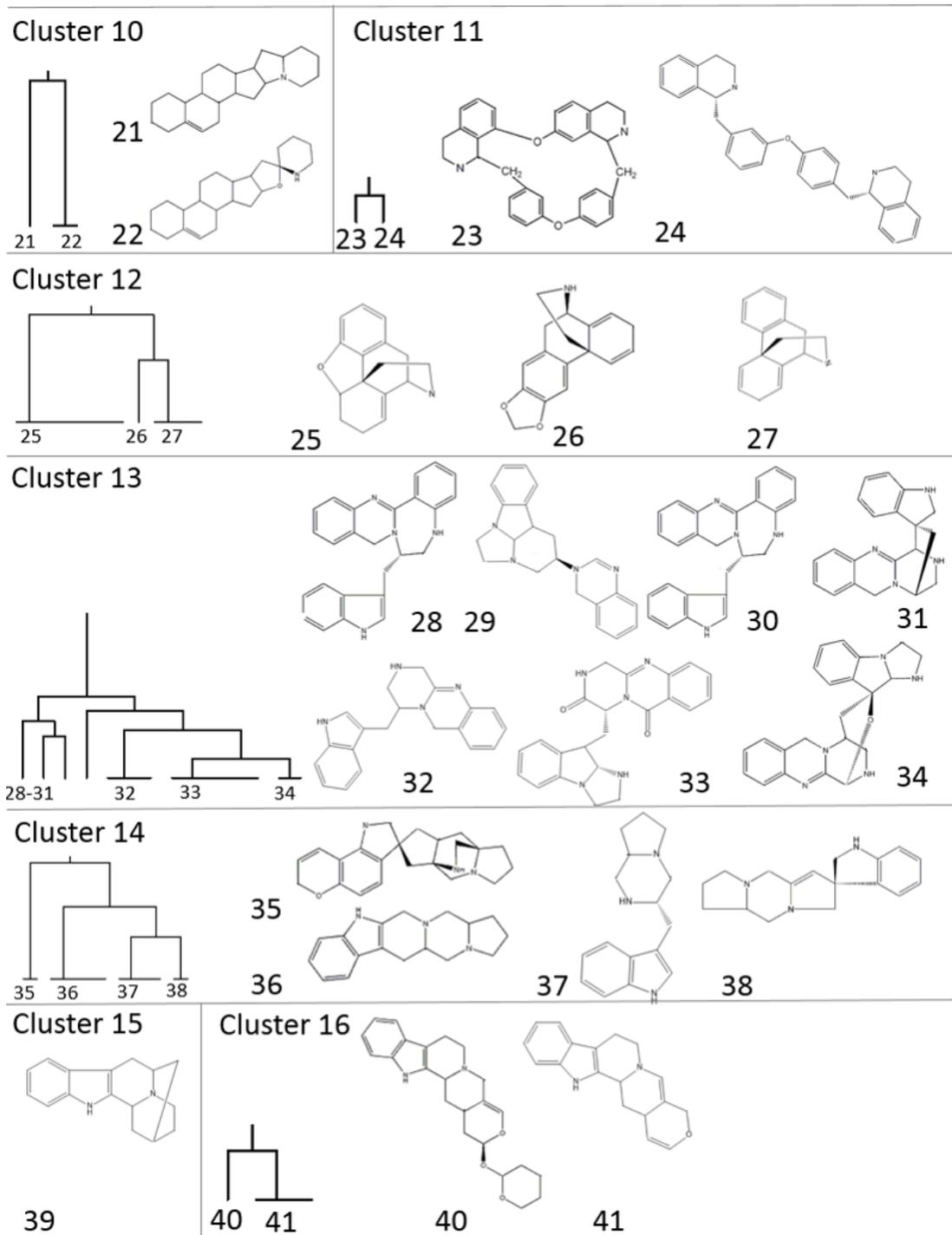


図 3.12 各クラスタ (10~16) とそれらを構成する環骨格

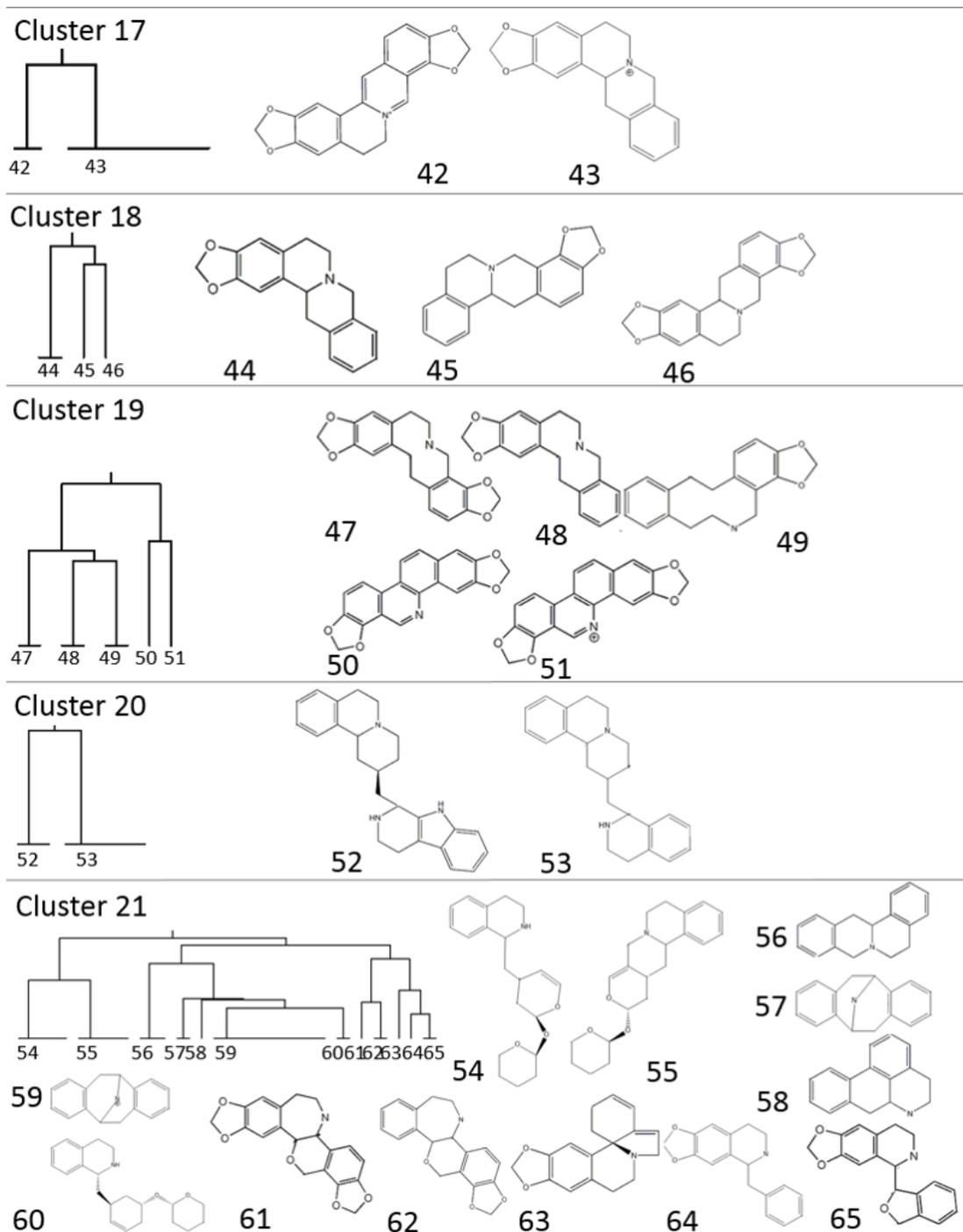


図 3.13 各クラスター (17~21) とそれらを構成する環骨格

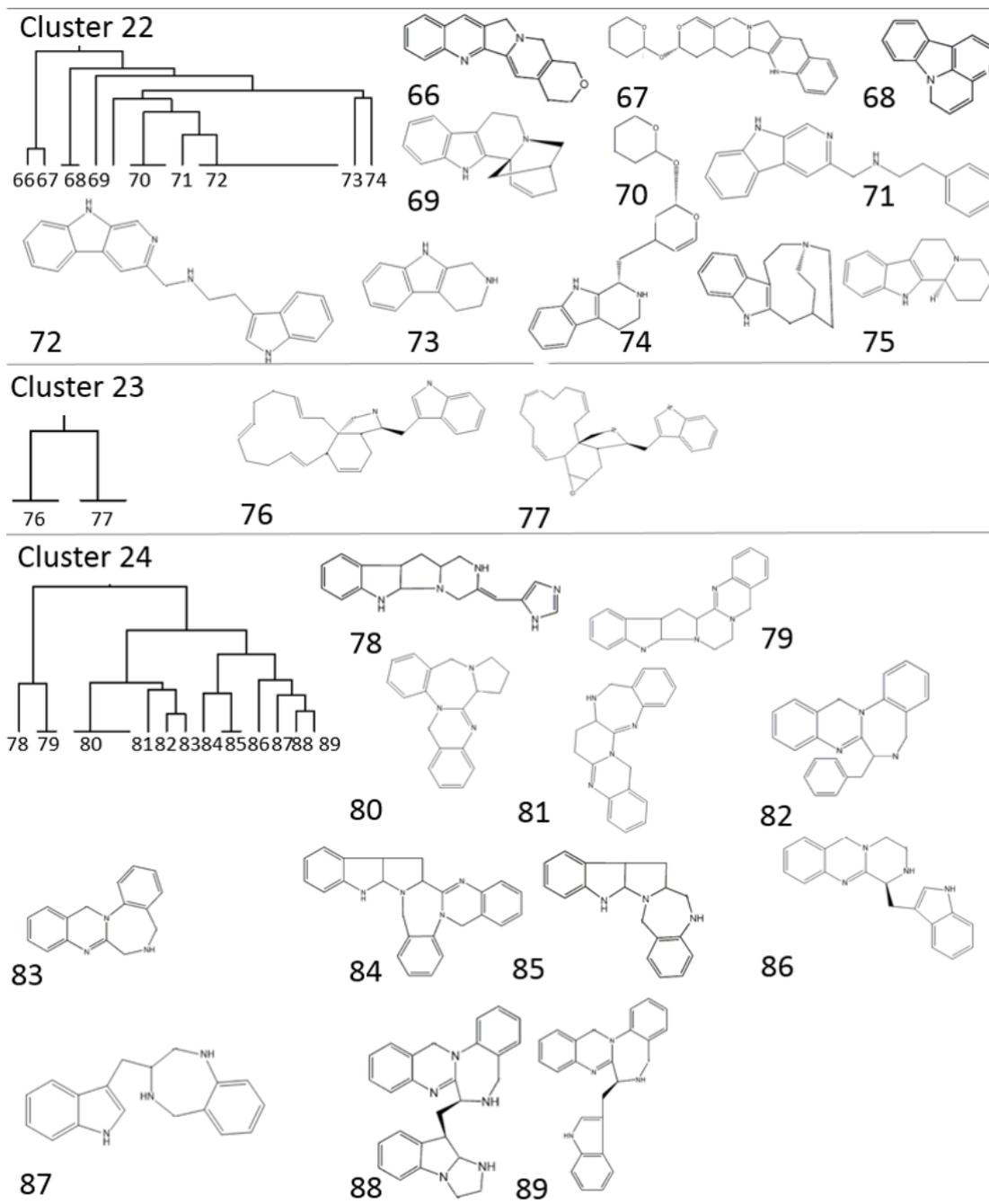


図 3.14 各クラスと (22-24) それらを構成する環骨格

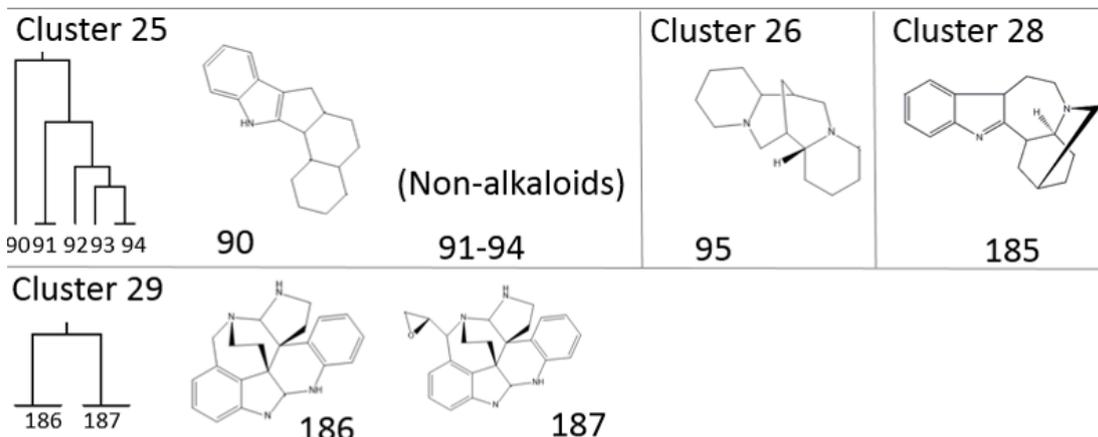


図 3.15 各クラスタ (25~29) とそれらを構成する環骨格

誘導体および Chanoclavin 誘導体と特徴づけることが出来る。Cyclopiazonate 誘導体は五員環を有する Ergot アルカロイドであり、Chanoclavin 誘導体は三員環を有する Ergot アルカロイドである。また、E は Chimonanthines、H は Iboga アルカロイドとの関連性があることが分かる。さらに、それぞれを特徴付ける環骨格は各サブクラスタを見ることで次のように確認することが出来る：Ergolines(サブクラスタ B および C)、Cinchona アルカロイド (サブクラスタ G)、Iboga アルカロイド (サブクラスタ H および I)、Lycopodium アルカロイド (サブクラスタ J および K)、Isoquinolines(サブクラスタ L)、Lupine アルカロイド (サブクラスタ M)、Quinolones(サブクラスタ N)、Quinolones(サブクラスタ O)、Acridines および Quinolones(サブクラスタ P)、Purine アルカロイド (サブクラスタ Q)、Benzyloisoquinoline アルカロイド (サブクラスタ T)。

Sparteine アルカロイドおよび Lupanine アルカロイド (環骨格 122、124、125)、Lupinine アルカロイド (環骨格 126) は Lupaine アルカロイドおよび Matrine アルカロイド (環骨格 120 および 121) の 3 つの多環系が全てサブクラスタ M に分類されている。

前述した結果からも確認できるように、このようにアルカロイド化合物は部分環構造あるいは環骨格に基づいて分類をした結果は、これまで他の研究で報告されたような一般的に定義された分類との合致が多くみられる。また、提案手法においてはこれらを系統的に分類することが可能であると結論付けることが出来る。

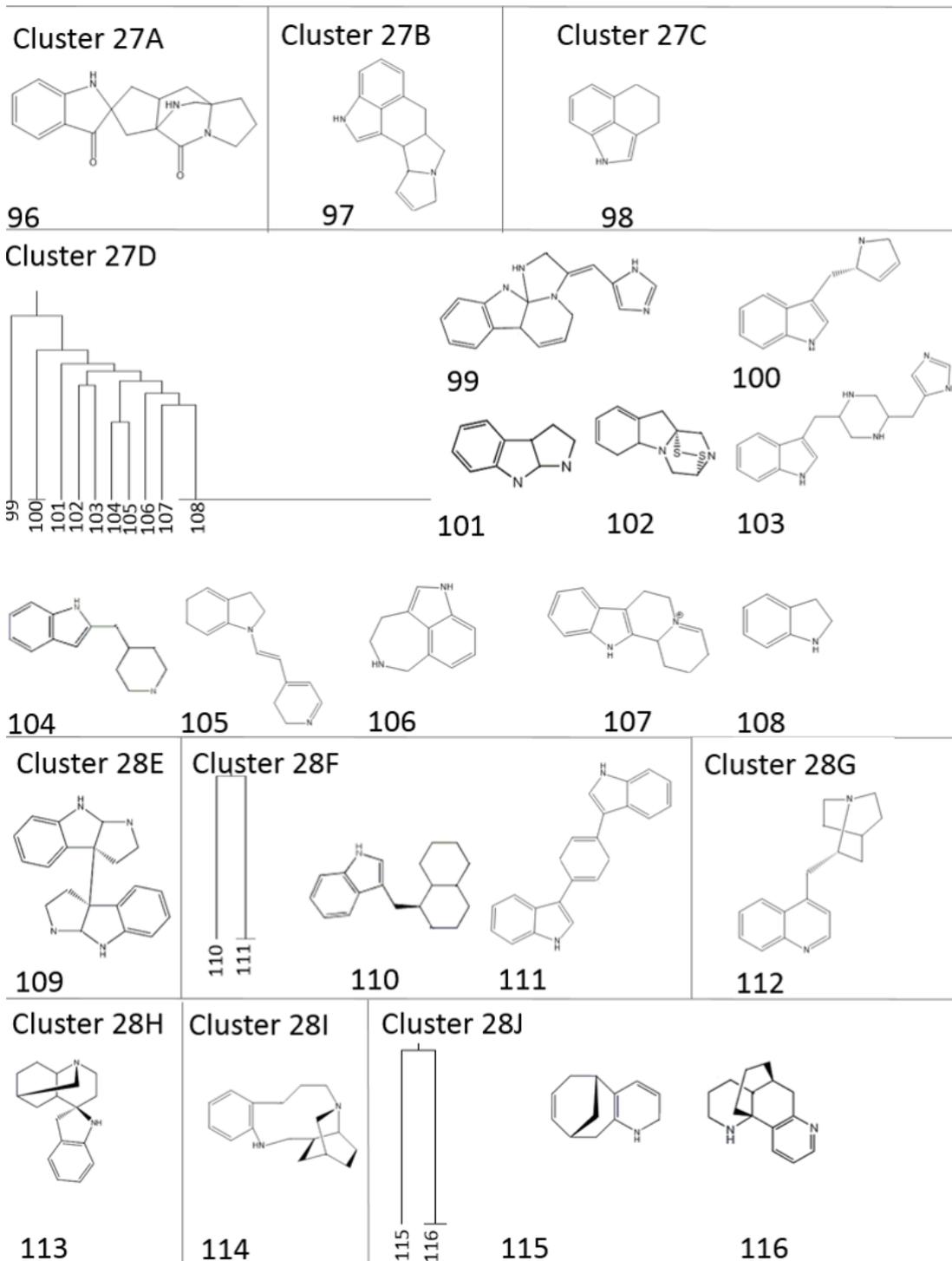


図 3.16 各クラス (27A~27J) とそれらを構成する環骨格

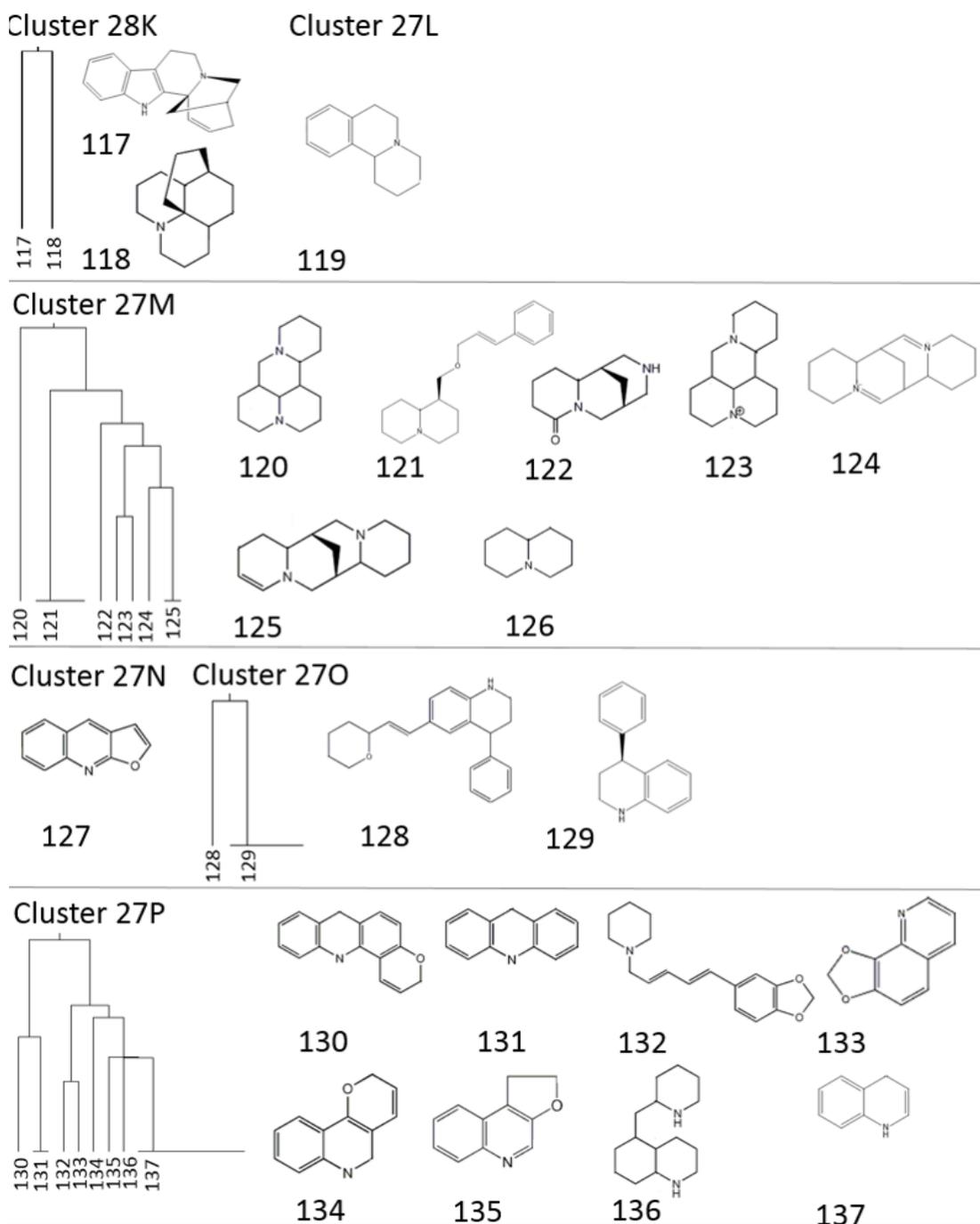


図 3.17 各クラスタ (27K~27P) とそれらを構成する環骨格

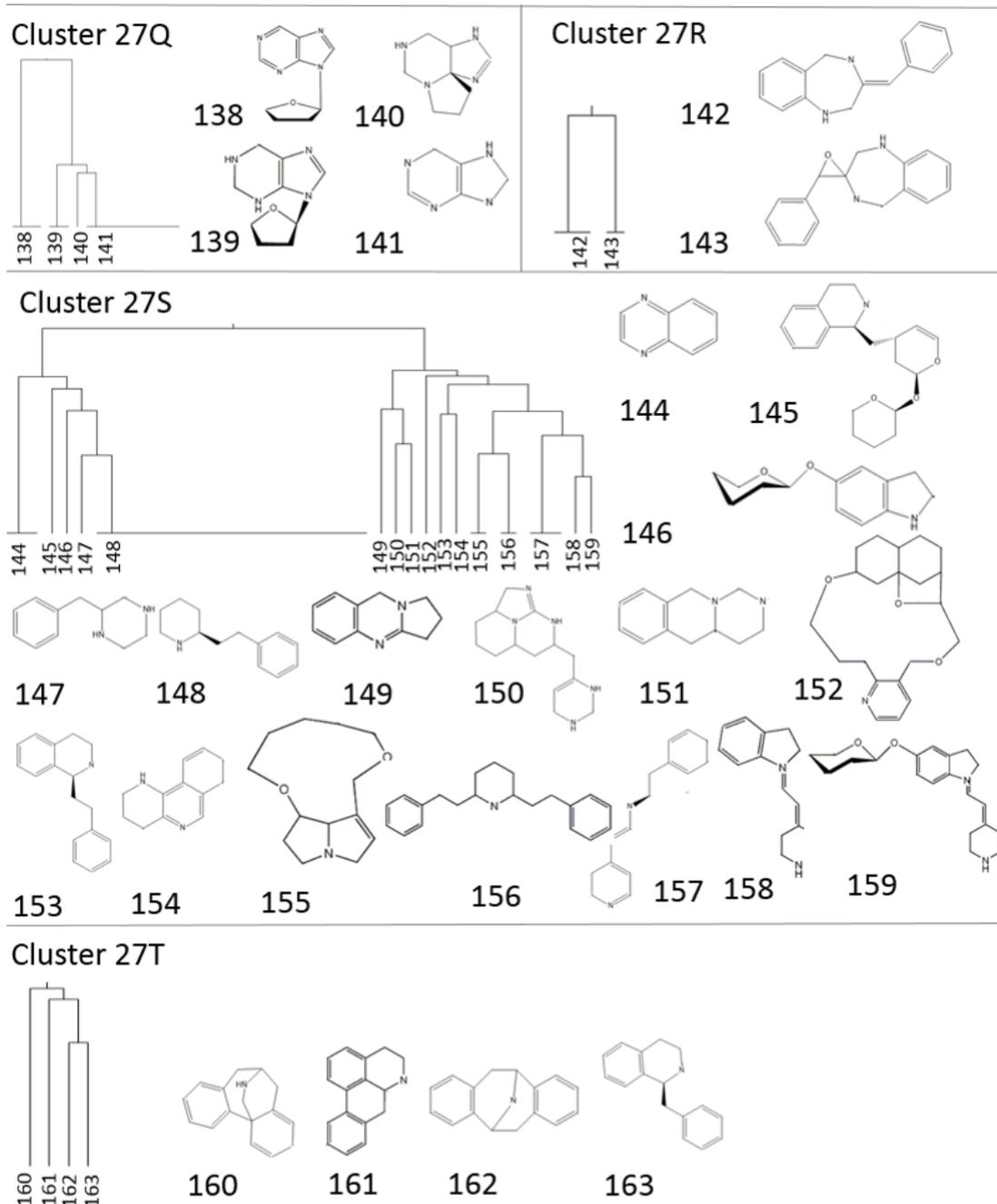


図 3.18 各クラスター (27Q~27T) とそれらを構成する環骨格

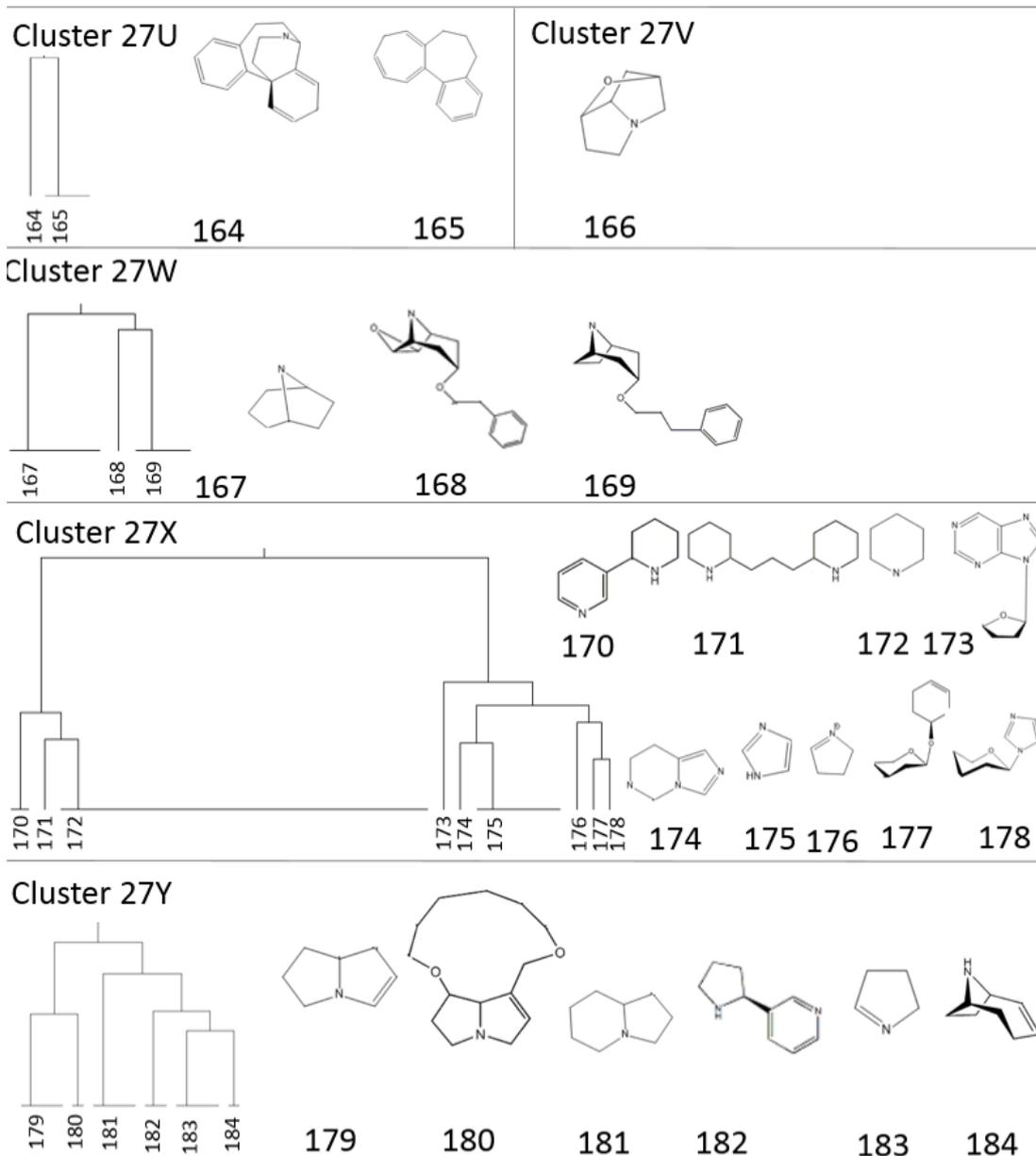


図 3.19 各クラスター (27U~27Y) とそれらを構成する環骨格

3.3.2 環構造に基づくアルカロイド代謝経路の代謝モジュール抽出

前節においては、アルカロイド化合物を部分環骨格に基づいて分類し、一般的なアルカロイド群との関連性を確認した。次にここでは、第2章で構築したアルカロイド代謝経路における環構造とモジュールとの関係を見出すことを試みる。

まず初めに、前節における化合物の分類をアルカロイド代謝経路上にマッピングすることで、分類結果と代謝経路との関係性を確認した。この結果を表としてまとめたものを図3.20に示す。この表の行は前節における環骨格に基づく分類の結果(29のクラスタと25のサブクラスタ)を、列は第2章で構築した31のアルカロイド代謝経路を表している。

この表からクラスタ13、クラスタ22、クラスタ24、クラスタ25、クラスタ27を除くクラスタは全て単一のパスウェイと関連していることが確認できる。このことから環骨格に基づくクラスタは、アルカロイド生合成経路と高い相関関係があるといえる。このような傾向はクラスタ27におけるサブクラスタでも同様にみられ、25のサブクラスタのうち10のサブクラスタは単一のアルカロイド生合成経路との関連性がある。そこでこの節では、前節におけるクラスタリング結果の観点から、アルカロイド代謝経路上で抽出されたモジュールに関して考察を行う。

図3.21に示すようにアルカロイド代謝経路1上では、6つのアミノ酸(Asparatic acid、Proline、Arginine、Glycine、Threonine、Phenylalanine)により様々な化合物を生成する。27のサブクラスタのうち5つのサブクラスタ(S、Q、W、X、Y)に属する化合物はアルカロイド代謝経路1に属する。その中でも特にサブクラスタX、V、Wはピリジン、ロリンおよびトロパンアルカロイドとそれぞれ関連している。

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32				
1																																			6	
2																																			14	
3				5																																
4																																				4
5				3																																
6				3																																
7				3																																
8				3																																
9				17																																
10																																				1
11																																				3
12																																				
13																																				
14																																				
15																																				
16																																				
17																																				
18																																				
19																																				
20																																				
21																																				
22																																				
23																																				
24																																				
25																																				
26																																				
28																																				
29																																				
A																																				
B																																				
C																																				
D																																				
E																																				
F																																				
G																																				
H																																				
I																																				
J																																				
K																																				
L																																				
M																																				
N																																				
O																																				
P																																				
Q																																				
R																																				
S																																				
T																																				
U																																				
V																																				
W																																				
X																																				
Y																																				
Z																																				

図 3.20 各アルカロイド代謝経路とクラスタの関係

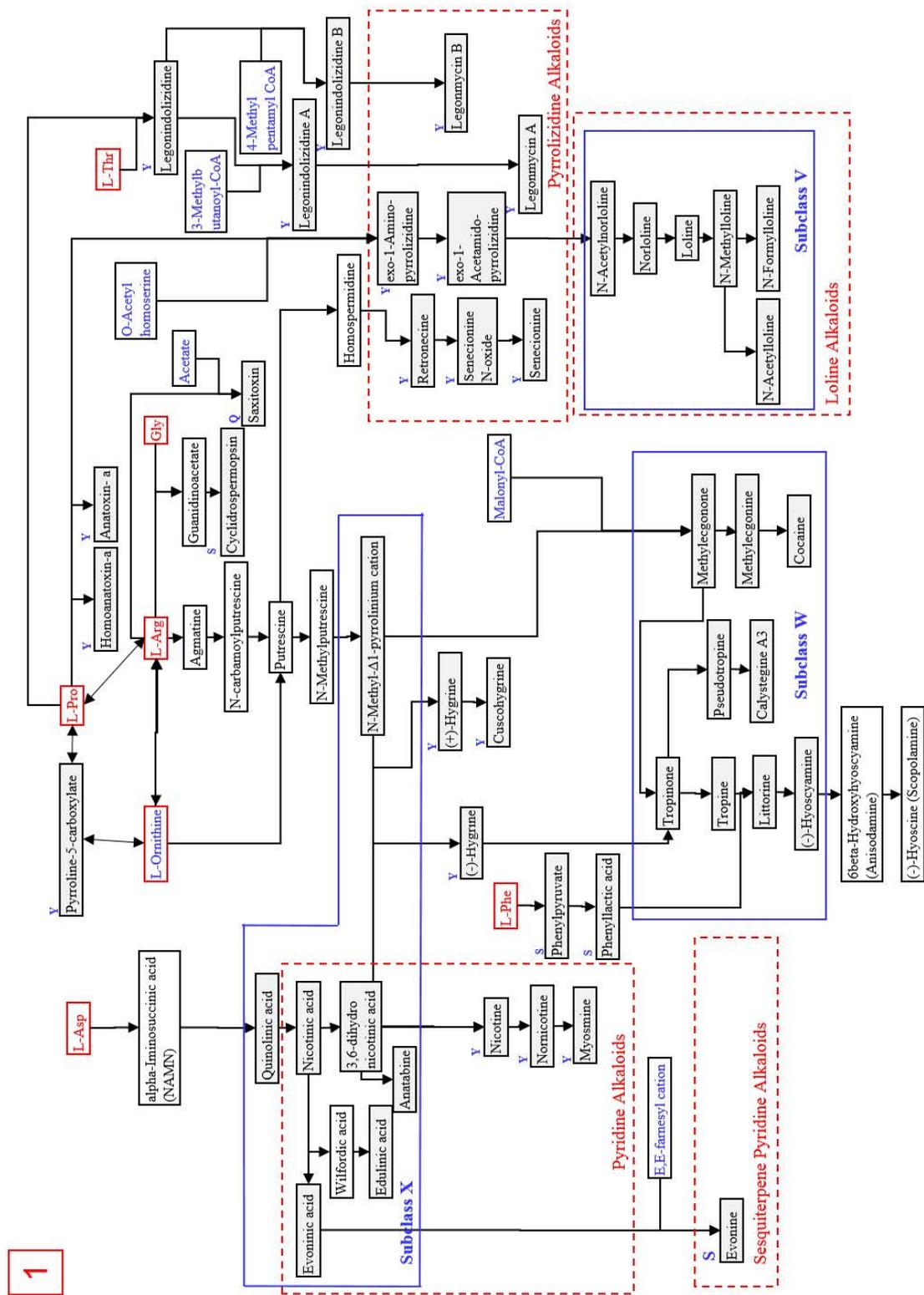


図 3.21 アルカロイド代謝経路へのクラスタのマップング結果 (代謝経路 1)

アルカロイド代謝経路 2~10 は L-tryptophan 誘導体アルカロイドに対応するが、Anthranilate および Formyl anthranilate を含んでいない。

L-Tryptophan および Isopentenyl diphosphate 誘導体はアルカロイド代謝経路 2(図 3.22) にマッピングされている。前節におけるクラスタリングの結果から、骨格構築のプロセスを網羅的に理解することが可能で、Isopentenyl diphosphate はサブクラスタ D(Indole 骨格) とサブクラスタ C(Chanoclavin 誘導体) の 2 段階でクラスタ 8(Ergoline 骨格) とつながることが確認出来る。

Geranyl diphosphate と L-Tryptophan によって生合成される Monoterpenoid indole のアルカロイド代謝経路を図 3.23 に示す。8 つのクラスタ (クラスタ 3、5、6、7、15、16、22、28) に反映されている環骨格は、Monoterpenoid indole アルカロイド代謝経路と関連している。この経路では初めに Secologanin および L-Tryptophan がクラスタ 22(Beta-carbolines) に属する Stictosidine に合成されたのち、クラスタ 28(Iboga アルカロイド)、サブクラスタ G(Cinchaona アルカロイド)、クラスタ 5~7(Vindoline 誘導体)、クラスタ 15、3(Poluneuridines および Stricosides) のように異なるクラスタに割り当てられた環骨格に分岐する。

L-Tryptophan 誘導体 および L-Tryptophan と L-Proline、Acetyl-CoA、Manonyl-CoA、Oxaloacetate、Dimethylallyl pyrophosphate(DMAPP)、Acetoacetyl-CoA、L-Histidine との組み合わせは、アルカロイド代謝経路 4~10 に示されるように多様なアルカロイド化合物を生合成する (図 3.24)。アルカロイド代謝経路 4 に存在する化合物は、Indole 骨格を主な特徴とするクラスタ 14 の環骨格を有している。アルカロイド代謝経路 5 においてはクラスタ 22(Beta-Carbolines)、クラスタ 23(Chatoglobosines)、およびクラスタ 29(Communesins) に対応し、アルカロイド代謝経路 6 はクラスタ 22(Beta-Carbolines) に対応している。また、アルカロイド代謝経路 5 において、L-Tryptophan はサブクラスタ D(Indole 誘導体) に割り当てられた骨格を構築したのちにクラスタ 22、23 および 29 の属する化合物を生合成する。

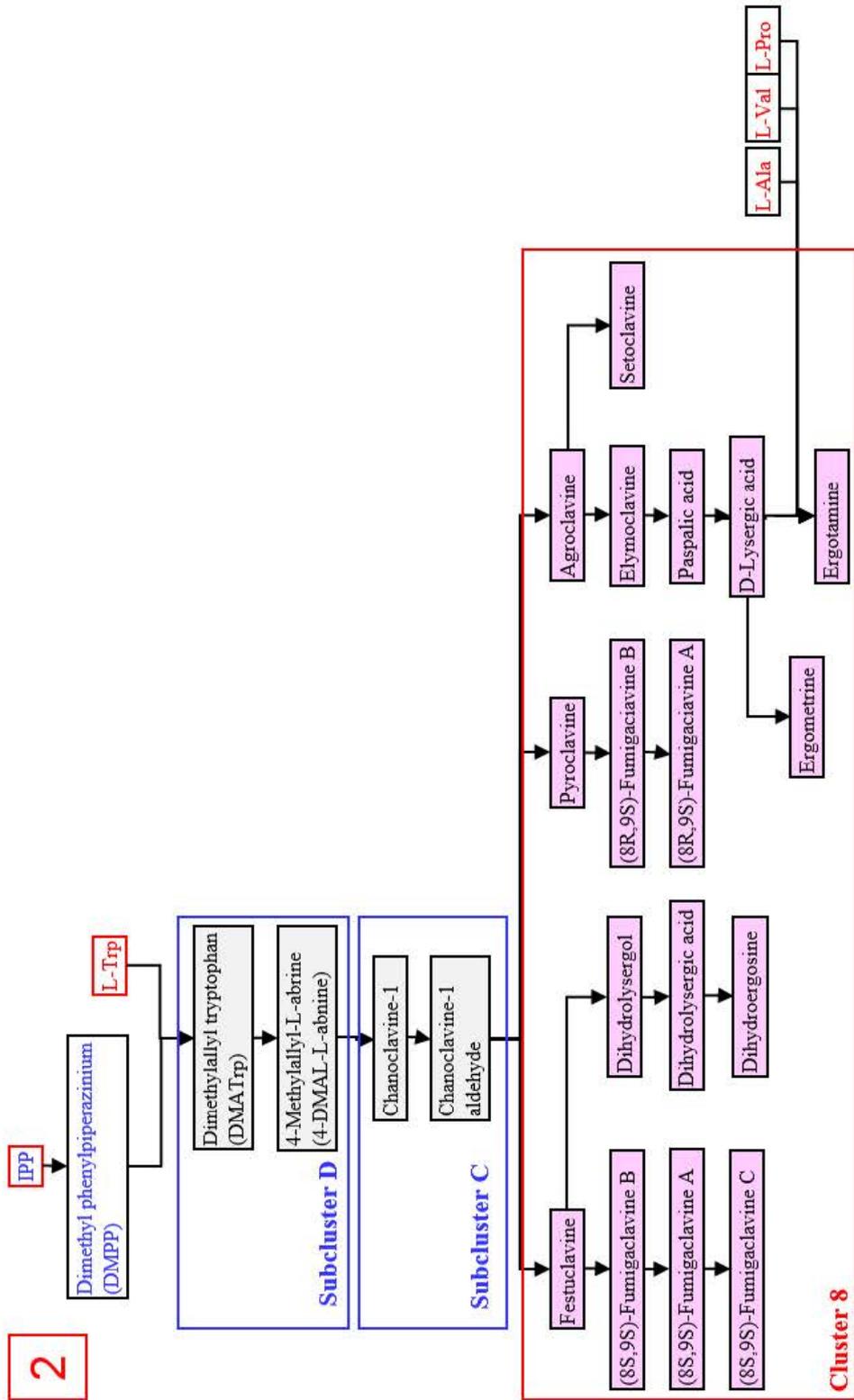


図 3.22 アルカロイド代謝経路へのクラスタのマップング結果 (代謝経路 2)

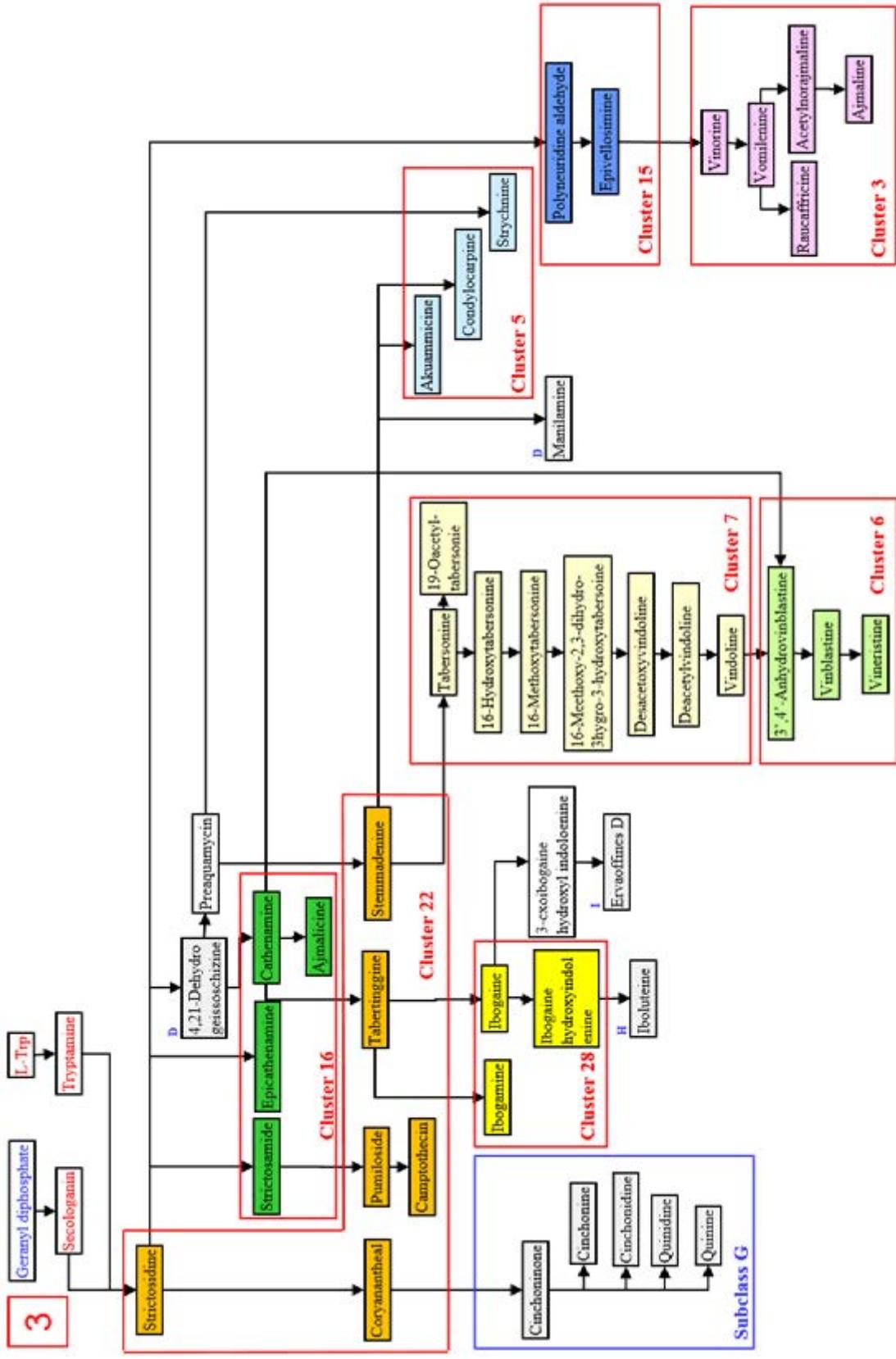


図 3.23 アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 3)

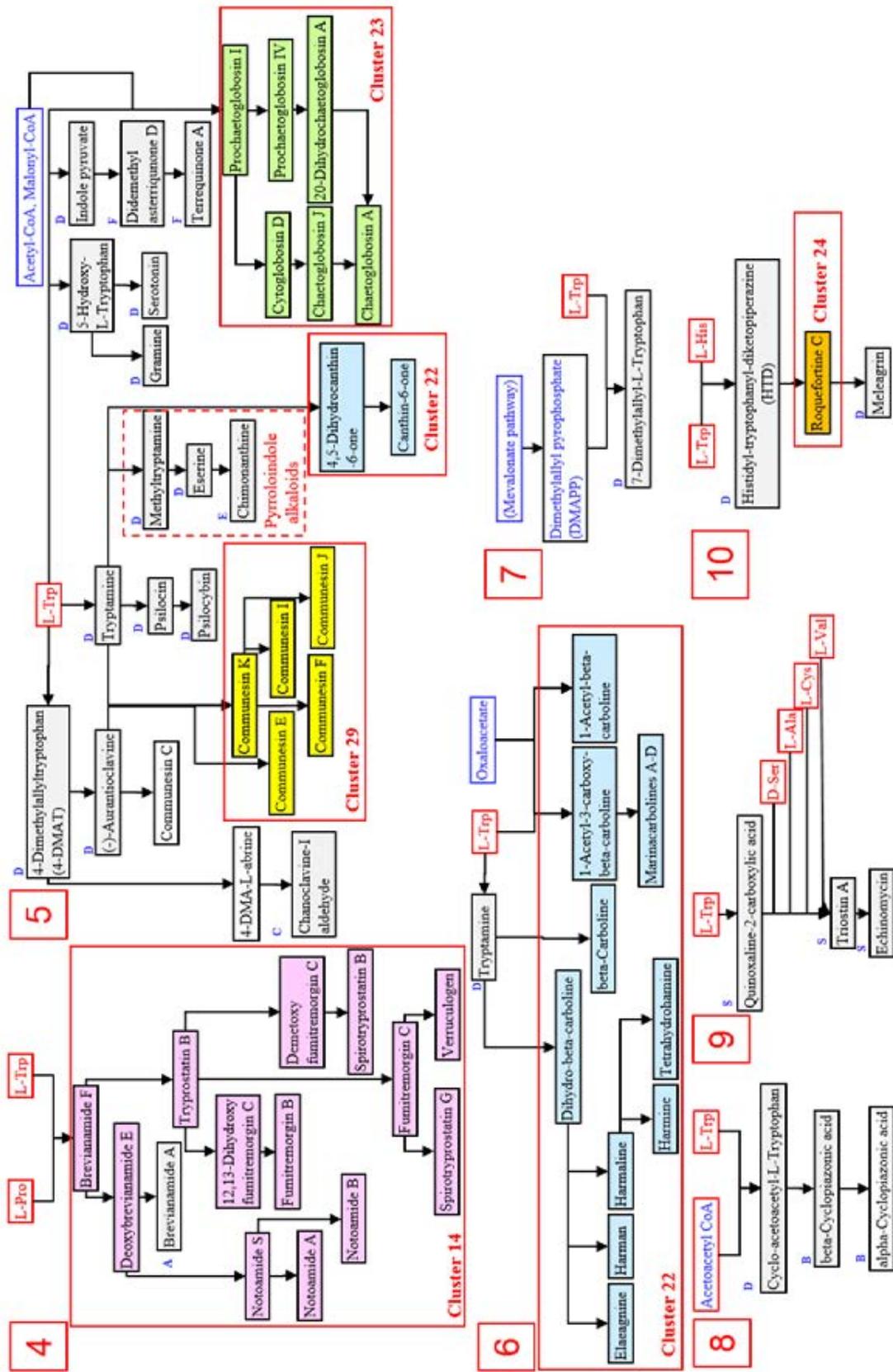


図 3.24 アルカロイド代謝経路へのクラスターのマッピング結果 (代謝経路 4~10)

アルカロイド代謝経路 11(図 3.25) は、Betalanin および Colchine の生合成を行う経路である。この経路において Betalanin 関連化合物を構成する主要な環骨格はサブクラスタ S に、Colchine 関連化合物を構成するの主要な環骨格はサブクラスタ U にそれぞれ属している。Betalanin および Colchine はそれぞれ L-tryptophan と L-Phenylalanine に由来する化合物である。また、Dopamine や Adrenaline は L-Tryptophan と L-Phenylalanine の両方に誘導される化合物であるが、このうちのほとんどの化合物が Betalanin 関連化合物と同様にサブクラスタ S に属している。サブクラスタ S に着目してみると、主に三員環の骨格を五員環および六員環のユニットと統合する。従って、アルカロイド代謝経路 11 は、比較的小さな複素環骨格を単位として有する生合成経路であると解釈することが可能である。

アルカロイド代謝経路 12(図 3.26) は、L-Tyrosine から分岐するアルカロイド生合成経路を示しており、この経路上に存在する化合物はクラスタ 11(Bis-isoquinolines)、クラスタ 12(Morphinans)、クラスタ 17 から 19(Berberines) およびクラスタ 21(Isoquinolines) と関連している。これらのクラスタ間においては、最初に L-Tyrosine から Quinolones と関連する環骨格を構築した後に、導出されたそれらの骨格が Bis-isoquinolines、Morphinans、Berberines それぞれと関連する 3 つの環骨格グループに分岐するという関係性があると説明することができる。

アルカロイド代謝経路 13(図 3.27) における Secologain および L-Tyrosine を出発物質とする化合物からも、アルカロイド代謝経路 12 の場合と同様に Isoquinolines に関連する環骨格を得られる。アルカロイド代謝経路 13 では、Iecac(クラスタ 20) がサブクラスター L に属する Isoquinolines から導出され、さらにクラスタ 21 に属する Isoquinolines を介することで最終的に Quinazolines(クラスタ 13) に誘導される。

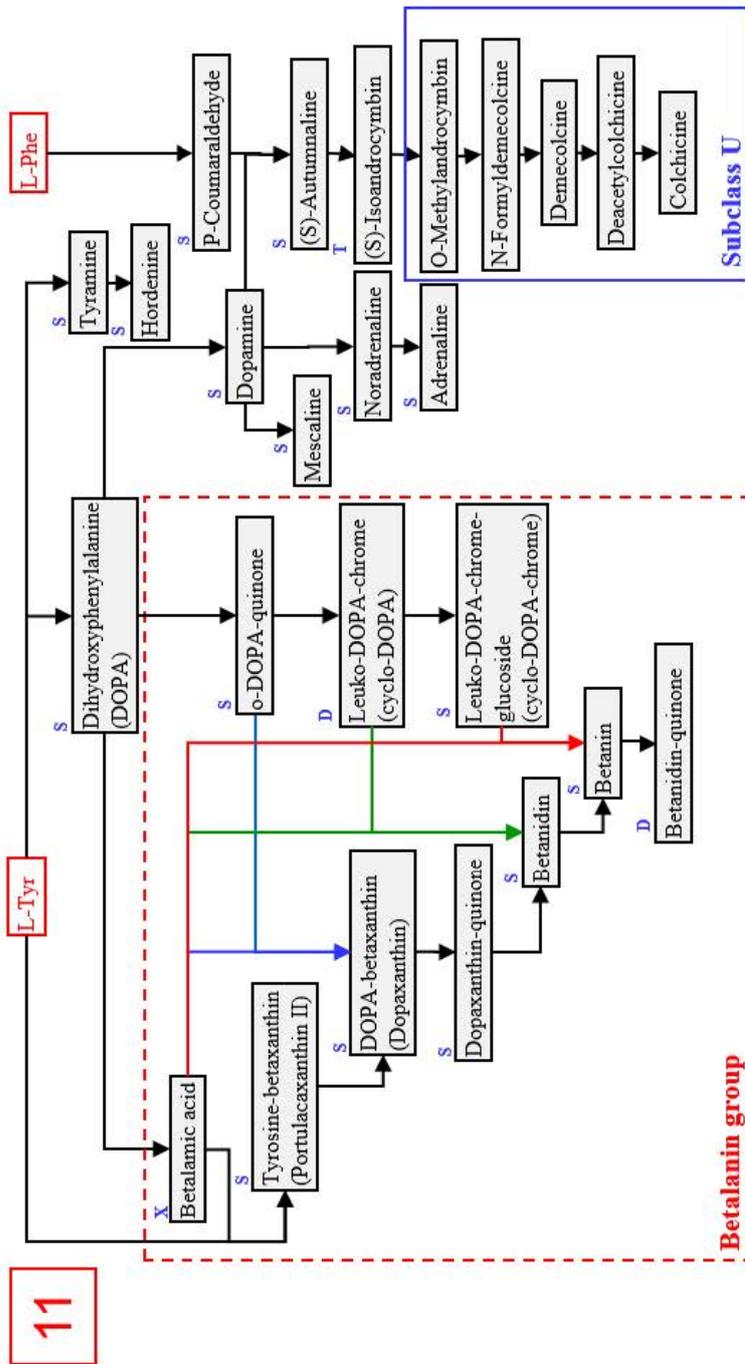


図 3.25 アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 11)

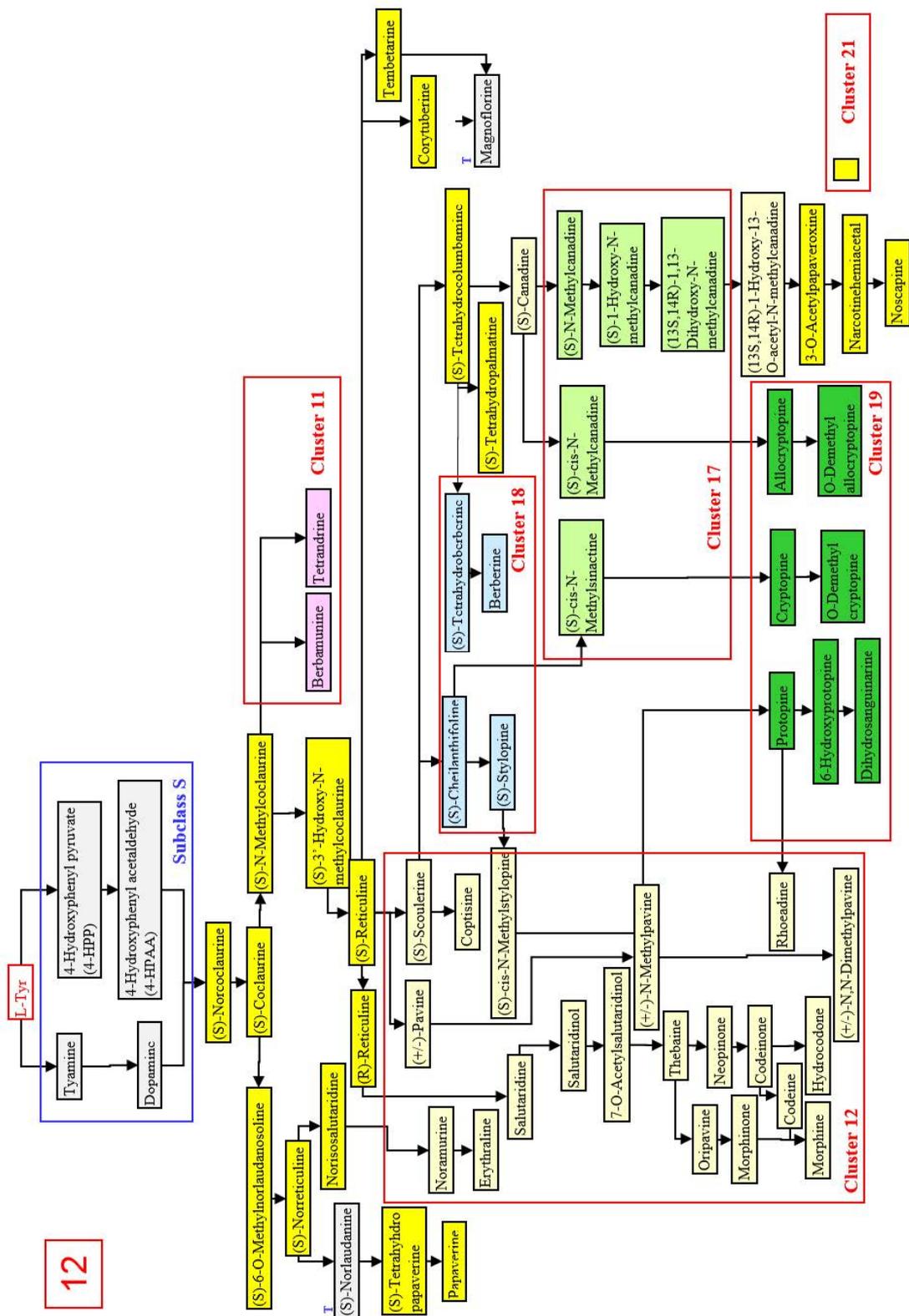


図 3.26 アルカロイド代謝経路へのクラスターのマッピング結果 (代謝経路 12)

L-Lysine はアルカロイド代謝経路 15(図 3.28) 上に存在する化合物の大部分がクラスタ 27 のサブクラスターおよびクラスタ 26(Quinolizidined) に属することからも分かるように、比較的単純な構造の環骨格を有する化合物を導出する。クラスタ 27 に属する化合物の数は、本研究において検証に用いているアルカロイド化合物の約半数にあたり、それらの環骨格は他の 29 クラスタと比べると小さく単純な構造である。また、アルカロイド代謝経路 15 において、L-Lysine 誘導体はサブクラスター Y(Isodolizidines)、サブクラスター L および M(Quinolizidines)、サブクラスター J、K、P、S および X(Lycopodium アルカロイドおよび Piperidines) によって特徴づけられている。

L-Alanine、Malonyl-CoA および Acetyl-CoA はサブクラスター X の環骨格を特徴とする一連の Piperidine アルカロイドを誘導し、また L-Histidine および L-Leucine も同様にサブクラスター X の構造によって特徴づけられる Imidazole アルカロイドを誘導する (図 3.29)。

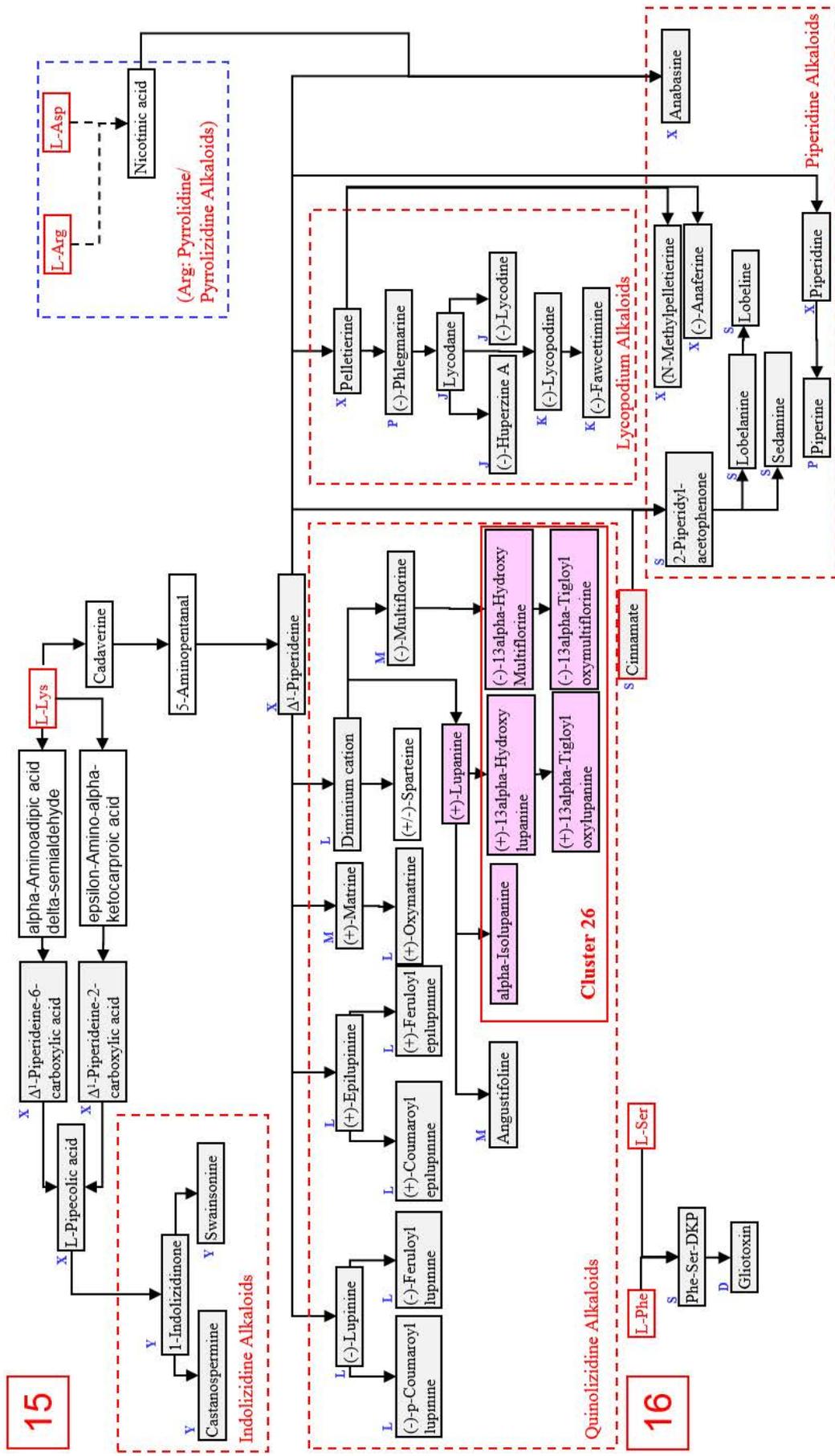


図 3.28 アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 15~16)

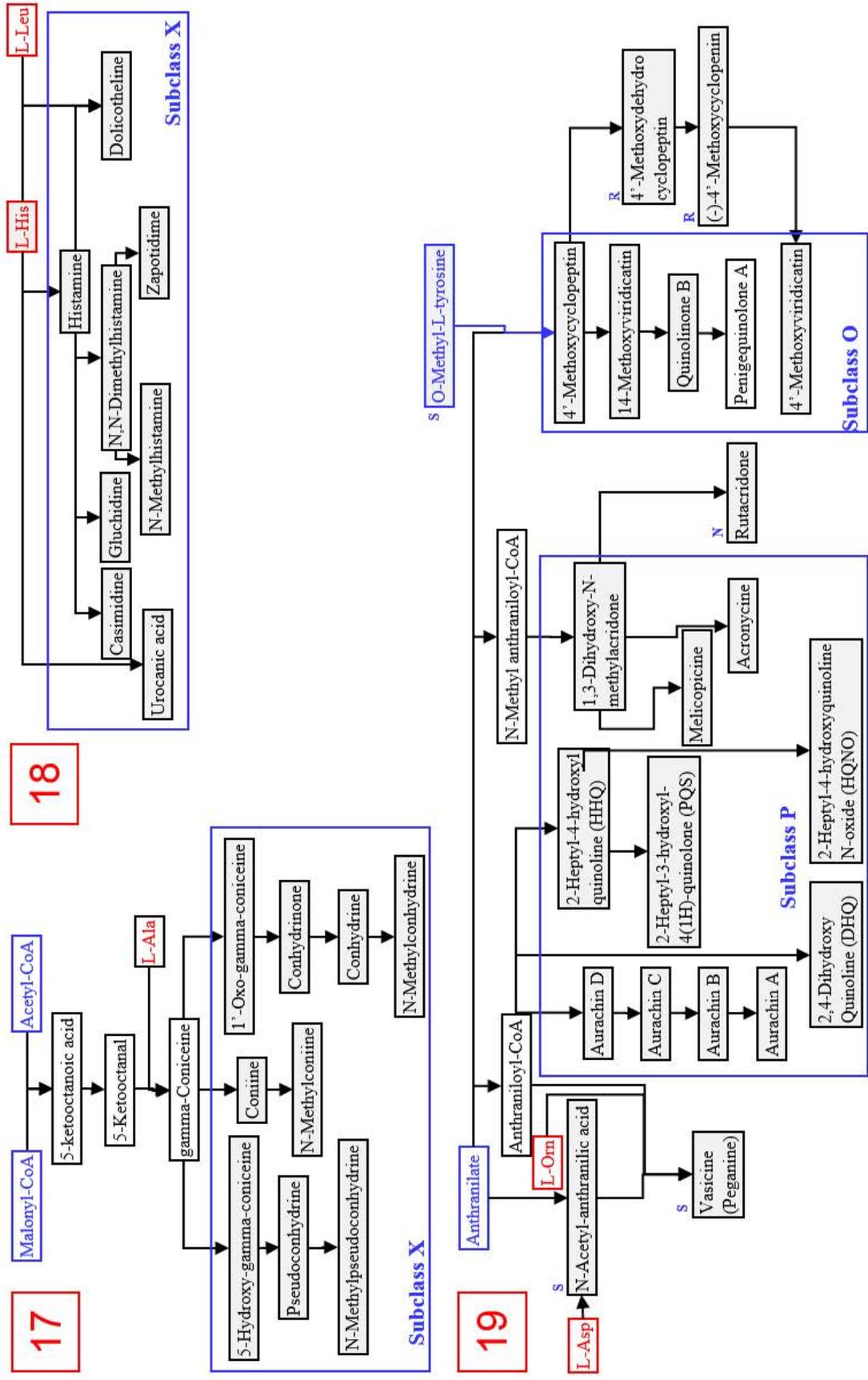


図 3.29 アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 17~19)

Anthranilate は多様なアルカロイドを生成するために様々な化合物と結合する性質を有している。アルカロイド代謝経路 19 から 30(図 3.29 および図 3.30) は Anthranilate との組み合わせによってそれぞれ次のような生合成経路を示す: O-Methyl-L-tryrosine(アルカロイド代謝経路 19)、L-Tryptophane(アルカロイド代謝経路 20)、L-Phenylalanine(アルカロイド代謝経路 21)、L-Trpthane および D / L-Alanine(アルカロイド代謝経路 22)、D-Trptophane と L-Alanine および L-Valine(アルカロイド代謝経路 23)、D/L-Alanine(アルカロイド代謝経路 24)、L-Tryptophan および Glycine(アルカロイド代謝経路 25)、Glycine(アルカロイド代謝経路 26)、L-Proline(アルカロイド代謝経路 27)、L-Glutamine(アルカロイド代謝経路 29)。また、アルカロイド代謝経路 28 は Anthranilate を由来として生合成される N-Methylanthranilate と Malonyl-CoA から誘導される生合成経路である。

以上のように Anthranilate は非常に多様な化合物と組み合わせる性質を有している。しかしながら、Anthranilate によって誘導される化合物はその由来こそ異なるものの、クラスタ 13 あるいは 24 に属する環骨格を有するという共通の特徴があることが確認できた。クラスタ 13 および 24 は Quinazolines 環骨格と、Roquefortines および Acetylazonaleins を含む非常に複雑な環骨格がそれぞれに対応している。

アルカロイド代謝経路 30 は Indole-diterpene アルカロイド整合性経路を示している。部分環構造による化合物分類の結果において、Indole-3-glycerophosphates および Geranylgeranyl diphosphate(GGPP) は初めにクラスタ 2 に割り当てられた環骨格を最初に誘導した後、より複雑な構造であるクラスタ 1 に属する環骨格を生合成する。

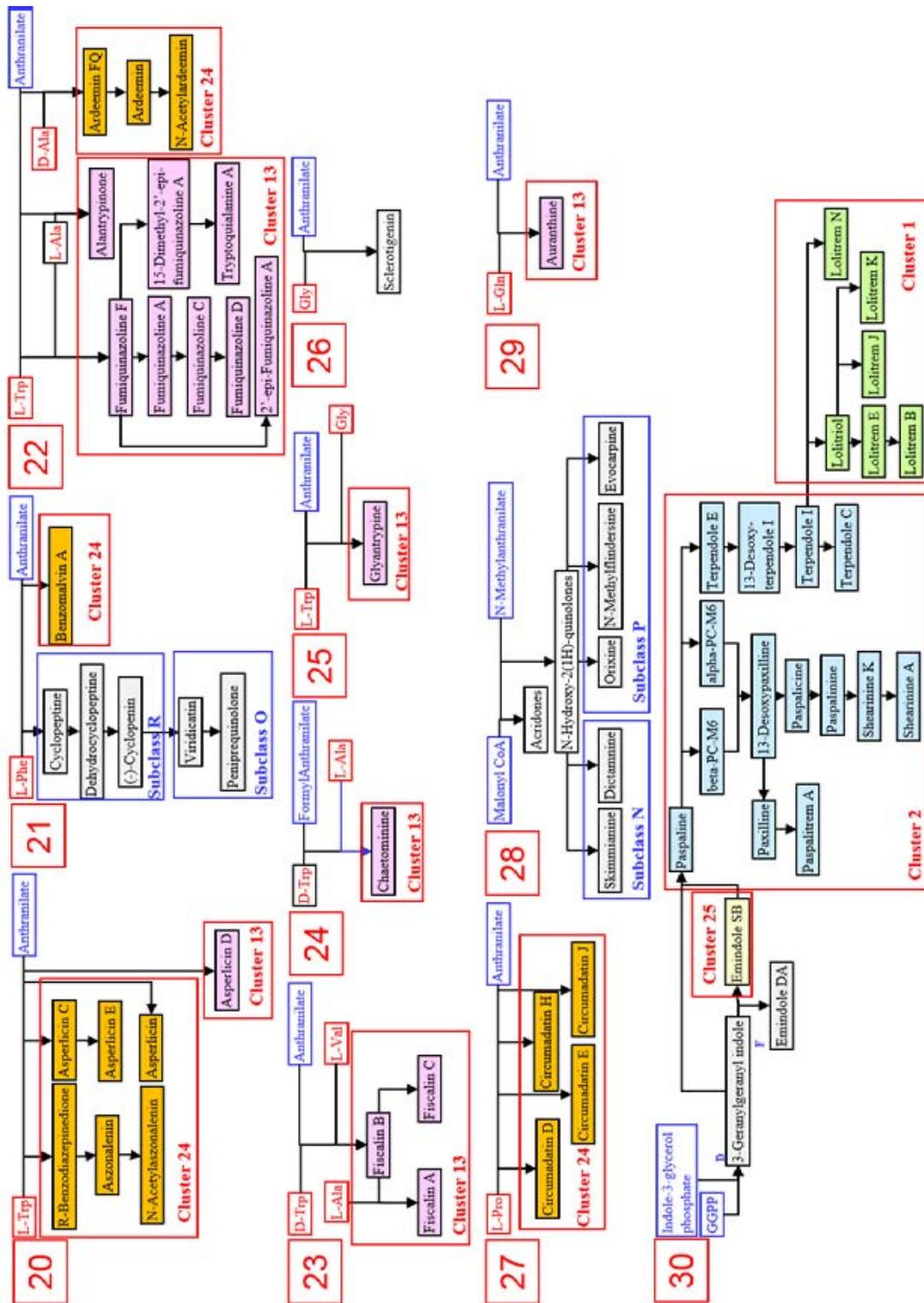


図 3.30 アルカロイド代謝経路へのクラスタのマッピング結果 (代謝経路 20-30)

アルカロイドにはその性質及び起源から、真正アルカロイド (True alkaloid)、不完全アルカロイド (Protoalkaloid)、疑アルカロイド (Pseudo-alkaloid) の 3 つの群に大別される。真正アルカロイドおよび不完全アルカロイドはアミノ酸に由来しているのに対して、疑アルカロイドはアミノ酸に由来していないアルカロイドである。アルカロイド代謝経路 31 および 32(図 3.31) は出発物質がそれぞれ Cholesterol、Adenine であるため、これらの経路は疑アルカロイドに対応していることが分かる。

アルカロイド代謝経路 31 は、Emiindoles(クラスタ 25)、Aglycon(クラスタ 10)、Glucoside of Tomatidines(クラスタ 9)、Tomatins および Chaconine(クラスタ 4) との関連性がみられ、これら 4 つの環骨格による生合成プロセスと定義することが出来る。また、アルカロイド代謝経路 32 の経路上に存在する化合物は、互いに類似性が非常に高い Purine alkaloids であり、それらの環骨格は全てサブクラスター Q に属している。

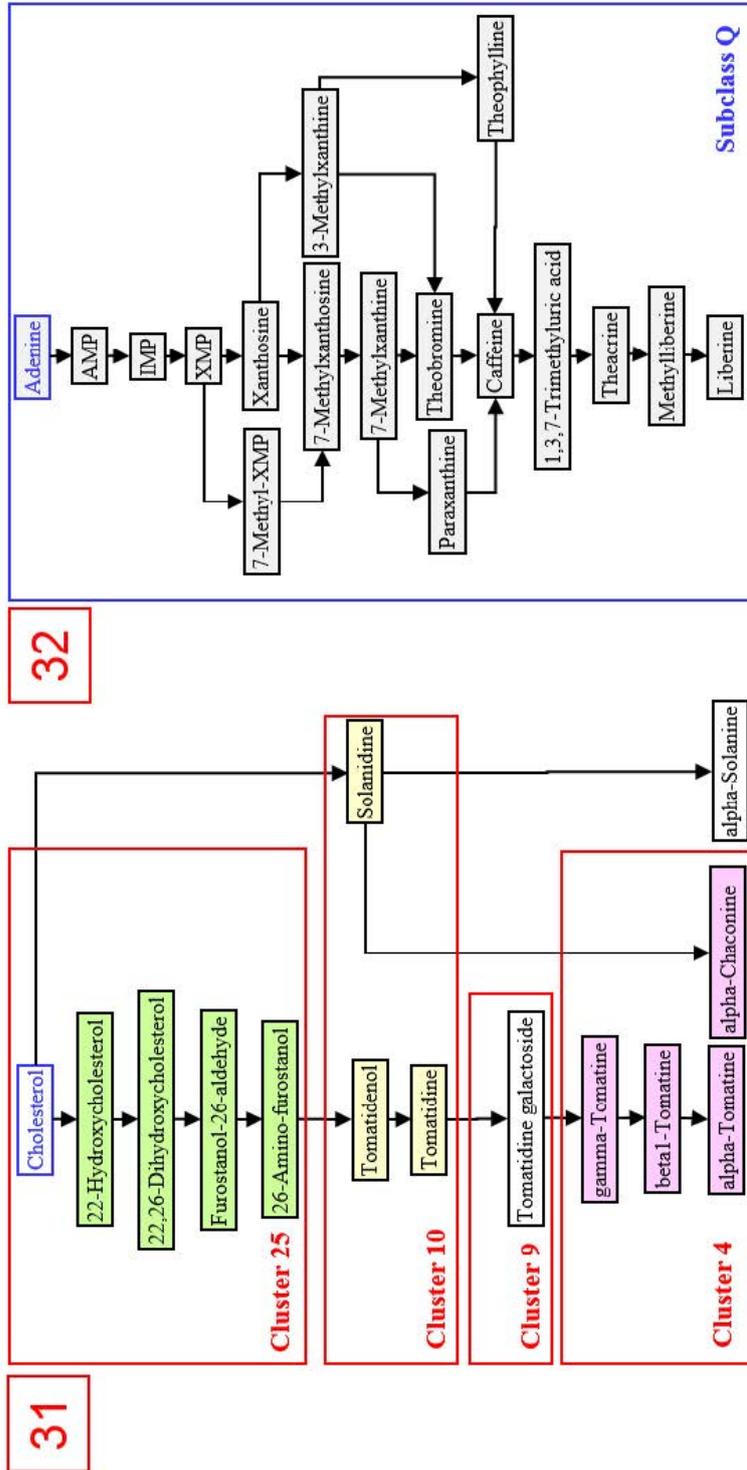


図 3.31 アルカロイド代謝経路へのクラスタのマップング結果 (代謝経路 31~32)

上述した結果からもわかるように、アルカロイド化合物を包括的に体系化し、アルカロイド化合物を部分環構造あるいは環骨格に基づいて分類した結果は、アルカロイド中の環骨格の構築原理を代謝経路を考慮した上で説明することが可能であると結論付けることができる。

3.3.3 KNApSAcK Database 中のアルカロイド化合物における分類

概要で記述した通り、二次代謝物の構造やその特性が既知であるにも関わらず、それが創出される代謝経路が決定されていない化合物は非常に多い。本研究では、KEGG における経路データベースと同様に科学文献の経路情報を収集することで 478 個の経路情報と結び付けられたアルカロイド化合物を得た。しかしながら、アルカロイド化合物の化学構造は広範囲に決定されており、KNApSAcK Database においてはアルカロイドに関連する窒素原子を持つ 12243 種類の化合物が現時点で蓄積されている。

本研究では、収集した代謝経路情報が既知のアルカロイド化合物に対して 187 の環骨格を定義したが、12243 の化合物を経路既知の環骨格に割り当てようとしたところ、全体の化合物のうち、3124 化合物 (25.5%) が定義した環骨格により構成されており、残りの 9119 化合物 (74.5%) はそうではないことが確認された。これはアルカロイドのみならず天然由来の化合物が創出される一連の代謝反応が、複雑な構造多様性を生み出していることを再確認する結果であるとともに、残りのアルカロイド化合物の環構造に基づいた骨格類似性を調べることにより、代謝経路情報のさらなる手がかりを得られることが期待される。

3.3.4 化合物の部分環構造の抽出に関する問題点と解決策

部分環構造に基づいた化合物の分類はバイオインフォマティクス、ケモインフォマティクス分野において非常に興味深いテーマである。しかしながら一方で、化合物から部分環構造を列挙するプロセスに膨大な計算コストが掛かるという問題点もある。化合物の不飽和度が高いほどに抽出される部分環構造の数は増加し、計算コストが膨大なものになってしまう。ここでは、この問題を解決するための一つの案

を提示する。

二次代謝物から環構造を抽出する問題はその二次代謝物中の全ての結合のうち、どの結合を選択するかという組み合わせ問題とみることが出来る。ここでは、二次代謝物から環構造を列挙する場合を例に挙げて説明をする。図 3.32(a) の化合物は 19 の結合から構成されており、それぞれの結合を使用するか否かの二通りを選択することを考える。このとき、結合の組み合わせの総数は $2^{19}=524,288$ 通りになり、これを場合分け二分木で表現したものが図 3.32(b) である。図 3.32(b) の場合分け二分木において各節点には結合を表すラベルが割り当てられており、その節点を使用する場合 (1-枝) とその節点を使用しない場合 (0-枝) で枝分かれしている。また、場合分け二分木の終端 (葉) は 0 または 1 の値をもっており、終端が 1 の場合はその結合の組み合わせが環構造となっていることを示している。図 3.32(a) の場合、 2^{19} 通りという膨大な結合の組み合わせに対して、列挙される環構造は A~I のわずか 9 通りしか存在しない。従って、全ての結合の組み合わせを生成した後に部分構造の条件を満たす組み合わせを探索するのは非常に効率が悪いことが分かる。

この問題を解決する手段として、Zero-suppressed binary decision diagram(ZDD) アルゴリズムの抽出に応用が考えられる。ZDD は北海道大学の湊真一らが考案した組み合わせ集合を効率良く表現することが出来るデータ構造であり、場合分け二分木を 2 つの規則に基づいて圧縮することで得られる [89, 90]。

以下に場合分け二分木を ZDD に圧縮する際の 2 つの規則を示す。

(a) 冗長節点の削除

1-枝が 0 の値を持つ終端を指している場合、この節点を削除して 0-枝の行き先に連結させる。ここでは、図 3.32(b) の場合分け二分木から図 3.32(c) の ZDD への圧縮過程を例に示す。節点 1 を使用し、節点 2 を使用する場合、3 を使用するか否かに関わらず行先の終端の値は 0 となっている。この場合、節点 3 の 1-枝は 0 の値を持つ終端を示しているため、節点 3 を削除して節点 2 の 1-枝と終端を連結させる。次に、節点 2 についても同様の手順で削除をし、節点 1 の 0-枝と節点 3 を連結させる (図 3.33)。

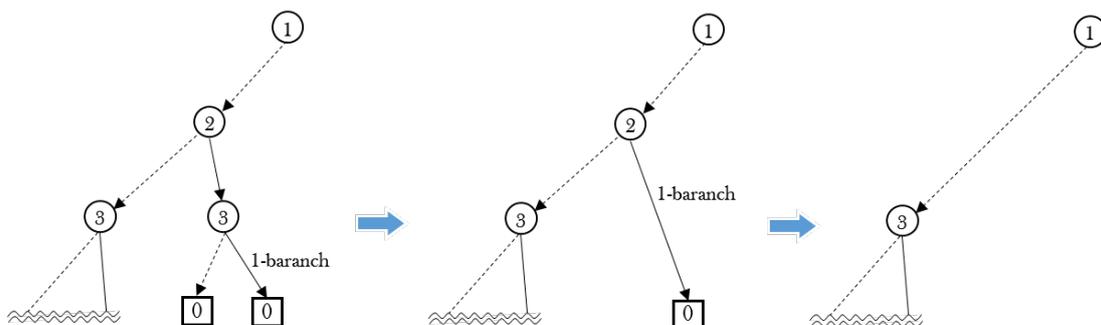


図 3.33 二分木から ZDD に圧縮する際の規則 (冗長節点の削除)

(b) 等価節点の共有

節点のラベルが同じで 0-枝、1-枝の行き先が同じである等価節点を共有する。ここでは、図 3.32(b) の場合分け二分木から図 3.32(c) の ZDD への圧縮過程を例に示す。節点二つある節点 12 はいずれも 0-枝と 1-枝の行き先が同じであり、共有することが可能である (図 3.34)。

ZDD の圧縮効率は適用する構造にもよるが、等価な節点が存在せず共有が全く行われないという最悪な場合でも、全ての組み合わせの総数を超えないという性質がある。また、上述したような既に構築されている場合分け二分木から規則を適用することで ZDD を生成する方法では計算時間やメモリを膨大に消費してしまう

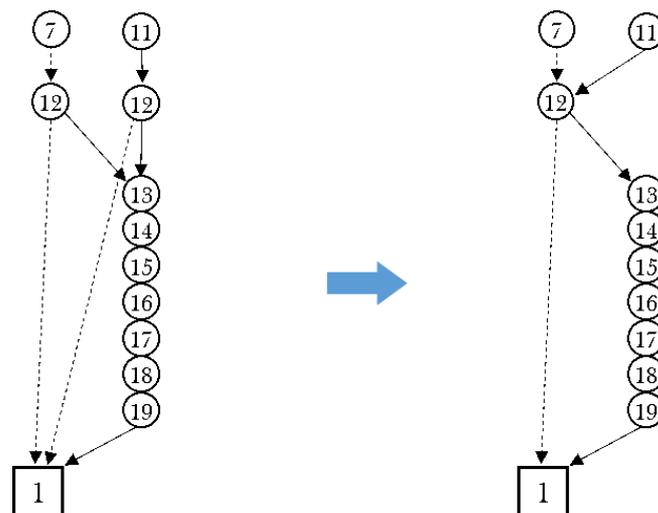


図 3.34 二分木から ZDD に圧縮する際の規則 (等価節点の共有)

が、ZDD は場合分け二分木から構築する以外に上に位置する節点からトップダウン的に 0-枝と 1-枝の先を作成することでも構築が可能である [91]。この場合、ある節点の行く先に環構造がある見込みがないと判定された時点でそれ以降の節点を作成しないことで効率的な構築を行う。この ZDD を応用したアルゴリズムを開発することで生物、化学の両分野において部分環構造に基づく分類の応用範囲がより広がることが期待される。

3.4 まとめ

本章では、第 2 章において構築したアルカロイド代謝経路の情報を用いることで、化合物の環構造に焦点を当てた分類を行うことで代謝経路からモジュールの抽出を試みた。本研究ではアルカロイド化合物に関する情報を体系化し、個々の化合物を代謝経路に関連付けるために化合物が有する環構造に焦点を当ててそれに基づく分類を行った。提案した分類手法をアルカロイド化合物の分類を行った結果、29 のクラスと 25 のサブクラス、また 187 種の環骨格を抽出した。そしてアルカロイド化合物の化学構造分類との妥当性を検証した。

二次代謝産物の生物活性が化学構造と高度に関連しているという仮説は、概要で

示したように 3次元グラフマッチアルゴリズム COMPLIG に基づく代謝産物間の 3次元構造類似性評価を用いることによって示されている [8]。例えば、イソキノリンアルカロイドは鎮痛作用と関連している。また、鎮静作用を含むキノロンアルカロイド、抗肥満作用を含むテルペニアルカロイド、抗腫瘍活性作用を有するインドールアルカロイド、舌下作用を含むステロイドアルカロイド、抗糖尿病薬として薬剤利用がなされるピロリジンおよびキノリジジンアルカロイドは、抗糖尿病薬として利用されるトロパンアルカロイドなどが化学構造と生物活性の関係性として報告された。これらの報告された関係は、本研究における環構造に焦点を当てた分類と一致している。さらに、環構造に焦点を当てた分類は、化合物を代謝経路を考慮した上で説明することが可能である。

また本研究で試みた生物学的活性および代謝産物の生合成経路の体系化は、アルカロイド化合物の進化プロセスの解釈に新しい知見を提供すると期待される。具体的には、第 1 章において関連研究で記述したように三次元化学構造の類似性を考慮することにより、代謝モジュール内を生理活性という情報でさらに細分化することが可能となり、二次代謝物-代謝経路または二次代謝物-生物活性間に関する網羅的な理解を推し進めることが期待される。

第 4 章 グラフ畳み込みニューラルネットワークを用いた二次代謝物の化学構造による出発物質予測

4.1 概要

二次代謝物が生物の生体内でどのような代謝反応によって生合成されるかという情報は、その物質を天然資源から安定的に供給したいという場合において非常に重要な情報であり、医薬・農業などの分野からもその情報を得ることが求められている。しかしながら、化学構造情報および性質が同定されている天然化合物であっても、それらがどの代謝経路から誘導されるかが不明であるものは非常に多い。

本論文の第 4 章において記述した、環骨格、部分環構造に基づく分類を活用することで、二次代謝物の構造情報と代謝経路の相互関係がすでに明らかにされている範囲においては代謝経路の推定が可能である。しかしながら、これは今までに確認されていない新規の化学構造に対して適用しても良い結果は得られないことは、第 4 章における KNApSAcK Database 上のアルカロイド化合物に対する検証結果からも確認できる。

そこで本章では、近年注目されているグラフ構造を対象とした機械学習手法であるグラフ畳み込みネットワークを用いた学習モデルを二次代謝物に適用することで、二次代謝物の構造情報からそれらがどの出発物質から誘導されるのか、即ち、いずれの代謝経路に由来する二次代謝物であるのかの予測を試みる。

4.2 手法

ここでは二次代謝物の構造情報から出発物質を予測するために用いる手法であるグラフ畳み込みネットワークについて説明をする。また、グラフ畳み込みネットワークに関する説明を記述する前に、これと深く関連するディープラーニング、および同様のコンセプトを有する畳み込みネットワークに関する説明も記述する。

4.2.1 ディープラーニング

生物は外界の物事を知覚し、記憶し、それらをもとに様々な思考・論理を行う。これらは脳神経系によるものである。生体の脳神経系を抽象化して考えると、多数の神経細胞が興奮性及び抑制性のシナプス結合で構成されたネットワークとしてモデル化することが出来る。ニューラルネットワークとは、生体の脳神経系の特徴を学習システムに応用したものである。生物は外界の物事を知覚し、記憶し、それらをもとに様々な思考・論理を行う。これらは脳神経系によるものである。生体の脳神経系を抽象化して考えると、多数の神経細胞が興奮性及び抑制性のシナプス結合で構成されたネットワークとしてモデル化することが出来る。ニューラルネットワークとは、生体の脳神経系の特徴を学習システムに応用したものである。

ニューラルネットワークにおいて、脳神経系の神経細胞（ニューロン）は次のようにモデル化され、ユニットとして扱われる。(図 4.1)

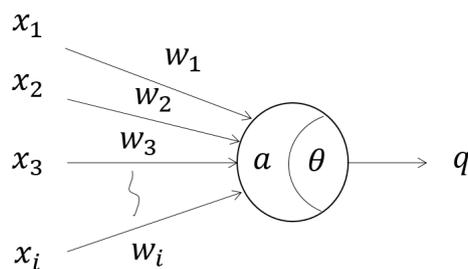


図 4.1 ニューロンモデル

各ユニットの入出力の関係は、出力関数 g と状態遷移関数 f によって定められる。このとき、 q はユニットの出力、 a はユニットの状態、 x は他のユニットあるいは外界からの入力、 w はユニット同士の結合強度（結合荷重）、 θ はユニットの発火閾値である。ユニットの出力を定める出力関数として、以下のようなステップ関数やシグモイド関数がかつては多用されていた。

ステップ関数：

$$g(x) = \begin{cases} 1 & (x \geq 0) \\ 0 & (x < 0) \end{cases} \quad (4.2.1)$$

シグモイド関数:

$$g(x) = \frac{1}{1 + \exp(-\frac{1}{T}x)} \quad (4.2.2)$$

ニューラルネットワークは、複数のユニットを結合することで構成される。ニューラルネットワークはユニットの配置や信号伝搬の方向が異なる様々な形態がある。その中でも最も単純かつ基本的な形態である順伝搬型ニューラルネットワークでは、複数のユニットが階層状に結合し、ネットワークに入力された信号が決められた単一方向に伝搬することによって、最終的な出力が決定する。入力層、出力層および1つ以上の中間層によって形成され、信号は入力層から中間層を介して出力層に向けて伝搬し、逆方向に伝搬することはない。以下に順伝搬型ニューラルネットワークでの構造の一例を示す。(図4.2)

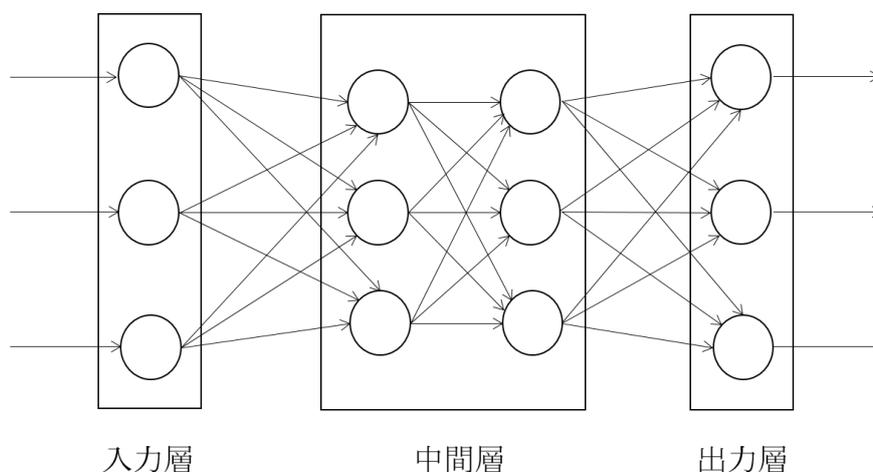


図 4.2 順伝播型ニューラルネットワークの構造例 (4 層)

層の数を増やしすぎると勾配消失問題により学習が困難になること、また、パラメータが増えることにより過学習が起これり汎化性能が低下することなどから、以前は3層構造のニューラルネットワークが一般的に用いられていた。しかし近年、それらの問題を低減する技術が開発されたことにより4層以上、多い場合は十

数層に及ぶニューラルネットワークを用いたディープラーニング (深層学習; Deep learning) が様々な分野で注目されている [92-94]。実際、ディープラーニングはこれまでに画像処理、音声信号処理、自然言語処理などの様々な分野で多くの成果を残している。

4.2.2 畳み込みニューラルネットワーク

近年、ディープラーニングが一般に周知されるようになった大きな要因の一つとして、順伝搬型ニューラルネットワークの一つである畳み込みニューラルネットワーク (Convolutional Neural Network) のモデルが画像処理分野で大きな成功を収めたことが挙げられる。

2012年、人工知能を用いた画像認識の精度を競う国際コンテストである Image Net Large Scale Visual Recognition Challenge (ILSVRC) において、トロント大学の研究者らがディープラーニングによって物体のカテゴリ認識を行い、圧勝した [95]。この時に用いられたモデルが畳み込みニューラルネットワークである。この結果が現在まで続く、画像処理分野において畳み込みニューラルネットワークのディープラーニングがスタンダードとされている発端ともいえる。

畳み込みニューラルネットワークの基本構造は入力層と出力層に加え、中間層として畳み込み層、プーリング層、全結合層から構成されている (図4.3)。ここでは、画像データに対する畳み込みニューラルネットワークにおける各層の概要を説明する。

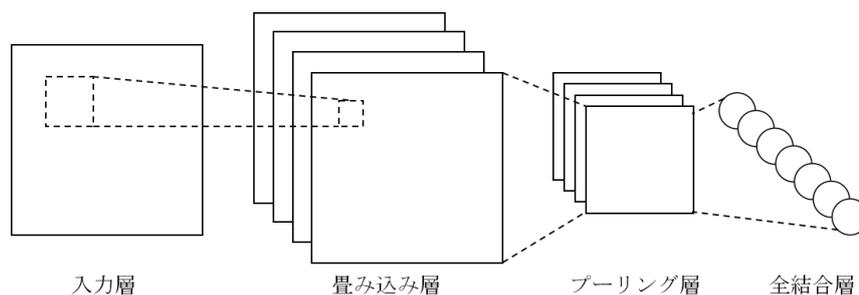


図 4.3 畳み込みニューラルネットワークの概念図

畳み込み層

畳み込み層では入力する画像データの一部を入力としてニューラルネットワークの出力を計算する。入力する小領域を移動させながら、画像全体をカバーするように対応する出力を求める畳み込み演算を行うことで、入力画像の局所的な情報を変換処理し、特徴を抽出するフィルタとして扱うことができる。一般的には一層につき複数のフィルタを持たせ、入力した画像をそれらのフィルタと活性化関数をそれぞれ適用した特徴マップに変換する。その結果として、特徴マップがフィルタと同様の数だけ出力される。

プーリング層

プーリング層は、畳み込み層から出力された特徴マップのサイズを圧縮 (ダウンサンプリング) して、代表値だけを残すような処理を行う。例を挙げると 8×8 の入力データに対して縦横比が $1/2$ になるようにダウンサンプリングする際、 2×2 の区間で代表値を決定し、 4×4 のデータを出力する (図 4.4)。代表値の決定方法は区間内で平均をとる平均プーリング、区間内で最大値となる値をとる最大プーリングなどがよく用いられる。

プーリングは、畳み込み層によって出力された特徴マップ内の特徴の位置感度を低下させる働きがある。即ち、元の画像データで考えたとき、画像内のある物体の位置が変わっていたり、あるいは変形していたりしても、その差異がプーリングによって無くなり、出力が不変なものとなる (移動不変性)。

全結合層

全結合層には、畳み込み層、プーリング層での処理を何度か繰り返し行った後、それらの出力データをの全て値として入力する。この層で全ての結果をまとめ上げることで、最終的に一次元ベクトルとなる。

全結合層での出力を確率関数として用いることで画像分類や物体認識を行えるようになる。

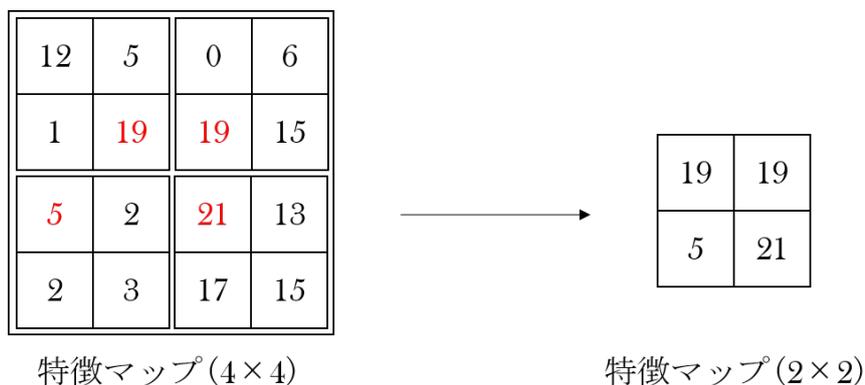


図 4.4 畳み込みニューラルネットワークにおけるプーリングの一例

4.2.3 グラフ畳み込みネットワーク

畳み込みニューラルネットワークにおける一連の処理がもたらす移動不変性あるいは位置不変性は分類問題などの応用を考慮する際に、非常に重要かつ有用な特性である。そこで、研究者らは画像データ以外のデータ、具体的にはグラフ構造に対しても畳み込みと同様の処理を行うことが可能であるかを考え、今日まで様々な手法が研究されてきた。その一つが本研究において使用するグラフ畳み込みネットワーク (Graph Convolutional Network) である。グラフ畳み込みネットワークとはグラフ構造上における畳み込み演算を定義し、畳み込みニューラルネットワークと類似した振る舞いをするモデルである。

グラフ畳み込みネットワークの基本処理は、部分的に見れば前述した畳み込みニューラルネットワークと等価である。これまでの研究でいくつかのグラフ畳み込みネットワークのアルゴリズムが提案されているが、本研究においては畳み込み層、プーリング層、集約層を基本構造としたグラフ畳み込みネットワークを用いている。

4.2.4 入力層

本研究におけるグラフ畳み込みネットワークモデルでは化合物の構造情報を入力としている。この際化合物はグラフとして表現しており、各原子にはそれぞれに対

応じた特徴ベクトルを割り当てる。原子数 S で構成される化合物の場合、ネットワークに入力される特徴ベクトルの数は S である。

初期の状態において、各原子の特徴ベクトルはその種類に応じた One-hot ベクトルとして表現している。One-hot ベクトルとはベクトルの全要素のうち 1 つの要素のみ値が 1 となり、その他の要素の値は 0 となるベクトルである。ここでは One-hot ベクトルの各要素は原子の種類を表現している。例えば、特徴ベクトルの各要素が $["C" "O" "H"]^T$ という元素種に対応しているとする、各原子の特徴ベクトルは $C = [1 0 0]^T$ 、 $H = [0 1 0]^T$ 、 $O = [0 0 1]^T$ というベクトルとなる。この初期状態の特徴ベクトルはネットワーク入力される際に、重み W_0 によって長さ M_0 の入力ベクトルに変換される。この時、 M_0 は手動で決定する必要があるハイパーパラメータである。

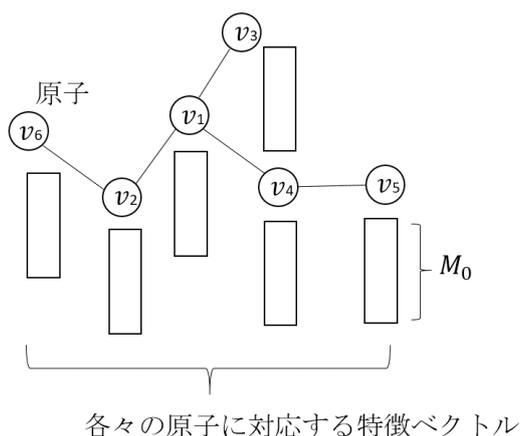


図 4.5 グラフ畳み込みネットワークへの入力として用いる各原子に特徴ベクトルが割り当てられた化合物

畳み込み層

畳み込み層では、まず化合物中の各原子に対して重み \mathbf{W} を用いた特徴量の線形変換を行い、その後に各原子の特徴ベクトルの更新を行う。任意の原子に対する特徴ベクトルの更新は任意の原子とそれに隣接している原子の特徴ベクトルを加算した後に活性化関数を適用し、この出力を新たな特徴ベクトルとして更新す

る。ここでは活性化関数として次式で表される ReLU 関数 (Rectified Linear Unit Function) を用いている。

$$\text{ReLU}(x) = \begin{cases} x & (x \geq 0) \\ 0 & (x < 0) \end{cases} \quad (4.2.3)$$

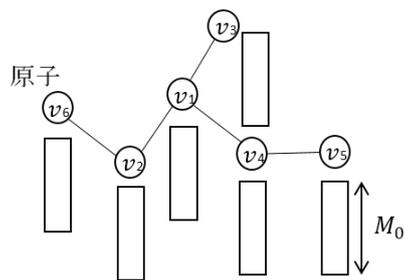
次にグラフ上の任意の原子に対する畳み込み処理を定式化したものを次式で示す。式は i 番目の原子の特徴ベクトルを更新する処理を表している。

$$\text{Conv}(v_i) = \text{ReLU}\left(\sum_{v_j \in N(v_i)} W_c v_j\right) \quad (4.2.4)$$

ここで v_i は i 番目の原子における特徴ベクトル、 $N(v_i)$ は i 番目の原子およびそれと隣接している原子の特徴ベクトルの集合、 W_c は線形変換における重みである。

また、畳み込み層における一連の流れを図 4.6 に示す。図 4.6 は原子数 S で構成されている化合物に対して畳み込み層を適用する際の実例である。ここでは 1 番目の原子に対して畳み込み処理を適用し、特徴ベクトル v_1 を更新する場合を例に挙げて説明する。1 番目の原子に対して畳み込みを行う場合、1 番目の原子およびそれと隣接している原子である 2 番目、3 番目、4 番目の特徴ベクトル v_1, v_2, v_3, v_4 を用いて、1 番目の原子の特徴ベクトルを更新する。畳み込み層に入力される各特徴ベクトルの長さが M_c とすると、特徴ベクトル v_1, v_2, v_3, v_4 はそれぞれ、重み W_c によって長さ M_{c+1} のベクトルに変換される。そして変換後のベクトルを各要素ごとに加算することで 1 つのベクトルにまとめ、ReLU 関数を適用することで得られた v_1^{conv} を新たな特徴ベクトルとして更新する。以上が 1 番目の原子に対して畳み込み処理を行う際の流れであり、畳み込み層においてはこの処理を S 原子全てに対して行う。

グラフ畳み込みネットワークにおける畳み込み層では、隣接している原子の特徴ベクトルを用いることで新たな特徴ベクトルの更新を行っている。この処理によって、隣接している原子間の相互作用を考慮した特徴ベクトルを生成することが可能である。



Conv

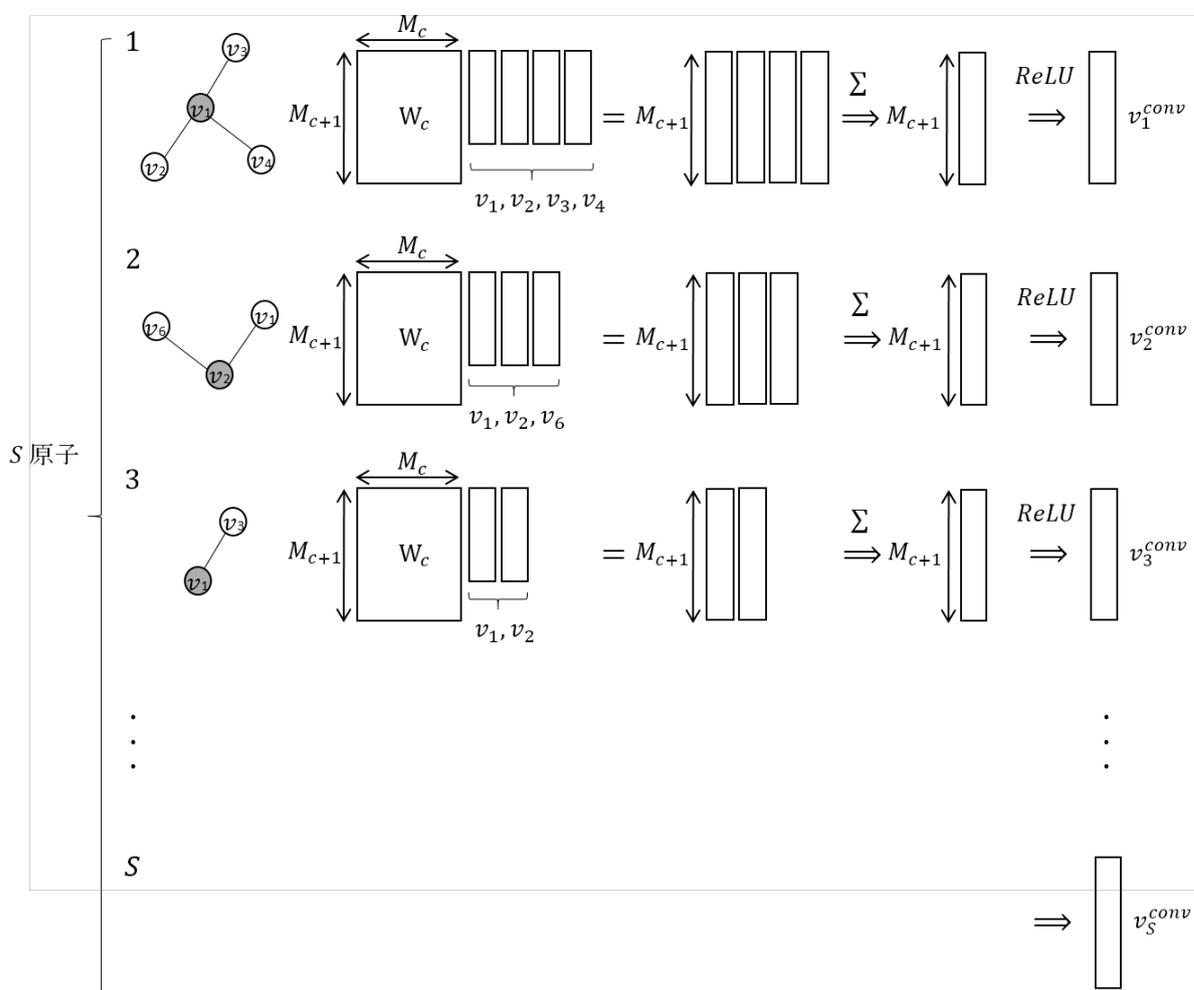


図 4.6 畳み込み層における処理の概念図

畳み込み処理は化合物中の S 原子全てに対してそれぞれ行われる

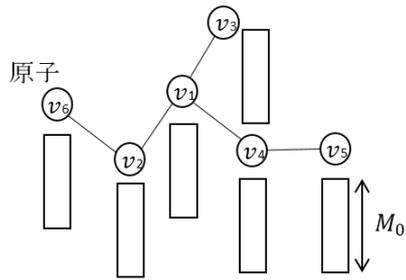
プーリング層

プーリング層では、化合物中の各原子に対してプーリングを適用することで特徴ベクトルの更新を行う。任意の原子に対する特徴ベクトルの更新は任意の原子とそれに隣接している原子の特徴ベクトルを比較することでベクトルの各要素における代表値を決定し、その後これら代表値を要素としたベクトルを新たな特徴ベクトルとして更新する。ここでは各要素における最大値を代表値とする最大プーリングを用いる。

プーリング層における一連の流れを図 4.7 に示す。図 4.7 は原子数 S で構成されている化合物に対して畳み込み層を適用する際の概念図である。ここでは、1 番目の原子に対してプーリング処理を適用する場合を例に挙げて説明する。1 番目の原子に対しプーリング処理を行う場合、1 番目の原子およびそれと隣接している原子である 2 番目、3 番目、4 番目の特徴ベクトルを用いて、1 番目の原子の特徴ベクトルを更新する。

特徴ベクトル $v_1^{conv}, v_2^{conv}, v_3^{conv}, v_4^{conv}$ を比較して各要素における代表値 (ここでは最大値) を決定する。ここでは説明を簡略化するためにプーリング層に入力される各特徴ベクトルの長さを $M_{c+1} = 4$ としている。特徴ベクトル $v_1^{conv}, v_2^{conv}, v_3^{conv}, v_4^{conv}$ の第 1 要素を比較すると、第 1 要素は v_4^{conv} の 4 が最大値となっている。同様に全ての要素に対して最大値を選択し、これらを要素とした $v_1^{pool} = [4 \ 8 \ 3 \ 4]^T$ を新たな特徴ベクトルとして更新する。以上が 1 番目の原子に対してプーリング処理を行う際の流れであり、プーリング層においてはこの処理を S 原子全てに対して行う。

畳み込みニューラルネットワークにおけるプーリング層と同様に特徴の位置感度を低下させる点は同じだが、グラフ畳み込みネットワークの場合、情報の圧縮 (ダウンサンプリング) を行わない点異なる。



Pool

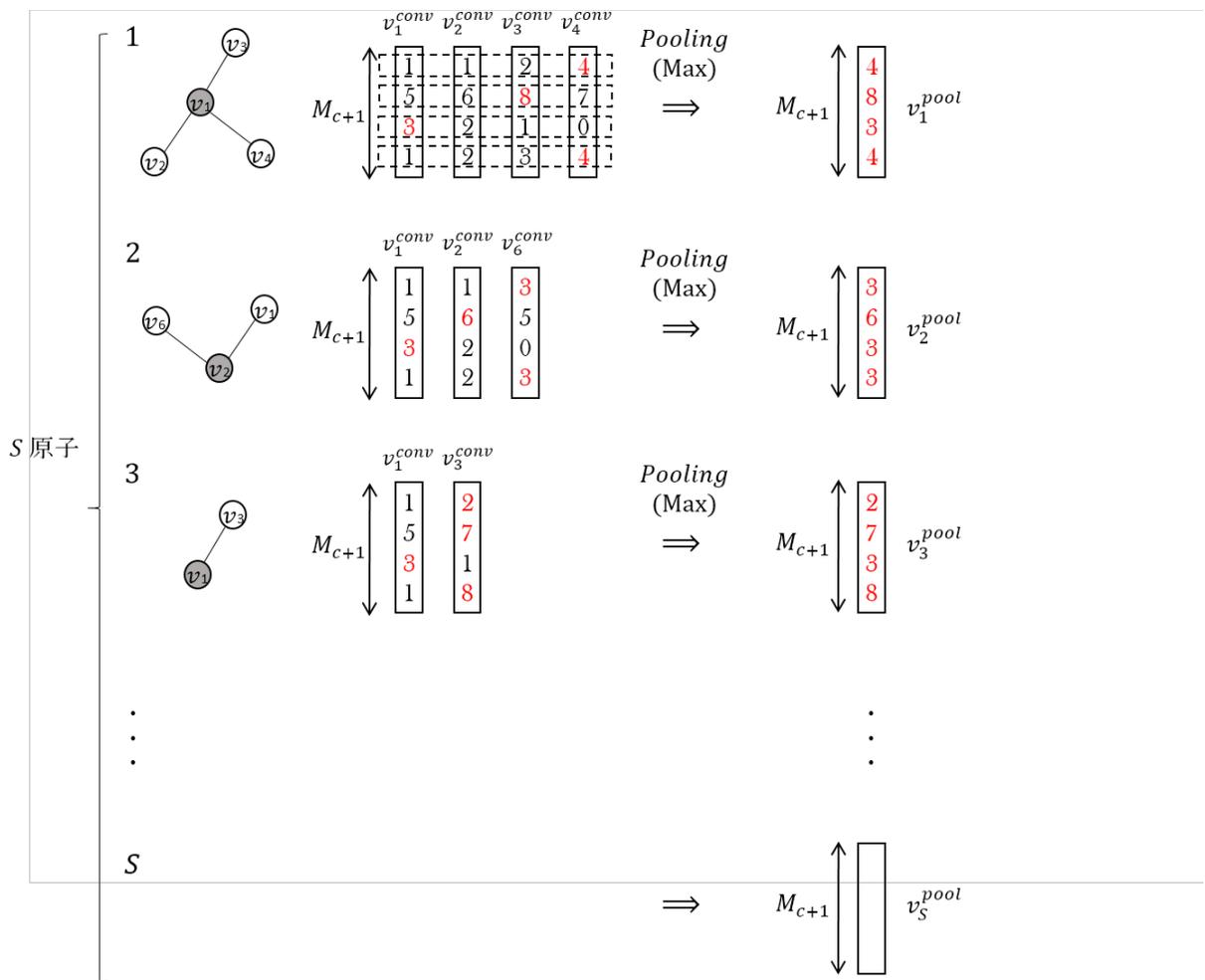


図 4.7 プーリング層における処理の概念図

プーリング処理は化合物中の S 原子全てに対してそれぞれ行われる

集約層

集約層は畳み込み、プーリングの処理を複数回繰り返した後、出力層の手前で適用される。本研究においては、グラフ畳み込みネットワークモデルでは、化合物中の各原子が特徴ベクトルを有しており、化合物中の原子数が S の場合、特徴ベクトルの数も S である。集約層では各原子の特徴ベクトルをまとめることで、最終的に化合物の特徴ベクトルを出力する。

また、グラフ畳み込みネットワークによる一連の流れを図 4.8 に示す。図 4.8 は原子数 S で構成される化合物に対して畳み込み層、プーリング層を複数回適用した後、集約層を適用する際の概念図である。ここでは初期に入力される特徴ベクトル数は S 、各特徴ベクトルの長さは M_0 としている。

入力された特徴ベクトルは畳み込み層、プーリング層を介することによって新たな特徴ベクトルに更新されていく。畳み込み層に入力された長さ M_0 の特徴ベクトルは畳み込み処理によって更新され、長さ M_1 の新たな特徴ベクトルとして出力される。畳み込み層で出力された特徴ベクトルはそのままプーリング層に入力され、プーリング処理によってまた新たな特徴ベクトルとして出力される。プーリング層では前述した通り、特徴ベクトルの長さを変化させる処理を行わないため、出力される特徴ベクトルの長さは M_1 のままである。畳み込みを複数回行う場合は、このプーリング層で出力された特徴ベクトルが次の畳み込み層に入力するベクトルとなる。

畳み込み層、プーリング層により一連の処理によって更新された特徴ベクトルの長さを M_f とすると、集約層ではこの特徴ベクトルを各要素ごとに加算することで 1 つのベクトルにまとめる。これによって出力されるものは M_f の長さをもつ単一ベクトルである。つまり、初期の入力の時点では S 原子それぞれの特徴ベクトルであったものを最終的には化合物自体の特徴ベクトルとして出力する。

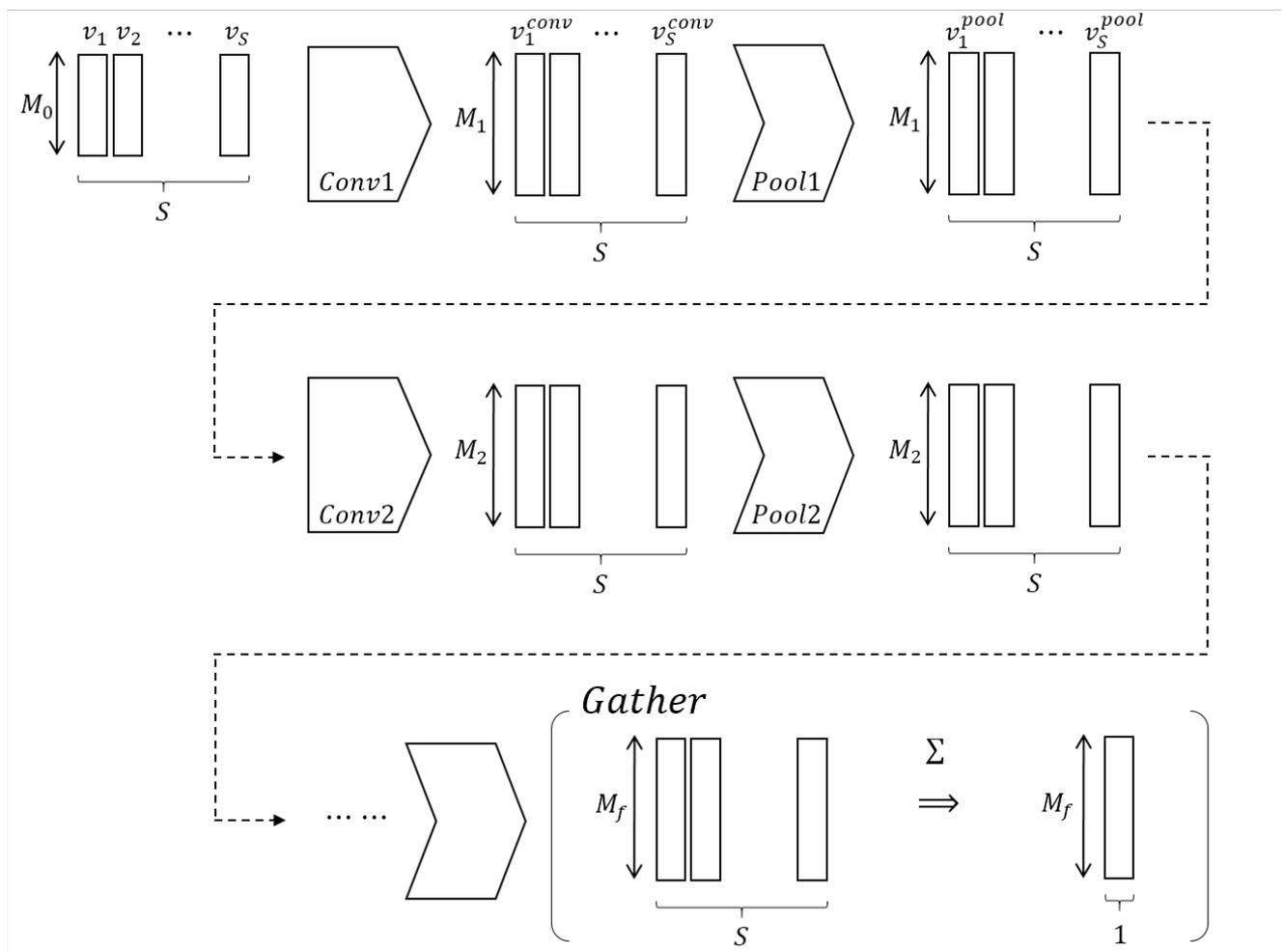
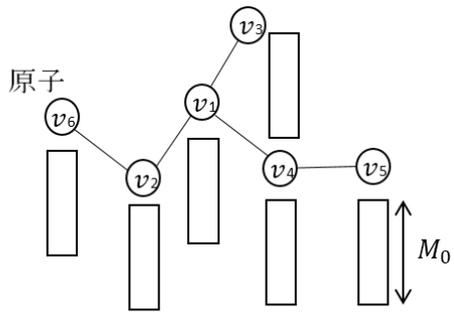


図 4.8 グラフ畳み込みネットワークによる処理の全体像. 特徴ベクトルを更新後、集約層においてベクトルを1つにまとめる.

4.3 グラフ畳み込みネットワークモデルにおける重みの学習

グラフ畳み込みネットワークでは集約層において出力された特徴ベクトルに対して、重み W_f の全結合ネットワークを適用することで最終的な予測値 $\hat{\mathbf{y}} = F(\mathbf{x}; W_1, W_2, \dots, W_f)$ を出力する。予測値 $\hat{\mathbf{y}}$ は畳み込み層における重み $\mathbf{W} = W_1, W_2, \dots, W_f$ によって決定される値である (図 4.9)。

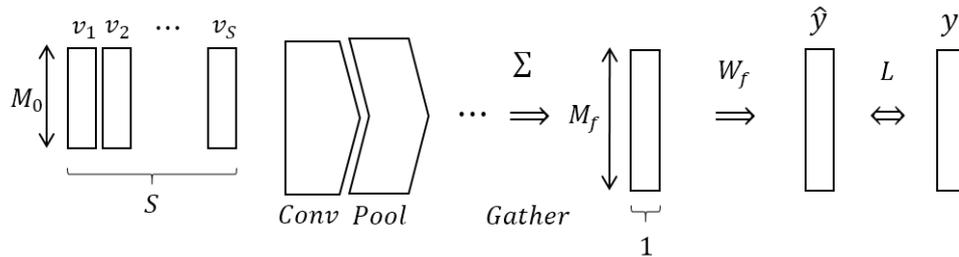


図 4.9 学習の全体像. 集約層で出力された特徴ベクトルは重み W_f によって学習の予測値に変換し、訓練データとの誤差を損失関数 L によって評価する。

ここで、訓練データ \mathbf{y} と予測値 $\hat{\mathbf{y}}$ の誤差を表す損失関数 $L(\mathbf{y}, \hat{\mathbf{y}})$ を定義する。グラフ畳み込みネットワークにおける学習の目標は、畳み込み層での重みを更新することによって、この損失関数 L を最適化することである。本研究では損失関数 L を最適化するための手法として勾配降下法を用いている。勾配降下法では、現時点における重み \mathbf{W} を損失関数の値が減少する方向、つまりは負の勾配方向に変化させる操作を何度も繰り返すことによって損失関数の最小化を試みる。現時点における重みを \mathbf{W}^t 、変化させた後の重みを \mathbf{W}^{t+1} とすると下式のように重みの更新を行う。ここで α^0 は学習係数と呼ばれ、重みの更新量の大きさを定める値である。 α^t が定数の場合は最急降下法と呼ばれるが、一般には学習の収束に応じて α^t を変化させる手法が提案されており、本研究では AdaGrad 法を用いて調整している。

$$\mathbf{W}^{t+1} = \mathbf{W}^t - \alpha^t \frac{\partial L}{\partial \mathbf{W}} \quad (4.3.1)$$

$$\text{where } \frac{\partial L}{\partial \mathbf{W}} = \left[\frac{\partial L}{\partial W_1} \quad \frac{\partial L}{\partial W_2} \quad \dots \quad \frac{\partial L}{\partial W_f} \right]^T \quad (4.3.2)$$

4.4 データセット

本章における学習に用いたデータは、第 2 章で構築した代謝経路上に存在する 469 のアルカロイド化合物の構造情報および由来する出発物質の情報である。予測対象である出発物質は本研究で考慮している出発物質のうち、出現頻度が 10 以上である 16 の出発物質を選択した。予測する出発物質およびそれらに属する化合物数を表 4.1 に示す。このデータセット中の化合物は単一の出発物質から導出されるものと複数の出発物質から導出されるものが混在しているため、マルチラベルのデータとなっている。

表 4.1 学習の対象とした出発物質とそれらによって誘導される二次代謝物数

Starting Material	Frequency
L-Tryptophan	160
L-Tyrosine	114
L-Proline	61
Anthranilate	59
L-Lysine	44
L-Arginine	42
Indole-3-glycerolphosphate	24
GGPP	24
L-Alanine	28
L-Aspartic acid	28
Secologanin	26
Malonyl-CoA	23
L-Phenylalanine	20
IPP	23
L-Histidine	12
Cholesterol	13

4.5 結果および考察

グラフ畳み込みネットワークを利用して化合物の構造データからその化合物の出発物質の予測を行う。本研究で利用したグラフ畳み込みネットワークの構成を図4.10に示す。

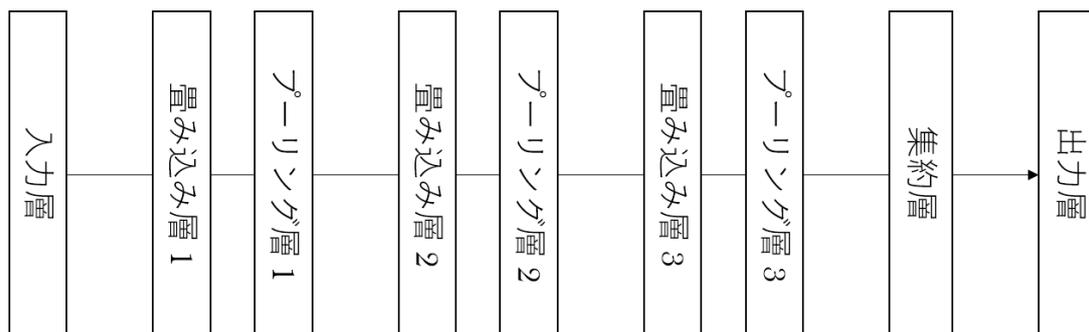


図 4.10 本研究におけるグラフ畳み込みネットワークの構成

本研究におけるモデルでは3度の畳み込み処理を繰り返した後に、集約層において化合物中の特徴ベクトルをまとめ上げることで、各化合物に対して単一の特徴ベクトルを出力層にて出力する。いずれの畳み込み層においても活性化関数としてReLU関数を使用している。また、特徴ベクトルの長さは75、その他の学習に関するハイパーパラメータについては過去の研究において推奨されている値を参考にし、 $\alpha^0 = 1e - 3$ 、 $\epsilon = 1e - 8$ 、 $\beta_1 = 0.9$ 、 $\beta_2 = 0.999$ としている。

ここでは、469の化合物に対して学習を行う際に469の化合物の内1つの化合物をテストデータとし、残りの468の化合物をトレーニングデータとした。これにより、学習されたモデルに対してLeave-one-out交差検証(LOOCV)を行うことでその性能を評価した。また、一つのトレーニングデータに対する学習の繰り返し回数(エポック数)は30に設定している。

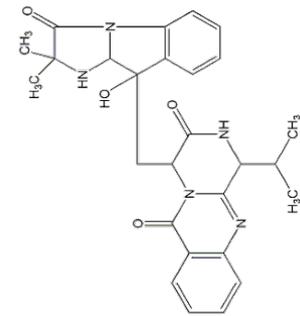
次に化学構造から出発物質を予測させた結果を示していく。表 4.2 は各出発物質に対する予測の精度を表している。この表において、TP(True Positive) は出発物質であると正しく予測した数、FP(False Positive) は出発物質であると誤って予測した数、TN(True Negative) は出発物質ではないと正しく予測した数、FN(False Negative) は出発物質ではないと誤って予測した数である。また、予測の精度として $(TP + TN)/N$ を表記している。N は全体の化合物数であり、ここでは N=469 である。

表 4.2 グラフ畳み込みネットワークによる出発物質の予測結果

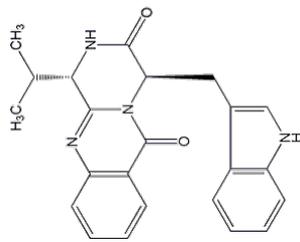
Starting Material	Frequency	TP	FP	TN	FN	$(TP+TN)/N$
L-Tryptophan	160	160	5	304	0	0.989
L-Tyrosine	114	114	0	355	0	1.000
L-Proline	61	60	3	405	1	0.991
Anthranilate	59	59	10	400	0	0.979
L-Lysine	44	44	13	412	0	0.972
L-Arginine	42	42	1	426	0	0.998
Indole-3-glycerolphosphate	24	24	0	445	0	1.000
GGPP	24	24	0	445	0	1.000
L-Alanine	28	27	16	425	1	0.964
L-Aspartic acid	28	28	4	437	0	0.991
Secologanin	26	25	0	443	1	0.998
Malonyl-CoA	23	23	3	443	0	0.994
L-Phenylalanine	20	20	24	425	0	0.949
IPP	23	23	2	444	0	0.996
L-Histidine	12	12	0	457	0	1.000
Cholesterol	13	13	0	456	0	1.000

この結果から、L-Alanine における FDR(False Discovery Rate) は約 37%、L-Phenylalanine における FDR は約 55% であり FP が多くなっている出発物質が見られるものの、各出発物質の精度の平均値をとると約 98.8% であり非常に高い精度での予測が行えていることが確認できる。グラフ畳み込みネットワークを用い出発物質の予測を高精度に行うことができた理由の一つとして、各アルカロイド化合物の化学構造と出発物質の構造との間に強い関係性があることが挙げられる。第 3 章においてアルカロイド化合物の環骨格と代謝経路の関係性について考察したが、類似した環骨格を有するアルカロイド化合物はそのほとんどが同一の代謝経路から生合成されていた。また、これまでも化合物の化学構造から何らかのラベル、物性値を予測する研究は行われてきた。しかしながら、それらの予測では化合物の構造情報をフィンガープリント法などによりベクトル化し、それを学習モデルの入力としてきたため、原子間の相互作用は一切考慮されていなかった。グラフ畳み込みネットワークでは、化合物中の各原子に特徴ベクトルを割り当てて学習を行うことで、これまでに考慮されていなかった原子間の相互作用を上手く取り入れている。本研究における出発物質予測もこのことが高い予測精度に繋がったものと思われる。

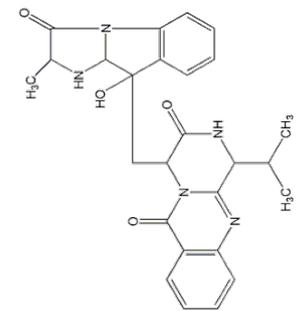
次にグラフ畳み込みネットワークによる出発物質の予測結果から、FP と判定された化合物の構造および代謝経路の関連性についての考察を行う。各出発物質の予測において FP と判定された化合物の構造を図 4.12 および図 4.20 に、また、FP と判定された化合物をマッピングしたアルカロイド代謝経路を図 4.21 および図 4.28 に示す。



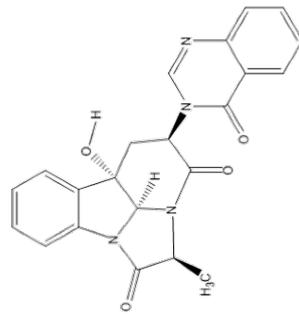
Fiscalin C



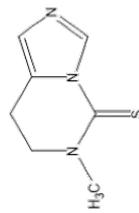
Fiscalin B



Fiscalin A



Chaetominine



Zapotidine

図 4.11 L-Tryptophan と誤って予測された化合物

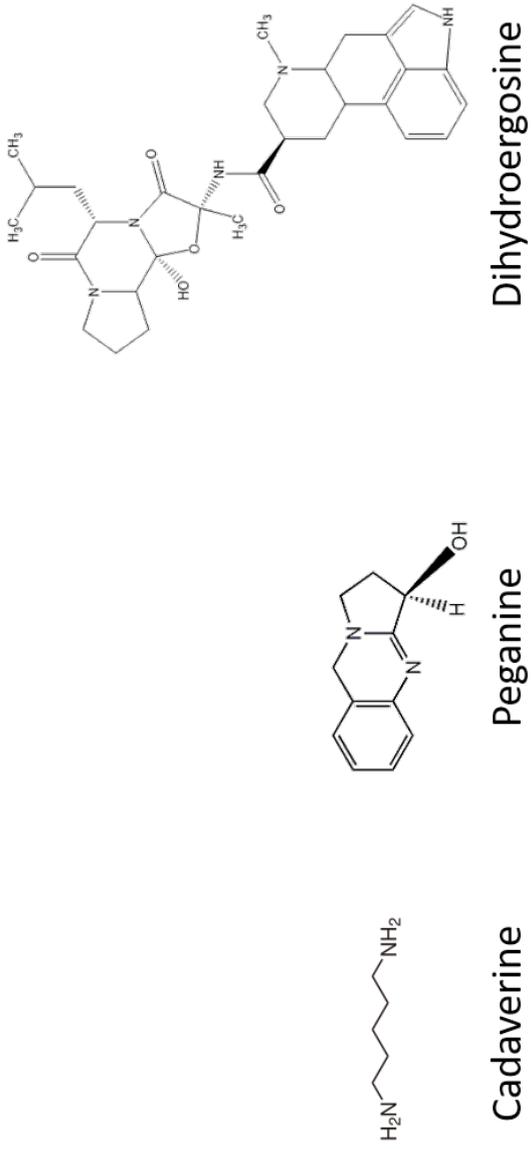


図 4.12 L-Proline と誤って予測された化合物

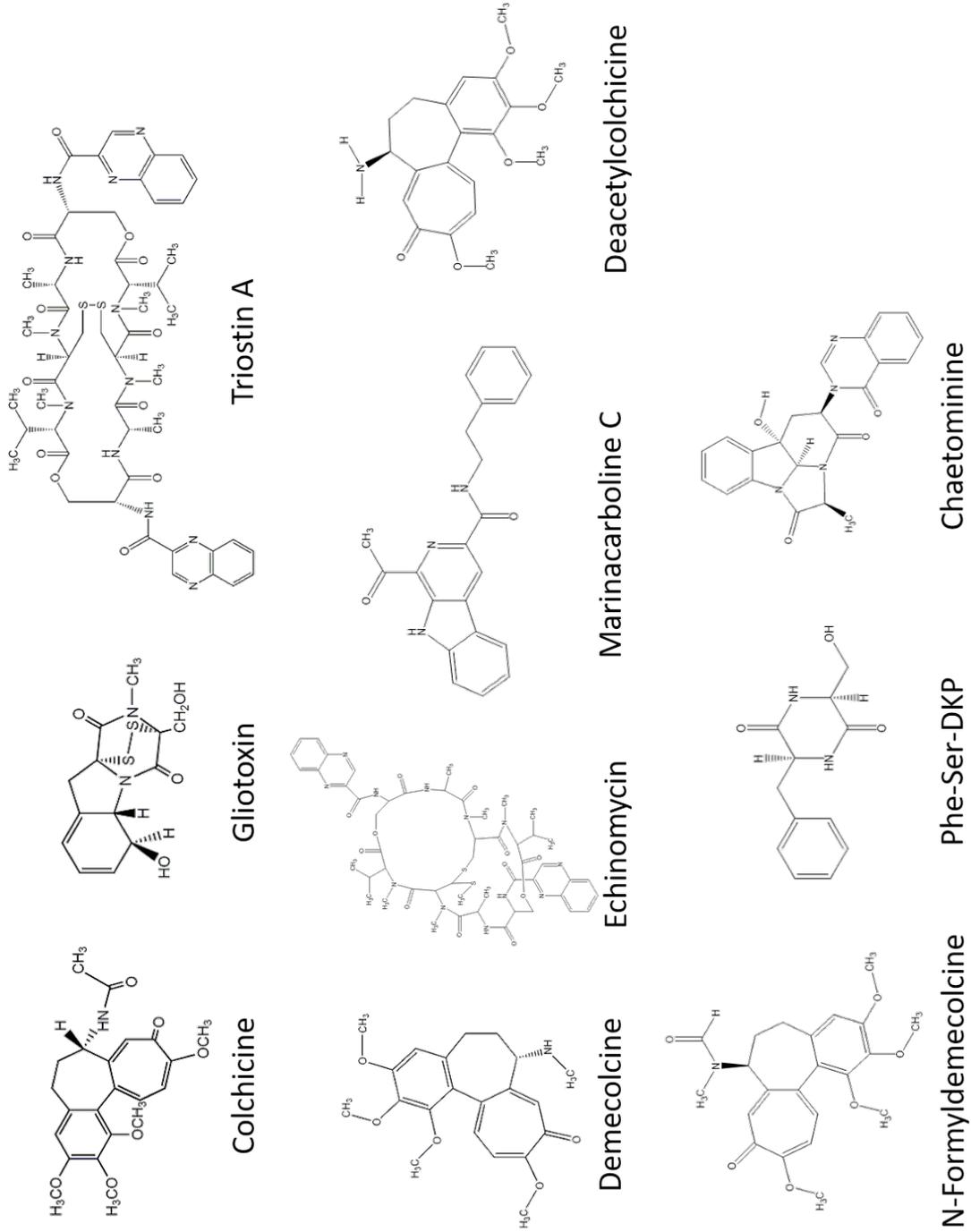


図 4.13 Anthranilate と誤って予測された化合物

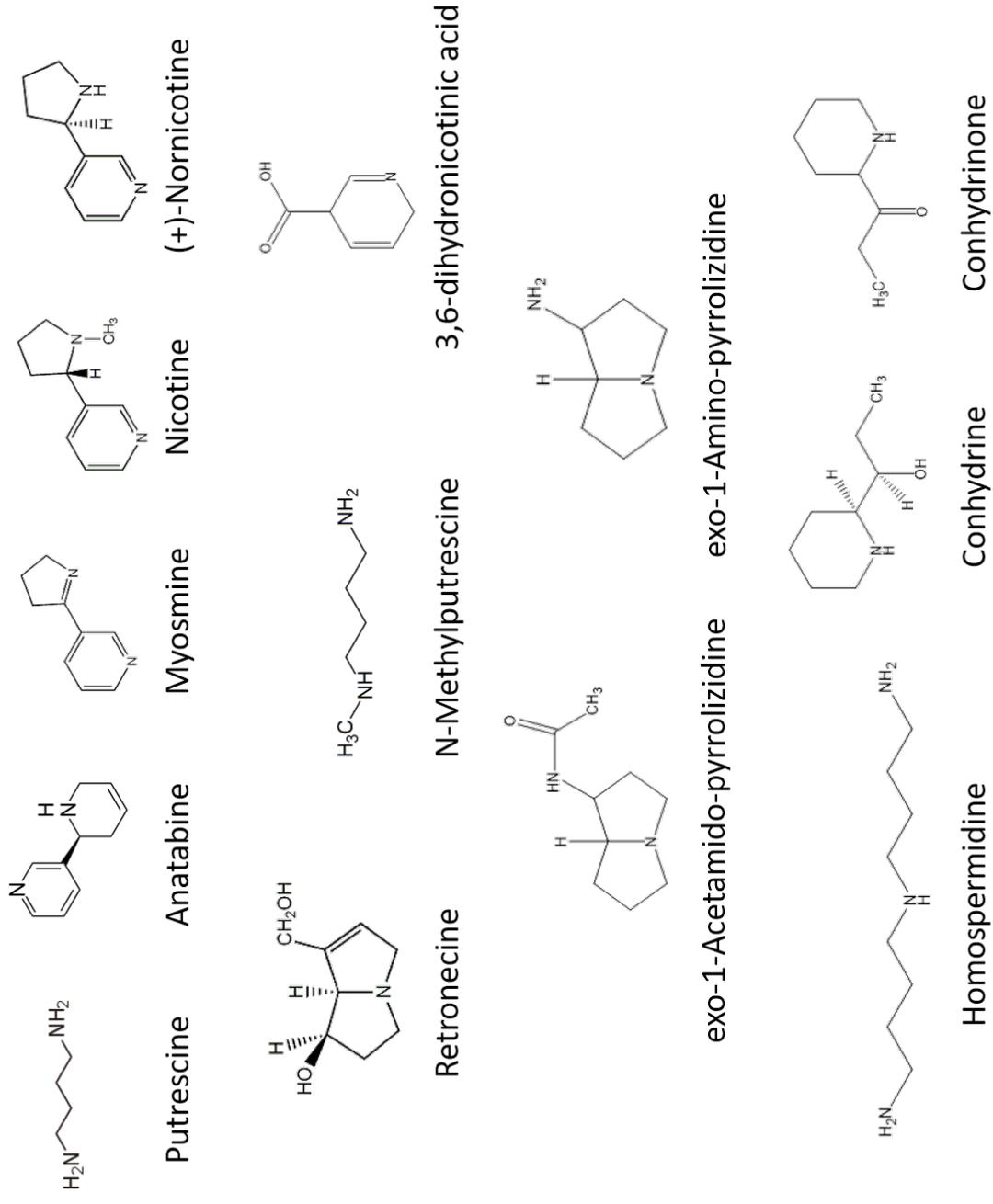
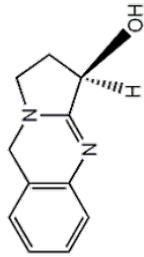


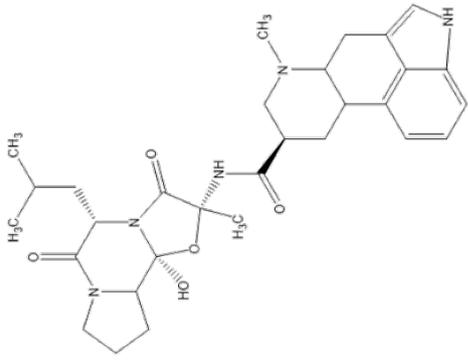
図 4.14 L-Lysine と誤って予測された化合物



Cadaverine



Peganine



Dihydroergosine

図 4.15 L-Arginine と誤って予測された化合物

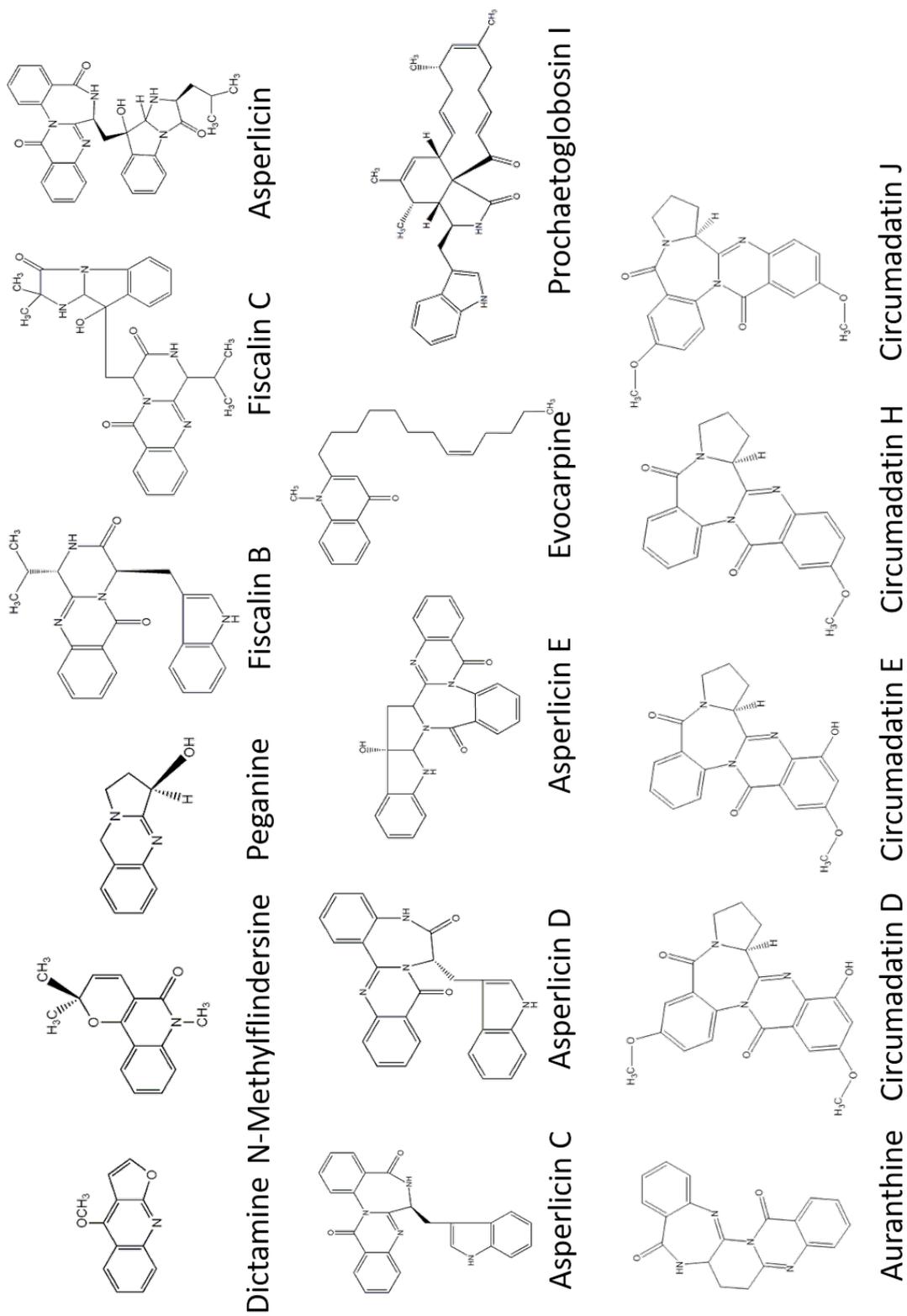


図 4.16 L-Alanine と誤って予測された化合物

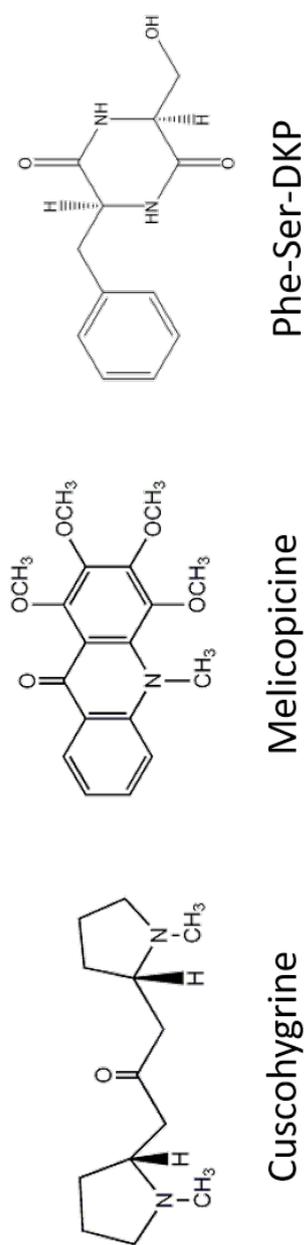


図 4.17 Malonyl-CoA と誤って予測された化合物

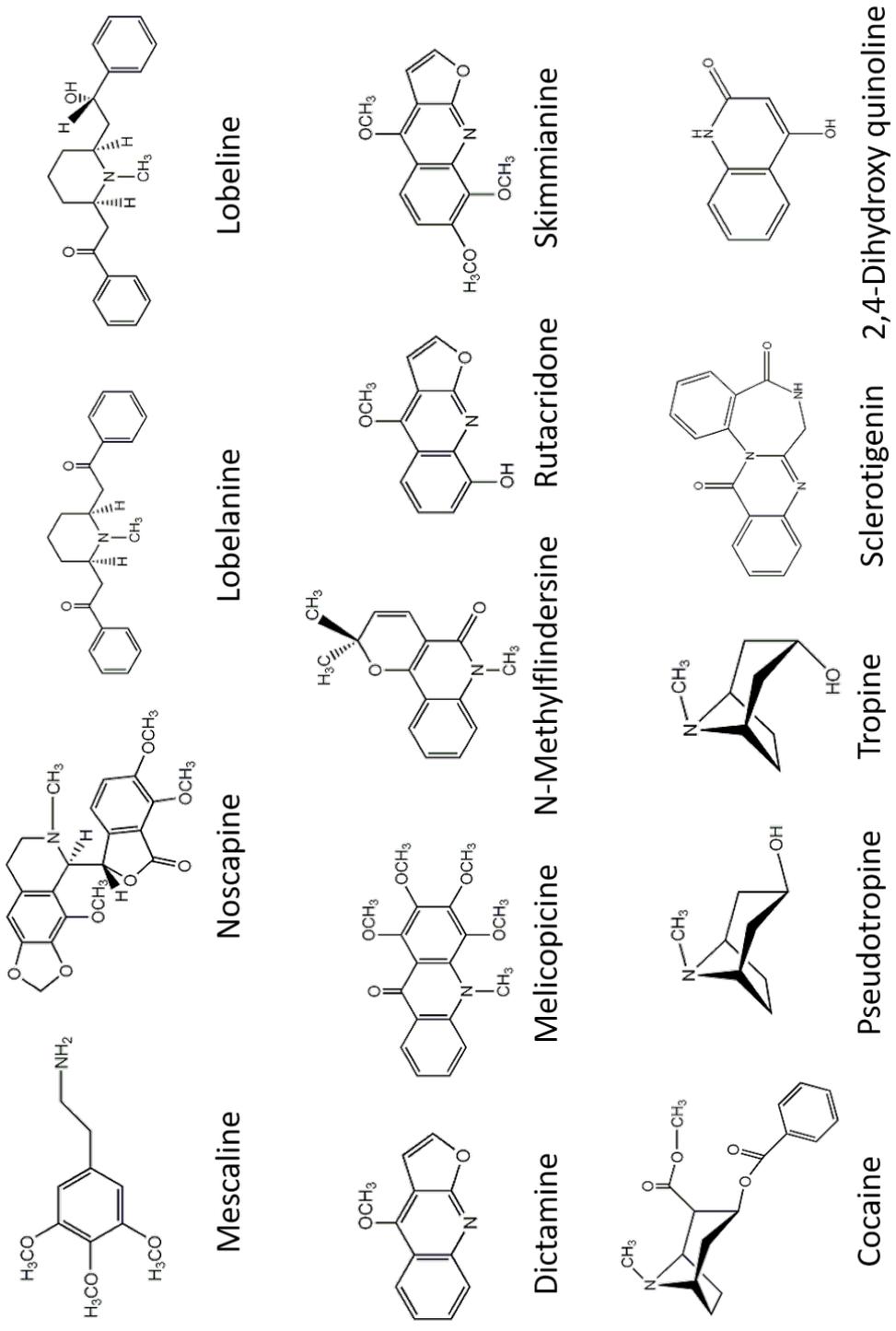
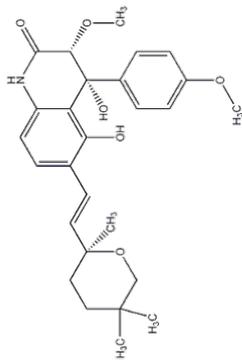
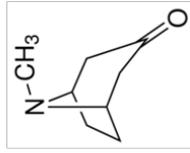


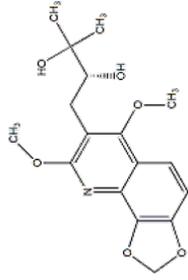
図 4.18 L-Phenylalanine と誤って予測された化合物



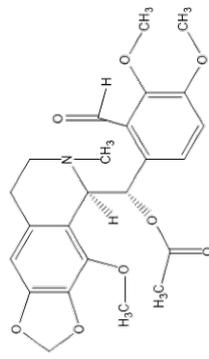
Penigequinolone A



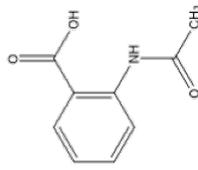
Tropinone



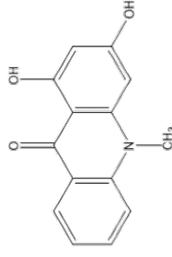
(+)-Orixine



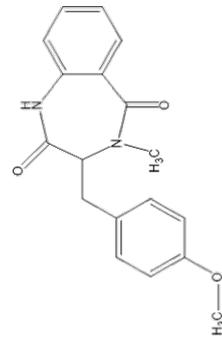
3-O-Acetylpapaveroxine



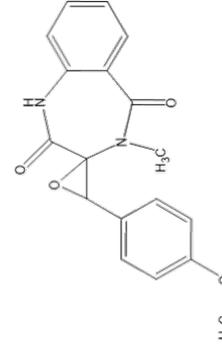
N-Acetyl-anthranilic acid



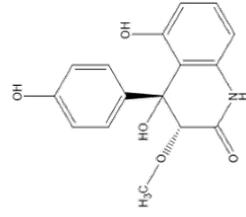
1,3-Dihydroxy-N-methylacridone



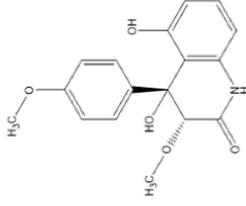
4'-Methoxycyclopeptin



(-)-4'-Methoxycyclophenin

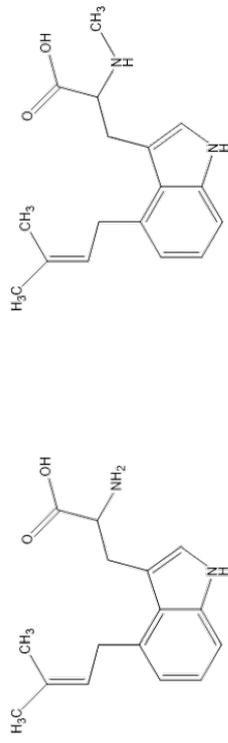


14-Methoxyviridicatin



Quinolinone B

図 4.19 L-Phenylalanine と誤って予測された化合物



4-Dimethylallyltryptophan

4-DMA-L-tryptophan

図 4.20 IPP と誤って予測された化合物

2

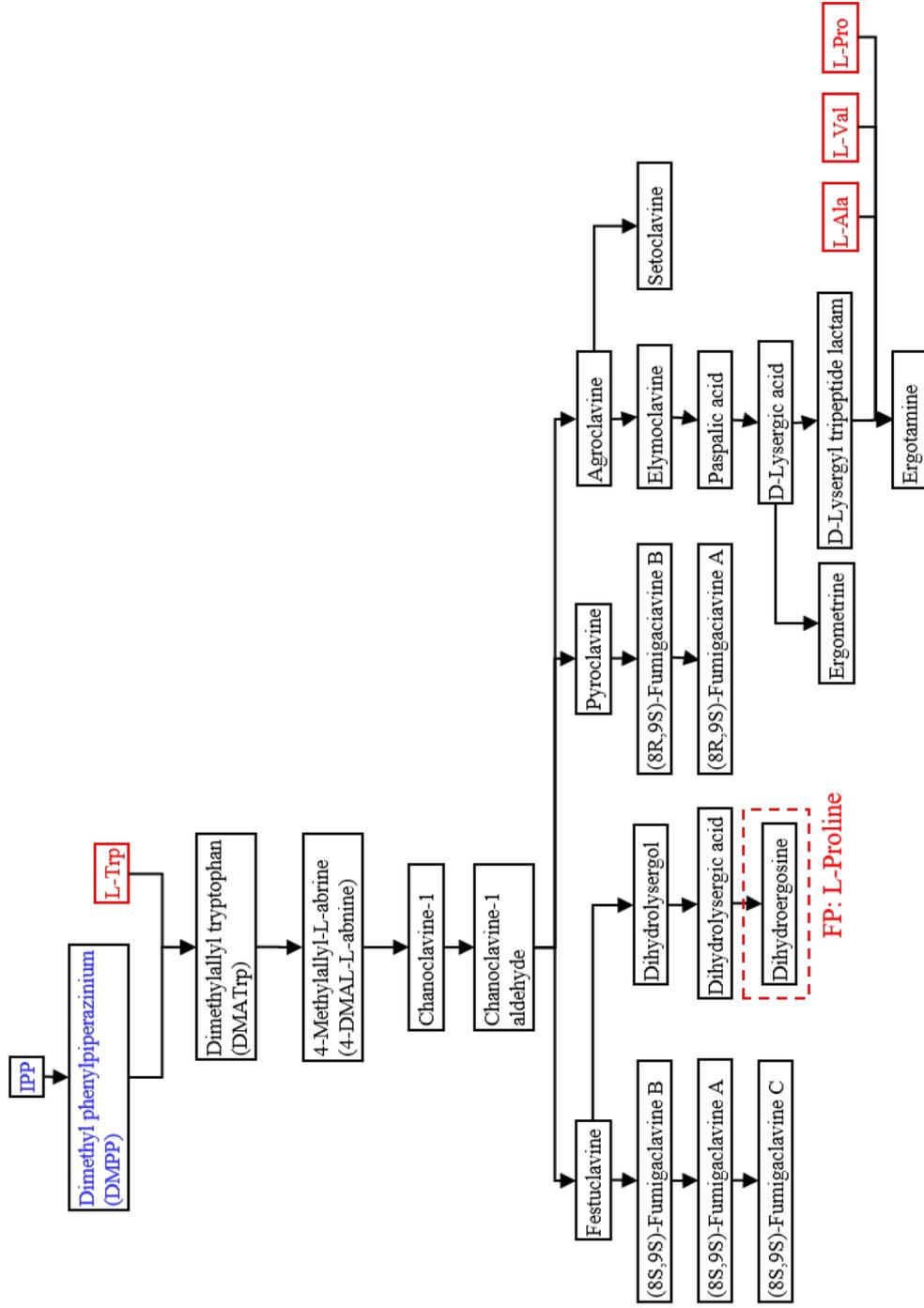


図 4.22 誤って予測された化合物の経路上へのマッピング (2)

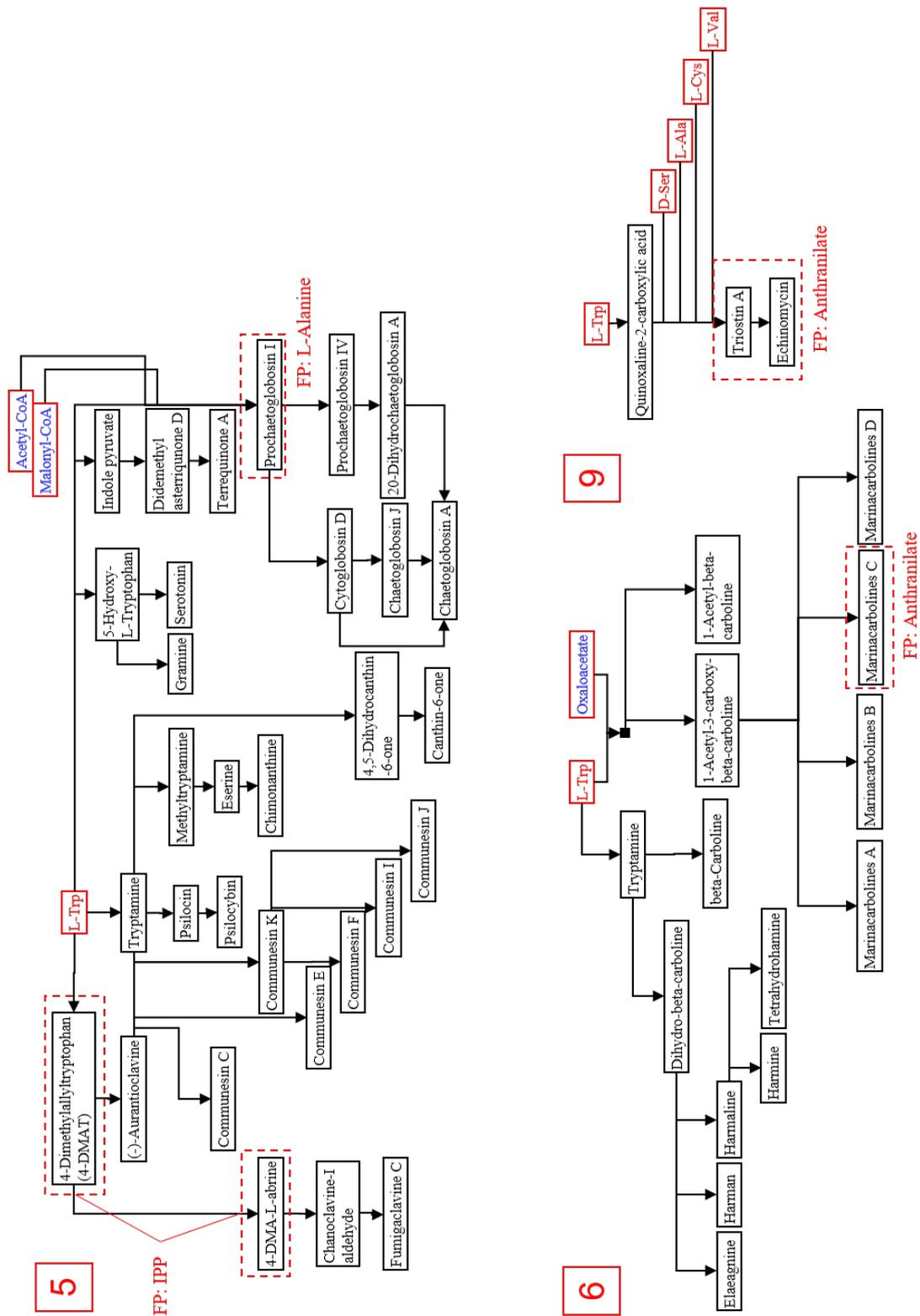


図 4.23 誤って予測された化合物の経路上へのマッピング (3)

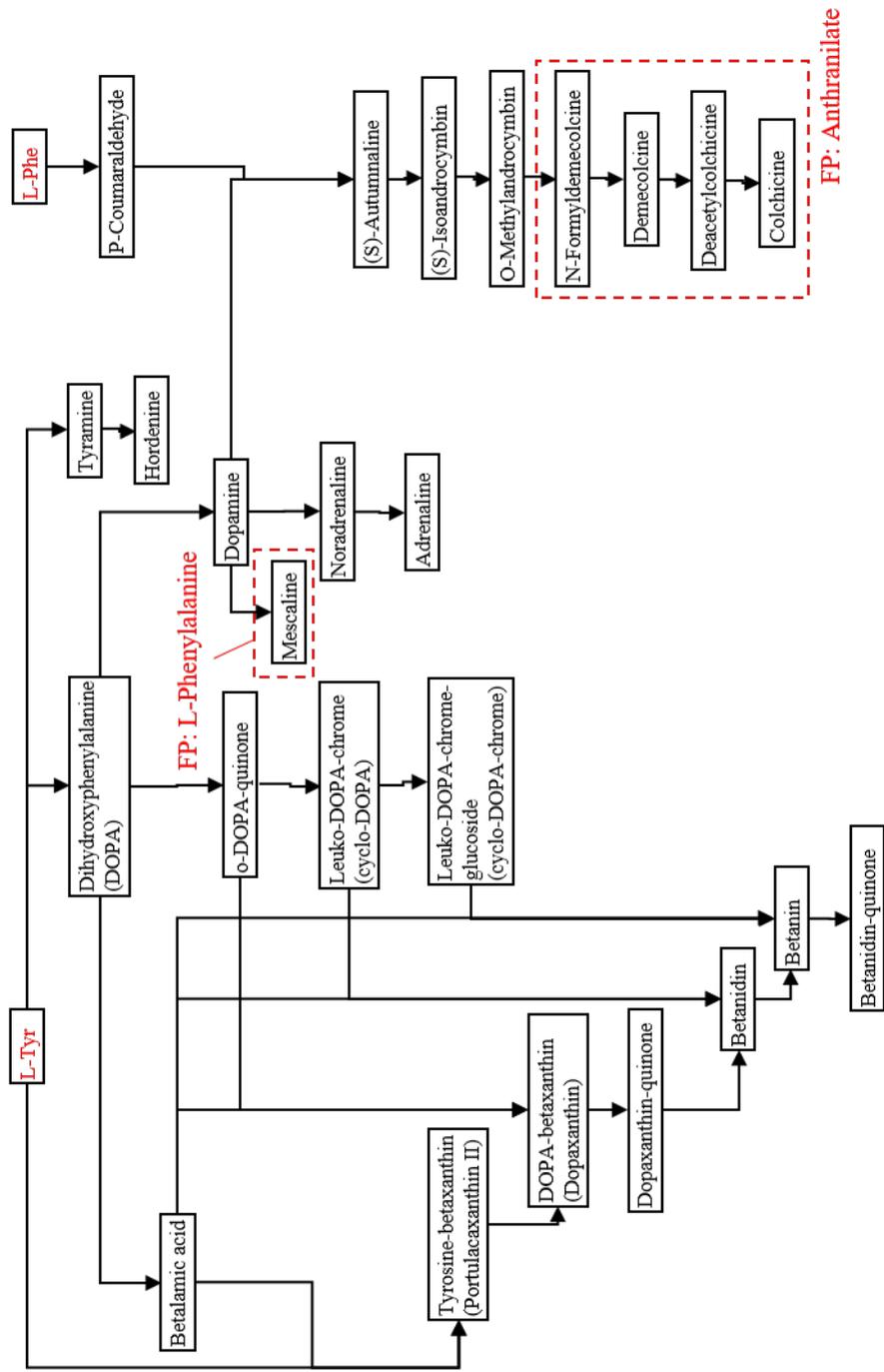


図 4.24 誤って予測された化合物の経路上へのマッピング (4)

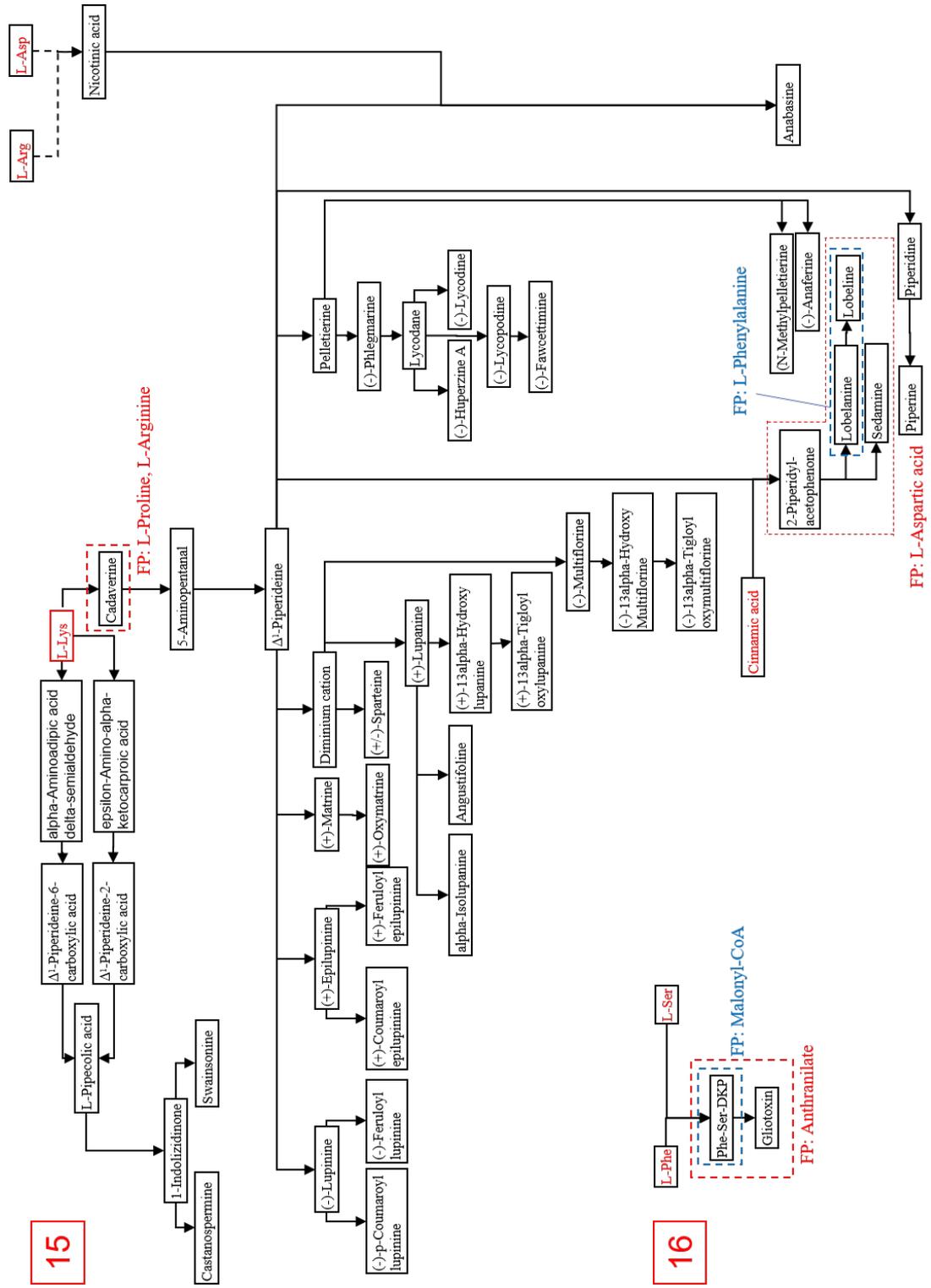


図 4.26 誤って予測された化合物の経路上へのマッピング (6)

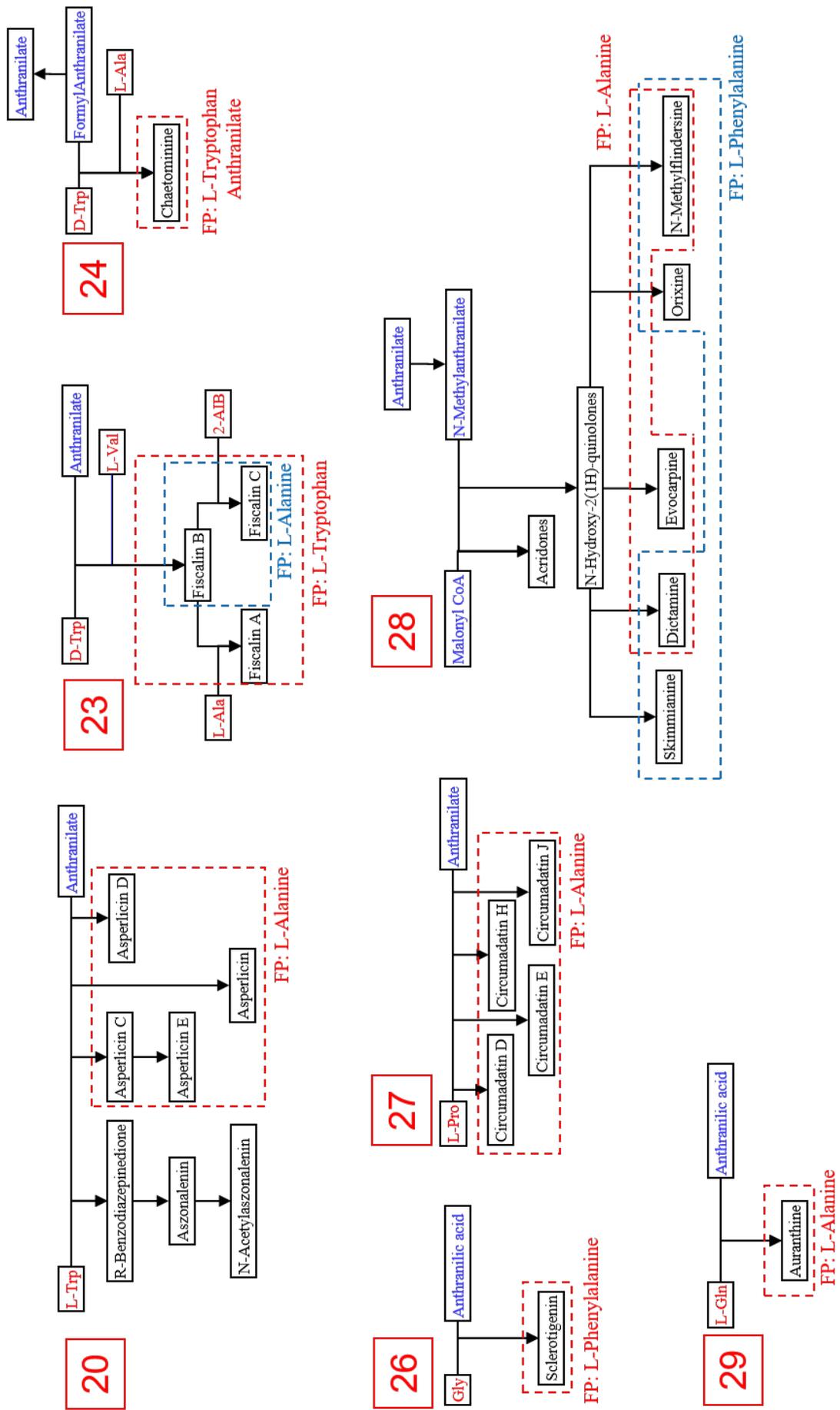


図 4.28 誤って予測された化合物の経路上へのマッピング (8)

図 4.21~図 4.28 では、枠に囲まれている部分が FP と判定された化合物を表している。例えば図 4.22 において、Dihydroerogosine の出発物質は IPP と L-Tryptophan であるが、予測結果では Proline が出発物質であるという誤った予測をしている。

まず、代謝経路全体を俯瞰すると図 4.21 における Tropinone、Tropine、Pseudotropine や図 4.23 における Triostin A および Echinomycin のように、多くの出発物質から誘導される化合物が誤って予測されている。また図 4.25 において示されているアルカロイド代謝経路 12 上の化合物は、その全てが L-Tyrosine にのみ誘導されるが、3-O-Acetylpapaveroxine および Noscapine は誤って予測されている。これは 3-O-Acetylpapaveroxine および Noscapine は、出発物質である L-Tyrosine から多くの代謝反応を経て誘導されている、つまり代謝経路上では L-Tyrosine から離れた位置にあるために誤って予測されている可能性がある。

図 4.12 で示したように出発物質が L-Tryptophan と誤って予測した化合物は 5 つ存在するが、代謝経路からも確認できるようにこれらは全て D-Tryptophan から誘導される化合物であり、この予測は必ずしも間違っただけのものとはいえない。

図 4.27 および図 4.28 から Anthranilate に関連する化合物の多くが誤った予測をしていることが確認できる。Anthranilate は様々なアミノ酸と組み合わせることによって多くのアルカロイドの前駆体となる物質である。しかしながら、第 3 章の結果から分かるように Anthranilate から誘導される化合物は、生合成される経路が異なっても類似した構造を有しており、このことが予測結果に影響を与えているものとみられる。また、出発物質の中で L-Phenylalanine が最も FP が多い出発物質であるが、L-Phenylalanine と誤って予測された化合物の本来の出発物質をみると L-Tyrosine 由来のものが多いことが確認できる。これは L-Phenylalanine と L-Tyrosine の構造が非常に酷似しているためであると考えられる。

また、Indole-3-glycerolphosphate、GGPP、L-Histidine、Cholesterol においては予測結果が FP、FN とともに 0 であり、グラフ畳み込みネットワークによって完璧に予測できたといえる。これらは経路上で他の出発物質と組み合わせっていないこと、また、それらを出発物質とする化合物の構造の特異性によって得られた結果であると考えられる。

4.6 まとめ

本章では、近年注目されているグラフ構造を対象とした機械学習手法であるグラフ畳み込みネットワークを用いたモデルを二次代謝物に適用することで、二次代謝物の構造情報からそれらがどの出発物質から誘導されるのか、即ち、いずれの代謝経路に由来する二次代謝物であるのかの予測を試みた。アルカロイド化合物の構造情報に対して学習手法を適用することで、出発物質予測に関するグラフ畳み込みネットワークの妥当性の検証を行った。

第2章において構築された代謝マップに含まれる469のアルカロイド化合物に対して、16の出発物質 (L-Tryptophan、L-Tyrosine、L-Proline、Anthranilate、L-Lysine、L-Arginine、Indole-3-glycerolphosphate、GGPP、L-Alanine、L-Aspartic acid、Secologanin、Malonyl-CoA、IPP、L-Phenylalanine、L-Histidine、Cholesterol) を予測したところ、約98%と非常に高い精度で予測を行うことができた。また、出発物質の予測結果のうち False Positive と判定されたものが多くみられたため、これらに関して、アルカロイド代謝経路および第3章における化合物の分類結果に基づき考察を行った。本章の結果から、化合物に対するグラフ畳み込みネットワークモデルを用いた学習は生物学、生化学の分野において非常に強力であると確認できたと同時に、今後は本研究で検証を行った出発物質の予測のみに留まらず、さらに応用分野を広げることで今まで困難であった問題を解決することが期待される。

第5章 総括

生物学と化学の両分野において、統合した知識に基づいて化学構造と二次代謝物の生物活性を包括的に理解することは重要な課題の一つであり、これまでにその理解を深めるための様々な研究が行われてきた。本研究においては化学構造と二次代謝物の生物活性の包括的に理解をデータサイエンスを通じて試み、特に代謝物とそれを創出する代謝経路の相互関係性について、第一の観点として、二次代謝物の構造情報から本研究において提案する環構造に基づく分類を行い、その分類を代謝経路上に投影することで代謝モジュールとして抽出すること、第二の観点として、グラフ畳み込みネットワークを用いた機械学習モデルを適用することで、二次代謝物の構造情報からそれらがどの出発物質から誘導されるのかの予測すること、以上の二つの観点からアプローチを行い妥当性を検討した。

本研究で示した結果は、そのものが生物学、化学、薬学をはじめとする様々な分野において活用されることで新しい知見を提供することが期待される。また、本研究において提案した手法を応用することで各分野間の知識の統合・体系化を促し、融合領域の研究を今後ますます発展させることになると思われる。

謝辞

本研究を進めるに当たり、奈良先端科学技術大学院大学 情報科学研究科 金谷重彦 教授に終始懇切丁寧なる御指導・御鞭撻を賜りました。また、研究者としてのあるべき姿をその身を持って示して下さいました。厚く御礼申し上げます。

また、研究発表及び修士論文作成に当たり、貴重な御助言を賜りました副指導教員である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。

本研究を進めるに当たり、大変有益なる御助言、御助力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 特任教授に深く感謝致します。

本研究を進めるに当たり、日頃から数多くの御意見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.ALTAFA-UL-AMIN 准教授に深く感謝致します。

本研究を進めるに当たり、門外漢で右も左も分からない私に多くのご教示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 准教授に深く感謝致します。

本研究を進めるに当たり、率直かつ的確な御意見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 黄銘 助教に深く感謝致します。

研究に関することのみならず、日常生活においても奈良先端科学技術大学院大学 計算システムズ生物学研究室に在籍する皆様ならびに修了生の方々には掛け替えのないものを賜り、心からの支えになって頂きました。また、母校である茨城工業高等専門学校 電子情報工学科の先生方、友人たちの存在無しに私が研究者の道歩むことはありませんでした。何よりもここに至るまでに祖父、祖母、父、母、兄には様々な支援・後押しをして頂きました。ここに深く感謝を申し上げます。

本論文はここでは記しきれない程、多くの方々のご協力なしには完成し得なかったものです。未だに浅学非才であるこの身に期待をし、ご支援をして下さった多くの皆様に心より御礼申し上げます。

研究業績

査読付き論文

1. Ryohei Eguchi, Naoaki Ono, Hisayuki Horai, Md. Altuf-Ul Amin, Aki Morita Hirai, Jun Kawahara, Shoji Kasahara, Tomoaki Endo, Shigehiko Kanaya, Classification of Alkaloid Compounds Based on Subring Skeleton (SRS) Profiling: On Finding Relationship of Compounds with Metabolic Pathways, Journal of Computer Aided Chemistry, Volume18, p58-75, 2017
2. Ryohei Eguchi, Mohammand Bozlul Karim, Pingzhao Hu, Tetsuo Sato, Naoaki Ono, Kanaya Shigehiko and 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, BMC Bioinformatics(Under Review)

査読付き国際学会発表

1. Ryohei Eguchi, Ono Naoaki, Hisayuki Horai, Md. Altuf-Ul Amin, Shigehiko Kanaya, Classification of alkaloid compounds based on subring skeleton(SRS), NIG International Symposium 2017, May 29th, 2017

参考文献

- [1] Steve Kelling, Wesley M Hochachka, Daniel Fink, Mirek Riedewald, Rich Caruana, Grant Ballard, and Giles Hooker. , Data-intensive science: a new paradigm for biodiversity studies. , *BioScience*, Vol. 59, No. 7, pp. 613–620, 2009.
- [2] Y Shinbo, Y Nakamura, Md Altaf-Ul-Amin, H Asahi, K Kurokawa, M Arita, K Saito, D Ohta, D Shibata, and S Kanaya. , Knapsack: a comprehensive species-metabolite relationship database. , *Plant metabolomics*, pp. 165–181, 2006.
- [3] Kensuke Nakamura, Naoki Shimura, Yuuki Otabe, Aki Hirai-Morita, Yukiko Nakamura, Naoaki Ono, Md Altaf Ul-Amin, and Shigehiko Kanaya. , Knapsack-3d: a three-dimensional structure database of plant metabolites. , *Plant and Cell Physiology*, Vol. 54, No. 2, pp. e4–e4, 2013.
- [4] Farit Mochamad Afendi, Taketo Okada, Mami Yamazaki, Aki Hirai-Morita, Yukiko Nakamura, Kensuke Nakamura, Shun Ikeda, Hiroki Takahashi, Md Altaf-Ul-Amin, Latifah K Darusman, et al. , Knapsack family databases: integrated metabolite–plant species databases for multifaceted plant research. , *Plant and Cell Physiology*, Vol. 53, No. 2, pp. e1–e1, 2011.
- [5] Shun Ikeda, Takashi Abe, Yukiko Nakamura, Nelson Kibinge, Aki Hirai Morita, Atsushi Nakatani, Naoaki Ono, Toshimichi Ikemura, Kensuke Nakamura, Md Altaf-Ul-Amin, et al. , Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the knapsack motor-cycle database. , *Plant and Cell Physiology*, Vol. 54, No. 5, pp. 711–727, 2013.
- [6] Sony Hartono Wijaya, Husnawati Husnawati, Farit Mochamad Afendi, Irmanida Batubara, Latifah K Darusman, Md Altaf-Ul-Amin, Tetsuo Sato, Naoaki Ono, Tadao Sugiura, and Shigehiko Kanaya. , Supervised clustering based on dpluso: Prediction of plant-disease relations using jamu formulas

- of knapsack database. , *BioMed research international*, Vol. 2014, , 2014.
- [7] Yukiko Nakamura, Farit Mochamad Afendi, Aziza Kawsar Parvin, Naoaki Ono, Ken Tanaka, Aki Hirai Morita, Tetsuo Sato, Tadao Sugiura, Md Altaf-Ul-Amin, and Shigehiko Kanaya. , Knapsack metabolite activity database for retrieving the relationships between metabolites and biological activities. , *Plant and Cell Physiology*, Vol. 55, No. 1, pp. e7–e7, 2014.
- [8] Yuki Ohtana, Azian Azamimi Abdullah, Md Altaf-Ul-Amin, Ming Huang, Naoaki Ono, Tetsuo Sato, Tadao Sugiura, Hisayuki Horai, Yukiko Nakamura, Klaus W Lange, et al. , Clustering of 3d-structure similarity based network of secondary metabolites reveals their relationships with biological activities. , *Molecular informatics*, Vol. 33, No. 11-12, pp. 790–801, 2014.
- [9] Peter J Facchini. , Alkaloid biosynthesis in plants: biochemistry, cell biology, molecular regulation, and metabolic engineering applications. , *Annual review of plant biology*, Vol. 52, No. 1, pp. 29–66, 2001.
- [10] Jörg Ziegler and Peter J Facchini. , Alkaloid biosynthesis: metabolism and trafficking. , *Annu. Rev. Plant Biol.*, Vol. 59, pp. 735–769, 2008.
- [11] Toni M Kutchan. , Alkaloid biosynthesis [mdash] the basis for metabolic engineering of medicinal plants. , *The plant cell*, Vol. 7, No. 7, p. 1059, 1995.
- [12] Fumihiko Sato, Takashi Hashimoto, Akira Hachiya, Ken-ichi Tamura, Kum-Boo Choi, Takashi Morishige, Hideki Fujimoto, and Yasuyuki Yamada. , Metabolic engineering of plant alkaloid biosynthesis. , *Proceedings of the National Academy of Sciences*, Vol. 98, No. 1, pp. 367–372, 2001.
- [13] Vincenzo De Luca and Benoit St Pierre. , The cell and developmental biology of alkaloid biosynthesis. , *Trends in plant science*, Vol. 5, No. 4, pp. 168–173, 2000.
- [14] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. , Kegg: new perspectives on genomes, pathways, diseases and drugs. , *Nucleic acids research*, Vol. 45, No. D1, pp. D353–D361, 2017.
- [15] Tadeusz Aniszewski. , *Alkaloids: Chemistry, Biology, Ecology, and Appli-*

- cations.* , Elsevier, 2015.
- [16] Paul M Dewick. , The shikimate pathway: aromatic amino acids and phenylpropanoids. , *Medicinal Natural Products: A Biosynthetic Approach, 3rd Edition*, pp. 137–186, 2009.
- [17] Sheng Huang, Jioji Tabudravu, Somayah S Elsayed, Jeanne Travert, Doe Peace, Ming Him Tong, Kwaku Kyeremeh, Sharon M Kelly, Laurent Trembleau, Rainer Ebel, et al. , Discovery of a single monooxygenase that catalyzes carbamate formation and ring contraction in the biosynthesis of the legonmycins. , *Angewandte Chemie International Edition*, Vol. 54, No. 43, pp. 12697–12701, 2015.
- [18] Jianqing Ruan, Mengbi Yang, Peter Fu, Yang Ye, and Ge Lin. , Metabolic activation of pyrrolizidine alkaloids: insights into the structural and enzymatic basis. , *Chemical research in toxicology*, Vol. 27, No. 6, pp. 1030–1039, 2014.
- [19] Jieyu Dai, Fan Zhang, and Jiang Zheng. , Retrorsine, but not monocrotaline, is a mechanism-based inactivator of p450 3a4. , *Chemico-biological interactions*, Vol. 183, No. 1, pp. 49–56, 2010.
- [20] Juan Pan, Minakshi Bhardwaj, Padmaja Nagabhyru, Robert B Grossman, and Christopher L Schardl. , Enzymes from fungal and plant origin required for chemical diversification of insecticidal loline alkaloids in grass-epichloë symbiota. , *PLoS one*, Vol. 9, No. 12, p. e115590, 2014.
- [21] Juan Pan, Minakshi Bhardwaj, Jerome R Faulkner, Padmaja Nagabhyru, Nikki D Charlton, Richard M Higashi, Anne-Frances Miller, Carolyn A Young, Robert B Grossman, and Christopher L Schardl. , Ether bridge formation in loline alkaloid biosynthesis. , *Phytochemistry*, Vol. 98, pp. 60–68, 2014.
- [22] Neill Kim, Olga Estrada, Benjamin Chavez, Charles Stewart, John CD’Auria. , Tropane and granatane alkaloid biosynthesis: a systematic analysis. , *Molecules*, Vol. 21, No. 11, p. 1510, 2016.
- [23] Ralph E Dewey and Jiahua Xie. , Molecular genetics of alkaloid biosynthe-

- sis in nicotiana tabacum. , *Phytochemistry*, Vol. 94, pp. 10–27, 2013.
- [24] Bin Cai, Balazs Siminszky, Joseph Chappell, Ralph E Dewey, and Lowell P Bush. , Enantioselective demethylation of nicotine as a mechanism for variable nornicotine composition in tobacco leaf. , *Journal of Biological Chemistry*, Vol. 287, No. 51, pp. 42804–42811, 2012.
- [25] Katarzyna M Romek, Gérald S Remaud, Virginie Silvestre, Piotr Paneth, and Richard J Robins. , Non-statistical ¹³C fractionation distinguishes coincident and divergent steps in the biosynthesis of the alkaloids nicotine and tropine. , *Journal of Biological Chemistry*, Vol. 291, No. 32, pp. 16620–16629, 2016.
- [26] Dongsheng Fan, Guo-Yuan Zhu, Ting Li, Zhi-Hong Jiang, and Li-Ping Bai. , Dimacrolide sesquiterpene pyridine alkaloids from the stems of tripterygium regelii. , *Molecules*, Vol. 21, No. 9, p. 1146, 2016.
- [27] Dengqun Liao, Pengfei Wang, Peng Sun Chan Jia, Jianjun Qi, Lili Zhou, Xian’ enLi. , Identification and developmental expression profiling of putative alkaloid biosynthetic genes in corydalis yanhusuo bulbs. , *Scientific reports*, Vol. 6, , 2016.
- [28] David L Burgoyne, Thomas K Hemscheidt, Richard E Moore, and Maria TC Runnegar. , Biosynthesis of cylindrospermopsin. , *The Journal of organic chemistry*, Vol. 65, No. 1, pp. 152–156, 2000.
- [29] Carolyn A Young, Christopher L Schardl, Daniel G Panaccione, Simona Florea, Johanna E Takach, Nikki D Charlton, Neil Moore, Jennifer S Webb, and Jolanta Jaromczyk. , Genetics, genomics and evolution of ergot alkaloid diversity. , *Toxins*, Vol. 7, No. 4, pp. 1273–1302, 2015.
- [30] Sarah L Robinson and Daniel G Panaccione. , Diversification of ergot alkaloids in natural and modified fungi. , *Toxins*, Vol. 7, No. 1, pp. 201–218, 2015.
- [31] Prashanthi Mulinti, Simona Florea, Christopher L Schardl, and Daniel G Panaccione. , Modulation of ergot alkaloids in a grass–endophyte symbiosis by alteration of mrna concentrations of an ergot alkaloid synthesis gene. ,

- Journal of agricultural and food chemistry*, Vol. 64, No. 24, pp. 4982–4989, 2016.
- [32] Yulia Bilovol and Daniel G Panaccione. , Functional analysis of the gene controlling hydroxylation of festuclavine in the ergot alkaloid pathway of *Neosartorya fumigata*. , *Current genetics*, Vol. 62, No. 4, pp. 853–860, 2016.
- [33] Priyanka Paul, Sanjay K Singh, Barunava Patra, Xueyi Sui, Sitakanta Patanaik, and Ling Yuan. , A differentially regulated *ap2/erf* transcription factor gene cluster acts downstream of a map kinase cascade to modulate terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. , *New Phytologist*, Vol. 213, No. 3, pp. 1107–1123, 2017.
- [34] Claire Parage, Emilien Foureau, Franziska Kellner, Vincent Burlat, Samira Mahroug, Arnaud Lanoue, Thomas Dugé de Bernonville, Monica Arias Londono, Inês Carqueijeiro, Audrey Oudin, et al. , Class ii cytochrome p450 reductase governs the biosynthesis of alkaloids. , *Plant physiology*, Vol. 172, No. 3, pp. 1563–1577, 2016.
- [35] Shiv S Pandey, Sucheta Singh, CS Vivek Babu, Karuna Shanker, NK Srivastava, Ashutosh K Shukla, and Alok Kalra. , Fungal endophytes of *Catharanthus roseus* enhance vindoline content by modulating structural and regulatory genes related to terpenoid indole alkaloid biosynthesis. , *Scientific reports*, Vol. 6, p. 26583, 2016.
- [36] Jianhua Zhu, Mingxuan Wang, Wei Wen, and Rongmin Yu. , Biosynthesis and regulation of terpenoid indole alkaloids in *Catharanthus roseus*. , *Pharmacognosy reviews*, Vol. 9, No. 17, p. 24, 2015.
- [37] Shu-Ming Li. , Evolution of aromatic prenyltransferases in the biosynthesis of indole derivatives. , *Phytochemistry*, Vol. 70, No. 15, pp. 1746–1757, 2009.
- [38] Hsiao-Ching Lin, Grace Chiou, Yit-Heng Chooi, Travis C McMahon, Wei Xu, Neil K Garg, and Yi Tang. , Elucidation of the concise biosynthetic pathway of the communesin indole alkaloids. , *Angewandte Chemie International Edition*, Vol. 54, No. 10, pp. 3004–3007, 2015.

- [39] Takahiro Mori, Shotaro Hoshino, Shusaku Sahashi, Toshiyuki Wakimoto, Takashi Matsui, Hiroyuki Morita, and Ikuro Abe. , Structural basis for β -carboline alkaloid production by the microbial homodimeric enzyme mcbb. , *Chemistry & biology*, Vol. 22, No. 7, pp. 898–906, 2015.
- [40] Martin E Tanner. , Increasing the diversity of known pictet-spenglerases. , *Chemistry & biology*, Vol. 22, No. 7, pp. 806–807, 2015.
- [41] Xinyu Liu and Christopher T Walsh. , Characterization of cyclo-acetoacetyl-l-tryptophan dimethylallyltransferase in cyclopiazonic acid biosynthesis: substrate promiscuity and site directed mutagenesis studies. , *Biochemistry*, Vol. 48, No. 46, pp. 11032–11044, 2009.
- [42] Xin Zhen, Ting Gong, Fu Liu, Pei-Cheng Zhang, Wan-Qi Zhou, Yan Li, and Ping Zhu. , A new analogue of echinomycin and a new cyclic dipeptide from a marine-derived streptomyces sp. ls298. , *Marine drugs*, Vol. 13, No. 11, pp. 6947–6961, 2015.
- [43] Marco I Ries, Hazrat Ali, Peter P Lankhorst, Thomas Hankemeier, Roel AL Bovenberg, Arnold JM Driessen, and Rob J Vreeken. , Novel key metabolites reveal further branching of the roquefortine/meleagrin biosynthetic pathway. , *Journal of Biological Chemistry*, Vol. 288, No. 52, pp. 37289–37295, 2013.
- [44] Fernando Gandía-Herrero and Francisco García-Carmona. , Biosynthesis of betalains: yellow and violet plant pigments. , *Trends in plant science*, Vol. 18, No. 6, pp. 334–343, 2013.
- [45] Kathy E Schwinn. , The dope on l-dopa formation for betalain pigments. , *New Phytologist*, Vol. 210, No. 1, pp. 6–9, 2016.
- [46] Mohammad Imtiyaj Khan and P Giridhar. , Plant betalains: chemistry and biochemistry. , *Phytochemistry*, Vol. 117, pp. 267–295, 2015.
- [47] N Bhagya and KR Chandrashekar. , Tetrandrine—a molecule of wide bioactivity. , *Phytochemistry*, Vol. 125, pp. 5–13, 2016.
- [48] Limei Chang, Jillian M Hagel, and Peter J Facchini. , Isolation and characterization of o-methyltransferases involved in the biosynthesis of glaucine

- in glaucium flavum. , *Plant physiology*, Vol. 169, No. 2, pp. 1127–1140, 2015.
- [49] Jillian M Hagel, Rupasri Mandal, Beomsoo Han, Jun Han, Donald R Dinsmore, Christoph H Borchers, David S Wishart, and Peter J Facchini. , Metabolome analysis of 20 taxonomically related benzyloisoquinoline alkaloid-producing plants. , *BMC plant biology*, Vol. 15, No. 1, p. 220, 2015.
- [50] Thilo Winzer, Marcelo Kern, Andrew J King, Tony R Larson, Roxana I Teodor, Samantha L Donninger, Yi Li, Adam A Dowle, Jared Cartwright, Rachel Bates, et al. , Morphinan biosynthesis in opium poppy requires a p450-oxidoreductase fusion protein. , *Science*, Vol. 349, No. 6245, pp. 309–312, 2015.
- [51] Yasuyuki Yamada, Tadashi Yoshimoto, Sayumi T Yoshida, and Fumihiko Sato. , Characterization of the promoter region of biosynthetic enzyme genes involved in berberine biosynthesis in *coptis japonica*. , *Frontiers in plant science*, Vol. 7, , 2016.
- [52] Thais Guaratini, Denise Brentan Silva, Aline Cavalli Bizaro, Lucas Rossi Sartori, Hans-Ulrich Humpf, Norberto Peporine Lopes, Letícia Veras Costa-Lotufo, and João Luis Callegari Lopes. , In vitro metabolism studies of erythraline, the major spiroalkaloid from *erythrina verna*. , *BMC complementary and alternative medicine*, Vol. 14, No. 1, p. 61, 2014.
- [53] Ulrich H Maier, Wolfgang Rödl, Brigitte Deus-Neumann, and Meinhard H Zenk. , Biosynthesis of *erythrina* alkaloids in *erythrina crista-galli*. , *Phytochemistry*, Vol. 52, No. 3, pp. 373–382, 1999.
- [54] Miguel A Torres, Elesha Hoffarth, Luiz Eugenio, Julia Savtchouk, Xue Chen, Jeremy S Morris, Peter J Facchini, and Kenneth K-S Ng. , Structural and functional studies of pavinic n-methyltransferase from *thalictrum flavum* reveal novel insights into substrate recognition and catalytic mechanism. , *Journal of Biological Chemistry*, Vol. 291, No. 45, pp. 23403–23415, 2016.
- [55] Isabel Desgagné-Penix and Peter J Facchini. , Systematic silencing of ben-

- zylisoquinoline alkaloid biosynthetic genes reveals the major route to paverine in opium poppy. , *The Plant Journal*, Vol. 72, No. 2, pp. 331–344, 2012.
- [56] Taiji Nomura and Toni M Kutchan. , Is a metabolic enzyme complex involved in the efficient and accurate control of ipecac alkaloid biosynthesis in *psychotria ipecacuanha*? , *Plant signaling & behavior*, Vol. 5, No. 7, pp. 875–877, 2010.
- [57] Taiji Nomura, Alfonso Lara Quesada, and Toni M Kutchan. , The new β -d-glucosidase in terpenoid-isoquinoline alkaloid biosynthesis in *psychotria ipecacuanha*. , *Journal of Biological Chemistry*, Vol. 283, No. 50, pp. 34650–34659, 2008.
- [58] Joshua A Baccile, Joseph E Spraker, Henry H Le, Eileen Brandenburger, Christian Gomez, Jin Woo Bok, Juliane Macheleidt, Axel A Brakhage, Dirk Hoffmeister, Nancy P Keller, et al. , Plant-like biosynthesis of isoquinoline alkaloids in *aspergillus fumigatus*. , *Nature chemical biology*, Vol. 12, No. 6, pp. 419–424, 2016.
- [59] Somnuk Bunsupa, Mami Yamazaki, and Kazuki Saito. , Quinolizidine alkaloid biosynthesis: recent advances and future prospects. , *Frontiers in plant science*, Vol. 3, , 2012.
- [60] Somnuk Bunsupa, Kousuke Hanada, Akira Maruyama, Kaori Aoyagi, Kana Komatsu, Hideki Ueno, Madoka Yamashita, Ryosuke Sasaki, Akira Oikawa, Kazuki Saito, et al. , Molecular evolution and functional characterization of a bifunctional decarboxylase involved in lycopodium alkaloid biosynthesis. , *Plant physiology*, Vol. 171, No. 4, pp. 2432–2444, 2016.
- [61] Hannu Hotti, Tuulikki Seppänen-Laakso, Mikko Arvas, Teemu H Teeri, and Heiko Rischer. , Polyketide synthases from poison hemlock (*conium maculatum* l.). , *The FEBS journal*, Vol. 282, No. 21, pp. 4141–4156, 2015.
- [62] Noriyasu Ishikawa, Hidenori Tanaka, Fumi Koyama, Hiroshi Noguchi, Clay CC Wang, Kinya Hotta, and Kenji Watanabe. , Non-heme dioxygenase catalyzes atypical oxidations of 6, 7-bicyclic systems to form the 6,

- 6-quinolone core of viridicatin-type fungal alkaloids. , *Angewandte Chemie*, Vol. 126, No. 47, pp. 13094–13098, 2014.
- [63] Yohei Katsuyama, Kirsten Harmrolfs, Dominik Pistorius, Yanyan Li, and Rolf Müller. , A semipinacol rearrangement directed by an enzymatic system featuring dual-function fad-dependent monooxygenase. , *Angewandte Chemie International Edition*, Vol. 51, No. 37, pp. 9437–9440, 2012.
- [64] Yi Zou, Zhajun Zhan, Dehai Li, Mancheng Tang, Ralph A Cacho, Kenji Watanabe, and Yi Tang. , Tandem prenyltransferases catalyze isoprenoid elongation and complexity generation in biosynthesis of quinolone alkaloids. , *Journal of the American Chemical Society*, Vol. 137, No. 15, pp. 4980–4983, 2015.
- [65] Stuart W Haynes, Xue Gao, Yi Tang, and Christopher T Walsh. , Assembly of asperlicin peptidyl alkaloids from anthranilate and tryptophan: a two-enzyme pathway generates heptacyclic scaffold complexity in asperlicin e. , *Journal of the American Chemical Society*, Vol. 134, No. 42, pp. 17444–17447, 2012.
- [66] Stuart W Haynes, Xue Gao, Yi Tang, and Christopher T Walsh. , Complexity generation in fungal peptidyl alkaloid biosynthesis: a two-enzyme pathway to the hexacyclic mdr export pump inhibitor ardeemin. , *ACS chemical biology*, Vol. 8, No. 4, pp. 741–748, 2013.
- [67] Ikuro Abe, Tsuyoshi Abe, Kiyofumi Wanibuchi, and Hiroshi Noguchi. , Enzymatic formation of quinolone alkaloids by a plant type iii polyketide synthase. , *Organic letters*, Vol. 8, No. 26, pp. 6063–6065, 2006.
- [68] Joseph M Jez, Marianne E Bowman, and Joseph P Noel. , Expanding the biosynthetic repertoire of plant type iii polyketide synthases by altering starter molecule specificity. , *Proceedings of the National Academy of Sciences*, Vol. 99, No. 8, pp. 5319–5324, 2002.
- [69] Matthew J Nicholson, Carla J Eaton, Cornelia Stärkel, Brian A Tapper, Murray P Cox, and Barry Scott. , Molecular cloning and functional analysis of gene clusters for the biosynthesis of indole-diterpenes in penicillium

- crustosum and p. janthinellum. , *Toxins*, Vol. 7, No. 8, pp. 2701–2722, 2015.
- [70] Koichi Tagami, Chengwei Liu, Atsushi Minami, Motoyoshi Noike, Tetsuya Isaka, Shuhei Fueki, Yoshihiro Shichijo, Hiroaki Toshima, Katsuya Gomi, Tohru Dairi, et al. , Reconstitution of biosynthetic machinery for indole-diterpene paxilline in aspergillus oryzae. , *Journal of the American Chemical Society*, Vol. 135, No. 4, pp. 1260–1263, 2013.
- [71] Takayuki Motoyama, Toshiaki Hayashi, Hiroshi Hirota, Masashi Ueki, and Hiroyuki Osada. , Terpendole e, a kinesin eg5 inhibitor, is a key biosynthetic intermediate of indole-diterpenes in the producing fungus chaunopycnis alba. , *Chemistry & biology*, Vol. 19, No. 12, pp. 1611–1619, 2012.
- [72] Sanjay Saikia, Emily J Parker, Albert Koulman, and Barry Scott. , Defining paxilline biosynthesis in penicillium paxilli functional characterization of two cytochrome p450 monooxygenases. , *Journal of Biological Chemistry*, Vol. 282, No. 23, pp. 16829–16837, 2007.
- [73] Matthew J Nicholson, Albert Koulman, Brendon J Monahan, Beth L Pritchard, Gary A Payne, and Barry Scott. , Identification of two aflatrems biosynthesis gene loci in aspergillus flavus and metabolic engineering of penicillium paxilli to elucidate their function. , *Applied and environmental microbiology*, Vol. 75, No. 23, pp. 7469–7481, 2009.
- [74] Sanjay Saikia, Emily J Parker, Albert Koulman, and Barry Scott. , Four gene products are required for the fungal synthesis of the indole-diterpene, paspaline. , *FEBS letters*, Vol. 580, No. 6, pp. 1625–1630, 2006.
- [75] Carolyn A Young, Silvina Felitti, Katherine Shields, German Spangenberg, Richard D Johnson, Gregory T Bryan, Sanjay Saikia, and Barry Scott. , A complex gene cluster for indole-diterpene biosynthesis in the grass endophyte neotyphodium lolii. , *Fungal Genetics and Biology*, Vol. 43, No. 10, pp. 679–693, 2006.
- [76] Carolyn A Young, Brian A Tapper, Kimberley May, Christina D Moon, Christopher L Schardl, and Barry Scott. , Indole-diterpene biosynthetic

- capability of epichloë endophytes as predicted by ltm gene analysis. , *Applied and environmental microbiology*, Vol. 75, No. 7, pp. 2200–2211, 2009.
- [77] M Itkin, U Heinig, O Tzfadia, AJ Bhide, B Shinde, PD Cardenas, SE Bobcoba, T Unger, S Malitsky, R Finkers, et al. , Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. , *Science*, Vol. 341, No. 6142, pp. 175–179, 2013.
- [78] Pablo D Cárdenas, Prashant D Sonawane, Jacob Pollier, Robin Vanden Bossche, Veena Dewangan, Efrat Weithorn, Lior Tal, Sagit Meir, Ilana Rogachev, Sergey Malitsky, et al. , Game9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. , *Nature communications*, Vol. 7, p. 10654, 2016.
- [79] Hiroshi Ashihara, Hiroshi Sano, and Alan Crozier. , Caffeine and related purine alkaloids: biosynthesis, catabolism, function and genetic engineering. , *Phytochemistry*, Vol. 69, No. 4, pp. 841–856, 2008.
- [80] Ana Luisa Anaya, Rocio Cruz-Ortega, and George R Waller. , Metabolism and ecology of purine alkaloids. , *Front Biosci*, Vol. 11, No. Suppl 1, p. 2354, 2006.
- [81] Ana Paula Santos and Paulo Roberto H Moreno. , Alkaloids derived from histidine: Imidazole (pilocarpine, pilosine). , In *Natural Products*, pp. 861–882. Springer, 2013.
- [82] Christopher T Walsh, Stuart W Haynes, Brian D Ames, Xue Gao, and Yi Tang. , Short pathways to complexity generation: fungal peptidyl alkaloid multicyclic scaffolds from anthranilate building blocks, 2013.
- [83] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. , Pubchem substance and compound databases. , *Nucleic acids research*, Vol. 44, No. D1, pp. D1202–D1213, 2015.
- [84] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. , Reoptimization of mdl keys for use in drug discovery. , *Journal of chemical information and computer sciences*, Vol. 42, No. 6, pp. 1273–1280, 2002.

- [85] Mahendra Awale and Jean-Louis Reymond. , Atom pair 2d-fingerprints perceive 3d-molecular shape and pharmacophores for very fast virtual screening of zinc and gdb-17. , *Journal of chemical information and modeling*, Vol. 54, No. 7, pp. 1892–1907, 2014.
- [86] Guy W Bemis and Mark A Murcko. , The properties of known drugs. 1. molecular frameworks. , *Journal of medicinal chemistry*, Vol. 39, No. 15, pp. 2887–2893, 1996.
- [87] Stephen R Heller and Alan D McNaught. , The iupac international chemical identifier (inchi). , *Chemistry International*, Vol. 31, No. 1, p. 7, 2009.
- [88] David Weininger, Arthur Weininger, and Joseph L Weininger. , Smiles. 2. algorithm for generation of unique smiles notation. , *Journal of Chemical Information and Computer Sciences*, Vol. 29, No. 2, pp. 97–101, 1989.
- [89] Shin-ichi Minato. , Zero-suppressed bdds for set manipulation in combinatorial problems. , In *Proceedings of the 30th international Design Automation Conference*, pp. 272–277. ACM, 1993.
- [90] Jun Kawahara, Takeru Inoue, Hiroaki Iwashita, and Shin-ichi Minato. , Tcs technical report. , 2011.
- [91] Jun Kawahara, Takeru Inoue, Hiroaki Iwashita, and Shin-ichi Minato. , Frontier-based search for enumerating all constrained subgraphs with compressed representation. , *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. 100, No. 9, pp. 1773–1784, 2017.
- [92] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. , Deep learning. , *Nature*, Vol. 521, No. 7553, pp. 436–444, 2015.
- [93] Li Deng, Dong Yu, et al. , Deep learning: methods and applications. , *Foundations and Trends® in Signal Processing*, Vol. 7, No. 3–4, pp. 197–387, 2014.
- [94] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. , Deep learning via semi-supervised embedding. , In *Neural Networks: Tricks of the Trade*, pp. 639–655. Springer, 2012.

- [95] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. , ImageNet Large Scale Visual Recognition Challenge. , *International Journal of Computer Vision (IJCV)*, Vol. 115, No. 3, pp. 211–252, 2015.
- [96] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. , Neural message passing for quantum chemistry. , *arXiv preprint arXiv:1704.01212*, 2017.
- [97] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. , Modeling relational data with graph convolutional networks. , *arXiv preprint arXiv:1703.06103*, 2017.
- [98] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. , Molecular graph convolutions: moving beyond fingerprints. , *Journal of computer-aided molecular design*, Vol. 30, No. 8, pp. 595–608, 2016.
- [99] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. , Low data drug discovery with one-shot learning. , *ACS central science*, Vol. 3, No. 4, pp. 283–293, 2017.

付録

Compound Name	Molecular formula
(-)-Hygrine	C8H15NO
(-)-Hyoscyne (Scopolamine)	C17H21NO4
(-)-Hyoscyamine	C17H23NO3
(+)-Hygrine	C8H15NO
3,6-dihydronicotinic acid	C6H7NO2
6beta-Hydroxyhyoscyamine (Anisodamine)	C17H23NO4
Aantoxin- a	C10H15NO
Agmatine	C5H14N4
alpha-Iminosuccinic acid (NAMN)	C4H5NO4
Anatabine	C10H12N2
Calystegine A3	C7H13NO3
Cocaine	C17H21NO4
Cuscohygrine	C13H24N2O
Cyclidrospermopsin	C15H21N5O7S
Edulnic acid	C11H11NO4
Evonine	C36H43NO17
Evoninic acid	C11H13NO4
exo-1-Acetamido-pyrrolizidine	C9H16N2O
exo-1-Amino-pyrrolizidine	C7H14N2
Guanidinoacetate	C3H7N3O2
Homoantoxin-a	C11H17NO
Homospermidine	C8H21N3
L-Arg	C6H14N4O2
L-Asp	C4H8N2O3

Compound Name	Molecular formula
Legonindolizidine A	C ₁₄ H ₂₀ N ₂ O ₃
Legonmycin A	C ₁₃ H ₂₀ N ₂ O ₃
Legonmycin B	C ₁₄ H ₂₂ N ₂ O ₃
L-Gly	C ₂ H ₅ NO ₂
Littorine	C ₁₇ H ₂₃ NO ₃
Legonindolizidine B	C ₁₅ H ₂₂ N ₂ O ₃
Loline	C ₈ H ₁₄ N ₂ O
L-Ornithine	C ₅ H ₁₂ N ₂ O ₂
L-Phe	C ₉ H ₁₁ NO ₂
L-Pro	C ₅ H ₉ NO ₂
L-Thr	C ₄ H ₉ NO ₃
Malonyl-CoA	C ₂₄ H ₃₈ N ₇ O ₁₉ P ₃ S
Methylecgonine	C ₁₀ H ₁₇ NO ₃
Methylecgonone	C ₁₀ H ₁₇ NO ₃
Myosmine	C ₉ H ₁₀ N ₂
N-Acetyllooline	C ₁₀ H ₁₆ N ₂ O ₂
N-Acetylnorlooline	C ₉ H ₁₄ N ₂ O ₂
N-carbamoylputrescine	C ₅ H ₁₃ N ₃ O
N-Formyllooline	C ₉ H ₁₄ N ₂ O ₂
Nicotine	C ₁₀ H ₁₄ N ₂
Nicotinic acid	C ₆ H ₅ NO ₂
N-Methyllooline	C ₉ H ₁₄ N ₂ O ₂
N-Methylputrescine	C ₅ H ₁₄ N ₂
N-Methyl- Δ 1-pyrrolinium cation	C ₅ H ₁₀ N ⁺
Norlooline	C ₇ H ₁₂ N ₂ O
Nornicotine	C ₉ H ₁₂ N ₂
O-Acetyl homoserine	C ₆ H ₁₁ NO ₄
Phenyllactic acid	C ₉ H ₁₀ O ₃

Compound Name	Molecular formula
Phenylpyruvate	C ₉ H ₈ O ₃
Pseudotropine	C ₈ H ₁₅ NO
Putrescine	C ₄ H ₁₂ N ₂
Pyrroline-5-carboxylate	C ₅ H ₇ NO ₂
Quinolinic acid	C ₇ H ₅ NO ₄
Retronecine	C ₈ H ₁₃ NO ₂
Saxitoxin	C ₁₀ H ₁₇ N ₇ O ₄
Senecionine	C ₁₈ H ₂₅ NO ₅
Senecionine N-oxide	C ₁₈ H ₂₅ NO ₆
Tropine	C ₈ H ₁₅ NO
Tropinone	C ₈ H ₁₃ NO
Urea	CH ₄ N ₂ O
Wilfordic acid	C ₁₁ H ₁₃ NO ₄
IPP	C ₅ H ₁₂ O ₇ P ₂
(8R,9S)-Fumigaciavine A	C ₁₈ H ₂₂ N ₂ O ₂
(8R,9S)-Fumigaciavine B	C ₁₆ H ₂₀ N ₂ O
(8S,9S)-Fumigaclavine A	C ₁₈ H ₂₂ N ₂ O ₂
(8S,9S)-Fumigaclavine B	C ₁₆ H ₂₀ N ₂ O
(8S,9S)-Fumigaclavine C	C ₂₃ H ₃₀ N ₂ O ₂
4-Methylallyl-L-abrine;4-DMAL-L-abnine	C ₁₇ H ₂₂ N ₂ O ₂
Agroclavine	C ₁₆ H ₁₈ N ₂
Chabickavube-1	C ₁₆ H ₂₀ N ₂ O
Chabickavube-1 aldehyde	C ₁₆ H ₁₈ N ₂ O
Dihydroergosine	C ₃₀ H ₃₉ N ₅ O ₅
Dihydrolysergic acid	C ₁₆ H ₁₈ N ₂ O ₂
Dihydrolysergol	C ₁₆ H ₁₈ N ₂ O
Dimethyl phenylpiperazinium (DMPP)	C ₁₂ H ₁₉ N ₂
Dimethylallyl tryptophan (DMATrp)	C ₁₆ H ₂₀ N ₂ O ₂

Compound Name	Molecular formula
D-Lysergic acid	C ₁₆ H ₁₆ N ₂ O ₂
Elymoclavine	C ₁₆ H ₁₈ N ₂ O
Ergometrine	C ₁₉ H ₂₃ N ₃ O ₂
Ergotamine	C ₃₃ H ₃₅ N ₅ O ₅
Festuclavine	C ₁₆ H ₂₀ N ₂
L-Ala	C ₃ H ₇ N ₂ O ₂
L-Pro	C ₅ H ₉ N ₂ O
L-Trp	C ₁₁ H ₁₂ N ₂ O ₂
L-Val	C ₅ H ₁₁ N ₂ O ₂
Paspalic acid	C ₁₆ H ₁₆ N ₂ O ₂
Pyroclavine	C ₁₆ H ₂₀ N ₂
Setoclavine	C ₁₆ H ₁₈ N ₂ O
16-Hydroxytabersonine	C ₂₂ H ₂₆ N ₂ O ₃
16-Meethoxy-2,3-dihydro-3-hydroxytabersonine	C ₂₂ H ₂₈ N ₂ O ₄
16-Methoxytabersonine	C ₂₂ H ₂₆ N ₂ O ₃
19-Oacetyl-tabersonine	C ₂₃ H ₂₆ N ₂ O ₄
3',4'-Anhydrovinblastine	C ₄₆ H ₅₆ N ₄ O ₈
4,21-Dehydrogeissoschizine	C ₂₁ H ₂₃ N ₂ O ₃ +
Acetylnorajmaline	C ₂₁ H ₂₆ N ₂ O ₃
Ajmalicine	C ₂₁ H ₂₄ N ₂ O ₃
Ajmaline	C ₂₀ H ₂₆ N ₂ O ₂
Akuammicine	C ₂₀ H ₂₂ N ₂ O ₂
Camptothecin	C ₂₀ H ₁₆ N ₂ O ₄
Cathenamine	C ₂₁ H ₂₂ N ₂ O ₃
Cinchonidine	C ₁₉ H ₂₂ N ₂ O
Cinchonine	C ₁₉ H ₂₂ N ₂ O
Cinchoninone	C ₁₉ H ₂₀ N ₂ O
Condylocarpine	C ₂₀ H ₂₂ N ₂ O ₂

Compound Name	Molecular formula
Corynantheal	C ₁₉ H ₂₂ N ₂ O
Deacetylvindoline	C ₂₃ H ₃₀ N ₂ O ₅
Desacetoxyvindoline	C ₂₃ H ₃₀ N ₂ O ₄
Epicathenamine	C ₂₁ H ₂₂ N ₂ O ₃
16-Epivellosimine	C ₁₉ H ₂₀ N ₂ O
Ervaoffines D	C ₂₀ H ₂₄ N ₂ O ₄
Geranyl diphosphate	C ₁₀ H ₂₀ O ₇ P ₂
Ibogaine	C ₂₀ H ₂₆ N ₂ O
Ibogaine hydroxyindolenine	C ₂₀ H ₂₆ N ₂ O ₂
Ibogamine	C ₁₉ H ₂₄ N ₂
Iboluteine	C ₂₀ H ₂₆ N ₂ O ₂
L-Trp	C ₁₁ H ₁₂ N ₂ O ₂
Manilamine	C ₁₈ H ₂₂ N ₂ O
Polyneuridine aldehyde	C ₂₁ H ₂₂ N ₂ O ₃
Preaquamycin	
Pumiloside	C ₂₆ H ₂₈ N ₂ O ₉
Quinidine	C ₂₀ H ₂₄ N ₂ O ₂
Quinine	C ₂₀ H ₂₄ N ₂ O ₂
Raucaffricine	C ₂₇ H ₃₂ N ₂ O ₈
Secologanin	C ₁₇ H ₂₄ O ₁₀
Stemmadenine	C ₂₁ H ₂₆ N ₂ O ₃
Strictosamide	C ₂₆ H ₃₀ N ₂ O ₈
Strictosidine	C ₂₇ H ₃₄ N ₂ O ₉
Strychnine	C ₂₁ H ₂₂ N ₂ O ₂
Tabersonine	C ₂₁ H ₂₄ N ₂ O ₂
Tabertingine	C ₁₉ H ₂₀ N ₂ O
Tryptamine	C ₁₀ H ₁₂ N ₂
Vinblastine	C ₄₆ H ₅₈ N ₄ O ₉

Compound Name	Molecular formula
Vindoline	C ₂₅ H ₃₂ N ₂ O ₆
Vineristine	C ₄₆ H ₅₆ N ₄ O ₁₀
Vinorine	C ₂₁ H ₂₂ N ₂ O ₂
Vomilenine	C ₂₁ H ₂₂ N ₂ O ₃
(-)-Aurantioclavine	C ₁₅ H ₁₈ N ₂
12,13-Dihydroxyfumitremorgin C	C ₂₂ H ₂₅ N ₃ O ₅
1-Acetyl-3-carboxy-beta-carboline	C ₁₄ H ₁₀ N ₂ O ₃
1-Acetyl-beta-carboline	C ₁₃ H ₁₀ N ₂ O
20-Dihydrochaetoglobosin A	C ₃₂ H ₃₈ N ₂ O ₅
4,5-Dihydrocanthin-6-one	C ₁₄ H ₁₀ N ₂ O
4-Dimethylallyltryptophan (4-DMAT)	C ₁₆ H ₂₀ N ₂ O ₂
4-DMA-L-abrine	C ₁₇ H ₂₂ N ₂ O ₂
5-Hydroxy-L-Tryptophan	C ₁₁ H ₁₂ N ₂ O ₃
7-Dimethylallyl-L-Tryptophan	C ₁₆ H ₂₀ N ₂ O ₂
Acetyl-CoA	C ₂₃ H ₃₈ N ₇ O ₁₇ P ₃ S
alpha-Cyclopiazonic acid	C ₂₀ H ₂₀ N ₂ O ₃
Beta-Carboline	C ₁₁ H ₈ N ₂
beta-Cyclopiazonic acid	C ₂₀ H ₂₂ N ₂ O ₃
Brevianamide A	C ₂₁ H ₂₃ N ₃ O ₃
Brevianamide F	C ₁₆ H ₁₇ N ₃ O ₂
Canthin-6-one	C ₁₄ H ₈ N ₂ O
Chanoclavine-I aldehyde	C ₁₆ H ₁₈ N ₂ O
Chaetoglobosin A	C ₃₂ H ₃₆ N ₂ O ₅
Chaetoglobosin J	C ₃₂ H ₃₆ N ₂ O ₄
Chimonanthine	C ₂₂ H ₂₆ N ₄
Communesin C	C ₃₁ H ₃₄ N ₄ O ₂
Communesin E	C ₂₇ H ₃₀ N ₄ O ₂
Communesin F	C ₂₈ H ₃₂ N ₄ O

Communesin I	C26H30N4O
Communesin J	C32H36N4O
Communesin K	C26H30N4
Cyclo-acetoacetyl-L-Tryptophan	C15H13N2O3-
Cytoglobosin D	C32H38N2O4
Deoxybrevianamide E	C21H25N3O2
Didemethylasterriquinone D	C22H14N2O4
Dihydro-beta-carboline	C12H12N2
Dimethylallyl pyrophosphate (DMAPP)	C5H12O7P2
Echinomycin	C51H64N12O12S2
Elaeagnine	C12H14N2
Eserine	C15H21N3O2
Fumigaclavine C	C23H30N2O2
Fumitremorgin B	C27H33N3O5
Gramine	C11H14N2
Harmaline	C13H14N2O
Harman	C12H10N2
Harmine	C13H12N2O
Indolepyruvate	C11H9NO3
L-Ala	C3H7NO2
L-Cys	C3H7NO2S
L-Pro	C5H9NO2
L-Trp	C11H12N2O2
L-Val	C5H11NO2
Malonyl-CoA	C24H38N7O19P3S
Marinacarboline A	C23H21N3O3
Marinacarboline B	C22H19N3O3
Marinacarboline C	C22H19N3O2

Compound Name	Molecular formula
Phenylpyruvate	C ₉ H ₈ O ₃
Pseudotropine	C ₈ H ₁₅ NO
Putrescine	C ₄ H ₁₂ N ₂
Pyrroline-5-carboxylate	C ₅ H ₇ NO ₂
Quinolinic acid	C ₇ H ₅ NO ₄
Retronecine	C ₈ H ₁₃ NO ₂
Saxitoxin	C ₁₀ H ₁₇ N ₇ O ₄
Senecionine	C ₁₈ H ₂₅ NO ₅
Senecionine N-oxide	C ₁₈ H ₂₅ NO ₆
Tropine	C ₈ H ₁₅ NO
Tropinone	C ₈ H ₁₃ NO
Urea	CH ₄ N ₂ O
Wilfordic acid	C ₁₁ H ₁₃ NO ₄
IPP	C ₅ H ₁₂ O ₇ P ₂
(8R,9S)-Fumigaciavine A	C ₁₈ H ₂₂ N ₂ O ₂
(8R,9S)-Fumigaciavine B	C ₁₆ H ₂₀ N ₂ O
(8S,9S)-Fumigaclavine A	C ₁₈ H ₂₂ N ₂ O ₂
(8S,9S)-Fumigaclavine B	C ₁₆ H ₂₀ N ₂ O
(8S,9S)-Fumigaclavine C	C ₂₃ H ₃₀ N ₂ O ₂
4-Methylallyl-L-abrine;4-DMAL-L-abnine	C ₁₇ H ₂₂ N ₂ O ₂
Agroclavine	C ₁₆ H ₁₈ N ₂
Chabickavube-1	C ₁₆ H ₂₀ N ₂ O
Chabickavube-1 aldehyde	C ₁₆ H ₁₈ N ₂ O
Dihydroergosine	C ₃₀ H ₃₉ N ₅ O ₅
Dihydrolysergic acid	C ₁₆ H ₁₈ N ₂ O ₂
Dihydrolysergol	C ₁₆ H ₁₈ N ₂ O
Dimethyl phenylpiperazinium (DMPP)	C ₁₂ H ₁₉ N ₂
Dimethylallyl tryptophan (DMATrp)	C ₁₆ H ₂₀ N ₂ O ₂

Compound Name	Molecular formula
Marinacarboline D	C ₂₄ H ₂₀ N ₄ O ₂
Methyltryptamine	C ₁₁ H ₁₄ N ₂
Prochaetoglobosin I	C ₃₂ H ₃₈ N ₂ O ₂
Prochaetoglobosin IV	C ₃₂ H ₃₈ N ₂ O ₃
Psilocin	C ₁₂ H ₁₆ N ₂ O
Psilocybin	C ₁₂ H ₁₇ N ₂ O ₄ P
Quinoxaline-2-carboxylic acid	C ₉ H ₆ N ₂ O ₂
Serotonin	C ₁₀ H ₁₂ N ₂ O
Terrequinone A	C ₃₂ H ₃₀ N ₂ O ₃
Tetrahydroamine	C ₁₃ H ₁₆ N ₂ O
Triostin A	C ₅₀ H ₆₂ N ₁₂ O ₁₂ S ₂
Tryprostatin B	C ₂₁ H ₂₅ N ₃ O ₂
Tryptamine	C ₁₀ H ₁₂ N ₂
Brevianamide F	C ₁₆ H ₁₇ N ₃ O ₂
Demetoxyfumtremorgin C	C ₂₁ H ₂₃ N ₃ O ₂
Deoxybrevianamide E	C ₂₁ H ₂₅ N ₃ O ₂
Fumtremorgin C	C ₂₂ H ₂₅ N ₃ O ₃
Histidyl-tryptophanyl-diketopiperazine (HTD)	C ₁₇ H ₁₇ N ₅ O ₂
L-His	C ₆ H ₉ N ₃ O ₂
L-Trp	C ₁₁ H ₁₂ N ₂ O ₂
Meleagrins	C ₂₃ H ₂₃ N ₅ O ₄
Notoamide A	C ₂₆ H ₂₉ N ₃ O ₅
Notoamide B	C ₂₆ H ₂₉ N ₃ O ₄
Notoamide S	C ₂₆ H ₃₃ N ₃ O ₃
Roquefortine C	C ₂₂ H ₂₃ N ₅ O ₂
Spirotryprostatin B	C ₂₁ H ₂₁ N ₃ O ₃
Spirotryprostatin G	C ₂₂ H ₂₃ N ₃ O ₄
Tryprostatin B	C ₂₁ H ₂₅ N ₃ O ₂

Compound Name	Molecular formula
Verruculogen	C ₂₇ H ₃₃ N ₃ O ₇
L-Tyr	C ₉ H ₁₁ N ₁ O ₃
Betalamic acid	C ₉ H ₉ N ₁ O ₅
(S)-Autumnaline	C ₂₁ H ₂₇ N ₁ O ₅
(S)-Isoandrocymbine	C ₂₁ H ₂₆ N ₁ O ₅ +
Adrenaline	C ₉ H ₁₃ N ₁ O ₃
Betanidin	C ₁₈ H ₁₆ N ₂ O ₈
Betanidin-quinone	C ₁₈ H ₁₂ N ₂ O ₈ -2
Betanin	C ₂₄ H ₂₆ N ₂ O ₁₃
Colchicine	C ₂₂ H ₂₅ N ₁ O ₆
Deacetylcolchicine	C ₂₀ H ₂₃ N ₁ O ₅
Demecolcine	C ₂₁ H ₂₅ N ₁ O ₅
Dihydroxyphenylalanine (DOPA)	C ₉ H ₁₁ N ₁ O ₄
DOPA-betaxanthin (Dopaxanthin)	C ₁₈ H ₁₈ N ₂ O ₈
Dopamine	C ₈ H ₁₁ N ₁ O ₂
Dopaxanthin-quinone	C ₁₈ H ₁₆ N ₂ O ₈
Hordenine	C ₁₀ H ₁₅ N ₁ O
Leuko-DOPA-chrome (cyclo-DOPA)	C ₉ H ₉ N ₁ O ₄
Leuko-DOPA-chrome-glucoside;Cyclo-DOPA-chrome-glucoside	C ₁₅ H ₁₈ O ₈
L-Phe	C ₉ H ₁₁ N ₁ O ₂
Mescaline	C ₁₁ H ₁₇ N ₁ O ₃
N-Formyldemecolcine	C ₂₂ H ₂₅ N ₁ O ₆
Noradrenaline	C ₈ H ₁₁ N ₁ O ₃
o-DOPAquinone	C ₉ H ₉ N ₁ O ₄
O-Methylandrocymbine	C ₂₂ H ₂₇ N ₁ O ₅
P-Coumaraldehyde	C ₉ H ₈ O ₂
Tyramine	C ₈ H ₁₁ N ₁ O
Tyrosine-betaxanthin (Portulacaxanthin II)	C ₁₈ H ₁₈ N ₂ O ₇

Compound Name	Molecular formula
(+/-)-N,N-Dimethylpavine	C22H28NO4+
(+/-)-N-Methylpavine	C21H25NO4
(+/-)-Pavine	C20H23NO4
(13S,14R)-1,13-Dihydroxy-N-methylcanadine	C21H24NO6+
(13S,14R)-1-Hydroxy-13-O-acetyl-N-methylcanadine	C23H26NO7+
(R)-Reticuline	C19H23NO4
(S)-1-Hydroxy-N-methylcanadine	C21H24NO5+
(S)-3'-Hydroxy-N-methylcoclaurine	C18H21NO4
(S)-Canadine	C20H21NO4
(S)-Cheilanthifoline	C19H19NO4
(S)-cis-N-Methylcanadine	C21H24NO4
(S)-cis-N-Methylsinactine	C21H24NO4
(S)-cis-N-Methylstylopine	C20H20NO4+
(S)-Coclaurine	C17H19NO3
(S)-N-Methylcanadine	C21H24NO4
(S)-Norcoclaurine	C16H17NO3
(S)-Norreticuline	C18H21NO4
(S)-Reticuline	C19H23NO4
(S)-Scoulerine	C19H21NO4
(S)-Stylopine	C19H17NO4
(S)-Tetrahydroberberine	C20H21NO4
(S)-Tetrahydrocolumbamine	C20H23NO4
(S)-Tetrahydropalmatine	C21H25NO4
(S)-Tetrahydropapaverine	C20H25NO4
3-O-Acetylpapaveroxine	C24H27NO8
6-O-Methylnorlaudanosoline	C17H19NO4
4-Hydroxyphenylacetaldehyde (4-HPAA)	C8H8O2
4-Hydroxyphenylpyruvate (4-HPP)	C9H7O4-

Compound Name	Molecular formula
6-Hydroxyprotopine	C ₂₀ H ₁₉ N ₃ O ₆
7-O-Acetylsalutaridinol	C ₂₁ H ₂₅ N ₃ O ₅
Alloccryptopine	C ₂₁ H ₂₃ N ₃ O ₅
Berbamunine	C ₃₆ H ₄₀ N ₂ O ₆
Berberine	C ₂₀ H ₁₈ N ₃ O ₄
Codeine	C ₁₈ H ₂₁ N ₃ O ₃
Codeinone	C ₁₈ H ₁₉ N ₃ O ₃
Coptisine	C ₁₉ H ₁₄ N ₃ O ₄
Corytuberine	C ₁₉ H ₂₁ N ₃ O ₄
Cryptopine	C ₂₁ H ₂₃ N ₃ O ₅
Dihydrosanguinarine	C ₂₀ H ₁₅ N ₃ O ₄
Dopamine	C ₈ H ₁₁ N ₃ O ₂
Erythraline	C ₁₈ H ₁₉ N ₃ O ₃
Glaudine	C ₂₂ H ₂₅ N ₃ O ₆
Hydrocodone	C ₁₈ H ₂₁ N ₃ O ₃
L-Tyr	C ₉ H ₁₁ N ₃ O ₃
Magnoflorine	C ₂₀ H ₂₄ N ₃ O ₄
Morphine	C ₁₇ H ₁₉ N ₃ O ₃
Morphinone	C ₁₇ H ₁₇ N ₃ O ₃
Narcotinehemiacetal	C ₂₂ H ₂₅ N ₃ O ₇
Neopinone	C ₁₈ H ₁₉ N ₃ O ₃
N-Methylporphyroxine	C ₂₁ H ₂₃ N ₃ O ₆
N-Methylcoclaurine	C ₁₈ H ₂₁ N ₃ O ₃
Noramurine	C ₁₈ H ₁₇ N ₃ O ₄
Norisosalutaridine	C ₁₈ H ₁₉ N ₃ O ₄
(S)-Norlaudanine	C ₁₉ H ₂₄ N ₃ O ₄ +
Noscapine	C ₂₂ H ₂₃ N ₃ O ₇
O-Demethylalloccryptopine	C ₂₀ H ₂₁ N ₃ O ₅

Compound Name	Molecular formula
O-Demethylcryptopine	C ₂₀ H ₂₁ NO ₅
Oripavine	C ₁₈ H ₁₉ NO ₃
Papaverine	C ₂₀ H ₂₁ NO ₄
Protopine	C ₂₀ H ₁₉ NO ₅
Rhoadine	C ₂₁ H ₂₁ NO ₆
Salutaridine	C ₁₉ H ₂₁ NO ₄
Salutaridinol	C ₁₉ H ₂₃ NO ₄
Sanguinarine	C ₂₀ H ₁₄ NO ₄
Tembetarine	C ₂₀ H ₂₆ NO ₄
Tetrandrine	C ₃₈ H ₄₂ N ₂ O ₆
Thebaine	C ₁₉ H ₂₁ NO ₃
Tyamine	C ₈ H ₁₁ NO
10,7' -O,O-Didemethylcephaeline	C ₂₆ H ₃₄ N ₂ O ₄
10-O-Demethylprotoemetine	C ₁₈ H ₂₅ NO ₃
10-O-Demethylcephaeline	C ₂₇ H ₃₆ N ₂ O ₄
3-O-Demethyl-2-O-methylalangiside	C ₂₅ H ₃₁ NO ₁₀
6-O-Methylpecoside	C ₂₈ H ₃₇ NO ₁₂
6-O-Methyl-N-deacetylpecoside	C ₂₇ H ₃₇ NO ₁₀
6-O-Methyl-N-deacetyloisopecoside	C ₂₆ H ₃₆ NO ₁₁ +
7' -O-Demethylcephaeline	C ₂₇ H ₃₆ N ₂ O ₄
7-O-Methylpecoside	C ₂₈ H ₃₇ NO ₁₂
7-O-Methyl-N-deacetylpecoside	C ₂₇ H ₃₇ NO ₁₀
9,7' -O,O-Didemethylcephaeline	C ₂₆ H ₃₄ N ₂ O ₄
9-O-Demethylprotoemetine	C ₁₈ H ₂₅ NO ₃
Alangimarckine	C ₂₉ H ₃₇ N ₃ O ₃
Alangiside	C ₂₅ H ₃₁ NO ₁₀
Cephaeline	C ₂₈ H ₃₈ N ₂ O ₄
Demethylalangiside	C ₂₄ H ₂₉ NO ₁₀

Compound Name	Molecular formula
Demethylisoalangsine	C ₂₄ H ₂₉ N ₁₀ O ₁₀
Deoxytubulosine	C ₂₉ H ₃₇ N ₃ O ₂
Dopamine	C ₈ H ₁₁ N ₂ O ₂
Emetine	C ₂₉ H ₄₀ N ₂ O ₄
Ipecoside	C ₂₇ H ₃₅ N ₁₂
Isolalangsine	C ₂₅ H ₃₁ N ₁₀ O ₁₀
L-Tyr	C ₉ H ₁₁ N ₃ O ₃
N-Deacetylipecoside	C ₂₅ H ₃₃ N ₁₁
N-Deacetylisoipecoside	C ₂₅ H ₃₃ N ₁₁
Protoemetine	C ₁₉ H ₂₇ N ₃ O ₃
Secologanin	C ₁₇ H ₂₄ O ₁₀
Tubulosine	C ₂₉ H ₃₇ N ₃ O ₃
L-Tyr	C ₉ H ₁₁ N ₃ O ₃
L-Ser	L-Serine
Fumisoquin A	C ₁₃ H ₁₆ N ₂ O ₅
Fumisoquin C	C ₁₃ H ₁₀ N ₂ O ₆
L-Pipecolic acid	C ₆ H ₁₁ N ₂ O ₂
Cadaverine	C ₅ H ₁₄ N ₂
Fawcettimine	C ₁₆ H ₂₅ N ₂ O ₂
Lycodine	C ₁₆ H ₂₂ N ₂
Lycopodine	C ₁₆ H ₂₅ N ₂ O
Phlegmarine	C ₁₆ H ₃₀ N ₂
(-)-Anabasine	C ₁₀ H ₁₄ N ₂
(-)-Anaferine	C ₁₃ H ₂₄ N ₂ O
Castanospermine	C ₈ H ₁₅ N ₄ O ₄
Lobelanine	C ₂₂ H ₂₅ N ₂ O ₂
(-)-Lobeline;Lobeline	C ₂₂ H ₂₇ N ₂ O ₂

Compound Name	Molecular formula
Pelletierine	C ₈ H ₁₅ NO
Piperine;1-Piperoylpiperidine;Bioperine	C ₁₇ H ₁₉ NO ₃
Sedamine;(-)-Sedamine	C ₁₄ H ₂₁ NO
Swainsonine;(-)-Swainsonine	C ₈ H ₁₅ NO ₃
Angustifoline	C ₁₄ H ₂₂ N ₂ O
Lupanine;(+) -Lupanine	C ₁₅ H ₂₄ N ₂ O
(-) -Lupinine	C ₁₀ H ₁₉ NO
(+) -Matrine;Matrine	C ₁₅ H ₂₄ N ₂ O
(-) -Multiflorine;Multiflorine	C ₁₅ H ₂₂ N ₂ O
(+) -Epilupinine	C ₁₀ H ₁₉ NO
alpha-Isolupanine	C ₁₅ H ₂₄ N ₂ O
(-) -13alpha-Hydroxy Multiflorine	C ₁₅ H ₂₂ N ₂ O ₂
(-) -13alpha-Tigloyl oxymultiflorine	C ₂₀ H ₂₈ N ₂ O ₃
Huperzine A;Kimpukan A	C ₁₅ H ₁₈ N ₂ O
Cinnamic acid;beta-Phenylacrylic acid	C ₉ H ₈ O ₂
Piperidine	C ₅ H ₁₁ N
(+) -Oxymatrine	C ₁₅ H ₂₄ N ₂ O ₂
N-Methylpelletierine	C ₉ H ₁₇ NO
alpha-Aminoadipic acid delta-semialdehyde;L-allysine	C ₆ H ₁₁ NO ₃
delta1-Piperideine-6-carboxylic acid	C ₆ H ₉ NO ₂
(+/-) -Sparteine	C ₁₅ H ₂₆ N ₂
delta1-Piperideine-2-carboxylic acid	C ₆ H ₉ NO ₂
epsilon-Amino-alpha-ketocarproic acid	C ₆ H ₁₁ NO ₃
1-Indolizidinone	C ₈ H ₁₃ NO
5-Aminopentanal	C ₅ H ₁₁ NO
delta1-Piperideine	C ₅ H ₉ N
(-) -p-Coumaroyllupinine	C ₁₉ H ₂₅ NO ₃
(-) -Feruloyllupinine	C ₂₀ H ₂₇ NO ₄

Compound Name	Molecular formula
(+)-Coumaroylepilupinine	C ₁₉ H ₂₅ N ₃ O ₃
(+)-Feruloylepilupinine	C ₂₀ H ₂₇ N ₃ O ₄
Diminium cation	C ₁₅ H ₂₄ N ₂
(+)-13alpha-Hydroxy lupanine	C ₁₅ H ₂₄ N ₂ O ₂
(+)-13alpha-Tigloyl oxylupanine	C ₂₀ H ₃₀ N ₂ O ₃
2-Piperidyl-acetophenone	C ₁₃ H ₁₇ NO
Lycodane	C ₁₆ H ₂₂ N ₂
Gliotoxin	C ₁₃ H ₁₄ N ₂ O ₄ S ₂
Phe-Ser-DKP	C ₁₂ H ₁₄ N ₂ O ₃
Coniine	C ₈ H ₁₅ N
N-Methylconiine	C ₉ H ₁₉ N
Pseudoconhydrine	C ₈ H ₁₇ NO
gamma-Coniceine	C ₈ H ₁₅ N
Conhydrine	C ₈ H ₁₇ NO
1'-Oxo-gamma-coniceine	C ₈ H ₁₃ NO
5-Ketooctanoic acid;5-Oxooctanoic acid	C ₈ H ₁₄ O ₃
5-Ketooctanal	C ₈ H ₁₄ O ₂
5-Hydroxy-gamma-coniceine	C ₈ H ₁₅ NO
N-Methylpseudoconhydrine	C ₉ H ₁₉ NO
Conhydrinone	C ₈ H ₁₅ NO
N-Methylconhydrine	C ₉ H ₁₉ NO
Histamine	C ₅ H ₉ N ₃
Urocanic acid	C ₆ H ₆ N ₂ O ₂
Dolicotheline	C ₁₀ H ₁₇ N ₃ O
N-Methylhistamine	C ₆ H ₁₁ N ₃
N,N-Dimethylhistamine	C ₇ H ₁₃ N ₃
Casimidine	C ₁₂ H ₂₁ N ₃ O ₅
Gluchidine	C ₃₀ H ₄₆ O

Compound Name	Molecular formula
Zapotidime	C7H9N3S
Acronycine	C20H19NO3
Melicopicine	C18H19NO5
Peganine;Vasicine;(-)-Linarine;(-)-Peganine;(-)-Vasicine	C11H12N2O
Robustine;(-)-Rutacridone	C12H9NO3
Aurachin A	C25H33NO3
Aurachin B	C25H33NO2
Aurachin C	C25H33NO2
2-Heptyl-4-hydroxyquinoline N-oxide	C16H21NO2
Penigequinolone A	C27H33NO6
O-Methyl-L-Tyrosine	C10H13NO3
2,4-Dihydroxy quinoline;DHQ	C9H7NO2
N-Acetyl-anthranilic acid	C9H9NO3
Anthraniloyl-CoA	C28H41N8O17P3S
Aurachin D	C25H33NO
2-Heptyl-4-hydroxy quinoline;HHQ	C16H21NO
2-Heptyl-3-hydroxyl-4(1H)-quinolone;PQS	C16H21NO2
N-Methyl anthraniloyl-CoA	
1,3-Dihydroxy-N-methylacridone	C14H11NO3
4'-Methoxycyclopeptin	C18H18N2O3
(-)-4'-Methoxycyclophenin	C18H16N2O4
4'-Methoxyviridicatin	C16H13NO2
14-Methoxyviridicatin	C16H15NO5
Quinolinone B	C17H17NO5
Dictamine;Dictamnine	C12H9NO2
N-Methylfindersine	C15H15NO2
Skimmianine	C14H13NO4
Aszonalenin	C23H23N3O2

Compound Name	Molecular formula
N-Acetylaszonalenin;LL-S490beta	C25H25N3O3
(-)-N-Acetylardeemin	C28H28N4O3
Fiscalin A	C26H27N5O4
Fiscalin B;(-)-Fiscalin B	C23H22N4O2
Fiscalin C	C27H29N5O4
(-)-Benzomalvin A;Benzomalvin A	C24H19N3O2
Asperlicin;(-)-Asperlicin	C31H29N5O4
Asperlicin C;(-)-Asperlicin C	C25H18N4O2
Asperlicin D	C25H18N4O2
Asperlicin E	C25H18N4O3
Evocarpine	C23H33NO
Viridicatin	C15H11NO2
Fumiquinazoline F;(-)-Fumiquinozoline F	C21H18N4O2
Alantrypinone;(+) -Alantrypinone	C21H16N4O3
Fumiquinazoline C	C24H21N5O4
Fumiquinazoline A	C24H23N5O4
Fumiquinazoline D	C24H21N5O4
Peniprequinolone;(-)-Peniprequinolone	C22H25NO5
Orixine;(+) -Orixine	C17H21NO6
Glyantrypine	C20H16N4O2
Cycloopenin;(-)-Cycloopenin	C17H14N2O3
Cyclopeptine	C17H16N2O2
Sclerotigenin	C16H11N3O2
Chaetominine	C22H18N4O4
R-Benzodiazepinedione	C18H15N3O2
Dehydrocyclopeptine	C17H14N2O2
2'-epi-Fumiquinazoline A	C24H23N5O4
15-Dimethyl-2' -epi-fumiquinazoline A	C25H25N5O4

Compound Name	Molecular formula
Tryptoquialanine A	C27H26N4O7
Ardeemin FQ	C21H18N4O2
Ardeemin	C26H26N4O2
Auranthine	C19H14N4O2
Circumadatin D	C21H19N3O5
Circumadatin E	C20H17N3O4
Circumadatin H	C20H17N3O3
Circumadatin J	C21H19N3O4
N-Hydroxy-2(1H)-quinolone	
Terpendole C;(-)-Terpendole C	C32H41NO5
Terpendole E	C28H39NO3
Terpendole I	C27H35NO5
Lolitrem E;Lolitrem C	C42H57NO7
Lolitriol	C37H49NO7
Paspaline	C28H39NO2
Paspalicine	C27H31NO3
Paspalinine	C27H31NO4
Paxilline	C27H33NO4
Paspalitrem A;3-Methyl-2-butenylpaspalinine	C32H39NO4
Lolitrem N	C37H49NO7
Shearinine A;(+) -Shearinine A	C37H45NO5
Emindole DA	C28H39NO
Lolitrem B	C42H55NO7
3-Geranylgeranyl indole	C28H39N
Emindole SB	C28H39NO
beta-PC-M6;beta-10-Hydroxy-12-demethyl-11,12-dehydropaspaline	C27H35NO3
alpha-PC-M6;alpha-10-Hydroxy-12-demethyl-11,12-dehydropaspaline	C27H35NO3
13-Desoxypaxilline	C27H33NO3

Compound Name	Molecular formula
Shearinine K	C37H47NO4
13-Desoxy-terpendole I	C27H35NO4
Lolitrem J	C39H51NO8
Lolitrem K	C37H47NO6
alpha-Chaconine	C45H73NO14
Solanidine	C27H43NO
Cholesterol;Cholesterin;Cholest-5-en-3beta-ol	C27H46O
gamma-Tomatine	C39H65NO12
Tomatidenol	C27H43NO2
22-Hydroxycholesterol;22(R)-Hydroxycholesterol	C27H46O2
Tomatidine	C27H45NO2
22,26-Dihydroxycholesterol	C27H46O3
Furostanol-26-aldehyde	C27H42O4
26-Amino-furostanol	C27H45NO3
Tomatidine galactoside	C33H55NO7
beta1-Tomatine	C45H75NO17
alpha-Tomatine	C50H83NO21
Adenine	C5H5N5
Caffeine	C8H10N4O2
Theobromine	C7H8N4O2
Xanthosine	C10H12N4O6
IMP;Inosine 5'-phosphate	C10H13N4O8P
7-Methylxantine	C6H6N4O2
Xanthosine 5'-monophosphate;XMP	C10H13N4O9P
7-Methylxanthosine	C11H15N4O6
AMP;Adenosine 5'-phosphate	C10H14N5O7P
Theacrine	C9H12N4O3
Paraxanthine	C7H8N4O2

Compound Name	Molecular formula
3-Methylxanthine	C ₆ H ₆ N ₄ O ₂
Methylxanthine	C ₉ H ₁₂ N ₄ O ₃
1,3,7-Trimethyluric acid	C ₈ H ₁₀ N ₄ O ₃
Theophylline	C ₇ H ₈ N ₄ O ₂
7-Methyl-XMP;7-Methylguanosine 5'-monophosphate	C ₁₁ H ₁₆ N ₅ O ₈ P
Liberine	C ₈ H ₁₀ N ₄ O ₃