Classification of Alkaloid Compounds Based on Subring Skeleton (SRS) Profiling: On Finding Relationship of Compounds with Metabolic Pathways

Ryohei Eguchi¹, Naoaki Ono¹, Hisayuki Horai², Md.Altuf-Ul Amin¹, Aki Morita Hirai¹, Jun Kawahara¹, Shoji Kasahara¹, Tomoaki Endo³, Shigehiko Kanaya¹*

(1) Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
(2) National Institute of Technology, Department of Industrial Engineering Ibaraki College, 866 Nakase, Hitachinaka, Ibaraki, 312-8508 Japan
(3) National Institute of Technology, Department of General Engineering, Sendai College, 4-16-1 Ayashi-Chuo, Aoba-ku, Sendai-shi, Miyagi 989-3128, Japan

(Received June 1, 2017; Accepted June 13, 2017)

Systematic representation of alkaloid biosynthetic pathways based on ring skeletons has been proposed because the skeleton nucleus of an alkaloid is the main criterion for determination in biosynthetic pathways. So the idea of ring skeletons was extended to apply classification of alkaloid compounds based on ring skeletons and to systematize alkaloid compounds and to examine the performance of this approach to predict biosynthetic pathways based on module elements. We constructed a 2-dimensional binary matrix corresponding to 2546 SRS and 478 pathway-known alkaloid compounds. Here, if ith substring skeleton is present in a target compound, the ith element was set to 1; otherwise, the ith element was set to 0. Relationship of alkaloid compounds with biosynthetic pathways are examined based on the dendrogram produced by Ward clustering method to the matrix. Of 12,243 alkaloid compounds accumulated in KNAPSAcK Core DB (http://kanaya.naist.jp/knapsack_jsp/top.html), 3,124 compounds (25.5%) correspond to the pathway-known ring skeletons (187 ring skeletons), but the remaining 9,119 (74.5%) compounds do not. By examining the sub-ring skeleton similarity of the remaining compounds, it might be possible to obtain clues of pathway information and systemization of all alkaloid compounds. Therefore, the present work focuses on comprehensive systematization of the alkaloid compounds and construction principles of ring skeletons in alkaloids based on subring skeleton profiling.

Key Words: subring skeleton profiling, alkaloids, cluster analysis

* skanaya@gtc.naist.jp
1. Introduction

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 [1]. 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 [2] etc. Secondary metabolites with known chemical structures are more than 3,000 terpenoids, 9,000 flavonoids, 1,600 isoflavonoids and 12,000 alkaloids [3, 4]. In 2008, we started to accumulate the relations between metabolites and producing species and retained them in KNApSAcK Core Database (DB, http://kanaya.naist.jp/knapsack.jsp/top.html), presently which are 101,500 species-metabolite relationships encompassing 20,741 species and 50,048 metabolites. Based on current statistics of the relationships between metabolites and species, it has been predicted that there are at least 1.06 million metabolites within all plants on this planet[5].

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 evolitional process. In the present study, we examined whether or not classification of alkaloids by ring skeletons can be related to metabolic pathways. Alkaloids are a large group of nitrogen-containing secondary metabolites produced by almost every variety of organisms such as bacteria, fungi, plants, and animals; organisms produce diverged polycyclic compounds including unsaturated and saturated bonds (Figure 1 shows some examples).

Aniszewski has established systematic representation of alkaloid biosynthetic pathways based on ring skeletons because the skeleton nuclei of the alkaloids are the main criterion for alkaloid precursor determination in biosynthetic pathways. L-lysine can produce skeletons for different alkaloid nuclei, for example, piperidine, indolizine, quinolizidine nuclei are respectively represented by skeletons C5N, C5NC3 and C5NC4 [6]. Here C5N means a skeleton with 6-membered ring including 1 nitrogen atom, C5NC3 means the heterocyclic ring skeleton composed by 6-membered ring including 1 nitrogen atom together with 5-member ring including 1 nitrogen, and C5NC4 means the heterocyclic ring skeleton composed by 6-membered ring including 1 nitrogen atom together with 6-member ring including 1 nitrogen. C5NC and C5NC4 possesses common C-N bonds in two heterocyclic systems. As shown in the synthetic pathways of quinolizine alkaloids from L-Lysine (L-Lys) to Lupanine (C2) and Multiflorine (C3) and to Lycodine (C6), three compounds C1, C2, and C3 have identical ring skeleton S1, whereas, C4 has the skeleton S2, and C5 and C6 have the skeleton S3. Here ring skeleton means ring structures without taking saturated or unsaturated chemical bonds into consideration. Thus, classification of alkaloids based on skeletons might provide important information for systematic understanding of chemical structures and also have important clues for predicting biosynthetic pathways. So the idea of ring skeletons was extended to apply classification of alkaloid compounds based on ring skeletons and to systematize alkaloid compounds and to examine the performance of this approach to predict biosynthetic pathways based by module elements. We also

Figure 1 Concept of classification based on subring skeleton structure
discuss structure-based classification of secondary metabolites as a multidisciplinary field combining chemoinformatics.

2. Methods

2.1 Data Set

Initially, we accumulated 478 polycyclic compounds whose metabolic pathways were available in references. We obtained pathway information from scientific literatures [6-74] and depicted the summary of 32 pathway maps in Table 1. We made alkaloid pathway maps based on starting compounds as amino acids, compounds related with amino acid biosynthesis (anthranilate, formyl anthranilate, indole-3-glycerol phosphate, O-methyl tyrosine) terpenes, compounds related with TCA cycle and fatty acids and nucleic acids because alkaloids are often divided into the true alkaloids which originate from amino acids and pseudoalkaloids that do not originate from amino acids, for example, terpene-like, steroid-like and purine-like alkaloids.

2.2 Subring skeleton profiling (SRS)

Table 1 1 Alkaloid metabolic pathways used in this study

| Starting substance (SS) | 3-phosphoglycerate | Pyruvate | Phosphoenolpyruvate | Oxaloacetate | alpha-Ketoglutarate | Terpenes | TCA cycle | Fatty acid | Nucleic acid | Refs |
|-------------------------|--------------------|----------|---------------------|--------------|---------------------|----------|----------|-----------|-------------|------|
| L-Tyr                   | L-Orn              | L-His    | L-Val               | L-Ile        | L-Asp               | L-Met    | L-Leu    | L-Lys     | L-Arg       |      |
| L-Phe                   | L-Val              | L-Leu    | L-Arg               | L-Lys        | L-Asp               | L-Met    | L-Leu    | L-Lys     | L-Arg       |      |
| L-Trp                   | L-Val              | L-Leu    | L-Arg               | L-Lys        | L-Asp               | L-Met    | L-Leu    | L-Lys     | L-Arg       |      |
| L-Asp                   | L-Met              | L-Leu    | L-Arg               | L-Lys        | L-Asp               | L-Met    | L-Leu    | L-Lys     | L-Arg       |      |

Figure 2 shows the subring skeleton profiling with the example of bizoisoquinoline alkaloid biosynthetic pathway from L-Tyr to Sanguinarine (C13). Out of 7 compounds (C7-C13), C12 and C13 have identical ring skeletons S9 and the others have individual ring skeletons (S4-S8). Next, subring skeletons were produced for individual skeletons (S4-S9). For example, S4 corresponds to nine subring skeletons (SRSs) (A-I in Step 1 of Figure 2). SRS matrix S was constructed by setting binary elements such that $s_{ij}$ = 1 or 0 respectively in case of presence or absence of the $j$th SRS in the $i$th compound (Step 2). Here we used InChI (the IUPAC International Chemical Identifier) for isomorphism check for SRSs [74, 75]. In Step 3, compounds were clustered based on hierarchical clustering methods. Molecular fingerprint techniques have been developed, that is, PubChem (881 bits) [76], CDK (1024 bits) [77], Extended CDK (1024 bits) [78], MACCS (166 bits) [79], Klekota-Roth (4860 bits) [80], Substructure (307 bits) [81], Estate (79 bits) [82], and atom pairs (780 bits) [83]. Those molecular fingerprints generally focuses on side-chain substructures of molecules. Alternatively, SRS profiling makes it possible to examine and compare compounds based on all possible SRS, so SRS profiling is useful for systematic...
understanding of building principle of rings systems.

3. Results

3.1 Subring skeleton Profiling in Alkaloids

Firstly, we extracted subring skeletons from 478 compounds and obtained 2546 unique SRSs. Figure 3 shows the distribution of the numbers of compounds along with the number of SRSs. It is interesting that there is no unique SRS in 478 compounds, that is, there is no SRS defined by single compounds. All subrings correspond to multiple alkaloid compounds. Thus, individual alkaloid compounds are related with each other by targeted unique SRSs.

Secondly, we represented individual compounds as 2546 dimensional binary vectors. Here, if $i$th subring skeleton is present in a target compound, the $i$th element was set to 1; otherwise, the $i$th element was set to 0. We applied ward clustering method to a matrix consisting 478 compounds and 2546 unique SRSs and tentatively separated 478 compounds into 29 clusters (Cluster ID = 1 to 29) as visualized by dendrogram in Figure 4. All compounds were assigned to 187 ring skeletons corresponding to 1-187 in Appendix Figure A.

We classified compounds into 6 groups designated by G1 (Clusters 1 and 2), G2 (Cluster 3), G3 (Cluster 4), G4 (Clusters 5-7), G5 (Clusters 8) and G6 (Cluster 9-29) as shown in Figure 4; G1 and G2 are related with indole-diterpenes involving paxilline, terpandoles and lolitrems (G1) and strictosides (G2). These consist of polycyclic...
compounds comprised of six to nine rings characterized by 5-6 member rings and only one nitrogen; Compounds in G3 (Cluster 4) are related with steroidal alkaloids with glucosides including alpha-solanine, alpha-choconine, and alpha-tomatine and their skeletal chemical structures are assigned with steroidal alkaloids; compounds included in G4 (Clusters 5-7) are characterized by vindoline derivatives with greater than or equal to 2 nitrogen atoms; G5 corresponds to Ergot alkaloids. The other alkaloid compounds belong to G6 (Clusters 9 to 29) and also have very diverged ring skeletons. Clusters 9 to 20 can be characterized as follows; aglycones and glucosides of Tomatidines (Clusters 9 and 10, respectively), bis-isoquinolines (Cluster 11), morphinans (Cluster 12), quinazoline derivatives (Cluster 13), indolizidine derivatives (Cluster 14), polyneuridines (Cluster 15), Ajmalicines (Cluster16), Berberine and relatives (Clusters 17 to 19). Ipecac alkaloids (Cluster 20), isoquinolines including ring skeletons of glaudines (Cluster 21), beta-carbolines (Cluster22), chateoglobosins (Cluster 23), ring skeletons roquefortines and acetylsazonaleins (Cluster 24), Emindoles (Cluster 25), quinolizidines (Cluster 26), iboganines (Cluster 28), and communesins (Cluster 29).

Thus the 28 clusters except cluster 27 can be characterized by ring skeletons based on the sub-ring skeleton profile proposed by the present study.

Cluster 27 consists of half of the compounds (233 compounds) examined in this study. So we further divide it into 25 sub-clusters to characterize ring skeletons as representatives of groups of compounds. The 25 sub-clusters in Cluster 27 are composed of relatively simple heterocyclic ring skeletons in comparison to other clusters. Five sub-clusters (A, B, C, D, E, F, and H) have indole type alkaloid structures based on ring skeletons, and especially characterized by brevianamides (A); cyclopiazonate derivates (B) and chanoclavin derivatives, that is, the former corresponds to a five-cyclic ergot alkaloids and the latter corresponds to tri-cyclic ergot alkaloids; chimonanthines (E); and iboga alkaloids (H). Furthermore, remarkable ring skeletons can be observed in individual sub-clusters such as, ergolines (B and C), cinchona alkaloids (G), iboga alkaloids (H and I), lycopodium alkaloids (J and K), isoquinolines (L), lupine alkaloids (M), quinolones (N), quinolones (O), acridines and quinolones (P), purine alkaloids (Q), and benzylisoquinoline alkaloids (T). It is noteworthy that all three polycyclic systems in lupine alkaloids, matrine type, (120 and 121 in skeleton ID of Figure 3) sparteine/lupanine type (122 124 and 125), and lupinine type (126) are clustered into sub-cluster 27M. In summary it can be concluded that, alkaloid compounds can be classified according to generally defined ring systems based on subring skeleton profiles.

3.2 Pathway construction based on Scientific Literature

We classified alkaloid compounds based on subring skeleton profile and thus associated general ring structures to groups of alkaloids. Next, we tried to apply this concept to find relations between ring structures and modules in alkaloid metabolic pathways. Initially, we mapped compounds onto 32 pathway maps (Figure 5) and summarized the relations between classification results and pathway maps in Table 2. All but four clusters (Clusters 13, 22, 24, 25 and 27) correspond to single pathway maps (PMs). Thus, clusters expressed by the subring skeleton profile are highly related with alkaloid
biosynthetic pathways. Such trends are also observed in case of the sub-clusters in Cluster 27 (10 in 25 sub-clusters). So in this section, we discuss on extracted modules on the pathway maps in terms of the clustering results.

A variety of compounds can be biosynthesized by six amino acids (aspartate, proline, arginine, glycine, threonine, phenylalanine) as shown in Figure 5A. The compounds of five sub-clusters (S, Q, W, X, Y) in Cluster 27 belong to PM1. In particular, sub-clusters X, V, W are associated with pyridine, loline, and tropane alkaloids, respectively. Thus, similarity of subring skeleton profiling makes it possible to group those remarkable chemical structures.

PMS2-10 correspond to L-tryptophan derivative alkaloids but do not include anthranilate and formyl anthranilate. L-Trp and isopentenyl diphosphate (IPP) derivatives are mapped in PM2 (Figure 5B). We can comprehensively understand the skeleton construction process by the clustering results, that is, IPP and L-Trp leads to Ergoline skeletones (Cluster 8) by two steps represented by sub-clusters D (indole skeletons) and C (chanoclavin derivatives).

Monoterpenoid indole alkaloid pathway biosynthesized by geranyl diphosphate (monoterpenoids) and L-Trp is depicted in Figure 5C. Ring skeletons reflected in eight clusters (Clusters, 3, 5, 6, 7, 15, 16, 22, 28) are related to monoterpenoid indole alkaloid pathway. Initially secologanin (monoterpenoids) and L-Trp lead to strictosidine assigned to Cluster 22 (beta-carbolines), then to diverge ring skeletons assigned to different clusters such as Cluster 28 (Iboga alkaloids) and sub-cluster G (Cinchona alkaloids), Clusters 5-7 (Vindoline derivatives), Clusters 15 and 3 (polyneuridines and stricosides, respectively).

Table 2 | Classification of alkaloid compounds in clusters wit pathway maps.
L-Trp derivatives and combination of L-Trp with L-Pro, Acetyl-CoA, Malonyl-CoA, oxaloacetate, dimethylallyl pyrophosphate (DMAPP), acetoacetyl-CoA and L-His lead to biosynthesize very diverse alkaloid compounds depicted in PMs4-10 (Figure 5C). PM4 comprise ring skeletons in Cluster 14 (mainly characterized by indole skeletons); PM5 in Clusters 22 (beta-carbolines), 23 (chatoglobosins), and 29 (communesins); PM6 in Clusters 22 (beta-carbolines). In PM5, L-Trp mainly leads to ring skeletons assigned in sub-cluster D (indole derivatives), then biosynthesize the compounds in Cluster 22, 23, and 29.
Figure 5E (PM11) shows betalanin and colchine biosynthetic pathways; major ring skeletons in betalanin-related compounds belong to sub-cluster S and major ring skeletons in colchicine-related compounds belong to sub-cluster U. Though betalains and colchicine are derived by L-Tyr and L-Phe, respectively, and adrenaline is derived by both L-Tyr and L-Phe, most of the compounds belong to sub-cluster S in the initial biosynthesis. Sub-cluster S integrates mainly mono-, di- and tri-cyclic skeletons with units of five and six heterocycles. Thus PM11 can be interpreted as biosynthetic pathways based on units with relatively small heterocyclic ring skeletons.
Figure 5F (PM12) shows L-Tyrosine derivate alkaloid biosynthetic pathways which is associated to Cluster 11 (bis-isooquinolines), Cluster 12 (morphinans), Clusters 17-19 (berberines), and Cluster 21 (isooquinolines). Those clusters can be explained by the relation that L-Tyr initially derives ring skeletons related with quinolones (Cluster 21), and those skeletons diverge into three-groups of skeletons (Clusters 11, 12 and 17-19). Secologanin and L-Tyr can also derive ring skeletons related with isooquinolines (Cluster 21 in Figure 5G; PM13). The ring skeletons of Iecac can be derived by those of isooquinoline designated by Subclass L. The ring skeletons designated in Cluster 31 (quinazlolines) can be derived via Clusters 20 and 21.

L-Lys derives relatively simple ring skeletons because most of the compounds belong to sub-clusters in Cluster 27 (Figure 5H; PM15) and to Cluster 26 (quinolizidines). The number of compounds included in Cluster 27 is half.
of the compounds analyzed in this work and the ring skeletons are relatively simple compared to those in other 28 clusters. In PM15, Lys derivatives can be characterized by subcluster levels of Cluster 27 such as subclasses Y (Isodolizidines), L and M (Quinolizidines), J, K, P, S and X (Lycopodium alkaloids and piperidines). L-Ala, Malonyl-CoA, and Acetyl CoA can derive a series of piperidine alkaloids characterized by the ring skeletons of Subclass X (PM17; Figure 5I) and L-His and L-Leu can derive imidazole alkaloids also characterized by the ring skeletons of Subclass X (PM18; Figure 6I).

Anthranilate has the ability of linking with a variety of compounds to produce diverged alkaloids. PMs19-30

(Figures 5I and J) represent biosynthetic pathways by combination of anthranilate with O-Methyl-L-tyrosine (PM19), L-Trp (PM20), L-Phe (PM21), L-Trp and D/L-Ala (PM22), D-Trp, L-Ala, and L-Val (PM23), D-Ala and L-Ala (PM24), L-Trp and Gly (PM25), Gly (PM26), L-Pro (PM27), and L-Gln (PM29) and by combination of N-Methylanthranilates with Malonyl CoA (PM28). Though the combinations are very diverse (PMs19-29), two consensus clusters (Clusters 13 and 24) was obtained related to biosynthesis of anthranilate derivative alkaloids. Clusters 13 and 24 correspond to ring skeletons of quinazolines and very complex ring involving roquefortines and acetylazonaleins, respectively.

Figure 5G Pathway Map (PM13,14)

Figure 5H Pathway Map (PM15, 16)
Biosynthetic pathway of Indole-diterpene alkaloids is depicted in PM30. On the basis of subring skeleton profiling, indole-3-glycerolphosphates and Geranylgeranyl diphosphate (GGPP) initially derive ring skeletons assigned in Cluster 2 and then more complex skeletons (Cluster 1) are biosynthesized so that complex skeletons can be constructed in biosynthesis of indole-diterpenes involving Paxilline, Terpendoles and Lolitrems.

There are three main types of alkaloids, true alkaloids, protoalkaloids and pseudo alkaloids. True alkaloids and...
protoalkaloids are derived from amino acids mentioned above, whereas pseudoalkaloids are not derived from amino acids. PM31 and PM32 (Figure 5K) correspond to pseudoalkaloids, and starting compounds are cholesterol and adenine, respectively. Four ring skeletons associated to Clusters 25 (Emiindoles), 10, 9 (aglycon and glucoside of Tomatidines, respectively) and 4 (Tomatins and Chaconine) can be defined in the biosynthetic process from Cholesterol to tomatine and chaconine. On the other hand, ring skeletons are very similar to each other in purine alkaloids because all compounds in PM32 belong to Subclass Q.

It can be concluded that we can comprehensively systematize the alkaloid compounds and explain construction principles of ring skeletons in alkaloids based on subring skeleton profiling.

4. Discussion

The pathways of all known metabolites are not known. In the present study, we surveyed the pathway information in scientific literatures as well as in KEGG pathway database (http://www.genome.jp/kegg/pathway.html) and obtained 478 heterocyclic alkaloid compounds. However, chemical structures of alkaloid compounds have been determined extensively and we have accumulated information of those compounds as well as other natural compounds. There are 12,243 compounds with nitrogen atoms related to alkaloids in the KNApSAcK Core Database. In the present study, we defined 187 ring skeletons in pathway-known alkaloids. We tried to assign 12,243 compounds to pathway-known ring skeletons and observed that 3,124 compounds (25.5%) correspond to the skeletons, but the remaining 9,119 (74.5%) compounds do not. By examining the subring skeleton similarity of the remaining compounds, it might be possible to obtain clues of pathway information and systemization of all alkaloid compounds.

Bioinformatics and Chemoinformatics play important roles in interpretation and understanding chemical and biological fields based on chemical and biological data, and in fast and efficient calculations for handling huge data sets. The enumeration process of unique SRSs is a very interesting subject. The complete set of SRSs for S4 in Figure 2 can be obtained by considering all combinations of edges included in S4, where the 1-branch (resp. 0-branch) means that the edge is connected (resp. not connected) in the set (Figure 6A). Thus, 2\(^19\) combinations of 1-branches and 0-branches for 19 edges are possible (Figure 6B). Zero-suppressed binary decision diagrams (ZDDs) [83] make it possible to reduce the enumeration process of SRSs by the two rules of sharing nodes (represented by blue arrow in Figure 6) and deleting nodes (red arrow). A ZDD is a labeled directed acyclic graph derived by reducing a binary decision tree graph. Decision making processes are carried out through binary input variables (Figure 6C). The deletion rule is that a node with a 1-branch directly pointing to the 0-terminal node can be deleted (red arrow) and the sharing rule is that the equivalent nodes possessing the same indices and the same 0- and 1-child nodes, can be shared (blue arrow). It is known that there is an algorithm that directly constructs a ZDD (Figure 6C) without creating a binary tree (Figure 6B) [84, 85], which reduces much computation time. If a ring skeleton can be converted into graphs based on ZDDs, the enumeration process is reduced to a conventional algorithm (Figure 7B). For example, the number of enumeration process decreases from 2\(^19\) to 9 in the example of Figure 7A. Thus, it is obvious that the ZDD algorithm can reduce the enumeration process of SRSs in compounds.

It has been demonstrated by using 3D structure similarity between metabolites based on heuristic graph...
match algorithm COMPLIG [86, 87] that activities of secondary metabolites are highly related with chemical structures, for example, isoquinoline alkaloids are related with analgesic agents; quinolone alkaloids are with sedatives, terpene alkaloids with antihypertension agents, steroid alkaloids with antitumor activities, terpene alkaloids with anti-inflammatory agents, and tujupa alkaloids are with antidiabetic agents. Those relations are consistent with the clustering analysis of SRS profiles. Furthermore SRS profiles make it possible to classify compounds to biosynthetic pathways (Figure 6). In the present study we focused on the SRS profiles to systematize alkaloids and to relate individual compounds to metabolic pathways. Applying clustering method to SRS profiles, the drastic change of ring skeletons across 29 clusters and 25 subclusters was clearly understood. Systematization of core structure, biological activity and biosynthetic pathway of metabolites will provide new insight into interpretation of evolution process for active compounds in nature.

Acknowledgements

This work was partially supported by the National Bioscience Database Center in Japan; the Ministry of Education, Culture, Sports, Science, and Technology of Japan (16K07223 and 17K00406); NAIST Big Data Project; the Platform Project for Supporting Drug Discovery and Life Science Research funded by Japan Agency for Medical Research and Development (AMED).

References and Notes

[1] Kossel A On the chemical composition of the cell.
Appendix Figure A: All ring skeletons in the present study. The representative chemical structures are illustrated for
Appendix Figure A (Continued)
Appendix Figure A (Continued)