

Metabolic pathway prediction based on inclusive relation between cyclic substructures

Kenichi Tanaka¹, Kensuke Nakamura¹, Tamio Saito², Hiroyuki Osada², Aki Hirai¹, Hiroki Takahashi¹, Shigehiko Kanaya¹, Md.Altaf-Ul-Amin^{1,*}

¹ Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan;

² Chemical Library Management and Informatics team, Advanced Science Institute, RIKEN, Wako, Saitama 351-0198

*E-mail: amin-m@is.naist.jp Tel: +81-743-72-5257 Fax: +81-743-72-5258

Received September 4, 2009; accepted October 6, 2009 (Edited by T. Abe)

Abstract Secondary metabolites are highly species-specific and play important roles in the survival of the producing organism within its natural habitat. Systematization of secondary metabolic pathways is necessary to understand species-specific metabolic pathways and to develop new drugs, etc. To attain this, we have made a database system called KNAPSAcK, which describes the relationships between metabolites and species. On March 21, 2009, KNAPSAcK had 34,852 metabolite entries and 68,819 metabolite-species pair entries. Though the chemical structures of around 50,000 secondary metabolites are known, information on their pathways is very limited. In this work, we have developed an algorithm to predict metabolic pathways on the basis of chemical structures of metabolites by exploiting the information contained in their cyclic substructures. Also, to handle a huge amount of these metabolites and to predict metabolic pathways automatically, we have developed a software tool called MetClassifier. MetClassifier is written in C and uses the OpenGL, GLUT (<http://www.opengl.org/resources/libraries/glut/>) and AntTweakBar (<http://www.antisphere.com/Wiki/tools:anttweakbar>) libraries. MetClassifier can be downloaded from the following URL (<http://kanaya.naist.jp/MetClassifier/>).

Key words: Metabolomics, pathway prediction, MetClassifier, KNAPSAcK, KEGG.

Secondary metabolites of the plant kingdom have long been important as leading precursors in the pharmaceutical industry (Simmond and Grayer 1999). Reconstruction of biopathways in plants plays key roles in effectively biosynthesizing those precursors, but rational engineering of secondary metabolic pathways in plants requires a thorough knowledge of the whole biosynthetic pathway and a detailed understanding of the regulatory mechanisms controlling the onset and flux of the pathways. Such information is not yet available for the vast majority of secondary metabolites though studies have progressed extensively. For example, chemical structures of around 50,000 secondary metabolites from the plant kingdom have been determined (Verpoote 1998; De Luca and St Pierre 2000) whereas only around 2,000 enzyme reactions are known. Some researchers have predicted more than 200,000 metabolites for the plant kingdom and the number of plant species is predicted to be around 400,000 in the world (Hostettmann 2000). Thus, experimental evidence is insufficient for assigning all metabolites to metabolic pathways.

There are several approaches for pathway prediction such as fingerprints (Tohsato and Nishimura 2008),

reaction rule-base (Langowski and Long 2002; Talafous et al. 1994; Ellis et al. 2006; Hou et al. 2004; Oh et al. 2007) and maximum common subgraph search (MCSS) (Kotera et al. 2008). The fingerprint-based approach predefines some important molecular fragments and determines which fragments are included in each metabolite as bit-strings consisting of 0's and 1's. This approach can measure similarity between two metabolites without much computational effort, but these fingerprints can't consider connectivity between each fragment, making it difficult for this approach to predict correct pathways. Rule-based approaches predefine reaction rule-base on the basis of predefined organic metabolic reactions and predict possible pathways. Prediction by these approaches has the limitation that it depends on the restricted types of rule-base. Also, there is the possibility that the result of prediction of unknown pathways is biased by the nature of known pathways. An MCSS-based approach does not require predefined information, but this approach is an NP-hard problem (Hattori et al. 2003). Therefore, an MCSS-based approach is computationally difficult and probably to reduce the burden of computation in one such approach, only 2,502,333 metabolite pairs were compared while

their target was 74,766,971 pairs collected from the KEGG database (Kotera et al. 2008) causing the possibility of generating incomplete results.

The building blocks for the secondary metabolism are strongly regulated and we observed the inclusive relationship between substrate and product metabolites at the cyclic substructure (defined in the next section) level. In the present study we have developed prediction of the biosynthetic pathways based on a method for the inclusive relation between the cyclic substructures of metabolites originating from identical species. Our strategy can predict pathway relations at the cyclic substructure level from 28,675 metabolites in the KNApSAcK database in about three minutes (Test machine spec: Intel core2 Duo processor T7600 2.33 GHz, 1 GB RAM).

Methods

The proposed algorithm

Chemical structures of metabolites are considered as molecular graphs. Before giving details of the algorithm, we define some terms utilized in the present study.

Definition 1: Molecular graph

A molecular graph is the representation of the chemical structure of a molecule as a graph where atoms are nodes and bonds are edges.

Definition 2: Cyclic subgraph

A cyclic subgraph is the maximal subgraph of a molecular graph where the degree of each node is two or more, generally these graphs are called 2-core graphs.

If no such subgraph exists then a null graph is considered as a cyclic subgraph. For example the cyclic subgraph corresponding to Thiothecce 460 of Figure 2 is a null graph.

Definition 3: Inclusive relation

A cyclic graph A has an inclusive relation with a cyclic graph B if B is a subgraph of A and we express this relation as $B \subset A$.

Definition 4: Parent graph

We say graph B is a parent graph of graph A if $B \subset A$ and there exists no other graph C such that $B \subset C \subset A$.

The flow-chart of the algorithm

The flowchart of the algorithm is shown in Figure 1 and it is divided into five major steps: (a) Data input, (b) Unique cyclic subgraph extraction, (c) Fingerprint matrix formation, (d) Parent matrix formation and (e) Output.

a) Data input

The input to this algorithm is chemical structures from a database. Let us select N chemical structures and represent them as a set of molecular graphs $G(N)$.

b) Unique cyclic subgraph extraction

From each molecular graph $g_i \in G(N)$, we extract cyclic subgraph cg_i , where $i=1, 2, \dots, N$. Cyclic subgraphs are extracted as follows. Step 1: Check whether the degree of each

node of g_i is 2 or more; if yes then g_i is a cyclic subgraph, call it cg_i and exit, or else remove all nodes with a degree smaller than 2. Step 2: Update all node degrees of g_i . Step 3: If presently the highest degree in g_i is 0 or 1 then exit considering cg_i a null graph, or else continue to Step 1. Here, we define $CG(N)$ as the set of all extracted cyclic subgraphs. By deducting the redundant elements from $CG(N)$, we determine the unique cyclic subgraph set and define it as $UCG(M)$. Obviously $M \leq N$.

c) Fingerprint matrix formation

For each cyclic subgraph, we check its inclusive relation with all other subgraphs by a substructure search and represent that as a fingerprint vector. For example, in Figure 1, ucg_3 has an inclusive relation with ucg_1 and ucg_4 but not with ucg_2 and ucg_5 . Therefore, the fingerprint vector of ucg_3 is [1, 0, 1, 0]. The fingerprint vectors of all the unique cyclic subgraphs constitute the fingerprint matrix. Let F be the fingerprint matrix and $F[A][B]=1$ implies $B \subset A$. The fingerprint matrix is not a diagonally symmetric matrix.

d) Parent matrix formation

For each cyclic subgraph ucg_i , we find parent subgraphs. Parent subgraphs of ucg_i are found by the following steps. Step 1: For each cyclic subgraph ucg_j , if $ucg_j \subset ucg_i$, then ucg_j is a candidate parent of ucg_i , or else ucg_j is not a parent of ucg_i . Step 2: For any other cyclic subgraph ucg_k , if the relation $ucg_j \subset ucg_k \subset ucg_i$ does not exist, then ucg_j is a parent of ucg_i . (In this part, we don't recalculate the inclusive relation by substructure search, but use the fingerprint matrix as described in the flowchart of Figure 1.)

For example (in Figure 1), ucg_2 has an inclusive relation with ucg_1 and therefore ucg_1 is a candidate of a parent of ucg_2 , but an intermediate subgraph ucg_4 exists such that $ucg_1 \subset ucg_4 \subset ucg_2$. So ucg_1 is not a parent but an ancestor of ucg_2 . Indeed, ucg_4 is a parent of ucg_2 . We store information about parents in the parent matrix. Let P be the parent matrix, and $P[A][B]=1$ implies B is a parent of A . Like the fingerprint matrix, the parent matrix is also not a diagonally symmetric matrix.

e) Output

The output of the algorithm is the parent matrix that can be represented as a parent-child directed network showing probable metabolic pathways at the cyclic structure level. This information can be extended to predict pathways at the chemical structure level.

System

For the automation of the above algorithm and other related processes, we have developed a software tool called MetClassifier. We have developed MetClassifier to handle a huge amount of metabolites for prediction of metabolic pathways automatically and to learn more about the complex world of metabolites. For our purposes, not only faster calculation is important but also an interactive user interface is essential, and therefore MetClassifier has been developed using easy to operate graphical user interface (GUI) facilities. MetClassifier uses chemical structural data format MDL Molfile and Sdfile (Symyx, <http://www.symyx.com/>). MetClassifier is written in C and uses the OpenGL, GLUT and AntTweakBar libraries. MetClassifier can be downloaded from

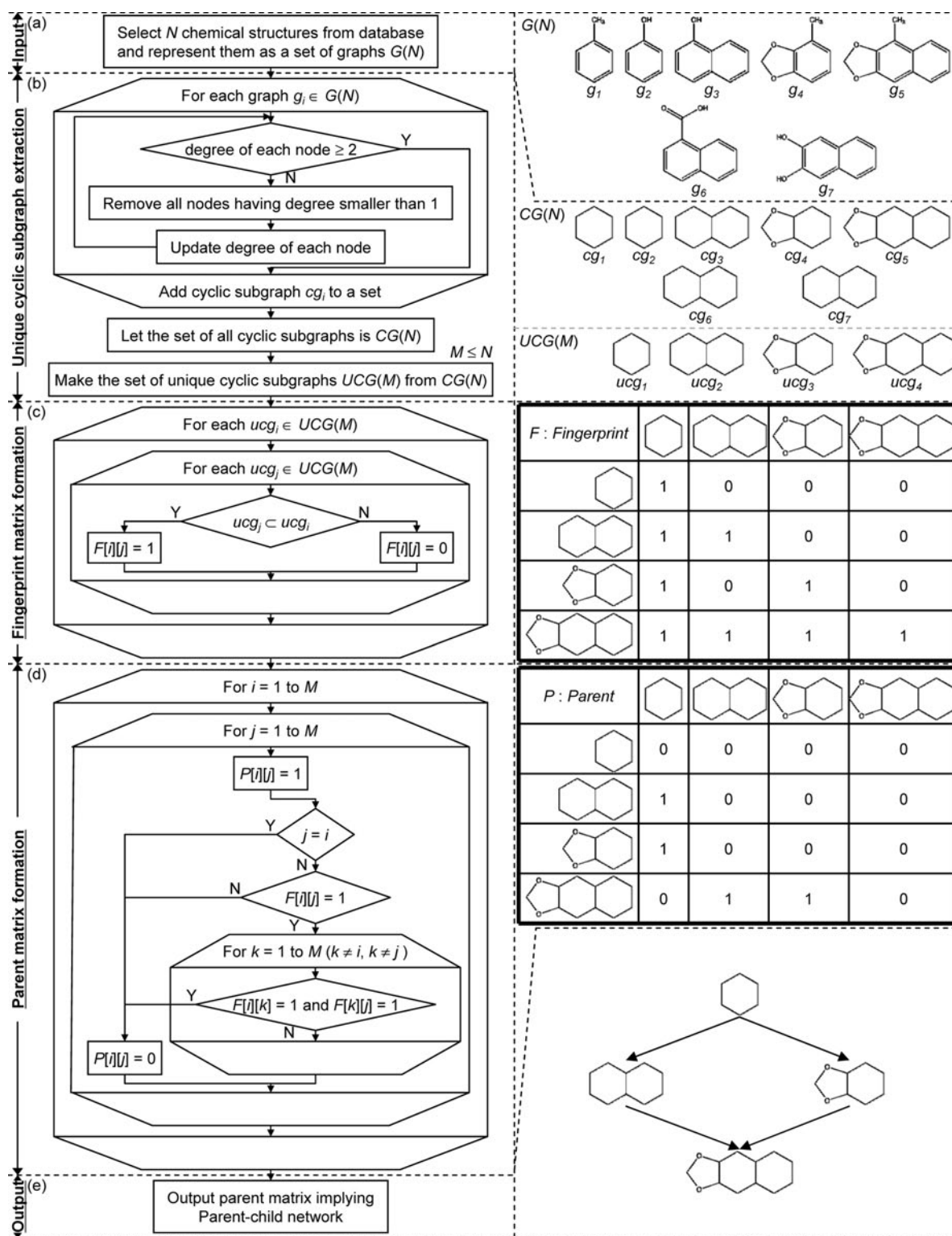


Figure 1. Flow-chart of the proposed algorithm along with a simple example.

the following URL (<http://kanaya.naist.jp/MetClassifier/>).

Dataset

To demonstrate the performance of the proposed algorithm, we applied it to two separate sets of metabolites collected from KEGG on May 29, 2009 (Kanehisa and Goto 2000) and

KNApSack on April 15, 2009 (Shinbo et al. 2006). Chemical structural datasets are necessary to predict pathways at the cyclic substructure level. We use molfiles as chemical structural datasets from KNApSack and KEGG. Furthermore, to evaluate our system and to discuss biological viewpoints, we use additional information on metabolite-species relation from

	Portensterol	Thiothece 460
Unconnected	$2^{(\# \text{ of edges})} = 2^{33} = 8,589,934,592$	
Connected	5,787,190	7,811
Cyclic	1	1

Figure 2. Chemical structures of Portensterol and Thiothece 460. Both metabolites include 33 bonds (excluding the bonds with hydrogen), but the number of connected subgraphs is very big in the case of complex structure specially containing cycles. In this case, the MCSS-approach requires enormous computational resources.

KNApSAcK and reaction data from KEGG.

Results and discussion

Important features of the proposed method

In general, metabolic pathways are predicted by estimating similarities between metabolites. In the case of the MCSS approach, to measure the similarity between two metabolites, first the maximum common connected substructure between their chemical structures is determined. However, this is computationally difficult when chemical structures are big and complex. For example, Figure 2 shows the cyclic structure of Portensterol (KNApSAcK ID=C00023762), and the non-cyclic structure of Thiothece 460 (C00023121). Both metabolites include 33 bonds (excluding the bonds with hydrogen). For both structures, the number of unconnected subgraphs is equal to $2^{\text{number of edges}}$, which is an enormous number. Therefore, MCSS focused on connected subgraphs but still it is very big in the case of complex structures specially containing cycles (as shown in Figure 2, 5,787,190 connected subgraphs are in Portensterol compared to 7,811 in Thiothece 460, which does not contain any cycle). The MCSS approach must find the maximum common connected subgraph between two metabolites, which theoretically requires comparisons proportional to the product of the numbers of connected substructures in the metabolites under consideration. However, the number of comparisons can be somewhat reduced using some constraints, but computational cost remains huge.

To solve this problem, we introduce the concept of the cyclic subgraph. Whereas a chemical structure contains numerous connected subgraphs, it contains only one cyclic subgraph. For almost all reactions, an inclusive relation exists between cyclic subgraphs of substrate and product, i.e. one cyclic subgraph contains another cyclic subgraph perfectly. Therefore, to establish the product-substrate relation between two metabolites, we search for an inclusive relation between their cyclic subgraphs,

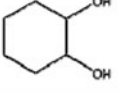
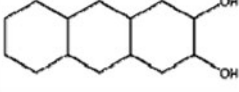
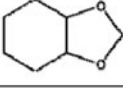
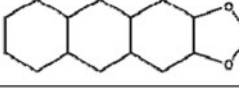
	
	
Tanimoto coefficient $= 8 / (8 + 0 + 1) = 0.89$	Tanimoto coefficient $= 16 / (16 + 0 + 1) = 0.94$

Figure 3. The Tanimoto coefficient is calculated as the ratio of the number of atoms included in the common structure to the number of atoms included in the union of both structures. The Tanimoto coefficient puts emphasis on the size of common features. So Tanimoto coefficients are different even though two reactions are very similar.

which can be performed by only one comparison, making the computation cost very low compared to the MCSS approach, which must perform many comparisons to determine the maximum common subgraph. As a result, our approach offers very fast computation.

Both fingerprint and MCSS approaches estimate chemical structural similarity by the Tanimoto coefficient. It should be noted that the Tanimoto coefficient is not appropriate for pathway prediction because of the following two reasons: the Tanimoto coefficient puts emphasis on the size of common features between two structures. Therefore, Tanimoto coefficients are different even though two reactions are very similar (Figure 3). If we set a lower threshold for pathway prediction, the possibility is high that false-positive pathways are included in the case of larger compounds; on the other hand, if we set a higher threshold then many true-positive pathways may be rejected in the case of smaller compounds. In known pathways, the extent of structural change between substrate and product varies, because of which the Tanimoto coefficient may vary a lot. For example, methylation and hydroxylation cause small change, but glycosilation and dimer formation cause large change between substrate and product. This means it is difficult to select a suitable threshold regarding the Tanimoto coefficient that can perform well for different types of pathway prediction. In our case, there is no necessity of selecting any threshold. In the proposed method, we do not focus on the extent of the difference between substrate and product but focus on inclusive relation of their cyclic subgraphs. Finally, this algorithm selects the most similar subgraphs as the parent subgraphs from all the subgraphs with inclusive relation, making our approach much more relevant to bio pathways.

Performance evaluation using data from KEGG

To demonstrate the performance of the proposed method in a simple way, we collected a part of map00942 from

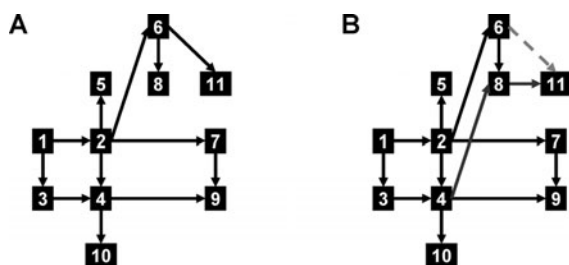


Figure 5. Predicted pathways. (A) pathways of Figure 4 and (B) the predicted pathways at cyclic structure level. The dotted red arrow in (B) is rejected by our prediction. Arrows from 4 to 8 and from 8 to 11 in (B) indicate new predictions. Our algorithm recognizes CID 8 as an intermediate structure between CID 6 and CID 11. In Figure 4, CID 6 has Cycles A, B and C. CID 8 has Cycles A, B, C and D. CID 11 has Cycles A, B, C, D, and E. So the pathway from CID 6 to CID 11 is rejected and another new pathway from CID 8 to CID 11 is suggested. The algorithm found another new pathway from CID 4 to CID 8.

intermediate structure between CID 6 and CID 11. In Figure 4, CID 6 has Cycles A, B and C. CID 8 has Cycles A, B, C and D. CID 11 has Cycles A, B, C, D, and E. Therefore, the pathway from CID 6 to CID 11 is rejected and another new pathway from CID 8 to CID 11 is suggested. The algorithm found another new pathway from CID 4 to CID 8.

We also applied the present method to all KEGG RPAIRs available at (<http://www.genome.jp/kegg/download/ftp.html>). In general, a reaction consists of multiple reactants but for this analysis we considered only 9,577 main pairs involving 5,701 metabolites. These 5,701 metabolites correspond to 585 unique cyclic substructures. At the cyclic structure level, 9,577 main pairs can be represented by 735 relations. The predicted pathways by the present method contain 1,198 relations, out of which there are 521 matches with KEGG RPAIR cyclic level relations and 214 KEGG RPAIR cyclic level relations were missed while 677 new relations were predicted.

Predicted pathways in the context of species-metabolite relation

The present method makes it feasible to enumerate topologically unique cyclic structures and determine inclusive relations between cyclic structures (called cyclic pairs). Out of 34,852 chemical structures contained in KNApSACk DB, we have identified 5,281 topologically unique cyclic structures. Among these, 16,602 pairs of inclusive relations were found. Also, 3,009 pairs of inclusive relations were found in at least one organism, and 67.0% of these cyclic pairs were found in only one species. We can observe the power-law distribution for the number of species and the number of cyclic pairs in each species (Figure 6). In a previous study (Shinbo et al. 2006), we observed the power-law distribution for the number of species and the number of metabolites in each species. It is noteworthy that the

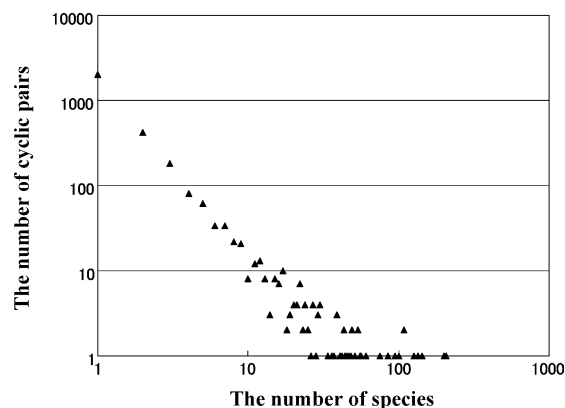


Figure 6. Power-law distribution for the number of species and the number of cyclic pairs in each species.

metabolic pathways of individual organisms also follow the power law, i.e., the probability $P(k)$ that a metabolite interacts with k other metabolites in the metabolic pathway decays as a power law, following $P(k) \sim k^{-r}$, where r is a constant (Ravasz et al. 2002).

It should be noted that in the present case, not so many cyclic pairs are shared in many species, that is, most of the cyclic pairs are common in few species. The cyclic pairs shared by the largest number of species are listed in Table 1. All of these processes are involved in biosynthesis of unique and vital secondary metabolites of plants, such as flavonoids (cyclic pair ID (CPID) 1, 2, 3, 4, 5, 7, 11, 13, 14, 17), alkaloids (CPID 6, 9, 15), gibberellins (CPID 8, 12) or lignans (CPID 16). Following are more detailed descriptions of these most frequently observed CPIDs.

The most common cyclic pairs are attributed to glycosilation of flavonoids (CPID 1, 2, 7, 11, 13, 14 and 17) catalyzed by hexose transferases (EC 2.4.1). For instance, one of the processes represented by CPID 1 is a flavonoid glucosilation reaction catalyzed by Flavonol-O³-Glucosyltransferase (EC 2.4.1.81, Sutter and Grisebach 1973). Two of the CPIDs (5 and 9) correspond to the formation of methylene dioxy bridges catalyzed by oxidoreductases. One such enzyme relatively well studied is (S)-canadine synthase (EC 1.14.21.5, Rueffer and Zenk 1994). Two of the most commonly found cyclic pairs (CPID 3 and 4) correspond to flavonoid backbone biosynthesis pathways. CPID 3 corresponds to the formation of medicarpin from vestiton or 7,2'-dihydroxy-4'-methoxyisoflavanol, catalyzed by pterocarpin synthase (EC 1.1.1.246, Guo et al. 1994, Bless and Barz 1988). CPID 4 corresponds to the formation of flavanones from chalcones catalyzed by chalcone-flavanone isomerases (EC 5.5.1.6, Moustafa and Wong 1967). CPID8 and 12 correspond to the formation of γ - and δ -lactones, respectively, catalyzed by gibberellin oxidases (EC 1.14.11, Gilmour et al. 1987). CPID 15 corresponds to the formation of

Table 1. List of cyclic pairs shared by more than 50 species.

CPID	# of species	from		to		Reaction type
		CID	Cyclic structure	CID	Cyclic structure	
1	208	933		2432		flavonoid glycosylation
2	201	933		2430		flavonoid glycosylation
3	142	927		1221		dehydration
4	134	736		933		isomerization
5	126	1221		1956		methylene dioxy bridge formation (oxidation)
6	109	499		1217		possibly reverse process
7	109	2432		3673		flavonoid glycosylation
8	100	886		1367		ring closure through oxidation
9	94	1202		1959		methylene dioxy bridge formation (oxidation)
10	85	22		137		possibly independent processes
11	75	927		2433		flavonoid glycosylation
12	61	886		1543		ring closure through oxidation
13	56	933		2271		flavonoid glycosylation
14	55	2432		3520		flavonoid glycosylation
15	54	1099		1202		oxidative biradical coupling
16	54	1303		1803		reverse process, ring opening through reduction
17	51	2432		3501		flavonoid glycosylation

aporphine-type alkaloids from reticuline, through oxidative *o,o'*-coupling or *o,p'*-coupling of biradical intermediates (Battersby *et al.* 1971). To date, more than 700 different kinds of aporphine-type alkaloids are

isolated (Guinaudeau *et al.* 1994). CPID 16 is an example of the reverse processes, that is, the less complicated ring structure (ring system id. 1303) is formed from more complicated ring structure (ring

system id. 1803). Pinoresinol/lariresinol reductase (Min et al. 2003) catalyzes this ring opening reaction of pinoresinol to form lariresinol. For CPID 6, the more complicated sparteine-type ring system is known to form by the coupling of three cadaverines (Binnig 1974). Therefore, the cyclic pair is speculated to represent another reverse process or a pair of independently formed ring systems. CPID 10 is an example where the inclusive relationship of cyclic systems may not be indicating the reaction pairs. In this case, the ring

systems are too simple with the result that 85 identified pairs are likely be independently formed compounds and not representing the reaction pairs. For this kind of simple compound, algorithms using more detailed structural information, such as Kotera's method (Kotera et al. 2009) should provide a better prediction of corresponding reaction pairs.

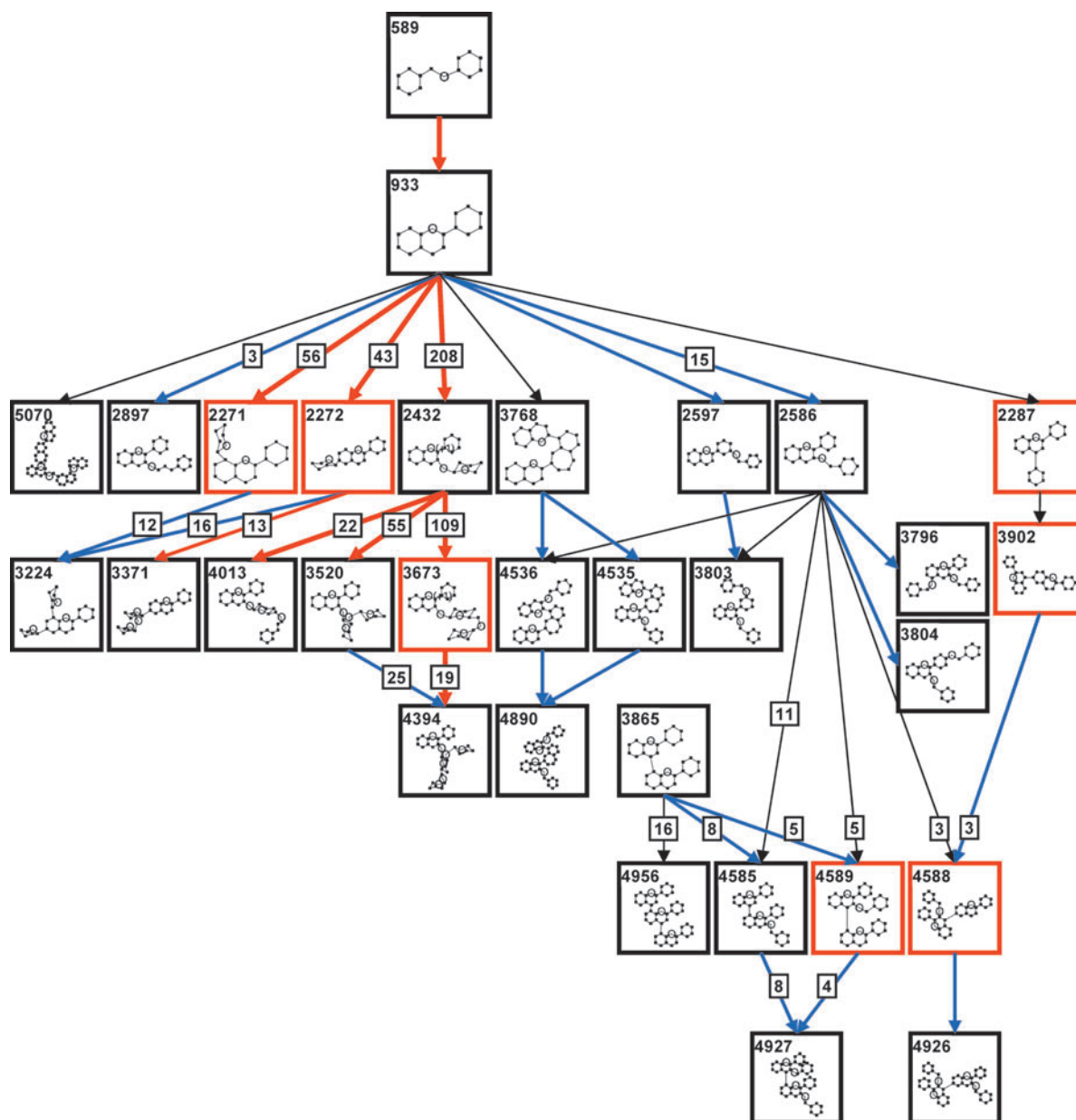


Figure 7. Out of 34,852 chemical structures contained in KNApSACk DB, we have identified 5,281 topologically unique cyclic structures, predicted 16,602 cyclic pair relations and extracted pathways concerning *Camellia sinensis*. Out of 35 types of unique cyclic structures in *C. sinensis*, 24 cyclic structures are associated with the largest cluster produced. To complement the results, intermediate cyclic structures are included if those are reported in the KNApSACk database but not yet associated to *C. sinensis*, indicated by the red surrounding boxes. Metabolic pathway relations supported by KEGG DB are shown by red arrows and explainable processes utilized for general modifications such as pyranosyl, caffeoyl, ferulyl, sinapyl and hydroxybenzoyl transferases are indicated by blue arrows. The number in a black box above an arrow indicates the number of species to which the concerned cyclic pair belongs calculated using information from the KNApSACk database.

Hierarchy of Cyclic pair relations in individual species

The hierarchy of cyclic pair relations in individual species can also be visualized using MetClassifier. As an example, we estimate the metabolic pathway of *Camellia sinensis* on the basis of inclusive relation of cyclic structures using data from KNApSAcK. Out of 35 types of unique cyclic structures in *C. sinensis*, 24 cyclic structures are associated with the largest cluster produced by MetClassifier (Figure 7). To complement the results, intermediate cyclic structures are included if those are reported in the KNApSAcK database but not yet associated with the targeted species indicated by red surrounding boxes in Figure 7, which are important for elucidating biosynthetic pathways while using an insufficient data set. Metabolic pathway relations supported by KEGG DB are shown by red arrows, and explainable processes utilized for general modifications such as pyranosyl, caffeoyl, ferulyl, sinapyl and hydroxybenzoyl transferases etc. are indicated by blue arrows in Figure 7. Therefore, it should be noted that MetClassifier effectively elucidates the metabolic pathways at the cyclic structure level because many of the cyclic pairs connected in Figure 7 correspond to reported metabolic pathways. In addition, when two metabolites corresponding to a predicted cyclic pair are available together in a number of species then it strengthens the validity of prediction because of their conserved nature across species. The number in a black box above an arrow in Figure 7 indicates the number of species to which the concerned cyclic pair belongs, calculated using information from KNApSAcK.

The bi- and triflavonoids constitute two major classes of complex C6-C3-C6 secondary metabolites. These compounds represent products of phenol oxidative coupling of flavones, flavonols, dihydroflavonols, flavanones, isoflavanones, auronones, auronols and calcones. Though the enzymes concerning to those processes have remained unknown, MetClassifier makes it possible to predict those processes, for example, CID 4956 may be derived from CID 3865 with the highest probability, and CID 4927 may be derived following a series of pathways involving CID 589, 933, 2586 and 4585. Thus, this method predicts metabolite synthesis pathways concerning cyclic substructures which are very relevant to bio-chemical reactions.

In the present paper, we focused on inclusive relations between cyclic substructures of substrate and product that are frequently observed in known metabolic pathways. We have also developed an algorithm to efficiently recognize inclusive relations and thus to predict metabolic pathways. Previously proposed MCSS-based approaches measure similarity between compounds by way of determining maximum common subgraphs, which is an NP-hard problem and hence their

computational cost is very high. On the other hand, our approach only check whether an inclusive relation exists or not between cyclic substructures by substructure search requiring low computational cost. Furthermore, we explain that the Tanimoto coefficient used in MCSS-based and fingerprint-based approaches is not suitable for pathway prediction, and we propose an algorithm that does not require selecting any threshold. Our method finds most similar cyclic substructure pairs on the basis of inclusive relation through recognizing parent subgraphs. By focusing on cyclic substructures and developing an algorithm without using any threshold, we achieve real-time operation for good pathway prediction while considering a large number of metabolites at a time.

Acknowledgements

This work was performed as one of the technology development projects of the "Green Biotechnology Program" supported by NEDO (New Energy and Industrial Technology Development Organization), Japan. This work was supported in part by a Grant-in-Aid for Science Research on Priority Areas, "Systems genomics," from the Ministry of Education, Culture, Sports, Science, and Technology of Japan and the BIRD project "Metabolome-Mass Spectral Database" from Japan Science and Technology Agency.

References

- Battersby AR, McHugh JL, Staunton J, Todd M (1971) Biosynthesis of the apparently "directly coupled" aporphine alkaloids. *Chem Commun* 985–986
- Binnig H (1974) The chemistry of sparteins. *Arzneimittelforschung* 24: 752–753
- Bless W, Barz W (1988) Isolation of pterocarpan synthase, the terminal enzyme of pterocarpan phytoalexin biosynthesis in cell suspension cultures of *Cicer arietinum*. *FEBS Lett* 235: 47–50
- De Luca V, St Pierre B (2000) The cell and developmental biology of alkaloid biosynthesis. *Trends Plant Sci* 5: 168–173
- Ellis L B M, Roe D, Wackett L P (2006) The University of Minnesota biocatalysis/biodegradation database: the first decade. *Nucl Acids Res* 34: D517–D521
- Gilmour SJ, Bleecker AB, Zeevaert JA (1987) Partial purification of gibberellin oxidases from spinach leaves. *Plant Physiol* 85: 87–90
- Guinaudeau H, Leboeuf M, Cave A (1994) Aporphinoid Alkaloids V. *J Nat Prod* 57: 1033–1135
- Guo L, Dixon RA, Paiva NL (1994) Conversion of vestiotone to medicarpin in alfalfa (*Medicago sativa* L.) is catalyzed by two independent enzymes. *J Biol Chem* 269: 22372–22378
- Hattori M, Okuno Y, Goto S, Kanehisa M (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* 125: 11853–11865
- Hostettmann K, Terreaux C (2000) Search for new lead compounds from higher plants. *Chimia (Aarau)* 54: 652–657
- Hou B K, Ellis L B M, Wackett L P (2004) Encoding microbial metabolic logic: predicting biodegradation. *J Ind Microbiol*

- Biotechnol* 31: 261–272
- Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl Acids Res* 28: 27–30
- Kotera M, Mcdonald AG, Boyce S, Tipton KF (2008) Eliciting possible reaction equations and metabolic pathways involving orphan metabolites. *J Chem Inf Model* 48: 2335–2349
- Langowski J, Long A (2002) Computer systems for the prediction of xenobiotic metabolism. *Adv Drug Delivery Rev* 54: 407–415
- Min T, Kasahara H, Bedgar DL, Youn B, Lawrence PK, et al. (2003) Crystal structures of pinoresinol-laricireinol and phenylcoumaran benzylic reductases and their relationship to isoflavone reductases. *J Biol Chem* 278: 50714
- Moustafa E, Wong E (1967) Purification and properties of chalcone-flavanone isomerase from soya bean seed. *Phytochemistry* 6: 625–632
- Oh M, Yamada T, Hattori M, Goto S, Kanehisa M (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J Chem Inf Model* 47: 1702–1712
- Ravasz E, Somera AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297: 1551–1555
- Rueffer M, Zenk MH (1994) Canadine synthase from *Thalictrum tuberosum* cell cultures catalyses the formation of the methylenedioxy bridge in berberine synthesis. *Phytochemistry* 36: 1219–1223
- Shinbo Y, Nakamura Y, Md.Altaf-Ul-Amin, Asahi H, Kurokawa K, Arita M, Saito K, Ohta D, Shibata D, Kanaya S (2006) KNApSACk: A comprehensive species-metabolite relationship database. *Biotchnol Agric Forestry* 57: 166–181
- Simmond M, Grayer R (1999) Plant drug discovery and development. In: Walton N, Brown D (eds) *Chemicals from Plants-Perspectives On Plant Secondary Products*, Imperial College Press, London, pp 215–249
- Sutter A, Grisebach H (1973) UDP-glucose: flavonol 3-O-glucosyltransferase from cell suspension cultures of parsley. *Biochim Biophys Acta* 309: 289–295
- Talafous J, Sayre L M, Mieyal J J, Klopman G (1994) META.2. A dictionary model of mammalian xenobiotic metabolism. *J Chem Inf Comput Sci* 34: 1326–1333
- Tohsato Y, Nishimura Y (2008) Metabolic Pathway Alignment Based on Similarity between Chemical Structures. *IPSI Transactions on Bioinformatics* 48: 736–745
- Verpoote R (1998) Exploration of nature's chemodiversity: the role of secondary metabolites as leads in drug development. *Drug Discov Today* 3: 232–238