Eliciting Possible Reaction Equations and Metabolic Pathways Involving Orphan Metabolites

Masaaki Kotera,* Andrew G. McDonald, Sinéad Boyce, and Keith F. Tipton

School of Biochemistry and Immunology, Trinity College, Dublin 2, Ireland

Received June 25, 2008

The development of metabolomics has resulted in the discovery of an increasing number of orphan metabolites, which are defined as compounds that are known to be present in living organisms but whose synthetic/degradation pathways are unknown. In this paper, we describe a procedure for identifying possible products and/or precursors of such orphan metabolites and for suggesting complete reaction equations and the corresponding EC (Enzyme Commission) number simultaneously. Chemical structure comparison is performed for a pair of compounds consisting of a reported substrate and its corresponding product and also for pairs of randomly selected compounds. Possible combinations of compounds registered in the KEGG database were used for generating putative enzyme reaction equations, which resulted in 77% of the reported equations being generated, as most of the remainder represent classes of compounds, rather than specific compounds, or contain Markush structures. The quality was checked using chemical structure comparison and the random-tree method, which gave 98% accuracy in suggesting EC subsubclasses for reported equations in cross-validation tests. The equations generated in this study can be seen using the Web-based program GREP (Generator of Reaction Equations & Pathways; http://bisscat.org/GREP/). The usefulness of our method for constructing possible metabolic pathways was demonstrated by mapping the generated equations for several groups of compounds, such as the betalain alkaloids. The possible development of our method so that alternative substrates for reported enzymes can be found and for annotating enzyme functions in genomic research is also discussed.

INTRODUCTION

Metabolomics can be used to identify compounds of unknown function that have an unknown synthesis/degradation pathway. By analogy with orphan genes in genome annotation, orphan metabolites may be defined as chemical compounds that are known to be present in a living organism but where it is not known how they are synthesized or degraded.[1] Poolman et al.[2] adopted a somewhat broader definition, using the term 'orphan metabolites', to include those metabolites that are involved in only one reaction.

Many orphan metabolites have been identified in research on plants and fungi, where they are often referred to as secondary metabolites. A secondary metabolite can be defined as a compound that is not directly involved in the normal growth, development, or reproduction of an organism and whose absence is not normally lethal. Secondary metabolites would be largely unnecessary for survival if the plants/fungi containing them never encountered predators, competitors, or environmental stress. Some secondary metabolites are known to function as toxins (poisons and venoms), defending the plant against predators, parasites, and diseases, providing a growth advantage over other species, and/or facilitating the reproductive processes (*e.g.* acting as coloring agents or in the production of attractive odors, *etc.*). Some of the most well-known secondary metabolites include the antibiotic penicillin (from the fungus *Penicillium notatum*), cocaine, which is present in the coca plant *Erythroxylon coca,* and tetrahydrocannabinol, which is from the plant *Cannabis sativa*. Such definitions of secondary metabolites were mooted at a time when little was known about these substances. In recent times, the functions of such compounds are becoming increasingly understood. For example, it is now known that tetrahydrocannabinol protects the plant from herbivores or pathogens and also from harmful UV−B irradiation.[3]

Secondary-metabolite profiling (chemotaxonomy) has been shown to be of great value in the classification and differentiation of fungal species.[4,5] Because some secondary metabolites, such as antibiotics, are of importance in drug development, the metabolism of such compounds has been investigated and is now better understood.[6] Despite this, many orphan metabolites remain, and metabolomic studies are resulting in a steady increase in the number of such compounds across all species. It is sometimes hard to identify enzyme activities involved in secondary metabolism because the enzymes acting on those metabolites often have low activities, generally three to 5 orders of magnitude lower than those of enzymes involved in primary metabolism, at least in plants.[7]

Knowledge of the functional groups present in chemical compounds can be used to find possible enzymes that act upon such orphan metabolites. These enzymes can be identified through EC (Enzyme Commission) numbers,[8] which are the most widely accepted classification scheme for reported enzyme reactions. Some enzymes act on a group of compounds that have a specific type of functional group, but the breadth of their substrate specificities is often

* Corresponding author e-mail: kot@kuicr.kyoto-u.ac.jp.

10.1021/ci800213g CCC: $40.75
Published on Web 11/21/2008

unknown. The BiSSCat (Biochemical Substructure Search Catalogue) database of chemical compounds, functional groups, and substructures (http://www.bisscat.org) was developed for the identification of chemical compounds that contain specified functional groups and substructures.[1] Although this system can also be used to predict possible enzymes that may act on a specified compound, there is no certainty about such predictions.

Here we report a substructure-based approach that can be used to identify possible products and/or precursors for orphan metabolites and to generate a plausible reaction. Searches are carried out to find compounds that are structurally related to a chosen target compound, and the structural differences are then checked to determine which of these has the potential to be a product (or precursor) of the target compound in an enzyme-catalyzed reaction. The process involves checking for chemical bonds or functional groups that are directly involved in the putative reaction and evaluating whether or not the remaining substructure(s) could be converted into, or derived from, known compounds, followed by construction of a putative reaction equation (which we refer to as a "generated equation") using a combination of reported compounds. If such an equation is found to be realistic, the EC subsubclass can then be assigned automatically. In our approach, the probability of a reaction occurring *in vivo* and the assignment of an appropriate EC subsubclass are performed simultaneously using a random forest data-mining and analysis method. Clearly, such an approach cannot guarantee that a predicted enzyme reaction actually occurs in any given species or tissue, but it should serve as a soundly based guide to the probable metabolism of orphan metabolites and to the design of rational experimental studies.

MATERIALS AND METHODS

Chemical-Compound and Enzyme-Reaction Data. The KEGG REACTION database (http://www.genome.jp/kegg/reaction/) was used as the initial data source for these studies. It contains all known enzyme-catalyzed reactions, taken from the IUBMB Enzyme Nomenclature list (http://www.enzyme-database.org/)[9,10] and also from the metabolic-pathway section of the KEGG PATHWAY database. In the KEGG release of February 2008, there were 7521 reactions, which included 3222 IUBMB reactions.[11] The KEGG database contains many orphan metabolites, most of which are phytochemical compounds. In the 2008 version, there were 5310 orphan metabolites in total, comprising 203 sesquiterpene lactones, 121 isoquinoline alkaloids, 119 indole alkaloids, 104 diterpenoids, 100 sesquiterpenoids, 87 flavones and flavonols, 86 quinones, *etc*.

Physicochemical Properties of Chemical Compounds. Using our methodology, we consider each chemical structure as a labeled graph with atoms as its nodes and covalent bonds as edges. In this context a graph is defined as a mathematical structure to model pairwise relations between objects.[12] Hydrogen atoms are excluded as nodes in the graph structure. All atoms except hydrogen atoms are distinguished by their elements and by their electrostatic and physicochemical properties, most of which are based on the "programmable atom-typer" program, PATTY.[13] Physicochemical properties are provided for each non-hydrogen atom rather than for the total structure of the chemical compound. For example, the PATTY method deconstructs the hydrophilic molecule ethanol (CH_3CH_2OH) into the "hydrophobic" ethyl group (CH_3CH_2-) and the "polar" oxygen atom of the hydroxy group (-OH). This approach was used for the classification of functional groups and substructures in our previous research[1] and is used here to carry out chemical-structure comparisons.

Functional Groups and Substructures. The BiSSCat database contains a collection of functional groups and substructures that are found in biochemical compounds.[1] Each substructure is computationally defined as a graph object, with non-hydrogen atoms and bonds described as nodes and edges, respectively. Each substructure is distinguished from others in terms of its elements (C, N, *etc*.), electrostatic and physicochemical properties (as described above), and topology. Functional groups found in reported enzyme-catalyzed reactions are also assigned. In this new approach, the substructure and functional-group data are used in a decision-tree process (described below) to determine the probability of a reaction involving an orphan metabolite actually taking place.

Compound Pairs, Reactant Pairs, Reported Pairs, and Unreported Pairs. Since an enzyme-catalyzed reaction usually involves multiple substrates and products, and the number of substrates and products involved can vary, it is useful to decompose each reaction into one or more substrate-product pairs. Each of the binary pairs with common structural moieties is defined as a reactant pair[14] and is registered in the RPAIR section of the KEGG database.[15] In this research, we designate those reactant pairs that occur in the KEGG database as "reported" reactant pairs. The February 2008 version of the KEGG database contains 15,050 compound entries, so the number of all possible combinations of all compounds is 113,243,725. Only a small proportion of these are reported reactant pairs (7342 pairs are registered in the RPAIR database, of which, 5189 are assumed to be of the "main" type, defined by Kanehisa et al.[15] as those that cannot be ignored in drawing metabolic pathway maps). The remaining compound pairs, which constitute the vast majority of pairs, are considered to be unreported, or generated, pairs. Most of these will be chemically unrealistic, but some are likely to be genuine reactant pairs that have not yet been reported. Since the aim of this approach was to identify possible reactions involving orphan metabolites, and in order to reduce the computational time, we excluded 38,476,754 chemical structure comparisons (explained later) because their maximum common substructures have less than five non-hydrogen atoms, based on their compositional formulas. 35,636,614 of these are pairs that have at least one compound having less than five non-hydrogen atoms, such as H_2O, CO_2, *etc*. As discussed later, exclusion of such simple compounds at this stage is not a fundamental limitation because they may be introduced later at the reaction-balancing stage. A further 2,840,140 of the pairs were excluded because they have less than five common non-hydrogen atoms, such as the pair C_3H_7NS and $C_6H_7O_4P$. Thus, the number of compound pairs to be considered in the chemical structure comparison process was 74,766,971. A chemical structure comparison is not feasible with such a large data set, since our graph comparison method is strict and does not use any heuristics. Therefore, the comparison

Scheme 1. Pseudocode for Score-Matching an Associated Node

```
Nodescore(V1, V2)
{
        Initial score = 1.0
        If elements (carbon, nitrogen, etc.) are different, return (0.0).
        If electrochemical properties are different, 0.1 is subtracted from the score.
        If orbitals (sp3, sp2, and sp) are different, 0.1 is subtracted from the score.
        If numbers of adjacent non-hydrogen atoms are different,
                the number of changes x 0.1 is subtracted from the score.
        If aromaticity is gained or lost, 0.1 is subtracted from the score.
        If a ring structure is created or broken, 0.1 is subtracted from the score.
        Return (score).
}
```

was performed first for all combinations of a substrate and a product that were found in reported reaction equations. After that, the remaining combinations of compound pairs were randomly subjected to the comparison methods, and the data were stored for analysis.

Chemical Structure Comparison. Comparing chemical structures is essential for determining whether or not any given pair of chemical compounds corresponds to a substrate and its product. This problem is handled as a two-dimensional graph-comparison problem. After 2-D chemical structure comparison is performed, the neighboring atoms belonging to the two compounds as well as the configuration of chiral centers are compared. It was not considered necessary to consider *cis/trans* isomerization in this aspect of the current study. If there is more than one stereochemical difference between the pair, they are regarded as being stereochemically improbable and are discarded. Since the purpose of the chemical structure comparison is to determine whether or not one compound can be converted into another in a single reaction, the compound pair was discarded if the conversion would involve more than two intermolecular chemical bonds being joined/cut. The approach we use, which is to find the maximum common substructure between a pair of compounds, is a modified version of the traditional association graph method.[16−18] In this application, the association graph (AG) of vertices and edges (V,E) for two compound graphs $G_1(V_1,E_1)$ and $G_2(V_2,E_2)$ is a new graph defined on the node set $V = V_1 \otimes V_2$ (a Cartesian product of V_1 and V_2) and the set of edges $E = V \otimes V$, where the associated nodes are adjacent if the original nodes are adjacent in both original graphs G_1 and G_2. The association graph AG possesses all possible node matches between two original graphs G_1 and G_2. Associated nodes are given matching scores by the procedure outlined in Scheme 1. The score of an edge in the association graph is a product of the scores of the two associated nodes. The association graph is then subjected to a maximum spanning tree search with a priority queue based on the edge score. This process is reiterated with every associated node being used as the start node. The tree with the highest score is then selected. This results in a single connected maximum common substructure for the two compounds.

In this paper, the maximum common substructure between two compounds A and B is described as A∩B. For example, "A∩B = 12" means that 12 non-hydrogen atoms are common between A and B. B\A and A\B apply when B or A, respectively, contain additional substructures that are not present in the other reactant. Taking the pair of compounds D-glucose (A) and D-glucose 6-phosphate (B) as an example, A∩B = 12 (corresponding to the C_6O_6 of the glucose residue), B\A = 4 (corresponding to the phosphate residue), and A\B = 0. When A∩B ≥ 5 for a compound pair, the Tanimoto (also known as the Jaccard) coefficient[19−21] was used as a method for scoring chemical structure-comparison results. This coefficient of groups A and B is the number of common members (A∩B) divided by the number of nonredundant members, A∪B (A∩B + B\A + A\B). This score ranges from 0 to 1, where 0 represents the absence of any common member and 1 applies when the two groups are identical. As an example, the Tanimoto coefficient for the pair D-glucose and D-glucose 6-phosphate is A∩B/A∪B = 12/(12 + 4 + 0) = 0.75.

Generating Possible Reaction Equations. The basic aim when generating a reaction equation from a set of compound pairs is to reverse the procedure used to deconstruct a reaction equation into a set of reactant pairs.[14] Two steps are involved in generating the equations. In step 1, the required number of compound pairs are assembled to produce a partial equation (as shown in Scheme 2). The partial equation is then classified in terms of the number of compound pairs involved, and a check is made to see which and how many atoms are missing on either side of the reaction equation. In step 2, the partial equations are compared to reported reaction equations, and then appropriate sets of additional compounds (including cofactors) are introduced to produce a balanced equation. For example, the partial equation "D-gluconolactone = D-glucose" is not balanced as 2 additional hydrogen atoms are required on the left-hand side of the equation for there to be the same number of each type of atom on each side of the equation. As a result, a donor/acceptor pair, such as "NAD(P)H + H^+ = NAD(P)$^+$" or "H_2O_2 = O_2", would need to be added. Similarly, "*S*-adenosyl-L-methionine = *S*-adenosyl-L-homocysteine" would be a possible balancing pair for a methyltransferase (EC 2.1.1.-) reaction. Clearly, this procedure can generate more than one balancing pair for some reactions, as in the oxidoreductase example above. Which balancing pair is used may depend on the species under consideration, but that would be a matter for experimental determination. Such generated equations are then subjected to the decision-tree method.

Decision-Tree Method. The decision-tree method, which is used as one of the classification methods in data-mining, was used to estimate the quality of the reaction equations generated. This method iteratively scans each attribute of a

group of entities to find the best branching criteria for the most accurate prediction, resulting in a multibranched tree-like data structure. This method should fit well to the EC classification scheme, which is also a tree-like data structure. For example, the presence or absence of water in a reaction equation is a good indicator (but not always correct) when deciding whether or not the equation corresponds to EC 3 (hydrolases).

The Gini coefficient is sometimes introduced in decision trees to decide which attribute is used to branch the group of entities.[22,23] This coefficient measures the impurity or inequality of a group and has been used in many different fields, including economics,[22,24,25] clinical science,[26,27] bacteriology,[28] and enzyme research.[29] Subtraction of the Gini coefficient (G) from 1

$$1 - G = \Sigma(n_i/n)^2 \quad (1)$$

where n = total number of entities, and n_i = number of the entities with i-th attribute, can be understood as a probability of two randomly selected entities in a group having the same attribute. The Gini coefficient ranges from 0 to 1, in which 0 means "pure" (all entities of the group have the same attribute) and 1 means "totally mixed" (every entity has a unique attribute).

In this study, entities are enzyme reaction equations, both reported and generated, and the targeted attribute to be classified is the EC subsubclass (up to the third digit of the EC number). Since there are cases where an equation may correspond to more than one EC number or subsubclass, we

Scheme 2. Pseudocode for Generating Partial Reaction Equations

```
GeneratePartialReactions(data)
{
  1.  Iterate for each starting compound A
  2.    Iterate for each compound B, where A≠B and A∩B ≥ 5
  3.      Function1 (A, B)  // to generate "A = B"
  4.      Iterate for each compound C, where A≠C, B≠C and A∩C ≥ 5
  5.        Function2(A, B, C) // to generate "A = B + C"
  6.          If B\A ≥ 2 and C\A ≤ 1, then
  7.            Iterate for each compound D, where A≠D, B≠D, C≠D and B∩D ≥ 5
  8.              Function3(A, B, C, D) // to generate "A + D = B + C"
  9.          If B\A ≤ 1 and C\A ≥ 2, then
 10.            Iterate for each compound D, where A≠D, B≠D, C≠D and C∩D ≥ 5
 11.              Function4(A, B, C, D) // to generate "A + D = B + C"
}
// function to generate a partial reaction "A = B" with a pair AB
Function1(A, B)
{
 12.  If A\B ≤ 1 and B\A ≤ 1, then the pair AB is approved (go to step 2).
 13.  If A\B ≤ 4 and B\A = 0, then the pair AB is approved (go to step 2).
 14.  If A\B = 0 and B\A ≤ 4, then the pair AB is approved (go to step 2).
 }
// function to generate a partial reaction "A = B + C" with pairs AB and AC
Function2(A, B, C)
{
 15.  If A∩B∩C ≥ 2, the pairs (AB & AC) are rejected.
 16.  If A\B\C ≥ 2, the pairs (AB & AC) are rejected.
 17.  If B\A ≥ 2 and C\A ≥ 2, the pairs (AB & AC) are rejected.
 18.  If B\A ≤ 1 and C\A ≤ 1, the pairs (AB & AC) are approved (go to step 2).
}
// function to generate a partial reaction "A + D = B + C" with pairs AB, AC and BD
Function3(A, B, C, D)
{
 19.  If A∩B∩D ≥ 1, the pairs (AB & AC & BD) are rejected.
 20.  If B\A\D ≤ 1 and D\B, the pairs (AB & AC & BD) are approved (go to step 2).
}
// function to generate a partial reaction "A + D = B + C" with pairs AB, AC and CD
Function4(A, B, C, D)
{
 21.  If A∩C∩D ≥ 1, the pairs (AB & AC & CD) are rejected.
 22.  If C\A\D ≤ 1 and D\C, the pairs (AB & AC & CD) are approved (go to step 2).
}
```

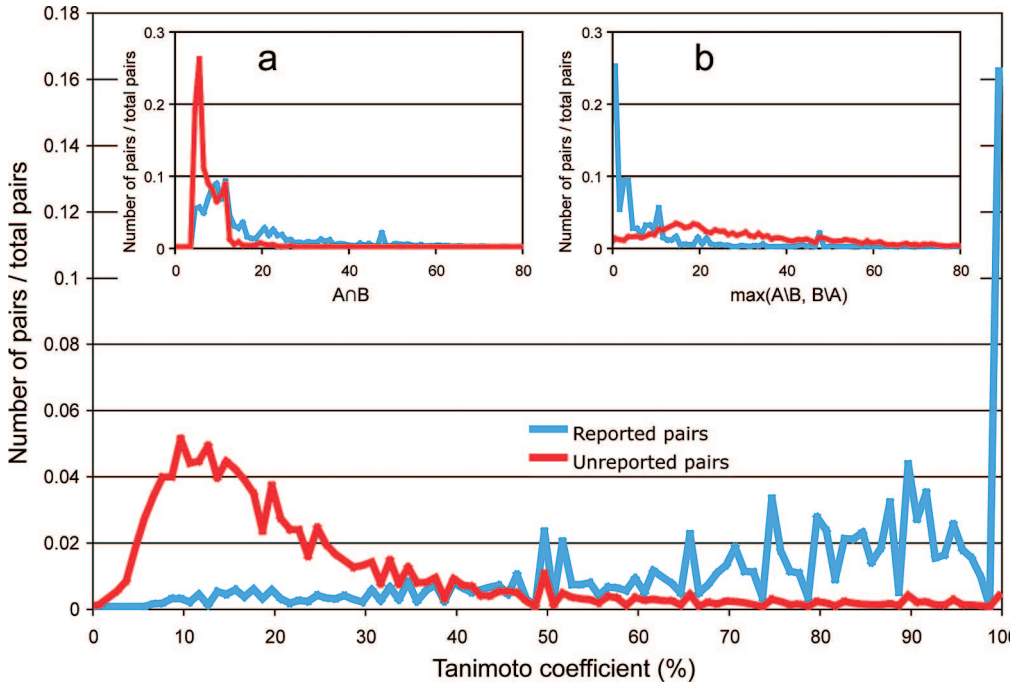

Figure 1. Distributions of the Tanimoto coefficient for reported and unreported reactant pairs. The vertical axis is normalized to the number of pairs/total number of pairs for the comparison of the two reactant pair types: reported and unreported. See text for the definition of the Tanimoto coefficient used in this study. The inserts show the distribution of the number of pairs/total pairs with the value of (a) A∩B and (b) the maximum value (A\B, B\A).

introduced a modified version of the Gini coefficient for our decision trees

$$G' = \Sigma(n_i/n)^2 \qquad (2)$$

where an entity can have more than one attribute. If all entities have only one attribute each, this coefficient ranges from 0 to 1, where 1 means "pure" and 0 means "totally mixed" (the reverse of the Gini coefficient). If the group contains entities with multiple attributes, this coefficient can exceed 1. A number of discrete questions concerning compositional formulas, chemical bonds, functional groups or substructures, and preserved atoms were applied to generate the tree, as shown in Table 1. These answers were not predefined prior to the procedures but were automatically generated in each iteration. The data are scanned, and the branching process based on the decision-tree results is repeated for each branched group of equations until the G' value no longer increases.

Random Forest. The random-forest method simultaneously executes many decision trees using random subsets of the target data. In order to classify a new entity, the entity is subjected to each of the trees in the forest, and the forest "chooses" the classification, based on a majority vote. In this approach, all of the equations registered in the KEGG database were considered to be positive examples. Truly negative examples are difficult to identify, because it is rare to find confirmed negative data in published form. Initially, we randomly assigned 10% of the generated reaction equations (after filtering out the reported equations) to the set of negative reaction equations. Similar approaches have been taken in several machine-learning methods that include decision trees.[30,31] The overall procedure used for random-forest generation in this study was as follows:

1. Reported equations were randomly divided into two groups: 95% were given the label of their own EC subsubclasses for generating a tree (training set), and 5% were given no label and were used for testing.

2. Unreported equations were randomly divided into two groups: 10% were given the label "*null*" instead of being assigned to an EC subsubclass for negative examples in a tree, and 90% were given no label.

3. A decision tree was then generated under these conditions.

4. Steps 1−3 were iterated 100 times.

RESULTS

Chemical Structure Comparison. Chemical structure comparisons have been completed for 2,502,333 compound pairs, of which 2,178,913 (87%) had A∩B ≤ 4 (these pairs are ignored in the following procedures), and 4145 (0.17%) were reported pairs. As shown in Figures 1−3, reported pairs of the "main" type and unreported pairs appeared to have significantly different characteristics, which can be used as clues in the identification of unreported but possible reactant pairs. The Tanimoto coefficient for reported pairs is generally greater than 50%, whereas for unreported pairs it is generally less than 50% (Figure 1). The insets in Figure 1 (a and b) show the distributions of A∩B and max(A\B, B\A), where max(A\B, B\A) is the larger number of A\B and B\A. These are clearly not as helpful as the Tanimoto coefficient in terms of identifying reactant pairs. In contrast, the distributions of min(A\B, B\A), corresponding to the smaller number of A\B and B\A, shown in Figure 2, can be a useful indicator of unreported and unlikely reactant pairs, which should be discarded, since only 1.6% of the reported pairs have a min(A\B, B\A) value ≥2. The majority (84%) of reported pairs have a min(A\B, B\A) value = 0, meaning that no groups other than hydrogen atoms are substituted when a bond is broken. Of the reported pairs, 14% have min(A\B,

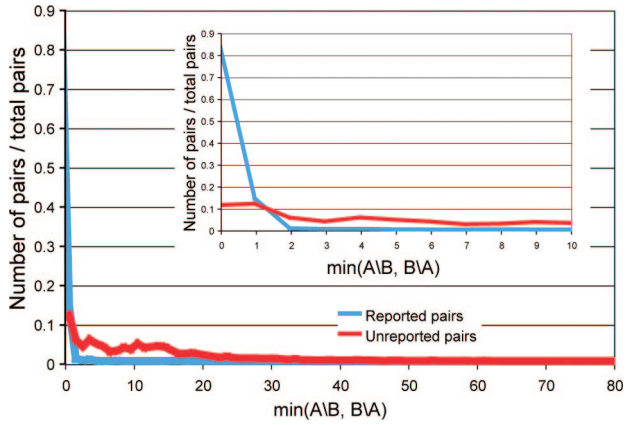

Figure 2. Distributions of min(A\B, B\A) for reported and unreported pairs. The inset is the expanded view of Figure 2. The vertical axis is normalized to the number of pairs/total pairs for the comparison of the two types of pairs. See Method for a definition of the min(A\B, B\A) value.

B\A) = 1, meaning that a group containing one non-hydrogen atom (such as -OH, =O, or -NH_2) is added when a bond is broken.

The number of chemical bonds created or destroyed is also essential for assessing the likelihood of a compound pair being a reactant pair. When we focus on the graph topology, the changes can be grouped into three types: (1) intermolecular joining/cutting, (2) ring formation/opening (intramolecular joining/cutting), and (3) hydrogenating/dehydrogenating. The numbers associated with each of these are written as a comma-separated triplet. For example, "*0,0,0*" means that none of these changes occur. This type of result would be found for racemase and epimerase reactant pairs (EC 5.1) as well as for those involved in *cis*−*trans* isomerase (EC 5.2) reactions. Most reactant pairs within the lyase class (EC 4) have one intermolecular bond that is broken with the concomitant production of an unsaturated bond, which corresponds to the triplet "*1,0,1*". Figure 3 shows the distribution of these three types of changes for reported and unreported pairs. The largest peak of reactant pairs corresponds to "*1,0,0*", in which one non-hydrogen−non-hydrogen bond is cut or substituted for another bond, corresponding to a variety of pairs taken mostly from EC 2 (transferases), EC 3 (hydrolases), and EC 6 (ligases). Most reactant pairs from EC 1.2 (oxidoreductases acting on C=O group) also correspond to "*1,0,0*" because of the additional involvement of water. The second largest peak for reported pairs corresponds to "*0,0,1*", in which one bond is hydrogenated/dehydrogenated (-CH_2OH ↔-CH=O, -CH_2−CH_2- ↔ -CH=CH-, *etc.*). Most of the oxidoreductases (EC 1.1, EC 1.3, EC 1.4, *etc.*) fall into this category. Most unreported pairs have two intermolecular joining/cutting points (with the largest peak corresponding to the triplet "*2,0,0*"). Reported pairs with this configuration include reactant pairs from among the oxygenases (EC 1.13). Mutases (EC 5.4) also fall into this category in the current study, as discussed later. Others, however, are thought to result from a two-step reaction or a misleading result caused by difficulties inherent in the maximum common subgraph-finding algorithm, as discussed below.

Our method addresses an order N^2 problem, so possible ways to avoid unnecessary calculations, without imposing arbitrary restrictions, should be considered. Selecting only compounds of similar size or with a similar number of carbon atoms might seem to be a reasonable way of reducing the data set to be compared, but it results in the loss of significant numbers of possible pairs (as shown in Figures 1 and 2), making it untenable. Selecting compounds within the same class also causes problems, for example, monosaccharides can be transferred from UDP-monosaccharides to a variety of compounds but those products are no longer classified as "monosaccharides". Genomics and proteomics have also been faced with the N^2 problem, leading to the development of a database containing frequently occurring motifs or precalculated alignment results. The development of a similar type of database containing multiple chemical substructures might ease this problem in the future.

Receiver Operator Characteristics (ROC) curves[32] plot the false-positive rate against the true-positive rate for different cutoff values. All compound pairs were classified according to the scores from Figures 13, and the cumulative true-positive (reported pairs) rate/false-positive (unreported pairs) rate was calculated in descending order of reported pairs/all pairs, as shown in Figure 4. This suggests that the Tanimoto coefficient and the number of chemical bonds created/destroyed provides an effective way of identifying reactant pairs. The ROC curve is improved when the Tanimoto coefficient, min(A\B, B\A), and the number of chemical bonds created/destroyed are considered simultaneously and should be improved still further when the whole reaction equation is taken into account.

Generated Equations. Our method has generated 133,030 reaction equations so far, of which, 2748 were reported in the latest version of the KEGG database. 1105 of the 5310 orphan metabolites were assigned to more than one generated reaction. KEGG includes some chemical compound entries with partial structural data or no structural data. Some of these compounds are large polymers such as proteins, glycans, and polynucleotides. Others represent a class of molecules rather than a specific compound, *e.g.*, "monoterpenyl diphosphate + H_2O = monoterpenol + diphosphate" (EC 3.1.7.3: monoterpenyl-diphosphatase). Reported reaction equations containing such compounds must, at least for the present, be excluded in the estimation of the coverage. The number of reactions remaining that were suitable for this study was 3581, representing coverage of approximately 77%. Equations were generated using the 2006 version of the KEGG database and were then compared to the latest (2008) version. It was hoped that this would validate and demonstrate the usefulness of this method for future prediction of unknown reactions. Newly added equations that contain newly registered compounds are, of course, not appropriate for this test. However, 100 of 136 (∼74%) new equations that contained no newly registered compounds were accurately generated by our procedure. Thus, the system summarized in Scheme 2 can successfully predict approximately 74−77% of reaction equations. Of the reaction equations that could not be generated, the largest number (46%) contained unbalanced equations (reactions for which only the main substrate and product are known) and therefore would be impossible to generate using any automated procedure. As discussed later, relaxing the constraints applied to the system could address many of these and other cases.

Equations are classified differently each time a decision tree is generated, and a tree is generated in such a way that

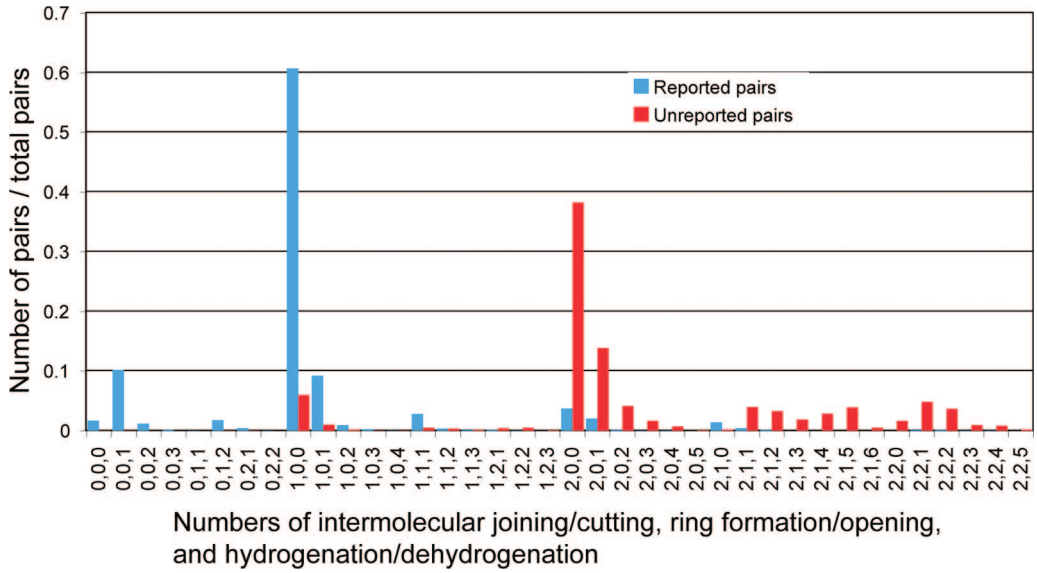

Figure 3. Distributions of the number of chemical bonds created/degraded for reported and unreported pairs. The numbers of three types of changes (intermolecular joining/cutting, ring formation/opening, and hydrogenation/dehydrogenation) are described by comma-separated triplets.

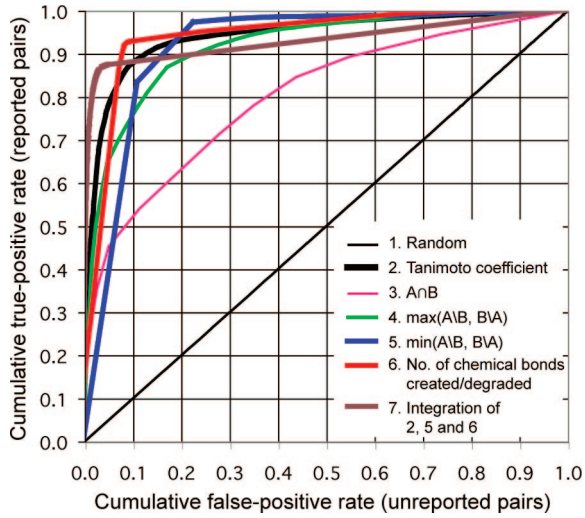

Figure 4. ROC curves of reported pairs against unreported pairs. If pairs are selected randomly from a group containing true-positive examples (reported pairs) and false-positive examples (unreported pairs), the proportion of selected true examples in total is expected to be the same as that of false examples (Line 1). A good criterion would enable one to select the highest number of true (reported) pairs with the lowest number of false pairs generated.

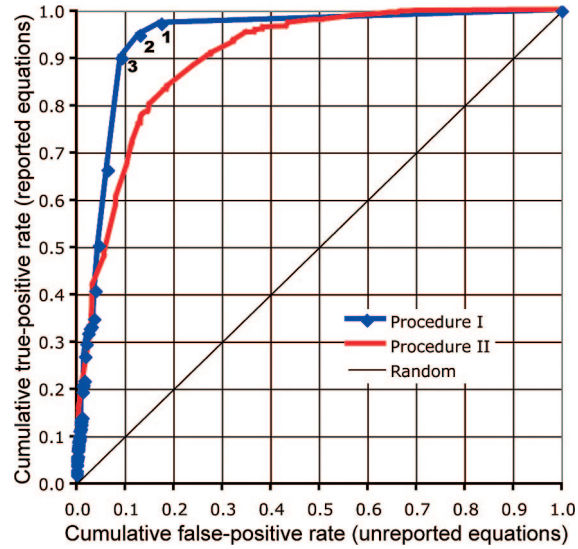

Figure 5. ROC curves of reported equations against unreported equations. Different procedures are applied to decide the likelihood of an equation occurring (see text).

nulls are separated as completely as possible from true examples. As "*null*" equations are selected randomly from unreported equations and cannot necessarily be regarded as truly negative, the majority (78%) of them were mingled with true examples, while the minority (22%) of nulls fell into exclusive groups (*i.e.*, nulls plus other unreported reactions). ROC curves are drawn for reported equations against unreported equations in order to estimate the likelihood of generated equations genuinely occurring (Figure 5). Two different evaluation procedures were applied for all iterations. Procedure I determines and applies the probability of an equation being a member of a group containing no null. Procedure II determines the average number of true examples divided by the number of (true examples + null). As shown in Figure 6, procedure I is more effective at excluding unreported equations. The points 1, 2, and 3 in Figure 5 correspond to the probabilities of an equation being a member of a group containing no null being more than 1, 2, and 3%, respectively. These results suggest that a generated equation can be regarded as being possible if it has even a small (≥3%) chance of being classified into a group without a null. Thus, the value of 3% was used as the threshold for a reaction to be regarded as being realistic and was applied in the following procedures.

Cross-validation tests were performed in order to estimate the accuracy of assigning EC subsubclasses to the generated equations. The probability of assigning correct EC subsubclasses to test sets in each decision tree varies widely, from 0.65 to 0.92 (average 0.79, standard deviation 0.048). Better predictions were obtained using one of the "majority vote" methods from the decision trees in a random forest. Two different procedures were tested. In the first procedure, the average number of each subsubclass in a group of each tree was calculated for each equation that was not used for

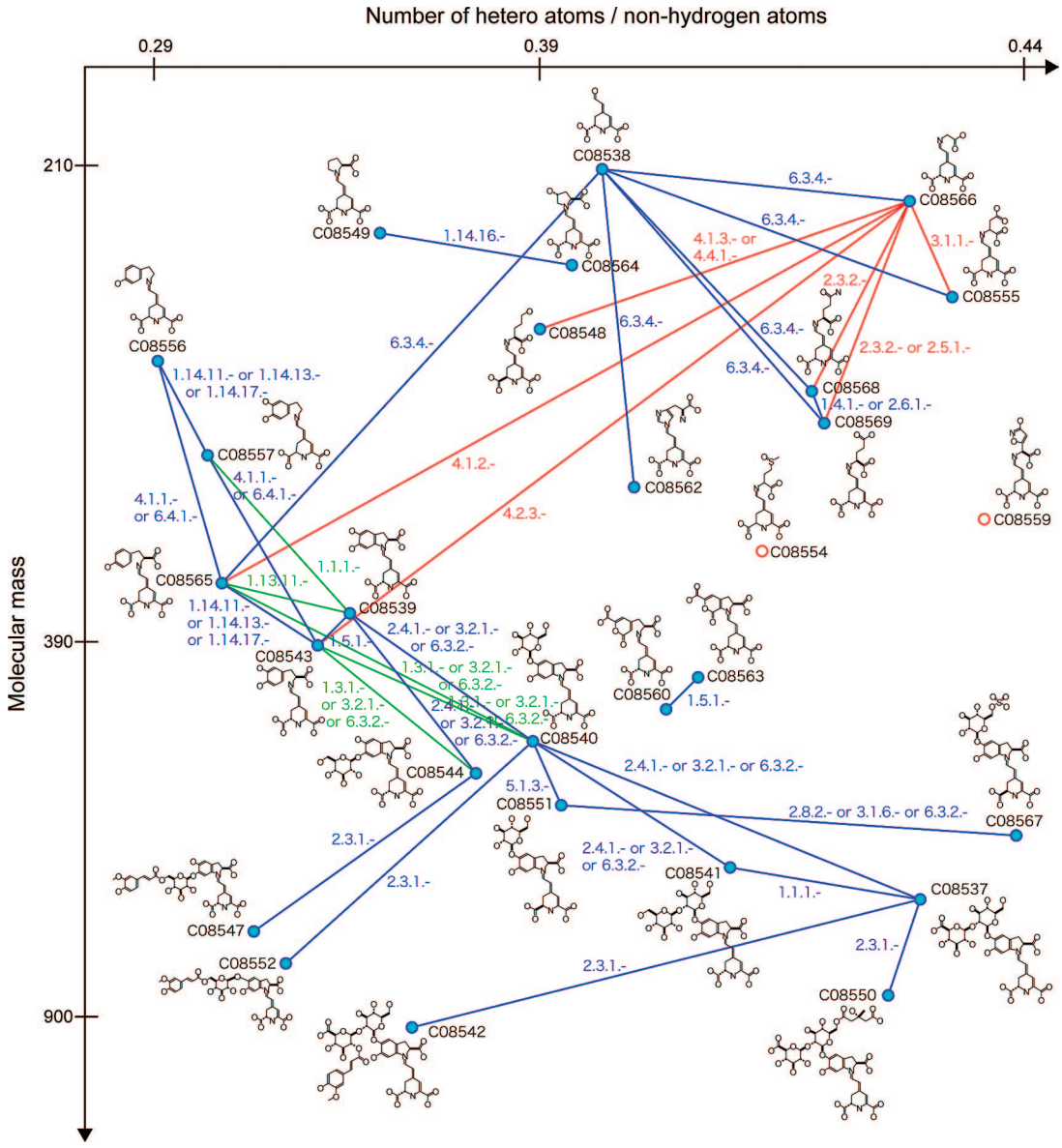

Figure 6. Generated metabolic map of betalain alkaloids. C-numbers are identifiers of compounds registered in the KEGG database. Vertical and horizontal axes represent approximate molecular mass and the number of heteroatoms (O, N, S, P, *etc*.) divided by the number of non-hydrogen atoms, respectively. Each circle represents a compound, the exact location of which may be slightly moved manually so that the lines and the circles do not overlap with each other. Blue circles are orphan metabolites involved in the generated equations in this study, and red circles represent those compounds that are not involved in any equations, either reported or generated. Each line between the circles represents the suggested reactant pair in this study. Of those, red lines correspond to pairs suggesting joining/cutting of a C−C bond, and green lines correspond to pairs suggesting more than one joining/cutting point.

generating the tree, and subsubclass was determined by "majority vote". In the second procedure, the probability of each subsubclass being present was calculated, and the subsubclass was determined by "majority vote". The resulting accuracy values were 0.73 and 0.98, respectively. The extraordinarily high accuracy of the second method was due to the robustness of random-forest iteration and the avoidance of the bias associated with the varying numbers of reactions in the different EC subsubclasses. The list of the possible equations generated in this study can be found at the GREP Web site.

Generated Pathways. One way to find possible metabolic pathways among the large number of generated equations would be to focus on a specified group of compounds. Figure 6 shows an example of generated metabolic maps that are well connected and relatively simple. Although they all share a common substructure, searching for the remaining substructure (Scheme 2) reveals only those pairs that may be directly connected in single reactions. Several compounds have many edges associated with them. These are thought to be important compounds in the metabolic system. Of these, betalamic acid (C08538) has the smallest molecular mass and appears to be a precursor of many different derivatives. Our method suggests that portulacaxanthin III (C08566), miraxanthin-II (C08555), vulgaxanthin-I (C08568), ulgaxanthin-II (C08569), musca-aurin VII (C08562), and portulacaxanthin II (C08565) are all synthesized from betalamic acid by ligase (EC 6.3.4) reactions that add common amino acids (glycine, aspartate, glutamine, glutamate, histidine, and tyrosine, respectively).

Despite the probability that the generated reactions actually occur, Figure 6 includes cases where likely paths are not generated. If all reactions starting from betalamic acid were to occur, it is reasonable to assume that indicaxanthin

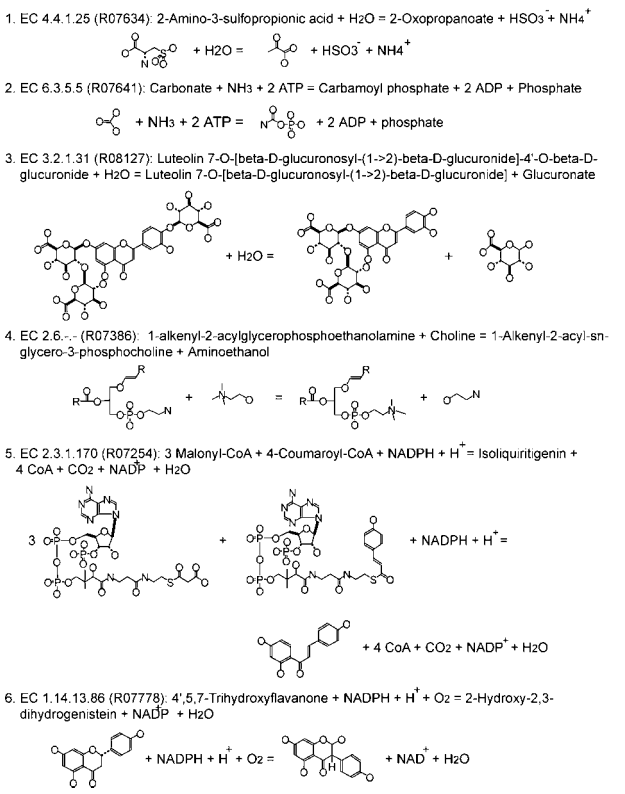

Figure 7. Examples of equations that could not be generated in this study. 1: where the compounds that should be added to the partial equation are very rare. 2: a main substrate or a main product (or both) has less than five non-hydrogen atoms. 3: a compound contains multiple identical or similar substructures. 4: cases where the maximum-common-substructure strategy fails. 5: multistep reactions. 6: cases where intramolecular transfer of a group takes place. R numbers are identifiers of reactions registered in the KEGG database. See text for details.

(C08549), portulacaxanthin I (C08564), and miraxanthin-V (C08557) might also be synthesized by adding proline, hydroxyproline, and dopamine, respectively. However, they were not generated in this study because our method could not find a common set of compounds that could be added to balance the equations for the first two and rejected the last one in the decision-tree process. In contrast, some unlikely paths are generated. Most of the paths starting from portulacaxanthin III (C08566) appear to be misleading results caused by the maximum-common-substructure strategy, which suggests that miraxanthin-II (C08555), vulgaxanthin-I (C08568), vulgaxanthin-II (C08569), dopaxanthin (C08543), and portulacaxanthin II (C08565) might be synthesized by creating a novel carbon−carbon bond on the glycine residue of portulacaxanthin III (C08566), as indicated by the red lines in Figure 6. These may be considered to be chemically unlikely. Suggested EC subsubclasses on these pairs are also hard to understand. There are other types of paths that seem unlikely to occur (green lines in Figure 6), because the pairs include multiple joining/cutting points, and there are alternative paths in each case.

The remaining blue lines indicate the pairs that remain probable when each equation is examined independently. However, each equation must be examined carefully in the context of the total pathway. There are three closed circuits consisting only of blue lines: (1) betalamic acid (C08538) ⇔ vulgaxanthin-I (C08568) ⇔ vulgaxanthin-II (C08569), (2) miraxanthin-III (C08556) ⇔ miraxanthin-V (C08557) ⇔ dopaxanthin (C08543) ⇔ portulacaxanthin II (C08565), and (3) betanin (C08540) ⇔ amaranthin (C08537) ⇔ bougainvillein-r-I (C08541). Of these, the pathway in the second circuit looks unnatural because the hydroxy group added in the first step from portulacaxanthin II (C08565) to dopaxanthin (C08543) is removed in the following step from miraxanthin-V (C08557) to miraxanthin-III (C08556). This pathway is convincing only when the compound miraxanthin-III (C08556) is needed but the step portulacaxanthin II (C08565) - miraxanthin-III (C08556) is blocked. These circuits may depend on the genes present in a specified organism, or they may work as compensatory pathways in mutant strains.

There are cases where several equations and EC subsubclasses are suggested for a reactant pair. These cases can be classified into the following two groups:

1. Reactions with a variety of possible cofactors. For example, if a partial reaction "R-CH(NH_2)-R′ = R-C(=O)-R′" is given, both EC 1.4 (oxidoreductases acting on the CH-NH_2 group of donors) and 2.6.1 (transaminases) are possible ["R-CH(NH_2)-R′ + H_2O + O_2 = R-C(=O)-R′ + NH_3 + H_2O_2" and "R-CH(NH_2)-R′ + 2-ketoglutaric acid = R-C(=O)-R′ + L-glutamic acid", respectively]. Oxidoreductases (EC 1) acting on the same functional group usually have a variety of cofactors (corresponding to the EC subsubclasses) [NAD(P)H + H^+/NAD(P)$^+$, H_2O_2/O_2, *etc.*]. Ligases (EC 6) also have a variety of cofactors (ATP/ADP + phosphate, ATP/AMP + diphosphate, GTP/GDP + phosphate, *etc.*), although the cofactors are not related to the classification of ligases. Examples of the first two cases can be seen in Figure 6 as the reactant pairs vulgaxanthin-I (C08568) - vulgaxanthin-II (C08569) and miraxanthin-III (C08556) - miraxanthin-V (C08557), respectively.

2. Some pairs corresponding to "*1,0,0*" could involve transferase (EC 2), hydrolase (EC 3), or ligase (EC 6) reactions. There is a tendency for the type of enzyme used to be dependent on the type of pathway involved; for example, the EC classes 2 and 6 tend to be used in synthetic pathways, while EC 1.13, EC 3, and EC 4 tend to be involved in degradation pathways, although there are many exceptions.

There is usually no particular reason to select one of these enzyme subsubclasses based on comparison of the structures of the main substrates and products. As well as the types of pathways, the contents of expressed enzyme genes in the specified organism and the dynamic balance of the total metabolic system of the cell must be considered in order to identify these kinds of enzymes.

Our method can be used to predict possible pathways, but it cannot prove that they actually exist in any specific organism. Such proof can only be obtained by direct experiment, and, in that context, this method should serve as a rational guide. It is also important to consider the possibility that there may be compounds missing from the data set used in this study. In naphthalene metabolism, for example, two epoxides [(1*R*,2*S*)-naphthalene epoxide and (1*S*,2*R*)-naphthalene epoxide] are known to be intermediates in the production of various derivatives such as 1-naphthol, 2-naphthol, *etc.*, but those pathways were not predictable using the old version of the KEGG database, because the intermediates had not been registered in the database at that time so the direct pathways from naphthalene to 1-naphthol,

Table 1. Questions Prepared for the Decision Tree

questions	answers
How many compounds are involved in the equation?	2, 3, 4, 5, 6, etc.
How many compound pairs are defined in the equation?	1, 2, or 3.
What kinds of atoms have to be added to the partial equation in order for it to balance?[a]	(C, H, O, N, S, P) = (0, 2, 0, 0, 0, 0), (0, 3, −1, 1, 0, 0), (1, 2, 0, 0, 0, 0), etc.
What compounds need to be added to the partial equation?	"NAD^+ = NADH + H^+", "ATP = ADP + phosphate", etc.
What compound is involved?	water, ATP, CO_2, etc.
What compound pair is involved?	"UDP-glucose = UDP", "ATP = ADP", etc.
Are the compositional formulas in a compound pair the same?	yes or no
What percentage of non-hydrogen atoms is preserved?[b]	100%, ≥90%, ≥80%, etc.
How many instances of intermolecular joining/cutting, ring formation/opening, and hydrogenation/dehydrogenation occur?[c]	"1,0,0", "0,0,1", etc.
What chemical bonds are changed?	
Level 1 (elements)	CC to CO, etc.
Level 2 (with orbitals)[d]	C:ar-O:sp3 to C:sp3-H, etc.
Level 3 (with the number of attached atoms)[e]	O:sp3x1-H to O:sp3x2-C:sp3x3, etc.
What functional groups are	
created or destroyed?	alcohol to aldehyde, etc.
conserved?	carboxylate, benzene ring, etc.
What substructures[f] are	
created or destroyed?	S0001727 to S0008573, etc.
conserved?	S0000214, etc.

[a] Positive and negative numbers indicate the number of corresponding elements that needs to be added to the right- and left-hand side of the equation, respectively. [b] Tanimoto coefficient (see text for definition). [c] The types of chemical bonds that are created/destroyed are grouped into three types: intermolecular joining/cutting, ring formation/opening, and hydrogenation/dehydrogenation. The number associated with each type is separated from its neighbor by a comma (see the first section of Results and Figure 3). [d] Element types (C, N, O, S, etc.) and orbitals [sp3, sp2, sp or ar (aromatic ring and heterocyclic ring with delocalized electrons)] are connected using a colon. [e] The number of attaching non-hydrogen atoms (x1, x2, x3 or x4) is added to the description at Level 2. [f] Substructure IDs registered in the BiSSCat database (Kotera et al., 2008).

2-naphthol, *etc.* were suggested instead. Such difficulties should be minimized by enlarging the compound data set used.

DISCUSSION

Finding the appropriate place(s) in cellular metabolism for the increasing number of orphan metabolites is a daunting task. Our approach, which uses substructure relationships and statistical-probability determinations, allows likely precursors of reaction products to be found and can predict the type of enzyme likely to be involved. As a result, it should act as a guide for simplifying the search for the metabolic processes involved. Other search strategies for determining possible pathways from chemical compound structures have been published.[33−39] These methods are similar to our method in that they use a rule-based approach, recognizing functional groups and their transformation patterns in organic compounds. However, our method has the advantage that transformation patterns need not be predefined. Boyer et al.[40] used the frequency with which atoms at reaction centers are known to be involved in xenobiotic transformations to predict possible reaction centers of "new" substructures. This approach is similar to that used in our BiSSCat system,[1] but their approach is limited to xenobiotic catabolism and does not provide any suggestions regarding the products of transferase or synthase reactions. Our method is applicable not only to xenobiotic degradation but also to synthetic pathways, and, in addition to suggesting the main products of a reaction, it also suggests possible reaction equations. Cross-docking approaches[41,42] also require the type of reactions to be specified and, apparently, the 3D structures of the proteins to be known. It is hoped that the freedom of our method from such constraints will prove advantageous in diverse secondary metabolism, where unexpectedly complicated reactions take place for diverse compounds.

There are reported reaction equations that were not generated in this study, which can be grouped into three categories.

1. Those where the reported equations contain incomplete structures.

Reactions containing compounds without any structural data cannot be predicted. Those containing compounds with partial structural data may be predictable, but it depends on whether or not the omitted substructures (often described as "R" and "*" for polymers) are consistent on both sides of the reaction equation. As an easy example, the equation "RCH_2OH + NAD^+ = RCHO + NADH + H^+" is predictable because the R group is defined consistently in both sides of the equation, but it would become unpredictable if the reaction was described as "ROH + NAD^+ = RCHO + NADH + H^+".

2. Cases where the constraints imposed to allow efficient operation of the system exclude certain possibilities, as listed below. Each of these could be accommodated relatively simply, but at the expense of computational time and storage capacity.

(a) Stoichiometrically unbalanced equations have been reported, where only the main substrate and product are known. While our method can often suggest ways to balance a reaction equation, *e.g.* by adding cofactors, the equation cannot be predicted when the compounds that should be added to the partial equation are very rare [such as the bisulfite ion and ammonia; see Reaction 1 of Figure 7]. This is a tradeoff problem between selectivity and sensitivity. The more rare cases are taken into account, the more false positives will be generated.

Table 2. Numbers of Compounds That Are Involved in Reported Equations, Generated Equations, and Orphan Compounds[a]

	compounds	reported[b]	of which generated[c]	orphan[d]	of which generated[e]
Steroids	350	147	116	203	20
estrogens and derivatives	30	21	14	9	0
androgens and derivatives	35	28	20	7	3
progestogens and derivatives	31	15	15	16	1
cholesterol and derivatives	24	22	21	2	1
phytosterols	37	16	12	21	4
steroidal alkaloids	32	0	0	32	0
steroid saponins	36	1	1	35	3
Terpenoids	967	199	136	768	128
C5 isoprenoids	2	2	2	0	0
C10 isoprenoids (monoterpenes)	136	75	70	61	26
C15 isoprenoids (sesquiterpenes)	28	7	5	21	8
C20 isoprenoids (diterpenes)	63	56	35	7	1
C25 isoprenoids (sesterterpenes)	1	0	0	1	0
C30 isoprenoids (triterpenes)	37	14	10	23	3
C40 isoprenoids (tetraterpenes)	34	23	8	11	5
monoterpenoids	78	31	29	47	32
sesquiterpenoids	105	5	1	100	20
sesquiterpene lactones	204	1	1	203	32
diterpenoids	109	5	3	104	4
diterpenoid alkaloids	64	2	1	62	8
triterpenoid saponins	67	4	0	63	0
carotenoids	37	18	9	19	10
Alkaloids	840	57	36	783	120
betalain alkaloids	28	0	0	28	26
indole alkaloids	130	11	7	119	9
isoquinoline alkaloids	140	19	14	121	13
peptide alkaloids	24	0	0	24	2
pyrrolidine and piperidine alkaloids	69	10	5	59	14
pyrrolizidine alkaloids	51	2	0	49	11
quinazoline alkaloids	14	0	0	14	2
quinoline alkaloids	49	1	1	48	11
quinolizidine alkaloids	37	1	0	36	7
tropane alkaloids	36	8	6	28	10
Phenolics	927	149	135	778	249
anthocyanins and anthochlors	43	13	12	30	24
benzofurans	24	1	0	23	14
chromones and chromenes	33	0	0	33	9
coumarins	65	9	7	56	13
flavones and flavonols	113	26	23	87	40
isoflavonoids and neoflavonoids	81	23	18	58	19
lignans	72	0	0	72	8
phenolic ketones	27	2	2	25	5
phenols and phenolic acids	55	25	25	30	11
phenylpropanoids	80	19	18	61	29
quinones	90	4	3	86	23
stilbenoids	41	3	3	38	13
tannins	35	1	1	34	1
Xantholnes	36	0	0	36	13

[a] Only some representative classes are listed. The classification is taken from the KEGG database and modified. [b] Number of compounds involved in at least one reported equation. [c] For which there is at least one equation suggested. [d] Number of orphan compounds. [e] For which there is at least one equation suggested.

(b) A main substrate or a main product (or both) has fewer than five non-hydrogen atoms (such as in Reaction 2 of Figure 7). This is only a problem because we decided to discard pairs having $A \cap B < 5$ in order to avoid comparing a large number of ubiquitous small substructures. Although this could be dealt with simply by removing that constraint, there are advantages in terms of computational time to retaining it, as long as we focus on identifying unknown metabolic reactions of relatively large compounds. In fact, only 1.8% (96 out of 5310) of orphan metabolites have fewer than five non-hydrogen atoms, of which 69 are inorganic compounds.

(c) A compound contains multiple identical or similar substructures, such as in the formation or degradation of a dimer, trimer, *etc*., or compounds containing repetitive substructures (such as Reaction 3 in Figure 7). This causes a problem because chemical-structure comparison and pair assembly (Scheme 2) are executed separately. There are cases where different substructures on one side of an equation are assigned to the identical substructure on the other side. This problem could be solved by comparing chemical structures dealing with the whole reaction equation or storing the second or third candidate results of binary chemical-structure comparison. However, both of these solutions require large amounts of calculation and data-storage space.

3. Cases where the maximum-common-substructure strategy fails. These cases, which appear to be relatively

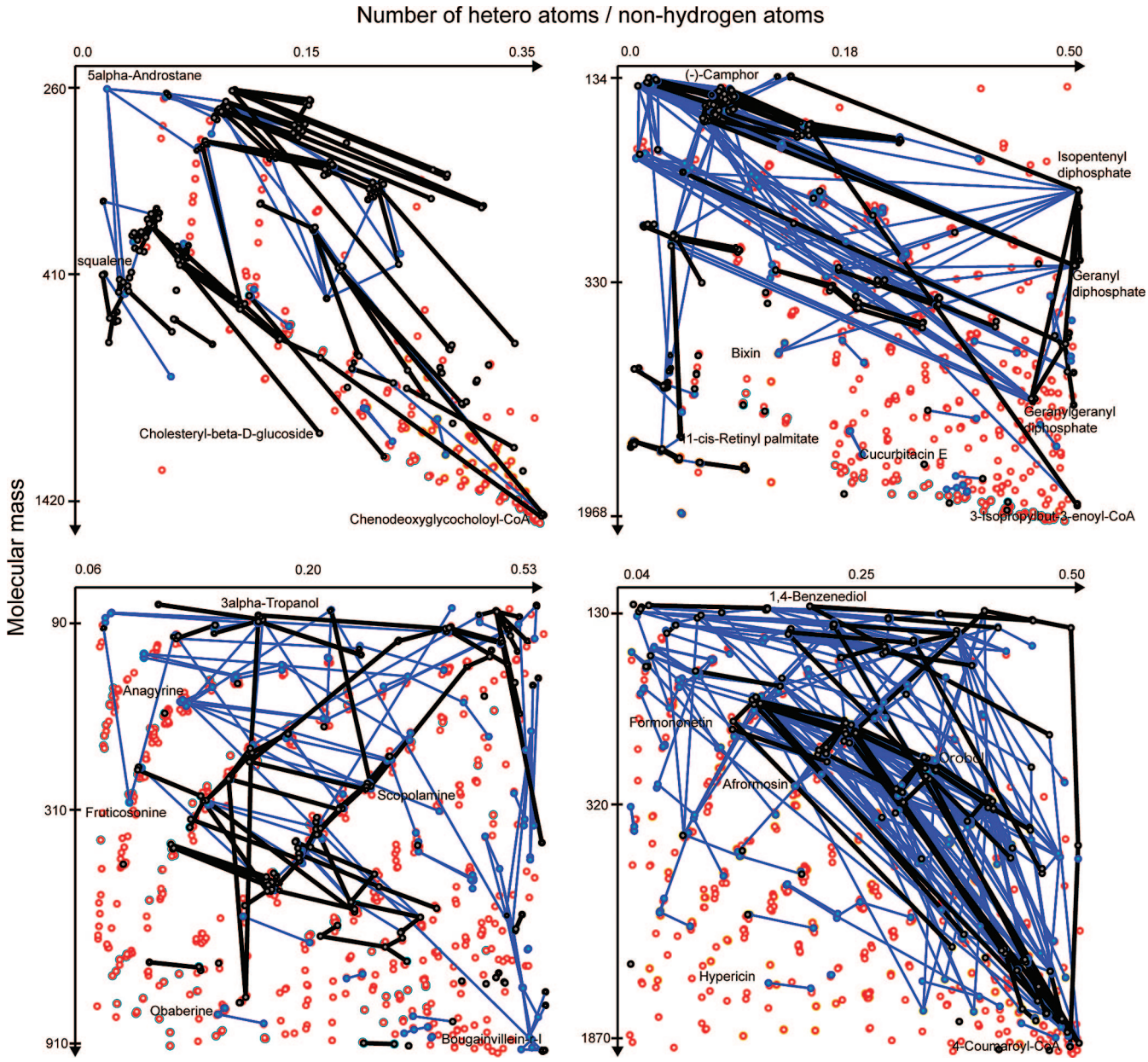

Figure 8. Generated metabolic maps of steroids (upper left), terpenoids (excluding steroids, upper right), alkaloids (lower left), and phenolics (including flavonoids, lower right). The vertical axis represents molecular mass, and the horizontal axis represents the number of heteroatoms (O, N, S, P, etc.) divided by the number of all non-hydrogen atoms, which roughly indicates hydrophilicity or hydrophobicity. Each circle represents a chemical compound, the exact location of which may be slightly moved so that the circles do not overlap with each other, if possible. Black circles are compounds involved in reported reactions. Blue circles are orphan compounds involved in the reactions generated in this study. Red circles are orphan compounds that are not involved in any equations, either reported or generated. Black lines indicate reported reactant pairs, and blue lines are reactant pairs that were suggested in this study.

uncommon in intermediary metabolism, can only be addressed by expansion of the present system.

(a) The maximum-common-substructure strategy is not always successful, as indicated in Reaction 4 in Figure 7. This reaction involves the exchange of an ethanolamine and a choline residue, but the maximum-common-substructure strategy would interpret this reaction as three successive methyl-group transfer reactions.

(b) Multistep reactions in which the product of one reaction becomes the substrate for the next reaction, with both reactions being catalyzed by the same enzyme are predictable so long as they occur with some frequency (such as dehydrogenation followed by decarboxylation), However, it is more difficult to predict a multistep reaction if it is rare or if it repeatedly incorporates small units, such as in Reaction 5 (Figure 7). Further development of our system will be necessary to deal with this problem, since this is a key type of reaction in flavonoid biosynthesis, and fatty-acid and polyketide synthesis both involve repetitive reactions.

(c) Intramolecular transfer of a group takes place (such as in Reaction 6 of Figure 7). This problem is caused in the process of finding a maximum spanning tree in the defined association graph, which finds only one connected substructure common to both compounds in the pair. This corresponds to the pattern "*2,0,0*" of chemical bonds created/destroyed (if a transferred group is also considered to be an intramolecular group, it should be "*0,2,0*"). Reactions are only predictable when the transferred group occurs in several

reported reactions, depending on the threshold set, such as phosphomutases (EC 5.4.2), aminomutases (EC 5.4.3), and hydroxymutases (EC 5.4.4). There are important cases, such as Reaction 5 (Figure 7) in flavonoid biosynthesis, where the transfer occurs in only one reported reaction. Our method can be expanded to find multiple common substructures; however, it would cause another tradeoff problem as well as an increase in computational time, since a compound usually has many isomers that differ in the positions of functional groups but that cannot be converted into each other in a single reaction.

Despite those equations that are unpredictable, our method gives some hints about pathways, especially those containing many orphan metabolites, as shown in Figure 6. Tables 1 and 2 contain a list of compound categories for some well-studied groups of secondary metabolites, *i.e.* steroids, terpenoids, alkaloids, and phenolics, with the numbers of compounds that are involved in reported equations and generated equations also given. The proportions of orphan metabolites in known compound groups differ in different categories, as do the proportions of orphan metabolites involved in generated equations. However, the former and the latter are independent of each other. An overview of the networks generated can be seen by putting them on a map in the same way as done for Figure 6 (Figure 8). In general, reactions involving small compounds have been more thoroughly investigated compared to those involving larger compounds. The connectivity, both reported and generated, also depends on the number of heteroatoms/non-hydrogen atoms, which provides a rough indication of hydrophilicity or hydrophobicity. The lower connectivity generated for hydrophobic compounds may be due to fewer intermediates being registered in the database (KEGG) we used, and the presence of many repetitive substructures, as discussed above. It is particularly difficult to connect large molecules in a hydrophobic area, because they contain only a few (or sometimes no) heteroatoms that are useful landmarks in finding related substructures in compound pairs.

Steroids are a well-investigated group, and most reported reactions are successfully generated in this study, while not many orphan metabolites are assigned in generated equations. Terpenoids located in hydrophobic and large-molecular-mass compounds (mainly carotenoids) are already well-known, while hydrophilic terpenoids (including most of the triterpenoid saponins) are not. Hydrophobic terpenoids with smaller molecular masses (the monoterpenoids, diterpenoids, *etc.*) are especially densely connected to each other by generated edges. Phenolics also have densely connected clusters in smaller-molecular mass and hydrophilic areas (especially flavones, flavonoids, isoflavonoids, and neoflavonoids). These densely connected clusters occur mainly because they share a common carbon backbone, and the only differences are in the locations of unsaturated bonds and/or small functional groups such as the hydroxy group, since there are several reported enzymes that catalyze intramolecular transfer of unsaturated bonds (EC 5.3.3) and hydroxy groups (EC 5.4.4). This does not necessarily mean that they connect directly to each other in single reactions as there may be unknown intermediates involved in the synthesis of these compounds. Such clusters do not appear in alkaloids, because alkaloids originate from many diverse backbones that are not structurally similar. The effects of loosening the

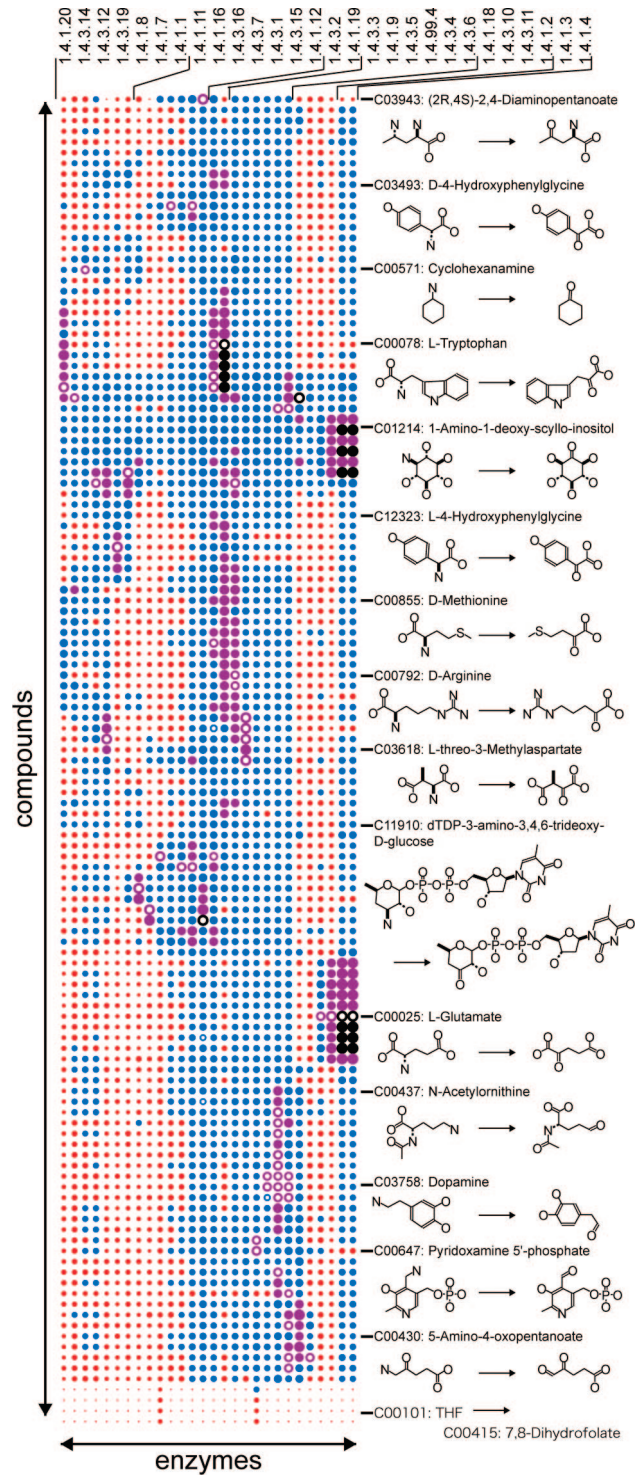

Figure 9. Suggested substrate specificity of enzymes in EC 1.4. Columns and rows represent EC numbers and compounds, respectively. C-numbers are identifiers of chemical compounds as used in the KEGG database. The diameter of each dot is proportionate to the likelihood of the compound (the row) being a substrate of the enzyme (column). Colors of the dots also represent the likelihood values as follows: black = 1.0; purple = 0.75—1.0; blue = 0.50—0.75; red = 0.0—0.50. Open circles indicate reported substrates. Each of the columns and rows is considered as a vector consisting of the likelihood scores, and the vectors are arranged using a data-optimization genetic algorithm, so that similar vectors are as close to each other as possible.

constraints might be evaluated for compounds with only one or no generated equation in order to see whether they might, in fact, be connected to other related compounds. Similarly

for compounds in densely connected clusters, stricter constraints might be used to assess the likelihood of individual edges.

The prediction of partial EC numbers is based solely on the overall reaction catalyzed and takes no account of the sequence, structure, evolution, or chemical mechanism of the putative enzyme. Our method is also designed to find inherently plausible reactions and pathways without any prior knowledge of the 2D or 3D structure of the enzyme involved. An advantage of this approach is that all enzymes, and not just those that have been characterized, can be included in the analysis. A slight modification of our method, *i.e.*, using full EC numbers instead of EC subsubclasses in the decision-tree processes, could help to identify possible alternative reactions of enzymes, supposing that such enzymes have wider specificity. It could also be used to identify competitive inhibitors. Figure 9 shows the likelihood scores of a compound being a substrate of enzymes in EC 1.4 (oxidoreductases acting on the CH-NH_2 group), generated by the random-forest method. This method should be improved by incorporating inhibitors as negative data and taking account of species differences in expressed enzyme activity[43] or paralogous enzyme genes. It could also be used to correlate genome annotation to predicted enzyme function. This might be especially useful in fungus metabolome research, because the genes required for a fungus to produce a given secondary metabolite are very frequently clustered, being adjacent to one another on the chromosome,[44−46] whereas eukaryotic genes involved in a single metabolic pathway are generally scattered throughout the genome.

CONCLUSIONS

This study is a guide to finding possible reactions for orphan metabolites in cellular metabolism, and the follow-up study must be conducted experimentally in order to confirm if the assignments made by the algorithm are valid. The number of orphan metabolites is still increasing, and it is expected that many more may be discovered. Our method should help to address this problem by suggesting possible reactions and pathways in which such orphan metabolites might be involved. It is also important to remember that some reactions are specific to a single organism, or group of organisms, and therefore the inferred equations will not necessarily be applicable to all species. This procedure could also contribute to environment genomics (also referred to as ecogenomics, metagenomics, or community genomics)[47,48] studies, which are based on the idea that a collection of genes sequenced from an environment could be analyzed in a way analogous to the study of a single genome.

ACKNOWLEDGMENT

We are grateful to Science Foundation Ireland (grant no. 07/IN.1/B930-Tipton) for support.

REFERENCES AND NOTES

(1) Kotera, M.; McDonald, A. G.; Boyce, S.; Tipton, K. F. Functional group and substructure searching as a tool in metabolomics. *PLoS One* **2008**, *3*, e1537.
(2) Poolman, M. G.; Bonde, B. K.; Gevorgyan, A.; Patel, H. H.; Fell, D. A. Challenges to be faced in the reconstruction of metabolic networks from public databases. *IEE Proc.-Syst. Biol.* **2006**, *153*, 379–384.
(3) Grotenhermen, F. Pharmacokinetics and pharmacodynamics of cannabinoids. *Clin. Pharmacokinet.* **2003**, *42*, 327–360.
(4) Frisvad, J. C.; Andersen, B.; Thrane, U. The use of secondary metabolite profiling in chemotaxonomy of filamentous fungi. *Mycol. Res.* **2008**, *112*, 231–240.
(5) Andersen, B.; Dongo, A.; Pryor, B. M. Secondary metabolite profiling of *Alternaria dauci, A. porri, A. solani,* and *A. tomatophila*. *Mycol. Res.* **2008**, *112*, 241–250.
(6) Rotem, J. *The Genus Alternaria: biology, epidemiology and pathogenicity*. American Phytopathological Society Press: St. Paul, MN, 1994.
(7) Hartmann, T. From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry* **2007**, *68*, 2831–2846.
(8) Tipton, K. F.; Boyce, S. Enzyme Classification and Nomenclature. In *Encyclopedia of Life Sciences*; John Wiley & Sons, Ltd.: Chichester, 2005.http://www.els.net doi: 10.1038/npg.els.0003893 (accessed June 4, 2008).
(9) Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. *Enzyme Nomenclature*; Academic Press: San Diego, CA, 1992.
(10) McDonald, A. G.; Boyce, S.; Moss, G. P.; Dixon, H. B. F.; Tipton, K. F. ExplorEnz: a MySQL database of the IUBMB Enzyme Nomenclature. *BMC Biochem.* **2007**, *8*, 14.
(11) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **2008**, *36*, D480−D484.
(12) Biggs, N.; Lloyd, E.; Wilson, R. *Graph Theory, 1736−1936*; Oxford University Press: Oxford, U.K., 1986.
(13) Bush, B. L.; Sheridan, R. P. PATTY: A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–762.
(14) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **2004**, *126*, 16487–16498.
(15) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–D357.
(16) Kuhl, F. S.; Crippen, G. M.; Friesen, D. K. A combinatorial algorithm for calculating ligand binding. *J. Comput. Chem.* **1984**, *5*, 24–34.
(17) Takahashi, Y.; Maeda, S.; Sasaki, S. Automated recognition of common geometrical patterns among a variety of three-dimensional molecular structures. *Anal. Chim. Acta* **1987**, *200*, 363–377.
(18) Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **2003**, *125*, 11853–11865.
(19) Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.
(20) Willett, P.; Barnard, J.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(21) Watson, G. A. An algorithm for the single facility location problem using the Jaccard metric. *SIAM J. Sci. Stat. Comput.* **1983**, *4*, 748–756.
(22) Gini, C. *Variabilita′ e mutabilita, 1912 (reprinted in Memorie di metodologica statistica)*; Pizetti, E., Salvemini, T., Eds.; Libreria Eredi Virgilio Veschi: Rome, 1955.
(23) Chandra, B.; Mazumdar, S; Arena, V; Parimi, N. Elegant decision tree algorithm for classification in data mining. Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops); 2002; 0-7695-1754-3/02,IEEE.
(24) Dorfman, R. A formula for the Gini Coefficient. *Rev. Econ. Stat.* **1979**, *61*, 146–149.
(25) Kleiber, C.; Kotz, S. A characterization of income distributions in terms of generalized Gini coefficients. *Soc. Choice Welfare* **2002**, *19*, 789–794.
(26) Haidich, A.; Ioannidis, J. The Gini coefficient as a measure for understanding accrual inequalities in multicenter clinical studies. *J. Clin. Epidemiol.* **2004**, *57*, 341–348.
(27) Kaufmann, T.; Schupfer, G.; Bauer, M. The Gini coefficient. A numerical grading for the degree of standardization of surgical subspecialities. *Der Anaesthesist* **2006**, *55*, 791–796.
(28) Harcha, B. D.; Correll, R. L.; Meech, W.; Kirkby, C. A.; Pankhurst, C. E. Using the Gini coefficient with BIOLOG substrate utilisation data to provide an alternative quantitative measure for comparing bacterial soil communities. *J. Microbiol. Methods* **1997**, *30*, 91–101.
(29) Graczyk, P. P. Gini Coefficient: a new way to express selectivity of kinase inhibitors against a family of kinases. *J. Med. Chem.* **2007**, *50*, 5773–5779.

(30) Qi, Y.; Bar-Joseph, Z.; Klein-Seetharaman, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 490–500.
(31) Zhang, L.; Wong, S.; King, O. D.; Roth, F. P. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinf.* **2004**, *5*, 38.
(32) Flach P. The many faces of ROC analysis in machine learning, ICML-04 Tutorial; **2004**. Notes available from http://www.cs.bris.ac.uk/~flach/ICML04tutorial/ (accessed July 20, 2008).
(33) Langowski, J.; Long, A. Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Delivwey Rev.* **2002**, *54*, 407–415.
(34) Klopman, G.; Tu, M. Structure-biodegradability study and computer automated prediction of aerobic biodegradation of chemicals. *Environ. Toxicol. Chem.* **1997**, *16*, 1829–1835.
(35) Talafous, J.; Sayre, L. M.; Mieyal, J. J.; Klopman, G. META.2. A dictionary model of mammalian xenobiotic metabolism. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1326–1333.
(36) Darvas, F. Predicting metabolic pathways by logic programming. *J. Mol. Graphics Modell.* **1988**, *6*, 80–86.
(37) Ellis, L. B. M.; Roe, D.; Wackett, L. P. The University of Minnesota biocatalysis/biodegradation database: the first decade. *Nucleic Acids Res.* **2006**, *34*, D517−D521.
(38) Hou, B. K.; Ellis, L. B. M.; Wackett, L. P. Encoding microbial metabolic logic: predicting biodegradation. *J. Ind. Microbiol. Biotechnol.* **2004**, *31*, 261–272.
(39) Oh, M.; Yamada, T.; Hattori, M.; Goto, S.; Kanehisa, M. Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J. Chem. Inf. Model.* **2007**, *47*, 1702–1712.
(40) Boyer, S.; Arnby, C. H.; Carlsson, L.; Smith, J.; Stein, V.; Glen, R. C. Reaction site mapping of xenobiotic biotransformations. *J. Chem. Inf. Model.* **2007**, *47*, 583–590.
(41) Hermann, J. C.; Marti-Arbona, R.; Fedorov, A. A.; Fedorov, E.; Almo, S. C.; Shoichet, B. K.; Raushel, F. M. Structure-based activity prediction for an enzyme of unknown function. *Nature* **2007**, *448*, 775–779.
(42) Favia, A. D.; Nobeli, I.; Glaser, F.; Thornton, J. M. Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. *J. Mol. Biol.* **2008**, *375*, 855–874.
(43) Barthelmes, J.; Ebeling, C.; Chang, A.; Schomburg, I.; Schomburg, D. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.* **2007**, *35*, D511–D514.
(44) Keller, N. P.; Hohn, T. M. Metabolic pathway gene clusters in filamentous fungi. *Fungal Genet. Biol.* **1997**, *21*, 17–29.
(45) Tag, A.; Hicks, J.; Garifullina, G.; Ake Jr, C.; Phillips, T. D.; Beremand, M.; Keller, N. P. G-protein signaling mediates differential production of toxic secondary metabolites. *Mol. Microbiol.* **2000**, *38*, 658–665.
(46) Yu, J. H.; Keller, N. Regulation of secondary metabolism in filamentous fungi. *Ann. Rev. Phytopathol.* **2005**, *43*, 437–458.
(47) Handelsman, J.; Rondon, M. R.; Brady, S. F.; Clardy, J.; Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **1998**, *5*, 245–249.
(48) Chen, K.; Pachter, L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol.* **2005**, *1*, 24.

CI800213G