BIOINFORMATICS

Vol. 30 ISMB 2014, pages i165–i174
doi:10.1093/bioinformatics/btu265

Metabolome-scale prediction of intermediate compounds in multistep metabolic pathways with a recursive supervised approach

Masaaki Kotera[1,†], Yasuo Tabei[2,†], Yoshihiro Yamanishi[3,4,†], Ai Muto[5], Yuki Moriya[6], Toshiaki Tokimatsu[6] and Susumu Goto[6,*]

[1]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan, [2]PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan, [3]Division of System Cohort, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan, [4]Institute for Advanced Study, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan, [5]Graduate School of Biological Sciences, Nara Institute of Science and Technology (NAIST), 8916-5 Takayama, Ikoma, Nara 630-0192, Japan and [6]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

ABSTRACT

Motivation: Metabolic pathway analysis is crucial not only in metabolic engineering but also in rational drug design. However, the biosynthetic/biodegradation pathways are known only for a small portion of metabolites, and a vast amount of pathways remain uncharacterized. Therefore, an important challenge in metabolomics is the *de novo* reconstruction of potential reaction networks on a metabolome-scale.

Results: In this article, we develop a novel method to predict the multistep reaction sequences for *de novo* reconstruction of metabolic pathways in the reaction-filling framework. We propose a supervised approach to learn what we refer to as '*multistep reaction sequence likeness*', i.e. whether a compound–compound pair is possibly converted to each other by a sequence of enzymatic reactions. In the algorithm, we propose a recursive procedure of using step-specific classifiers to predict the intermediate compounds in the multistep reaction sequences, based on chemical substructure fingerprints/descriptors of compounds. We further demonstrate the usefulness of our proposed method on the prediction of enzymatic reaction networks from a metabolome-scale compound set and discuss characteristic features of the extracted chemical substructure transformation patterns in multistep reaction sequences. Our comprehensively predicted reaction networks help to fill the metabolic gap and to infer new reaction sequences in metabolic pathways.

Availability and implementation: Materials are available for free at http://web.kuicr.kyoto-u.ac.jp/supp/kot/ismb2014/

Contact: goto@kuicr.kyoto-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Metabolic pathway analysis is crucial not only in systematic metabolic engineering (Toya and Shimizu, 2013) but also in rational drug discovery (Ramautar *et al.*, 2013). For example, 48.6% of cancer drugs are either natural products or their direct derivatives (Newman and Cragg, 2012), and many pharmaceutically useful compounds are produced by microbes, fungi and plants (Nakabayashi and Saito, 2013). It is estimated that plants produce at least 1 060 000 metabolites (Afendi *et al.*, 2012), and the total number of natural products is undoubtedly much larger if microbes and fungi are also considered. However, the biosynthetic/biodegradation pathways are known only for a small portion of metabolites, and a vast amount of pathways remain uncharacterized even in human (Sreekumar *et al.*, 2009). For example, International Union of Biochemistry and Molecular Biology recognizes only ~6000 enzymatic reactions (McDonald and Tipton, 2014). Therefore, *in silico* prediction of unknown pathways is expected to support the experimental characterization, leading to benefit not only drug discovery and health care but also agricultural and environmental issues.

There have been many successful studies on computational reconstruction of metabolic pathways in organisms or in specific conditions of cellular processes. The most traditional approach is '*reference-based framework*', where enzyme genes are mapped to appropriate positions in the predefined reference pathways using orthologous and other information across different organisms or conditions (Bono *et al.*, 1998). This framework is dependent on the predefined reference pathways, i.e. the collection of characterized substrate–product relationships that have been described in the literature or experimentally validated. Thus, this is not applicable to predicting unknown substrate–product relationships or completely new metabolic pathways.

In contrast, a variety of computational methods have been developed for *de novo* reconstruction of new metabolic pathways based on chemical structure data of metabolites. The goal is to elucidate novel reactions (absent from the reference pathway maps in the reference-based framework) based on our current knowledge about known reactions and chemical transformations (to be used as training data). The previous methods are mainly classified into '*compound-filling framework*' and '*reaction-filling framework*'. The compound-filling framework generates the chemical structures of the intermediates even if they are not present in databases. The users input the start (source) compound and/or the goal (target) compound, and the methods predict intermediates and reactions between the two compounds

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

(Darvas, 1988; Ellis *et al.*, 2008; Greene *et al.*, 1999; Moriya *et al.*, 2010; Talafous *et al.*, 1994).

On the other hand, the reaction-filling framework does not generate unknown chemical compounds, but use the compounds that are already present in databases. The users input a group of compounds, or vast amount of compounds in databases, and the methods predict the connectivity among the compounds, i.e. substrate–product pairs in a reaction (Hatzimanikatis *et al.*, 2005; Kotera *et al.*, 2008a, 2013a,b; Nakamura *et al.*, 2012; Tanaka *et al.*, 2009). The previous *de novo* metabolic pathway reconstruction studies handled single reactions independently. Although metabolic network involves some sequential reactions whose chemical transformation patterns are conserved (Muto *et al.*, 2013), these conserved patterns have not been considered.

In this article, we develop a novel method to predict the multistep reaction sequences for *de novo* reconstruction of metabolic pathways in the reaction-filling framework. We propose a supervised approach to learn what we refer to as '*multistep reaction sequence likeness*', i.e. whether a compound–compound pair is possibly converted to each other by multiple enzymatic reactions, as an extension of the previous work (Kotera *et al.*, 2013b). In the algorithm, we propose a recursive procedure of using step-specific classifiers to predict the intermediate compounds in the multistep reaction sequences, based on chemical substructures. We further demonstrate the usefulness of our proposed method on the prediction of enzymatic reaction networks from a metabolome-scale compound set and discuss characteristic features of multistep reaction sequences. Our comprehensively predicted reaction networks help to fill the metabolic gap and to infer new reaction sequences in metabolic pathways.

2 MATERIALS

2.1 Enzymatic reactions and reactant pairs

We retrieved enzymatic reactions and the associated chemical compounds from the KEGG database (Kanehisa *et al.*, 2012). Chemical compounds are given IDs consisting of the letter 'C' and the following five-digit numerals, and the structures are described as graphs where nodes and edges represent atoms and bonds, respectively. Hydrogen atoms are not explicitly represented as nodes but are included in the accompanying atoms. For example, the compounds D-glucose (C00031) and D-glucose 6-phosphate (C00092) consist of 12 and 16 nodes, respectively.

KEGG describes reactions not only as conventional reaction equation format but also as '*reactant pair*' format (RPAIR; Kotera *et al.*, 2004), representing substrate–product relationships with conserved chemical moiety in the reaction. For example, the pair D-glucose (12 nodes) and D-glucose 6-phosphate (16 nodes) conserves 12 nodes (corresponding to the glucose residue) during the reaction (see http://www.kegg.jp/entry/RP00060). As of January 2014, KEGG RPAIR stores 14 386 reactant pairs.

2.2 Positive/negative dataset for single reactions

In this study, the 'main' type of reactant pairs [the compound–compound pairs in the KEGG RPAIR database; see Kotera *et al.* (2013b)] were regarded as the positive examples, and the remaining all possible pairs of compounds were regarded as the negative examples for predicting enzymatic-reaction likeness. Each compound–compound pair has to be described in both forward and backward directions, avoiding the loss of the similarity in backward reactions. Considering these, the number of all possible compound pairs is $n(n–1) = O(n^2)$, where n is the number of compounds. We regard these positive/negative substrate–product relationships as the gold standard data.

It is known that most compound pairs in the negative examples are structurally dissimilar. In other words, it is easy to predict that dissimilar compound pairs are unlikely to be converted to each other by single enzymatic reactions. Thus, incorporation of structural dissimilar pairs in the prediction would overestimate the prediction accuracy in the performance evaluation. To avoid such trivial predictions, we removed compound pairs whose Tanimoto coefficient (Jaccard coefficient) are <0.5 from the gold standard data and constructed the filtered data consisting of compound pairs whose structures are similar to some extent. Note that classification is more difficult for the filtered data compared with the full data.

2.3 Positive/negative dataset for *k*-step reaction sequences

In the field of enzymology, the term '*multistep reaction*' sometimes means a series of chemical transformations catalyzed by an enzyme. In this article, we do not deal with *multistep* in that context. To avoid the confusion, we use the term '*k*-step reaction sequences' or '*k*-step sequences' for a series of chemical transformations catalyzed by k enzymes within a metabolic pathway. One-step sequences correspond to single reactions.

We prepared the positive and negative examples of k-step reaction sequences as follows (where $k = 2, 3, 4$).

(1) k-step reaction sequences consisting of $k + 1$ compounds were generated using k reactant pairs sharing common compounds. For example, from the three reactant pairs 'C_1–C_2', 'C_1–C_3' and 'C_3–C_4' (where C_n represents a compound ID), two 2-step sequences 'C_2–C_1–C_3' and 'C_1–C_3–C_4' and a 3-step sequence 'C_2–C_1–C_3–C_4' were generated. The first and the last compounds in the sequence were referred to as the start and the goal compounds, respectively.

(2) Using the RPAIR database, the conservation ratio of the atoms from the start compound to the goal compound was calculated for each k-step sequence. For example, consider that C_1, C_2, C_3 and C_4 consists of 18, 20, 23 and 24 nodes (i.e. atoms other than hydrogen atoms), respectively, and the pair 'C_1–C_2' conserve 18 nodes, the conservation ratio of the step 'C_1–C_2' is $18/18 = 1.0$ and that of 'C_2–C_1' is $18/20 = 0.9$, respectively. When the 17 nodes from the conserved nodes in 'C_2–C_1' are conserved in 'C_1–C_3', the conservation ratio of the step 'C_2–C_1–C_3' is $17/20 = 0.85$. In this study, the k-step sequences with the monotonic increase of the numbers of the nodes and the conservation ratio ≥ 0.5 were regarded as positive examples. The remaining k-step sequences were regarded as negative examples.

As a result of the data filtering, the numbers of positive examples of 1-, 2-, 3- and 4-step sequences were 10 852, 4294, 8073 and 15 112, respectively, and the numbers of negative examples in those are 518 854, 75 170, 258 883 and 1 138 634,

respectively. Note that the definitions of positive/negative examples in 1-step and the longer-steps were different, resulting in the numbers of positive/negative examples not in monotonic increase.

2.4 Vector representation of chemical structures

We obtained chemical structures of compounds (metabolites) from the KEGG (Kanehisa *et al*., 2012) and KNApSAcK (Afendi *et al*., 2012) databases. Chemically identical compounds with the same structures (duplicates) were removed, so structures of all compounds in the dataset were unique. We described each compound by using KEGG Chemical Function and Substructures (KCF-S; Kotera *et al*., 2013a), which was designed based on the numbers of functional group and other named biochemical substructures in a molecule (Kotera *et al*., 2008b). We represented each compound by an integer vector of length 53 679 in which the occurrence of a substructure is coded as an integer value.

For a comparison study, we also tested many other chemical fingerprints by using Chemistry Development Kit (CDK; Steinbeck *et al*., 2003) and the descriptor defined by Nakamura, Sakakibara and colleagues (Nakamura *et al*., 2012), referred to as 'NS-descriptor' in this study. The NS-descriptor is an integer vector, and the other fingerprints are binary vectors. We calculated eight fingerprints/descriptors: CDK extended fingerprint, CDK graph-only fingerprint, CDK hybridization fingerprint, E-state fingerprint, Klekota–Roth fingerprint, Molecular ACCess System (MACCS) fingerprint, PubChem fingerprint and NS-descriptor, and their dimensions were 1022, 1024, 1024, 71, 4860, 164, 879 and 1346, respectively, where the feature elements absent from the compound set are not considered.

2.5 Reaction module

Metabolic network involves some sequential reactions whose chemical transformation patterns are conserved, and these conserved sequences are referred to as 'reaction modules' (Muto *et al*., 2013). They consist of purely chemical data without incorporating any enzyme data (genes and proteins). As of January 2014, there are 34 manually curated reaction modules that are given IDs consisting of the letters 'RM' and the following three-digit numerals (such as RM001; see http://www.genome.jp/kegg/reaction/rmodule.html) and up to 3016 conserved reaction patterns. In this study, KEGG reaction modules were used for the analysis of the chemical substructures characteristic to k-step sequences.

3 METHODS

We address the problem of metabolome-scale metabolic pathway reconstruction in the reaction-filling framework. In this section, we present a general approach to evaluate the enzymatic-reaction likeness of any pair of two compounds and to estimate potential intermediate chemical structures between the two compounds.

3.1 Feature vector representation of compound–compound pairs

We represent a compound C by a D-dimensional integer vector (an integer vector of length D) as $\Phi(C) = (c_1, c_2, ..., c_D)^T$, where $c_k \in \mathbb{Z}$, $k = 1, ..., D$ and each element corresponds to the number of times a given chemical substructure from a library of defined substructures occurs in the compound. To characterize any pair of two compounds C and C', we introduce two kinds of operations for the descriptors as follows:

$$(\Phi(C) \wedge \Phi(C')) = (\min(c_1, c'_1), \min(c_2, c'_2), ..., \min(c_n, c'_n))$$

and

$$(\Phi(C) \ominus \Phi(C')) = (\max(c_1 - c'_1, 0), \max(c_2 - c'_2, 0), ..., \max(c_n - c'_n, 0)).$$

where $\min(c_k, c'_k)$ is a function that returns c_k if $c_k \leq c'_k$ and otherwise returns c'_k, and $\max(c_k, c'_k)$ is a function that returns c_k if $c_k \geq c'_k$ and otherwise returns c'_k. Note that $(\Phi(C) \wedge \Phi(C'))$ is an operation that captures common chemical substructures between $\Phi(C)$ and $\Phi(C')$, whereas $(\Phi(C) \ominus \Phi(C'))$ is an operation that captures chemical substructures present in $\Phi(C)$ and absent in $\Phi(C')$.

To represent any compound–compound pair using the above operations, we define two types of feature vectors as follows:

$$\Phi(C, C') = (\Phi(C) \wedge \Phi(C'), \Phi(C) \ominus \Phi(C'), \Phi(C') \ominus \Phi(C))^T$$

and

$$\overline{\Phi(C, C')} = (\Phi(C) \ominus \Phi(C'), \Phi(C') \ominus \Phi(C))^T.$$

We shall refer to $\Phi(C, C')$ and $\overline{\Phi(C, C')}$ as 'diff-common feature vector' and 'diff-only feature vector', respectively (Fig. 1a). Both feature vectors can handle reversible reactions, and they are designed to capture substructure changes around the reaction center in the conversion of a chemical compound to another compound. In addition, the diff-common feature vector is designed to capture conserved substructures kept in the conversion of a chemical compound to another compound.

3.2 Multistep reaction sequence-likeness prediction

3.2.1 Enzymatic-reaction likeness We make a brief review of the previous method to predict the enzymatic-reaction likeness, i.e. whether a compound–compound pair is possibly converted to each other by an enzymatic reaction (Kotera *et al*., 2013b), which is solved by the following supervised classification problem.

Using the feature vectors $\Phi(C, C')$ and $\overline{\Phi(C, C')}$ for compounds C and C', we apply a linear model to estimate a linear function $f(C, C') = \mathbf{w}^T \Phi(C, C')$, where $\mathbf{w}$ is a weight vector. The enzymatic-reaction likeness between C and C' is predicted by thresholding the value of $f(C,C')$. The weight vector $\mathbf{w}$ is estimated such that it can correctly predict the enzymatic-reaction likeness of compound–compound pairs.

A limitation of the previous method is that the method is only applicable to single reactions. Thus, in this study, we generalize it for the use of multistep reaction sequences as described in the following sections.

3.2.2 Multistep reaction sequence likeness Here we propose an extension of the enzymatic-reaction likeness from single reactions to multistep sequences. Let k be the number of reaction steps.

A k-step reaction sequence is defined as a series of reactions in which a chemical compound is known to be converted to another compound by multiple enzymatic reactions, and the corresponding $(k-1)$ intermediate compounds are missing. To evaluate the enzymatic-reaction likeness of k-step reaction sequence, we estimate a linear function $f_k(C, C')$ that would predict whether a chemical compound C is converted to another compound C' by k enzymatic reactions.

Using the feature vectors $\Phi(C, C')$ and $\overline{\Phi(C, C')}$ for compounds C and C', we propose to learn a linear function $f_k(C, C') = \mathbf{w}_k^T \Phi(C, C')$, where $\mathbf{w}_k$ is a weight vector, and the weight vector $\mathbf{w}_k$ is estimated such that it can correctly predict k-step reaction sequence likeness.

Given a collection of compound–compound pairs and their labels $(\Phi(C_i, C_j), y_{ijk})$, where $y_{ijk} \in \{+1, -1\}$ $(i = 1, ..., n, j = 1, ..., n, i \neq j)$ and $y_{ijk} = +1$ (resp. $y_{ijk} = -1$) indicates a positive pair (resp. a negative pair) in the k-step reaction, we estimate the weight vector $\mathbf{w}_k$ by linear

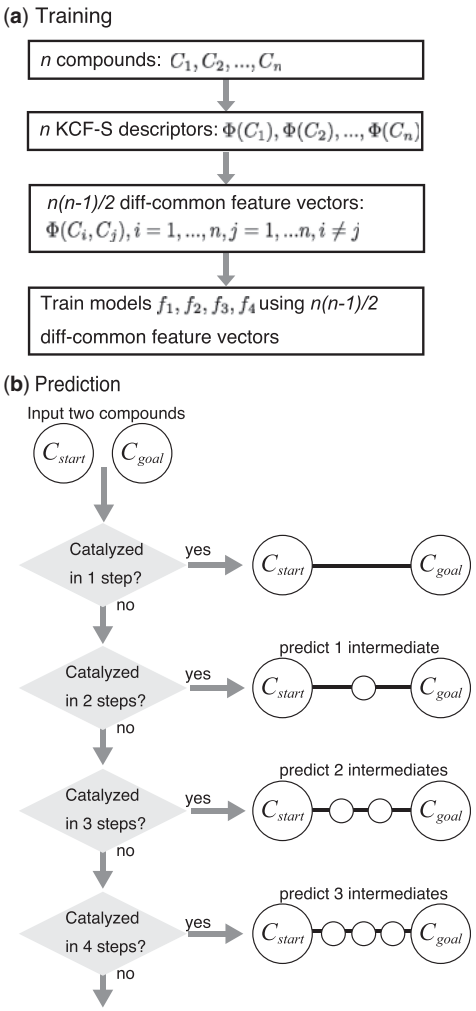

Fig. 1. Overview of training models and predictions of new compound–compound pairs. (**a**) Flowchart of training models using diff-common feature vectors. The same procedure is conducted for diff-only feature vectors as well. See Sections 3.1 and 3.2 for more details. (**b**) Flowchart of predicting the k-step reaction sequences. The k-th step is predicted by whether $f_k(C_{start}, C_{end}) > 0$. See Sections 3.1 and 3.3 for more details

Support Vector Machine (SVM) formulated by the following unconstrained optimization problem:

$$\min_{\mathbf{w}_k} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{i-1} M_{ij} + \sum_{j=i+1}^{n} M_{ij} \right\},$$

where

$$M_{ij} = \max\{1 - y_{ij}\mathbf{w}_k^T \Phi(C_i, C_j), 0\}^2.$$

To enhance the interpretability of linear models, the weight vector is optimized with L_1-regularization as follows:

$$\min_{\mathbf{w}_k} ||\mathbf{w}_k||_1 + C \sum_{i=1}^{n} \left\{ \sum_{j=1}^{i-1} M_{ij} + \sum_{j=i+1}^{n} M_{ij} \right\},$$

where $||\cdot||_1$ is L_1 norm (the sum of absolute values in the vector), and C is a hyper-parameter. L_1-regularization has an effect of making the weights of uninformative features zeros without loss of classification accuracy. L_1-regularized linear SVM is referred to as L1SVM.

For example, the prediction for 1-step, 2-step, 3-step and 4-step reaction sequence likeness can be performed as follows:

Enzymatic reaction-likeness prediction

We construct a function $f_1(C, C') = \mathbf{w}_1{}^T \Phi(C, C')$ based on a learning set of compound–compound pairs in known single reactions. We then apply f_1 to a given compound–compound pair to predict whether the two compounds in the pair are interconvertible by one enzymatic reaction.

2-step reaction sequence-likeness prediction

We construct a function $f_2(C, C') = \mathbf{w}_2{}^T \Phi(C, C')$ based on a learning set of compound–compound pairs in known 2-step sequences. We then apply f_2 to a given compound–compound pair to predict whether the two compounds in the pair are interconvertible by two enzymatic reactions.

3-step reaction sequence-likeness prediction

We construct a function $f_3(C, C') = \mathbf{w}_3{}^T \Phi(C, C')$ based on a learning set of compound–compound pairs in known 3-step reaction sequences. We then apply f_3 to a given compound–compound pair to predict whether the two compounds in the pair are interconvertible by three enzymatic reactions.

4-step reaction sequence-likeness prediction

We construct a function $f_4(C, C') = \mathbf{w}_4{}^T \Phi(C, C')$ based on a learning set of compound–compound pairs in known 4-step reaction sequences. We then apply f_4 to a given compound–compound pair to predict whether the two compounds in the pair are interconvertible by four enzymatic reactions.

3.3 Intermediate compound prediction in the multistep reaction sequences

Given a pair of start (source) compound C_{start} and goal (target) compound C_{goal} in the k-step reaction sequence, we attempt to estimate potential intermediate compounds $C_{inter}^{(1)}, C_{inter}^{(2)}, \ldots, C_{inter}^{(k-1)}$ between the start compound C_{start} and the goal compound C_{goal}. Note that there are $(k–1)$ intermediate compounds between the start compound C_{start} and the goal compound C_{goal} in the k-step reaction sequence (Fig. 2).

Suppose that we have a chemical database storing a huge number of chemical compounds, and we consider selecting potential compounds from the database for the intermediate compounds in the k-step reaction sequence. The j-th intermediate compound $C_{inter}^{(j)}$ is considered convertible from the start compound C_{start} by single reactions (1-step sequences) and is also considered convertible from the goal compound C_{goal} by $(k + 1– j)$-step sequence. Therefore, we propose the following candidate score to select an appropriate compound for the j-th intermediate compound $C_{inter}^{(j)}$ $(j = 1, 2, \ldots, (k-1))$ by integrating individual reaction sequence-likeness evaluation functions $f_1, f_2, \ldots, f_{(k-1)}$ in a recursive manner:

$$s_k^{(j)}(C) = f_j(C_{start}, C) + f_{k+1-j}(C, C_{goal}).$$

In practice, high-scoring compounds in the database are predicted to be candidates for the intermediate compounds.

For example, we propose the candidate scores for the 2-step, 3-step and 4-step reaction sequences as follows:

2-step reaction sequence with one intermediate compound

The intermediate compound is connected with the start compound by one step and with the goal compound by one step, so we propose the following candidate score for the intermediate compound:

$$s_2(C) = f_1(C_{start}, C) + f_1(C, C_{goal}).$$

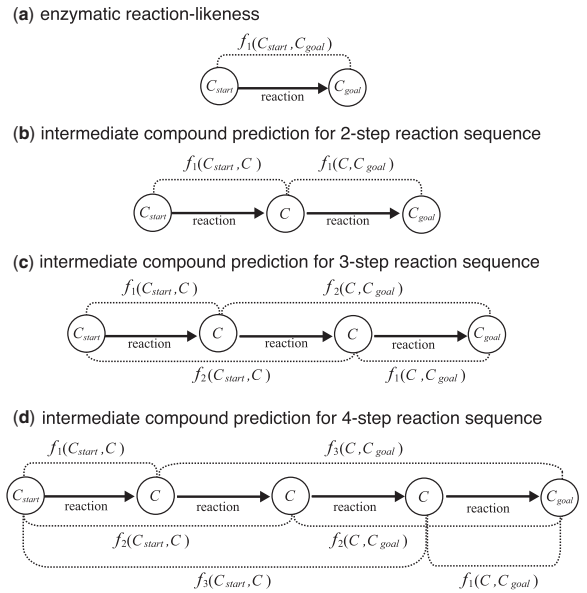

Fig. 2. k-step reaction sequences and intermediate compound prediction

3-step reaction sequence with two intermediate compounds

The first intermediate compound is connected with the start compound by one step and with the goal compound by two steps, so we propose the following candidate score for the first intermediate compound:

$$s_3^{(1)}(C) = f_1(C_{start}, C) + f_2(C, C_{goal}).$$

The second intermediate compound is connected with the start compound by two steps and with the goal compound by one step, so we propose the following candidate score for the second intermediate compound:

$$s_3^{(2)}(C) = f_2(C_{start}, C) + f_1(C, C_{goal}).$$

4-step reaction sequence with three intermediate compounds

The first intermediate compound is connected with the start compound by one step and with the goal compound by three steps, so we propose the following candidate score for the first intermediate compound:

$$s_4^{(1)}(C) = f_1(C_{start}, C) + f_3(C, C_{goal}).$$

The second intermediate compound is connected with the start compound by two steps and with the goal compound by two steps, so we propose the following candidate score for the second intermediate compound:

$$s_4^{(2)}(C) = f_2(C_{start}, C) + f_2(C, C_{goal}).$$

The third intermediate compound is connected with the start compound by three steps and with the goal compound by one step, so we propose the following candidate score for the third intermediate compound:

$$s_4^{(3)}(C) = f_3(C_{start}, C) + f_1(C, C_{goal}).$$

Practical application

In practice, we do not know how many reaction steps are there between the start compound C_{start} and the goal compound C_{goal}, so we propose the following recursive procedure (Fig. 1b):

(1) If $f_1(C_{start}, C_{goal}) > 0$, C_{start} and C_{goal} are predicted to be converted to each other.

(2) If $f_1(C_{start}, C_{goal}) \leq 0$ and $f_2(C_{start}, C_{goal}) > 0$, the intermediate compound is predicted with $s_2(C)$.

(3) If $f_2(C_{start}, C_{goal}) \leq 0$ and $f_3(C_{start}, C_{goal}) > 0$, the intermediate compounds are predicted with $s_3^{(1)}(C)$ and $s_3^{(2)}(C)$.

(4) If $f_3(C_{start}, C_{goal}) \leq 0$ and $f_4(C_{start}, C_{goal}) > 0$, the intermediate compounds are predicted with $s_4^{(1)}(C)$, $s_4^{(2)}(C)$ and $s_4^{(3)}(C)$.

(5) To be continued in a recursive manner.

3.4 Experimental evaluation protocol

3.4.1 Cross-validation experiment on reaction sequence-likeness prediction We perform the following 5-fold cross-validation. (i) We randomly split compound–compound pairs in the gold standard reaction data into five subsets of roughly equal sizes. We regard known start–goal compound pairs (the first and the last compounds in the sequence) as positive examples and the other compound–compound pairs as negative examples. (ii) We take each subset as a test set and the remaining four subsets as a training set. (iii) We learn a predictive model based only on the training set. (iv) We compute the prediction scores for compound–compound pairs in the test set. (v) Finally, we evaluate the prediction accuracy over the 5-folds.

We evaluate the prediction performance by the receiver operating characteristic (ROC) curve, which is a plot of true-positive rates as a function of false-positive rates based on various thresholds, and the precision-recall (PR) curve, which is a plot of precision (positive predictive value) as a function of recall (sensitivity) based on various thresholds. We summarize the performance by the area under the ROC curve (AUC) score, where 1 is for a perfect inference and 0.5 is for a random inference, and the area under the PR curve (AUPR) score, where 1 is for a perfect inference and the ratio of positive examples in the gold standard data is for a random inference.

In this study, we perform the above cross-validation experiments for 1-, 2-, 3- and 4-step reaction sequences, separately (see Section 4.1). In each case, we repeat the cross-validation experiment five times, and computed the averages of the AUC scores and the AUPR scores over the five cross-validation folds. The parameters involved in the methods are optimized with the AUC score and the AUPR score as the objective functions.

3.4.2 Self-rank test on intermediate compound prediction We conduct a self-rank test to simulate the intermediate compound prediction. The procedure of the self-rank test is as follows: (i) we take intermediate compounds in known k-step reaction sequences and regard them as missing intermediate compounds (to be tested), (ii) we compute the candidate scores for all candidate compounds in the chemical database and the intermediate compounds being tested, (iii) we rank the intermediate compounds based on the candidate scores among all candidate compounds plus themselves and (iv) we repeat the above steps for all the k-step reaction sequences. Note that we test one intermediate compound for the 2-step reaction, two intermediate compounds for the 3-step reaction and three intermediate compounds for the 4-step reaction.

In this study, we conduct the above self-rank test for 2-, 3- and 4-step reaction sequences, separately (see Section 4.2). A self-rank of 1 is a perfect prediction, indicating that the method is able to assign the test compound to the original position in the k-step reaction sequence. In the case of random prediction, the self-rank follows the uniform distribution on the interval from 1 to the number of candidate compounds in the chemical database.

3.5 Baseline method

3.5.1 Reaction sequence-likeness prediction The most straightforward method for the reaction sequence-likeness prediction is a similarity-based approach, assuming that the start compound and the goal compound are likely to have similar chemical structures. Actually, a substrate compound and a product compound in an enzyme reaction tend to have a big conserved substructure, and their different regions tend to be small (Kotera *et al.*, 2008a), so the start compound and the goal compound in the k-step reaction sequences are expected to have high chemical structure similarity. We use weighted Jaccard similarity for binary fingerprints and cosine correlation coefficient for real-valued descriptors as chemical structure similarity measures of two compounds. A direct strategy is therefore to predict the k-step reaction sequence-likeness between the start and goal compounds whenever the chemical structure similarity value between the start and goal compounds is above a threshold to be determined. We refer to this approach as BASELINE.

3.5.2 Intermediate compound prediction In a similar manner as the reaction sequence-likeness prediction, we define a baseline method for the intermediate compound prediction, assuming that the intermediate compounds are likely to have similar chemical structures both with the start compound and with the goal compound. The candidate score in the k-step sequence is defined as the sum of the chemical structure similarity between the candidate and start compounds and the chemical structure similarity between the candidate and goal compounds. We refer to this approach as BASELINE.

4 RESULTS

4.1 Performance evaluation on multistep reaction sequence-likeness prediction

We tested the proposed L1SVM method on its ability to predict the multistep reaction sequence likeness of given compound–compound pairs from their chemical fingerprint/descriptor data by performing 5-fold cross-validation experiments (see Section 3.4.1 for more details). We evaluated the performance of the method for 1-, 2-, 3- and 4-step reaction sequence likeness, separately. We also compared the performance between nine chemical fingerprints/descriptors: CDK extended fingerprint, CDK graph-only fingerprint, CDK hybridization fingerprint, E-state fingerprint, Klekota–Roth fingerprint, MACCS fingerprint, PubChem fingerprint, NS-descriptor and KCF-S descriptor (see Section 2.4 for more details). Note that KCF-S is the descriptor we proposed to use in this study.

Table 1, 2, 3 and 4 show the resulting AUC and AUPR scores for the 1-, 2-, 3- and 4-step reaction sequence-likeness predictions, respectively (full tables are shown in Supplementary Material). Among the nine chemical fingerprints/descriptors, the KCF-S descriptor achieved the highest prediction accuracy in any kinds of enzymatic reactions. The L1SVM method clearly outperformed the BASELINE method regardless of fingerprints, which suggests that supervised learning is meaningful for reaction prediction. These results suggest that the proposed L1SVM method with the KCF-S descriptor is expected to be useful in practice.

The AUC and AUPR scores of BASELINE were relatively high regardless of fingerprints in the case of single reactions, which validated the fact that a core substructure is conserved between a substrate and a product in a reactant pair. On the other hand, the AUC and AUPR scores of BASELINE were low regardless of fingerprints in the case of k enzymatic reactions where $k = 2, 3, 4$. This result suggests that a core substructure conserved from the start to the goal compounds tends to be smaller in the k-step sequences, compared with 1-step reactions. The diff-common feature vector worked better than the diff-only feature vector in most cases. This result also implies that both substructure transformation patterns and core substructures are important in the k-step sequence prediction.

4.2 Performance evaluation on intermediate compound prediction

We tested the proposed recursive L1SVM method with the diff-common feature vector on its ability to infer intermediate compounds in the k-step sequences. We evaluated the performance by conducting a self-rank test, which simulates the situation where we want to detect a series of intermediate compounds between a start compound and a goal compound (see Section 3.4.2 for more details). We evaluated the performance of the proposed method for 2-, 3- and 4-step sequences, separately.

Figure 3 shows the distributions of the computed self-ranks for the 2-, 3- and 4-step sequences, where the self-rank scores are shown on a log scale with base 10 in each panel, and the left box-plot corresponds to the random inference, the middle box-plot corresponds to the BASELINE method and the right box-plot corresponds to the proposed L1SVM method, respectively. Note that there are 1, 2 and 3 intermediate compounds in the 2-, 3- and 4-step sequences, respectively.

In both BASELINE and L1SVM, the self-rank distributions have a large peak at high ranks at a significant level (the P-value is almost zero), which means that both the methods are capable of predicting most known intermediate compounds correctly, compared with the random inference. The L1SVM method

Table 1. Cross-validation on the 1-step reaction sequence likeness (enzymatic-reaction likeness)

Chemical fingerprints/descriptors	Diff-common L1SVM		Diff-only L1SVM		Baseline		Random	
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
CDK extended	0.6917	0.0603	0.6742	0.0468	0.6161	0.0289	0.5000	0.0199
MACCS	0.6837	0.0489	0.6582	0.0342	0.5914	0.0189	0.5000	0.0199
PubChem	0.7170	0.0531	0.7026	0.0422	0.6752	0.0307	0.5000	0.0199
NS-descriptor	0.8858	0.2134	0.8429	0.0968	0.6566	0.0446	0.5000	0.0199
KCF-S descriptor	0.9659	0.3943	0.9610	0.2801	0.6945	0.0755	0.5000	0.0199

Table 2. Cross-validation on the 2-step reaction sequence-likeness prediction (with one intermediate compound)

Chemical fingerprints/descriptors	Diff-common L1SVM AUC	AUPR	Diff-only L1SVM AUC	AUPR	Baseline AUC	AUPR	Random AUC	AUPR
CDK extended	0.7747	0.1730	0.7178	0.1352	0.4815	0.0576	0.5000	0.0665
MACCS	0.7474	0.1418	0.6634	0.1152	0.4465	0.0502	0.5000	0.0665
PubChem	0.7674	0.1589	0.7270	0.1357	0.5732	0.0710	0.5000	0.0665
NS-descriptor	0.8898	0.2937	0.8673	0.2651	0.6187	0.0937	0.5000	0.0665
KCF-S descriptor	0.9411	0.4493	0.9419	0.4473	0.6621	0.0635	0.5000	0.0665

Table 3. Cross-validation on the 3-step reaction sequence-likeness prediction (with two intermediate compounds)

Chemical fingerprints/descriptors	Diff-common L1SVM AUC	AUPR	Diff-only L1SVM AUC	AUPR	Baseline AUC	AUPR	Random AUC	AUPR
CDK extended	0.8103	0.1436	0.7542	0.0959	0.5474	0.0368	0.5000	0.0367
MACCS	0.7608	0.0986	0.6770	0.0713	0.4959	0.0309	0.5000	0.0367
PubChem	0.8097	0.1239	0.7656	0.0910	0.6365	0.0489	0.5000	0.0367
NS-descriptor	0.9284	0.2638	0.9028	0.1989	0.7069	0.0807	0.5000	0.0367
KCF-S descriptor	0.9624	0.4232	0.9585	0.4094	0.6621	0.0635	0.5000	0.0367

Table 4. Cross-validation on the 4-step reaction sequence-likeness prediction (with three intermediate compounds)

Chemical fingerprints/descriptors	Diff-common L1SVM AUC	AUPR	Diff-only L1SVM AUC	AUPR	Baseline AUC	AUPR	Random AUC	AUPR
CDK extended	0.8577	0.1062	0.7867	0.0649	0.5863	0.0172	0.5000	0.0156
MACCS	0.7663	0.0582	0.6898	0.0351	0.5187	0.0141	0.5000	0.0156
PubChem	0.8536	0.0818	0.7962	0.0481	0.6590	0.0234	0.5000	0.0156
NS-descriptor	0.9535	0.2058	0.9304	0.1341	0.7521	0.0436	0.5000	0.0156
KCF-S descriptor	0.9772	0.3283	0.9837	0.3202	0.7039	0.0315	0.5000	0.0156

usually outperforms the BASELINE method in terms of the average of the computed self-ranks. An exception was observed using the median of the self-ranks in the case of the second intermediate compound of 4-step sequence, but the corresponding averages of the self-ranks in BASELINE and L1SVM are 249 and 146, respectively. These results suggest that potential intermediate compounds tend to be strongly correlated with the start and the goal compounds on metabolic pathways in terms of chemical transformation patterns.

4.3 Biochemical interpretation of the extracted substructures specific to *k*-step sequences

We analyzed the characteristics of the extracted substructures as follows. First, *k*-step-specific substructures were defined as the substructures whose obtained weights were above zero only in the L1SVM for the *k*-step sequences. Second, among all *k*-step sequences, those that contain the *k*-step-specific substructures in the start or the goal compounds were selected. Third, the obtained *k*-step sequences were ranked according to the average weights of the *k*-step-specific substructures. Finally, the obtained *k*-step sequences were compared with the reaction modules and the conserved reaction patterns.

As the result, the numbers of the obtained *k*-step sequences with *k*-step-specific substructures were 13 264, 3630, 3877 and 5276 for $k = 1$, 2, 3 and 4, respectively. Among those, the numbers of the *k*-step sequences that correspond to the conserved reaction patterns were 0, 507, 114 and 44, respectively.

Figure 4 shows some correctly predicted examples of *k*-step reaction sequences that corresponds to reaction modules, with the *k*-step-specific substructures and reaction centers highlighted in green and red, respectively. It was observed that the *k*-step-specific substructures generally do not contain reaction centers, i.e. the substructures that changes during reactions. This is understandable because the substructures that contain reaction centers are so common in metabolic pathways that they cannot be used to distinguish *k*-step reaction sequence likeness.

Although the *k*-step-specific substructures were *involved in* the conserved substructures in the start and the goal compounds, the

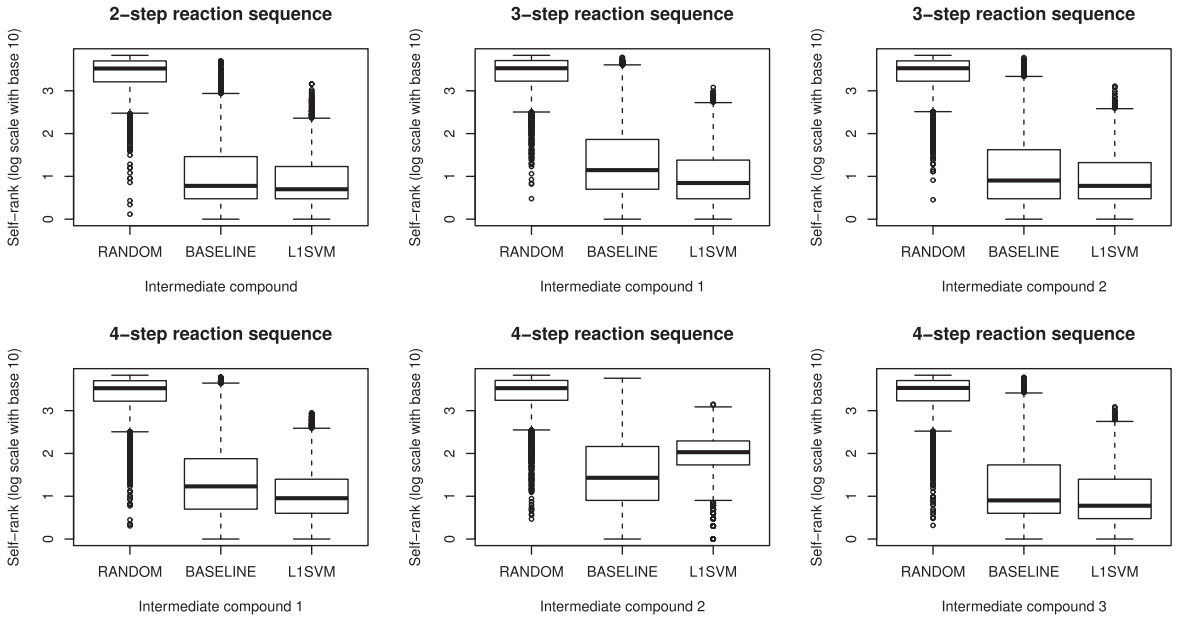

Fig. 3. Self-rank distributions for the intermediate compounds in the 2-step reaction sequences (upper left), 3-step reaction sequences (upper middle and upper right) and 4-step reaction sequences (bottom left, bottom middle and bottom right)

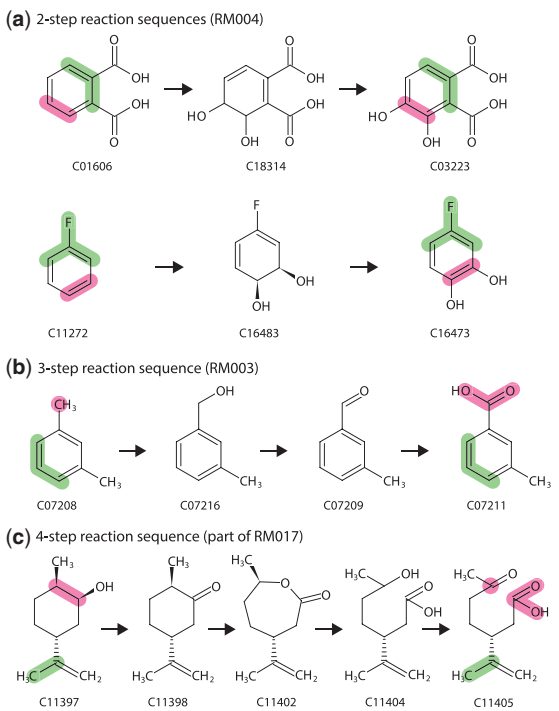

Fig. 4. Extracted substructures specific to *k*-step reaction sequences (green) and reaction center-related substructures (red)

k-step-specific substructures were not necessarily equal to the conserved substructures. For example, the first two 2-step reaction sequences in Figure 4a correspond to a reaction module RM004 [*dihydroxylation of aromatic ring, type 1* (*dioxygenase and dehydrogenase reactions*)]. As shown in the figure, carbon atoms in the substituted aromatic rings were extracted as the *k*-step-specific substructures in these sequences. However, although RM004 contains 13 reaction sequences, it appeared that only the simplest reaction sequence, i.e. from benzene (C01407) to catechol (C00090), did not have any *k*-step-specific substructures. A possible interpretation for these observations would be that benzene ring, not substituted aromatic rings, was too common to be used to distinguish *k*-step reaction likeness.

As another example of 2-step sequences, the extracted substructures from sequence 'Indole-3-acetonitrile (C02938) - Indole-3-acetamide (C02693) - Indole-3-acetate (C00954)', part of RM031 (*oxime to acetate conversion*), also was in conserved substructure (pyrrole ring) rather than the reaction center-related substructures (nitrile and carboxylate).

Degradation of aromatic compounds consists of three types of reaction modules, preprocessing, dihydroxylation and cleavage, and they can be classified into aerobic or anaerobic types (Muto *et al.*, 2013). Characteristic substructures were extracted from all 3-step sequences in RM003, a preprocessing module. Also, in this case, extracted substructures were not on reaction centers but on conserved substructures, as shown in Figure 4b. The anaerobic equivalent, RM015 (*methyl to carboxyl conversion on aromatic ring, anaerobic*), did not include characteristic substructures.

Some other representative modules, i.e. the dihydroxylation module RM004 and the following cleavage [3-step sequences 'biphenyl (C06588) - *cis*-2,3-dihydro-2,3-dihydroxybiphenyl (C06589) - biphenyl-2,3-diol (C02526) - 2-hydroxy-6-oxo-6-phenylhexa-2,4-dienoate (C01273)' and '4-chlorobiphenyl (C06584) - *cis*-2,3-dihydro-2,3-dihydroxy-4'-chlorobiphenyl (C06585) - 2,3-dihydroxy-4'-chlorobiphenyl (C06586) - 2-hydroxy-6-oxo-6-(4'-chlorophenyl)-hexa-2,4-dienoate (C06587)'], and RM017 (*ring cleavage via Baeyer–Villiger oxidation*; Figure 4c), was also extracted. There were some more sequences [e.g. 'benzamide (C09815) - benzoate (C00180) - 4-hydroxybenzoate (C00156)

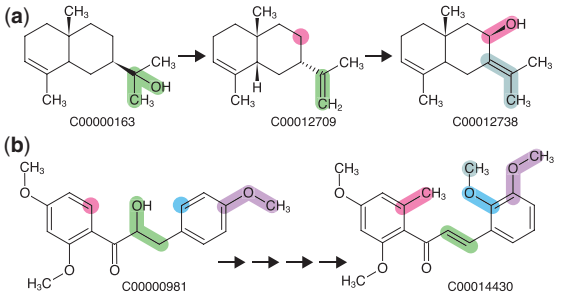

Fig. 5. Examples of falsely predicted reaction sequences. Colors represent structural changes during the reaction sequences. (**a**) The intermediate was possibly correct, but the number of steps was possibly wrong. Stereoisomerization was not considered. (**b**) Not including the distinction of geometric isomers (in purple), the number of steps was possibly correct. The intermediates were possibly wrong

- 3,4-dihydroxybenzoate (C00230) - 3-carboxy-*cis*,*cis*-muconate (C01163)'] that were not included in the previously known modules but were revealed to contain characteristic substructures. These results suggest that more intensive investigation would help the manual curation of reaction modules.

4.4 Novel prediction of multiple reaction sequences

Having confirmed the usefulness of our method, we conducted a comprehensive reaction prediction for all possible compound pairs. We trained a predictive model using all known reactant pairs in the gold standard data, and novel multiple reaction sequences were predicted using the KNApSAcK database. Start and goal compounds for this prediction were prepared by extracting compound pairs that are not too similar and not too dissimilar (weighted Jaccard coefficient between 0.6 and 0.7). The computation time was ~4 h using 40 threads in two CPUs.

The number of predicted 1-, 2-, 3- and 4-step sequences were 2 499 982, 297 295, 18 164 and 16 128, respectively. The advantage of reaction-filling approach is the quick calculation for this huge amount of pathways, which has not been possible in compound-filling approach to date. It is difficult to confirm how many of these are true positive—it was naturally observed that there were some predicted sequences whose intermediate is possibly correct, but the number of steps is possibly wrong, and *vice versa*. Examples are shown in Figure 5. Stereoisomers and geometric isomers were not distinguishable, which is a common disadvantage of using vectors including KCF-S and other fingerprints. Even though our proposed method enabled quick calculation for metabolome-scale compound sets with better AUC and AUPR, there is still room to improve, especially AUPR, for more practical use.

5 DISCUSSION

This study provided a general method to predict the number of reactions to connect two metabolites. The more the number of known reactions increases, the better the predictive performance would become. However, the recursive strategy chooses the smallest number of steps with the fewest numbers of intermediates for given start–goal compounds. There are some known cases where different organisms synthesize the same compounds using different pathways with different number of steps. The further extension will be needed to obtain possible longer pathways with keeping the computation efficiency.

This study used a predefined set of chemical substructures (KCF-S); however, it is known that some metabolic pathways use their characteristic chemical substructures. This may imply that when the users want to predict pathways for a specific group of metabolites, using the common substructures in multiple metabolites (Kotera *et al.*, 2011) would detect the metabolite-group-specific substructures, leading to the improvement of the specific pathways.

The preparation of positive and negative examples is crucial in this study. In the study of enzymatic reaction likeness (1-step likeness), distinction of positive/negative is relatively clear; positive if a compound pair corresponds to a known substrate–product pair and negative otherwise. In the study of multiple reaction sequence likeness, reversibility of reaction may affect the distinction of positive/negative. Enzymatic reactions are generally reversible *in vitro*, but they are sometimes irreversible *in vivo* depending on the physiological condition. These reversibilities are merely described in databases, making it difficult to distinguish positive/negative multistep reaction sequences.

Recent metabolomics studies enable metabolite-driven approaches for understanding previously unknown biosynthetic mechanisms at the gene level for genome-sequenced plants (Nakabayashi and Saito, 2013). We believe that this study will contribute to the understanding and the identification of metabolites and genes in the biosynthetic machinery.

ACKNOWLEDGEMENT

Computational resources were provided by the Bioinformatics Center and the Supercomputer System, Institute for Chemical Research, Kyoto University.

Funding: The Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Science and Technology Agency (JST) National Bioscience Database Center and the Japan Society for the Promotion of Science; MEXT/JSPS Kakenhi (25108714 and 24700140). The Program to Disseminate Tenure Tracking System, MEXT, Japan, the JST PRESTO program and Kyushu University Interdisciplinary Programs in Education and Projects in Research Development.

Conflict of Interest: none declared.

REFERENCES

Afendi,F. *et al.* (2012) KNApSAcK family databases: integrated metabolite—plant species databases for multifaceted plant research. *Plant Cell Physiol.*, **53**, e1.

Bono,H. *et al.* (1998) Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res.*, **8**, 203–220.

Darvas,F. (1988) Predicting metabolic pathways by logic programming. *J. Mol. Graph.*, **6**, 80–86.

Ellis,L. *et al.* (2008) The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res.*, **36**, W427–W432.

Greene,N. *et al.* (1999) Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. *SAR QSAR Environ. Res.*, **10**, 299–314.

Hatzimanikatis,V. *et al.* (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21**, 1603–1609.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kotera,M. *et al.* (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.

Kotera,M. *et al.* (2008a) Eliciting possible reaction equations and metabolic pathways involving orphan metabolites. *J. Chem. Inf. Model.*, **48**, 2335–2349.

Kotera,M. *et al.* (2008b) Functional group and substructure searching as a tool in metabolomics. *PLoS One*, **3**, e1537.

Kotera,M. *et al.* (2011) MUCHA: multiple chemical alignment algorithm to identify building block substructures of orphan secondary metabolites. *BMC Bioinformatics*, **12**, S1.

Kotera,M. *et al.* (2013a) KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction in chemical bioinformatics. *BMC Syst. Biol.*, **7** (**Suppl. 6**), S2.

Kotera,M. *et al.* (2013b) Supervised *de novo* reconstruction of metabolic pathways from metabolome-scale compound sets. *Bioinformatics*, **29**, i135–i144.

McDonald,A. and Tipton,K. (2014) Fifty-five years of enzyme classification: advances and difficulties. *FEBS J.*, **281**, 583–592.

Moriya,Y. *et al.* (2010) PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.*, **38**, W138–W143.

Muto,A. *et al.* (2013) Modular architecture of metabolic pathways revealed by conserved sequences of reactions. *J. Chem. Inf. Model.*, **53**, 613–622.

Nakabayashi,R. and Saito,K. (2013) Metabolomics for unknown plant metabolites. *Anal. Bioanal. Chem.*, **405**, 5005–5011.

Nakamura,M. *et al.* (2012) An efficient algorithm for *de novo* predictions of biochemical pathways between chemical compounds. *BMC Bioinformatics*, **13**, S8.

Newman,D. and Cragg,G. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.*, **75**, 311–335.

Ramautar,R. *et al.* (2013) Human metabolomics: strategies to understand biology. *Curr. Opin. Chem. Biol.*, **17**, 841–846.

Sreekumar,A. *et al.* (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, **457**, 910–914.

Steinbeck,C. *et al.* (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.

Talafous,J. *et al.* (1994) A dictionary model of mammalian xenobiotic metabolism. *J. Chem. Inf. Comput. Sci.*, **34**, 1326–1333.

Tanaka,K. *et al.* (2009) Metabolic pathway prediction based on inclusive relation between cyclic substructures. *Plant Biotechnol.*, **26**, 459–468.

Toya,Y. and Shimizu,H. (2013) Flux analysis and metabolomics for systematic metabolic engineering of microorganisms. *Biotechnol. Adv.*, **31**, 818–826.