Bioinformatics, 31(6), 2015, 905–911
doi: 10.1093/bioinformatics/btu750
Advance Access Publication Date: 13 November 2014
Original Paper

Systems biology

M-path: a compass for navigating potential metabolic pathways

Michihiro Araki[1,*], Robert Sidney Cox III[2], Hiroki Makiguchi[3], Teppei Ogawa[3], Takeshi Taniguchi[4], Kohei Miyaoku[5], Masahiko Nakatsui[2], Kiyotaka Y. Hara[1] and Akihiko Kondo[2,*]

[1]Organization of Advanced Science and Technology, Kobe University, Kobe 657-8501, [2]Department of Chemical Science and Engineering, Graduate School of Engineering, Kobe University, Kobe 657-8501, [3]Mitsui Knowledge Industry (MKI) Co., Osaka 530-0005, [4]MCHC R&D Synergy Center, Inc., Yokohama 227-8502 and [5]Mitsubishi Chemical Group Science and Technology Research Center (MCRC) Inc., Yokohama 227-8502, Japan

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 16, 2014; revised on October 6, 2014; accepted on November 9, 2014

Abstract

Motivation: Construction of synthetic metabolic pathways promises sustainable production of diverse chemicals and materials. To design synthetic metabolic pathways of high value, computational methods are needed to expand present knowledge by mining comprehensive chemical and enzymatic information databases. Several computational methods have been already reported for the metabolic pathway design, but until now computation complexity has limited the diversity of chemical and enzymatic data used.

Results: We introduce a computational platform, M-path, to explore synthetic metabolic pathways including putative enzymatic reactions and compounds. M-path is an iterative random algorithm, which makes efficient use of chemical and enzymatic databases to find potential synthetic metabolic pathways. M-path can readily control the search space and perform well compared with exhaustively enumerating possible pathways. A web-based pathway viewer is also developed to check extensive metabolic pathways with evaluation scores on the basis of chemical similarities. We further produce extensive synthetic metabolic pathways for a comprehensive set of alpha amino acids. The scalable nature of M-path enables us to calculate potential metabolic pathways for any given chemicals.

Availability and implementation: The web tool and viewer are available for free at http://bp.scitec.kobe-u.ac.jp/m-path/aa/.

Contact: araki@port.kobe-u.ac.jp and akondo@kobe-u.ac.jp

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

1 Introduction

Recent developments in synthetic biology and metabolic engineering have led to the construction of synthetic metabolic pathways for efficient production of various natural and non-natural chemicals. Synthetic metabolic pathways have been mostly constructed by heterologous expression of natural or non-natural enzymes, to catalytically convert metabolites into target chemicals and intermediates (Keasling *et al.*, 2010; Lee *et al.*, 2012; Martin *et al.*, 2009; Nielsen and Keasling, 2011; Soh and Hatzimanikatis, 2010; Stephanopoulos, 2012). Knowledge bases of enzymatic reactions and metabolic pathways provide troves of information to find key enzymes (Medema *et al.*, 2012) and to identify heterologous enzymes from orthologous genes in different organisms. New enzymatic activities can also be engineered by comparing related enzyme functions (Keasling et al., 2010; Lee et al., 2012). As databases of enzymatic reactions and compounds have increased in size, computational methods have become necessary to identify the key enzymatic reaction steps for efficient synthetic pathway design (Kanehisa et al., 2008; Schomburg et al., 2013).

There are two major computational approaches in designing synthetic metabolic pathways. One is solely dependent on experimentally verified information of enzymes and reactions in current knowledge bases. Synthetic pathways can be designed using a network search from starting to target compounds (Chou et al., 2009; Handorf et al., 2005; Kumar et al., 2012; McClymont and Soyer, 2013; Noor et al., 2010; Rodrigo et al., 2008; Xia et al., 2011). These methods are effective in finding known heterologous enzymatic reactions but ignore any pathway with an unknown reaction step. Another approach is based on chemical structures (Carbonell et al., 2012, 2014; Cho et al., 2010; Hatzimanikatis et al., 2005; Henry et al., 2010; Yim et al., 2011). Reaction rules are derived from the chemical structures of substrates and products in known enzymatic reactions and applied to produce possible metabolic pathways including putative compounds and enzymatic reactions. This approach can suggest previously unknown enzymatic reactions in synthetic metabolic pathways. However, the space of possible solutions grows quickly when considering the size of reaction and compound data that is contained in modern databases. Previous methods for synthetic pathway design have been limited to a small number of either reactions or compounds to avoid combinatorial explosion (Carbonell et al., 2014; Hatzimanikatis et al., 2005; Nakamura et al., 2012; Yim et al., 2011). These methods may overlook the importance of rare or poorly characterized enzymatic reactions or unknown metabolites.

Here, we introduce a computational platform, M-path, for synthetic pathway design, which makes efficient use of extensive enzymatic reaction and chemical compound databases. We developed an iterative random algorithm to design possible synthetic metabolic pathways. M-path could control the search space in finding potential synthetic metabolic pathways to a given chemical. A web-based platform was developed to check extensive metabolic pathways by ranking scores based on chemical similarities and suggests enzymes to be engineered. M-path further allows us to compute putative synthetic metabolic pathways for a set of given chemicals. We chose 6903 alpha amino acid compounds from the PubChem database to predict possible metabolic pathways and found about 40 000 putative metabolic pathways for 3543 compounds.

2 Materials and Methods

2.1 Definition of chemicals and enzymatic reactions

Chemical data are from both KEGG (Kyoto Encyclopedia of Genes and Genomes) (17 091 compounds) and PubChem (47 686 910 compounds) databases (Kanehisa et al., 2008; NCBI Resource Coordinators, 2014). Chemical structures were decomposed into lists of atom and bond types to create feature vectors of 318 atom and bond feature types by referring SYBYL MOL2 format (Supplementary Fig. S1 and Table S1): the numbers of primary, secondary and tertiary carbons were counted, and each covalent bond in a structure was recorded as pairs of atom types. Enzymatic reaction data are from KEGG (9097 reactions). The KEGG reactant–-product reaction pairs were extracted from the KEGG RPAIR database. There were 7403 main reactions, once we removed leaving group small molecule and cofactor reactions. The reactions were separated into their component steps to get 7246 substrate–product pairs. We converted the chemical structures for each of these pairs as chemical feature vectors.

The substrate and product were represented with the *chemical feature vector*, which is the count of each atom and bond type:

$$\mathit{Substrate}:\ Cs = \{d1s, d2s, \ldots, dNs\}$$

$$\mathit{Product}:\ Cp = \{d1p, d2p, \ldots, dNp\}$$

Enzymatic reactions are then defined as the structural differences between substrates and products, which can be represented by differences between two chemical feature vectors, the *reaction feature vectors*:

$$\mathit{Forward\ Reaction}:\ Rs\_p = \{d1p - d1s, d2p - d2s, \ldots, dNp - dNs\}$$

$$\mathit{Reverse\ Reaction}:\ Rp\_s = \{d1s - d1p, d2s - d2p, \ldots, dNs - dNp\}$$

We have implemented an option taking into considerations the directions of reactions for part of reactions by using reported results from thermodynamics on the basis of Gibbs free energy of formation using group contribution method (Jankowski et al., 2008). The thermodynamics data used in this work are shown in Supplementary Table S2, and we used this option in this work.

2.2 M-path algorithm

2.2.1 Definition of pathway feature

M-path is based on linear programming to find possible combinations of reaction feature vectors, which sum to produce a desired *pathway feature vector* (Fig. 1, Step 1). The pathway feature vector is the difference in chemical feature vectors between the specified start (S) and the target (T) compounds, analogous to the reaction feature vector (Ps_t):

$$\mathit{Pathway\ Feature}:\ Ps\_t = \{d1t - d1s, d2t - d2s, \ldots, dNt - dNs\}$$

2.2.2 Find combinations of reaction features by solving set covering problem

To find combinations of reaction features which sum to the pathway feature vector, we took a random iterative approach and solved a deterministic set covering problem using integer linear programming in each cycle. To this end, glpsol (http://www.gnu.org/software/glpk/) was applied to obtain single solution for one cycle (Fig. 1, Steps 2–4). Given a subset R of all reaction feature vectors, the solution provided by glpsol consists of a list of up to K reaction feature vectors that sum to the pathway feature vector:

$$\text{maximize}: \sum_{i=1}^{N} Ri^{T} Xi$$

$$\text{subject to}: P = \sum_{j=1}^{K} RjXj$$

$$\sum_{j=1}^{K} Xj \le K$$

$$Xj \in \{0, 1, 2, \ldots, K\}$$

where R represents the subset of all reaction feature vectors as matrix rows, X represents the vector of variables to be solved, K is the maximum number of reaction steps and N is the length of each feature vector.

In each iteration of the calculation procedure, the random subset R is resampled from the set of all reaction feature vectors. M-path stores each solution from every cycle and keeps only unique solutions for the next algorithm step. If no solution is found then the

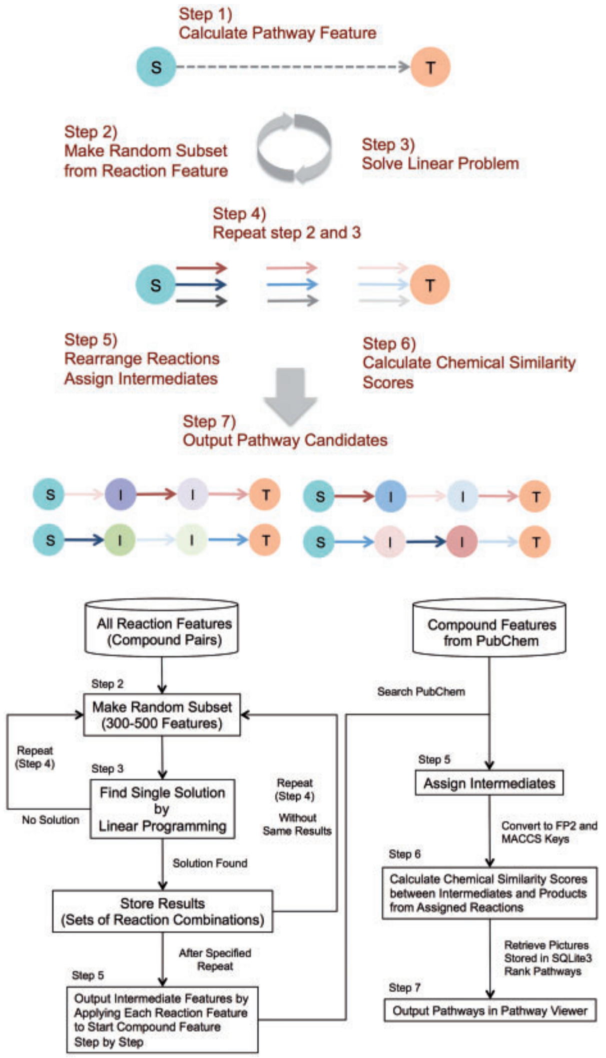


Fig. 1. (Upper) Illustrated view of the M-path algorithm. Arrows show either pathway (dotted) or reaction (solid) feature vectors. Circles indicate chemical feature vectors for start (S), target (T) and intermediate (I) compounds. Different colors show different feature vectors. M-path first calculates pathway feature from start to target compounds (Step 1). Random subsets of reactions are sampled (Step 2) to calculate combinations of reaction features that fill in the pathway feature (Step 3). Steps 2 and 3 are repeated within preset cycles to output sets of reaction combinations (Step 4). The combinations of reaction features are rearranged and intermediates are assigned (Step 5), then pathway scores are calculated to output pathway candidates (Steps 6 and 7). (Lower) Procedure of the M-path algorithm. See details in Materials and Methods

iteration is quit and the next cycle proceeds. After the specified number of iterations, each unique solution of reaction feature vectors $\{R_1, R_2,\ldots, R_k\}$ is kept. We find that the computational time of glpsol is highly dependent on the number of reaction feature vectors used for calculation at one time and that 300–500 reaction vectors are appropriate for inputs in this work to obtain results within practical time. Options such as the reaction dataset (R), the number of iteration cycles and the maximum number of reaction steps (K) allow us to control the computational time and the variety of the results.

2.2.3 Rearrange reaction order to produce pathways

Once possible combinations of reaction features are obtained, pathways are made by ordering the reaction feature vectors, then matching intermediates to each pathway (Fig. 1, Step 5). Each linear programming solution returned an unordered set of K reaction feature vectors (with K generally set to 2–6 reaction steps), which generate a set of $K!$ possible orderings of the reaction steps to make the pathway. For each of these orderings of reaction feature vectors, the intermediate chemical feature vectors are calculated by adding each reaction feature vector to the previous chemical feature vector from start compounds. Pathways including a chemical feature vector with negative values are discarded.

2.2.4 Compound assignment for intermediates

Intermediates are matched to retrieve compounds by comparing to a chemical feature database of known compounds obtained from KEGG and PubChem databases (Fig. 1, Step 5), since the conversion of chemical data into feature vectors accompanied with loss of chemical connectivity information (Supplementary Fig. S2). To allow for fast searching, we developed a chemical database where chemical feature vectors are converted into shorter strings of letters using MD5 message-digest algorithm and then stored together with other data in MongoDB (http://www.mongodb.org/), a document-oriented database for string searching and dynamic resizing (Supplementary Data S1). Multiple compounds sharing the same chemical feature vector such as chemical isomers are often found these are saved to assign a chemical similarity score.

2.2.5 Rank pathways with chemical similarity measures

We use a scoring method by chemical similarity comparison to rank the resulting pathways (Fig. 1, Step 6). Each step in a pathway consists of a list of possible enzymatic reactions (multiple candidates with the same reaction feature vector) and a list of possible reaction intermediates (multiple candidates with the same chemical feature vector). The chemical similarity score is calculated for every combination of enzymatic reaction step and intermediate. We use a chemical similarity score (Tanimoto co-efficient) using the FP2 and MACCS keys from the Open Babel toolbox to differentiate both compounds and reactions with same feature vectors (Willett *et al.*, 1998) (http://openbabel.org/wiki/Tutorial:Fingerprints). For each pathway, the intermediates are represented by fingerprint vectors containing all the FP2 and MACCS keys. The intermediate fingerprint vectors are then compared with the fingerprint vectors of the products from the assigned concrete reactions. The similarity scores are assigned to both compounds and reactions in the pathway candidates. The total score for each pathway is the average Tanimoto coefficient of the reaction similarity scores. The resulting pathways are stored in SQLite3 and implemented on pathway viewer to check each pathway candidate by visual inspection (see details in Supplementary Fig. S3 and Tutorial at http://bp.scitec.kobe-u.ac.jp/m-path/aa/) (Fig. 1, Step 7).

2.2.6 Introduction of hub compounds

In metabolic networks, highly connected compounds play a critical role in linking compounds each other (Barabási and Oltvai, 2004). As described above, the maximum number of reaction steps is a key factor for computational time. The performance of M-path calculation can be significantly improved by introducing such highly connected hub compounds as first intermediates. For implementing the hub compounds on M-path, we introduced tentative reaction features between start and/or ‘hub’ compounds as first reactions.

$$Hub\,compound : Ch = \{d1h, d2h, \ldots, dNh\}$$

$$Hub\,reaction : Rs\_h = \{d1h - d1s, d2h - d2s, \ldots, dNh - dNs\}$$

We defined 139 compounds involved in eight or more reactions in our reaction data as hub compounds to be incorporated in M-path calculation (Supplementary Fig. S4 and Table S3). The reactions between the start compound (such as glucose) and hub compounds (*Rhub*) are introduced as first steps, so that we can start M-path calculations without information of an explicit start compound. The calculation of linear programming was performed in a same manner except adding the following constraints:

$$P = Rh + \sum_{j=1}^{K-1} RjXj$$

$$Rh \in Rhub$$

2.3 Possible pathway calculations for alpha amino acids

We apply M-path to expand the synthetic pathways for 6903 alpha amino acids from PubChem database. All reaction data are used and maximum reaction steps are set to 3 and 4. A chemical database (viewer) is first constructed for each amino acid with pathway candidates found by M-path calculation. From the results of the two calculations (with maximum 2 and 3 steps from the hub compounds), we collated the resultant pathways and ordered them by the *Mscore*.

$$Mscore = Max\_pathway\{Min\_reaction\{Chemical\ similarity\}\}$$

When the *Mscore* differed between the two calculations, we took the largest Mscore as representative of the ‘best’ synthetic pathway. Pathway candidates for compounds appeared in steps 3 and 4 are often found to have overlap reactions in each other. We select reactions (edges) and compounds (nodes) in the pathway candidates for compounds with higher *Mscores* (0.7 and 0.8 or larger) to integrate all pathway data in the form of a network using Cytoscape Web (http://cytoscapeweb.cytoscape.org/) as shown in the Network 0.7 and 0.8 in the M-path website (http://bp.scitec.kobe-u.ac.jp/m-path/aa/).

3 Results

3.1 M-path performance

The reaction feature vectors represent the chemical changes that occur in each enzymatic step. We found 4196 unique reaction feature vectors including both forward and reverse reactions (Materials and Methods). We then counted how many reactions were found with the same reaction feature vector, assuming a common reaction feature vector implies similarity in enzymatic reaction (Supplementary Fig. S5 and Table S3). The number of unique substrate–product pairs with the same reaction feature vector is termed the frequency score for the reaction feature vector and is a measure of the frequency of similar reactions in the databases. For example, there are 1276 reaction feature vectors with frequency score of 2 or more, while the rest of reactions are unique. Reactions with high-frequency score include prevalent biochemical reactions such as phosphorylation, CoA-related reactions and reactions by dehydrogenase. In the M-path algorithm, the frequency score sets the reaction library size for calculation. When the frequency score threshold is set to 2, the 1276 most frequent reaction features are used, while setting the threshold to 0 includes all reactions.

The synthetic metabolic pathways between two compounds can be classified into three categories (Supplementary Fig. S6); PATH_0 with only known (KEGG) enzymatic reactions and compounds, PATH_1 with putative enzymatic reactions between known compounds and PATH_2 with putative enzymatic reactions and compounds. PATH_0 can be found in KEGG pathway database. The basic idea in finding putative enzymatic reactions in PATH_1 and 2 is to choose enzymatic reactions with the same reaction feature vector and use a chemical structure similarity score to evaluate them (Supplementary Fig. S7). Chemical similarity is often used as a measure of biological activity, which is expected to reflect the feasibility of engineering enzymes to catalyze the putative reactions as shown in previous reports on catalytic (substrate) promiscuity (Bar-Even and Salah Tawfik, 2013; Bar-Even *et al.*, 2011; Khersonsky *et al.*, 2011; Nobeli *et al.*, 2009; Rahman *et al.*, 2014; Schulenburg and Miller, 2014). The assignment of compounds is also an issue to yield PATH_2 because the chemical feature itself does not include enough connectivity information to reproduce a chemical structure. For this purpose, we constructed a database of 45 million PubChem compounds to assign chemical structures using chemical features with string compression to allow compound assignment in practical time (Supplementary Fig. S2 and Supplementary Data S1).

An M-path iteration uses a randomly selected subset from all reaction feature vectors, then uses linear programming to solve for a subset of these reaction feature vectors which sum to the pathway feature. Each iterative cycle (Fig. 1, Steps 1–4) uses a different reaction subset, so that the number of new results is expected to go to zero as the iteration number increases. To illustrate this, we applied M-path to calculate a pathway from glucose to succinate. As shown in Fig. 2, we obtained most of the results in the early cycles. This results show that M-path can find almost all possible solutions (reaction combinations) and the iterative cycles required for obtaining all solutions are highly dependent on the number of reaction feature vectors.

Each combination of reactions is then arranged to produce pathway candidates. The potential reactions and intermediate compounds are assigned with the similarity scores to aid checking the feasibility of each possible pathway (Fig. 1, Steps 5–7). The resulting

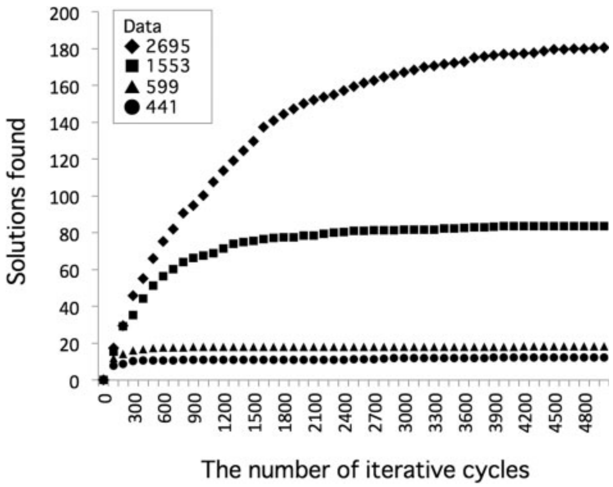


Fig. 2. Accumulated numbers of reaction combinations for the pathway features between glucose and succinate found in every 100 iterative cycles according to the number of reaction data (Fig. 1, Step 4). Each shape of marker corresponds to each number of reaction data

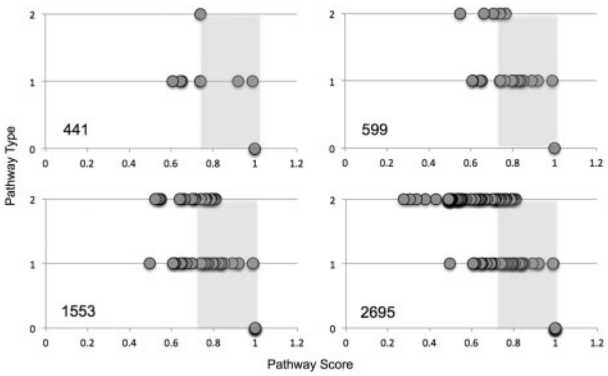


Fig. 3. Pathway scores for pathway candidates (circles) observed in each pathway type for each number of reaction data. PATH_0 consists of only known enzymatic reactions and compounds, PATH_1 includes putative enzymatic reactions and known compounds and PATH_2 includes putative enzymatic reactions and compounds. PATH_0, with pathway scores 1.0, can be found in KEGG pathway database

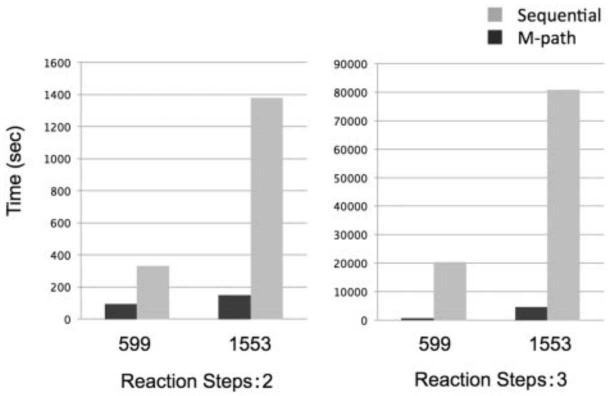


Fig. 4. Comparison of computational time between M-path (black bar) and sequential (gray bar) methods in terms of the number of reaction steps and reaction data. Each number shows the number of reaction data used for calculation

pathway candidates are presented on the pathway viewer with the similarity scores between putative and known chemical structures for each reaction (Supplementary Fig. S3). In the pathway viewer, the pathway candidates were listed according to the Mscores, and each reaction in the pathway could be checked by viewing chemical structure data. Appropriate chemical and reaction candidates can be selected by the user. We implemented additional functions on the viewer to provide an output (XGMML) file including reaction and compound information for subsequent analyses such as flux balance analysis once we fixed the pathway candidates (see details in Tutorial at http://bp.scitec.kobe-u.ac.jp/m-path/aa/).

We further checked the dependency of the results of the pathway candidates on the variety of reaction data in M-path. As expected, an increasing number of reaction data produces a larger number of PATH_1 and 2 pathways (Fig. 3). The results are promising to find new synthetic pathways, though many candidates include false-positive data. We empirically found that the pathway with score below 0.7 often included false positives due to the assignment of infeasible enzymatic reactions. This indicated that chemical similarity can be used to filter pathway assignments for feasibility. However, there are still false-positive pathways even for pathways with high scores, though these can be removed by the user in the pathway viewer.

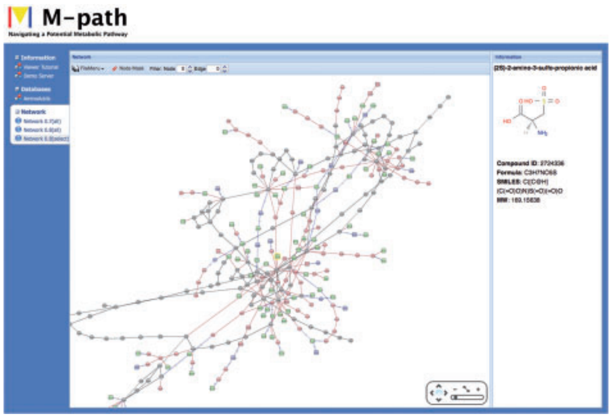
Most previous methods are 'sequential (exhaustive)'. These extend pathways from a start compound step-by-step based on reaction and chemical rules until reaching a target compound. The sequential method is effective with a limited number of compounds or reactions (Noor *et al.*, 2010; Yim *et al.*, 2011) but would not be applicable for pathway design using larger number of reaction and chemical data. To evaluate this point, we implemented the sequential method (depth-first search) on the same platform to compare the computational time of M-path with the sequential method. As a result in standard test (from glucose to succinate), M-path algorithm outperformed the sequential method where we found better performance (∼50-fold or more) in M-path in case of larger number of reaction data and reaction steps (up to five steps) (Fig. 4 and Supplementary Fig. S8).

We next ask if M-path can recapitulate expert designs of synthetic pathways reported for 1,4-butanediol (Yim *et al.*, 2011) and 3-hydroxypropionate (Henry *et al.*, 2010) (Supplementary Fig. S9). The key process in designing the synthetic metabolic pathway for 1,4-butanediol from 4-hydroxybutanoate is to find putative reactions and the intermediate compound (4-hydroxybutanal). M-path can assign them with high score by choosing known reactions (C00136_C01412 and C01412_C06142) and compound data (CID93093), respectively. The synthetic metabolic pathway for 3-hydroxypropionate (3-HP) from alpha-Alanine can be also reproduced by assigning aminomutase reactions in the first step. The reaction features of the mutase reactions tend to be null vectors in our definition, but the feature vector is not fixed by the algorithm and can be extended to substructures. We added 18 features totalling 336 to count the number of carbon atoms with respect to the number of carbon and hydrogen neighbors to discriminate the null reactions. As a result, the known reaction (C01186_C01142) was found as putative reaction in the first step. For other reported pathways to 3-HP, putative reactions from beta-Alanyl-CoA (C02335) to 3-Hydroxypropionyl-CoA (C05668) are assigned to reactions from beta-Alanine (C00099) to 3-HP (C01013), but the similarity scores are very low because of the difference in the existence of CoA moiety. The reaction-rule-based approaches in previous methods cannot explicitly suggest enzymes and genes, whereas M-path directly relates putative reactions with known enzymes (EC numbers) and genes. These results indicate that M-path is useful for making a decision to select plausible pathways for practical applications.

3.2 Expanding the scope of amino acid pathways

The potential of M-path was tested to calculate an expanded metabolic network for the production of alpha amino acids. We targeted 6903 alpha amino acids from PubChem and were able to assign putative pathways for 3543 of them. From the ordered list, we focused on the 99 compounds with an *Mscore* of 0.875 or greater. We found 50 amino acids with an *Mscore* of 1.0. From the remaining compounds, we chose 49 representatives with feasible single reaction steps (e.g. corresponding to a similar chemical change in a known reaction). We assigned the list of corresponding enzyme classification numbers to each reaction, and identified commonly occurring reaction types including transamination, side-chain linking (e.g. gamma-glutamyl transfer), methylation, acetylation, phosphorylation and degradation pathways (Supplementary Tables S5 and S6).

To classify the list of amino acid derivatives, we attached the synthetic pathways onto the core amino acid metabolic network curated by KEGG. For this purpose, we used the published KEGG pathway: biosynthesis of amino acids. We extracted the list of



(2S)-2-amino-3-sulfo-propionic acid

Compound ID: 2724336
Formula: C3H7NO5S
SMILES: C([C@H](C(=O)O)N)S(=O)(=O)O
MW: 169.15636

Fig. 5. Expanded metabolic network of amino acid production in the network viewer. The core L-amino acid biosynthetic network annotated by KEGG (gray). Naturally reachable (PATH_0) compounds are shown as pink circles. Blue edges are putative reaction steps predicted by M-path. The green squares show putative L-amino acids not found in KEGG, with blue squares as their intermediates such as keto acids

compounds and their connections from the component KEGG modules with manual curation. We reconstructed the KEGG network as the core backbone. We made the core amino acid network, colored gray, with the nodes that are amino acids and the other metabolites shown (Fig. 5). We aligned each synthetic pathway from the list of 99 amino acid derivatives to the core network. Reachable compound nodes (*Mscore* = 1.0, not represented in the core backbone) were colored red. Putative compound nodes (*Mscore* < 1.0, compounds with no red or core edges) were colored green if corresponding to an alpha amino acid and blue if an intermediate. Likewise, edges to reachable nodes were colored red, and putative edges were colored blue.

M-path found several derivatives of natural amino acids that were reachable by 1, 2 or 3 enzymatic steps from the core amino acid network. The algorithm correctly predicted 50 amino acid derivatives contained in KEGG but are not part of the core reference pathway (Supplementary Table S5). Most commonly, the side chains of lysine, glutamic acid, cysteine and serine were found to participate in several 'side-chain linking' derivative reactions. Several small molecules were found to attach to these amino acids including carboxylic acids such as acetic acid, formic acid, succinic acid and also small primary amines. Other pathways to natural targets were classified, including amino transfer reactions, catabolic degradation pathways and aromatic ring substitutions. M-path is thus useful for recapitulating known metabolic pathways.

From the L-amino acids described in PubChem but not contained in KEGG, we selected 49 predicted pathways (Supplementary Table S6). The largest number of amino acid derivatives was found in the glutamate synthesis pathway (Supplementary Fig. S10). The core network describes flux from the citric acid cycle (via 2-oxoglutarate), amidation to glutamine and the phosporylation of glutamate on the pathway for proline and arginine syntheses. Eleven examples of KEGG annotated (known) gamma-glutamyl reactions, which are outside the core reference pathway, include the attachment of an amino acid nitrogen to the carboxylic side-chain (e.g. in glutathione synthesis E.C. 6.3.2.2) and attachment of amines or cyano groups for nitrogen utilization (e.g. putrescene uptake in *Escherichia coli*) (Supplementary Fig. S11).

M-path assigned putative reactions for six gamma-glutamyl linking compounds to produce new amino acids not annotated in KEGG (Supplementary Fig. S12), by using the putrescene linking reaction (E.C. 6.3.1.11) to attach amines with different carbon chain lengths and to modify gamma-glutamyl amino attachments (E.C. 2.3.2.2, E.C. 2.3.2.14). M-path further assigned a carbon skeleton rearrangement reaction from glutamate to methyl-aspartate. Putative reactions via the same mechanism suggest two pathways to beta-methyl-asparagine, a compound with no biological annotation in KEGG or PubChem (Supplementary Fig. S13). These natural and putative glutamate derivatives illustrate how M-path can be applied to expand metabolism of amino acid synthesis, which could lead to search extensive spaces in designing synthetic metabolic pathways for other chemicals as well.

4 Discussion

We present the M-path platform to explore latent synthetic metabolic pathways of putative compounds and reactions to expand the scope of metabolic pathways. To find possible metabolic pathways for any given chemicals, one needs to consider the tradeoff between computational feasibility and the size of chemical and reaction data. Specifically, the successful design of synthetic metabolic pathways based on reaction rules is highly dependent on how chemicals and reactions are represented.

In M-path, the representation of chemicals and reactions in the form of feature vectors enables us to make efficient use of extensive chemical and reaction data. The reaction feature vectors have the ability to cover all enzymatic reactions from KEGG database and differentiate them on the basis of the chemical structures. Even though the abstraction of chemical structures to chemical feature vectors ignores information on chirality, isomer and substructure, these information can be added as additional features.

M-path uses an iterative random approach and linear programming to avoid the combinatorial explosion in exploring possible metabolic pathways. The random nature of M-path gives us a chance to obtain some results even if the number of possible metabolic pathways cannot be exhaustively enumerated in computational time. Moreover, using hub reactions as the first step makes it possible to find appropriate start compounds for any target compounds.

The evaluation of resulting metabolic pathways is also an issue in the design of synthetic metabolic pathways. Metabolic pathways have been reported to expand the scope in nature by introducing an enzyme with the alteration of substrate specificity (Bar-Even and Salah Tawfik, 2013; Bar-Even *et al.*, 2011; Khersonsky *et al.*, 2011; Nobeli *et al.*, 2009; Rahman *et al.*, 2014; Schulenburg and Miller, 2014). A compound with similar chemical structure to a substrate in a known enzymatic reaction could be a candidate for new substrate during enzyme evolution. In the same manner, M-path designs synthetic pathways including putative reactions with similar reaction chemistry to known enzymatic reactions. The chemical similarities between assigned compounds in M-path calculations and substrates or products in known enzymatic reactions can thus be an index for evaluating the applicability of the enzymes and pathways.

M-path takes advantage of the chemical similarity to calculate average scores for each reaction using chemical fingerprints to differentiate the resulting pathways. The results here show that the chemical similarity score is one evalution index. Additional methods for pathway evaluation have been proposed (Cho *et al.*, 2010; Hatzimanikatis *et al.*, 2005; Yim *et al.*, 2011), but M-path platform is first designed to allow user to check results without excluding all possibilities in the process of making decision. For this purpose, the pathway viewer visualizes all possible metabolic pathways and to output user-selected results in the form of either pathway or network data for subsequent analyses such as flux balance analysis (Orth *et al*., 2010; Yim *et al*., 2011).

Due to the increasing size of available data, it is preferable to implement scalable options for dealing with the data size without significant loss of information in the design platform. M-path allows various options to control the design space. For example, we can readily extend additional information such as chirality, isomer and substructure in the form of chemical feature vectors by adding the difference in SMILES strings as descriptors in feature vectors (Supplementary Table S7). Conversely, we can downsize the chemical features by hierarchical integration of atom and bond types as proposed in previous methods (Cho *et al*., 2010). The reaction feature vector can also be re-defined in higher resolution integrating auxiliary compounds such as co-factors into chemical feature vector. M-path can also be expanded beyond the specific chemical and reaction data used in this study. The set of reaction feature vectors is increased up to 20 000 by using a manually curated BRENDA database in the current version of M-path, and newer PubChem data can also be updated. The scalability of M-path will lead to more extensive and precise design of synthetic metabolic pathways.

Funding

This work was supported by Grant-in-Aid for Scientific Research (KAKENHI) on Innovative Areas [No. 24119508 and 24119711]; Mitsubishi Chemical Group Science and Technology Research Center; Takeda Science Foundation; Creation of Innovation Centers for Advanced Interdisciplinary Research Areas (Innovative Bioproduction Kobe, iBioK) and the commission for Development of Artificial Gene Synthesis Technology for Creating Innovative Biomaterial from the Ministry of Economy, Trade and Industry (METI), Japan.

Conflict of Interest: none declared.

References

Barabási,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*., **5**, 101–113.

Bar-Even,A. and Salah Tawfik,D. (2013) Engineering specialized metabolic pathways—is there a room for enzyme improvements? *Curr. Opin. Biotechnol*., **24**, 310–319.

Bar-Even,A. *et al.* (2011) The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, **50**, 4402–4410.

Carbonell,P. *et al.* (2012) Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Syst. Biol*., **6**, 10.

Carbonell,P. *et al.* (2014) XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Res.* **42**(Web Server issue):W389–W94.

Cho,A. *et al.* (2010) Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst. Biol*., **4**, 35.

Chou,C.-H. *et al.* (2009) FMM: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic Acids Res*., **37**, W129–W134.

Handorf,T. *et al.* (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol*., **61**, 498–512.

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

Henry,C.S. *et al.* (2010) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol. Bioeng*, **106**, 462–473.

Jankowski,M.D. *et al.* (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J*., **95**, 1487–1499.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res*., **36**, D480–D484.

Keasling,J.D. (2010) Manufacturing molecules through metabolic engineering. *Science*, **330**, 1355–1358.

Khersonsky,O. *et al.* (2011) Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions. *Biochemistry*, **50**, 2683–2690.

Kumar,A. *et al.* (2012) MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics*, **13**, 6.

Lee,J.W. *et al.* (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol*., **8**, 536–546.

Martin,C.H. *et al.* (2009) Synthetic metabolism: engineering biology at the protein and pathway scales. *Chem. Biol*., **16**, 277–286.

McClymont,K. and Soyer,O.S. (2013) Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways. *Nucleic Acids Res*., **41**, e113.

Medema,M.H. *et al.* (2012) Computational tools for the synthetic design of biochemical pathways. *Nat. Rev. Microbiol*., **10**, 191–202.

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

NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **42**(Database issue): D7–D17.

Nielsen,J. and Keasling,J.D. (2011) Synergies between synthetic biology and metabolic engineering. *Nat. Biotechnol*., **29**, 693–695.

Nobeli,I. *et al.* (2009) Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol*., **27**, 157–167.

Noor,E. *et al.* (2010) Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Mol. Cell*, **39**, 809–820.

Orth,J.D. *et al.* (2010) What is flux balance analysis? *Nat. Biotechnol*., **28**, 245–248.

Rahman,S.A. *et al.* (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, **11**, 171–174.

Rocha,I. *et al.* (2010) OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst. Biol*., **4**, 45.

Rodrigo,G. *et al.* (2008) DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics*, **24**, 2554–2556.

Schomburg,I. *et al.* (2013) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res*., **41**, D764–D772.

Schulenburg,C. and Miller,B.G. (2014) Enzyme recruitment and its role in metabolic expansion. *Biochemistry*, **53**, 836–845.

Soh,K.C. and Hatzimanikatis,V. (2010) DREAMS of metabolism. *Trends Biotechnol.* **28**, 501–508.

Stephanopoulos,G. (2012) Synthetic biology and metabolic engineering. *ACS Synth. Biol*., **1**, 514–525.

Willett,P. *et al.* (1998) Chemical similarity searching. *J. Chem. Inf. Model*., **38**, 983–996.

Xia,D. *et al.* (2011) MRSD: a web server for metabolic route search and design. *Bioinformatics*, **27**, 1581–1582.

Yim,H. *et al.* (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat. Chem. Biol*., **7**, 445–452.