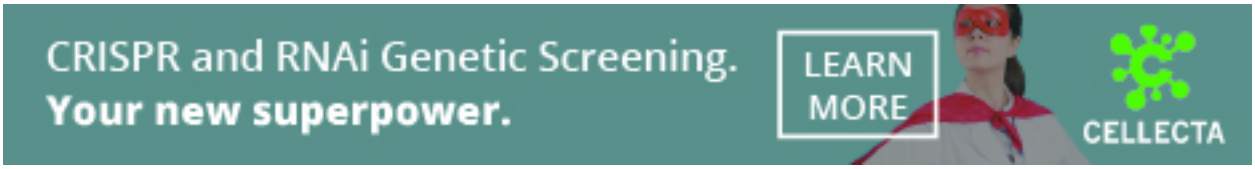
In Silico Atomic Tracing by Substrate-Product Relationships in *Escherichia coli* Intermediary Metabolism

Masanori Arita

Genome Res. 2003 13: 2455-2466
Access the most recent version at doi:10.1101/gr.1212003



In Silico Atomic Tracing by Substrate–Product Relationships in *Escherichia coli* Intermediary Metabolism

Masanori Arita

Department of Computational Biology, Faculty of Frontier Sciences, The University of Tokyo and PRESTO, JST, 277-8561 Kashiwa, Japan

We present a software system that computationally reproduces biochemical radioisotope-tracer experiments. It consists of three main components: A mapping database of substrate–product atomic correspondents derived from known reaction formulas, a tracing engine that can compute all pathways between two given compounds by using the mapping database, and a graphical user interface. As the system can facilitate the display of all possible pathways between any two compounds and the tracing of every single carbon, nitrogen, or sulfur atom in the metabolism, it complements and bridges other metabolic databases and simulations on fixed models.

[The software and data files for the *Escherichia coli* metabolism (MOL-format files for molecular structures, reaction formulas for enzymes, and mapping information at the atomic level) are available at http://www.metabolome.jp/.]

Complete genomes are now available for a variety of species, and the integration of genomic data with other biological information will yield a deeper understanding of the mechanisms of life. Pathway databases, the primary reference for the metabolism of annotated genomes, manage biochemical data in symbolic form for information retrieval and deduction (Karp 2001). They systematically integrate genomic, proteomic, and metabolomic data and offer reconstructed pathways for annotated genomes. The next step in this ongoing undertaking demands computer-aided discovery of metabolic networks. In particular, computer prediction of metabolic pathways from a set of enzymatic reactions is essential for the reaction-based reconstruction of metabolic maps.

This seemingly simple requirement cannot be addressed by merely arranging reactions by their reactant names or atomic compositions, because the resultant sequence of reactions may fail to conserve structural moiety. To better understand the difficulties involved, consider two transferase reactions whose reaction mechanism is very common in metabolism. As used here, metabolite *X* reaches *Y* if there is a sequence of reactions from *X* to *Y* in which at least one atom in *X* is transferred to *Y*. A sequence of reactions is called a pathway if at least one substrate in the first reaction can reach at least one product in the last reaction. When *Y* is the final product reached in the pathway, we state that the pathway reaches *Y*. For acronyms of compounds such as ATP, see Table 1.

Transaldolase (EC 2.2.1.2)

In the catalytic reaction by transaldolase, the structure of Sed7P is split in the middle (Fig. 1). One part combines with GA3P to form Fru6P, and the other becomes Ery4P. Therefore, if a pathway under investigation reaches Sed7P, it can reach both Fru6P and Ery4P by adding the transaldolase reaction to the end of the pathway. However, for a pathway to GA3P, only Fru6P becomes reachable when the transaldolase reaction is added.

E-MAIL arita@k.u-tokyo.ac.jp; FAX +81-4-7136-3974.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1212003. Article published online before print in October 2003.

Serine-Pyruvate Aminotransferase (EC 2.6.1.51)

In the catalytic reaction by serine-pyruvate aminotransferase, only the amino group of L-serine is exchanged with the keto group of pyruvate (Fig. 2). Suppose a pathway under investigation reaches L-serine. If the atoms reached in the pathway include nitrogen, then L-Ala becomes reachable by adding the reaction by serine-pyruvate aminotransferase to the end of the pathway. If, on the other hand, nitrogen is not reached, only pyruvate is reachable.

Thus, the reachability of a pathway depends on two factors: (1) the atom under focus, and (2) the conserved structural moiety in the reactions. In deducing or predicting pathways from reactions, it is therefore essential to consider the molecular structure of metabolites to guarantee that some moiety is conveyed through the reactions.

There are reasons to build a database of the structural correspondence between metabolites at the atomic scale. Hereafter, let us call the atomic correspondents *(atomic) mappings*. First, mappings are consistent in each reaction; structural changes are chemical, and there is no probabilistic arbitrariness. Second, the data size is not very large; in bacteria, there are only a few thousand known variations of enzymatic reactions. Moreover, as the mappings for these reactions may share equivalent information, the description length can be compressed (see Results). Finally and most important, this information is indispensable in reliable pathway reconstruction and may be used subsequently to predict enzyme/reaction-based metabolic maps. Heretofore, the realization of a mapping database has been confronted by such limitations as the standardization of metabolite structures with chirality and the detection of atomic mappings between metabolites for all known reactions.

Here we present a software system that computationally reproduces biochemical radioisotope-tracer experiments. It consists of three main components: a mapping database of substrate–product atomic correspondents derived from known reaction formulas, a tracing engine that can compute all pathways between two given compounds by using the mapping database; and a graphical user interface. In this system, metabolite structures are represented using graphs and atomic mappings are computationally detected by a maximum common subgraph (MCS) algorithm and precompiled in the database. The detection procedure is au-

Table 1. List of Acronyms for Compounds

Acronym	Compound
ATP	Adenosine 5′-triphosphate
ADP	Adenosine 5′-diphosphate
AMP	Adenosine 5′-monophosphate
CO_2	Carbon dioxide
CoA	Coenzyme A
Gal	D-Galactose
DNA	Deoxyribonucleic acid
Ery4P	D-Erythrose 4-phosphate
Fru16BP	D-Fructose 1,6-bisphosphate
Fru6P	D-Fructose 6-phosphate
GA3P	D-Glyceraldehyde 3-phosphate
Gly2P	D-Glycerate 2-phosphate
Gly3P	D-Glycerate 3-phosphate
KetoGlu	alpha-keto-glutarate
L-Ala	L-Alanine
L-Asp	L-Aspartate
L-Glu	L-Glutamate
NAD^+/NADH	Nicotinamide adenine dinucleotide
$NADP^+$/NADPH	Nicotinamide adenine dinucleotide phosphate
NH_3	Ammonia
O_2	Oxygen
OAA	Oxaloacetate
PEP	Phosphoenol pyruvate
Pi	Orthophosphate
PPi	Pyrophosphate
Rib5P	D-Ribose 5-phosphate
SAM	S-Adenosyl-L-methionine
SAH	S-Adenosyl-L-homocysteine
Sed7P	D-Sedoheptulose 7-phosphate
UDP	Uridine 5′-diphosphate
Xyl5P	D-Xylulose 5-phosphate

tomated, but the user has the option to explicitly instruct the software program about which atomic positions are to match in the substrate–product relationships. This flexibility largely eliminates the need for error-prone manual data input and helps improve the quality of molecular structures and the computed mapping information.

The use of mapping information in representing metabolic networks offers a new perspective different from conventional graph representation and analysis of metabolism (Jeong et al. 2000). Using our software system, we report the global characteristics of the metabolism of *Escherichia coli*. This paper introduces the qualitative knowledge representation of metabolic networks without addressing quantitative issues that pertain to the network, for example, flux control or reaction efficiency.

RESULTS

Graph Representation

Each metabolite structure is represented using graphs; hydrogen atoms are implicitly represented by the number of chemical bonds and atomic valences. A mapping between compound *X* and *Y* is a list of atomic position pairs, one from *X* and the other from *Y*. Hydrogen atoms are not considered in the mapping computation. The entire metabolic network is represented as a directed graph in which nodes and edges correspond to metabolites and their mappings, respectively. Hereafter, this graph is referred to as the metabolic graph.

In this representation, one enzymatic reaction corresponds to a set of edges (i.e., mappings). For example, consider three transferase reactions in the pentose-phosphate pathway (Fig. 3):

EC 2.2.1.1: Rib5P + Xyl5P = GA3P + Sed7P
EC 2.2.1.2: GA3P + Sed7P = Fru6P + Ery4P
EC 2.2.1.1: Fru6P + GA3P = Ery4P + Xyl5P

In the conventional knowledge representation (Goto et al. 1999; Jeong et al. 2000; Küffner et al. 2000; Wagner and Fell 2001; Ma and Zeng 2003), each reaction (or its equivalent data object) links two substrates with two products, leaving their structural relationships unknown. In the proposed representation, on the other hand, metabolites are linked only when they are structurally related (Fig. 3, left/right panel). By referencing metabolite structures, each reaction is automatically decomposed to a set of mappings, each of which describes the atomic correspondents between two metabolites. For example, the reaction by transaldolase (EC 2.2.1.2) comprises three (not four) mappings, as shown in different colors in Figure 1. Likewise, two reactions by transketolase (EC 2.2.1.1) comprise three mappings each (Fig. 4). Note that the mappings between Xyl5P and GA3P in these two reactions are equivalent. They are therefore represented by the same graph edge in the metabolic graph. In general, different reactions may use equivalent mappings. For example, the mapping between ATP and ADP appears throughout the metabolism. Because such mappings are shared in the metabolic graph, its edge density becomes lower than that of the conventional knowledge representation; the node degree in the proposed graph represents the number of structural changes for each metabolite. In the conventional representation, on the other hand, it represents the number of occurrences in the set of reactions. The procedure for computing mappings is described in the Methods section.

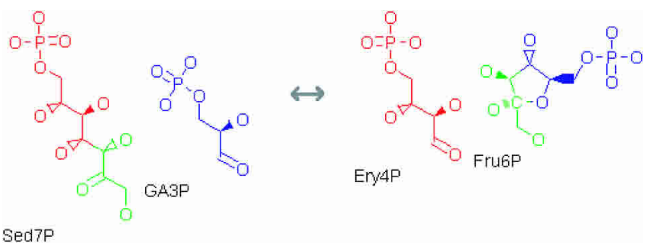


Figure 1 The reaction by transaldolase (EC 2.2.1.2).

Figure 2 The reaction by serine–pyruvate aminotransferase (EC 2.6.1.51).

With the mapping information, the tracing engine can follow any single carbon (or nitrogen or sulfur) atom in the metabolism. To list all possible pathways between compounds *X* and *Y*, we compute all graph paths between them in the order of length, and then we verify each computed graph path using the mapping information to determine whether *X* reaches *Y*. The procedure requires listing all graph paths, a criterion that is not satisfied by the standard shortest-path algorithm (Dijkstra 1959). Instead, we adopted the *k*-shortest paths algorithm that can compute any *k*-th shortest path (Eppstein 1998). Because metabolic networks contain loops, the number of graph paths may be infinite. The actual computation to find longer graph paths is performed on demand, and theoretically there is no limit on the maximum length of pathways that can be computed. The correctness of pathways depends solely on the correctness of the mapping information.

For example, consider pathways between Xyl5P and Ery4P

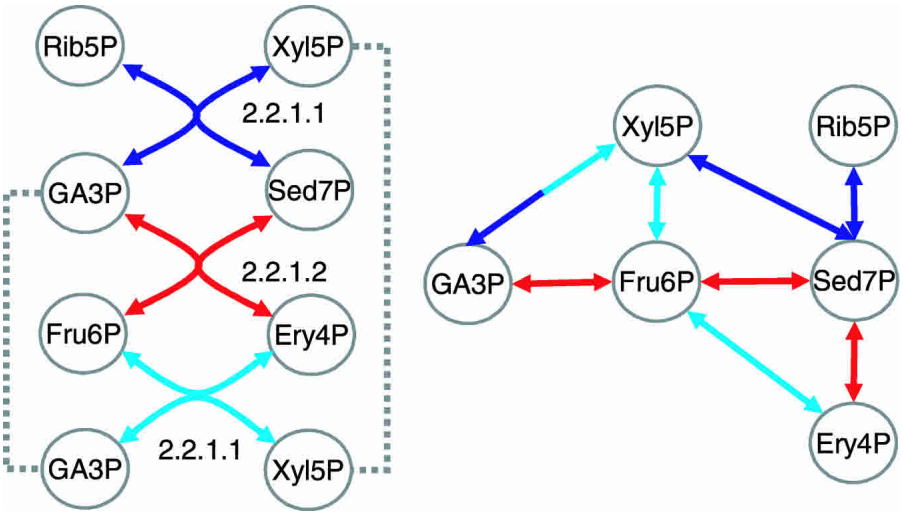


Figure 3 Three reactions in the pentose–phosphate pathway. Two are catalyzed by transketolase (EC 2.2.1.1) and the other by transaldolase (EC 2.2.1.2). Dotted lines connect equivalent metabolites. (*Left* panel) The conventional description. (*Right* panel) Proposed graph representation in which each reaction corresponds to three mappings.

in Figure 3. In the proposed metabolic graph (right panel), graph paths are verified in the following order:

1. Xyl5P -EC2.2.1.1→ Sed7P -EC2.2.1.2→ Ery4P
2. Xyl5P -EC2.2.1.1→ Fru6P -EC2.2.1.1→ Ery4P
3. Xyl5P -EC2.2.1.1→ Sed7P -EC2.2.1.2→ Fru6P -EC2.2.1.1→ Ery4P
4. Xyl5P -EC2.2.1.1→ GA3P -EC2.2.1.2→ Fru6P -EC2.2.1.1→ Ery4P, and so forth.

The first three graph paths are rejected because Xyl5P does not reach Ery4P along them. The fourth path, however, conserves three carbon atoms in Xyl5P, and therefore forms a pathway (Fig. 5). The pathway between Xyl5P and Ery4P thus requires at least three reactions, although the two metabolites appear to be only two reactions apart in either graph representation. In the conventional illustration (Fig. 6), because the same compound appears at multiple positions, the fourth path can be displayed in two ways. Note that not all the computed pathways are biologically active or important. The tracing engine only exhausts possibilities based on the mapping database, and users are expected to select candidate pathways.

Statistics of Mappings

The curated and precompiled data set represents the metabolism of 2791 reactions in 1751 EC sub-subclasses. It is essentially a subset of reactions from the Enzyme Nomenclature (http://us.expasy.org/enzyme; ENZYME database) and was prepared to cover reactions that appear in the metabolic maps (GIF pictures) in the KEGG and EcoCyc databases, and the basic metabolism in the Roche Biochemical Pathways chart (Michal 1999; Kanehisa et al. 2002; Karp et al. 2002). The major change in our curation process consists of the addition of chirality information for metabolites and the balancing of the left/right side of reaction formulas (see Methods). We expect that the curated set represents major reactions in the basic metabolism and we call it the reference metabolism. All data were cross-checked with published information (Budavari 1996; Imabori et al. 1998). In the data compilation step, atomic mappings were computed for 2764 reactions; the remaining 27 were not compiled into the graph because their formulas were unbalanced or their mappings undetectable. Reactions involved in macromolecule metabolism (e.g., protein, DNA, and tRNA) were either rewritten so that mappings were properly computed or removed from the data set (see Methods).

For carbon metabolism, the reconstructed reference network consisted of 2100 metabolites and 3172 mappings. For nitrogen metabolism, it comprised 1063 metabolites and 1546 mappings. The number of metabolites in the nitrogen network is smaller because fewer metabolites contained nitrogen atoms. Both networks were very sparse; on average there were only 1.5 mappings per metabolite, indicating that most metabolites underwent structural changes in only two or fewer different patterns in the reference metabolism.

The gene annotation for *E. coli* was obtained from the KEGG database (Kanehisa et al. 2002). The total 727 annotated metabolic genes accounted for 1193 reactions in 605 EC sub-subclasses. Among them, mapping information was computed for 1178 reactions. The network consisted of 1119 metabolites and 1492 mappings for carbon metabolism, and of 623 metabolites and 757 mappings for nitrogen metabolism. The numbers are subject to change depending on the annotation and the reactions defined for each EC sub-subclass, but, in general, we can extrapolate that the current annotation of the *E. coli* metabolism represents half of the structural

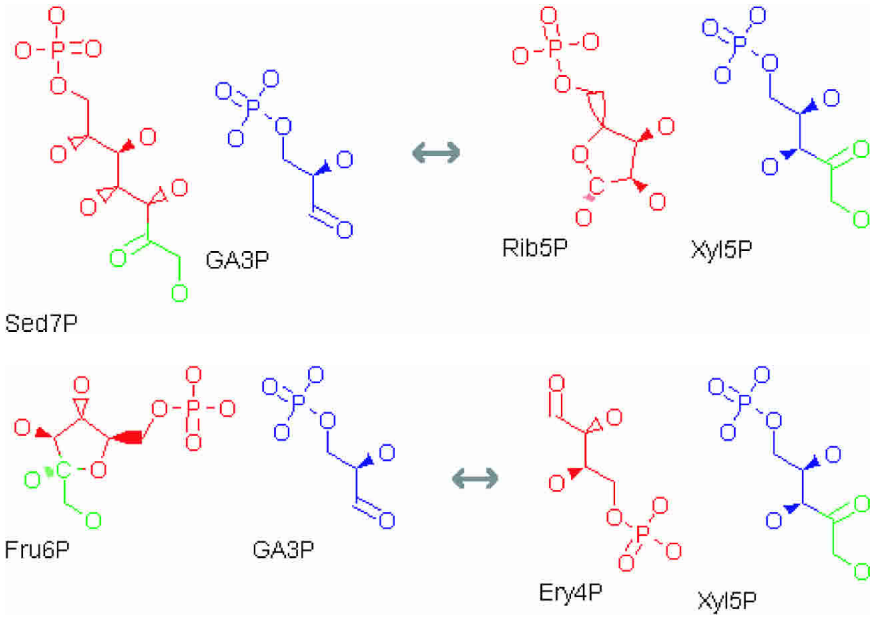


Figure 4 Two reactions by transketolase (EC 2.2.1.1).

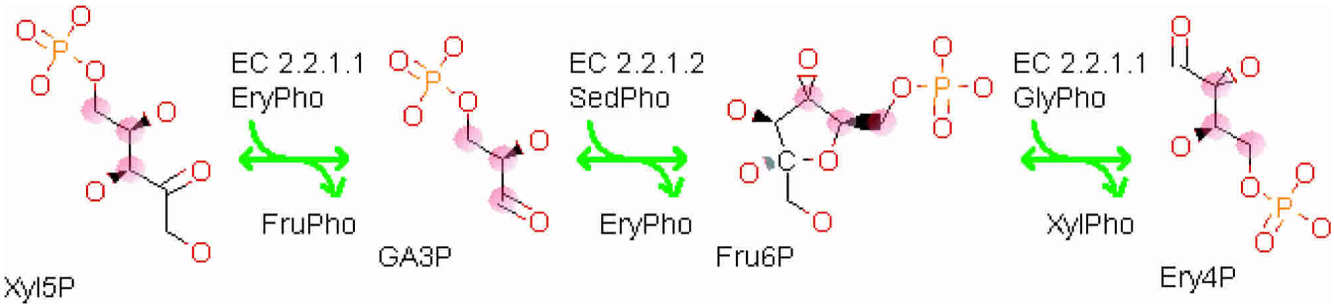


Figure 5 A valid pathway from Xyl5P to Ery4P. Positions highlighted in red show the conserved moiety (carbon atoms) in this pathway.

changes (not the number of reactions) in the reference metabolism.

The distribution of reactions and their mappings in each EC class is shown in Figures 7 and 8, respectively. The distribution for the *E. coli* metabolism roughly coincided with that for the reference metabolism. Compared with published results (Ouzounis and Karp 2000), EC class 3 (hydrolase) was significantly underrepresented in our data set. This is because hydrolases are often involved in a reaction whose substrate-product relationship is unknown (e.g. peptidase and nuclease). Such reactions of unspecified reactants were not registered in our database. In terms of computability of the mapping information, EC class 6 (ligase) presented a similar problem because no exact atomic mapping can be determined for polymerizing reactions. The problem for ligase was resolved, however, by rewriting reaction formulas to be able to compute their mappings (see Methods). In both figures, the number of mappings showed a different distribution from that of reactions because the average number of mappings for each reaction varies among EC classes (see Discussion).

Accuracy of the Mapping Computation

In the mapping database, only carbon, nitrogen, and sulfur atoms were registered; other atoms were considered only in the computation of mappings. One reason was the reduction of the data size, and the other was the difficulty in determining correct mappings for oxygen atoms. Because water molecules are often involved in catalysis, it was frequently impossible to determine correct mappings for this atomic element. (Note that the mapping for hydrogen atoms was not computed.) The correctness of the results was manually verified for metabolically important elements, that is, carbon, nitrogen, and sulfur.

Incorrect mappings were corrected by explicitly telling the software program the atomic positions to be paired. The number of mappings that required operator instructions was <2% of the total mappings (63 of 3478). Failed mappings were classified into six categories. The number in parentheses as follows is the number of currently specified instructions in each category.

1. Isomeration and dimerization of sugars (27). Because chirality was not considered in the MCS algorithm, it was often impossible to detect the correct alignment between sugar backbones. Examples include transformation from Fru16BP to Fru6P (EC 2.7.1.11), from D-glucose to D-fructose (EC 5.3.1.5), and from D-glucose to maltose (EC 2.4.1.8).
2. Complex cyclization or rearrangement (13). Transformations such as the synthesis of lanosterol (EC 5.4.99.7) or the lyase reaction of porphyrinogen (EC 4.2.1.75) were too complicated to detect with our MCS procedure (see Methods).
3. Carbon skeletal rearrangement (5). The rearrangement from a linear chain into a branched form requires explicit positioning. Examples include transformation from succinyl-CoA to methylmalonyl-CoA (EC 5.4.99.2), and from L-Glu to L-threo-3-methylaspartate (EC 5.4.99.1).
4. Shift of chemical groups (7). Because the MCS algorithm calculates the size of a common subgraph as its number of nodes, the shift of a large chemical group was not correctly detected. Examples include transformation from Gly2P to Gly3P (EC 5.4.2.1), and from chorismate to prephenate (EC 5.4.99.5).
5. Transfer of small moiety (6). Because small moieties can match at multiple positions, we had to specify their correspondents in some transferase reactions. Examples include transformation from acetyl-CoA to acetate, and from lactoyl-CoA to lactate (EC 2.8.3.1). The treatment of small moieties is discussed in the Methods section.
6. Symmetric or dimeric structures (5). When 2 equivalent molecules are coupled to form one compound, it is often difficult to find their correct alignments. The same problem arises

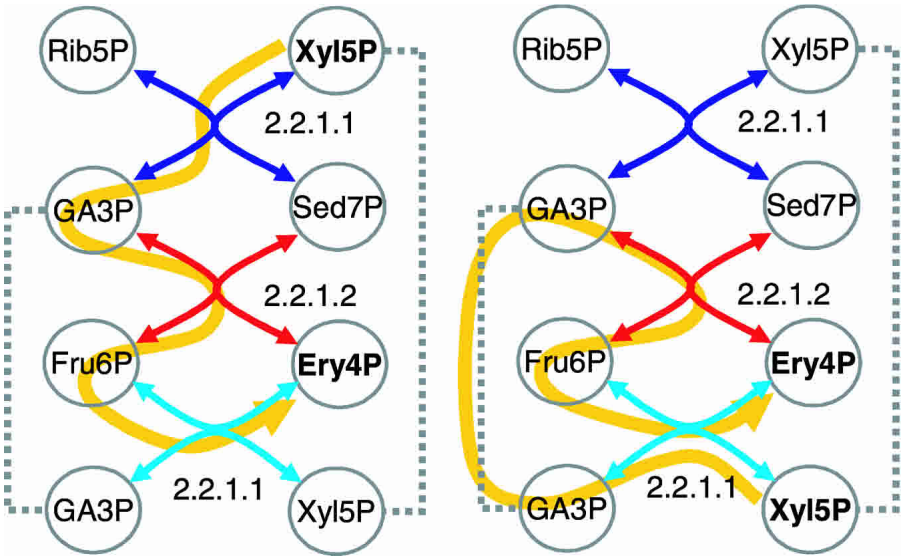


Figure 6 Two displays of the same pathway from Xyl5P to Ery4P.

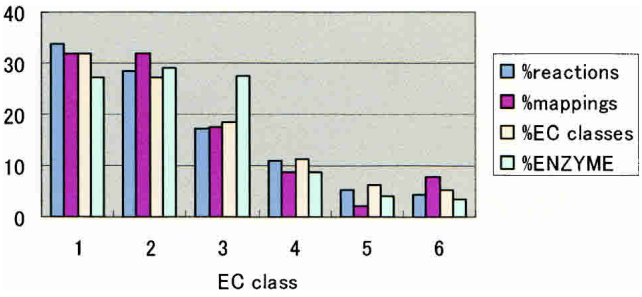


Figure 7 The number of reactions, mappings, and EC sub-subclasses in the reference metabolism against the total number of EC sub-subclasses in the LIGAND database. The blue, red, and yellow bars (%reactions, %mappings, and %EC classes, respectively) signify the percent contribution of each EC class for all reactions, mappings, and EC sub-subclasses in the reference metabolism (the six classes total 100%); the skyblue bar (%reactions) signifies the percent contribution of the EC sub-subclasses in the LIGAND database.

when the molecule is almost symmetric. Examples include transformation from 3-dehydro shikimate to shikimate (EC 1.1.1.25), urate to allantoin (EC 1.7.3.3), and 3-hydroxyanthranilate to cinnavalininate (EC 1.11.1.6).

Note that the number of instructions is small considering the catalytic variety of enzyme reactions. For most reactions, rearranging metabolite orders in a reaction suffices for finding correct mappings. For details, see the Methods section.

Central Metabolites

Table 2 lists the metabolites that appear most often in the reference metabolism. The list of high-ranking metabolites roughly agrees with the similarly computed result using 2921 reactions from the ENZYME database (Mummaneni and Kazic 2002). The observed reshuffling of metabolites comes from our data curation process, that is, standardization of metabolite names and the rewriting of multiple names such as "NADP(+),NADPH". UDP was underrepresented in our data set because it is often involved in reactions for peptidoglycan and glycolipid biosynthesis (EC 2.4.1.-), which was not represented in our data set. Similarly, the underrepresentation of SAM and SAH is due to their methyl transfer to large molecules that do not appear in basic metabolism, and whose structures could not be identified in our curation.

In Table 3, our data set for *E. coli* metabolism is compared with the high-ranking metabolites in 744 reactions of the *E. coli* small-molecule metabolism (Ouzounis and Karp 2000). For these metabolites, we also counted their occurrence in all reactions except for transport reactions and two-component phosphorylation reactions in the current EcoCyc database (http://www.ecocyc.org; EcoCyc database version 7.1). In comparison, NAD and NADP were significantly overrepresented in our data set because of our curation, especially the duplication for multiple names such as "NAD(P)" (see Methods).

In the proposed metabolic graph, the list of high-degree nodes is quite different from the lists in Tables 2 and 3, because the node degree indicates the number of structural changes of metabolites rather than their occurrences in reactions. Table 4 shows the most variably transformed metabolites, that is, nodes of high degree in the metabolic graph of the reference metabolism. When all mappings in the reference metabolism were enumerated, most variable were CO_2, involved in various decarboxylation reactions, and SAM, involved in various methyl-transfer reactions. Both metabolites transfer only one carbon atom in their reactions. When the size of transferred moiety was restricted to two or greater, the rank of the two compounds was immediately lowered and more functionally central metabolites took on high positions, for example, acetyl-CoA, CoA, pyruvate, D-glucose, and acetate. When the mapping size was restricted to three or more, the rank of acetyl-CoA and acetate was also lowered. Our approach thus effectively excludes coenzymes and inorganics, and characterizes biochemically central metabolites by focusing on the number of structural changes. Although CoA stands for coenzyme A, its centrality in metabolism is clear from its proximity to acetyl-CoA. The position of CoA remained high because it is involved in diverse transferase reactions.

In Silico Radioisotope-Tracer Experiment

The transformation of all reactions into a single graph facilitates application of graph-theoretic techniques for pathway reconstruction. The exclusion of specified compounds and reactions from the reconstruction is realized by hiding their corresponding nodes and edges in the metabolic graph. For further flexibility, any weight of positive value can be assigned to edges in the metabolic graph. Because pathway computation is performed by the *k*-shortest paths algorithm, by changing the edge weights, pathways can be searched for different purposes: Typical examples include pathways preferring or avoiding specified metabolites, or pathways containing as many specified reaction types as possible. Also realized is the function to trace specific atoms (carbon, nitrogen, or sulfur) because the mapping information is at the atomic scale.

All operations are performed through the graphical user interface of the software program. The software is especially useful for studying pathways that are not illustrated in metabolic maps. For example, consider tracing a nitrogen atom from l-Asp to L-Ala in *E. coli*. This microorganism grows on L-Asp as its sole nitrogen source, and NH_3 is mainly assimilated in the reaction either by glutamate dehydrogenase (EC 1.4.1.2) or glutamine synthetase (EC 6.3.1.2; Goux et al 1995). Suppose we are interested in the route of L-Asp catabolism to assimilate its nitrogen into other amino acids. In this example, we set the edge weights in the metabolic graph as follows:

2 . . . reactions that appear in the *E. coli* genome annotation, and
10 . . . otherwise.

These edge weights are user dependent, and here we chose arbitrary values only to preferentially select reactions annotated in *E. coli*. The relevant metabolic subgraph constructed from about 20 shortest pathways is shown in Figure 9. Blue arrows indicate EC reactions that appear in the *E. coli* annotation in the KEGG database, and green dashed arrows indicate reactions in the reference metabolism but not in *E. coli*. The shortest pathway was formed with reactions in the current annotation, but the pathways that bypass L-Glu require the reaction of either beta-

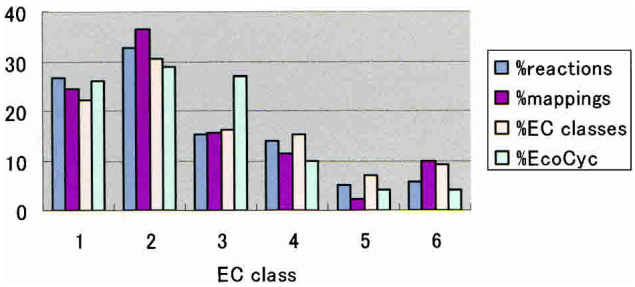


Figure 8 The number of reactions, mappings, and EC sub-subclasses in the *E. coli* metabolism against the total number of EC sub-subclasses in the EcoCyc database. Colors conform to those in Fig. 7.

Table 2. Most Frequently Used Metabolites in the Reference Metabolism

Reference set		Mummaneni and Kazic (2002)	
Compound	Occurrences	Compound	Occurrences
H_2O	961	$H_(_2_)O$	932
NAD^+	344	ATP	318
NADH	336	O_2	298
ATP	311	$NAD(^+)$	254
O_2	271	NADH	245
$NADP^+$	249	$NADP(^+)$, NADPH	233
NADPH	246	ADP	227
Pi	231	Phosphate	215
ADP	231	CoA	200
CO_2	219	$CO_(_2_)$	197
NH_3	171	Diphosphate	171
CoA	165	$NH_(_3_)$	164
PPi	154	UDP	135
L-Glu	96	S-Adenosyl-L-methionine	128
AMP	91	S-Adenosyl-L-homocysteine	120
Pyruvate	90	AMP	98
KetoGlu	88	$H_(_2_)O_(_2_)$	89
H_2O_2	83	Pyruvate	85
SAM	73	2-Oxoglutarate	79
SAH	69	L-Glutamate	78

Compound names on the right conform to Mummaneni and Kazic (2002). Originals are written in ASCII text characters and parentheses indicate subfixes.

alanine-pyruvate transaminase (EC 2.6.1.18) or aspartate 4-decarboxylase (EC 4.1.1.12), neither of which is found in the current annotation. In the next section, we discuss this network more in detail.

In this manner, pathways can be searched using not only reactions annotated in a particular species, but also reactions in the reference metabolism. This capability corresponds to hypothesizing alternative pathways using the reference metabolism.

DISCUSSION

Importance of Mapping Information

The ability to logically represent and compute pathways sheds new light on pathway analysis. However, because the symmetry of molecular structures must be considered in deciding whether a sequence of reactions forms a pathway, standardization of molecular names, including chirality information, is essential for this task. On the other hand, a variety of numerical methods is available when the network to be analyzed is determined (Stephanopoulos et al. 1999; Edwards et al. 2001). The missing link between the vast amount of known genomic sequences and mathematical analysis (including simulation) of a fixed model is provided by reconstruction software systems that can select candidate pathways in the metabolic network.

The pioneer in this field is Mavrovouniotis, who presented a sound algorithm for pathway reconstruction (Mavrovouniotis 1992). Although his method can output all pathways among a given set of metabolites, the level of abstraction is at the substrate level rather than the atomic. To verify the results of radioisotope-tracer experiments, individual investigators have been burdened with the task of following atomic positions because no mapping database was available. Our proposed software system is designed to eliminate this burden and provides valuable information for research and education. Ideally, pathway reconstruction should consider mapping information for all (including currently unknown) reactions in the metabolism. Although our data set is far short of complete, its size has grown to serve as a primer to analyze the basic metabolism in bacteria.

It must be noted that, in isolation, our pathway reconstruction results do not lead to new discoveries in biology. Information on genomes and proteomes, subsequent mathematical models for quantitative analysis, and finally, genetic and biochemical experiments are required to fully exploit and validate the results. In that sense, our approach complements and bridges the function of pathway databases and simulation models for hypothesis-driven system-oriented research.

Characters of EC Classes

In Figures 7 and 8, the population of mappings in six EC classes appears inconsistent with the number of reactions. However, their character is revealed by taking the average number of mappings per reaction. Table 5 enumerates all atomic mappings for each EC class allowing duplication.

When a reaction is completely intramolecular with one substrate and one product, all atoms in the substrate go to the product and the number of mappings becomes one. The low value in EC 5 indicates that structural changes in this class are relatively intramolecular. This result is certain because the class stands for isomerase. The values for EC 2 (transferase) and EC 3 (hydrolase) were around three, indicating that they share similar molecular mechanisms in terms of structural changes, that is, transfer of a structural moiety from one substrate to another. Indeed, EC 3 is considered a transfer reaction to water molecules. Although EC 1 (oxidoreductase) can also be considered a transfer of either a hydrogen or oxygen atom between substrates, its value became 2.7. This inconsistency is attributable to the type of atom transferred: Many metabolites transfer only hydrogen atoms in EC 1 (e.g., oxidoreduction using NAD), which do not explicitly appear as a mapping in our analysis; hydrogen atoms are implicitly represented in the molecular structure and not considered in the

Table 3. Most Frequently Used Metabolites in the *E. coli* Metabolism

E. coli set		Ouzounis and Karp (2000)	
Compound	Occurrences	Compound	Occurrences
H_2O	392	H_2O	205 (272)
ATP	186	ATP	152 (179)
NAD^+	157	ADP	101 (114)
NADH	153	Phosphate	100 (110)
ADP	133	Pyrophosphate	89 (103)
Pi	123	NAD	66 (91)
PPi	92	NADH	60 (82)
CO_2	84	CO_2	54 (71)
$NADP^+$	72	H^+	53 (74)
NADPH	70	AMP	49 (56)
O_2	65	NH_3	48 (55)
CoA	61	NADP	48 (60)
NH_3	60	NADPH	45 (59)
L-Glu	52	Coenzyme A	44 (54)
Pyruvate	49	L-Glutamate	43 (47)
AMP	48	Pyruvate	41 (46)
Acetyl-CoA	33	Acetyl-CoA	29 (33)
KetoGlu	32	O_2	26 (36)
Gal	25	2-Oxoglutarate	24 (29)
Acetate	22	S-Adenosyl-L-Methionine	23 (30)

Compound names on the right conform to Ouzounis and Karp (2000).

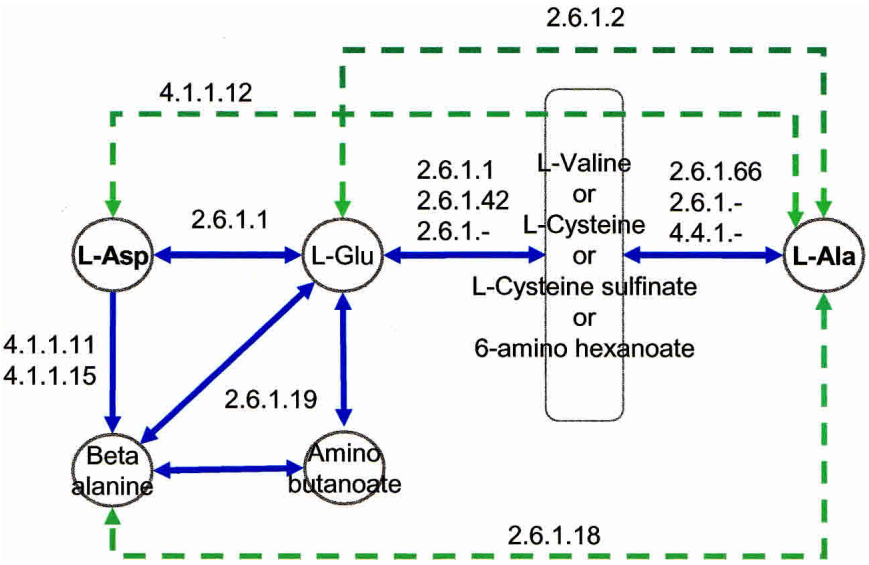


Figure 9 Relevant subgraph between L-Asp and L-Ala. Blue arrows signify reactions annotated for *E. coli* in the KEGG database, and green dashed arrows are reactions not annotated in *E. coli*. Numbers beside the arrows are ECs for enzymes.

MCS algorithm. Therefore, its value (lower than three) is an artifact of our computation.

The definition of EC 4 (lyase) is "enzymes cleaving C-C, C-O, C-N and other bonds by means other than hydrolysis or oxidation. They differ from other enzymes in that 2 substrates are involved in one reaction direction while only one participates in the other direction" (http://www.chem.qmul.ac.uk/iubmb/enzyme; IUBMB enzyme nomenclature). From this definition, we can extrapolate that the number of mappings per reaction becomes two, that is, the value when a single substrate is cleaved into two products. The observed increase in the value (2.4) is attributable to small, cleavage-associated molecules such as H_2O.

The number of mappings became larger for EC 6 (ligase), because an average of six metabolites (three substrates and three products) is involved in ligase reactions. Usually two structural moieties, one a phosphate from ATP, are transferred in a reaction, and the expected number of mappings becomes five, which is close to the observed value.

Interpretation of Coenzymes

An important aspect of the proposed representation is the impartial treatment of coenzymes. Coenzymes (or cofactors) are explained as "nonprotein organic cofactors that are required …, for an enzymic reaction to proceed" (Smith et al. 2000; abbreviation by the author) and typical examples include vitamins such as NAD and riboflavin. Sometimes SAM, L-glutamine, and other frequently appearing organics are also regarded as coenzymes. In fact, coenzymes are a concept used for the convenient understanding of the metabolism; no clear distinction can be made biochemically, and they should not be distinguished from other metabolites.

Some approaches for pathway reconstruction ignored coenzymes in their analyses without clearly defining the coenzymes (Wagner and Fell 2001; Ma and Zeng 2003). The temptation to ignore coenzymes originates in their frequent appearance in reactions. Our approach effectively resolves this problem by focusing on the structural changes of metabolites rather than their frequencies. It also realizes the analysis of the biosynthesis/degradation of coenzymes because they are not unconditionally ignored.

Comparison Against the Reference Metabolism

Coverage of the reference metabolism is an important requirement for reconstruction tools in the comparative analysis of metabolism. If our data set were based on the *E. coli* metabolism alone, the reconstructed network would be an underestimate because many *E. coli* genes remain functionally unidentified or hypothetical (Ouzounis and Karp 2000).

Although the *E. coli* metabolism is already well characterized, the possibility of discovering new pathways exists; in fact, there is evidence for missing pathways (Goux et al. 1995; Rohmer 1999). The reference metabolism is introduced, in a sense, to augment possible deficits in the current annotation.

For example, *E. coli* can grow on L-Asp as its sole nitrogen source. The nuclear magnetic resonance (NMR) result of its culture on L-(N^{15})Asp, however, showed a higher intake of N^{15} by L-Ala than L-Glu (Goux et al. 1995). This indicates an uncharacterized pathway from L-Asp to L-Ala that bypasses L-Glu under conditions of nitrogen deficiency. There are several pathways that can explain this effect (Fig. 9). Aspartate 4-decarboxylase (EC 4.1.1.12), found in bacteria as well as mammals, catalyzes the decomposition of L-Asp into L-Ala and CO_2 (Rathod and Fellman 1985). Such decarboxylation reactions consume protons and also generate an intracellular pH gradient that can drive ATP synthesis (Abe et al. 2002). Although decarboxylation is not as prevalent as aminotransfer within a cell, the contribution of the decarboxylation to the pathway from L-Asp to L-Ala cannot be dismissed.

Beta-alanine-pyruvate transaminase (EC 2.6.1.18), found in

Table 4. Most Frequently Transforming Metabolism in the Reference Metabolism (High-Degree Nodes in the Metabolic Graph)

Compound Map size	Degree (rank)		
	0<	1<	2<
CO_2	99	0	0
SAM	77	7 (73)	7 (69)
Acetyl-CoA	56	56 (1)	16 (14)
Pyruvate	47	45 (3)	37 (2)
CoA	46	46 (2)	46 (1)
D-Glucose	34	34 (4)	34 (3)
Acetate	30	30 (5)	0
UDP-D-glucose	25	25 (6)	25 (4)
Malonyl-CoA	25	22 (7)	14 (18)
ATP	21	21 (8)	21 (5)
L-Glu	20	19 (10)	19 (7)
AMP	20	20 (9)	20 (6)
UDP Gal	18	18 (11)	18 (8)
Succinyl-CoA	18	17 (12)	17 (9)
L-Asp	18	16 (14)	16 (12)
L-Lysine	17	16 (14)	16 (12)
Glycine	17	16 (14)	13 (22)
Gal	17	17 (12)	17 (9)
OAA	16	13 (27)	13 (22)
Formate	16	0	0

The numbers in parentheses indicate the ranking of metabolites.

Table 5. The Number of Reactions, Mappings, and Mappings per Reaction for Each EC Class in the Reference Metabolism

EC class	1	2	3	4	5	6
Reactions	315	387	181	166	61	68
Mappings	848	1269	544	396	75	344
Mappings/reaction	2.7	3.3	3.0	2.4	1.2	5.1

Bacillus cereus, catalyzes beta-alanine and pyruvate into 3-oxopropanoate and L-Ala. It also catalyzes beta-alanine and OAA into malonic semialdehyde and L-Asp (http://www.brenda.uni-koeln.de; BRENDA database). Although the existence of its activity can explain the intake of N^{15} into L-Ala (Fig. 10), the current annotation for *E. coli* does not contain this function. A closely related activity annotated in *E. coli* is beta-alanine-oxoglutarate transaminase (EC 2.6.1.19), coded in the genes *goaG* (b1302) and *gabT* (b2662). However, this enzyme uses L-Glu as an amino acceptor, and L-Ala was shown not to be an acceptor in *Pseudomonas fluorescens* (http://www.brenda.uni-koeln.de; BRENDA database). Its substrate specificity in *E. coli* is crucial in assessing the existence of this bypass. Our preliminary laboratory experiment revealed that neither *goaG* nor *gabT* has an activity as beta-alanine-pyruvate transaminase, and a further experiment is underway. In summary, our study can indicate possible pathways from L-Asp to L-Ala. Their proof, however, requires biochemical and genetic analysis under various growth conditions.

As indicated by the statistic that 3172 carbon mappings can cover 2764 reactions, adding extra reactions does not always introduce new mappings in the metabolic graph; most metabolites allow only a limited number of structural changes, and the same mapping information is reused in many reactions. Our graph representation is a compact description of seemingly diverse metabolic reactions, and we are cognizant of the disadvantage of using the reference metabolism for prediction: Extra reactions often did not introduce structurally new conversions, and their addition may not contribute to resolving discrepancies between tracing results obtained at the wet bench and results reproduced in silico.

It is therefore desirable that at least minor structural changes, such as oxidation, phosphorylation, and rearrangement, be exhaustively applied and hypothesized as putative enzymatic reactions to demonstrate the power of computation (Arita et al. 2000). This function is theoretically possible with structural information on compounds, but its realization is hampered when chirality changes in a structural conversion. Efforts are underway in our laboratory to implement such improvements in the next release of the software system.

METHODS

Data Preparation

Molecular structures and reaction data were prepared in MOL-format (http://www.mdli.com/; MDL Information Systems) and in reaction formulas, respectively. For the reference metabolism, reactions were manually collected from the metabolic maps in the KEGG and EcoCyc databases, and in the Roche Pathway chart regardless of their EC assignments (Michal 1999; Kanehisa et al. 2002; Karp et al. 2002). Molecular structures of substrates and products in the reactions were collected from the LIGAND database (Goto et al. 1999). All data underwent the following curation process.

Standardization of Molecular Names

In the LIGAND database, the same compound may be multiply registered depending on its three-dimensional configurations (e.g., glucose, D-glucose, alpha-D-glucose, and beta-D-glucose), and reaction formulas are written using these variants. To reduce such redundancies, we cross-checked all files for their names and molecular structures against the literature (Budavari 1996). Molecular names conformed to the dictionary published by Tokyo Kagaku Dozin (Imabori et al. 1998). For example, pyruvate was renamed pyruvic acid.

Spontaneously interchangeable configurations such as the α- and β-form of sugars were merged. Because chiral configurations (D and L) were specified only for graphical views in the LIGAND database, we manually supplied configuration data in the MOL-format files.

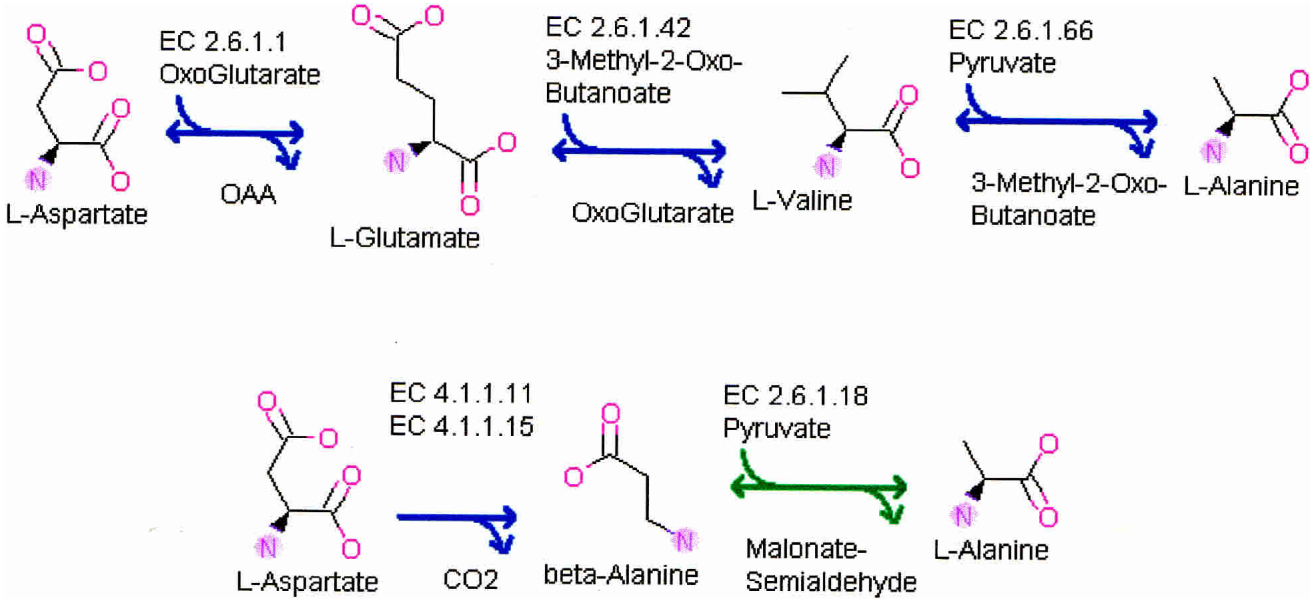


Figure 10 Two pathways from L-Asp to L-Ala in the *E. coli* metabolism. Positions highlighted in purple show the conserved nitrogen atoms in these pathways.

```
ENTRY       EC 1.1.1.1
NAME        alcohol dehydrogenase
            aldehyde reductase
            ADH
            alcohol dehydrogenase (NAD)
            aliphatic alcohol dehydrogenase
            ethanol dehydrogenase
            NAD-dependent alcohol dehydrogenase
            NAD-specific aromatic alcohol dehydrogenase
            NADH-alcohol dehydrogenase
            NADH-aldehyde dehydrogenase
            primary alcohol dehydrogenase
            yeast alcohol dehydrogenase
CLASS       Oxidoreductases
            Acting on the CH-OH group of donors
            With NAD+ or NADP+ as acceptor
SYSNAME     alcohol:NAD oxidoreductase
REACTION    an alcohol + NAD = an aldehyde or ketone + NADH2
SUBSTRATE   alcohol
            NAD
PRODUCT     NADH
            ketone
            aldehyde
COFACTOR    Zinc
COMMENT     A zinc protein. Acts on primary or secondary alcohols or
            hemi-acetals; the animal, but not the yeast, enzyme acts
    also on cyclic secondary alcohols.
REFERENCE   .... (Other fields continue below.)
```

Figure 11 Entry 1.1.1.1 in the LIGAND database.

Symmetry Detection

Many structures are symmetric, and their symmetry must be detected prior to pathway computation. For each compound structure, topologically equivalent atoms were detected using two well-known procedures in computational chemistry: the detection of aromatic rings and of topologically equivalent sites. Aromatic rings are recognized by applying the Hückel rule, that is, a ring becomes aromatic if it is planar (that is, all of the atoms in the ring must be sp2-hybridized) and has $4n + 2$ Pi electrons, for all small rings in a structure. For ring detection, the algorithm for finding the shortest cycle basis was implemented (Horton 1987). For efficiency, large rings (size >10) are not considered in our implementation. For example, the aromatic outer ring of heme is not detected in our approach. For details on these procedures, see Arita (2000a).

Preparation of Reaction Formulas

Figure 11 shows an example entry in the LIGAND database. Reaction formulas were initially collected from the REACTION field in the database, and their reactant names were rewritten into the aforementioned standardized names. Multiple names were separately described. For example, the reaction for glutamate dehydrogenase (EC 1.4.1.3) was rewritten as follows.

Original: L-Glu + H_2O + NAD(P) = KetoGlu + NH_3 + NAD(P)H
Curated: NAD^+ + L-Glu + $H_2O \rightarrow$ NADH + KetoGlu + NH_3
$NADP^+$ + L-Glu + $H_2O \rightarrow$ NADPH + KetoGlu + NH_3

The direction of reactions, described with "=" or "→", conforms to the arrow directions in the Roche Pathway chart (Michal 1999).

In some cases, a molecule without configuration (e.g., serine) was merged with its configurated form (e.g., L-serine) so that it could constitute proper known pathways. When additional reactants were written in the COMMENT field (in addition to the SUBSTRATE and PRODUCT fields) in the LIGAND database, we manually translated the information into reaction formulas and added them to the data set.

In a reaction for polymerization, it is inherently impossible to compute one-to-one atomic mappings; the size of a polymer is indeterminate. As a compromise, variable regions of polymers were substituted with their corresponding monomers or dimers in our data set. Consequently, the notation for polymers, typically described as $(\ldots)_n$, was changed to its corresponding monomeric or dimeric expression. For example, the reaction by DNA ligase (EC 6.5.1.1) was changed as follows:

Original:
ATP + $(DNA)_N$ + $(DNA)_M$ = AMP + PPi + $(DNA)_{N+M}$
Curated:
ATP + DNA + H_2O = AMP + DNA + PPi

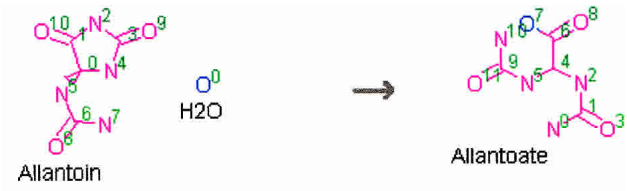
In this example, a water molecule is added only to balance the number of oxygen atoms between both sides. Thus, the function for ligase or protease is not accurately represented in our data set, and the occurrence of H_2O is biased.

Many reactions contain generic names such as "alcohol." Such abstract names were substituted with their corresponding specific names: for example, methanol, ethanol, propanol, and butanol. Specific names for each abstract name were picked from the metabolic maps in the aforementioned literature and databases, and from the set of compound names appearing in our curated reactions. Although the substitution underestimates the spectrum of enzyme specificity, it is a necessary procedure to use obtained mappings for pathway computation.

Compounds like phosphatidyl-choline include abstract chemical groups, typically described with an R notation. Although ideally all variable parts should be substituted with all possible instances, we left the R part in lipids untouched. For this reason, the lipid metabolism is not accurately represented in our data set.

Each reaction was computationally checked for the atomic balance between its left- and right-hand side. When unbalanced, appropriate compounds extrapolated from the reaction category of enzymes were supplied. When one oxygen atom was missing on either side, a water molecule was added by default. Although this complement may be enzymologically inaccurate (i.e., it may be that a molecular oxygen should be added), this procedure does not affect our mapping database because hydrogen and oxygen atoms are not registered. For this reason, the balance of hydrogen atoms was not thoroughly checked; our approach also did not consider ionization of compounds and charges.

In principle, we curated all reactions so that the connectiv-




```
SOURCE   Allantoin
TARGET   Allantoate
0,4; 1,6; 2,10; 3,9; 4,5; 5,2; 6,1; 7,0; 8,3; 9,11; 10,8;
SOURCE   H2O
TARGET   Allantoate
0,7;
```

Figure 12 Mapping information between allantoin and allantoate.

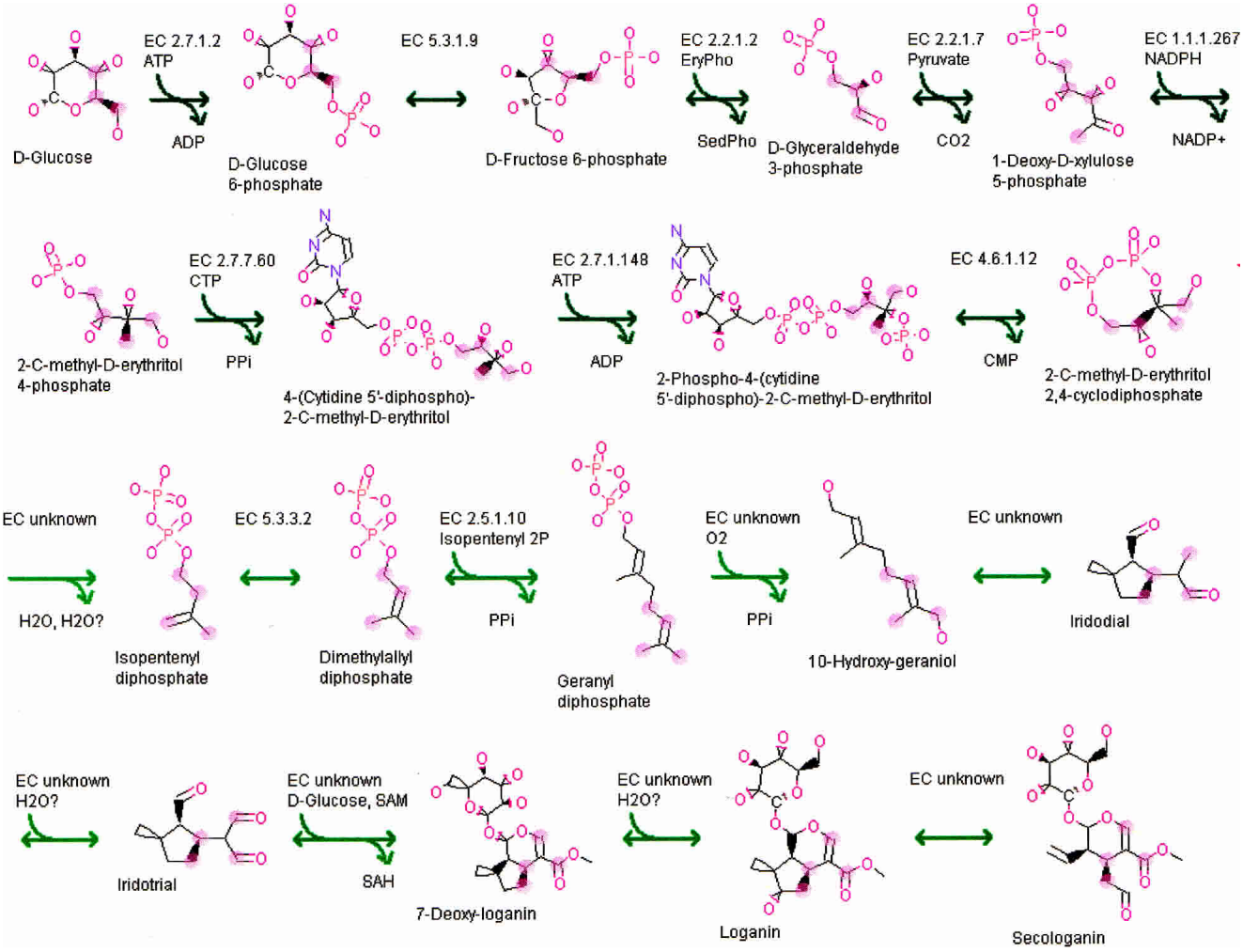


Figure 13 The pathway from D-glucose to secologanin. Positions highlighted in red show the conserved carbon atoms in this pathway.

ity of known pathways would be fully covered. This may have resulted in some over- or underestimation, especially in terms of chiral specificity. Finally, the curated reactions were cross-checked with all reactions in the EcoCyc database (Karp et al. 2002) so that no reactions were missing in our data set. The curated data files (molecular structures, reaction formulas) can be freely downloaded from http://www.metabolome.jp/.

Computation of Mappings

Finding Common Substructures

Because the general problem of finding the MCS between graphs belongs to the class NP-hard (Garey and Johnson 1979), a heuristic algorithm is used in our software system. To find a common subgraph, we use a variant of the Morgan algorithm (Wipke and Dyott 1974) together with a branch-and-bound procedure (Arita 2000b). The Morgan algorithm is a standard method to generate a unique description of compound structures. It initially assigns an integer label to every node in the given graph and updates the node labels iteratively so that topologically equivalent nodes have the same labels. In the initial phase, every node obtains a label corresponding to its degree (i.e., the number of connected edges). In subsequent iterative steps, every node label is updated as the sum of its current label and the labels of adjacent nodes. After the nth iteration, each node obtains a label that contains the information of nodes that are within n steps around it. The computed labels are used as a necessary condition to find topologically equivalent nodes in the graph structure. Obviously, the algorithm does not work for graphs whose nodes are connected by the same number of edges, called regular graphs. Such graphs, however, usually do not appear in chemical compounds (fullerene, or carbon buckyball, is a notable exception.)

In our approach, a Morgan algorithm that considers atoms and chemical bonds is used to find small common subgraphs between two given graph structures. Small subgraphs found in this manner are then extended stepwise to detect the MCS by a branch-and-bound procedure. Throughout the MCS algorithm, chirality is not considered.

Order Rearrangement of Reactants

Structural similarity alone is insufficient for detecting the correct atomic mappings in a reaction. In the reaction of serine-pyruvate aminotransferase (Fig. 2), for example, all involved metabolites have similar structures. To indicate which reactants are to be matched, we manually rearranged the order of compounds on both sides in a reaction so that correct mapping was computed. In the example reaction, the formula was written as follows:

L-serine + pyruvate = L-Ala + hydroxy-pyruvate

rather than

L-serine + pyruvate = hydroxy-pyruvate + L-Ala.

Mapping Detection

A mapping between two metabolites is defined as a list of atomic position pairs. Because there is no standard for referring to the atomic positions of compounds, the position numbers are defined as the line numbers for atoms in the MOL-format file of compounds. Therefore, mapping information depends on our structure files and is not transferable to structure data in other databases. (However, our software system is supported with a function to compute mapping information on demand for a given reaction.) Figure 12 shows the mapping between allantoin and allantoate in the reaction by allantoinase (EC 3.5.2.5): 11 atomic positions in allantoin are paired with their corresponding positions in allantoate. Note that the mapping is one to one, and molecular symmetry is not considered in the mapping data itself. In this example, although allantoate is symmetric, the mapping does not describe the atomic correspondents for the mirror pattern. The symmetry is considered in the pathway computation stage.

Structural comparison is applied pairwise for both sides of a reaction: first substrate and first product, second substrate and second product, and so forth. At each comparison step, the MCS between metabolites is computed and registered as one mapping. Then, leftover structures are collected and compared again until all atoms are matched. In the process, comparison of inorganic phosphates, CO_2, and small metabolites of less than three atoms (excluding hydrogen atoms) is deferred because such molecules could easily match several positions when compared with a larger molecule. For a clearer demonstration of this procedure, several examples are shown:

EC 3.5.4.4 reaction: Adenosine + $H_2O \rightarrow$ inosine + NH_3

In this reaction, adenosine is compared against inosine; comparison of H_2O against NH_3 is deferred. After comparing the first pair, two positions, one for NH_3 and the other for H_2O, are detected. In the second round, these positions are paired with the deferred molecules and the matching process is completed.

EC 4.1.1.49 reaction: ADP + PEP + CO_2 = ATP + OAA

In this reaction, first ADP is compared against ATP and PEP against OAA. This leaves two phosphate moieties and two single-carbon moieties. In the second round, they are paired to complete the matching.

EC 6.3.4.1 reaction: ATP + XMP + $NH_3 \rightarrow$ AMP + GMP + pyrophosphate

In this reaction, first ATP is compared against AMP and XMP against GMP, deferring comparison of NH_3 against PPi. In the next round, PPi and the leftover moiety in ATP are compared, completing the matching process.

Thus, the order rearrangement greatly contributes to the accuracy of computed substrate–product relationships without losing information of the original reaction. Characteristic moieties such as purine or pyrimidine are detected first, and small molecules such as NH_3 and H_2O are then paired with their corresponding positions left in the moieties. As described in the Results section, symmetric molecules represent negative examples of this procedure. Let us consider the conversion of urate to allantoin (EC 1.7.3.3). In our database, the conversion was described using an intermediate, 5-hydroxy isourate (5-HI), as follows:

Urate + O_2 = 5-HI + H_2O
5-HI + H_2O = CO_2 + (R)(-)-allantoin
5-HI + H_2O = CO_2 + (S)(+)-allantoin

Although 5-HI was introduced to bridge urate and allantoin, the MCS algorithm alone could not find the correct mapping. (There is more than one perfect match between 5-HI and allantoin.) For this reason, the mapping between 5-HI and CO_2 was explicitly ordered in the computation. They are paired in the first round, and the comparison of H_2O against allantoin is deferred. In the next round, the leftover moiety in 5-HI is matched with allantoin. The description using intermediates was effective, however, for citrate hydro-lyase (EC 4.2.1.3) and enzyme complexes like pyruvate dehydrogenase (EC 1.2.4.1). These reactions did not require explicit instructions (data not shown).

Computation of Pathways

Graph Construction

The set of mappings was implemented as a single network in which nodes and edges corresponded to compounds and the computed mappings, respectively. Mappings obtained from reactions with "=" and "→" were considered reversible and unidirectional (left to right), respectively. Reversible mapping was transformed to two directed graph edges and unidirectional mapping to one graph edge. The reaction by transaldolase (Fig. 1), for example, was transformed to six edges in carbon metabolism. The reaction by serine-pyruvate aminotransferase (Fig. 2) was transformed to four edges in the carbon metabolism and two edges in the nitrogen metabolism. The degree distributions in the Results section counted the number of outgoing edges; incoming edges and self-directed edges were not counted.

Computation of Pathways

Paths in the metabolic graph are computed by the *k*-shortest paths algorithm (Eppstein 1998). The algorithm enumerates all paths that may contain loops, but in our software system looping paths are removed before validation as a pathway. Therefore, metabolic pathways in which the same compound is encountered multiple times are not output in our implementation. This restriction excludes cyclic pathways such as those in the tricarboxy acid (TCA) or urea cycle, but because most enzymatic reactions are reversible, the restriction effectively excludes meaningless pathways using the same reaction more than once.

Validation of Pathways

Not all graph paths correspond to valid biochemical pathways because migration between mappings is not reflected in the graph structure. Therefore, to reconstruct pathways, one must compute transition through mappings to guarantee that at least one atom in the source compound reaches the target. The validation of pathways proceeds as follows. For each graph edge *E* from compound *X* to *Y,* candidate positions in *X* are mapped to positions in *Y* using the mapping that corresponds to *E*. Then the positions in *Y* are distributed to their equivalent sites within *Y* using its symmetry information. This process is repeated for all graph edges in the given path starting from the source compound. If at least one position in the target compound is marked, the graph path is considered valid. For the graphical display of positions, the positions in the target must be traced back to the source compound so that the system can determine exactly which atoms can reach the target.

For example, Figure 13 shows the biosynthetic pathway of secologanin. The number of carbons in D-glucose transferring to secologanin in this pathway is three, but in most compounds four positions are highlighted. This randomization effect is attributable to the equivalent positions in dimethylallyl diphosphate (and in geranyl diphosphate and iridotorial). Positions are distributed in both directions in the pathway, although some arrows are unidirectional. Positions are not distributed in the condensation of 2 isopentenyl diphosphate moieties in EC 2.5.1.10. (If distributed, a total of eight positions are highlighted in geranyl diphosphate.) The randomization effect at the equivalent positions was confirmed in an NMR study (Y. Yamazaki, M. Kitajima, M. Arita, H. Takayama, H. Sudo, M. Yamazaki, N. Aimi, and K. Saito, in prep.).

ACKNOWLEDGMENTS

I thank Yukiko Nakanishi (Intec Web and Genome Informatics Corporation) for her continuous support in data curation and management, and Dr. Aya Itoh (Institute of Advanced Biosciences, Keio University) for measuring the catalytic activity of

proteins from the *goaG* and *gabT* genes. I also thank Ursula Petralia for editing this manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Abe, K., Ohnishi, F., Yagi, K., Nakajima, T., Higuchi, T., Sano, M., Machida, M., Sarker, R.I., and Maloney, P.C. 2002. Plasmid-encoded asp operon confers a proton motive metabolic cycle catalyzed by an aspartate-alanine exchange reaction. *J. Bacteriol.* **184:** 2906–2913.

Arita, M. 2000a. Metabolic reconstruction using shortest paths. *Simulation Pract. Theory* **8:** 109–125.

Arita, M. 2000b. Graph modeling of metabolism. *J. Jpn. Soc. Artif. Intell. (JSAI)* **15:** 703–710.

Arita, M., Asai, K., and Nishioka, T. 2000. Reconstructing metabolic pathways with new enzyme classification. In *Proceedings German Conf. Bioinformatics (GCB'00)*, pp. 99–106. Heidelberg, Germany.

Budavari, S., O'Neil, M.J., Smith, A., Heckelman, P.E., and Kinneary, J.F. eds. 1996. *Merck index: An encyclopedia of chemicals, drugs, and biologicals,* 12th ed. Merck, NJ.

Dijkstra, E.W. 1959. A note on two problems in connection with graphs. *Numerische Mathematik* **1:** 269–271.

Edwards, J.S., Ibarra, R.U., and Palsson, B.O. 2001. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19:** 125–130.

Eppstein, D. 1998. Finding the k shortest paths. *SIAM J. Comput.* **28:** 652–673.

Garey, M.R. and Johnson, D.S. 1979. *Computers and intractability.* W.H. Freeman, New York.

Goto, S., Nishioka, T., and Kanehisa, M. 1999. LIGAND database for enzymes, compounds, and reactions. *Nucleic Acids Res.* **27:** 377–379.

Goux, W.J., Strong, A.A.D., Schneider, B.L., Lee, W.-N.P., and Reitzer, L.J. 1995. Utilization of aspartate as a nitrogen source in *Escherichia coli*. *J. Biol. Chem.* **270:** 638–646.

Horton, J.D. 1987. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J. Comput.* **16:** 358–366.

Imabori, K., Yamakawa, T., Inoue, K., Suzuki, K., Toyoshima, S., Hoshi, M., Ooshima, Y., Sekiyama, Y., Hatanaka, H., and Watanabe, K., eds. 1998. *Dictionary of biochemistry,* 3rd ed. Tokyo Kagaku Dozin, Tokyo.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.-L. 2000. The large-scale organization of metabolic networks. *Nature* **407:** 651–654.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30:** 42–46.

Karp, P.D. 2001. Pathway databases: A case study in computational symbolic theories. *Science* **293:** 2040–2044.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., and Gama-Castro, S. 2002. The Ecocyc database. *Nucleic Acids Res.* **30:** 56–58.

Küffner, R., Zimmer, R., and Lengauer, T. 2000. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics* **16:** 825–836.

Ma, H. and Zeng, A.-P. 2003. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19:** 270–277.

Mavrovouniotis, M.L. 1992. Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.* **36:** 1119–1132.

Michal, G., ed. 1999. *Biochemical pathways: An atlas of biochemistry and molecular biology*. Wiley & Spektrum, Heidelberg, Germany.

Mummaneni, A. and Kazic, T. 2002. The architectural dynamics of cellular biochemistry and molecular biology. In *Proceedings of the 6th World Multiconf. Systemics Cybern. Inform.* Available online at https://www.the-agora.org/repository/sci2002.ps.

Ouzounis, C.A. and Karp, P.D. 2000. Global properties of the metabolic map of *Escherichia coli*. *Genome Res.* **10:** 568–576.

Rathod, P.K. and Fellman, J.H. 1985. Identification of mammalian aspartate-4-decarboxylase. *Arch. Biochem. Biophys.* **238:** 435–446.

Rohmer, M. 1999. The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Nat. Prod. Rep.* **16:** 565–574.

Smith, A.D., Datta, S.P., Smith, G.H. Campbell, P.N., Bentley, R., McKenzie, H.A., Bender, D.A., Harris, A.J., Goodwin, T.W., Parish J.H., et al. eds. 2000. *Oxford dictionary of biochemistry and molecular biology*, rev. ed. Oxford University Press, New York.

Stephanopoulos, G.N., Aristidou, A.A., and Nielsen, J. 1999. *Metabolic engineering.* Academic Press, San Diego.

Wagner, A. and Fell, D. 2001. The small world inside large metabolic networks. *Proc. R. Soc. London B Biol. Sci.* **268:** 1803–1810.

Wipke, W.T. and Dyott, T.M. 1974. Stereochemically unique naming algorithm. *J. Amer. Chem. Soc.* **96:** 4834–4842.

WEB SITE REFERENCES

http://us.expasy.org/enzyme; ENZYME database.
http://www.brenda.uni-koeln.de; BRENDA database.
http://www.chem.qmul.ac.uk/iubmb/enzyme; IUBMB enzyme nomenclature.
http://www.ecocyc.org; EcoCyc database.
http://www.mdli.com; MDL Information Systems.
http://www.metabolome.jp; ARM database.

Received January 23, 2003; accepted in revised form July 23, 2003.