ARTICLE

BIOTECHNOLOGY and BIOENGINEERING

Construction of an *E. Coli* Genome-Scale Atom Mapping Model for MFA Calculations

Prabhasa Ravikirthi,[1] Patrick F. Suthers,[2] Costas D. Maranas[2]

[1]Department of Cell and Developmental Biology, The Pennsylvania State University, University Park, Pennsylvania

[2]Department of Chemical Engineering, The Pennsylvania State University, 112 Fenske Laboratory, University Park, Pennsylvania 16802; telephone: 814-863-9958; fax: 814-865-7846; e-mail: costas@psu.edu

Received 18 October 2010; revision received 17 January 2011; accepted 18 January 2011

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/bit.23070

ABSTRACT: Metabolic flux analysis (MFA) has so far been restricted to lumped networks lacking many important pathways, partly due to the difficulty in automatically generating isotope mapping matrices for genome-scale metabolic networks. Here we introduce a procedure that uses a compound matching algorithm based on the graph theoretical concept of pattern recognition along with relevant reaction information to automatically generate genome-scale atom mappings which trace the path of atoms from reactants to products for every reaction. The procedure is applied to the *i*AF1260 metabolic reconstruction of *Escherichia coli* yielding the genome-scale *i*sotope *m*apping model imPR90068. This model maps 90,068 non-hydrogen atoms that span all 2,077 reactions present in *i*AF1260 (previous largest mapping model included 238 reactions). The expanded scope of the isotope mapping model allows the complete tracking of labeled atoms through pathways such as cofactor and prosthetic group biosynthesis and histidine metabolism. An EMU representation of imPR90068 is also constructed and made available.

Biotechnol. Bioeng. 2011;xxx: xxx–xxx.

© 2011 Wiley Periodicals, Inc.

KEYWORDS: metabolic flux analysis; isotope mapping; isotopomers; atom mapping matrices; elementary metabolite units; genome-scale reconstructions

Introduction

Metabolic flux analysis (MFA) (Vallino and Stephanopoulos, 1993) has emerged as a critical tool to understand the physiological state of a cell (Bailey, 1991; Nielsen, 2003; Stephanopoulos and Vallino, 1991). Using isotopically labeled substrates with different labeling patterns, experimental techniques such as NMR (Adelbert et al., 1998; Kelleher, 2001) and GC-MS (Wittmann and Heinzle, 2002) are used to measure the amounts of different isotope forms of select metabolites. The fluxes in a metabolic network are directly coupled to the relative isotopic abundances of different metabolites through a system of nonlinear algebraic equations (Schmidt et al., 1999). Details of the same can be found in literature in a recent review (Kim et al., 2008). Briefly, these nonlinear equations are constructed using mapping matrices that trace the path of each atom and subsequently each isotopomer (isotope isomer) in a metabolic reaction. This information was initially represented using atom mapping matrices (AMM) (Zupke and Stephanopoulos, 1994) that track the transfer of carbon atoms from reactants to products. This concept was subsequently generalized in the form of isotopomer mapping matrices (IMM) (Schmidt et al., 1997) that enumerate all possible product isotopomers that can be created from each reactant isotopomer.

Two separate computational challenges arise during flux elucidation based on MFA. The first challenge involves the automated generation of isotope mapping matrices for genome-scale metabolic reconstructions while the second involves the efficient solution of the corresponding system of nonlinear equations for the unknown fluxes while accounting for measurement error. The challenge of flux elucidation has been previously addressed using a variety of computational techniques including the cumomer concept (Wiechert et al., 1999), theoretical bondomer (van Winden et al., 2002), the elementary metabolite units (EMU) framework (Antoniewicz et al., 2007a), FluxCalc (Suthers et al., 2007), and handling of measurement errors (Antoniewicz et al., 2006). However, the application of these methods has been restricted to models that were at least an order of magnitude smaller than genome-scale reconstructions as a consequence of the aforementioned challenges. Typical isotope mapping models contain 25–50 reactions (Kim et al., 2008), 76 reactions (Antoniewicz et al., 2007b), or 238 reactions (Suthers et al., 2007), which is the largest to-date model

Correspondence to: Costas D. Maranas

Additional Supporting Information may be found in the online version of this article.

© 2011 Wiley Periodicals, Inc.

(developed in our group). A key shortcoming of using lumped metabolic abstractions to perform flux elucidation is that they may erroneously lead to the conclusion that the available NMR, GC/MS, or MS/MS data is sufficient for unique flux elucidation (Chang et al., 2008). The inferred metabolic fluxes may then inherently reflect the biases/assumptions built-in during the lumped metabolic map creation step. In addition, utilizing a genome-scale model for simulation/strain design purposes and a separate lumped metabolic model for flux elucidation could complicate the seamless integration/transfer of results.

Motivated by these shortcomings, here we introduce a genome-scale *E. coli* isotope mapping model. This challenge is formidable, as it requires a detailed account of atom transitions for all 90,068 atoms in 2,077 reactions present in the metabolic reconstruction, *i*AF1260 (Feist et al., 2007). Atom mappings are obtained for each reaction by tracing the origin and destination of atoms through each individual reaction in the metabolic network. Tracing atoms from reactants to products requires the ability to topologically superimpose the structures of reactant and product molecules. This involves the identification of all "common" substructures between the two molecules. Thus, in addition to tracing isotopically labeled carbon atoms (typically the choice in MFA experiments) the path of O, N, P, S atoms as well as of metal/non-metal ions are also traced as part of the algorithm. Even though the immediate utility of imPR90068 is in the carbon mappings, the model is poised to take advantage of advances in labeling choices and detection. For instance, ^{15}N isotopes have been recently utilized in techniques such as kinetic flux profiling (KFP) (Yuan et al., 2006, 2008) and non-targeted tracer fate detection (NFTD) (Hiller et al., 2010) to elucidate metabolic fluxes. NFTD can also be used with other stable isotopes like ^{33}S or ^{18}O (Hiller et al., 2010).

Techniques relying on pattern recognition concepts from graph theory, which have been extensively employed in cheminformatics (Gillet et al., 1998; Raymond and Willett, 2002; Willett, 1995), can be used to topologically align and compare a reactant with a product molecule. These techniques essentially apply two mathematical operations on the molecular graphs of the two compounds to be aligned. The first mathematical operation combines the two molecular graphs into a single association graph (AG). The second operation identifies the largest clique (i.e., connected graph) within the AG. The maximum common subgraph (MCS) approach (Hattori et al., 2003b), formulates the edges of the AG based on the bond connectivity but without considering the bond-type data (single, double bond, etc.) of the two compounds involved. The NP-complete nature of all graph isomorphism problems (Raymond and Willett, 2002) adds to the difficulty of generating genome-scale atom mappings. Furthermore, several symmetry considerations about metabolites need to be addressed before the final reaction atom mappings are formulated. These include equivalent oxygen atoms (such as those present in carboxyl and phosphate groups) and rotationally symmetric molecules (e.g., succinate), which result in scrambling of isotope labeling. Additionally metabolites containing a prochiral carbon center (e.g., citrate) or metabolites with a center of inversion but lacking a rotational axis of symmetry restrict the mapping degeneracy.

So far, graph isomorphism techniques have only been used to contrast pairs of compounds (Hattori et al., 2003a) or trace just carbon atoms (Mu et al., 2007) within the KEGG/LIGAND database (Goto et al., 1998, 2002). In addition, the atom transitions listed in KEGG are inadequate for flux analysis using MFA since alternative atom transitions are not explicitly listed when symmetric molecular sub-structures or symmetric molecules are present in the reaction. Alternatively, compound matching based on an algorithm that tallies the connectivity (i.e., number of atoms connected to a given atom) of atoms in the compared compounds (Wipke and Dyott, 1974), has been used to trace atoms across reactions (Arita, 2003; Flower, 1998). However, this procedure requires the manual reordering of metabolites in reactions and has scaling limitations (i.e., it cannot detect rings of size greater than ten such as heme) (Arita, 2004).

We chose to overcome these limitations and generate mappings for the latest metabolic reconstruction of *E. coli* (Feist et al., 2007) by first representing molecular chemical structures as graphs defined by a set of vertices (the atoms) connected by edges (the bonds). Subsequently, the MCS method (Hattori et al., 2003a) coupled with a modified branch and bound algorithm for clique finding (Coen and Joep, 1973) is customized to automatically generate genome scale atom mappings. Finally the mappings obtained for each reaction are pruned to retain only the biochemically relevant ones.

Results

The proposed procedure used to generate imPR90068, requires as input the stoichiometry of all reactions present in the metabolic network and data encoding the chemical structure of all metabolites involved in the network in the form of MDL mol files. MDL is a file format created by MDL Information Systems containing atom and bond information of the participating compounds. The method described can be applied to any genome-scale metabolic model and is amenable to the straightforward inclusion of additional reactions not present in the original organism models as well as user-supplied metabolite structures. During the automated procedure, a library of atom mappings and recurring motifs is generated which can be leveraged for future isotope mapping efforts. The four steps of the procedure (see Fig. 1) are described in detail in the Methods section (see Supplementary Appendix A). The end result of the atom mapping process is the isotope mapping model imPR90068 for the *E. coli* strain K-12 that spans 1,039 metabolites, 2,077 reactions and contains a total of 1.37×10^{157} isotopomers (with 8.34×10^{93} ^{13}C isotopomers). The atom mappings

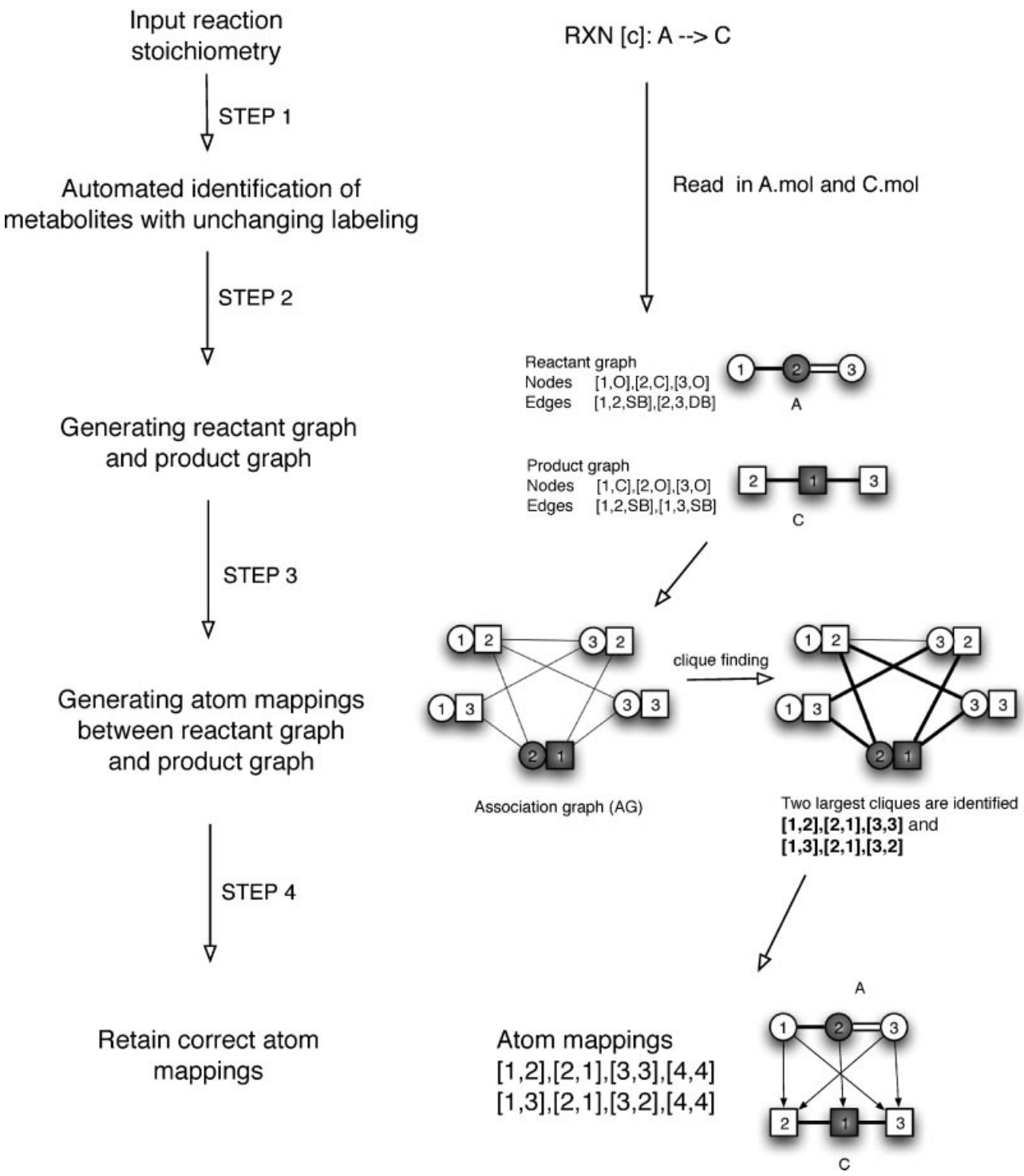

Figure 1. Steps 1-4 are applied to a general reaction A => C. The molecular structures of A and C are shown in Step 4. Grey circles and squares indicate carbon atoms (C) and white denote oxygen (O) atoms. (Step 2) The atoms of reactant graph are shown as circles and that of product graph are shown as squares. (Step 3) The nodes of the AG are pairs of nodes from reactant and product graphs, and grey lines are the edges of the AG [see Supplementary Appendix A for details]. The two cliques identified are the largest set of vertices that are completely connected to each other in the AG and are shown as thick black lines. (Step 4) The atom mappings are shown as lines (atom traces) between reactant and product molecular structures. From the visual representation we see that two alternate mappings exist due to symmetry of A and C molecules.

were generated for each reaction separately using the Lion-XJ computational cluster of the High Performance Computing Group consisting of Dell PowerEdge 1950 servers with dual 3.0 GHz Intel Xeon E5450 Quad-Core Processors and 32 GB of ECC RAM. The identification of all possible atom mappings for reactions containing fewer than 25 reactant atoms took between 10 and 25 min of CPU time (Hattori et al., 2003b) report average running times of approximately 11 h per comparison when comparing two random chosen compounds from the KEGG/LIGAND database. Reactions with more than 25 atoms required between 4 and 40 h to run. The average CPU time taken was approximately 25 h per reaction. Because the analysis of each reaction can proceed independently of others, we typically had in excess of 300 running simultaneously. Atom mappings were generated for every reaction in the network tracing all non-hydrogen elements including C, N, O, P, S, and metal/non-metal ions. The EMU representation (Antoniewicz et al., 2007a) was implemented using Python modules. Briefly, given a set of mass isotopomer measurements and a set of source metabolites, this implementation calculates network fluxes through an EMU representation. The details of the procedure used to identify all EMU species and variables are outlined in Suthers et al. (2010).

Reactant to Product Atom Mapping Examples

The metabolic network *i*AF1260 contains 304 exchange reactions, 690 transport reactions and 1,387 metabolic reactions (Feist et al., 2007). As many as 653 reactions contained compounds with at least one kind of symmetry (i.e., equivalent resonance atoms, prochiral centers, rotational axis, or center of inversion). For 91 of these reactions the atom mapping generated were refined due to the structural geometry of the participating metabolites. For example, reactions containing prochiral metabolites (i.e., including C atoms bonded to two stereo-heterotopic groups) react in vivo stereo-specifically. Therefore, their atom mappings were pruned to only biochemically feasible ones.

The atom mappings for the 690 transport reactions, which account for 12,325 of the traced atoms, were generated in a straightforward manner as the molecular graphs remain invariant upon transport. For example, the atom mappings for the arginine/agmatine antiport reaction, which is a reversible inner membrane transport reaction, are retained as arginine and agmatine as they simply transported from the cytosol to the periplasmic space without any bond modifications:

$$\text{ARGAGMt7pp} \quad \text{agm[c]} + \text{arg-L[p]} \Leftrightarrow \text{agm[p]} + \text{arg-L[c]}$$

The atom mappings for the remaining 1,387 metabolic reactions, containing 77,619 of the mapped atoms, were created by iteratively applying for every reaction the proposed workflow (see Steps 1–4 in Fig. 1). During this process, five frequently occurring reaction motifs were automatically identified and stored in a database (see Table I). The atom mappings of these five reaction motifs, which occur in 424 different reactions, were simply copied from the reaction motif library (Table I).

Table I. List of frequently occurring reaction motifs.

Reaction motif	# of occurrences in iAF1260	# of atoms mapped
$atp + h_2o \rightarrow adp + h + pi$	162	32
$atp + h_2o \rightarrow amp + h + ppi$	65	32
$adp + h_2o \rightarrow amp + h + pi$	5	32
$nad + h \leftrightarrow nadh$	110	44
$nadp + h \leftrightarrow nadph$	82	48

The following examples detail the challenges faced during the atom mapping procedure (Fig. 1) and illustrate the handling of large molecules and various symmetries that produce/restrict isotopic scrambling. The first example is citrate oxaloacetate-lyase reaction, abbreviated as CITL in *i*AF1260:

$$\text{CITL} : \text{Citrate} \Leftrightarrow \text{Acetate} + \text{Oxaloacetate}$$

The molecular information of citrate, acetate, and oxaloacetate is extracted from the corresponding MDL mol files. The reactant and product graphs constructed from this molecular information yield an association graph with 85 nodes (each node of an AG consists of a pair of atoms) containing 16 cliques corresponding to all alternative atom mappings (see Supplementary Appendix A for details). The chemical structures of the mapped metabolites and their possible alternative mappings are shown diagrammatically in Figure 2. The stereo-specific enzyme that catalyzes the citrate oxaloacetate-lyase (or citrate lyase) reaction in *E. coli* is known to produce acetate from the pro-S arm of the prochiral citrate molecule (Dagley and Dawes 1955). Therefore the structurally feasible but biochemically irrelevant formation of acetate from the pro-R carboxymethyl is ignored (Step 4). The equivalent oxygen atoms on the carboxyl groups result in eight alternative mappings. The set of mappings shown in Figure 2a are the required *reaction mapping* of reaction CITL (Supplementary Information).

The presence of rotationally symmetric molecules causes additional scrambling of isotopic labeling. This is illustrated using the taurine dioxygenase reaction (see Fig. 3), which is abbreviated as TAUDO in *i*AF1260:

$$\text{TAUDO[c]} : \text{akg} + \text{o2} + \text{taur} \rightarrow \text{aacald} + \text{co2} + \text{h} + \text{so3} + \text{succ}$$

Due to the presence of the rotationally symmetric succinate moiety, equivalent oxygen atoms on carbon dioxide, O_2, and sulfite group, 96 alternate mappings are generated. These atom mappings, which trace 7 carbon, 1 nitrogen, 10 oxygen, and 1 sulfur atoms between 7 metabolites (Fig. 3), are stored as a reaction mapping under the reaction name, TAUDO.

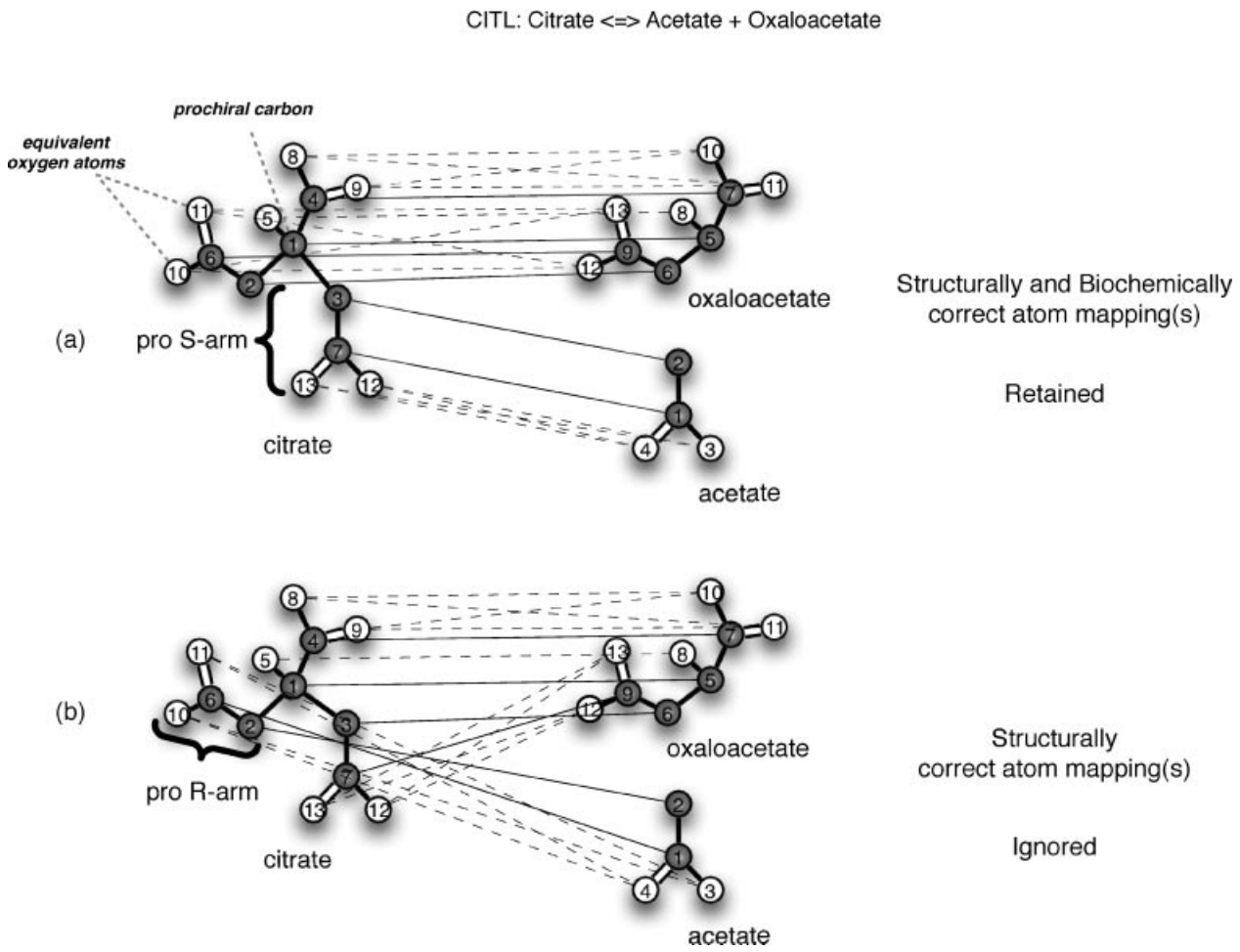

Figure 2. Retaining biochemically valid atom mappings. **a**: The stereo-specific enzyme citrate oxaloacetate-lyase that catalyzes this reaction forms acetate from the pro S carboxymethyl group of citrate. **b**: The structurally equivalent but biochemically infeasible alternative mapping generated during Step 3 is eliminated in Step 4. The equivalent oxygen atoms are kept track to identify equivalent EMUs for predicting labeling distributions.

Size Statistics and Mapping Degeneracy of imPR90068

The genome-scale mapping model imPR90068 generated for the *E. coli* encodes the complete list of reactions in *i*AF1260 (Feist et al., 2007) as a library of 2,077 reaction mappings (see supplementary information for the mapping files). Each reaction mapping contains multiple atom mappings that trace all reactant atoms to all product atoms in the respective reaction. The model contains a total of 20,872 alternate atom mappings that trace the fate of 90,068 atoms through a network of 2,077 reactions and 1,039 metabolites. These atom mappings trace the path of C, O, N, P, S atoms as well as Ag, As, Ca, Cd, Cl, Co, Cu, halogens, Fe, Hg, K, Mg, Mn, Na, Ni, Se, W, Zn ions. Detailed information on atoms traced is provided in Table II. Figure 4 depicts the COG classifications of the reactions in imPR90068 and imPS1485 (Suthers et al., 2007) revealing a dramatic increase in the number of reactions present. In addition to a sevenfold change in the total number of reactions, over 800 new metabolites are present with 45 new biomass components accounting for lipids, cell wall components, nucleotide synthesis, and tRNA species. The imPR90068 model can accommodate as many as 174 different carbon sources signifying different labeling opportunities.

The classification of all 1,387 metabolic reactions in imPR90068 based on the number of alternative mappings (per reaction) is shown in Table III (also see Supplementary Figure). Among these, 734 reactions contain a single mapping alternative implying that the atoms in these reactions are uniquely mapped from reactants to products. The majority of these 734 reactions with no mapping degeneracy are isomerization, displacement or substitution reactions typically containing less than three reacting species. The remaining reaction mappings are degenerate to various degrees and contain multiple alternative atom transitions from reactants to products due to symmetry(ies) present in the reaction operator (Table III). A general downward trend is observed in the number of reactions with

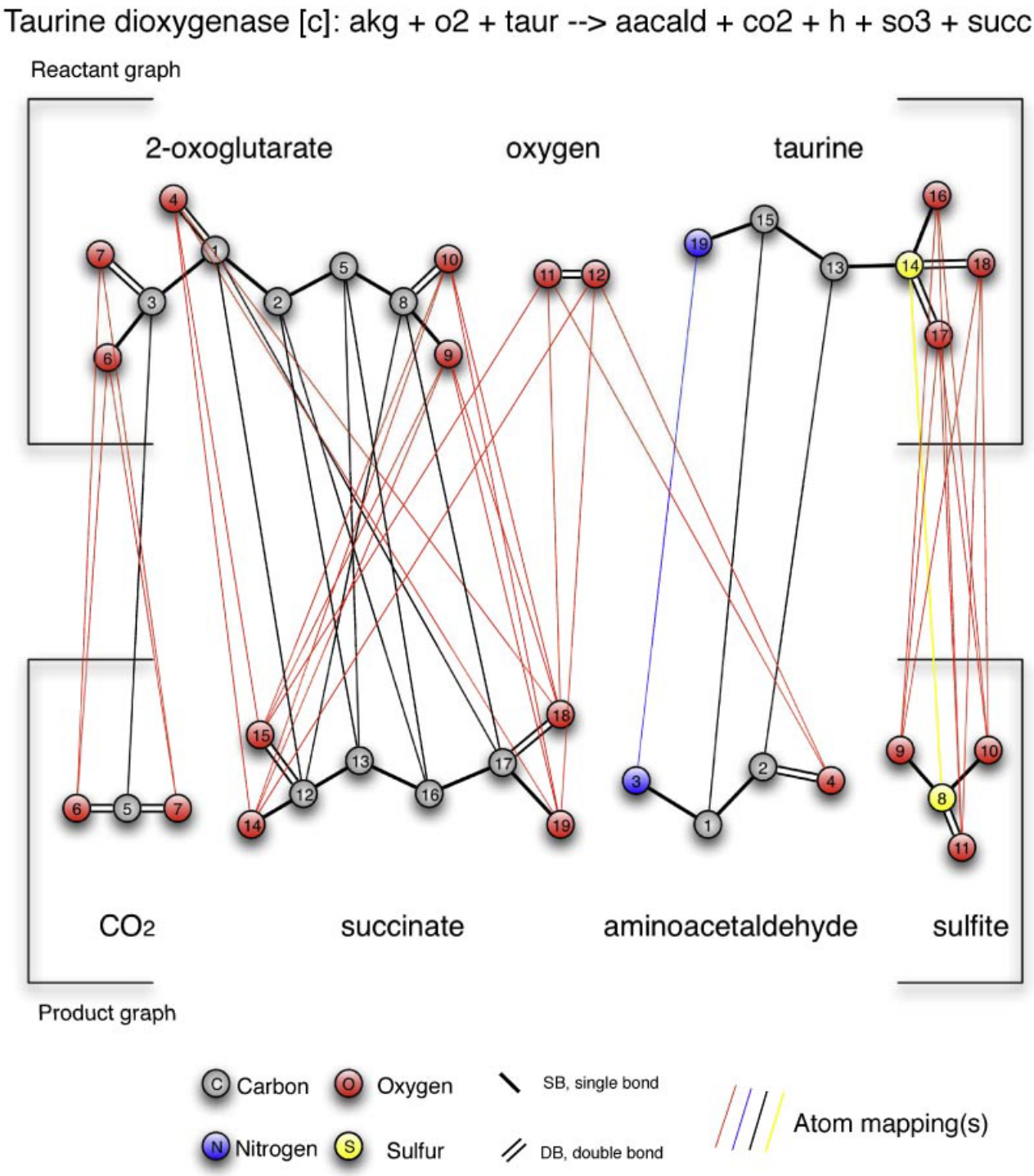

Figure 3. Atom mappings of Taurine dioxygenase. The reaction taurine dioxygenase appears as shown in *i*AF1260 (top). All feasible atom mappings between reactant graph and product graph are shown as a set of lines connecting reactant atoms to product atoms (bottom). Note that carbon atoms (1,2,5,8) in 2-oxoglutarate map to either carbon atoms (12,13,16,17) in succinate or in reverse to (17,16,13,12). The mappings are explicitly listed in the corresponding atom mapping file of the reaction available as Supplementary Information. Gray, red, blue, and yellow lines connect carbon, oxygen, nitrogen, and sulfur atoms, respectively. The product compounds have been rearranged in the product graph for visual clarity. The reactant and product atoms have been sequentially numbered in accordance with the reaction representation in *i*AF1260 and the atom numberings are in accordance with the numbering present in the corresponding MDL mol files of the compounds in the supplemental information of *i*AF1260.

Table II. Total number of most-prevalent atoms and their respective isotopomers.

Atom type	Total # of atoms traced	Total # of isotopomers
Carbon	49,539	8.34×10^{93}
Oxygen	29,061	1.61×10^{60}
Phosphorous	3,280	1.00×10^{4}
Nitrogen	2,386	2.58×10^{7}
Sulfur	409	4.09×10^{3}
Others[a]	265	4.05×10^{3}
Total	90,068	1.37×10^{157}

[a]Includes Ag, As, Ca, Cd, Cl, Co, Cu, halogens, Fe, Hg, K, Mg, Mn, Na, Ni, Se, W, Zn.

increasing reaction mapping degeneracy with 528, 256, 155 reactions containing respectively 2–8, 9–128, 129–1024 alternative mappings (Supplementary Figure). Spikes are observed at 17–32 alternatives due to the presence of phosphate groups (24 alternative mappings) and similarly at 257–512 due to the presence of diphosphate groups (288 alternative mappings).

Table III also identifies which atom type (or combination of atoms) is responsible for the degeneracy in the mapping. The individual reactions containing a modest number of mappings (i.e., from two to eight) are primarily degenerate either due to equivalent carbons or due to equivalent oxygens and less likely due to the presence of both equivalent carbons and oxygens (71% due to either only C or only O and 22% due to both C and O). The reactions containing equivalent O (either standalone or in combination with other equivalent atoms such as C, N) are predominantly due to oxygen atoms in the carboxyl groups. Degeneracy due to equivalent C (or N) arise as a result of rotational symmetry of the reacting species (e.g., succinate, D-mannitol, fumarate). Furthermore, reaction mapping degeneracy arising from both C and O scrambling are fairly ubiquitous through the model. For example in reaction TAUDO (see Fig. 3), reactant 2-oxoglutarate can be mapped to the symmetric product succinate, in four possible ways. Two of these mappings arise when carbon atoms (1,2,5,8) from 2-oxoglutarate map to either carbons (12,13,16,17) or in reverse (17,16,13,12) in succinate. The other two degenerate mappings are due to resonance-stabilized oxygen atoms 9,10

Table III. Distribution of alternate atom mappings of reactions present in imPR90068.

Alternatives (degeneracy)	Total # of reactions	# of reactions with equivalent C, O, N, or P								
		C only	O only	N only	C, O	C, N	O, N	O, P	C, O, N	C, O, P
1	734									
2	232	138	105	4	30	3			1	0
3–4	117	17	48		41	2	2		4	1
5–8	179	41	66		58	1	2		7	3
9–16	71	2	31		33	1	1	1	1	0
17–32	121	9	68		31	1			4	7
33–64	35	1	14		14				3	2
65–128	29	0	16		9			2	1	1
129–256	19	0	5		6			8	0	0
257–512	126	1	107		4			3	3	1
513–1,024	10	0	2		2			2	3	1

The breakdown of degenerate reactions with respect to equivalent carbon (C), oxygen (O), nitrogen (N), and phosphorous (P) atoms is also shown.

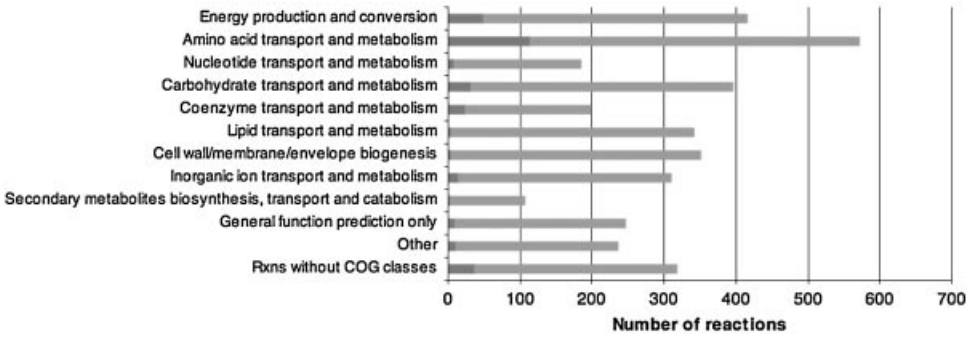

Figure 4. Classification of reactions according to respective clusters of orthologous groups (COGs). Dark grey bars are reactions present in imPS1485 while light grey bars represent new reactions present in the genome-scale imPR90068.

in 2-oxoglutarate that can be mapped onto equivalent oxygen atom pairs 14,15 or 18,19 in succinate.

Oxygen atoms are by far the most highly contributing to alternate mappings (i.e., 44% of all degenerate reactions). This is not surprising given the prevalence of phosphate, sulfur, and carboxyl groups containing multiple equivalent oxygen atoms. Often, multiple atoms (e.g., C, O, N, or P) simultaneously contribute in the mapping degeneracy. Phosphorous atoms accompanied by equivalent oxygen atoms (due to the presence of resonating phosphate groups) are involved in reactions with large numbers of mappings (i.e., more than 64). There exist 10 reactions with number of mappings in the range of 513–1,024. These reactions contain four or more reacting molecules usually with multiple symmetric metabolites and are involved in cofactor and prosthetic group biosynthesis, murein recycling, and nucleotide synthesis/salvage pathways. For example, in the asparigine synthetase reaction ASNS2, six molecules containing five reaction operators (two carboxyl groups and three phosphate groups) bring the reaction mapping degeneracy to 864 alternatives.

New Reactions/Metabolites in imPR90068

The introduced isotope mapping model imPR90068 contains mappings for reactions that were previously lumped or completely absent from isotope mapping models (even in imPS1485). These new additions include 68 reactions involved in the metabolism of 17 different amino acids (all but Asparagine, Glutamine, and Glutamic acid), 65 reactions involved in central metabolism, 153 reactions in nucleotide biosynthesis and salvage pathways, 225 reactions in glycerophospholipid metabolism, 160 reactions in cofactor and prosthetic group biosynthesis and 181 reactions in alternate carbon metabolism (see Fig. 4). The inclusion of all biotransformations spanned by the genome-scale model implies that alternate metabolic routes can now fully be taken into account during flux elucidation using MFA. For example, in imPR90068, the xylose isomerase catalyzed reaction XYLI2 that reversibly isomerizes D-glucose to D-fructose combined with the fructose transport reaction FRUpts2pp which converts phosphoenolpyruvate (PEP) to pyruvate during the transport of D-fructose, creates a pathway from glucose to pyruvate alternate to glycolysis. Other alternate glucose metabolism entries include amylo-maltase (AMALT1-4), maltodextrin glucosidase (MLTG1-5), and α- and β-galactosidase (GALS3, LACZ, LACZpp) reactions. In addition, a growth on 174 different carbon sources is possible using imPR90068 as opposed to only glucose and a few amino acids using imPS1485. As many as 45 biomass components absent from imPS1485 are now part of the model. These metabolites include cofactors (e.g., CoA), amino acids (e.g., His and Trp), riboflavin, murein, and inorganic ions (e.g., Fe + 3). It is important to note that new reactions in imPR90068 are not necessarily far away from central metabolism. Even under aerobic glucose growth conditions, as many as 35 new reactions are added to central metabolism that are part of Citric Acid Cycle, Glycolysis/Gluconeogenesis, Oxidative Phosphorylation, Pentose Phosphate Pathway, and Pyruvate Metabolism.

Notably, imPR90068 accounts for not only all reactions but also all metabolites present in *i*AF1260. Nearly 800 new metabolites are present in imPR90068 that were absent in imPS1485. These newly added metabolites link parts of metabolism previously treated before as separate. For example, (see Fig. 5) the added metabolite AICAR (5-Amino-1-(5-Phospho-D-ribosyl)imidazole-4-carboxamide) directly participates in purine meabolism and the histidine pathway. It is also indirectly linked to thiamine metabolism (through metabolite AIR), glycine, serine and threonine metabolism (through glycine) and in alanine, aspartate, and glutamate metabolism (through glutamate). Thus, the incorporation of a single additional metabolite in imPR90068 enables for the first time the ability to fully describe histidine and purine metabolism as well as account for interactions between many pathways.

Reduced and EMU Based Representation of imPR90068

Armed with a complete database of all atom mappings implied by the genome-scale model *i*AF1260, it is straightforward to select only the mappings which are relevant for a given isotope labeling experiment. The numbers of isotopomers present upon labeling various atoms present in the model are detailed in Table II. For example, by labeling only carbons we find that the 932 carbon-containing metabolites (with a total of 20,935 carbon atoms) yield 8.34×10^{93} ^{13}C isotopomers. We can tailor the set of considered isotopomers to the specifics of the system under consideration by removing all reactions/mappings that are suppressed under the experimental conditions. For example, under aerobic glucose minimal media conditions 752 blocked/suppressed reactions can be removed from the model leaving 793 metabolites containing 33,026 tractable carbon atoms and 3.02×10^{62} isotopomers.

An even more compact representation of the isotope mapping relations can be achieved using the EMU representation (Antoniewicz et al., 2007a). We have developed Python scripts that given the atom mapping matrices of imPR90068, the labeled substrate, and measured fragments the EMU representation is automatically generated. The EMU representation of imPR90068 for aerobic labeled glucose minimal media conditions and using the 31 amino acid fragments listed in Table I of Suthers et al. (2007) is provided as supplemental material. Table IV highlights the savings afforded by the EMU representation. The 17,346 carbon isotopomers of imPS1485 are reduced to 1,215 EMU species and 3,912 mass isotopomers (Suthers et al., 2010). All 10^{93} carbon isotopomers in imPR90068 are reduced to 1,068,431 EMU species and 6,066,011 mass isotopomers. This is still a very large model size that will require customized implementations using nonlinear optimization

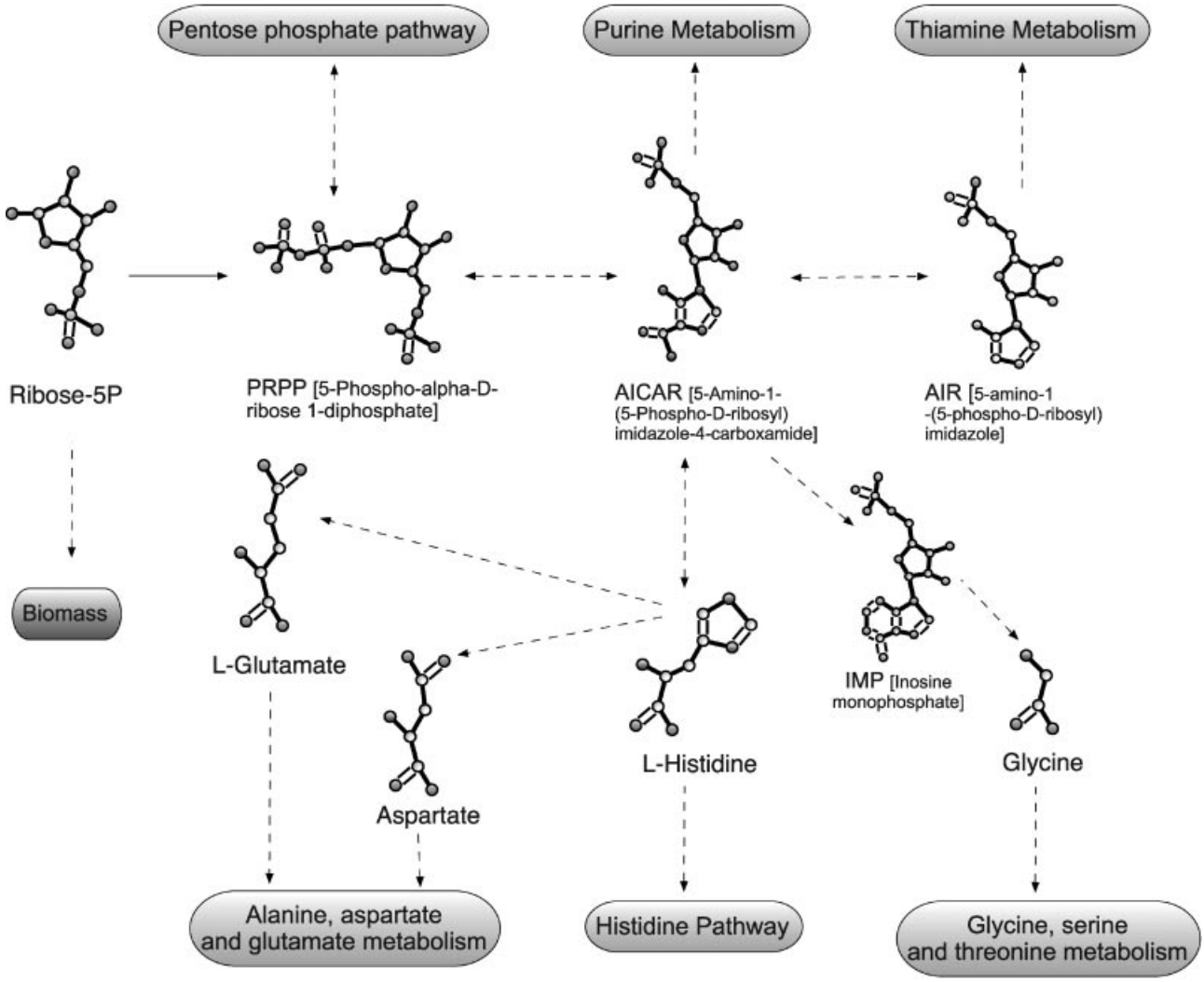

Figure 5. An example of the expanded scope of the genome-scale isotope mapping model imPR90068. In imPS1485 Ribose-5P production was directly routed to biomass as a stand-in substitute for histidine. In imPR90068 R5P downstream conversion is linked to other amino acid synthesis pathways.

Table IV. Comparison of the sizes of imPS1485 and imPR90068 isotope mapping models of *E. coli*.

		EMU model		EMU reduced model	
Isotope mapping model	^{13}C isotopomers	EMU species	EMU mass isotopomers	EMU species	EMU mass isotopomers
Allowing for all uptakes with a transport mechanism					
imPR90068	8.3×10^{93}	1,068,431	6,066,011	621,622	2,787,563
imPS1485	17,346	1,215	3,912	762	2,438
Aerobic glucose minimal growth medium with all blocked reaction removed					
imPR90068	3.02×10^{62}	748,841	3,425,985	473,712	1,978,917
imPS1485	3,584	909	2,911	486	1,538

solvers such as CONOPT (Drud, 1994, 2007) for flux elucidation. It is important to emphasize that for many MFA applications the complete set of reactions/metabolites may not be needed. We anticipate that users will pro-actively retain only parts of imPR90068 relevant to the set of measured fluxes and adopted labeling choices.

Discussion

This paper introduced the computational infrastructure for tracing all atoms present in every reaction in the *i*AF1260 metabolic reconstruction of *E. coli* from reactants to products to create a genome-scale mapping database.

This automated procedure can be efficiently leveraged for genome-scale models of other organisms to create isotope mapping databases. Common reactions already present in *i*AF1260 can be directly culled from the imPR90068 reaction-mappings database thus significantly reducing the effort needed to construct other organism-specific mapping models. The potential to improve our understanding of flux allocation in different organisms is alluded by the gap in the size of genome scale versus isotope mapping models. For example, there exists a 50-fold difference in the size of the genome-scale reconstruction of *Bacilus subtilis* that spans 1,020 reactions (Oh et al., 2007) and its current isotope mapping model (Dauner et al., 2001) that accounts for only 25 reactions (all from central metabolism). It is expected that incorporating reactions into the mapping model already present in the genome-scale model could shed light onto metabolic pathway usage patterns.

The incorporation of more than 1,100 new reactions involved in central metabolism, amino acid synthesis, alternate carbon metabolism and other parts of *E. coli* metabolism together with the inclusion of more than 800 metabolites compared to the previous largest imPR1485 model (Suthers et al., 2007) integrates metabolic flows between all pathways (see Fig. 5 for an example). However, the ability to elucidate fluxes using the full complement of reactions and metabolites present in genome-scale level reconstructions comes at the expense of requiring additional labeling data. While lumped isotope models (Antoniewicz et al., 2007b; Kim et al., 2008; Suthers et al., 2007) typically require the analysis of spectra (i.e., NMR or GC/MS) for only about 20–50 fragments, using the totality of mapped isotopomers in imPR90068 will likely require significantly higher numbers of carefully chosen labeled fragments. This makes even more pertinent the use of methods such as OptMeas (Chang et al., 2008; Suthers et al., 2010) to pinpoint minimal measurement sets and compact isotope representations such as EMU (Antoniewicz et al., 2007a) for complete flux elucidation. We anticipate that the development of systematic reaction step aggregation techniques (e.g., SLIPs (Quek, 2009)) that avoid any loss of information will lead to substantial reduction in the size of the problems that need to be solved for flux elucidation. Even though imPR90068 tracks the fate of non-carbon atoms through reactions, we expect that carbon atom mapping information to be at present the most useful in the context of MFA calculations. This is because significant uncertainties exist in the description of isotopomer scrambling caused by non-carbon atoms (e.g., exchange of oxygen atoms with water) that may erase any labeling information. Such isotopomer scrambling is also a concern for carbon atoms. Significant effort was spent during the construction of imPR90068 to account for chiral and prochiral metabolites, rotationally symmetric molecules both with a rotational axis or with a center of inversion, resonance stabilized equivalent atoms (in total 31% of all reactions) and partial exchange of oxygen with water (e.g., during aldolase catalyzed reactions) whenever supporting reaction mechanism information was available.

Finally, the use of molecular graph representations at a genome-scale level can be used to study the synthesis problem in metabolic networks (Hatzimanikatis et al., 2005). The ability to map atom origins and destinations without the use of any pre-defined reaction rules based on EC reaction classification (Tipton and Boyce, 2000) can be useful in elucidating novel chemistries.

References

Adelbert B, Christoph R, Dietmar E, Duilio A, Georg F, Wolfgang E. 1998. Elucidation of novel biosynthetic pathways and metabolite flux patterns by retrobiosynthetic NMR analysis. FEMS Microbiol Rev 22(5): 567–598.

Antoniewicz MR, Kelleher JK, Stephanopoulos G. 2006. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. Metab Eng 8(4):324–337.

Antoniewicz MR, Kelleher JK, Stephanopoulos G. 2007a. Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. Metab Eng 9(1):68–86.

Antoniewicz MR, Kraynie DF, Laffend LA, Gonzalez-Lergier J, Kelleher JK, Stephanopoulos G. 2007b. Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. Metab Eng 9(3):277–292.

Arita M. 2003. In silico atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. Genome Res 13(11):2455–2466.

Arita M. 2004. The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci USA 101(6):1543–1547.

Bailey JE. 1991. Toward a science of metabolic engineering. Science 252(5013):1668–1675.

Chang Y, Suthers PF, Maranas CD. 2008. Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments. Biotechnol Bioeng 100(6):1039–1049.

Coen B, Joep K. 1973. Algorithm 457: Finding all cliques of an undirected graph. Commun ACM 16(9):575–577.

Dagley S, Dawes EA. 1955. Citridesmolase: Its properties and mode of action. Biochim Biophys Acta 17:177–184.

Dauner M, Bailey JE, Sauer U. 2001. Metabolic flux analysis with a comprehensive isotopomer model in *Bacillus subtilis*. Biotechnol Bioeng 76(2):144–156.

Drud AS. 1994. CONOPT—A large-scale GRG code. Informs J Comput 6(2):207–216.

Drud A. 2007. CONOPT. A/S, Bagsvaerd. Denmark: ARKI Consulting and Development.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121.

Flower DR. 1998. On the properties of bit string-based measures of chemical similarity. J Chem Inf Comput Sci 38(3):379–386.

Gillet VJ, Wild DJ, Willett P, Bradshaw J. 1998. Similarity and dissimilarity methods for processing chemical structure databases. Comput J 41(8): 547–558.

Goto S, Nishioka T, Kanehisa M. 1998. LIGAND: Chemical database for enzyme reactions. Bioinformatics 14(7):591–599.

Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M. 2002. LIGAND: Database of chemical compounds and reactions in biological pathways. Nucleic Acids Res 30(1):402–404.

Hattori M, Okuno Y, Goto S, Kanehisa M. 2003a. Development of a chemical structure comparison method for integrated analysis of

chemical and genomic information in the metabolic pathways. J Am Chem Soc 125(39):11853–11865.
Hattori M, Okuno Y, Goto S, Kanehisa M. 2003b. Heuristics for chemical compound matching. Genome Inform 14:144–153.
Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ. 2005. Exploring the diversity of complex metabolic networks. Bioinformatics 21(8):1603–1609.
Hiller K, Metallo CM, Kelleher JK, Stephanopoulos G. 2010. Nontargeted elucidation of metabolic pathways using stable-isotope tracers and mass spectrometry. Anal Chem 82(15):6621–6628.
Kelleher JK. 2001. Flux estimation using isotopic tracers: Common ground for metabolic physiology and metabolic engineering. Metab Eng 3(2): 100–110.
Kim HU, Kim TY, Lee SY. 2008. Metabolic flux analysis and metabolic engineering of microorganisms. Mol Biosyst 4(2):113–120.
Mu F, Williams RF, Unkefer CJ, Unkefer PJ, Faeder JR, Hlavacek WS. 2007. Carbon-fate maps for metabolic reactions. Bioinformatics 23(23): 3193–3199.
Nielsen J. 2003. It is all about metabolic fluxes. J Bacteriol 185(24):7031–7035.
Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R. 2007. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. J Biol Chem 282(39):28791–28799.
Quek L-E. 2009. Large-scale metabolic flux analysis for mammalian cells: A systematic progression from model conception to model reduction to experimental design. The University of Queensland. Brisbane, QLD 4072, Australia. 313p.
Raymond JW, Willett P. 2002. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. J Comput Aided Mol Des 16(7):521–533.
Schmidt K, Carlsen M, Nielsen J, Villadsen J. 1997. Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. Biotechnol Bioeng 55(6):831–840.
Schmidt K, Nielsen J, Villadsen J. 1999. Quantitative analysis of metabolic fluxes in *Escherichia coli*, using two-dimensional NMR spectroscopy and complete isotopomer models. J Biotechnol 71(1–3):175–189.
Stephanopoulos G, Vallino JJ. 1991. Network rigidity and metabolic engineering in metabolite overproduction. Science 252(5013):1675–1681.
Suthers PF, Burgard AP, Dasika MS, Nowroozi F, Van Dien S, Keasling JD, Maranas CD. 2007. Metabolic flux elucidation for large-scale models using 13C labeled isotopes. Metab Eng 9(5–6):387–405.
Suthers PF, Chang YJ, Maranas CD. 2010. Improved computational performance of MFA using elementary metabolite units and flux coupling. Metab Eng 12(2):123–128.
Tipton K, Boyce S. 2000. History of the enzyme nomenclature system. Bioinformatics 16(1):34–40.
Vallino JJ, Stephanopoulos G. 1993. Metabolic flux distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. Biotechnol Bioeng 41(6):633–646.
van Winden WA, Wittmann C, Heinzle E, Heijnen JJ. 2002. Correcting mass isotopomer distributions for naturally occurring isotopes. Biotechnol Bioeng 80(4):477–479.
Wiechert W, Mollney M, Isermann N, Wurzel M, de Graaf AA. 1999. Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. Biotechnol Bioeng 66(2): 69–85.
Willett P. 1995. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. J Mol Recognit 8(5):290–303.
Wipke WT, Dyott TM. 1974. Stereochemically unique naming algorithm. J Am Chem Soc 96(15):4834–4842.
Wittmann C, Heinzle E. 2002. Genealogy profiling through strain improvement by using metabolic network analysis: Metabolic flux genealogy of several generations of lysine-producing corynebacteria. Appl Environ Microbiol 68(12):5843–5859.
Yuan J, Fowler WU, Kimball E, Lu W, Rabinowitz JD. 2006. Kinetic flux profiling of nitrogen assimilation in *Escherichia coli*. Nat Chem Biol 2(10):529–530.
Yuan J, Bennett BD, Rabinowitz JD. 2008. Kinetic flux profiling for quantitation of cellular metabolic fluxes. Nat Protoc 3(8):1328–1340.
Zupke C, Stephanopoulos G. 1994. Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrixes. Biotechnol Prog 10(5):489–498.