Bioinformatics Advance Access published October 12, 2007

Systems biology

Carbon fate maps for metabolic reactions

Fangping Mu[1], Robert F. Williams[2], Clifford J. Unkefer[2], Pat J. Unkefer[2], James R. Faeder[3] and William S. Hlavacek[1,4*]

[1]Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

[2]National Stable Isotope Resource, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

[3]Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15260, USA

[4]Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: Stable isotope labeling of small-molecule metabolites (e.g., ^{13}C-labeling of glucose) is a powerful tool for characterizing pathways and reaction fluxes in a metabolic network. Analysis of isotope labeling patterns requires knowledge of the fates of individual atoms and moieties in reactions, which can be difficult to collect in a useful form when considering a large number of enzymatic reactions.

Results: We report carbon-fate maps for 4,605 enzyme-catalyzed reactions documented in the KEGG database. Every fate map has been manually checked for consistency with known reaction mechanisms. A map includes a standardized structure-based identifier for each reactant (namely, an InChI™ string); indices for carbon atoms that are uniquely derived from the metabolite identifiers; structural data, including an identification of homotopic and prochiral carbon atoms; and a bijective map relating the corresponding carbon atoms in substrates and products. Fate maps are defined using the BioNetGen™ language (BNGL), a formal model-specification language, which allows a set of maps to be automatically translated into isotopomer mass-balance equations.

Availability: The carbon fate maps and software for visualizing the maps are freely available (http://cellsignaling.lanl.gov/FateMaps/).

Contact: wish@lanl.gov

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

1 INTRODUCTION

Stable isotope labeling of small-molecule metabolites combined with analytic techniques for measuring isotopomer distributions has a number of important applications, including the determination of chemical structures, characterization of enzymatic reaction mechanisms, and elucidation of metabolic pathways (Birkemeyer *et al*., 2005; Boros *et al*., 2003; Kopka, 2006). Isotope labeling is likely to become more important with the continued development of methods for high-throughput metabolite profiling. These methods, which are largely based on mass spectrometry (MS), enable isotopomer distributions to be assayed on a scale that is useful for estimating mass fluxes in networks with hundreds of reactions (Hellerstein, 2003; Hellerstein and Murphy, 2004; Nielsen, 2003; Sauer, 2004; Sauer, 2006). Flux estimates based on methods incorporating isotopomer data have been found to be significantly improved compared to results obtained using other procedures, such as conventional metabolic flux analysis (MFA), in microorganisms, plants and animals (Emmerling *et al*., 2002; Fischer and Sauer, 2005; Hua *et al*., 2003; McCabe and Previs, 2004; Ratcliffe and Shachar-Hill, 2006). In the future, we can hope to see accurate determination of reaction fluxes in large reconstructed metabolic subnetworks, which should significantly improve, for example, our ability to manipulate the metabolism of plants (Ratcliffe and Shachar-Hill, 2006) and microorganisms (Nielsen, 2003) for useful purposes, such as biofuel production.

Analysis of isotopomer distributions in metabolic networks (e.g., for the purpose of estimating fluxes) requires tracing of atoms from substrates to products, which can be accomplished through a variety of mathematical formalisms, given knowledge of atom fates (Schmidt *et al*., 1997; Szyperski, 1995; Szyperski, 1998; Wiechert, 2001; Zupke and Stephanopoulos, 1994). In tracer-based flux profiling studies, this knowledge is usually assembled *ad hoc*, and the knowledge is stored in different formats across studies. Atom fate maps are not included in commonly used archives of metabolic knowledge, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Goto *et al*., 2002), MetaCyc (Krieger *et al*., 2004), BRENDA (Schomburg *et al*., 2004), or the biochemical pathways wall chart of Roche Applied Science (Michal, 1999). Only recently, has a large collection of fate maps been assembled (Arita, 2003). These maps were used to study the connectivity of the metabolic network of *Escherichia coli* (Arita, 2004). Interpretation of the maps, which account for 2,764 reactions involving 2,100 metabolites, depends largely on MDL® Mol files downloaded from the KEGG COMPOUND database, and their use for flux estimation purposes requires reformatting to obtain isotopomer mapping matrices (Schmidt *et al*., 1997) or one of the other data structures used in the available frameworks for isotopomer flux balance analysis.

Here, we add to the currently available collections of carbon-fate maps 1) by considering a greater number of reactions documented in the KEGG database [e.g., nearly twice as many reactions as considered by Arita (2003)], 2) by using a systematic structure-based system for naming and referencing metabolites and their carbon atoms (http://www.iupac.org/inchi/), 3) by accounting for prochiral carbon atoms, which are present in 17% of the metabolites considered, and 4) by encoding the fate maps in a formal model-specification language, the BioNetGen™ language (BNGL;

*To whom correspondence should be addressed.

Blinov *et al*., 2006; Faeder *et al*., 2005; Hlavacek *et al*., 2006). As a result of the latter feature, maps can be automatically interpreted by BioNetGen™ (Blinov et al., 2004; http://bionetgen.lanl.gov/) to obtain mass-balance equations and simulate stationary or dynamical labeling patterns. We demonstrate this capability through simple examples. Our database, which is freely available, includes carbon-fate maps for 4,605 reactions involving 3,713 metabolites.

2 METHODS

2.1 Reaction data and preprocessing

On June 8, 2006, the KEGG LIGAND database (Goto, et al., 2002) contained records for 6,577 reactions, which we downloaded. After inspection of these records, we eliminated 1,972 reactions from further consideration. A reaction was eliminated if it involved a polymeric reactant with a variable or unspecified number of monomer subunits, such as starch, or a reactant designated to contain a generic R group. Examples of such reactions include KEGG entries R04195 and R01345. We also eliminated a reaction if structural information (i.e., an MDL® Mol file) could not be found in the KEGG COMPOUND database for a reactant (e.g., C00947 or C00193 in R01487) or if the reaction was annotated as missing information (e.g., R00419). We ultimately assigned carbon fate maps for 4,605 reactions.

2.2 Metabolite data and generation of structure-based identifiers and carbon atom indices

The reactions of interest involve 3,713 metabolites. We downloaded MDL® Mol files from the KEGG COMPOUND database for these metabolites and converted the 2D chemical structure information encoded in these files into International Chemical Identifier (InChI™) strings using the 1.01 release of the InChI™ software (http://www.iupac.org/inchi/) – a technical manual distributed with the software documents the conventions and features of InChI™ strings (Stein et al., 2006). In generating InChI™ strings, we used the SUU program execution flag to force reporting of stereocenters in the stereochemistry layer of a string. An InChI™ string provides a unique identifier for a metabolite, one recognized by the International Union of Pure and Applied Chemistry (IUPAC), and a unique index for each non-hydrogen atom in a metabolite. By convention, atoms represented in an InChI™ string are indexed in Hill order (i.e., carbon atoms first). Thus, the carbon atoms in a metabolite are referenced by integer indices 1 to *n*, where *n* is the number of carbon atoms. In addition to generating an InChI™ string from an MDL® Mol file, the InChI™ software automatically generates and reports auxiliary information, including information about the constitutional equivalence of atoms and a map relating the atom indices in the InChI™ string to the atoms represented in the input MDL® Mol file.

To check the quality of structural information obtained from KEGG, we extracted MDL® Mol files and InChI™ strings for metabolites from the *BioMeta* database (Ott and Vriend, 2006), which provides corrections of constitution and stereochemistry for a number of chemical structures misreported in KEGG. In cases where the *BioMeta* and KEGG InChI™ strings were found to differ, we used the structural information from the *BioMeta* database.

2.3 Finding common substructures of substrate-product pairs

In identifying the parts of reactants affected and unaffected by reaction, we used SIMCOMP (http://web.kuicr.kyoto-u.ac.jpa/simcomp/) (Hattori, et al., 2003). This software tool implements graph comparison methods to identify the maximum common substructures of two chemical compounds. With this tool, we identified the common substructures of all possible substrate-product pairs for each reaction of interest. There is one pair for an A $\rightarrow$ B reaction, namely (A, B), and four pairs for an A+B $\rightarrow$ C+D reaction, namely (A, C), (A, D), (B, C) and (B, D). The inputs accepted by SIMCOMP are two KCF-formatted files specifying chemical structures. We generated KCF-formatted files from MDL® Mol files using a Perl script distributed with SIMCOMP (envatom2.pl). The output of SIMCOMP includes a list of matching atoms in the common substructure of the two input structures.

2.4 Distinguishing homotopic and prochiral carbon atoms

It is important to distinguish between two kinds of constitutionally equivalent atoms: those that are homotopic and those that are prochiral (Figure 1). Homotopic atoms (e.g., carbon atoms 1 and 2 in Figure 1A) are indistinguishable with respect to stable isotope labeling, whereas prochiral atoms (e.g., carbon atoms 1 and 2 in Figure 1B) are distinguishable, as their spatial environments differ (e.g., see Figure 1C). Because enzymes are chiral catalysts, prochiral atoms, even though they are constitutionally equivalent, can be selectively transformed in enzyme-catalyzed reactions (Eliel, 1982). An example is conversion of citrate to isocitrate. For this reaction, a partial mapping of carbon atoms from citrate to isocitrate is shown in Figure 1 (between panels 1C and 1D). Thus, when prochiral carbon atoms are present in a metabolite, we need to identify which of these atoms are reactive to define a correct mapping of atom fates. From the auxiliary information associated with an InChI™ string, we obtain a list of sets of constitutionally equivalent atoms. To determine if constitutionally equivalent carbon atoms are homotopic or prochiral, we use substitution and addition criteria (Eliel, 1982). In other words, we replace each constitutionally equivalent carbon atom with a distinct isotope and compare the 3D structures of the different isotopomers – these comparisons are aided by using InChI™ strings derived from MDL® Mol files with 3D atomic coordinates. The structures of the different isotopomers are identical if the substituted atoms are homotopic, whereas they are different (e.g., mirror images) if the substituted atoms are prochiral. 3D structures are generated from KEGG or *BioMeta* MDL® Mol files using MolConverter (http://www.chemaxon.com/marvin/doc/user/molconvert.html). For each metabolite found to contain prochiral carbon atoms, 3D atomic coordinates are added to the auxiliary information associated with the metabolite's InChI™ string. This structural information, which is required to determine the relative positions of prochiral carbon atoms, is considered in the manual specification of fate maps (see below).

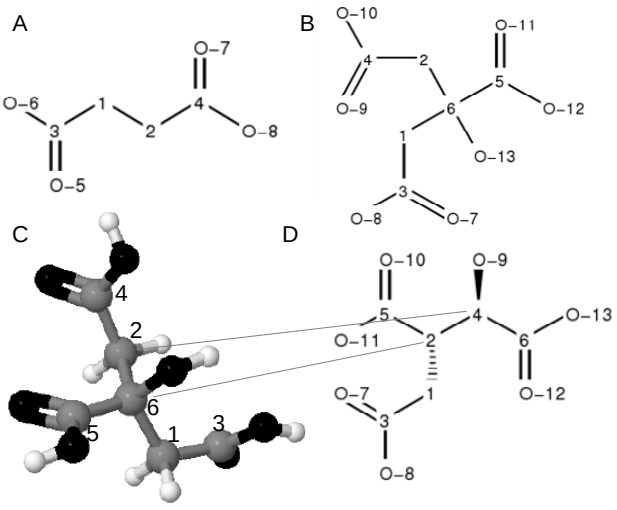

Fig. 1. Chemical structures of metabolites with homotopic and prochiral pairs of carbon atoms. A) 2D structure of succinate, which contains homotopic carbon atoms (1 and 2; 3 and 4). B) 2D structure of citrate, which contains prochiral carbon atoms (1 and 2; 3 and 4). C) 3D structure of citrate, in which the relative positions of the prochiral carbon atoms can be

seen. D) 2D structure of isocitrate. In the enzymatic conversion of citrate to isocitrate, atoms 2 and 6 of citrate map to atoms 4 and 2 of isocitrate, respectively, as indicated by the lines connecting the corresponding atoms. Atom 1 in citrate is constitutionally equivalent to atom 2 but is not reactive because of its distinct local environment.

2.5 Manual specification of carbon-fate maps

Using SIMCOMP outputs and the 3D chemical structures of prochiral metabolites for guidance, we manually defined reaction centers, as in earlier work (Mu *et al*., 2006), and an initial carbon fate map for each reaction of interest. To aid in the latter step, we wrote a special-purpose Java™ program (available upon request) that converts SIMCOMP outputs to BNGL-encoded fate maps. We also wrote and used FateMapViewer, which is included in the Supplementary material. This Java™ program visualizes fate maps using routines available in the Chemistry Development Kit (CDK) (Steinbeck *et al*., 2003). FateMapViewer accesses information in the fate map database, which is comprised of two Microsoft® Excel spreadsheets, through a Java™ application programming interface, JExcelApi (http://jexcelapi.sourceforge.net/). Maps were refined manually through text editing at the level of BNGL as needed. Information about reaction mechanisms was obtained primarily from textbooks (Berg *et al*., 2001; Silverman, 2000) and the web site of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (http://www.chem.qmul.ac.uk/iubmb/enzyme/). The primary literature was consulted as needed.

2.6 Notational conventions

Metabolites and carbon-fate maps are represented using BNGL (Blinov *et al*., 2006; Faeder *et al*., 2005; Hlavacek *et al*., 2006), an executable model-specification language. In this language, in general, text-encoded graphs are used to represent structured objects and graph-rewriting rules, which are called reaction rules, are used to represent (biochemical) transformations of objects.

In BNGL, an object has a name and a collection of named component parts, each of which can be associated with a set of attributes. Here, the objects are taken to be metabolites, which are named using InChI™ identifiers. Identifiers are delimited by quotation marks. The component parts of a metabolite are taken to be carbon atoms, which are named by appending atom indices appearing in the metabolite's InChI™ identifier to letter C's. For reaction rules that will be processed by BioNetGen™, homotopic carbon atoms must be relabeled such that they have the same name to ensure that they are appropriately treated as equivalent. We recommend that a pair of homotopic carbon atoms with indices i and j, where $i < j$, be represented as Ci_j. As usual in BNGL, component names are placed immediately after a metabolite (object) name and enclosed in parentheses. In the database, to identify reaction centers without using color, we use lower case letters to represent carbon atoms inside reaction centers, and we use upper case letters to represent carbon atoms outside reaction centers. *NB*: lower case letters should be converted to upper case letters before reaction rules are processed by BioNetGen™, as rule interpretation is case sensitive. Use of component attributes is not needed to define carbon-fate maps; however, attribute labels, which are prefixed by tilde characters, can be added as needed for a specific application to indicate mass numbers, which is important for distinguishing between ^{12}C and ^{13}C.

In BNGL, a reaction rule represents a generalized reaction that permits multiple reactants (e.g., isotopomers). It is comprised of lists of text-encoded (sub)graphs, each representing a collection of objects (e.g., the set of all glucose molecules labeled in any possible way with ^{13}C). In the rules considered here, each subgraph is identical to a graph for a metabolite up to attribute labels for mass numbers. A rule is divided into left- and right-hand sides by a transformation symbol, which indicates directionality. In the database, we have arbitrarily used '<->' (the symbol usually used for a bidirectional reaction) to divide the left- and right-hand sides of each reaction rule – this convention should be understood to be neutral with respect to reversibility of reactions, which we have not considered. In all cases we have preserved the ordering of reactants found in KEGG reaction records. In BNGL, component names appearing in a reaction rule may be augmented with mapping indices, which are prefixed by percent signs. The purpose of these indices is to explicitly indicate how the components of objects map from left to right and *vice versa*. Name sharing of a mapping index indicates correspondence between two components. Plus signs are used to separate reactants and indicate stoichiometry. The molecularity of a reaction leading from substrates to products is given by $p + 1$, where p is the number of plus signs on the substrate side of the reaction. Thus, for example, the map for a reaction with stoichiometry $2A \rightarrow B$ is represented as follows: $s_A(c1\%x_1,...,cn\%x_n) + s_A(c1\%x_{n+1},...,cn\%x_{2n})$ <-> $s_B(c1\%y_1,...,cm\%y_{2n})$, where s_A and s_B are quote-delimited InChI™ strings for metabolites A and B, $c \in \{c,C\}$ is a single character that precedes an atom index and indicates by case whether a carbon atom is in a reaction center, n is the number of carbon atoms in A, $m = 2n$ is the number of carbon atoms in B, and x_i, $y_i \in [1,...,m]$ are integer indices that indicate how the carbon atoms of A map to the carbon atoms of B. Atoms that share the same index map to each other.

3 RESULTS

3.1 Overview of database of carbon-fate maps for metabolic reactions

Using the approach described above, we have developed a database of carbon-fate maps for a significant fraction of the enzyme-catalyzed reactions documented in the KEGG database. The database is available as two Microsoft® Excel spreadsheets, which are included in the Supplementary material. The first spreadsheet provides information about 3,713 metabolites, 641 of which contain prochiral carbon atoms, and the second provides information about 4,605 reactions, including fate maps and references to the metabolites participating in these reactions.

Table 1 An entry for a metabolite in the database. A description of the fields is given in the text. For this metabolite, Fields 5, 6 and 8 are empty.

Field	Value
1	InChI=1/C4H6O4/c5-3(6)1-2-4(7)8/h1-2H2,(H,5,6)(H,7,8)
2	AuxInfo= 1/1/N:1,2,3,4,5,6,7,8/E:(1,2)(3,4)(5,6,7,8)/gE:(1,2)/rA:8nCCCCOOOO/rB:s1;s1;s2;d3;s3;d4;s4;/rC:3.7321,.25,0;4.5981,-.25,0;2.866,-.25,0;5.4641,.25,0;2.866,-1.25,0;2,.25,0;5.4641,1.25,0;6.3301,-.25,0;
3	KEGG:C00042; Succinate; Succinic acid; Butanedionic acid; Ethylenesuccinic acid
4	(1,2);(3,4)
7	http://www.genome.jp/dbget-bin/www_bget?compound+C00042

For each metabolite, the following fields are recorded in the database (see Table 1 for an example): 1) the InChI™ identifier of the metabolite, which includes up to six layers of information; 2) auxiliary information associated with the InChI™ identifier, including at least the information generated automatically by the InChI™ software when producing an identifier from an MDL® Mol file; 3) the KEGG COMPOUND accession number, common name of the metabolite, and synonyms; 4) lists of constitutionally equivalent carbon atoms (both prochiral and homotopic); 5) lists of prochiral carbon atoms; 6) comments; 7) links to the metabolite's

KEGG COMPOUND entry and *BioMeta* entry (if available); and 8) atomic coordinates of the 3D chemical structure (included only if a metabolite is prochiral). Lists of atoms in the fourth and fifth fields are enclosed in parentheses and separated by semicolons. In these fields, atoms are referenced by the unique atom indices of the InChI™ identifier, given in the first field. Table 1 illustrates the information recorded in the database for succinate, the structure of which is depicted in Figure 1A.

For each reaction, the following fields are recorded in the database (see Table 2 for an example): 1) the name of the enzyme that catalyzes the reaction; 2) the corresponding Enzyme Commission (EC) number, if available from KEGG; 3) the carbon-fate map we have defined for the reaction using InChI™ identifiers to name reactants and BNGL to encode the mapping of carbon atoms; 4) comments; 5) the KEGG LIGAND accession number; and 6) a link to the reaction's KEGG LIGAND entry. Table 2 illustrates the information recorded in the database for the reaction depicted in Figure 2. A fate map includes an identification of carbon atoms in reaction centers and all the information required for display by FateMapViewer, our software tool for visualizing carbon fate maps.

Table 2 An entry for a reaction in the database. A description of the fields is given in the text. For this reaction, Field 4 is empty. Figure 2 illustrates a FateMapViewer visualization of the carbon-fate map defined for this reaction in Field 3. The three metabolites participating in this reaction are indicated by the InChI™ strings included in the map.

Field	Value
1	D-Fructose-1,6-bisphosphate D-glyceraldehyde-3-phosphate-lyase
2	4.1.2.13
3	"InChI=1/C6H14O12P2/c7-4-3(1-16-19(10,11)12)18-6(9,5(4)8)2-17-20(13,14)15/h3-5,7-9H,1-2H2,(H2,10,11,12)(H2,13,14,15)/t3-,4-,5+,6?/m1/s1"(C1%1,C2%2,C3%3,c4%4,c5%5,C6%6) <-> "InChI=1/C3H7O6P/c4-1-3(5)2-9-10(6,7)8/h4H,1-2H2,(H2,6,7,8)"(c1%5,C2%2,C3%6)+"InChI=1/C3H7O6P/c4-1-3(5)2-9-10(6,7)8/h1,3,5H,2H2,(H2,6,7,8)/t3-/m0/s1"(c1%4,C2%1,C3%3)
5	R01068
6	http://www.genome.jp/dbget-bin/www_bget?rn:R01068

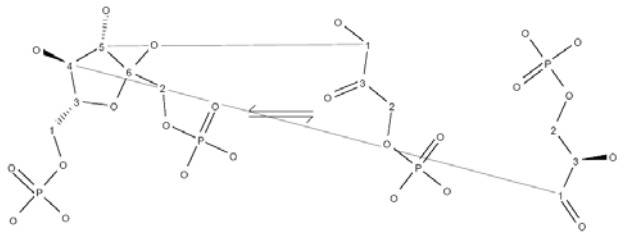

Fig. 2. Illustration of the carbon-fate map defined in Table 2. Lines indicate the fates of carbon atoms in reaction centers. Mapping of other carbon atoms is suppressed for clarity. FateMapViewer can be used to display carbon-fate maps as shown here. The enzyme that catalyzes this reaction and the metabolites involved are identified by their common names in the caption of Figure 3.

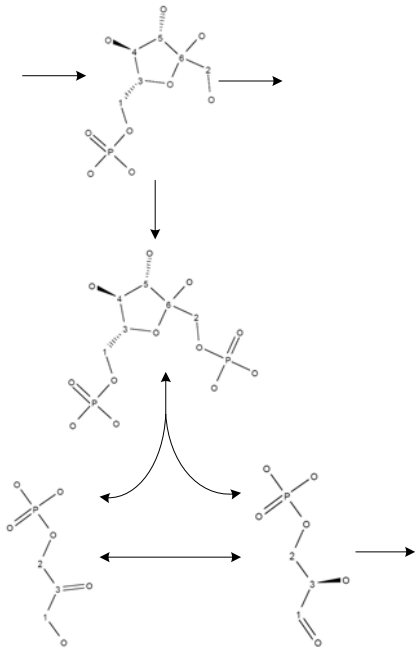

Fig. 3. A simple reaction network. The following abbreviations are used for enzymes catalyzing reactions in the network: PFK for D-fructose-6-phosphate 1-phosphotransferase, FBA for D-fructose-1,6-bisphosphate D-glyceraldehyde-3-phosphate-lyase, and TPI for D-glyceraldehyde-3-phosphate ketol-isomerase. The following abbreviations are used for metabolites in the network: F6P for D-fructose 6-phosphate, F16P for D-fructose 1,6-bisphosphate, DHAP for dihydroxyacetone phosphate, and T3P for D-glyceraldehyde 3-phosphate. Abbreviations are further defined in a file (.rtf format) included in the Supplementary material. Fluxes in and out of the system are denoted as follows: vi represents a flux feeding into the F6P pool, vs_F6P represents the flux that drains this pool for biomass production, and vs_T3P represents the combination of fluxes that drain the T3P pool.

For clarity, let us briefly note some of the nomenclatural features of the carbon-fate map defined in Table 2 – more details are provided in Section 2.6. This map characterizes the reaction F16P $\leftrightarrow$ DHAP + T3P (*cf.* Figures 2 and 3), where F16P, DHAP and T3P denote KEGG compounds C00354, C00111 and C00118, respectively. In the map, the reactants are named using InChI™ strings, which are comprised of layers (Stein et al., 2006). Layers and sublayers are separated by '/' signs. The main layer of each string considered here consists of three sublayers, the first of which gives a chemical formula (e.g., the chemical formula of F16P, $C_6H_{14}O_{12}P_2$, is denoted C6H14O12P2). The second sublayer gives the connectivity of non-hydrogen atoms, which are represented by integers. This sublayer describes a traversal of a molecular graph that visits all edges (bonds) connecting non-hydrogen atoms. Dashes are used to separate integers as needed, and parentheses and commas are used to describe branching. The next sublayer describes bonds to hydrogen atoms. A stereochemistry layer is included in the string for F16P and T3P, but not DHAP. Unneces-

sary layers of an InChI™ string are omitted by convention. Besides InChI™ strings, the map of Table 2 includes an explicit mapping of carbon atoms from F16P to DHAP and T3P and *vice versa*. In the mapping, which is specified according to the conventions of BNGL, each of the six carbon atoms in F16P is paired with a carbon atom in either DHAP or T3P. A pair is indicated by name sharing of a mapping index. Mapping indices appear after percent signs, which follow carbon names. Thus, the carbon atom in F16P named 'C1' is paired with the carbon atom in T3P named 'C2,' and the carbon atom in F16P named 'C2' is paired with the carbon atom in DHAP named 'C2.' The integers used to name carbon atoms follow from the canonical numbering scheme of the InChI™ generation process (Stein et al., 2006).

Figure 4 shows the statistics of reactions in the database according to EC number assignment. EC1 has the most reactions, whereas EC6 has the least reactions. A total of 571 reactions that are not associated with an EC number in the KEGG database are also included in the fate-map database.

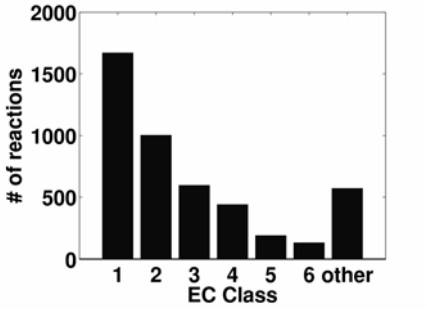

Fig. 4. Number of reactions in the database sorted into major EC classes. The last category includes reactions unassociated with an EC number in KEGG.

3.2 Example application of the database: deriving isotopomer mass-balance equations

Isotopomer mass-balance equations are essential for estimating reaction fluxes from measurements of the relative abundances of isotopomers (Sauer, 2004; Sauer, 2006). To specify such equations, we must know the fates of individual atoms in reactions. This type of knowledge is encoded in the carbon-fate maps of our database, which can be used (in multiple ways) to derive isotopomer mass-balance equations for analysis of data from experiments with ^{13}C-labeled compounds. To illustrate how, we will consider the simple reaction network of Figure 3, which includes three glycolytic reactions with maps in the database.

Using matrix methods (Schmidt *et al*., 1997; Zupke and Stephanopoulos, 1994), the information encoded in the three fate maps for the reactions of Figure 3 can be used to obtain balance equations that can be used to track labeling of the metabolites in all possible ways. These ordinary differential equations (ODEs) are conveniently written in the form exemplified below:

V_DHAP*d(DHAP_100)/dt = (F16P_000010 + F16P_000110 + F16P_001010 + F16P_001110 + F16P_100010 + F16P_100110 + F16P_101010 + F16P_101110) * v_FBA_f + (T3P_100) * v_TPI_b - (v_TPI_f + v_FBA_b) * DHAP_100 (1)

In this equation, the pool size of each metabolite is represented as V_*M*, where *M* is the abbreviation of a metabolite common name. The relative abundance of a metabolite fraction is represented as *M_b*, where *b* is a bit string. In the string, a '0' represents ^{12}C, and a '1' represents ^{13}C. Each position in the string corresponds to a particular carbon atom; atoms are ordered by their InChI™ string indices from left to right. A flux is represented as v_*E*, where *E* is the abbreviation of an enzyme's common name. Reversible reactions have fluxes in forward and backward directions, denoted by '_f' and '_b,' respectively. For a metabolite with *N* carbons, there are 2^N possible patterns of ^{13}C labeling. Thus, for the simple reaction network of Figure 3, 144 balance equations are required to account for all possible patterns. These equations are included in the Supplementary material in a plain-text file. However, for a given experiment, not all of these equations are relevant, as the labeling pattern of the feed metabolite (F6P) can restrict the labeling patterns of downstream metabolites.

BioNetGen™ can be used to automatically generate the balance equations relevant for a given experiment directly from fate maps. For example, in the scheme of Figure 3, if feed of labeled metabolite into the system (through flux vi) begins at time t=0 and consists of F6P with ^{13}C at position 1, then only 10 of the 144 equations are relevant. The BioNetGen™ input file that generates these equations (.bngl format) is included in the Supplementary material. This file is comprised mostly of carbon-fate maps extracted from the database – note that these maps have been edited for readability (mainly, InChI™ identifiers have been replaced by the abbreviations given in the caption of Figure 3) and to the meet the requirements indicated in Section 2.6 for proper processing by BioNetGen™, namely lower-case letters (c's) used to identify the carbon atoms in reaction centers have been replaced by upper-case letters (C's). The BioNetGen™ input file also specifies a simulation of labeling dynamics; a MATLAB® M-file specifying the same simulation, which is an output of BioNetGen™, is included in the Supplementary material for comparison. In the simulation, continuous feed of F6P labeled at position 1 begins at time t=0 and reaction fluxes (including vi) and the total concentrations of metabolites remain fixed. For t<0, the system is in a steady state and there is no label in the system.

To further demonstrate the utility of the database, we have built a model that can be used to analyze isotope labeling patterns and determine fluxes (through fitting) in a metabolic network of a size that is relevant for practical applications, i.e., comparable in size to networks considered in published work on isotopomer flux balance analysis (Sauer, 2006). The model accounts for labeling of 29 metabolites involved in 30 enzyme-catalyzed reactions of the central carbon pathway of *E. coli* after feeding a mix of unlabeled and ^{13}C-labeled glucose to a steady-state growth culture. The reachable network of isotopomers includes 355 species, which are connected by 1,043 reaction links. The model/simulation specification is provided in the Supplementary material as a BioNetGen™ input file (.bngl format) and as a BioNetGen™-generated MATLAB® M-file (.m format). It should be noted that the former file is far more readable than the latter, as the BioNetGen™ input file consists mostly of carbon-fate maps obtained largely through copy-paste operations from the database. (As before, maps were edited for readability and consistency with the conventions detailed in Section 2.6.) BioNetGen™ can process the input file to generate and solve isotopomer balance equations with correct automatic handling of molecular symmetry (e.g., as found in succinate and malate) and prochirality (e.g., as found in citrate). The model can

be simulated using conventional procedures. For example, the built-in ODE solver of BioNetGen™, which is based on CVODE (Hindmarsh *et al.*, 2005), is adequate.

4 DISCUSSION

We have built a database of executable carbon-fate maps for a large collection of metabolic reactions. This resource should be useful for pathway characterization. The ability to account for individual atoms in a mathematical model is essential for analysis of data from isotope tracer experiments and for quantitative estimation of fluxes in a metabolic network. Isotopic labeling, especially ^{13}C labeling, is increasingly being used for this purpose (Sauer, 2004; Sauer, 2006). Analysis of label distribution data depends on isotopomer mass-balance equations (or the equivalent), and our database provides the information required to obtain such equations.

Compound identification and atom identification were systematically considered in design of the database. We use InChI™ strings to represent metabolites. The InChI™ system is the most recent entry in the field of cheminformatics and has a number of noteworthy advantages compared with other naming systems (Coles *et al*., 2005; Prasanna *et al*., 2005; Richard *et al*., 2006). InChI™ provides a precise, robust, IUPAC-approved tag for a chemical substance, and it is derived from a structural representation of that substance in a way designed to be independent of how the structure may be drawn. InChI™ also provides a unique atom index for each non-hydrogen atom in a chemical, which depends only on the molecular structure, and the index is part of the InChI™ string. Compared with registry systems, InChI™ does not depend on the existence of a database of unique substance records to establish the next available sequence number for any new chemical substance being assigned an InChI™, and InChI™ generation and handling software are free and open source. It should be possible to link the carbon indices directly to other chemical information systems employing InChI™ identifiers, such as NMRShiftDB (Steinbeck *et al*., 2003) and METLIN (Smith *et al*., 2005).

Molecular symmetry and prochirality are important for carbon tracing. Homotopic carbons are not distinguishable in labeling experiments, whereas prochiral carbons are distinguishable and may have different reactivities because enzymes are chiral catalysts. Using the constitutionally equivalent atoms reported by the InChI™ software, the homotopic and prochiral carbons can be identified. For prochiral atom pairs, we examined 3D structures to distinguish them. Prochirality was not accounted for in earlier libraries of carbon-fate maps (Arita, 2003). Our library is also improved by incorporation of recent corrections of KEGG structural data (Ott and Vriend, 2006).

In summary, the carbon-fate maps reported here ease the task of high-throughput flux profiling of reactions in metabolic networks in two ways: 1) high-quality unambiguously defined carbon-fate maps have been assembled for diverse and numerous enzyme-catalyzed reactions, and 2) the maps are provided in a format amenable to modeling and computation. With these maps, models capable of tracking individual atoms can now be readily built not only for small metabolic networks but also for large networks. However, the usefulness of models for large networks is uncertain at this time because of experimental limitations and the curse of dimensionality. In the future, fate maps could be defined for other atoms, such as nitrogen, sulfur, phosphorus, oxygen and hydrogen. Updates of our database will be announced at the BioNetGen™ web site.

ACKNOWLEDGEMENTS

We thank G. Matthew Fricke for help with visualization of carbon fate maps, Jin Yang for help with MATLAB®, and Dimitrii Tchekhovskoi for help with InChI™ software and identifiers. This work was supported by NIH, through grants GM35556 and GM76570, and by DOE, through contract DE-AC52-06NA25396 and the Genomics:GTL program.

REFERENCES

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

Arita, M. (2004) The metabolic world of *Escherichia coli* is not small, *Proc. Natl. Acad. Sci. U.S.A.*, 101, 1543-1547.

Berg, J.M., Tymoczko, J.L. and Stryer, L. (2001) *Biochemistry* W. H. Freeman and Company, New York.

Birkemeyer, C., Leudemann, A., Wagner, C., Erban, A. and Kopka, J. (2005) Metabolome analysis: the potential of in vivo labeling with stable isotopes for metabolite profiling, *Trends Biotechnol.*, 23, 28-33.

Blinov, M.L., Faeder, J.R., Goldstein, B. and Hlavacek, W.S. (2004) BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains, *Bioinformatics* 20, 3289-3291.

Blinov, M.L., Yang, J., Faeder, J.R. and Hlavacek, W.S. (2006) Graph theory for rule-based modeling of biochemical networks, *Lect. Notes Comput. Sci.*, 4230, 89-106.

Boros, L.G., Cascante, M. and Lee, W.-N.P. (2003) Stable isotope-based dynamic metabolic profiling in disease and health—tracer methods and applications. In Harrigan, G.G. and Goodcare, R. (eds), *Metabolic profiling: its Role in Biomarker Discovery and Gene Function Analysis*. Kluwer Academic Publishers, Boston, 141-170.

Coles, S.J., Day, N.E., Murray-Rust, P., Rzepa, H.S. and Zhang, Y. (2005) Enhancement of the chemical sematics web through the use of InChI identifier, *Org. Biomol. Chem*., 3, 1832-1834.

Eliel, E.L. (1982) Prostereoisomerism (prochirality), *Top. Curr. Chem*., 105, 1-76.

Emmerling, M., Dauner, M., Ponti, A., Fiaux, J., Hochuli, M., Szyperski, T., Wüthrich, K., Bailey, J. and Sauer, U. (2002) Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*, *J. Bacteriol*., 184, 152-164.

Faeder, J.R., Blinov, M.L., Goldstein, B. and Hlavacek, W.S. (2005) Rule-based modeling of biochemical networks, *Complexity*, 10, 22-41.

Fischer, E. and Sauer, U. (2005) Large-scale in vivo flux analysis shows rigidity and sub-optimal performance of *Bacillus subtilis* metabolism, *Nat. Genet*., 37, 636-640.

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

Hattori, M., Okuno, Y., Goto, S. and Kanehisa, M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, *J. Am. Chem. Soc*., 125, 11853-11865.

Hellerstein, M. (2003) In vivo measurement of fluxes through metabolic pathways: the missing link in functional genomics and pharmaceutical research, *Annu. Rev. Nutr.*, 23, 379-402.

Hellerstein, M. and Murphy, E. (2004) Stable isotope-mass spectrometric measurements of molecular fluxes in vivo: emerging applications in drug development, *Curr. Opin. Mol. Ther*., 6, 249-264.

Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R., Shumaker, D.E. and Woodward, C.S. (2005) SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. *ACM Trans. Math. Software* 31, 363-396.

Hlavacek, W.S., Faeder, J.R., Blinov, M.L., Posner, R.G., Hucka, M. and Fontana, W. (2006) Rules for modeling signal-transduction systems, *Sci. STKE*, 2006, re6.

Hua, Q., Yang, C., Baba, T., Mori, H. and Shimizu, K. (2003) Responses of the central carbon metabolism in *Escherichia coli* to phosphoglucose isomerase and glucose-6-phosphate dehydrogenase knockouts, *J. Bacteriol.*, 185.

Kopka, J. (2006) Current challenges and developments in GC-MS based metabolite profiling technology, *J. Biotechnol.*, 124, 312-322.

Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S.Y. and Karp, P.D. (2004) MetaCyc: A multiorganism database of metabolic pathways and enzymes, *Nucleic Acids Res.*, 32, D438-442.

McCabe, B.J. and Previs, S.F. (2004) Using isotope tracers to study metabolism: application in mouse models, *Metab. Eng.*, 6, 25-35.

Michal, G. (1999) *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. Wiley & Sons, New York.

Mu, F., Unkefer, P.J., Unkefer, C.J. and Hlavacek, W.S. (2006) Prediction of oxidoreductase-catalyzed reactions based on atomic properties of metabolites, *Bioinformatics*, **22**, 3082-3088.

Nielsen, J. (2003) It is all about metabolic fluxes, *J. Bacteriol.*, 185, 7031-7035.

Ott, M.A. and Vriend, G. (2006) Correcting ligands, metabolites, and pathways, *BMC Bioinformatics*, 7:517.

Prasanna, M.D., Vondrasek, J., Wlodawer, A. and Bhat, T.N. (2005) Application of InChI to curate, index, and query 3-D Structures, *PROTEINS: Structure, Function, and Bioinformatics*, 60, 1-4.

Ratcliffe, R.G. and Shachar-Hill, Y. (2006) Measuring multiple fluxes through plant metabolic networks, *Plant J.*, 45, 490-511.

Richard, A.M., Gold, L.S. and Nicklaus, M.C. (2006) Chemical structure indexing of toxicity data on the internet: moving toward a flat world, *Curr. Opin. Drug Discov. Devel.*, 9, 314-325.

Sauer, U. (2004) High-throughput phenomics: experimental methods for mapping fluxomes, *Curr. Opin. Biotechnol.*, 15, 58-63.

Sauer, U. (2006) Metabolic networks in motion: ^{13}C-based flux analysis, *Mol. Syst. Biol.*, 2, 62.

Schmidt, K., Carlsen, M., Nielsen, J. and Villadsen, J. (1997) Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices, *Biotechnol. Bioeng.*, 55, 831-840.

Schomburg, I., Chang, A. and Schomburg, D. (2004) BRENDA, enzyme database: updates and major new developments, *Nucleic Acids Res.*, 32, D431-433.

Silverman, R.B. (2000) *The organic chemistry of enzyme-catalyzed reactions*. Academic Press, San Diego.

Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R. and Siuzdak, G. (2005) METLIN: a metabolite mass spectral database, *Ther. Drug Monit.*, 27, 747-751.

Stein, S.E., Heller, S.R. and Tchekhovskoi, D.V. (2006) The IUPAC Chemical Identifier - Technical Manual. Distributed with the InChI 1.01 software release, available at http://www.iupac.org/inchi/

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. and Willighagen, E.L. (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics, *J. Chem. Inf. Comput. Sci.*, 43, 493-500.

Steinbeck, C., Kuhn, S. and Krause, S. (2003) NMRShiftDB—constructing a chemical information system with open source components, *J. Chem. Inf. Comput. Sci.*, 43, 1733-1739.

Szyperski, T. (1995) Biosynthetically directed fractional ^{13}C-labeling of proteinogenic amino acids: an efficient analytical tool to investigate intermediary metabolism, *Eur. J. Biochem.*, 232, 433-448.

Szyperski, T. (1998) ^{13}C-NMR, MS and metabolic flux balancing in biotechnology research, *Q. Rev. Biophys.*, 31, 41-106.

Wiechert, W. (2001) ^{13}C metabolic flux analysis, *Metab. Eng.*, 3, 195-206.

Zupke, C. and Stephanopoulos, G. (1994) Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrices, *Biotechnol. Prog.*, 10, 489-498.