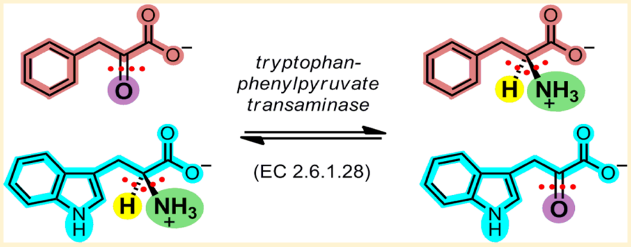
JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Accurate Atom-Mapping Computation for Biochemical Reactions

Mario Latendresse,* Jeremiah P. Malerich, Mike Travers, and Peter D. Karp

SRI International, 333 Ravenswood Ave., Menlo Park, California 94025, United States



ABSTRACT: The complete atom mapping of a chemical reaction is a bijection of the reactant atoms to the product atoms that specifies the terminus of each reactant atom. Atom mapping of biochemical reactions is useful for many applications of systems biology, in particular for metabolic engineering where synthesizing new biochemical pathways has to take into account for the number of carbon atoms from a source compound that are conserved in the synthesis of a target compound. Rapid, accurate computation of the atom mapping(s) of a biochemical reaction remains elusive despite significant work on this topic. In particular, past researchers did not validate the accuracy of mapping algorithms. We introduce a new method for computing atom mappings called the minimum weighted edit-distance (MWED) metric. The metric is based on bond propensity to react and computes biochemically valid atom mappings for a large percentage of biochemical reactions. MWED models can be formulated efficiently as Mixed-Integer Linear Programs (MILPs). We have demonstrated this approach on 7501 reactions of the MetaCyc database for which 87% of the models could be solved in less than 10 s. For 2.1% of the reactions, we found multiple optimal atom mappings. We show that the error rate is 0.9% (22 reactions) by comparing these atom mappings to 2446 atom mappings of the manually curated Kyoto Encyclopedia of Genes and Genomes (KEGG) RPAIR database. To our knowledge, our computational atom-mapping approach is the most accurate and among the fastest published to date. The atom-mapping data will be available in the MetaCyc database later in 2012; the atom-mapping software will be available within the Pathway Tools software later in 2012.

■ INTRODUCTION

The atom mapping of a chemical reaction is a bijection of the reactant atoms to the product atoms that specifies the terminus of each reactant atom. This paper considers only biochemical reactions, although the approach presented can be applied to chemical reactions in general.

In systems biology, the accurate atom mapping of biochemical reactions is fundamental for many applications. For example, atom mappings can be used to compute the number of carbon atoms conserved when a source compound is transformed to a target compound in an engineered metabolic pathway to determine the efficiency of the pathway. Atom mappings can also be used to track atoms in the isotope labeling experiments that have been commonly used to elucidate metabolic pathways for decades. Atom mappings can also be used to visually color corresponding atoms in reaction diagrams to make reaction mechanisms more visually apparent.

Some previous works mapped only some atom species, such as carbon, nitrogen, and sulfur. We aim at accurately computing the atom mappings for all nonhydrogen atoms in biochemical reactions, not just a specific set of species. Considering all atom species greatly complicates achieving high accuracy because some atom species, such as oxygen, are more difficult to trace in biochemical reactions. The version of our algorithm presented herein does not map hydrogen atoms because we used the KEGG RPAIR database[1] for comparison with our mapping results and that database does not map hydrogen atoms. Accordingly, it would not have been possible to validate the quality of our mapping of hydrogens. Also, mapping specific hydrogen atoms has much less utility than mapping nonhydrogen atoms. Hydrogen transfer reactions are very fast, and consequently, the polarized O−H, N−H, and S−H bonds exchange readily with the aqueous environment, negating the potential value of mapping these hydrogen atoms. On the other hand, most of the nonpolar C−H bonds in biomolecules are unreactive. Therefore, an accurate mapping of carbon atoms often provides enough information to understand the mapping of hydrogens bonded to carbon.

Historically, several metrics have been used to compute atom mappings. The two most popular ones are the Maximum Common Subgraph (MCS)[2−6] and bond edit-distance.[7] First et al.[7] implement bond edit-distance using a Mixed-Integer Linear Programming (MILP) technique, and as far as we know, this was the first publication applying MILP to the atom-mapping problem. Most past publications on this topic have focused on how to compute these metrics efficiently for chemical structures; as far as we know, the accuracy of these methods has not been assessed. These methods do not require advanced chemistry modeling, which is an advantage for simplicity.

In this work, we present a minimum weighted edit-distance (MWED) metric that can be computed efficiently, which yields more atom mappings that are valid than do previous techniques. More precisely, integer weights are assigned to almost all bonds of all reactants and products in a reaction.

Received: May 7, 2012

These integers represent the propensity of the bonds to break or form. In atom mapping, some bonds are made or broken or change type (e.g., single bond to double bond). We associate a cost when a bond changes type, breaks, or forms, based on the weights assigned to the bonds involved. The weighted edit-distance of atom mapping is the sum of the costs of the bonds broken, made, or changing type. We show that the MWED metric yields more chemically accurate mappings than does the unweighted bond-edit metric. With only a small amount of added complexity, the MWED metric increases the amount of chemical knowledge in modeling chemical reactions. In other words, although our modeling requires a small amount of chemical knowledge, it does not rely on advanced chemical knowledge such as electron density maps, consideration of reaction intermediates, or transition states. In addition, we will show that MWED can avoid the complexity of stereochemistry, in almost all cases, to compute the chemically valid atom mappings.

The contributions of our work are as follows. (1) We present a MILP-based atom-mapping method with the following novel properties: (a) It includes a method for matching rings. (b) It uses a complete weighted edit-distance metric on bonds. (c) Its maximization expression is different from First et al.[7] (d) It uses a much simpler approach for stereochemistry than did First et al.,[7] thereby simplifying the implementation (we report some cases in which the simplified method fails). (2) In contrast to most previous work, we evaluate the accuracy of our method empirically with respect to a manually curated collection of atom mappings; the accuracy is shown to be high. (3) In contrast to most previous work, we evaluate the fraction of solutions for which multiple mappings are obtained—our method yields multiple mappings for a very small fraction of reactions, which is significant because it means manual review of the output of our method to select a correct mapping is required in only a small number of cases. (4) We apply the method to a large and widely used database of metabolic pathways and reactions, MetaCyc,[8] thereby enhancing the value of that database.

The use of a MILP technique has the major advantage of portability to multiple existing MILP solvers. That is, a MILP formulation is a general description of an optimization problem that can be solved by multiple MILP solvers.[9,10] Moreover, the speed of MILP solvers is steadily increasing, which translates to immediate gains in speed for solving all problems that use a MILP formulation. Some MILP solvers use multiple-core processors for increased speed, thereby reducing the time needed to find optimal mapping. Implementing an *ad hoc* multiple-core algorithm is complex, requiring a major effort in software development.

The disadvantage of a MILP technique is the potential complexity of the formulation of the models, which can lead to inefficiency when solving them. Nevertheless, our work has shown that some heuristics can be added to the MILP formulation computing atom mappings to reduce its size and decrease the time to solve it. In fact, we show that the MILP technique compares very favorably in computation speed with MCS.

We have applied our approach to the reactions of the MetaCyc database[8] (see Application to MetaCyc section). The comparison of the computed atom mappings with the manually curated KEGG RPAIR database (described in Comparison to KEGG RPAIR) indicated an error rate of 0.9%. The section Solving Speed, discusses the computational speed of our approach. The next section, Computational Modeling, describes our computational approach.

COMPUTATIONAL MODELING

We present a MILP formulation to find the MWED atom mapping for a given reaction. In general, a MILP formulation has the form

$$\begin{aligned} &\max\ C^{\mathrm{T}}X \\ &AX \leq B \\ &X > 0 \end{aligned} \tag{1}$$

The vector X of variables must satisfy constraints expressed by the matrices A and B, and the maximization of the objective function is based on a vector of coefficients C (C^{T} is the transpose of vector C). Such a formulation can be solved by an integer linear program solver such as Solving Constraint Integer Programs (SCIP)[9] or CPLEX.[10]

In a MILP, some variables have integer values and some variables have continuous (i.e., noninteger) values. When a variable takes only the values 0 or 1, it is called a binary variable. In the following formulations, all integer variables are binary in the solution because of the specified constraints. They are not declared binary, however, to let the solver represent the variables in the most efficient manner.

Let A_{r} be the set of all atoms in the reactant compounds, A_{p} the set of all atoms in the product compounds, B_{r} the set of bonds in the reactant compounds, and B_{p} the set of bonds in the product compounds. A bond is represented as a tuple (a,x) where a and x are atoms. We denote by $t(a,b) \in \{1,2,3\}$ the bond type between a and b where 1 is for a single bond, 2 is for a double bond, and 3 is for a triple bond. We denote by $s(a)$ the species (e.g., carbon) of atom a. Note that the number of atoms of each species in A_{r} is the same in A_{p}. Otherwise, no chemically valid bijection could exist between A_{r} and A_{p}. An atom a of A_{r} can be mapped only to an atom x of A_{p} if and only if $s(a) = s(x)$. The possible mapping of a to x is controlled by the binary variable m_{ax}; that is, a is mapped to x if and only if $m_{ax} = 1$. Because every chemically valid mapping is a bijection, we have the following constraints

$$\forall\ a \in A_{\mathrm{r}},\quad \sum_{x \in A_{\mathrm{p}}, s(x)=s(a)} m_{ax} = 1 \tag{2}$$

Constraints 2 ensure that each atom on the reactant side maps to exactly one atom on the product side. We also add the following constraints

$$\forall\ x \in A_{\mathrm{p}},\quad \sum_{a \in A_{\mathrm{r}}, s(x)=s(a)} m_{ax} = 1 \tag{3}$$

Constraints 3 ensure that each atom on the product side receives exactly one atom from the reactant side. There are $|A_{\mathrm{r}}| + |A_{\mathrm{p}}|$ constraints for all m variables.

We define variables $e_{abxy} \in [0,1]$ to control the mapping of bond $(a,b) \in B_{\mathrm{r}}$ to bond $(x,y) \in B_{\mathrm{p}}$; that is, we have $e_{abxy} = 1$ if and only if $m_{ax} = 1$ and $m_{by} = 1$. If e_{abxy} is a defined variable, the variable e_{bayx} is not defined because it is essentially the same bond. There are a maximum of $|B_{\mathrm{r}}| \times |B_{\mathrm{p}}|$ e variables. Notice that there are no e variables for cases where a bond is broken or made. This is correct because, as we show below, the cases where a bond could be made or broken do not provide a gain for the objective function. The previous constraint applies to all existing bonds, so we have the following constraints

$$\forall\ e_{abxy}(m_{ax} \le e_{abxy} \wedge m_{by} \le e_{abxy}) \tag{4}$$

The objective function of the models could directly compute the minimum cost of breaking and forming bonds, but computing the maximum gain of maintaining the bonds between the atoms is more efficient because a typical compound has fewer bonds than nonbonds. For each variable, the objective function records the gain of mapping a to x and b to y. That is, for each e_{abxy}, we associate a coefficient g_{abxy}, which we now discuss how to compute.

We consider three bond types: single, double, and triple. The gain g_{abxy} depends on two main factors: the type of bonds of (a,b) and (x,y) and the species of the atoms a, b, x, and y. For each pair of atom species a and b, we attribute a *propensity* value for a single bonded (a,b) to break or form.

We also attribute a propensity of a double bonded (a,b), if that bond exists among those atoms, to change to a single bonded (a,b), or vice versa. Table 1 indicates the propensities of bonds to break or form. A chemist (the second coauthor) has determined the table content and adjusted it according to computational experimentation. The propensity of a bond to react is inversely proportional to the value given in the table; that is, the lower the value assigned to a bond, the higher its propensity to react. For example, the propensity is 400 for a single bond to break or form between two carbon atoms, which means that C−C bonds have a very low propensity to react. The table also shows the propensity of a double bond to become a single bond or vice versa.

Table 1. Bond Values To Compute Parameters g_{abxy} in the Objective Function and for $T(x,y)$ of Eq 5[a]

	C	O	N	P	H	S
C	400|24	48*|8	56*|8	48	72	48*
O	48*|8	16|8	8|72	8*|72	4	8|72
N	56*|8	8|72	16	8	8	24
P	48	8*|72	8	na	na	8
H	72	4	8	na	na	8
S	48*	8|72	24	8	8	16

[a] An entry for atoms x and y with a single integer c represents the value of a single bond being made or broken between atoms x and y; that is, $T_1(x,y) = c$. For an entry with two integers $c|k$, the first integer c has the same meaning, that is, $T_1(x,y) = c$, and the second integer k is the value of a single bond becoming a double bond or vice versa between atoms x and y; that is, $T_{12}(x,y) = k$. An entry with na is a nonapplicable case as these bonds do not exist in biochemical compounds. See the Special Bond Values section for values assigned for entries marked by an asterisk (∗). Notice that (1) the table is symmetric along its main diagonal because a bond value for (a,b) is the same as a bond value for (b,a) and (2) this table does not show values for triple bonds because eq 5 handles them.

We denote by $T_1(a,b)$ the propensity of a single bond breaking or forming between a and b and by $T_{12}(a,b)$ the propensity of a single bond changing to a double bond or vice versa between a and b. The values for $T_1(a,b)$ and $T_{12}(a,b)$ are directly given by Table 1. We denote by $c(a,b) \in$ {up,down,none} the stereochemistry of bond (a,b). Finally, we denote by $T(a,b)$ the propensity of completely making or breaking the bond (a,b) whatever its type. Its value is defined by

$$T(x, y) = \begin{cases} T_1(x, y), \textbf{ if } t(x, y) = 1 \\ T_1(x, y) + T_{12}(x, y), \textbf{ if } t(x, y) = 2 \\ T_1(x, y) + 2T_{1,2}(x, y), \textbf{ if } t(x, y) = 3 \end{cases} \tag{5}$$

We want the gain g_{abxy} to be maximum when the bond type, including stereochemistry, between (a,b) and (x,y) does not change, whereas we want the gain to be zero when the bond (a,b) breaks or when the bond (x,y) is formed. Naturally, the objective function does not need to record zero gain, which explains that variables e_{abxy} are not defined when there is no bond between a and b or between x and y. We want the gain g_{abxy} to be less than the maximum when the bond type changes. More precisely we define

$$g_{abxy} = \begin{cases} T(a, b), \textbf{ if } t(a, b) = t(x, y), c(a, b) = c(x, y) \\ T(a, b) - 1, \textbf{ if } t(a, b) = t(x, y), c(a, b) \ne c(x, y) \\ T(a, b) - |t(a, b) - t(x, y)|T_{12}(a, b), \textbf{ if } t(a, b) > t(x, y) \\ T(x, y) - |t(a, b) - t(x, y)|T_{12}(x, y), \textbf{ if } t(a, b) < t(x, y) \end{cases} \tag{6}$$

The case when stereochemistry changes, that is $t(a,b) = t(x,y), c(a,b) \ne c(x,y)$, is discussed in more detail in the Stereochemistry section.

Note that because we are not mapping the hydrogen atoms, the column and row for H in Table 1 are not used when computing g_{abxy}. However, next we show how to take into account hydrogen atoms in the objective function on the basis of the values for H in Table 1.

The values in Table 1 are related to the propensity (or probability) of the bond types breaking or forming in biochemical reactions. Bond dissociation energy values are not appropriate because these values represent heterolytic electron movement. Most biochemical reactions proceed by homolytic pathways. [For example, the bond energy of an O−H bond (water: 118 kcal/mol) is higher than an alkyl C−H bond (ethane: 101 kcal/mol). However, the O−H bond of water reacts readily in biological systems, while alkyl C−H bonds react comparatively rarely.] An initial set of propensities was estimated by a survey of biochemical and enzyme catalyzed reactions. We applied the MWED method to MetaCyc reactions and inspected the results. For reactions that gave incorrect mappings, adjustments to the propensity values were made to give the correct mapping. The values in Table 1 were established after three rounds of iteration, and these values provide atom mappings for the MetaCyc reaction set with a low error rate (see Comparison to KEGG RPAIR section).

Biochemical compounds contain numerous hydrogen atoms; however, only a relatively small percentage of those atoms participate in biochemical reactions. As mentioned, hydrogen atoms are not mapped. However, as we do for bonds between hydrogen and nonhydrogen atoms, we include all changes of bonding between hydogen and nonhydrogen atoms, especially carbon, when computing the objective function because doing so gives more accurate atom mappings.

In the MILP formulation, hydrogens are not mapped; that is, there are no m and e variables for them. But, the broken and made bonds to hydrogens are taken into account in the objective function by adding terms $h_{ax}m_{ax}$, where $h_{ax} > 0$ is the cost of adding or removing hydrogen atoms when mapping atom a to atom x: h_{ax} is the sum of the $T(x,H)$ and $T(a,H)$ values, as given by Table 1, for the bonds x—H and a—H broken or made when a is mapped to x.

Finally, the objective function to maximize is

$$\sum_{abxy} g_{abxy}e_{abxy} - \sum_{ax} h_{ax}m_{ax} \tag{7}$$

Notice that the objective function has e variables with coefficients based on the *gain* of mapping bonds, whereas the m variables have coefficients that measure the *cost* of mapping atoms. It would also be possible to use a gain coefficient for the m variables, but it is typically better to use a cost because there are fewer m variables with a nonzero cost than a nonzero gain.

Figure 1 shows a reaction and a description of the e and m variables that are needed to represent all possible mappings.

This completes our discussion of MILP modeling to compute the atom mappings of any reaction. The next two sections present two variations to this modeling that reduce the number of variables of the models in order to reduce the time needed to solve them.

Specific Compound Mappings. Some reactant pairs (e.g., NADP to NADPH) are typically interconverted in biochemical reactions. If such a pair of compounds occurs, one as a reactant and the other as a product, we match them directly before any other processing is done. Doing so eliminates the need to represent the mapping of their bonds and atoms in the MILP formulation and reduces the size of the formulation, which typically decreases the time needed to solve it. We have applied this technique to the pair of molecules presented in Table 2.

Ring Mapping Technique. The majority of biochemical reactions involve compounds with ring structures. For example, MetaCyc (version 16.0) has 7851 reactions (76%) with at least one ring structure in their substrates. In the great majority of cases, ring structures are conserved between reactants and products, thus providing an opportunity to reduce the search for optimal atom mapping. Ring structures can be mapped directly to other ring structures without considering each of their atoms and bonds.

The technique of this section, which we call *ring mapping technique* (RMT), is applied to reduce the size of the MILP formulation and consequently to hopefully increase the speed in finding optimal solutions. This technique is also helpful with aromatic rings without having to identify them: directly mapping a ring does not need to consider each bond type in it, which could be problematic because the exact positioning of double and single bonds in an aromatic ring is arbitrary.

This RMT is applied only if it can be determined that there is at least one possible one-to-one mapping of the ring structures from reactants to products. This determination is verified by a program before the formulation of the MILP is done. More precisely, all potential one-to-one ring mappings are precomputed by a program that generates the MILP formulation.

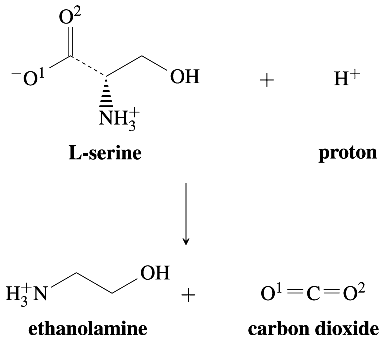


Figure 1. A carbon-lyase reaction (EC 4.1.1.-) with its atom mapping described by a broken bond shown as a dashed line and the mapping of the two oxygen atoms of L-serine tagged with superscripts 1 and 2 to the corresponding tagged oxygen atoms of carbone dioxide. Only one such possible mapping is shown, as these two oxygen atoms are indistinguishable. Another atom mapping is obtained by switching the two oxygen atoms of carbon dioxide, but that mapping is considered equivalent because these atoms are indistinguishable. There are 19 m binary variables to represent the one-to-one mapping of atoms: (1) 9 m variables for all possible mappings of the three carbon atoms of L-serine to the three carbon atoms of ethanolamine, (2) 9 more m variables for all possible mappings of the three oxygen atoms of L-serine to the three oxygen atoms of ethanolamine, and (3) 1 more m variable for the mapping of the nitrate atom. There are 14 e nonbinary variables to represent the mappings of the bonds: (1) 4 e variables for all possible mappings of the two C—O bonds from L-serine to the two C—O bonds of ethanolamine, (2) 9 more e variables for all possible mappings of the three C—C bonds of L-serine to the three C—C bonds of ethanolamine, and (3) 1 more e variable for the mapping of the N—C bond. The objective function to maximize is the weighted gain of keeping each bond unbroken/made minus the cost of removing or adding H atoms according to eq 6. The total gain of the this mapping is 491 based on: (1) $T_1(C,C) = 400$ to have kept one C—C bond unbroken/made, (2) $T_{12}(C,O) + T_1(C,O) = 56$ to have kept one C═O bond unbroken/made, (3) $T_1(C,O) = 48$ to have kept one C—O bond unbroken/made, (4) $T_1(C,N) - 1 = 55$ to have mapped one C—N bond to a C—N down bond, (5) $T_{12}(C,O) = 8$ to have mapped one C—O bond to a C═O bond, (6) $-T(C,H) = -72$ to have to add a H atom to C when mapping the C—N bond, and (7) $-T(O,H) = -4$ to have to add a H atom to O when mapping C—O$^-$ to C═O.

Table 2. List of Pairs of Compounds for Which Specific Direct Mappings Were Conducted before Formulating the MILP[a]

compound 1	compound 2	number of reactions in MetaCyc
fad	fadh2	34
nadp	nadph	963
nad	nadh	726

[a] These direct mappings decreased the size of the MILP formulations and increased the speed for finding their solutions.

A reactant ring is tentatively mapped to a product ring if and only if they are similar. Two rings are similar if (1) they have the same sequence of atom species and (2) for every atom of the rings, their neighboring atoms have at most one different atom species. Condition (1) is necessary, otherwise the mapping could not be valid. Condition (2) ensures that no major chemical transformation occurs around the ring, which otherwise could be a sign that the ring is broken and then reformed. Notice that given two rings, more than one mapping between them is possible due to rotation.

For example, Figure 2 shows a reaction with two rings on both sides of the reaction. Each ring on one side is similar to exactly one ring on the other side: only one mapping of the rings is possible in that case. Although, for each side of the reaction, the two rings are attached, that is, two rings share two carbon atoms, each ring is mapped independently.

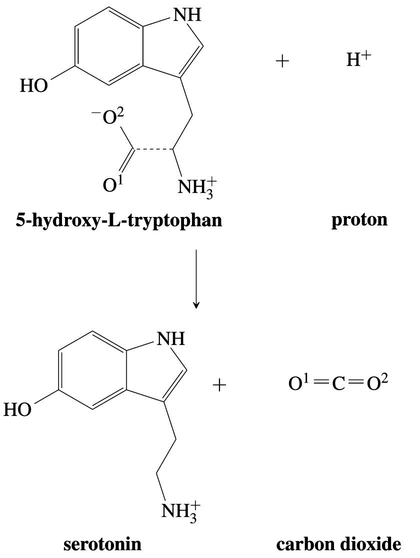


Figure 2. Reaction aromatic-L-amino-acid decarboxylase has two reactant rings that are similar to two product rings. The RMT is used in that case. Only 2 binary *r* variables are used to represent this possible ring mapping and 19 *m* binary variables to control the mapping of individual atoms: (1) 9 *m* variables for all possible mappings of the three carbon atoms of 5-hydroxy-L-tryptophan, not in a ring, to the three carbon atoms, not in a ring, of serotonin, (2) 9 more *m* variables for all possible mappings of the three oxygen atoms of 5-hydroxy-L-tryptophan to the three oxygen atoms of serotonin, and (3) 1 more *m* variable for the mapping of the nitrate atom of 5-hydroxy-L-tryptophan, not in the ring, to the nitrate atom, not in the ring, of serotonin. There are 21 nonbinary *e* variables to represent all possible bond mappings. They are used in the objective function with the *m* variables to minimize the cost of breaking and forming bonds. The model is feasible because the two rings do not need to be separated to get a valid reaction. A single bond is broken in 5-hydroxy-L-tryptophan and shown as a dashed line. Notice that the mapping of the two oxygen atoms of carbon dioxide, superscripted with integers 1 and 2, can actually be done in two different equivalent ways.

Similar ring structures are not found by the MILP formulation but are determined by a program that generates the MILP formulation. That program also determines if the technique of this section can be applied. The **technique of this section is applied** if and only if (1) each ring on one side of the reaction has at least one similar ring on the other side of the reaction and (2) the number of rings are the same on each side of the reaction. If this condition is met, the technique is used to potentially confirm, by the MILP solver, that such a mapping is possible. Because a reactant ring can be similar to many product rings, many potential ring mappings must be considered to find the optimal one. These considerations are done by the MILP solver on the basis of the generated formulation.

For the following formulation, we assume that if there is at least one mapping of the similar rings from reactants to products, then no rings are broken or made. The section below discusses what to do if this assumption proves false.

We denote by r_{AB} a binary ring-mapping variable from reactant to product where $A = (a_i)$, $1 \leq i \leq n$ and $B = (b_j)$, $1 \leq j \leq n$ are equal length sequences of $n = |A| = |B|$ atoms. The only bonds considered in A are (a_i,a_{i+1}) and (a_n,a_1); similarly for B, the only bonds considered are (b_j,b_{j+1}) and (b_n,b_1). We denote these bonds by $\rho(A)$ and $\rho(B)$ and by $\rho((v,w),A,B)$ the mapping of bond (v,w) of A to a bond of B. We have $r_{AB} = 1$ if and only if atom a_1 is mapped to atom b_1, atom a_2 is mapped to atom b_2, and so on. The set of rings in the reactant compounds are denoted R_r, and the set of rings in the product compounds are denoted R_p. The set of similar rings of A is denoted $S(A)$.

Using the *r* variables reduces the number of binary variables *m*; reducing the number of binary variables *m*, which in turn, typically reduces the time to solve the corresponding MILP model. Different approaches exist for reducing the number of binary variables *m*, but we use the following.

All the *m* variables involved in the ring mappings are declared continuous on [0,1], but all *r* variables remain binary. The objective function still uses the corresponding *e* variables, but the *r* variables are not used in the objective function. The mapping is fully described by the *m* variables. Note that this approach reduces the number of binary variables but not the total number of variables.

In this approach, the ring variables are constrained among themselves. The main constraints are that every ring must be mapped exactly once

$$\forall\ A \in R_r(\sum_{B\in S(A)} r_{AB} = 1) \tag{8}$$

The following constraints tighten the model by requiring every product ring to receive exactly only one reactant ring

$$\forall\ B \in R_p(\sum_{A\in S(B)} r_{AB} = 1) \tag{9}$$

Finally, we need to relate the ring variables to the *e* variables. This is done by observing that a bond in a reactant ring must be mapped to exactly one bond in a product ring. The following constraints express that condition

$$\forall\ A \in R_r(\ \forall\ (v,\ w) \in \rho(A)(\sum_{\substack{B\in S(A)\\(x,y)=\rho((v,w),A,B)}} e_{vwxy} = r_{AB})) \tag{10}$$

Notice that constraints 4 relate the *m* and *e* variables, so that constraints 10 implicitly relate the *r* and *m* variables.

Naturally, a reaction that performs a cyclization will not be considered by the RMT because no one-to-one mapping of the rings is possible because one more ring exists in the product(s). For example, Figure 3 shows a reaction that performs a cyclization of a sugar. RMT is not applied to this reaction: no one-to-one mapping of the ring structure from product to reactant can be done. The atom mapping computed for this reaction by our method is based on the basic formulation without RMT.

Table 3 shows the speedup of using the RMT on the MetaCyc reactions. The speedup is at least 3 but can be an order of magnitude larger on reactions that have a large number of rings and on which the solver took a long time to solve by using the basic formulation without RMT. In particular, there are many cases that took more than 10 min to solve using the basic formulation, but took less than 20 s with the RMT.

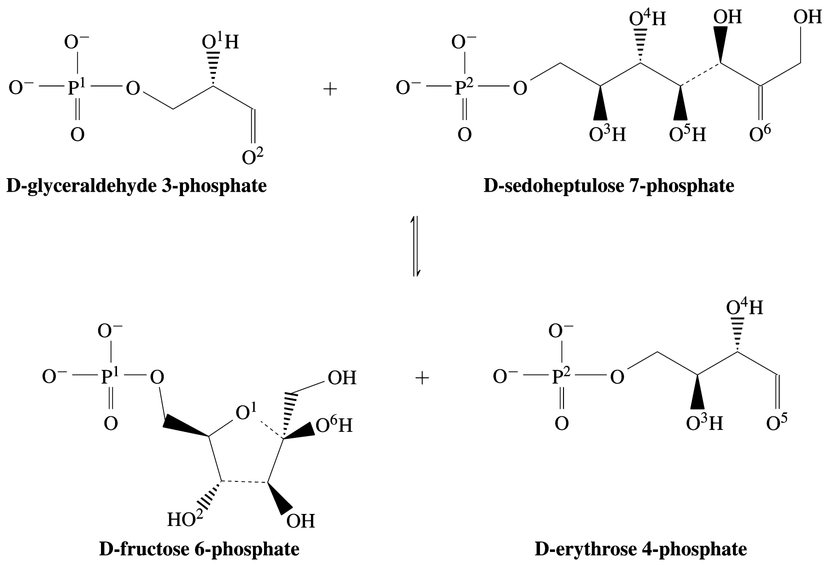


Figure 3. Transaldolase catalyzes the transfer of dihydroxyacetone from a ketose to an aldose. In this case, the dashed bond of D-sedoheptulose-7-phosphate is broken to give a dihydroxyacetone equivalent which adds to D-glyceraldehyde 3-phosphate, producing D-fructose 6-phosphate in its cyclic, furanose form. In this transaldol reaction, a cyclic product is produced from acyclic reactants; therefore, the RMT is not applied to this reaction.

Table 3. Statistics on Time To Solve Models and Number of Binary Variables with and without Ring Variables in Models When Applied to MetaCyc[a]

number of rings	average time with ring variables (sec)	average time no ring variables (sec)	speedup	number of binary variables with ring variables	number of binary variables with no ring variables	ratio of increase in number of binary variables	number of cases in MetaCyc
1	0.46	1.54	**3.34**	284	427	**1.50**	1350
2	2.63	8.85	**3.36**	518	865	**1.66**	884
3	0.91	2.99	**3.28**	495	977	**1.97**	1332
4	1.37	5.76	**4.20**	564	1242	**2.20**	1004
5	0.80	10.12	**12.64**	1588	3605	**2.27**	405
6	1.59	7.32	**4.60**	1219	3022	**2.46**	380
7	2.95	27.72	**9.39**	1079	2791	**2.58**	226
8	17.68	758.74	**42.90**	2803	6222	**2.21**	175
9	85.44	1222.42	**14.30**	4526	9816	**2.16**	183
10	2.53	308.03	**121.75**	1165	3915	**3.36**	120
11	2.65	69.42	**26.19**	1203	3855	**3.20**	25

[a] Each row gives, over five randomly selected models, the average time to solve the models, the average speedup, the average number of binary variables in the models with and without ring variables, the increase factor in the number of binary variables, and the number of cases (i.e., reactions) having this number of rings. We stop at 11 rings because there are less than five reactions in MetaCyc with compounds with more than 11 rings.

When the Ring Assumption Is False. When the ring assumption is false for a reaction, two cases might arise when solving its model: (1) The model is infeasible because the rings cannot be mapped to satisfy the other constraints. (2) The rings can be mapped and there are feasible solutions, but the optimal solution found is less optimal than breaking and reforming at least one ring.

The second case is quite rare, and we currently have no reasonable example of it. We did not encounter it when computing the atom mappings on MetaCyc (see Application to MetaCyc section) and in comparing its atom mappings with the KEGG RPAIR database.

In the first case—when the MILP model is not feasible—the solver reports it, and a new basic model, without using the RMT, can be generated and solved. It is simpler to proceed in that way instead of trying to detect all possible infeasible scenarios due to the *r* variables while generating the model.

For example, Figure 4 presents a reaction that is peculiar: two similar rings are present in the reactant and product compounds, that is, each ring can be mapped independently from reactant to product, where the right reactant ring containing N needs to be rotated to match the similar product ring . But the reaction breaks the bicycle, with the two cyclicly shared carbon atoms of the reactant becoming nonshared in one product ring. Figure 5 shows another example where each individual ring on one side of the reaction is similar to one ring on the other side of the reaction. The reaction does not separate the rings, but a rearrangement occurs on them. In that case, it appears as if the RMT can be applied, and indeed, such a technique is tried. But the MILP solver detects that the mapping of rings is not feasible. In these cases, the models with ring variables are infeasible, and the basic models, that is, without trying to map rings, are instead created and solved to find the optimal atom mapping.

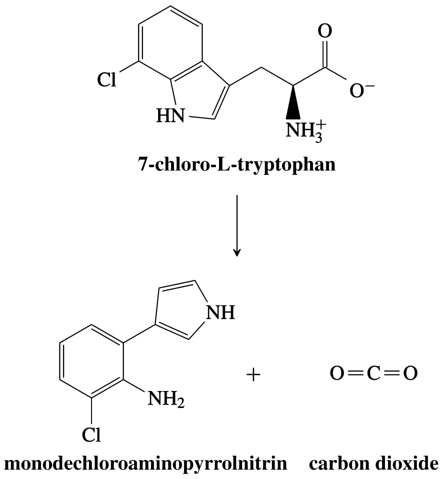


Figure 4. Reaction monodechloroaminopyrrolnitrin synthase (no official EC number); i.e., rxn-11795 of MetaCyc, with peculiar ring structures.

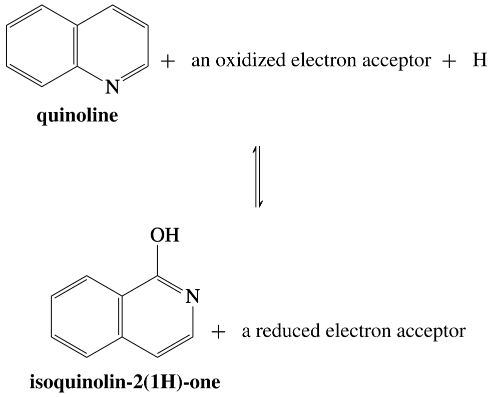

+ an oxidized electron acceptor + H₂O
N
quinoline
OH
N
+ a reduced electron acceptor
isoquinolin-2(1H)-one

Figure 5. Reaction quinoline 2-oxidoreductase (EC 1.3.99.17) shows a peculiar case of ring structure rearrangement: each reactant ring of the bicycle can be mapped to exactly one product ring, but the bicycle rings, as a whole, cannot be matched.

When applying the RMT to 10,262 reactions of MetaCyc (see Computational Modeling section), we encountered 48 infeasible cases. To be solved, they were reformulated using the basic modeling without the RMT.

Stereochemistry. We handle the modeling of stereochemistry, insofar as atom mapping is concerned, in a simple way: mapping bond a to bond b involving changing its stereochemistry has the cost $T(a,b) - 1$. This cost can happen in three ways: an up bond mapped to a down bond, an up bond mapped to a bond with no stereochemistry, or a down-bond mapped to a bond with no stereochemistry. This special cost is specified in the gain of mapping bonds in eq 6. That is, although the two bonds have the same atom species and type, the cost of mapping a bond that would require a change in its stereochemistry is not zero but almost equals the cost of breaking that bond.

This simple general technique does not entirely model the cost of atom mappings in the presence of stereochemistry. To do so would require modeling all possible rotations and symmetries of compounds—and symmetries of segments of compounds that separate from compounds when bonds are broken—which is very complex to do.

Figure 13 shows an example of a reaction where our approach for stereochemistry fails. These cases are rare.

Our stereochemistry technique depends on the drawing of compounds using down and up bonds. On the basis of that fact, the proposed technique could fail in some cases, although it is possible to mitigate that issue by drawing the stereochemistry of bonds of compounds between reactants and products as similar as possible. However, application to MetaCyc (see Application to MetaCyc section), for which no modification was done to any drawing of compounds, of this simple general cost technique of stereochemistry shows that it is very effective in practice. This effectiveness is due to the MWED metric used that seeks to find the minimum cost atom mapping.

Special Bond Values. Table 1 gives the general bond values used to compute gains and costs in the objective function. There are some exceptions noted by asterisks in that table. This section lists those special bond values and justifies them on the basis of experimental evidence.

NTP Compounds. According to Table 1, the propensity of all P—O bonds in ATP, for example, would all be the same, leading to many alternative optimal mappings. However, experiments have demonstrated that the broken bonds from ATP to ADP or ATP to AMP are most often the same.[11] The candidate bonds likely to break are shown in Figure 6. We assign lower bond propensity values to the actual reactive bonds in nucleoside triphosphate compounds, so that breakage of those bonds is favored. However, there are exceptions: it has been experimentally determined[12] that for dGTP, the bond broken is at P_β. This has also been observed for dUTP, dCTP, and dTTP. For these cases, the enzyme changes, which the P—O bond is the most reactive. This example illustrates a difficultly in performing atom mapping on enzyme-catalyzed reactions: an enzyme may activate a reactant in such a way that the reaction diverges from the mechanism observed under different conditions.

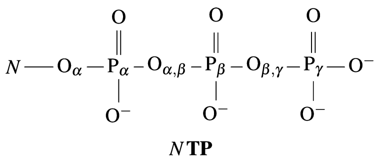


Figure 6. Special treatment for bond values for compounds with a triphosphate group (e.g., ATP). In most cases, the bonds P_α—$O_{\alpha,\beta}$ and $O_{\beta,\gamma}$—P_γ have a higher propensity to break compared with other P—O bonds. These two bonds receive propensity values 7 times lower than the general values assigned to the other bonds. Exceptions to this rule are applied to compounds dGTP, dCTP, dTTP, and dUTP: only the bond P_β—$O_{\alpha,\beta}$ is assigned a value 7 times lower than all other P—O bonds for these compounds.

Compound Phosphoenolpyruvate (PEP). It has been shown experimentally[13−15] that the synthase reaction, as shown in Figure 7, favors breaking the bond between C and O in phospho*enol*pyruvate (PEP) so that the oxygen of water is not incorporated in phosphate. A lower value than the value for bond P—O is assigned to the C—O bond in PEP. This assignment applies to PEP regardless of the reaction.

Bond Patterns x—C=O, x—C—O, and x—CH—O. For some cases, we consider neighboring atoms around a reaction site (functional group) to reduce the number of solutions to a single, optimal atom mapping. Here, we take advantage of the understood and predictable reactivity of functional groups such as the ester (O—C=O bond pattern) and acetal (O—C—O

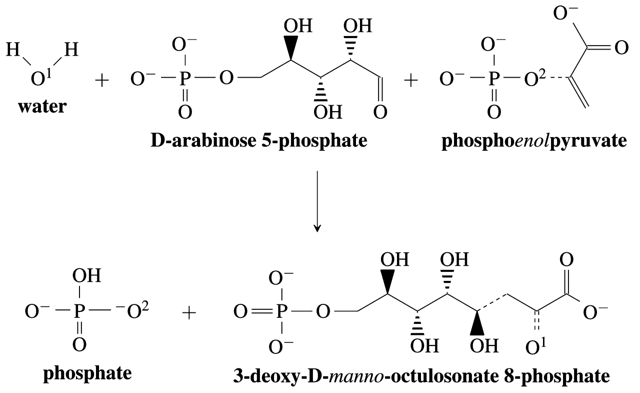


Figure 7. Reaction 3-deoxy-8-phosphooctulonate synthase (DAHP, EC 2.5.1.55). The dashed lines show the bonds broken and made. Experiments show that oxygen O^1 of water is transferred to 3-deoxy-D- *manno*-octulosonate 8-phosphate, whereas the fate of oxygen O^2 of PEP is inorganic phosphate.[13] Special treatment of the C−O and P−O bond values in PEP is necessary to achieve this solution.

bond pattern). We extend this reactivity to the related bond patterns *x*−C=O, *x*−C−O, and *x*−CH−O, where *x* is a N, C, or S atom. For these bond patterns, we assign a lower value than $T(x,\mathrm{C})$. For the pattern *x*−C=O, the bond value of *x*−C is $T_1(x,\mathrm{C}) - 4$, it is $T_1(x,\mathrm{C}) - 2$ for the pattern *x*−C−O, and it is $[T_1(x,\mathrm{C})/2]$ for pattern *x*−CH−O.

The effect of this special bond value assignment is illustrated by the hydrolysis of 1-O-alkyl-2-acetyl-sn-glycerol (Figure 8).

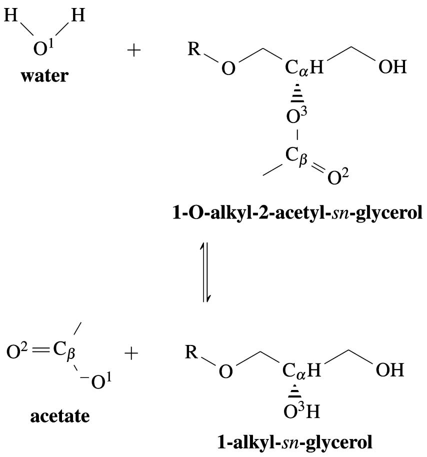


Figure 8. Reaction acetylalkylglycerol acetylhydrolase (EC 3.1.1.71). The cleavage of the C_β−O^3 bond is the correct mechanism to form acetate instead of breaking the C_α−O^3 bond. The C_β−O^3 bond is more reactive because of the presence of the C=O double bond. When the cost of the C_β−O^3 bond compared to that for the general bond value for C−O is lowered, the correct atom mapping is the single solution.

The correct atom-mapping results from water attacking C_β, breaking the C_β−O^3 bond. An alternative incorrect mapping would result from water attacking C_α, breaking the C_α−O^3 bond. Without the special treatment of the ester group (i.e., O−C=O bond pattern), both of these atom mappings would be returned as solutions. Reducing the bond value of the O−C=O bond discriminates between these two solutions, and only the former is returned.

Considering All Optimal Solutions. In general, a model can have more than one optimal solution. That is, more than one atom mapping might exist that satisfies the model and has the same minimal objective value. It is essential to find all optimal solutions because stopping at the first optimal solution found and declaring it to be the correct solution would be incorrect. It is also important to distinguish the case of multiple solutions because our MILP formulation cannot properly identify a single correct actual mapping when multiple actual atom mappings occur for a given reaction.

These atom mappings might differ by (1) the bonds made and broken, or (2) by the symmetry of the compounds, the presence of multiple identical molecules (i.e., stoichiometry above 1 for at least one molecule), or the indistinguishable atoms involved in the bonds made and broken. The presence of several optimal atom mappings found is not incorrect because if they differ because of point 2, they are essentially the same atom mapping.

We want to address the following questions regarding multiple atom mappings:

1. How do we find all optimal solutions; that is, all optimal atom mappings, using a LP solver?
2. If several optimal solutions exist, how can we determine the equivalent ones due to the symmetry of the compounds or the indistinguishability of some of the atoms and molecules?
3. Are all optimal solutions chemically valid?
4. If the last question cannot be fully answered, are the chemically valid solutions a subset of the optimal solutions?

We address the first two questions in the following two sections (Computing All Optimal Solutions, Determining Equivalent Atom Mappings) and address the last two questions in the Application to MetaCyc section.

Computing All Optimal Solutions. Typically, a MILP solver gives one optimal solution, in our case, one atom mapping, although some solvers (e.g., CPLEX) offer the ability to find all *optimal* solutions. Some other solvers (e.g., SCIP) find only all *feasible* solutions. For such solvers, we present a simple technique to find all optimal atom mappings. We proceed in two sequential steps to find all atom mappings:

1. Generate a MILP problem to compute one optimal solution. Let *v* be the optimal value found.
2. Generate another MILP problem with the same constraints plus the constraint $f = v$, where f is the objective function created in step 1. There is no objective function to optimize because the model verifies only the feasibility of the constraints. Use the solver to find all feasible solutions: they are all optimal solutions with value *v*.

Determining Equivalent Atom Mappings. Once all optimal atom mappings are found, for one reaction, we can partition them in equivalence classes so that two atom mappings are in the same class if they have the same bonds broken and if indistinguishable atoms and molecules are taken into account.

For example, in Figure 9, the reaction has four identical thioacrolein compounds as reactant. If we assume that these compounds are distinguishable, any atom mapping can be transformed into another one by interchanging any two thioacroleins. But assuming that these compounds are indistinguishable, the number of different atom mappings is reduced to one (see also Figure 10).

Our atom mapper removes such equivalent atom mappings from the set of solutions it returns. This algorithm to detect the

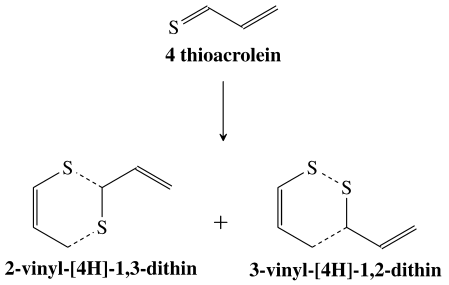


Figure 9. Spontaneous reaction of MetaCyc (RXN-8902, no EC number because there is no enzyme to catalyze this reaction). The four reactants thioacrolein can be composed in six ways to give the two products generating six atom mappings. But assuming that the four thioacroleins are undistinguishable, only one atom mapping exists. The bonds made are represented as dashed lines.

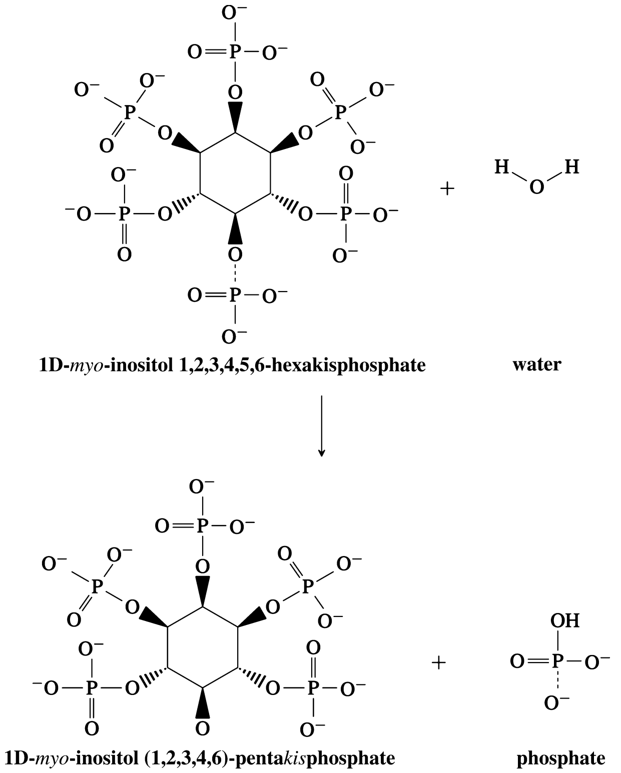

1D-*myo*-inositol 1,2,3,4,5,6-hexakisphosphate
water
+
1D-*myo*-inositol (1,2,3,4,6)-penta*kis*phosphate
phosphate
+

Figure 10. Compounds involved in reaction 5-phytase (EC 3.1.3.72) have many indistinguishable O^- atoms attached to Ps. For this reaction, 384 atom mappings were found, that is, 384 optimal solutions were found by the linear solver, but all of them are equivalent. That is, only one essential atom mapping exists, the one indicated by the bonds made and broken using dashed lines. If stereochemistry had not been taken into account, many more equivalent atom mappings would have been found because the cycle of six carbon atoms could have been mapped in many different ways by rotation and symmetries contributing as a multiplicative factor to the number of optimal solutions. This would have resulted in thousands of equivalent atom mappings, although still only one would have been considered essential.

equivalent atom mappings is applied after the MILP solver has found all atom mappings. The algorithm computes the set of bonds broken and made for each atom mapping, and if two atom mappings do not differ on the basis of them, they are considered equivalent. In most cases, this computation reduces the set of atom mappings to one. If not, it is still possible that some atom mappings are equivalent because some bonds might have broken or formed between indistinguishable atoms. We defined two atoms as indistinguishable if (1) they are leaf atoms of the same species, same bond type, and same charge attached to the same atom (e.g., the O atoms in O=C=O) and (2) they are equivalent atoms on the basis of the symmetry of the molecule. Two bonds, from two atom mappings, are equivalent when the atoms involved are the same or pairwise indistinguishable. Two atom mappings are equivalent if they share the same set of bonds broken and made, on the basis of equality or equivalence.

APPLICATION TO METACYC

We have applied the computational approach of the Computational Modeling section to the MetaCyc database,[8] version 16.0. MetaCyc is a multiorganism literature-based metabolic database containing, for version 16.0, a total of 10,262 biochemical reactions. The atom mappings were not computed on 2526 MetaCyc reactions either because the reaction is a generic reaction (i.e., one or more of its substrates are compound classes), because MetaCyc lacks a chemical structure for at least one substrate, or (for a small number of cases) because the reaction was not balanced. For 54 reactions, the number of optimal atom mappings was greater than 1000, and 181 atom mappings were not necessarily computed to optimality because the linear solver exceeded the time limit of 3 h to ensure optimality. In total, for 7501 reactions, all optimal solutions——that is, atom mappings——were computed. Two sections (Comparison to KEGG RPAIR and Comparison with DREAM) compare the computed MetaCyc atom mappings to the atom mappings with the manually curated KEGG RPAIR database and with the mappings derived from the approach of First et al.[7]

Computing All Optimal Solutions. When the computational approach of the Computational Modeling section is applied to MetaCyc, for most reactions several optimal solutions (several atom mappings) are obtained.

Table 4 shows the percentages of MetaCyc reactions according to their number of atom mappings found. The first row shows the percents without removing equivalent atom mappings, that is the real number of solutions returned by the solver, whereas the second row shows the percents once the atom mappings have been reduced to the equivalent classes described in the Computational Modeling section. The second row is the most important to consider because it represents the number of different reaction mechanisms found. Multiple atom mappings are undesirable because they are likely to contain at least one inexact atom mapping. In the case of MetaCyc,

Table 4. Percentages of 7501 MetaCyc Reactions Having One or Multiple Atom Mappings[a]

	number of atom mappings					
	1	2	3−4	5−8	9−64	>65
All	30%	25%	18%	10%	13%	4%
Class	98%	1.8%	19	3	1	0

[a] The first row gives the percentages according to all solutions directly found by the solver without considering equivalent solutions. The second row shows the percentages according to the number of equivalence classes as computed by the procedure described in the Computational Modeling section. For the second row, the last four integers are the numbers of reactions.

multiple optimal solutions were infrequently computed; otherwise, a manual curation is needed to remove the incorrect atom mappings.

Comparison to KEGG RPAIR. We compared the computed MetaCyc atom mappings with the manually curated KEGG RPAIR database, version 58.0 (June 2011). Among the 7501 reactions of MetaCyc for which we computed atom mappings, we were able to compare 2446 reaction atom mappings from the KEGG RPAIR database. The main reason for this partial comparison is that only 4018 MetaCyc reactions have a defined corresponding reaction in KEGG, and for some of those reactions, the compounds could not be completely matched.

The comparisons described in this and the following section consisted of a two-phase process in which we first compared the mappings computationally. When mappings were found not to match, a chemist in our group reviewed the mappings to determine if they truly differed (e.g., differences based on symmetry were ignored); if so, the mapping that was likely to be correct was determined on the basis of chemical expertise and available literature. KEGG RPAIR was usually found to have the correct mapping but not always.

Because some MetaCyc reactions had more than one optimal atom mappings, the comparison was based on whether or not, for each reaction, the single atom mapping defined in KEGG RPAIR was one of the optimal solutions found for MetaCyc. Among the 2446 compared reactions, 42 reactions (1.7%) had more than one atom mapping, 38 reactions had 2 atom mappings, and 4 reactions had three atom mappings.

Among the 2446 reactions compared, we found that 25 MetaCyc reactions did not have the KEGG atom mapping. For one case, experimental evidence indicated that the MetaCyc atom mapping was correct and the KEGG atom mapping was incorrect. Two other cases were similar, although no conclusive experiments have been reported. In the remaining 22 cases, the atom mappings for MetaCyc were incorrect, which is an error rate of 0.9%.

Figure 7 shows one example of an incorrect atom mapping found in the KEGG RPAIR database. The correct reaction mechanism displaces the oxygen of water in compound 3-deoxy-D-*manno*-octulosonate 8-phosphate, although the KEGG RPAIR database shows the oxygen displaced into Pi.

Figure 12 presents an example of incorrect atom mapping computed for MetaCyc, with the correct mechanism shown in Figure 11. The main difference is that the correct mechanism breaks two C−C bonds. But breaking such bonds entails a high cost, and a more optimal solution was found on the basis of the MWED, which turned out to be incorrect. The high cost of making and breaking a C−C bond is needed to get many other mappings correct, and lowering that cost to get this correct atom mapping would create many more incorrect atom mappings. Taking into account more neighboring atoms of C−C bonds is potentially what is needed for our approach to obtain the correct atom mapping for this reaction.

The comparison also revealed that properly chosen cost values for bonds broken and made on the basis of atom species rarely need precise modeling of stereochemistry, which is complex to model exactly. In fact, only two cases among the 22 atom mappings in error were due to incorrect handling of stereochemistry. Figure 13 shows one of the only two reactions that was not handled correctly by our technique to account for stereochemistry.

Comparison with DREAM. The recent work by First et al.[7] uses MILP to compute multiple atom mappings on the basis of a bond edit-distance metric. Their approach is similar to ours but differs in regard to the following points: (1) our objective function uses bond costs, based on the species of atoms (i.e., we use MWED), (2) our modeling is simpler because we do not model stereochemistry with constraints but only with bond weights, (3) our MILP models use the RMT to increase speed, whereas First et al.[7] uses several other general MILP techniques (e.g., tightening constraints), (4) we compare our atom mappings with KEGG RPAIR database, (5) we do not map hydrogen atoms, and (6) we apply our approach only to biochemical reactions.

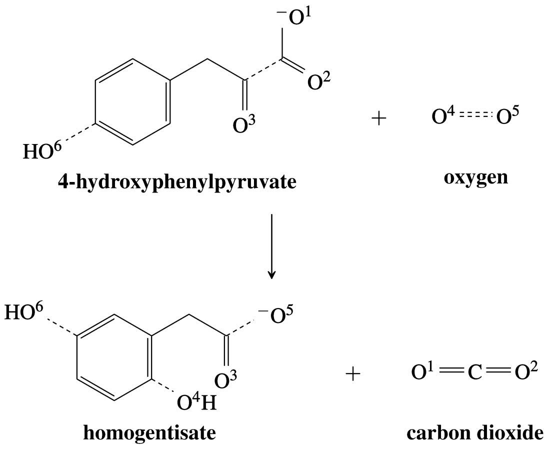


Figure 12. Reaction of 4-hydroxyphenylpyruvate dioxygenase (EC 1.13.11.27). The atom mapping computed for this reaction is incorrect. The optimal solution breaks and makes a bond to displace HO^6, which does not occur in the correct atom mapping. The cost of breaking or making one C−C bond is much higher than the cost of breaking or making one C−O bond. This optimal solution breaks only one C−C bond, which makes it less costly than the correct solution shown in Figure 11 that breaks two C−C bonds and makes one C−C bond.

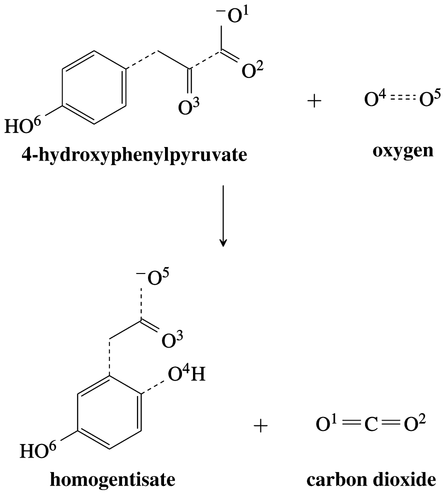


Figure 11. Reaction of 4-hydroxyphenylpyruvate dioxygenase (EC 1.13.11.27). The dashed lines show the bonds broken and made. Oxygen atoms are numbered to clearly show the details of the atom mapping. We believe that this is the correct atom mapping on the basis of chemical knowledge. Two C−C bonds are broken, and one C−C bond is made, which is a high cost atom mapping according to our MWED. Indeed, an incorrect atom mapping was computed for this reaction as shown in Figure 12.

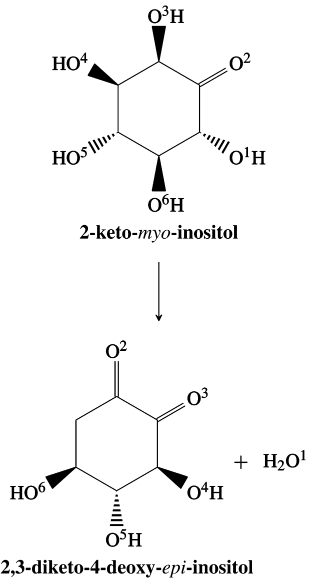


Figure 13. Atom mapping returned for *myo*-inosose-2-dehydratase (EC 4.2.1.44) is an example of an incorrect mapping due to stereochemistry. The mapping found cleaves the C−O^1 bond to form water rather than the correct C−O^3 bond. The incorrect solution rotates the compound around an axis at 30 degrees (between O^2/O^3 and O^5/O^6), inverting the stereochemistry at the C−O^4, C−O^5 and C−O^6 bonds. This is one of only two incorrect cases found among 7501 MetaCyc reactions for which atom mapping was computed. The other case (not shown) is the reaction of UDP-glucose-hexose-1-phosphate uridylyltransferase (EC 2.7.7.12).

The modeling of stereochemistry in First et al.[7] is precise but more complex than ours. They have not evaluated its effectiveness in finding the correct atom mappings.

Furthermore, our method computes only optimal solutions; it does not consider sub-optimal solutions. Weighting bonds according to Table 1 was implemented to avoid the necessity of considering sub-optimal solutions. Sub-optimal solutions increase the likelihood of obtaining atom mappings that are not chemically accurate. Our aim is to reduce the number of optimal solutions (ideally to a single solution), in contrast to First et al.[7], which includes sub-optimal solutions to ensure a chemically valid solution is contained within the set.

Directly comparing computational speed is difficult because First et al.[7] used a different MILP solver—the ILOG CPLEX solver (version 12.1) on a single threaded 2.66 GHz processor. We solved 10 of our own models using CPLEX 12.4 on a 4-core processor (CPLEX uses all the cores available). We observed a speed-up factor of 4 to 165, compared with SCIP 2.1.0, for these 10 cases when computing one optimal solution but found very little speedup when computing all optimal solutions. Because CPLEX version 12.4 differs from version 12.1, comparing the solving speeds is difficult.

The main difference between First et al.'s approach and ours is the metric used: we use a *weighted* edit-distance metric, whereas they use an unweighted edit-distance (number of bonds broken and made). Indeed, they reported that taking into account bond order (e.g., single bond becoming double bond) changed the atom-mapping solutions of four reactions among a sample of 71 for the KEGG database.

Eric First kindly provided us with the atom mappings of 6188 reactions computed according to the method described in.[7] That approach seeks to compute all optimal solutions and in some cases to compute suboptimal solutions near the optimal ones. The approach is based on the assumption that manual curation extracts the correct solution. For the purposes of this comparison, we were provided a single atom mapping per reaction—a fact that should be kept in mind in the following analysis.

We were able to compare 1709 of DREAM atom mappings against the KEGG RPAIR database. The other reactions could not be compared for the reasons cited in the Comparison to KEGG RPAIR section for the comparison between MetaCyc and KEGG RPAIR. We were able to identify 249 atom mappings among 1709 (14%) that were different from the KEGG RPAIR atom mappings and were manually verified to be chemically incorrect. Most cases involved oxygen atoms that were not correctly mapped but also cases where C−C bonds were broken where no breakage should have occurred. An example is shown in Figure 14. This result is not surprising because the metric used—the minimum number of bonds broken and made—is four in this case, which is the same as the exact atom mapping. If all atom mappings were computed, the exact atom mapping would be found among them. Our approach, using a MWED, typically reduces the number of optimal solutions and quite often to only one; in a large number of cases, as demonstrated with MetaCyc, that solution is the correct one.

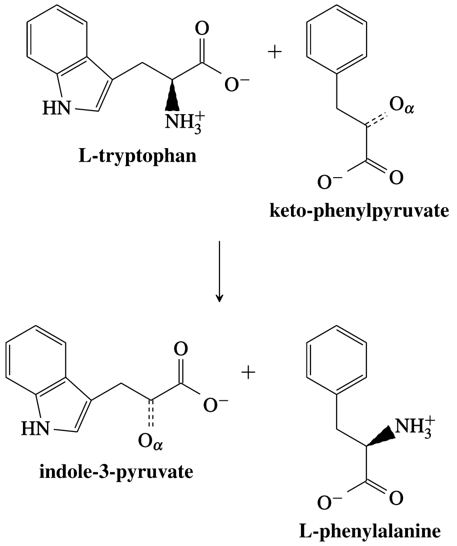


Figure 14. Reaction of tryptophan-phenylpyruvate transaminase (EC 2.6.1.28). The exact reaction mechanism involves the exchange of the O_α of keto-phenylpyruvate with the NH_3^+ group of L-tryptophan. This mechanism involves the breakage of two bonds and the formation of two bonds. One of the DREAM solutions involves the breakage of two C−C bonds and the formation of two C−C bonds; that is an incorrect mechanism but has the same optimal value on the basis of the number of bonds broken and made (i.e., the minimum bond edit-distance). Our computational approach uses a cost on breaking C−O bonds much lower than any other C−C bonds, avoiding this incorrect atom mapping and resulting in only one atom mapping, the correct one.

Solving Speed. We have used the single threaded SCIP solver (version 2.1.0) on a 2.66 GHz, 10GB Intel Linux workstation to solve all the MILP models for MetaCyc. The time used to solve the MetaCyc reaction atom mapping models varied from less than one tenth of a second to 3 h. The solver could not solve 181 models to optimality in 3 h or less of CPU time. The time variation is mostly due to the variation in the numbers of atoms and bonds involved in the reactions and the number of optimal solutions.

Table 5 presents the percentage of solved atom-mapping models under some time limits. Almost all models are solved in less than 30 min and a large percentage in less than 1 min. For the 181 reactions, it would have taken more than 3 h to compute all optimal solutions, most likely because of the number of compound symmetries and indistinguishable atoms. For all MetaCyc reactions for which all optimal atom mappings were computed, the average time to find the first optimal mapping was 26.6 s, and the average time to find all optimal atom mappings was 72.3 s.

Table 5. Percentages of Atom-Mapping Models That Could Be Solved under a Time Limit (seconds) Using the MILP Solver SCIP 2.1.0[a]

	solved under a time limit (sec)				
	< 0.1s	< 1s	< 10s	< 60s	< 1800s
1	50%	71%	89%	94%	98%
n	46%	70%	85%	91%	97%

[a] A total of 7682 reactions of MetaCyc were involved, which includes 7501 reactions with optimal solution(s) and 181 reactions with no optimal solution found after 3 h. The first row is for finding the first optimal atom mapping, and the second row is for finding all optimal atom mappings (i.e., all optimal solutions).

Note that these timing results are greatly influenced by the solver used, and faster solvers could solve the models in shorter time. We also solved 10 of our models using the ILOG/CPLEX solver, version 12.4, on a 2.66 GHz 4-core processor (CPLEX uses all cores when solving a model). The 10 models were randomly selected among the models that took at least 1 s to be solved when searching for one optimal solution using the SCIP solver. The ILOG/CPLEX solver ranged from 4 to 165 times faster than SCIP when searching for one optimal solution. However, the speed of CPLEX was not substantially faster when searching for all optimal solutions.

RELATED WORK

The KEGG RPAIR database[1] is the largest biochemical reaction atom-mapping database known to us. The atom mappings were manually curated, helped by a computational tool[16] to find common subgraphs. The level of accuracy of that database is unknown to us. Also, the database does not state which mappings are supported by experimental data, if any.

Ravikirthi et al.[3] have computed atom mappings for all reactions of the iAF1260 model using a MCS (Maximum Common Subgraph) technique supplemented by manual curation. The computational technique used is expensive, taking an average of 25 h of CPU time per reaction. Several hundred processors were used in parallel to complete the computation in a reasonable amount of time. In contrast, our computational approach takes 26.6 s, on average, to find the first optimal solution for 7501 MetaCyc reactions, and no manual curation is required.

Körner and Apostolakis,[5] using a wMCES computational technique, computed the atom mappings for 5630 reactions of the KEGG database. The wMCES is similar to MCS, but C-X bonds, where X can be C, N, O, or S, have a weight different than one. Stereochemistry is not handled by their approach. Multiple atom mappings were computed for 14.9% of the reactions for the KEGG database. This is in contrast to our method where multiple atom mappings were computed for 2.1% of 7501 reactions of MetaCyc. They do not report an error rate for KEGG, but a companion paper, Apostolakis et al.,[6] gives the following results for the computed atom mappings applied to the BioPath database: (1) 13.5% of the reactions had multiple atom mappings (7.6% with two atom mappings, and 5.9% with more than two atom mappings) and (2) 1.6% of the reactions did not have the same atom mapping as the one in BioPath, with a true error rate of 0.9% once the discrepancies were analyzed manually. The 13.5% of reactions with multiple atom mappings make their result less accurate than ours because these contain, for each of these reactions, at least one atom mapping that does not match the one in the BioPath database, which occurs, for our approach, for 2.1% of 7501 reactions in MetaCyc.

Arita[4] has computed the atom mappings for 3478 reactions, recording only the mappings of carbon, nitrogen, and sulfur atoms. The computation uses a MCS computational technique and a manual preparation of every reaction to help the computation find the correct mapping. The manual preparation is applied to all reactions and is potentially a reordering of the reactants so that the first reactant should be matched with the first product and so on. About 2% of the computed atom mappings were found to be incorrect, considering only carbon, nitrogen, and sulfur atoms, and a manual postprocessing was undertaken to correct them. No timing of the computational approach was reported in the paper. Only one atom mapping is computed for each reaction, although the symmetries of compounds are taken into account when computing atom tracing in pathways. These atom mappings were applied to *E. coli*.[17]

The MCS problem has been addressed by many researchers (see Ehrlich et al.[18] for a review). McGregor[2] devised one of the first algorithms to conduct MCS in the context of atom mapping.

CONCLUSION AND FUTURE WORK

We have developed accurate and efficient MILP modeling to compute atom mappings for biochemical reactions. The method does not map hydrogen atoms, but it does consider hydrogens when computing the optimal atom mappings.

Our modeling is based on a MWED metric that accounts for atom species between bonds and in some cases (e.g., C−C and N−C bonds) the neighboring atoms for these bonds. We have also shown techniques, namely, directly mapping ring structures (i.e., the RMT) and mapping specific compounds pairs (e.g., nadp to nadph) to reduce the size of the models, typically also reducing the time to solve them.

We tested this approach on 7501 reactions of MetaCyc and found that this computational approach is faster by 3 orders of magnitude when compared with the MCS approach reported by Ravikirthi;[3] at the same time, the accuracy of the algorithm is very close to the accuracy of the manually curated KEGG RPAIR database.

Our computational approach can give more than one atom mapping when indistinguishable atoms/molecules and symmetries are considered. However, multiple mappings should be minimized because they represent ambiguity. Through careful tuning of our bond weights, we were able to substantially reduce the occurrence of multiple mappings. Multiple atom mapping solutions occurs about 2.1% of the time when applied to 7501 reactions of MetaCyc, and 1.8% had only two atom mappings. Manual curation is required to remove the potential incorrect multiple atom mappings that still remain.

We compared the atom mappings of 2446 reactions of the KEGG RPAIR database and our MetaCyc atom mappings and found that 25 reactions did not have the KEGG atom mapping. We discovered that three atom mappings in KEGG RPAIR were incorrect, whereas 22 atom mappings were computationally in error in our approach, for an error rate of 0.9%. The comparison also revealed that properly chosen cost values for bonds broken and made on the basis of atom species rarely need precise modeling of stereochemistry, which is complex to model.

We showed that the MWED approach is more accurate than the simpler unweighted edit-distance metric by comparing 1709 atom mappings as computed by the technique presented in First et al.[7] and the KEGG RPAIR database. We were able to identify 249 atom mappings, among 1709 (14%) that were different from the KEGG RPAIR atom mappings and were manually verified to be chemically incorrect. Most cases involved oxygen atoms that were not correctly mapped, but cases also included broken C−C bonds where none should have occurred.

Our computational approach presented does not provide a complete curation-free approach but to the best of our knowledge is the most accurate that has been published to date. We have also shown that a curation-free approach cannot be achieved unless enzymatic activity is also taken into account.

In future work, we intend to model atom mappings of generic reactions of MetaCyc, compute bond propensity values more precisely on the basis of positive and negative examples of atom mappings, that is, by using a machine learning approach, model more precisely stereochemistry, and refine the modeling to compute only essential atom mappings, that is, by reducing the number of equivalent atom mappings directly computed by the solver.

AUTHOR INFORMATION

Corresponding Author

*E-mail: latendre@ai.sri.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by the Office of Science (BER), U.S. Department of Energy under Award Number DE-SC0004878. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

(1) Yamada, T.; Hattori, M.; Oh, M. A.; Goto, S.; Kanehisa, M. RPAIR: A Database of Chemical Transformation Patterns in Enzymatic Reactions. Genome Informatics 2005, International Conference, Poster and Software Demonstrations, Yokohama, Pacifico, Japan, December 19−21, 2005; http://www.jsbi.org/journal1/giw05poster/ (accessed September 1, 2012).

(2) McGregor, J. J.; Willett, P. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Model.* **1981**, *21*, 137−140.

(3) Ravikirthi, P.; Suthers, P. F.; Maranas, C. D. Construction of an *E. coli* genome-scale atom mapping model for MFA calculations. *Biotechnol. Bioeng.* **2011**, *108*, 1372−1382.

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

(5) Körner, R.; Apostolakis, J. Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *J. Chem. Inf. Model.* **2008**, *48*, 1181−1189.

(6) Apostolakis, J.; Sacher, O.; Körner, R.; Gasteiger, J. Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. *J. Chem. Inf. Model.* **2008**, *48*, 1190−1198.

(7) First, E. L.; Gounaris, C. E.; Floudas, C. A. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *J. Chem. Inf. Model.* **2012**, *52*, 84−92.

(8) Caspi, R.; et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **2012**, *40*, D742−753.

(9) Achterberg, T. SCIP: Solving constraint integer programs. *Math. Program. Comput.* **2009**, *1*, 1−41.

(10) IBM ILOG CPLEX, version 12.4. http://www.ilog.com/products/cplex/ (accessed September 1, 2012).

(11) Cohn, M. Mechanisms of enzymic cleavage of some organic phosphates. *J. Cell. Comp. Physiol.* **1959**, *54*, 17−31.

(12) Weber, D. J.; Bhatnagar, S. K.; Bullions, L. C.; Bessman, M. J.; Mildvan, A. S. NMR and isotopic exchange studies of the site of bond cleavage in the MutT reaction. *J. Biol. Chem.* **1992**, *267*, 16939−16942.

(13) Furdui, C.; Zhou, L.; Woodard, R. W.; Anderson, K. S. Insights into the mechanism of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase (Phe) from *Escherichia coli* using a transient kinetic analysis. *J. Biol. Chem.* **2004**, *279*, 45618−45625.

(14) Hedstrom, L.; Abeles, R. 3-Deoxy-d-manno-octulosonate-8-phosphate synthase catalyzes the C−O bond cleavage of phosphoenolpyruvate. *Biochem. Biophys. Res. Commun.* **1988**, *157*, 816−820.

(15) DeLeo, A. B.; Sprinson, D. B. Mechanism of 3-deoxy-d-arabino-heptulosonate 7-phosphate (DAHP) synthetase. *Biochem. Biophys. Res. Commun.* **1968**, *32*, 873−877.

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

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

(18) Ehrlich, H.-C.; Rarey, M. Maximum common subgraph isomorphism algorithms and their applications in molecular science: A review. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 68−79.