# Automatic Determination of Reaction Mappings and Reaction Center Information. 1. The Imaginary Transition State Energy Approach

Robert Körner and Joannis Apostolakis*

Institute for Informatics, Ludwig-Maximilians-Universität München, Amalienstrasse 17, Munich D-80333, Germany

Received November 22, 2007

Chemical reactions transform the reactant molecules by deleting existing and forming new bonds. The identification of these so-called reacting bonds is important for studying the reaction mechanism and for applications in metabolomics, e.g. for interpreting substrate labeling experiments. Here, we introduce an approach which suggests the simplest possible reaction center at the heavy atom level, with high accuracy. In contrast to current methods the approach is motivated by a simple theoretical model based on a crude approximation of the reaction energetics, and takes the complete reacting system into account. Finally, it recovers all optimal solutions to the problem while removing all symmetry-related, redundant solutions. We apply the method on the complete KEGG database of biochemical reactions, and compare our approach with previous methods. The resulting reaction centers are represented as imaginary transition states, which are molecule-like representations of reaction mechanisms. We provide the statistics of the calculations on the KEGG database and discuss some examples for the different types of alternative solutions found.

## 1. INTRODUCTION

Enzymes catalyze biochemical reactions that transform the substrates into products by deleting existing and forming new bonds. In order to study the reaction mechanism of enzymatically catalyzed reactions it is necessary to identify the reaction center, i.e. the reacting atoms and bonds. This then automatically determines the fate of each atom during the reaction. While single reactions can be studied in high detail with specific experiments and time intensive computational methods, a theoretically motivated high throughput approach for the prediction of the reaction center is missing. This would be useful for comparison of reaction mechanisms at the genomic level and for providing initial models for more exact experiments and calculations.

Furthermore, enzymes regulate the metabolic flow in the cell, by catalyzing metabolic reactions. By measuring the reaction rates of the different enzymatic reactions in the cell it should be possible to understand and simulate the metabolic flow and how it depends on the environment and the state of the cell.[1,2] The different reaction rates can be deduced from measurement of the concentration change of isotopically labeled metabolites. In order to be able to relate the measured concentration changes to single reaction rates it is necessary to know the fate of labeled atoms in a reaction, i.e. where they end up in the products. The complete mapping for each atom of the reactants to an atom of the products of a given reaction is called the atomic reaction mapping (ARM).

Atomic reaction mappings are not uniquely defined by the reaction itself and depend on the exact reaction mechanism. While two different reaction mechanisms may lead to the same reaction mapping, the opposite is relatively unusual. Enzymatic reactions in particular are specific with respect to the reaction mapping they induce.

Neglecting the possibility of trivial reactions (where the same bond is broken and formed in the same reaction) and not considering symmetry related ARMs (different ARMs that can be transformed into each other by symmetry operations on the reactants or the products), there exists a one-to-one relation between reaction mapping and the graph operation that describes the reaction. The latter is defined as an operation that deletes certain edges and forms others between existing vertices, in such a way that the reactant graph is transformed to the product graph. The graph operation corresponding to a given chemical reaction can be represented by a so-called imaginary transition state[3] (ITS), a molecule-like representation of a reaction. This work is concerned with the computational prediction of the ARM. The reason that we will keep both ITS and ARMs in the presentation in spite of their equivalence is that ITS is a more mechanistic representation, relevant for the reaction itself, whereas ARMs are directly related to atomic tracing in biochemical networks.

Already in 1980 it has been suggested that most chemical reactions follow the principle of minimal chemical distance[4] (PMCD), i.e. they follow the shortest path (e.g., the graph operation with the fewest bond deletions and creations) for transforming the reactant graph to the product graph. From the beginning this principle has been formulated as an empirical finding. A number of related methods have been suggested in the literature for obtaining the atomic reaction mapping based on graph similarity. Current methods generally perform atom based matchings between substrate−product pairs (e.g., Arita,[5] Hattori et al.[6]/Kotera et al.[7]) on the basis of which they then identify the complete mapping.

Based on an empirical similarity method developed by Hattori et al.,[6] Kotera et al.[7] have devised an algorithm which requires manual identification of substrate product pairs (called reactant pairs). These are defined as "pairs of

* Corresponding author e-mail: apostola@bio.ifi.lmu.de.

10.1021/ci7004324 CCC: $40.75 
Published on Web 06/06/2008

compounds that have atoms or atom groups in common on two sides of a reaction".[7] On the reactant pairs an empirical, atom-based matching is performed to identify the reacting parts of the molecules. In order to take chemistry correctly into account they introduce 68 different atom types that describe the hybridization and the chemical environment of the atoms in the molecules. The mappings obtained were used to build the first RPAIR database in the KEGG; however, the procedure has now been superseded by a semiautomated approach. The exact procedure for building RPAIRs is now summarized as follows ". . . the KEGG RPAIR (reactant pair) database, which is a curated database of substrate-product pairs with atomic alignments indicating correspondences of atoms before and after the reaction. This database is generated first by manually decomposing each reaction formula in the KEGG REACTION database into a set of biologically meaningful reactant pairs, and then by atomic alignments with the SIMCOMP[6] program followed by manual curation." (http://www.genome.jp/kegg/document/help_bget_rpair.html, accessed Oct 24, 2007).

The principle of identifying the reaction mappings on reactant pairs is not optimal, as it does not take the complete reacting system into account. Methods that are based on the identification of reactant pairs, in general, have problems with a number of reactions as mentioned by Arita.[5] Further, it is difficult to formulate a chemically intuitive principle motivating the choice of pair based approaches. On the other hand, such methods gain efficiency because they only handle relatively small graphs in the graph comparison step. Akutsu[8] has suggested two related polynomial algorithms for specific types of reactions. As such, they are quite fast; however, they lack in generality.

We introduce here a simple framework from which an optimization problem is rigorously derived, whose solutions are the different hypotheses for the atomic reaction mapping and the corresponding reaction center. We give a thorough description of the theory and a methodological comparison to existing methods with a few examples that aim at demonstrating the suggested approach. We show how almost all of the problematic cases mentioned by others are treated correctly in the new approach. Further, in contrast to current methods and manual annotations, the approach suggested here finds all optimal solutions, which often represent chemically reasonable alternatives, to what is found in manually annotated reaction databases. A few examples are described to show the relevance of both the question of alternative mappings and the relevance of ITS as a first structural model for enzymatically catalyzed reactions. Interestingly enough even though the framework is based on three assumptions which do not generally hold for enzymatic reactions, comparison with a curated database is presented in the second manuscript which demonstrates the high accuracy of the results.

## 2. METHODS

**A Simple Framework for Reaction Mapping Prediction.** We suggest a general and intuitive theoretical framework which contains previous graph based approaches as a special case. It is based on the following three assumptions:

1. The preferred reaction mechanism converts the reactants to the products and has the lowest activation energy (low temperature assumption).

2. The reaction proceeds with a single transition state (single transition state assumption).

3. The activation energy for the transition state is given as the sum of the stabilities (activation energies) of each reacting bond (additivity assumption).

While assumptions 1 and 2 are physically legitimate, in the sense that they are formally correct for specific types of reactions (single transition state reactions at low temperature), the additivity assumption leads to a very crude approximation of the true activation energy. In order to avoid confusion with the true transition state energy, we will denote the quantity arising from the additivity assumption as the imaginary transition state energy (ITSE), since its minimization leads to the identification of the atomic reaction mapping and therefore the ITS for the reaction. The ITSE can be easily computed as the sum of the weights of reacting bonds. We explicitly point out the fact that the single transition state assumption (assumption number 2) is not generally true for enzymatic reactions, however. this assumption is necessary for obtaining a well defined optimization problem. When the assumptions do not hold, the approach is approximate. The quality of the approximation is assessed empirically by the comparisons shown in the second manuscript.

By making a further simplification, namely that all bonds have the same stability (weight), the minimization of the ITSE recovers the original PMCD: the correct ITS according to the principle of minimal chemical distance minimizes the number of reacting bonds. The advantage of graph based approaches such as the PMCD is that they are defined over a large though enumerable configuration space. Therefore, since the corresponding problem on graphs is NP hard, it is at least possible to use systematic algorithms that sample the complete space to find the globally optimal solution to the problem. Furthermore, the use of branch and bound algorithms can achieve significant improvements in efficiency by rigorously ruling out large parts of the search space during the search.

The general (weighted) form of the ITSE approach as stated here can be solved optimally with a variety of the branch and bound algorithm for weighted matching of graphs. The introduced weights are important, because they take into account the stability of the bonds and additionally allow direct consideration of bond order changes. In existing methods, higher order bonds are usually treated as single bonds (bond order change is not penalized). By giving a lower weight to a match of a double to a single bond than to a match of two double bonds the difference in weight corresponds to the penalty in changing the bond order.

Finally while it is possible to treat bonds to hydrogens in the same way as bonds to other atoms, we have chosen here to present results based only on the heavy atom graphs and to obtain hydrogen bond change as change in number of bonds to heavy weight atoms. This is equivalent to giving a very low weight to the hydrogens. Including the hydrogens in the calculation increases the size of the molecular graphs by almost a factor of 2, which again has an even more detrimental effect on computational efficiency.

**Reduction of the Optimization Problem to Weighted Graph Matching.** The optimization problem resulting from the additivity assumption can be reduced to a weighted maximum common subgraph problem. The ITSE of a particular reaction mechanism is the negative sum of the stability of all reacting bonds, i.e. all bonds that are either broken or formed. By assigning a weight to each bond in the reactants and the products corresponding to the bond stability, the solution to the ITSE approach is the minimal weight graph operation for the reaction. The weight of a graph operation is the sum of the weights of the deleted and formed edges. The identification of the minimum weight graph operation can be reduced to the solution of the weighted maximum common edge induced subgraph problem. To show this we note that we allow only operations that have three types of edges: broken edges, retained edges, and formed edges. The weight of the transformation $w_T$ is equal to the sum of the weights of the broken bonds $w_B$ and the weights of the formed bonds $w_F$: $w_T = w_B + w_F$. Further, any bond in the reactant graph is either conserved or broken in the transformation. Therefore, the weight of the bonds in the reactant graph $w_R = w_B + w_C$, and, correspondingly, for the weight of the product graph we have $w_P = w_{CF} + w_C$. From this we obtain $w_T = w_R + w_P - 2*w_C$. Since the first two terms are independent of the graph operation, it is clear that the minimum weight operation conserves the maximum weight subset of consistent edges. Finding the latter corresponds to the weighted maximum common edge induced subgraph problem (wMCES).

An efficient branch and bound algorithm (called RASCAL) for solving the unweighted MCES problem has recently been suggested.[9] In this work we extend RASCAL to the weighted MCES and use it to solve the optimization problem ensuing from the ITSE model. The resulting reaction mechanisms are given in a molecule-like representation of the reaction, the ITS.[3] The ITS representation of the reaction has a direct mechanistic interpretation as a first approximation for the transition state of the reaction: It corresponds to the simplest possible associative transition state, i.e. one in which all new bonds are formed before (or simultaneously with) the breaking of bonds. The required simplicity indicates that no bonds are formed during the reaction which are not present in either the substrate or the product and that no bonds present both in the substrate and the product are broken and reformed.

**Weighted Maximum Common Edge Subgraph Identification.** As already mentioned, we are looking for weighted maximum common edge induced subgraphs (wMCES) between the reactants and products. The wMCES directly identifies the bonds that are conserved in the reaction. All unmatched bonds on the reactant side and the product side are deleted or formed by the reaction, respectively.

The MCES problem is similar to the better known maximum common subgraph problem (MCS). The MCS problem is the identification of the largest possible consistent injective mapping *f()* from vertices in a graph *G* to vertices in a second graph *G′*. A mapping is consistent when for any pair of vertices in G with projections to G′ the pair is connected if and only if the corresponding image of the pair is connected in G′. In MCES the edges are mapped as opposed to vertices in MCS. In practice, this is achieved by converting the original graphs to line graphs and then looking for the MCS between the line graphs.[9] Line graphs have a vertex corresponding to each edge in the original graph. Two line graph vertices are connected, if and only if the corresponding edges in the original graph are incident to the same vertex. A mapping of vertices between two line graphs corresponds to a mapping of bonds in the original graph. In the line graph comparison we do not use the information on atom type hybridization or bond type, so that for example single bonds can be mapped on double bonds. Thus $\pi$-bond formation or deletion needs to be taken care of via the weighting scheme.

```
Algorithm MCS (G, P, depth, LB)
LB = max (depth, LB)
if |P| == 0 or depth + |P| <= LB
      return LB
repartition (P)
for m in getSmallestPartition (P){
     P' = neighbours (m) ⋂ P
     LB = max (LB, MCS (G, P', depth + 1, LB))
}
return LB
```

**Figure 1.** Simplified algorithm for identification of (all) maximum common subgraphs. G is the association graph, P is the set of independent partitions, depth is the recursion depth, LB is the lower bound used for bounding the search. Repartition (P) resorts the partitions in a greedy way in order to reduce the size of the smallest partition P.

The most efficient algorithms for solving the MCS problem first construct an association graph and then identify cliques in that graph. Cliques in the association graph correspond to isomorphic subgraphs between the original graphs. The clique finding problem can be solved for example with the Bron Kerbosch algorithm.[10]

The algorithm we use here for determining the MCS is based on the same principles and corresponds to a simplification of previous work by Raymond et al.[9] Their algorithm called RASCAL is a branch and bound approach for finding maximum cliques in the association graph. RASCAL uses a number of different strategies for efficiently limiting the size of the search space, while guaranteeing optimality. However, in our experience, not all of those strategies actually achieve significant enhancement in efficiency and were therefore removed from the implementation. The methods which were removed were the Kikusts pruning, the determination of the chromatic number of the graph as an upper bound for the clique size, and the equivalence class pruning.[11] For a closer discussion of these methods the reader is referred to the original work.[9]

The main idea of the RASCAL algorithm (see also Figure 1) is to first partition the nodes of the association graph into independent sets (partitions) $P_{1..k}$. Each partition $P_i$ can maximally contribute one vertex to a clique, since any two vertices from an independent set are not connected. The number of partitions thus provides an upper bound for the size of the maximum clique. The algorithm then starts recursively building a clique: It first picks the smallest active partition $P_i$, and for each vertex $m$ in that partition checks whether it can be reassigned to another partition (i.e., if there is another partition, unconnected to $m$). If such a partition is found the vertex is reassigned to it. If all nodes of a partition could be reassigned, the current number of active partitions can be decreased by one, thus decreasing the upper bound

```
Algorithm WMCS (G, P, CCW, LB)
LB = max (CCW, LB)
if |P| == 0 or (CCW + sumWeights (P)) <= LB
      return LB
repartition (P)
for m in getSmallestPartition (P){
     P' = neighbours (m) ∩ P
     LB = max (LB, WMCS (G, P', CCW + weight (m), LB))
}
return LB
```

**Figure 2.** Simplified algorithm for identification of (all) maximum weight common subgraphs. CCW is the current clique weight, and sumWeights (P) returns the sum of the maximal weights in all partitions in P.

for the current branch. If, after repartitioning, the partition is not empty the remaining nodes are used sequentially as candidates for the next member of the clique. Once a node has been selected the recursion goes one level deeper, meaning that the active partitions for the next level are formed. Active partitions are formed by the conjunction of the partitions of the current level with the neighbor list of the last selected node. Thus at each level the partitions contain only vertices that are consistent with the currently selected clique. Empty partitions formed by the conjunction again reduce the upper bound.

As soon as the lower bound, given by the size of the largest clique found so far, is not smaller than (or larger than in the case that we are looking for all maximum common subgraphs) the upper bound, which is given by the sum of the current depth and the number of still available partitions, the algorithm backtracks.

We have extended the RASCAL algorithm to a weighted MCS algorithm shown in Figure 2, with the main differences to the original algorithm being: (1) The upper bound is the sum of the current cliques weight (CCW) with the sum of the maximum node weight in each partition. (2) Repartitioning is performed such that the upper bound does not increase, i.e. each vertex is only tested for repartitioning into partitions whose maximal weight is higher than the vertex weight. While weighted maximum common subgraph identification is in general a significantly more complex optimization problem, in this particular case the algorithm solves it in comparable time as the MCS problem. The reason lies in the choice of the weights: Matches of bonds to nonbonds are assigned a weight of zero, and therefore the number of vertices in the association graph does not increase. In the matching algorithm two line graph nodes are matched if the atom elements forming the bond are identical. The bond multiplicity only affects the weight of the match. The weight further depends on the element types forming the bond, and we use a weight of $W_{CC} = 1.5$ for CC $\sigma$-bonds, $W_{CX} = 0.48$ for $C-N_{amine}$, $C-O_{ester}$, and $C-S_{thioester}$ bonds, and 1 for all other bonds. Furthermore, the weight of the bond is increased by $W_{\pi} = 0.02$ for each additionally mapped $\pi$-bond. The choice of weights is qualitative and has been set to model the high stability of single CC bonds, the relative instability of amide/ester like bonds, and the high volatility of additional $\pi$-bonds. By setting the weights such that the sum of a C−O bond (1.0), a $\pi$-bond (0.02) and a $C-N_{amine}$ bond (0.48) exactly equal the weight of a single CC bond leads to retention of relevant alternatives in transamination reactions.

Any weighting parameters that satisfy the following conditions will lead to the same results with respect to the reaction mappings:

- $W > W_{CC} > 2*W$ (The cc bond is more stable than other bonds but costs less than breaking two other normal bonds.)
- $0 < W\pi \ll W$, $W_{CC}$, $W_{CX}$ ($\pi$-bonds are practically negligible compared to all other bonds.)
- $W + W_{CX} + W_{\pi} = W_{CC}$ (transamination condition).

With the chosen weights the mapping of a carbonyl (C═O) bond on a second (C═O) bond has a weight of 1.02, while the mapping of a C═C bond on a C−C has a weight of 1.5. In the latter case the C═C bond could have obtained a maximum value of 1.52 (by mapping it to another C═C bond), which means that the cost of converting a double bond to a single bond is 0.02. Thus, the weights directly correspond to the cost of not matching the bonds that induce the weight in the first place. Bonds that are not matched are reacting bonds, i.e. either deleted or formed during the reaction.

While this weighting scheme need not be optimal, it is shown to lead to very good results on the BioPath database.

**Atomic Mapping.** After the MCES has been identified, the corresponding atomic mapping needs to be determined. The mapping of the bonds implied by the MCES does not necessarily define the complete atomic mapping. For example when a single atom is transferred from one moiety to another, it does not retain any of its bonds and is thus not included in the MCES. Further, single mapped bonds between atoms of identical type (element) yield two possible mappings for the corresponding atoms. To resolve these problems we map all remaining atoms according to atomic labels (element name), using the MCS algorithm again: From the original graphs we remove all bonds that were not mapped in the MCES. The remaining graphs are isomorphic, under the condition that the original reaction is balanced (i.e., there is the same number of atoms of every element type on each side of the reaction), since all inconsistent edges have been removed. Thus there is at least one bijective atomic mapping from the reactants to the products. Bonds that were removed on the reactant side are deleted bonds, while those removed on the product side are created bonds, and they are labeled accordingly.

The case for unbalanced reactions is more complicated. In an unbalanced equation some atoms appearing on the left side do not appear on the right side of the reaction and/or vice versa. The minimal description of the reaction results by following the described procedure for balanced reactions, whereby all atoms that cannot be mapped in the second step (atomic mapping) are simply removed.

Alternatively the atoms that appear only on one side of the equation can be added to the other side, such that the reaction becomes balanced. According to the procedure for detecting new bonds, all bonds between atoms that existed on only one side will appear as reacting bonds. To remain consistent with the minimality principle for the number of reacting bonds we prefer to assume that the structure of the groups which originally appear on only one side of the reaction does not change. While this is clearly an assumption, it often allows the identification of the missing compound(s) in unbalanced reactions.

**Hydrogen Atom Mapping.** The mapping of hydrogen atoms has not been included in the MCS procedure for a number of reasons:

(1) Hydrogen atom exchange (as protons) with the environment is very likely and generally not explicitly included in the reaction itself.

(2) The protonation and tautomeric state of the reactants are not necessarily known and additionally depend on the reaction conditions.

(3) While nonreacting hydrogen atoms are trivially mapped to the heavy atom they are bound to, reacting hydrogen atoms have no reference bond on one side of the reaction and are thus difficult to map: All hydrogen atoms with a broken bond are exchangeable against each other, and the same is true for all hydrogen atoms with a formed bond. This would lead to a combinatorial number of possible reaction mappings.

(4) Including hydrogen atoms in the original mapping makes the procedure significantly slower as it increases the number of atoms by a factor of approximately 2.

Instead the number of reacting hydrogen atoms is calculated in a second step as the sum of the valence change of heavy atoms in the obtained mappings. When a number of different mappings are obtained, they are sorted in order of increasing number of reacting hydrogen atoms.

It is, however, important to note that the method itself can also be used to map hydrogen atoms, in principle. The significant higher computational time compared to the low possible improvement in results has lead us to choose to map them in a second step.

The reaction mapping relates every atom from the reactants to the corresponding atom in the products. In Figure 3 the reaction mapping for reaction R00026 (Figure 3a) is shown in two different formats (Figure 3b,c). In the first case the reaction mapping is implied by the equivalent position of the atoms (the relative positions do not change). Normally a complete mapping needs to be defined, which however soon becomes difficult to visualize. The reacting bonds are highlighted in red and blue for formed and broken bonds, respectively. In the second case, which corresponds to the KEGG format for the reaction mappings, the reaction is broken up into reaction pairs, and for each reaction pair the fate of the atoms is indicated through the use of the different colorings, which highlight only the reacting part of the molecules. An explicit mapping of the atoms is given for each reaction pair in the Kegg Chemical Function (KCF) structure files.

The reacting bonds are obtained in the following way: The reactant graph $G_R$ and the product graph $G_P$ are related by a bijective mapping of the vertices, and therefore we can write $G_R = (V,E)$, $G_P = (V,E')$. Then the ITS graph is $G_{ITS} = (V,E_{ITS})$, with $E_{ITS} = E \cup E'$, the formed bonds $E_F = E_{ITS} \backslash E$, and the deleted bonds $E_D = E_{ITS} \backslash E'$. Broken and formed bonds are called reacting bonds.

The wMCES produces a number of mappings, many of which are however related to each other by symmetry. The redundant (i.e., symmetry related) reaction mappings are removed from the results. The two alternative ITS for reaction R00026 shown in Figure 4 are not isomorphic and are therefore retained as true alternatives. The first alternative corresponds to the reaction mapping defined in the RPAIRs of the KEGG database. The second alternative, however, is chemically more interesting for two reasons. First, the C1 which is the target of the nucleophilic attack by the water is more polarized (two C−O bonds) than the C4 attacked in the first reaction (only one C−O bond). Second, the first

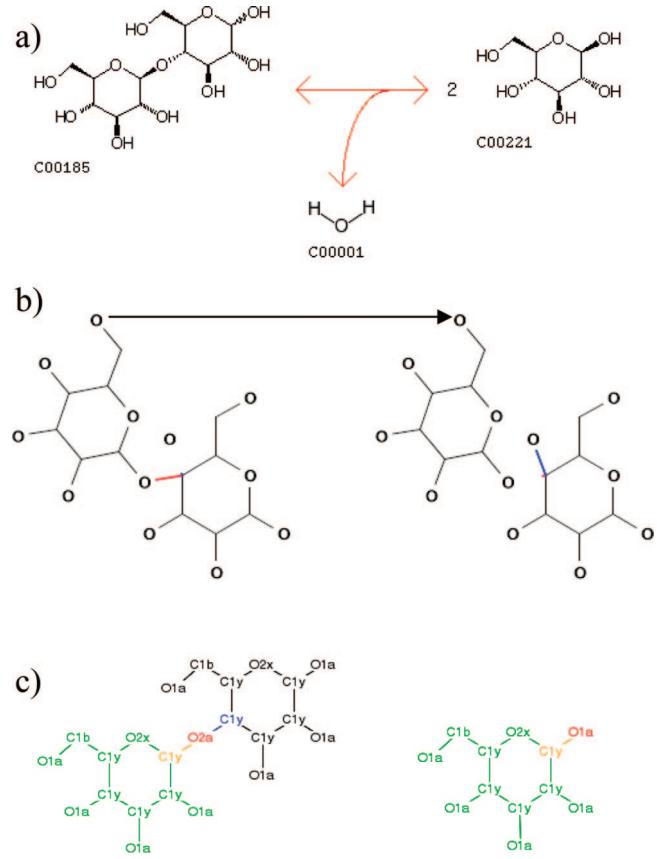


**Figure 3.** Representation of reaction mappings at the example of reaction R00026 from the KEGG database: Hydrolysis of cellobiose to two molecules of glucose. (a) Reaction as taken from the KEGG /LIGAND Web site.[12] (b) Reaction mapping between reactants and products. In the figure, mapped pairs of atoms in the reactants and the products have equivalent positions, related to each other by the translation vector shown for a single pair. The mapping shown here corresponds to a reaction mechanism where the water attacks the $C_4$ of the second ring, while the anomeric C1 retains the oxygen it was bound to in the cellobiose molecule. (c) Reaction pair representation for the same reaction, taken from KEGG/LIGAND database. Normally, for a single reaction, a number of reaction pairs are given, which show the fate of atoms in pairs of reactant- and product molecules. For this reaction only a single reaction pair is given in the database, and it shows the mapping of one glucose ring to the cellobiose molecule. This RPAIR found in the KEGG database corresponds to the same reaction center as the mapping in the middle panel.

alternative would normally lead to an inversion of the stereochemistry at C4, unless a particular mechanism to avoid this were in action (e.g., double inversion). In the second alternative the inversion would happen at the anomeric C which under physiological conditions racemizes to a mixture of α-D-glucose and β-D-glucose. Nevertheless both reaction mechanisms are in principle possible, and it is therefore important to take both into account as long as the reaction mechanism has not been experimentally elucidated for the particular enzyme.

## 3. RESULTS AND DISCUSSION

**3.1. Robustness, Accuracy, and Comparison to Previous Approaches and Databases.** As a first test of the efficiency and robustness of the method, the reaction mapping has been applied to the complete KEGG database (~6700) reactions. The algorithm could not be applied in 927 cases (see also Table 1), due to missing compound structures. A

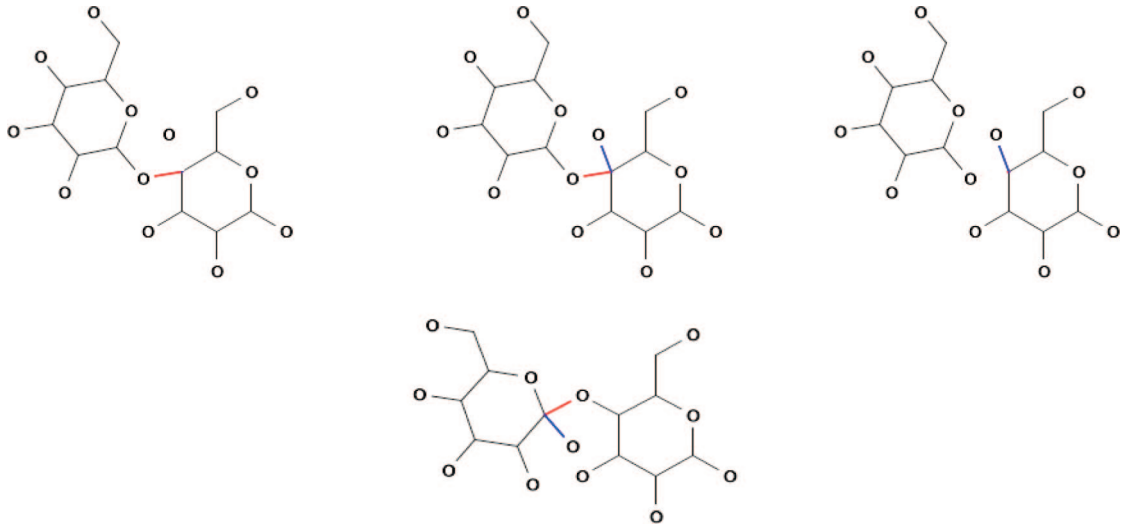

**Figure 4.** Generation of ITS for reaction R00026 from the KEGG database. Top, left: reactants in heavy atom graph notation. Top, right: products in heavy atom graph notation. Top middle: ITS as obtained by aligning reactants and products. Reacting bonds present only on the reactant or product side are colored red and blue accordingly. Bottom: alternative ITS for the same reaction.

**Table 1.** Statistics over the Results on the KEGG Database[a]

| **reactions** | **6670** | | |
|---|---|---|---|
| OK | 5007 | timeout | 105 |
| unbalanced | 623 | heap space | 8 |
| | | no compound | 927 |
| **solutions** | 5630 | **no solution** | 1040 |

| no. of alternatives per solution | reactions |
|---|---|
| 1 | 4793 |
| 2 | 629 |
| 3 | 81 |
| 4 | 39 |
| 5 | 12 |
| >5 | 76 |
| total | 5630 |

[a] The upper part of the table reports the number of reactions that were mapped without any problems and the number of reactions that were not stoichiometrically balanced as well as the reactions that could not be mapped due to timeout, memory space, or because the structures of the reactants (or products) were not available. In the lower part of the table the number of reactions with a single, two, or more alternatives are reported.

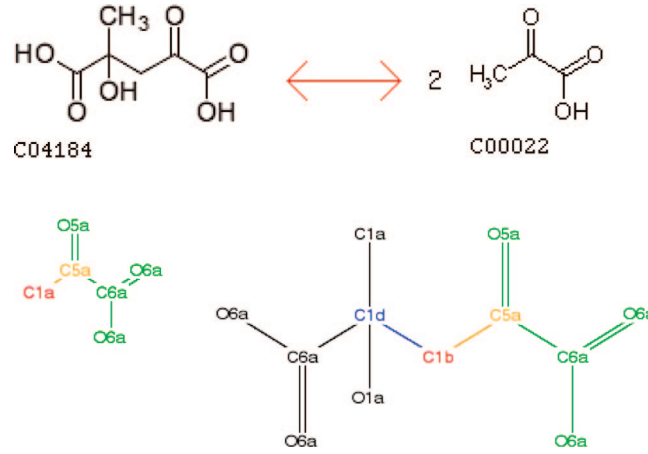


**Figure 5.** Top: 4-hydroxy-4-methyl-2-oxoglutarate pyruvate-lyase reaction (R00007). Bottom: the single reaction pair given for this reaction in the KEGG database.

number of compounds in the KEGG are only mentioned by name, with no chemical structure given. This is most notably the case for the so called glyco compounds, which are oligosaccharides, defined only at the level of monosaccharide building units and their connectivity. In further 105 cases the algorithm did not terminate after 6 h of computation, in contrast to the large majority of reactions (>4000) which are mapped in less than a second. These 105 reactions generally contain a large number of reacting atoms, some up to 350 atoms. Since the determination of the MCS is a NP hard problem, it is not surprising that computation time increases dramatically with the size of the problem. The large number of reactions that do show alternative solutions is due to the fact that in the different reactions, certain subreactions appear very often. For example in the KEGG more than 300 reactions involve hydrolysis of ATP to ADP + Pi. For this ATP hydrolysis the algorithm always finds two solutions. In one the water attacks the gamma phosphate, in the other the beta phosphate. Therefore the corresponding reactions all show two alternative solutions.

The reaction mappings in the KEGG database are given as a set of reaction pairs for each reaction. Each reaction pair contains the mapping between the two molecules in the reaction pair. Overall, this format is difficult to parse for both humans and machines. Furthermore, due to the way these mappings are produced, they are often incomplete. Reactions with stoichiometric coefficients larger than one are only implicitly coded. That means that only one reaction pair is given, even if the two or more molecules of the same type have different fate. For example reaction R00007 (Figure 5), the 4-hydroxy-4-methyl-2-oxoglutarate pyruvate-lyase reaction, is a retro aldol condensation which converts 4-hydroxy-4-methyl-2-oxoglutarate to two pyruvates. Only a single RPAIR (A01456) is given which describes the source of one of the pyruvates, while the rest of the mapping (not symmetrically related) is only to be understood implicitly. While this appears not to have been a problem for the Kotera et al.[7] study, it makes an automated comparison of our reaction mappings with the reaction pairs and their mappings given in the KEGG difficult.

**Table 2.** Problematic Cases according to Arita 2003[a]

| | reaction type | EC | reaction | *N* | correct |
|---|---|---|---|---|---|
| 1 | isomerization | 2.7.1.11 | R00767[b] | 1 | Y |
| | of sugars | 5.3.1.5 | R00307 | 2 | Y |
| | | 5.3.1.5 | R00877 | 2 | Y |
| | | 5.1.3.18 | R00889 | 1 | N |
| | | 2.4.1.8 | R06040 | 0 | G |
| 2 | cyclization/rearrangement | 5.4.99.7 | R03199 | 1 | PY |
| | | 4.2.1.75 | R03165 | 1 | Y |
| 3 | C-skeletal rearrangement | 5.4.99.2 | R00833 | 1 | Y |
| | | 5.4.99.1 | R00262 | 2 | Y |
| 4 | shift of chemical groups | 5.4.2.1 | R01518 | 1 | Y |
| | | 5.4.99.5 | R01715 | 1 | Y |
| 5 | transfer of small moiety | 2.8.3.1 | R01449 | 1 | Y |
| | | 2.8.3.1 | R05508 | 1 | Y |
| 6 | symmetric or dimeric structure | 1.1.1.25 | R02414 | 1 | Y |
| | | 1.7.3.3 | R02106 | 1 | Y |
| | | 1.11.1.6 | R02670 | 1 | Y |

[a] *N* is the number of alternatives found for the particular reaction. The last column refers to the correctness of the obtained solution: Y: one of the solutions corresponds to the correct mapping. N: none of the solutions corresponds to the correct mapping. G: compound structure not given. PY.: The mapping found by the algorithm is chemically reasonable; however, the reaction is too complex to assess its correctness by chemical intuition. [b] We have found a number of reactions (R00767 R00769 R00770 R01843 R01846 R01847 R01848 R03236 R03237 R03238 R03238) for this EC, which are all treated identically by the algorithm.

In order to test the correctness of the method we visually inspected the results obtained for a number of cases where other algorithms have been reported to fail: In the most comprehensive validation of a method Arita[5] has enumerated 6 different types of reactions which lead to wrong results with his empirical algorithm (see Table 2). Practically all the examples we found corresponding to the EC codes mentioned by Arita were treated correctly by our algorithm (except R06040, which fails due to missing compounds, i.e., the SMILES strings of the reacting compounds are not given in the KEGG database). However in the original paper of Arita an example of the first reaction type in Table 2 is given which concerns stereochemical isomerization of sugars. Arita does not give the correct EC codes for this type of reaction. Typical for this problem are epimerase reactions such as GDP mannose epimerase EC 5.1.3.18 (R00889). Those reactions are also not treated correctly by our algorithm, as we do not include stereochemical information in the mapping. Therefore, epimerase reactions appear as nonreactions in our results. Nevertheless the comparison shows that the suggested algorithm is significantly more robust and accurate than previous algorithms.

In the companion manuscript[14] we have compared the results obtained with the methods described in this manuscript with manually annotated reaction mappings, as they exist in the BioPath database;[15] the BioPath database is available at http://www.molecular-networks.com/databases/biopath.html. The results of that study show that of 1542 reactions in the database the algorithm could map all but 4 reactions. In 24 cases (~1.6%) the reaction mapping annotated in the database was not included among the solutions suggested by the algorithm. However, in 10 of those, literature has provided independent support for the result of the ITSE method.[14] Thus in less than 1% of the cases the algorithm did not find a correct mapping, according to the manual annotation of the BioPath database. To our knowledge this comparison is the first independent validation provided to date for the prediction of reaction mappings. The only other systematic validation of reaction mapping algorithms we are aware of has been performed by Arita[5] and was internal, i.e. he compared against his own intuition.

**3.2. Discussion of Single Examples.** The prediction of the reaction mappings often leads to different alternatives. Here we discuss single examples, which serve to highlight the mechanistic relevance of reaction mapping elucidation.

**Weighting of Bond Reactivities.** The example in Figure 6a−c exemplifies the effect of bond stability. Neglecting bond stabilities (i.e., assuming that all bonds are equally stable) leads to two possible ITS which minimize the number of reactive bonds for the ketose isomerization reaction (R00771). In the reaction center depicted by the left ITS (6b) a single C−O bond is broken, while in the ITS depicted on the right side (6c) a C−C bond is broken and formed. In nature this reaction can take place with a mechanism consistent with the first reaction center.[16] In general it is energetically significantly more difficult to break a C−C bond than a C−O bond. In cases such as this, taking bond reactivities into account is sufficient for discriminating between two reaction mappings and their corresponding mechanisms. Reactivities are modeled in the weighted MCES version of the algorithm by weighting the matches of different bonds according to bond stability. Indeed the solution shown in 6c) is found only in the unweighted MCES run. Using a slightly higher weight for C−C bond matches than C−O bond matches already solves the problem for the case shown here.

**Number of Reacting Hydrogen Atoms.** The arabinose phosphotransferase reaction (R01573; Figure 6d−f) involves opening of the sugar ring. The two variants shown in Figure 6e,f contain the same type of bond changes at the heavy atom level; however, they differ with respect to the position to which the phosphate is attached. Further, the reacting bonds form a connected subgraph of alternating broken and formed bonds in the right variant (Figure 6f). This leads to a lower number of reacting hydrogen atoms, which allows the identification of this variant as the more probable mechanism. The type of unconnected reaction seen in 6e is typical of reactions consisting of more than one step (complex reactions). While we do not take hydrogen atoms explicitly into account, the number of bonds to hydrogen atoms broken or formed is obtained in a postprocessing step from the sum of the change of heavy atom valences (i.e., the number of bonds every heavy atom makes to other heavy atoms). Therefore for this reaction the preferred solution suggested by the algorithm is reaction 6f.

**Chemically Reasonable Alternatives.** In many cases two or more of the suggested alternatives are equivalent with respect to the type of reacting bonds and atoms, so that no distinction can be made between them. The actual reaction (mechanism) then depends on the enzyme catalyzing the reaction. For example, the transformation of homocysteine to methionine (R00650; Figure 6g−i) is a methylation reaction, with the methyl being provided by S-methionyl adenosine. However, the ITSE approach suggests a second alternative, where the homocysteine substitutes the methionyl group in S-methionyl adenosine, as shown in Figure 6i. Chemically, both alternatives represent possible reactions, and in theory additional experiments are necessary to identify

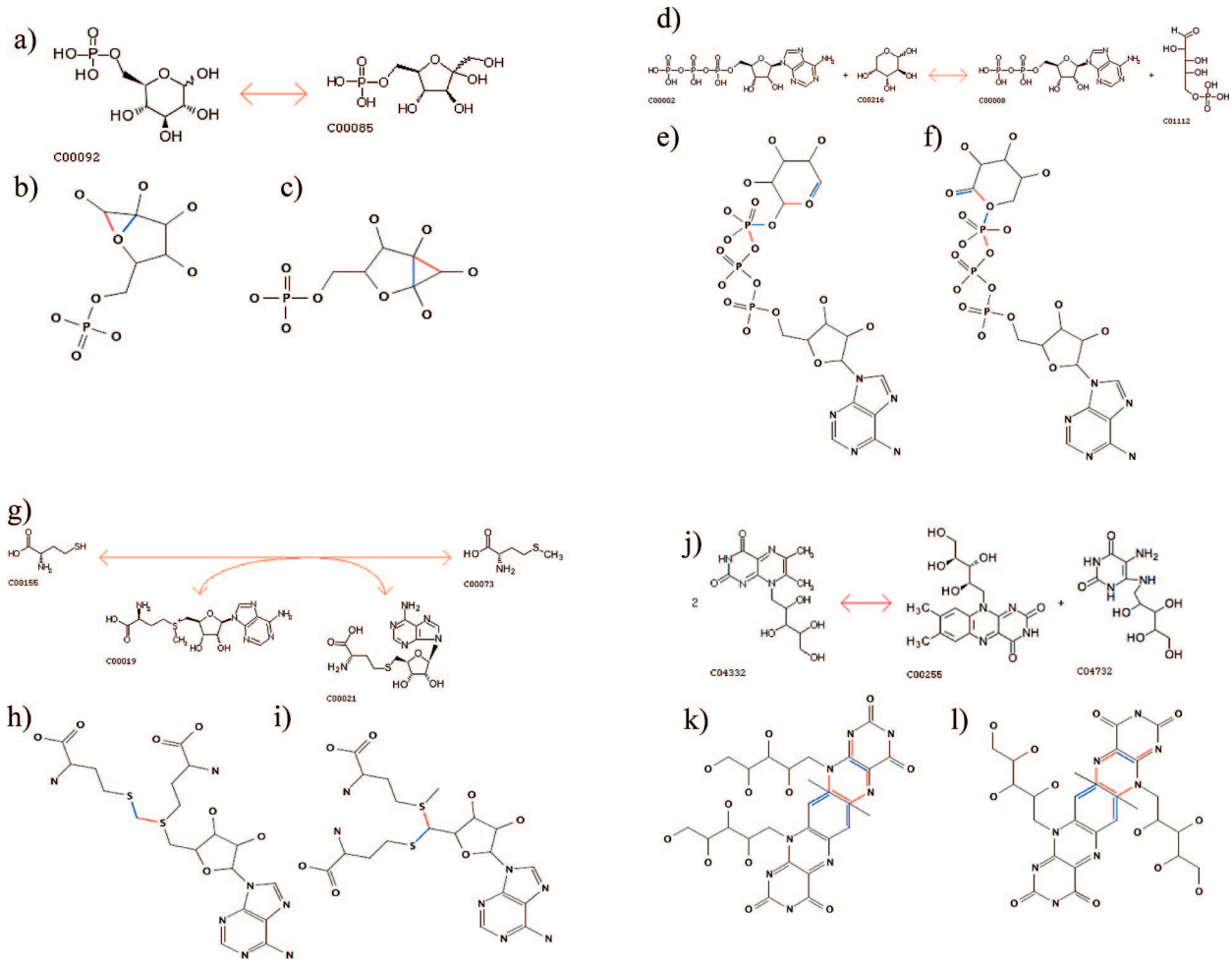


**Figure 6.** ITS representation of the alternative mechanisms for different reactions from the KEGG/LIGAND database:[12,13] (a) KEGG reaction formula for ketose isomerase (R00771), (b) ITS for retro aldol based mechanism, and (c) ITS for unnatural C−C bond rearrangement mechanism. (d) KEGG reaction formula for arabinose phosphotransferase (R01573), (e) mechanism with discontinuous reaction center, (f) continuous reaction center mechanism, (g) KEGG reaction formula for homocysteine S-methyl transferase (R00650), (h) methyl transfer, and (i) homocysteine/methionine substitution. Red (blue) bonds in the ITS representation are broken (formed) by the reaction (j) KEGG reaction formula for riboflavin synthase reaction (R00066), (k) cis ITS conformation, and (l) trans ITS conformation.

the correct mechanism for this reaction. The value of the algorithm is the identification of the second alternative as a chemically viable solution, something which is generally not taken into account in other algorithms or reaction centers given in curated databases.

One of the referees suggested that the alternative given for this reaction is not interesting, as the reaction has been studied with sufficient experiments. Subsequent additional search in the literature identified a recent paper[17] in which it is shown that SAM4 accepts also the (R,S) diastereomer of the S-adenosyl methionine as opposed to only the (S,S) diastereomer. The difference in stereochemistry is found at the sulfonium atom which is not planar. This difference is exactly within the two reaction center alternatives, and the different structure may lead to a change in mechanism. Our main point here is that when new substrates are still being identified, especially substrates with significant changes at the reaction center, we cannot assume that the reaction mechanism has been settled.

An example where the two possible alternatives have already been studied in the literature is shown in the next example. The riboflavin synthesis shown in Figure 6j−l shows two alternatives, the cis (Figure 6k) and the trans (Figure 6l) based transfer of a C4 moiety. Interestingly, the reaction does take place over a pentacyclic intermediate, closely related to the structure of the ITS for the reaction[18] (Figure 6k,l), indicating the structural relevance of ITS for deducing reaction intermediates or true transition states. While the cis conformation is in principle also possible, quantum chemical calculations at the B3LYP/6−31 g level of theory indicate that the trans conformation is energetically preferred. Furthermore, modeling the pentacyclic trans-intermediate shows a good fit into the binding site.[19] NMR data also support the trans intermediate.[17] The riboflavin synthesis reaction again exemplifies the relevance of atomic reaction mappings or ITS for mechanistic studies on the catalytic mechanism of enzymes. The reaction mappings obtained from the two alternatives differ and suggest a direct isotopic labeling experiment that could answer the question of the cis or the trans intermediate conclusively.

Concluding, we have to note that whenever reaction center alternatives are suggested for a given enzymatic reaction, a particular enzyme, and corresponding substrates, we assume that only one of those alternatives is correct. Therefore, by definition all others are false positives. The question is whether trained biochemists will think of all the possibilities themselves, and whether given these alternatives they will be able to identify the correct one without additional

experiments. This type of analysis is extremely subjective, and we have therefore left it to the reader to decide with the help of a few examples shown in Figure 6 and the more detailed classification of the types of alternatives found for the BioPath found in the second manuscript in the subsection "Alternative Reaction Mappings".

## 4. CONCLUSION

In this work we have presented a simple model of chemical reactions. Starting from the model we rigorously derived a discrete optimization problem. Previous work has used similar optimization principles for the prediction of atomic reaction mappings; however, the introduction of weights for describing bond stabilities leads to a qualititative improvement in the prediction of the reaction mappings. Based on the reaction mapping and reaction center information it is possible to narrow down the possible reaction mechanisms to those that are compatible with the corresponding reaction mapping. For the efficient solution of the optimization problem we have extended an existing branch and bound algorithm for finding common subgraphs of maximum weight. With respect to previous high throughput methods the suggested approach is:

• Based on the minimization of an energy-like quantity (the ITSE): This allows the use of bond stabilities to identify the most probable reaction center.

• Optimal as opposed to empirical e.g. Arita[5] and Kotera et al.:[7] When the algorithm terminates, it is guaranteed to have found all reaction mappings/mechanisms that are optimal with respect to the imaginary transition state energy. In contrast, the algorithm by Arita[5] depends on the order of mappings performed between single reactant/product pairs. In the algorithm by Kotera, the search for optimal mappings is simply cut off when it takes too long.

• Applied on a genomic scale, and at the complete heavy atom level: The most widely validated algorithm to date is the one by Arita.[5] However, Arita did not consider the mapping of oxygen atoms and only compared the results against his own chemical intuition. The application of our approach to the problematic cases in Arita[5] recovered the correct solution in all but two cases.

• Able to identify all optimal reaction mappings: Atomic mappings are assumed to be generally valid for a particular reaction, while the identification of different chemically reasonable alternatives for the same reaction suggests that they may well depend on the particular enzyme catalyzing the reaction.

The approach we describe here has been validated against the manual mappings found in the BioPath[15] database in cooperation with the curators of BioPath. An extensive analysis of these results is reported in the companion paper.[14] The relevance of the suggested theoretical framework for chemical biology lies in the fact that it provides a simple and highly efficient approach for the modeling of metabolic reactions.

## ACKNOWLEDGMENT

The authors wish to thank the initiators and developers of the KEGG database for providing it openly to the research community, Andreas Spitzmueller for the generation of the ITS representations, and Ralf Zimmer for general support.

## REFERENCES AND NOTES

(1) Kell, D. B. Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* **2004**, *7*, 296–307.
(2) Schuster, S.; Dandekar, T.; Fell, D. A. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **1999**, *17*, 53–60.
(3) Fujita, S. Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205–212.
(4) Jochum, C.; Gasteiger, J.; Ugi, I. The principle of minimal chemical distance (PMCD). *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495–505.
(5) Arita, M. In Silico Atomic Tracing by Substrate-Product Relationships in Escherichia coli Intermediary Metabolism. *Genome Res.* **2003**, *13*, 2455–2466.
(6) 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.
(7) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions. *J. Am. Chem. Soc.* **2004**, *126*, 16487–16498.
(8) Akutsu, T. Efficient Extraction of Mapping Rules of Atoms from Enzymatic Reaction Data. *J. Comput. Biol.* **2004**, *11*, 449–462.
(9) Raymond, J.; Gardiner, E.; Willett, P. RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631–644.
(10) Bron, C.; Kerbosch, J. Finding all cliques of an undirected graph. *Commun. ACM* **1973**, *16*, 575–577.
(11) Marialke, J.; Körner, R.; Tietze, S.; Apostolakis, J. Graph-Based Molecular Alignment (GMA). *J. Chem. Inf. Model.* **2007**, *47*, 591–601.
(12) Goto, S.; Okuno, Y.; Hattori, M.; Nishioka, T.; Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **2002**, *30*, 402–404.
(13) Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **2004**, *32*, D277–D280.
(14) Apostolakis, J.; Sacher, O.; Körner, R.; Gasteiger, J. Automatic Determination of Reaction Mappings and Reaction Center Information II. Validation on a Biochemical Reaction Database. *J. Chem. Inf. Model*. **2008**, *48*, 1190–1198.
(15) Reitz, M.; Sacher, O.; Tarkhov, A.; Trümbach, D.; Gasteiger, J. Enabling the exploration of biochemical pathways. *Org. Biomol. Chem.* **2004**, *2*, 3226–3237.
(16) Seemann, J. E.; Schulz, G. E. Structure and mechanism of L-fucose isomerase from Escherichia coli. *J. Mol. Biol.* **1997**, *273* (1), 256–68.
(17) Vinci, C. R.; Clarke, S. G. Recognition of Age-Damaged (R,S)-Adenosyl-L-methionine by Two Methyltransferases in the Yeast Saccharomyces cerevisiae. *J. Biol. Chem.* **2007**, *282*, 8604–8612.
(18) Illarionov, B.; Eisenreich, W.; Bacher, A. A pentacyclic reaction intermediate of riboflavin synthase. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 7224–7229.
(19) Zheng, Y.-J.; Jordan, D. B.; Liao, D.-I. Examination of a reaction intermediate in the active site of riboflavin synthase. *Bioorg. Chem.* **2003**, *4*, 278–287.

CI7004324