Automatic Determination of Reaction Mappings and Reaction Center Information. 2. Validation on a Biochemical Reaction Database

Joannis Apostolakis,*,† Oliver Sacher,‡ Robert Körner,† and Johann Gasteiger‡,§

Ludwig-Maximilians-Universität München, Institut für Informatik, Amalienstrasse 17,
80333 München, Germany, Molecular Networks GmbH, Henkestrasse 91, 91052 Erlangen, Germany, and
Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, 91052 Erlangen, Germany

Received November 22, 2007

The correct identification of the reacting bonds and atoms is a prerequisite for the analysis of the reaction mechanism. We have recently developed a method based on the Imaginary Transition State Energy Minimization approach for automatically determining the reaction center information and the atom−atom mapping numbers. We test here the accuracy of this ITSE approach by comparing the predictions of the method against more than 1500 manually annotated reactions from BioPath, a comprehensive database of biochemical reactions. The results show high agreement between manually annotated mappings and computational predictions (98.4%), with significant discrepancies in only 24 cases out of 1542 (1.6%). This result validates both the computational prediction and the database, at the same time, as the results of the former agree with expert knowledge and the latter appears largely self-consistent, and consistent with a simple principle. In 10 of the discrepant cases, simple chemical arguments or independent literature studies support the predicted reaction center. In five reaction instances the differences in the automatically and manually annotated mappings are described in detail. Finally, in approximately 200 cases the algorithm finds alternate reaction centers, which need to be studied on a case by case basis, as the exact choice of the alternative may depend on the enzyme catalyzing the reaction.

INTRODUCTION

The information on reaction centers and atom-atom mapping numbers is important for a number of reasons: They are required for the correct interpretation of atom tracing experiments, they can offer important information on the actual mechanism of a reaction, and they can therefore be used as a phenomenological representation of the reaction. Furthermore, by using the reaction center information it is possible to suggest transition state analogues, which can be used as enzyme−inhibitors.[1]

We have recently developed an Imaginary Transition State Energy Minimization approach, which identifies the reaction center with all reacting bonds in a single step for the complete reaction and derives the reaction mapping from that information.[2]

In this manuscript we validate the ITSE method by comparing its results with manually annotated reaction center information and reaction mappings from the BioPath database,[3] that is accessible online.[4] BioPath is the electronic version of the Biochemical Pathways wall charts[5] published by Boehringer Mannheim (now Roche) and edited by G. Michal. The corresponding atlas on Biochemical Pathways served as an additional source for the database.[6] The version of the BioPath database used (version 1.0 of 2006) in this study consisted of 1542 mapped reactions and 1175 molecules and contains the complete endogenous metabolism contained in the Boehringer Mannheim wall chart. One of the key features of BioPath is the manual encoding of reaction mapping information. This information consists of atom−atom mapping numbers (corresponding to reaction mappings) and marked reaction centers. The reaction mapping has been performed manually by students trained with standard coding rules for the input of organic and bioorganic reactions into a database. During the development of BioPath significant effort has been invested in checking the consistency of the reaction equations: e.g. the input of the reaction mapping information was verified by the program MN.CHECK.[7]

The comparison reported here is to our knowledge the most extensive validation of an atomic reaction mapping method reported up to now. The results strongly support the predictive value of the method and at the same time validate the consistency of the manual annotation performed in the generation of the BioPath database. By using the BioPath reaction mappings as a standard of truth we can further assess the improvement obtained by taking into account bond stabilities. It is shown that the principle of minimizing an energy-like quantity significantly improves prediction quality, compared to the simple bond matching approaches that only take graph structure into account.

We first report the overall statistics of the comparison and then address single examples of reactions that show inconsistencies or are otherwise of interest. For all examples where the ITSE algorithm has not found the reaction mapping denoted in the BioPath, we discuss the different reaction mappings from the point of view of the possible reaction mechanism. Furthermore, we have performed a literature

* Corresponding author e-mail: joannis.apostolakis@bio.ifi.lmu.de.
† Institut für Informatik.
‡ Molecular Networks GmbH.
§ Universität Erlangen-Nürnberg.

10.1021/ci700433d CCC: $40.75
Published on Web 06/06/2008

search, and in most cases we were able to find literature supporting either the ITSE solution or the BioPath mapping.

METHOD

While a detailed description of the mapping method is given in the companion manuscript,[2] we will mention here its main principles and characteristic features which are important for this work. The approach is based on the following three assumptions:

1. The valid 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).

The minimization of the ITSE corresponds to a weighted maximum common subgraph (MCS) problem, on graphs (line graphs) that are derived from the chemical graphs of the reactants and the products. In a line graph the nodes correspond to the bonds in the original molecules, and therefore the maximum common induced subgraph (MCS) on line graphs corresponds to a weighted maximum matching of bonds.

In the matching algorithm two bonds can only be 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 the weights as they are given in the companion manuscript, i.e. a weight of 1.5 for CC σ-bonds, 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 0.02 for each additionally mapped π-bond. 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. The line graph matching identifies the reacting bonds and is then used in a second step to obtain the complete atomic reaction mapping. For more details we refer to the companion manuscript.[2]

As detailed in the companion manuscript the mapping of hydrogen atoms has not been included in the MCS procedure for a number of technical reasons. As hydrogen atom exchange with the environment is very likely and generally not explicitly included in the reaction itself, we have followed the example of previous work, in not explicitly determining it. 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. This is equivalent to assigning a very low weight to hydrogen atom bonds.

Stereochemistry has been neglected in the determination of the mapping. Theoretically, taking stereochemistry into account is possible for MCS approaches. However, the two stage approach we take here complicates this. Again, we have followed the example of previous work and have also ignored it for simplicity.

Table 1. Statistics on the Comparison of Reaction Mappings Stored in the BioPath Database (with Manually Mapped Reaction Information) and Derived with the ITSE Method[a]

	BioPath	ITSE
number of mapped reactions	1542	1538
not mapped	-	4
uniquely mapped	1542	1331
two alternatives	-	117
more than two alternatives	-	90
identically mapped	1501	
no identical mapping	37	
nonchemical symbols in reaction	12	
mesomerically related mapping	1	
true discrepancies	24	
correct according to literature search/chemical rules	10	10
not found in literature search	4	

[a] In the upper part of the table the statistics pertaining to the BioPath and ITSE mappings independently are given. In the middle part the comparison of the mapping results in overall is given. In the lower part of the table the 37 cases where differences in the mapping were found are further split up after visual inspection.

The algorithm used is a branch and bound approach that returns all optimal mappings and can guarantee the global optimality of the found solutions. In general, the different alternatives are symmetry related representations of a chemically equivalent reaction mapping. A single representative is kept for each equivalence class. The mapping that has been manually annotated in the BioPath database is treated as an additional alternative and is compared to the other solutions in order to identify the cases where the manually annotated mapping was not found by the algorithm.

Finally, for reasons of consistency with the BioPath representation of reaction centers, we have used the description of marking reacting bonds for the figures in the result section as opposed to the ITS based representation used in the companion manuscript.

RESULTS

For each of the reactions a single atomic reaction mapping is given in the BioPath database. This is compared to the solutions of the algorithm. In the BioPath database hydrogen atoms are encoded in the same way as other elements, i.e. bonds to hydrogen atoms that are broken or built during a reaction are also part of the reaction center. Although all hydrogen atoms are encoded in the BioPath database, the validation of the ITSE method was performed by comparing only heavy atoms neglecting the hydrogen atoms for the reasons given in the methods section. For the comparison, each reaction with marked reaction center information from the ITSE method was compared with the corresponding manually mapped reaction of the BioPath database. The results of the comparison are presented in Table 1.

Of the reactions in the BioPath database, 231 contained nonchemical symbols such as ACP for acyl carrier protein or E for enzyme. These are normally substituted by the element with which these substances form a bond to the other reactants (usually C). We will refer to these reactions as nonchemical (NC) reactions, since they do not correspond to a standard chemical reaction equation. Nevertheless, we have performed the mapping for these reactions for completeness.

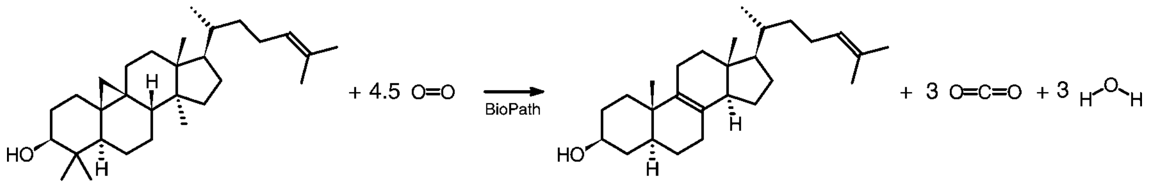

Figure 1. Reaction RXN00048 of BioPath. This figure simplifies the reaction equation for better understanding: identical molecules are summarized by using coefficients, and no reaction center information is shown.

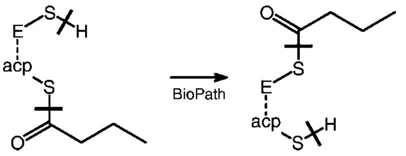

Figure 2. Reaction RXN00025 of BioPath catalyzed by beta-ketoacyl-acyl-carrier-protein synthase I (EC number 2.3.1.41). The reaction center is marked with lines perpendicular to bonds: thick lines represents bonds that are broken or built during the reaction.

The vast majority (~1480) of reactions takes less than 1 min for mapping. However, the mapping of 4 reactions (RXN00048, RXN00084, RXN01206, RXN01236) could not be automatically determined even after 24 h of CPU time. Three of these produce an excessive amount of solutions at the line graph matching level (>500,000) due to self-similarity or identity between reactants. For example RXN00048 is a reaction that summarizes several reaction steps (see Figure 1).

Furthermore, it contains 4.5 oxygen molecules on the reactant side which can arbitrarily be matched to the corresponding oxygen atoms on the product side. The nine O atoms thus already lead to 9! (362,880) solutions, which need to be tested for symmetry to identify unique mappings. These four reactions which could not be mapped within a reasonable time account for the difference between the number of mapped reactions found in BioPath and obtained from the ITSE method (Table 1).

Of the remaining reactions, in 1501 cases (97.5%) the ITSE approach found the reaction mapping annotated in the BioPath. Of the remaining 37 cases, 12 reactions were nonchemical (NC) reactions, which we do not consider as true discrepancies as the correct chemistry of the reactants is not given in the reaction equation. The type of problem typically seen in NC reactions is best explained with the example of reaction RXN00025 (see Figure 2).

In the input of the algorithm the nonchemical symbols E and ACP are substituted by C atoms. Thus, the algorithm sees the same molecule on each side of the reaction and concludes that no reaction takes place. This type of problem is a technical artifact, and therefore we do not consider these reactions as true discrepancies between our results and the database. In contrast, the remaining 25 reactions were further analyzed to identify possible causes for the discrepancy. To this end, the reactions were inspected, and, where possible, the corresponding literature was consulted. In some cases, previous work showed significant evidence for a particular mechanism. In other cases, the authors merely suggested a particular mechanism which was taken as independent support for one of the suggested reaction mappings.

Of 25 remaining reactions, in one case (RXN00495), visual inspection showed that the difference between the BioPath mapping and our solution consisted only in the mesomeric structure of an aromatic ring system, and thus the two mappings can be considered identical.

In ten cases the mapping suggested by the algorithm is wrong: In three cases (RXN00616, RXN00834, and RXN00903) the neglect of stereochemical information in the automatic mapping procedure leads to the suggestion of a simpler reaction than the one actually taking place. Three further reactions are the dioxygenase reaction RXN00047 and two transamination reactions (RXN00172, RXN00509) which are discussed in detail below. The last four cases (RXN00064, RXN00620, RXN01073, RXN01355) are reactions with little in common: RXN00064 for which the manual mapping indicates a slightly more complicated mechanism, involving the change of two more π-bonds, but which corresponds to the putative reaction mechanism;[8] reaction RXN00620, which is a multistep reaction in the biosynthesis of molybdopterin; reaction RXN01073, which is a deaminase; and reaction RXN01355, which is an imidazolase forming part of the xanthine degradation pathway. In an additional four cases (RXN00415, RXN00648, RXN00761, RXN01149), we found no information in the literature on the atomic mappings. Finally, in the remaining 10 cases (RXN00053, RXN00184, RXN00225, RXN00261, RXN00400, RXN00801, RXN01010, RXN01051, RXN01074, RXN01500) we have either found independent support for the mapping suggested by the algorithm or simple chemical rules that strongly support the correctness of the mapping (Table 2).

Reaction RXN00400 in Table 2 contains in the Biopath mapping a change of two bonds (one deleted, one formed) which can be avoided by a simple rotation of the substrate, due to its symmetry. Reactions RXN00801 and RXN01051 contain in BioPath some additional changes in bonds which are not consistent with the reductase activity of the reaction.

The above comparison of the two mappings allows a simultaneous validation of both the database and the reaction mapping algorithm: Agreement between the reaction mappings indicates on one hand that the automatically generated mappings correspond to chemical knowledge and intuition and are therefore at least reasonable. On the other hand, the concordance shows that the manual annotation is consistent with simple chemical principles and does not depend on subjective criteria. This is particularly obvious in this case since the BioPath database contains a number of similar reactions which have been annotated by different persons. It is important to note that the agreement of the two independent methods for a particular reaction mapping is no proof of its correctness, since both algorithm and annotator may be led to the same wrong result by chance. Furthermore, in many cases the reaction suggested by either algorithm or annotator may be chemically valid; however, the enzyme may catalyze a more complex route for the reaction, which leads to a different reaction mapping.

Table 2. List of Reactions Where the Solution Found by the Algorithm Has Independent Support

BioPath ID	enzyme	N_{alt}	EC	literature/reason
RXN00053	isochorismate synthase	1	5.4.99.6	Kozlowski et al. 1995[9]
RXN00184	2-oxoglutarate-4-dioxygenase	1	1.14.11.2	Myllyharju and Kivirikko 1997[10]
RXN00225	acetyl-coa acetyltransferase	1	2.3.1.9	Holliday et al. 2005, 2007[11]
RXN00261	myo-inositol oxygenase	1	1.13.99.1	Xing et al. 2006[12]
RXN00400	dextransucrase	1	2.4.1.5	symmetry
RXN00801	cob(ii)alamin reductase	1		additional reactions in BioPath
RXN01010	cephalosporinsynthase	2		Valegard et al. 1998[13]
RXN01051	aquacobalamin reductase	1	1.6.99.8	additional reactions in BioPath
RXN01074	N-acylneuraminate-9p synthase	2	4.1.3.20	Gunawan et al. 2004[14]
RXN01500	threonine synthase	1	4.2.99.2	Omi et al. 2003[15]

In order to assess the improvement obtained by using the ITSE method compared to simple graph based algorithms, we performed an unweighted reaction mapping, with the same algorithm, by simply matching bonds between identical elements. In that comparison the agreement with the BioPath database is significantly lower, showing discrepancies in 69 cases, of which only 8 cases correspond to NC reactions. An extensive literature search for these reactions was not necessary since they mainly contain CC bond breakages in transamination or transesterification reactions of different types. Thus, we can say that in the unweighted case 52 cases are clearly wrong as compared to 14 wrong or unresolved cases for the ITSE method. This corresponds to an improvement in the error rate by a factor of almost 4.

In the following, some cases in which the solutions provided by the algorithm do not agree with the reaction mapping found in the BioPath database are discussed in more detail.

Example 1: 4-Hydroxyphenylpyruvate Dioxygenase (RXN00047). 3-(4-Hydroxyphenyl)pyruvate reacts with oxygen under the influence of the enzyme 4-hydroxy-phenylpyruvate dioxygenase (HPD, EC code 1.13.11.27) to homogentisate and carbon dioxide (see Figure 3).

Present chemical knowledge assumes that HPD incorporates a 1,2-hydride shift (also known as an NIH shift)[16,17] during the oxidative decarboxylation of the alpha-oxo acid and that is also the way this reaction is stored in the BioPath database.

The first reaction step is the attack of an oxygen atom at the alpha-carbonyl group of 3-(4-hydroxyphenyl)pyruvate forming a peroxy acid that reacts to an epoxide under annihilation of the aromatic system. The highly strained 3-membered ring is opened by a proton forming a hydroxyl group. The next reaction step is the transfer of the acetoxy group to form a more stable carbocation. Finally, the aromatic system is recovered by loss of a proton (see Figure 4).

The reaction mechanism experimentally verified cannot be found among the solutions obtained by the ITSE method, although the ITSE method produces five different alternatives. The solution that comes closest to the reaction mechanism is shown in Figure 5. The only difference is the position of the acetyl-group (−CH2−COO) shifted across the aromatic system. In the experimental reaction mechanism, the hydroxyl group of the reactant is not changed during the reaction and is found in the meta position to the acetyl group in the product molecule. In the solution found by the ITSE method, the hydroxyl group is found in the ortho position to the acetyl group.

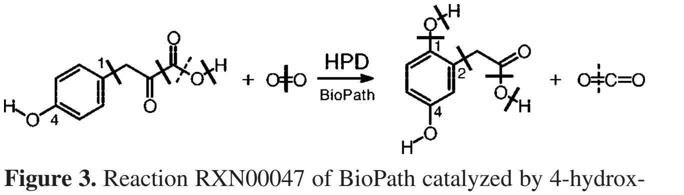

Figure 3. Reaction RXN00047 of BioPath catalyzed by 4-hydroxyphenylpyruvate dioxygenase (EC number 1.13.11.27). The reaction center is marked with lines perpendicular to bonds: thick lines represent bonds that are broken or built during the reaction, and dashed lines show bonds that change in bond order. The numerical labels indicate selected atom−atom mapping numbers.

Example 2: Isochorismate Synthase (RXN00053). The enzyme isochorismate synthase (EC 5.4.4.2) catalyzes the reversible reaction of chorismate to isochorismate. The mechanism of this reaction (RXN00053) is stored in BioPath as transferring the carboxy group across the conjugated system through a 1,3-shift. In this case, only one bond has to be broken and one bond is formed during the reaction (see Figure 6a).

The ITSE method found a different solution for this reaction. Here, the hydroxyl group performs a 1,5-shift across the conjugated system and causes a cascading shifting of single and double bonds of the conjugated ring system. Therefore, in this case, an additional four bonds change their bond order in the reactant and product molecule besides one bond to the hydroxyl group that is broken and built (see Figure 6b).

The reason for the preference of this reaction over the one coded in the BioPath database lies in the weights used for C−C and C−O bonds: A bond is broken when it is not part of the MCES, and a C−O bond is more easily broken than a C−C bond since the corresponding weights for matching are 1.0 and 1.5, respectively.

Interestingly, up to 2003 the enzyme isochorismate synthase was contained in the class “EC 5.4.99 Transferring Other Groups” of the enzyme classification system. In 2003 it was reassigned to the category “EC 5.4.4 Transferring hydroxy groups”.[18] The new class is consistent with the reaction mechanism implied by the ITSE mapping, while the former class is consistent with the reaction mapping in the BioPath database. Accordingly, we found literature supporting the mechanism which is consistent with the ITSE result for this enzyme.[19]

All coding errors of reactions found in BioPath which are mentioned in this publication have been corrected in the meantime.

Example 3: Transamination Reactions. In a transamination reaction, an amino acid reacts with an alpha-keto acid under the catalysis of a transaminase or aminotransferase with the coenzyme pyridoxal-phosphate (Figure 7). Although the mechanism of this reaction type looks like a double

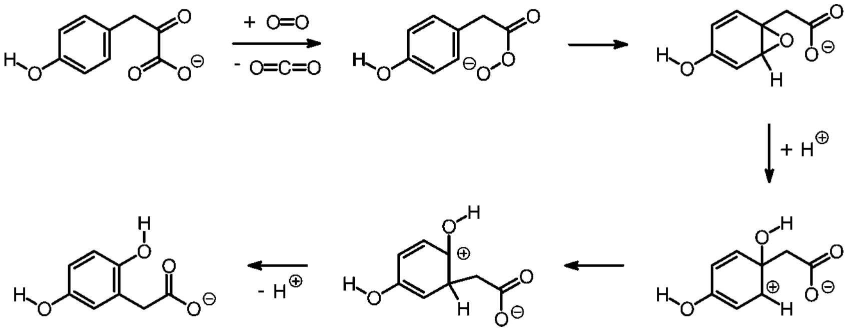

Figure 4. Mechanism of the reaction of 3-(4-hydroxyphenyl)pyruvate to homogentisate.

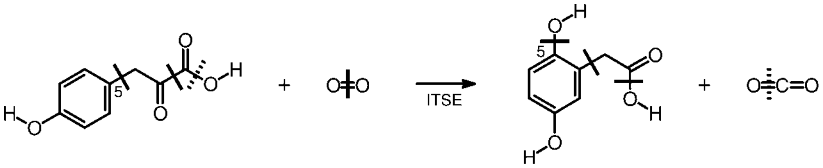

Figure 5. One of altogether five solutions found by the ITSE method for reaction RXN00047.

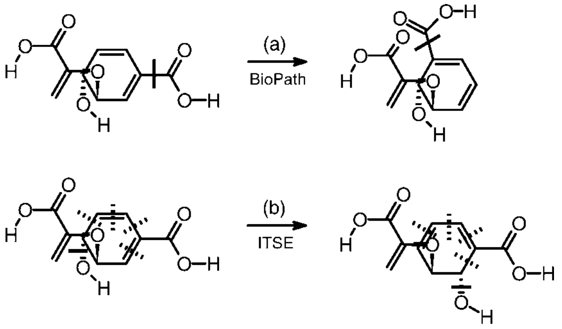

Figure 6. Reaction RXN00053 in the BioPath database (a) and the solution of the ITSE method (b).

replacement of an amino group and a carbonyl group, it is not that simple. The amino acid (**A**) first reacts with a Schiff base of PLP (**B**) to an aldimine derivative (**D**) and through a tautomerism step further to a ketimine (**F**). The next hydrolysis step (**V**) releases the oxo acid (**H**). In the second part of the so-called "ping-pong-bi-bi" mechanism[20] these steps are repeated in reverse direction (reactions steps **V**−**I**) with another oxo acid to make another amino acid (see Figure 7).

Because of this mechanism the oxygen atoms of the two oxo acids (see Figure 8) are different; usually one is derived from a water molecule. Therefore, a transamination reaction in BioPath is coded with one molecule of water on each side of the reaction equation, although this could be reduced to a shorter reaction equation (see Figure 8).

The reaction mappings for transamination reactions in the BioPath correspond to the mechanism shown in Figure 8. The ITSE method finds the correct mapping corresponding to the reaction with and without water (e.g., RXN00609 and RXN00669, respectively). These examples demonstrate the relevance of the weighted MCS: Unweighted MCS yields a reaction where an acetate moiety is transferred from the amino acid to the oxo acid, by breaking and forming a C−C bond. The correct transamination reaction is not found. Taking bond stability into account leads to 2−4 different candidate solutions (depending on whether the solvent is included in the reaction equation); one of these solutions corresponds to the reaction mapping in the BioPath database.

In the case of RXN00172, however, the similarity of the reactants allows an even simpler reaction to take place in which an oxygen atom is transferred (directly or via solvent) from one reactant to another, by breaking and forming a single C−O bond (see Figure 9).

Since the correct reaction requires breaking and forming a C=O bond as well as further bond changes, the principle of additive (non negative) weights for bond stabilities will always lead to the (wrong) solution we find here. Therefore, while not impossible to correct (e.g., by taking higher order information into account) this example demonstrates a certain limitation of the additivity assumption. Reaction RXN00559 is similar in principle showing the same degree of substrate similarity, which again allows two single bond changes that lead from reactants to products.

Example 4: Procollagen-Proline Dioxygenase (RXN00184). The enzyme procollagen-proline dioxygenase catalyzes the reaction of procollagen L-proline, 2-oxoglutarate, and an oxygen molecule to procollagen *trans*-L-4-hydroxyproline, succinate, and carbon dioxide under the influence of Fe^{2+} and ascorbate.

The catalyzing enzyme with the EC number 1.14.11.2 is important in the post-translational modification of collagen.

The ITSE method suggests an elimination of carbon dioxide by breaking the bond between the acid group and the 2-oxo group and changing the bond order in the acid group (Figure 10a).

In BioPath, the reaction center information of this reaction shows the elimination of the 2-oxo group and its subsequent oxidation to carbon dioxide (Figure 10b).

Studies on the active site of the enzyme were done by Myllyharju and Kivirikko[10] and Myllylä et al.[21] They describe the arrangement of 2-oxoglutarate, the oxygen molecule, and the amino acids of the active site around the iron ion (see Figure 11).

Therefore, the elimination of the carboxylic acid is more likely than the elimination of the 2-oxo group. In this case the ITSE method found a reaction that was wrongly coded in the BioPath database (this has been corrected in the meantime).

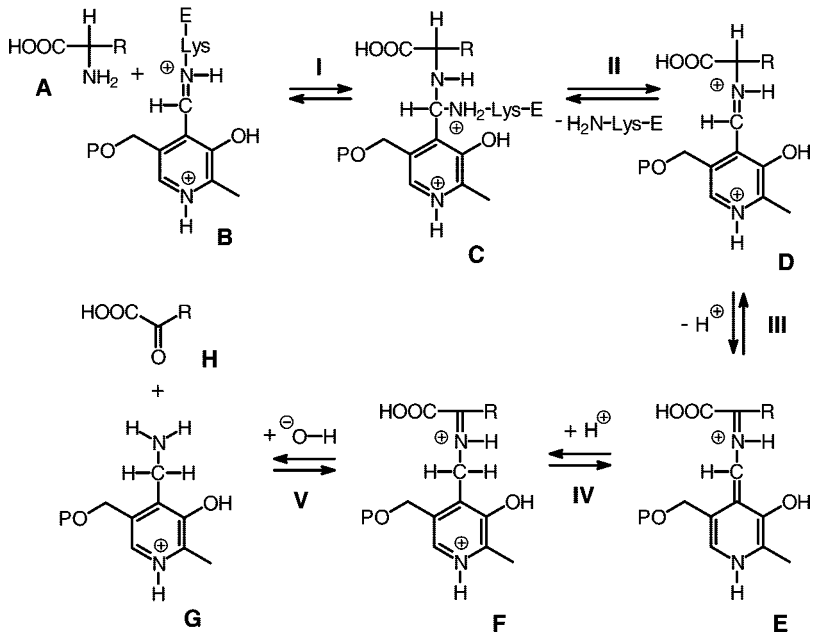

Figure 7. Mechanism of a transamination reaction.

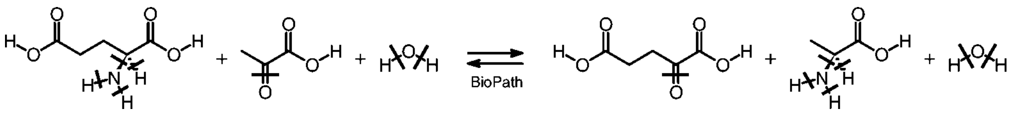

Figure 8. A typical transamination reaction (RXN00067) in the BioPath database: L-glutamate and pyruvate react to 2-oxoglutarate, and L-alanine is catalyzed by alanine transaminase.

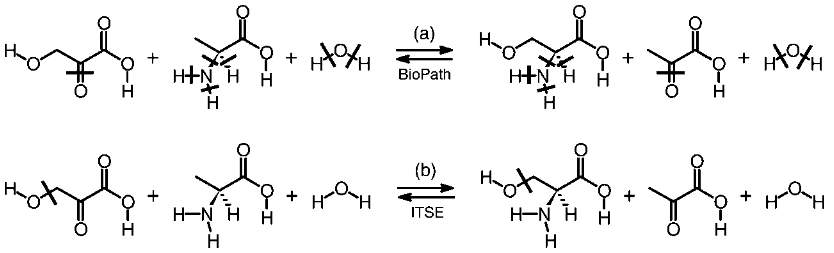

Figure 9. Reaction RXN00172 as it is stored in the BioPath database (a) and mapped with the ITSE method (b).

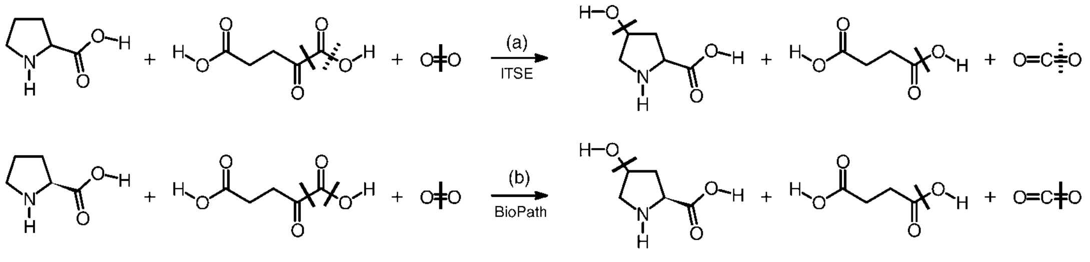

Figure 10. Reaction RXN00184 mapped with the ITSE method (a) and as it is stored in the BioPath database (b).

Example 5: Acetyl-CoA C-Acetyltransferase (RXN00225). In this example, two molecules of acetyl-CoA react to acetoacetyl-CoA and coenzyme A. This reaction is catalyzed by acetyl-CoA C-acetyltransferase (EC number 2.3.1.9). In the BioPath database both reactants are coded with a make/break flag between the sulfur atom and the carbonyl-carbon atom. On the product side also two bonds of the acetoacetyl-CoA molecule are marked as being made during the reaction (see Figure 12a). However, the ITSE method only marks one bond of both reactants and on the product side only the bond between the 1,3-dioxo fragment (see Figure 12b).

The exact reaction mechanism is found in the literature[22,23] and in the database of enzyme reaction mechanisms called MACiE[11] as entry M0077.[24] Here, the user has access to individual reaction steps of this reaction. In this case, the reaction is split into six individual reaction steps describing in detail the reaction mechanism under the influence of the amino acids of the binding pocket. One step of the reaction sequence in which the new carbon−carbon bond is built is shown in Figure 13.

Again, the ITSE method detected a reaction that was a wrongly coded reaction in the BioPath database (and has meanwhile been corrected).

Alternative Reaction Mappings. As shown in Table 1, in 207 cases the algorithm finds two or more alternatives

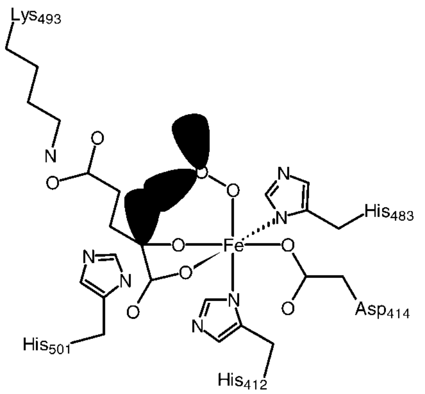

Figure 11. Binding site of prolyl 4-hydroxylase as described by Myllyharju.

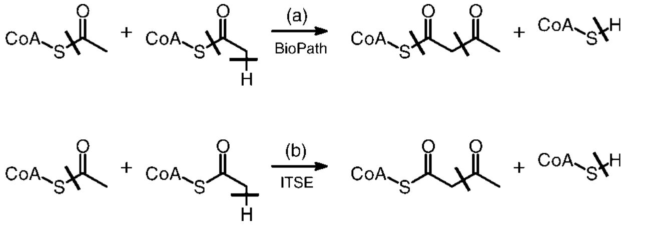

Figure 12. Reaction RXN00225 in the BioPath database (a) and mapped with the ITSE method (b).

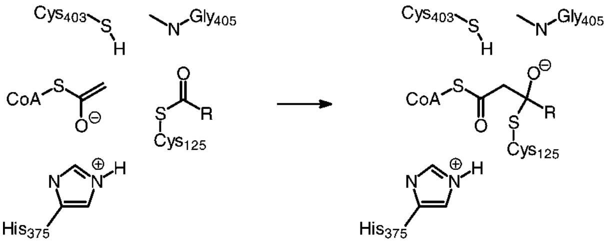

Figure 13. Reaction step no. 5 in the reaction sequence building acetoacetyl-CoA from two molecules of acetyl-CoA (the image corresponds to the image of MACiE for entry M0077).

for the reactions. A large number of those cases share the same reasons:

- In 68 cases, the reaction takes place under ATP hydrolysis, and the alternative reaction mapping shows a different attack of water on the P−O bond of ATP from the β- or γ-phosphate side.
- In 32 cases, the reaction is a transamination of the type discussed under example 3 (transamination reactions). Here the ITSE method finds at least two different alternatives, e.g. one including a water molecule and one involving direct transamination, as discussed for the reactions RXN00609 and RXN00669.
- In 20 cases, the ITSE method suggests alternatives that are not compatible with the stereochemistry of the molecules. Nineteen of those are reactions involving sugars.
- Seventeen reactions belong to the reactions already discussed because they resulted in a number of alternatives, none of which corresponded to the reaction mapping in the BioPath database.
- Eight reactions are transamidation reactions which take place under consumption of ATP. In the mapping of the BioPath reaction RXN00979, for example, the ATP hydrolysis and the transamidation appear as independent reactions. The ITSE result implies an alternative mechanism that shows that the coupling could be obtained by allowing the oxygen atom, freed in the transamidation reaction, to attack the pyrophosphate of ATP. This is consistent with the pyrophosphate being transferred to the nicotinic acid moiety as shown in Figure 14.

A similar mechanism is found for reaction RXN00356 (cytidine triphosphate synthase); however, here the BioPath reaction mechanism also suggests the transfer of the phosphate to the pyrimidine ring of CTP. For this reaction this turns out to be the mechanism suggested in the literature.[25]

Finally, eight more reactions are prenyl transferase reactions, which in some cases (e.g., RXN00933) have the additional prenyl group added at the end of the pyrophosphate activated chain (see Figure 15), while in others (e.g., RXN01480) it is added at the beginning, where it effectively substitutes the pyrophosphate (see Figure 16).

It is interesting to note, that while the ITSE method finds both alternatives, and scores them with an equal weight, the latter mechanism involves the breaking of fewer bonds to hydrogen atoms. For a similar reaction, Fujihashi et al. suggests that the crystal structure of undecaprenyl diphosphate synthase provides support for the latter mechanism.[26]

The remaining 62 reactions cannot be easily grouped into classes and are being studied independently with respect to the relevance of the alternative mechanism suggested by the ITSE.

CONCLUSION

To our knowledge, we have presented the most extensive validation to date comparing reaction center and reaction mapping predictions for more than 1500 manually annotated biochemical reactions. The mapping of hydrogen atoms has not been included in the comparison, and also the stereochemistry was neglected although this information is available in the BioPath database.

The energy minimization principle behind the ITSE method appears significantly superior to previous approaches that identify reaction mappings by matching atoms according to graph isomorphism. It significantly reduces the error rate by a factor of almost 4. The agreement of the predicted reaction mappings with the manual annotations is considerable, with only 24 cases showing disagreement. We have found independent support in the literature for either the automatically derived (10 cases) or the manual reaction mapping (10 cases) in a total of 20 of the 24 cases. The automatically produced reaction mappings thus helped identify a number of reactions that had been misrepresented in the database.

While in general it appears that the algorithm leads to good results even in reactions which do not proceed in a single step, it is clear to us that for more complex (multistep) reactions our approximation will break down. As suggested by one reviewer, secondary metabolism in plants will probably contain numerous such examples.

It is important to note that agreement between automatic and manual mapping does not prove the correctness of the mapping. It is possible that both manual annotator and the ITSE method lead to the same wrong reaction model. The observed high agreement in the mappings, over the complete database does, however, demonstrate that the chemical intuition that has gone into the curation of the database is consistent and at the same time that the

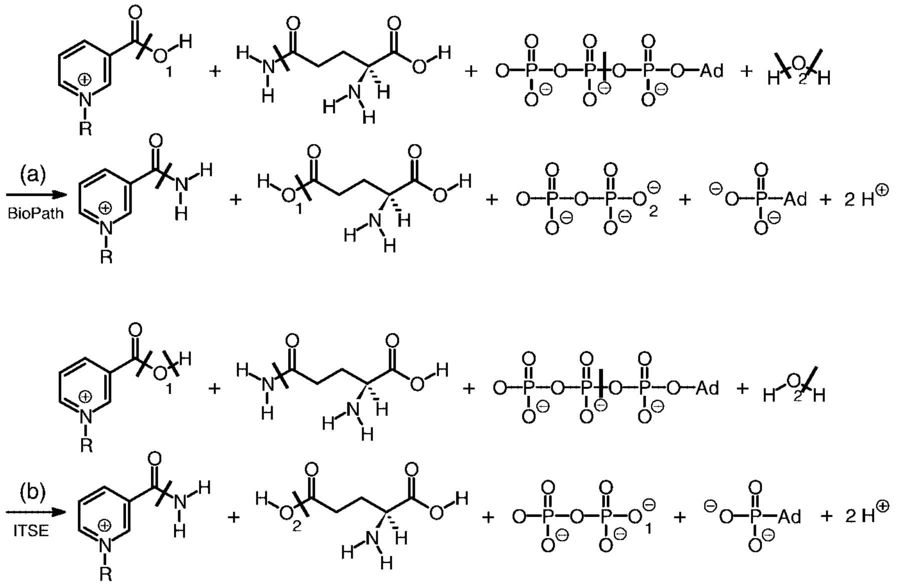

Figure 14. Reaction RXN00979 in the BioPath database (a) and the alternative reaction mapping found by the ITSE (b). Ad represents the adenosine residual, and R stands for the adenine dinucleotide rest.

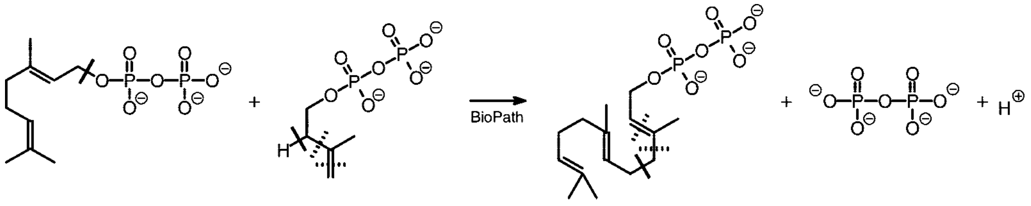

Figure 15. Reaction RXN00933 in the BioPath database.

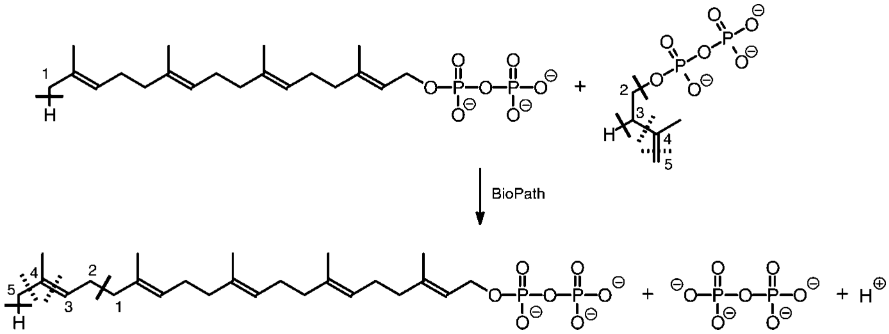

Figure 16. Reaction RXN01480 in the BioPath database.

recently introduced ITSE method leads to plausible hypotheses for the atomic mappings of biochemical reactions. The consistent reaction mapping in the BioPath database is remarkable in view of the fact that a number of different people have provided it. The validation of the manually assigned reaction mapping in the BioPath database with the algorithm presented here has shown the quality of the reaction center assignment and has allowed the elimination of some errors in the mapping. Thus, the BioPath database can now boast of a highly curated assignment of reaction centers which is so crucial for a deeper understanding of the mechanism of enzyme-catalyzed reactions. Furthermore, it allows highly interesting applications as shown for the search of transition state analogs.[1]

The comparison between the manually annotated database and the results of the algorithm shows that it is the first suggested approach with comparable accuracy to trained humans. Furthermore the method enumerates all optimal solutions when more than one solution exists. Each of the solutions corresponds to a different reaction mechanism which leads also to a different mapping. In many cases, it is difficult, if not impossible, to clearly identify the correct alternative on the basis of chemical intuition only, as at least some of the alternatives are chemically viable, making the choice of the reaction path a question that needs to be answered in the context of the enzyme catalyzing the reaction.

ACKNOWLEDGMENT

Oliver Sacher and Johann Gasteiger appreciate the initiation of the Biochemical Pathways project by Spektrum Akademischer Verlag and the collaboration with Professor Guido Moerkotte and Dr. Carl-Christian Kanne, University of Mannheim, Germany (projects no. 08 C 5850 0, 08 C 5879). The development of the database was continued within the BFAM project initiated by Professor Hans-Werner Mewes. We are grateful to our students who assiduously searched the literature and input the molecules and reactions into the BioPath database. We would like to thank the Bundesministerium für Bildung und Forschung for financial support (project no. 031U112A, 031U212A, 031U112D, and 031U212D).

REFERENCES AND NOTES

(1) Reitz, M.; von Homeyer, A.; Gasteiger, J. Query Generation to Search for Inhibitors of Enzymatic Reactions. *J. Chem. Inf. Model.* **2006**, *46*, 2333–2341.

(2) Körner, R.; Apostolakis, J. Automatic Determination of Reaction Mappings and Reaction Center Information: The Imaginary Transition State Energy approach. *J. Chem. Inf. Model.* **2008**, *48*, 1181−1189.

(3) Reitz, M.; Sacher, O.; Tarkhov, A.; Trümbach, D.; Gasteiger, J. Enabling the exploration of biochemical pathways. *Org. Biomol. Chem.* **2004**, *2*, 3226–3237.

(4) The BioPath database is accessible on the internet at the following address: http://www.molecular-networks.com/biopath/index.html (accessed November 16, 2007).

(5) Biochemical Pathways Wall Chart, Michal, G., Ed.; Boehringer Mannheim (now Roche), Germany. It can also be accessed on the internet at http://www.expasy.org/tools/pathways/ (accessed November 16, 2007).

(6) Michal, G. Biochemical Pathways - An Atlas of Biochemistry and Molecular Biology; Spektrum Akademischer Verlag: Heidelberg, 1999.

(7) MN.CHECK is maintained and distributed by Molecular Networks GmbH, Erlangen, Germany. http://www.molecular-networks.com (accessed November 16, 2007).

(8) Zhao, Y.; Schenk, D. J.; Takahashi, S.; Chappell, J.; Coates, R. M. Eremophilane Sesquiterpenes from Capsidiol. *J. Org. Chem.* **2004**, *69*, 7428–7435.

(9) Kozlowski, M. C.; Tom, N. J.; Seto, C. T.; Sefler, A. M.; Bartlett, P. A. Chorismate-Utilizing Enzymes Isochorismate Synthase, Anthranilate Synthase, and p-Aminobenzoate Synthase: Mechanistic Insight through Inhibitor Design. *J. Am. Chem. Soc.* **1995**, *117*, 2128–2140.

(10) Myllyharju, J.; Kivirikko, K. I. Characterization of the iron- and 2-oxoglutarate binding sites of human prolyl 4-hydroxylase. *EMBO J.* **1997**, *16*, 1173–1180.

(11) (a) Holliday, G. L.; Bartlett, G. J.; Almonacid, D. E.; O'Boyle, N. M.; Murray-Rust, P.; Thornton, J. M.; Mitchell, J. B. O. MACiE: a database of enzyme reaction mechanisms. *Bioinformatics* **2005**, *21*, 4315–4316. (b) Holliday, G. L.; Almonacid, D. E.; Bartlett, G. J.; O'Boyle, N. M.; Torrance, J. W.; Murray-Rust, P.; Mitchell, J. B. O.; Thornton, J. M. MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res.* **2007**, *35*, D515−D520.

(12) Xing, G.; Barr, E. W.; Diao, Y.; Hoffart, L. M.; Prabhu, K. S.; Arner, R. J.; Reddy, C. C.; Krebs, C.; Bollinger, J. M., Jr. Oxygen Activation by a Mixed-Valent, Diiron(II/III) Cluster in the Glycol Cleavage Reaction Catalyzed by myo-Inositol Oxygenase. *Biochemistry* **2006**, *45*, 5402–5412.

(13) Valegard, K.; et al. Structure of a cephalosporin synthase. *Nature* **1998**, *394*, 805–808.

(14) Gunawan, J.; Simard, D.; Gilbert, M.; Lovering, A. L.; Wakarchuk, W. W.; Tanner, M. E.; Strynadka, N. C. Structural and Mechanistic Analysis of Sialic Acid Synthase NeuB from Neisseria meningitidis in Complex with Mn2+, Phosphoenolpyruvate, and N-Acetylmannosaminitol. *J. Biol. Chem.* **2004**, *280*, 3555–3563.

(15) Omi, R.; Goto, M.; Miyahara, I.; Mizuguchi, H.; Hayashi, H.; Kagamiyama, H.; Hirotsu, K. Crystal Structures of Threonine Synthase from Thermus thermophilus HB8, conformational change substrate recognition and mechanism. *J. Biol. Chem.* **2003**, *278*, 46035–46045.

(16) Guroff, G.; Daly, J. W.; Jerina, D. M.; Renson, J.; Witkop, B.; Udenfriend, S. Hydroxylation-induced migration: the NIH shift. Recent experiments reveal an unexpected and general result of enzymatic hydroxylation of aromatic compounds. *Science* **1967**, *157*, 1524–1530.

(17) The function of 4-hydroxyphenylpyruvate dioxygenase is described on Wikipedia. http://en.wikipedia.org/wiki/4-hydroxyphenylpyruvate_dioxygenase (accessed November 16, 2007).

(18) According to the IUBMB Biochemical Nomenclature the entry of enzyme EC 5.4.99.6 was moved to EC 5.4.4.2 in 2003. http://www.chem.qmul.ac.uk/iubmb/enzyme/EC5/4/99/6.html (accessed November 16, 2007).

(19) Knaggs, A. R. The biosynthesis of shikimate metabolites. *Nat. Prod. Rep.* **2001**, *18*, 334–355.

(20) Cleland, W. W. The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations. *Biochim. Biophys. Acta* **1963**, *67*, 104–137.

(21) Myllylä, R.; Günzler, V.; Kivirikko, K. I.; Kaska, D. D. Modification of vertebrate and algal prolyl 4-hydroxylases and vertebrate lysyl hydroxylase by diethyl pyrocarbonate. Evidence for histidine residues in the catalytic site of 2-oxoglutarate-coupled dioxygenases. *Biochem. J.* **1992**, *286*, 923–927.

(22) Modis, Y.; Wierenga, R. K. Crystallographic analysis of the reaction pathway of Zoogloea ramigera biosynthetic thiolase. *J. Mol. Biol.* **2000**, *297*, 1171–1182.

(23) Mathieu, M.; Modis, Y.; Zeelen, J. Ph.; Engel, C. K.; Abagyan, R. A.; Ahlberg, A.; Rasmussen, B.; Lamzin, V. S.; Kunau, W. H.; Wierenga, R. K. The 1.8 Å crystal structure of the dimeric peroxisomal 3-ketoacyl-CoA thiolase of Saccharomyces cerevisiae: implications for substrate binding and reaction mechanism. *J. Mol. Biol.* **1997**, *273*, 714–728.

(24) The MACiE database is accessible on the internet at the following address: http://www.ebi.ac.uk/thornton-srv/databases/MACiE/ (accessed November 16, 2007).

(25) Endrizzi, J.; Kim, H.; Anderson, P. M.; Baldwin, E. P. Crystal Structure of Escherichia coli Cytidine Triphosphate Synthetase, a Nucleotide-Regulated Glutamine Amidotransferase/ATP-Dependent Amidoligase Fusion Protein and Homologue of Anticancer and Antiparasitic Drug Targets. *Biochemistry* **2004**, *43*, 6447–6463.

(26) Fujihashi, M.; Zhang, Y.-W.; Higuchi, Y.; Li, X.-Y.; Koyama, T.; Miki, K. Crystal structure of cis-prenyl chain elongating enzyme, undecaprenyl diphosphate synthase. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4337–4342.

CI700433D