Finding relevant biotransformation routes in weighted metabolic networks using atom mapping rules

Torsten Blum*, Oliver Kohlbacher*

Abstract: Computational analysis of pathways in metabolic networks has numerous applications in systems biology. While graph theory-based approaches have been presented that find biotransformation routes from one metabolite to another in these networks, most of these approaches suffer from finding too many routes, most of which are biologically infeasible or meaningless. We present a novel approach for finding relevant routes based on atom mapping rules (describing which educt atoms are mapped onto which product atoms in a chemical reaction). This leads to a reformulation of the problem as a lightest path search in a degree-weighted metabolic network. A key component of the approach is a new method of computing optimal atom mapping rules.

1 Introduction

Cellular Metabolism consists of a complex network of chemical reactions, connected by small molecules, and operating together to convert the nutrients into energy and biomass components essential for life. It has been shown that real metabolic networks are much more variable compared to the set of pathways defined in biochemical textbooks [Co99]. The existence of alternative pathways enables robustness of the cellular system against perturbations like mutations. Therefore the knowledge of all feasible routes transforming a source metabolite into a target metabolite can help the understand the metabolism better or to decide wether particular enzymes or intermediates are essential in the process. However, experimental determination of pathways is laborious and time-consuming. So far, there is no high-throughput method for this task. Hence, computational approaches for detecting plausible pathways in a metabolic network at genome-scale are needed.

Computational tools for metabolic pathway analysis are becoming more important since the complete genomic information of the putative genes, alongside the functional annotation, is available for an ever increasing number of organisms. Starting from the gene-enzyme relations, one can use the enzyme-reaction as well as the reaction-compound relations (extracted from pathway databases like KEGG [Ka96], EcoCyc [Ke05], MetaCyc [Ca06] and BRENDA [SCS02]) to reconstruct an organism-specific metabolic network.

The recent approaches, described in the literature, can be roughly divided into constraint-based methods inferring closed or stoichiometrically balanced pathways and graph theory-

*Department for Simulation of Biological Systems, WSI/ZBIT, Eberhard Karls University Tübingen, Sand 14, D-72076 Tübingen, Germany, (contact: blum@informatik.uni-tuebingen.de)

based methods searching for linear biotransformation routes. A stoichiometrically balanced pathway is defined as a metabolic subnetwork in which the net production and consumption of all compounds is zero. Excepted are the source and target compounds and a predefined set of pool metabolites. A biotransformation route is simply defined as a linear sequence of chemical reactions where a source compound is converted into a target compound step by step. Both terms are related to each other since a stoichiometrically balanced pathway always contains at least one main biotransformation route.

Each approach has its advantages and disadvantages and differs in its potential range of applications. Constraint-based methods represent the metabolic network as stoichiometric matrix [SDF99, SLP00]. The rows and columns represent the metabolites and reactions. The pathways are inferred using convex analysis [Ro70], a branch of mathematics for analyzing a set of linear equations given a set of constraints. The advantage of the method is that it is mathematically well-defined and enables biotechnological analysis of pathways where the focus is to increase the yield of industrial important metabolites. However, the underlying calculation represents an NP-hard computational problem and it seems to be impossible to use the method for a network at genome-scale [KS02, KSGG03]. In practice, the computational complexity is reduced by the extensive use of constraints like the predefined distinction between internal and external (pool) metabolites and the restriction on a subset of the reactions available in a given organism.

In graph theory-based approches, the metabolic network is mapped onto a mathematical graph [Ar00, Ra05, CCWH06]. An advantage is the availability of already established graph-algorithms (avoiding NP-hard calculations) which can be used for genome-scale network analysis [AS06]. An interactive navigation through metabolic networks is possible, simply by using path finding algorithms between a given source and target, without the need for user-defined constraints [HWGW02]. But searching without information other than the connectivity, i.e. two successive reactions are connected if they have a metabolite in common, often delivers meaningless results. Using this type of naive or “blind” search, approximately 500,000 different routes with at most nine reactions between glucose and pyruvate could be identified in a study [KZL00]. The problem with such a search is mainly based on the high degree of nodes corresponding to pool metabolites like water, ATP, NADH and so on. But ignoring these network hubs cannot be a satisfying solution since their choice is not always obvious. The main problem is that even such a typical side metabolite as ADP acts as a real intermediate in several pathways. Another example is pyruvate where its role as a main or side metabolite is not clear in all reactions [HWGW02]. A better idea is to incorporate the structural information of the metabolites. The PathwayHunter tool [Ra05] uses chemical fingerprints to guide a shortest path search between structurally similar metabolites. Another promising idea is to trace the flow of atoms in a shortet path search using atom mapping rules [Ar00, Ar03]. Given a chemical reaction, an atom mapping rule defines which atom of an educt compound is transferred to which atom of a product compound. This is helpful for detecting biochemically unfeasible shortest paths in which no atom is transferred from the source to the target. The main problems are that the shortest path search tend to go through pool metabolites despite the atom trace. The structural information, necessary for atom mapping calculation, is not given or incomplete for a fraction of the compounds participating in the reactions stored in pathway databases. Those compounds often are described only by a string name or represent general molecules like ”an alcohol”. Furthermore, the automated and efficient calculation of atom mapping rules, given thousands of reactions in a database like KEGG, is not easy.

In a degree-weighted graph representing a metabolic network [CCWH06] each node is assigned a weight equal to its degree. Searching for the lightest path significantly reduces the probability of finding irrelevant routes containing pool metabolites as intermediates. An advantage is that the structural information of the compounds is not needed but the lightest path search fails for routes containing network hubs as intermediates or for particularly long routes.

In this work, we present a novel graph theory-based approach for finding feasible biotransformation routes which integrates atom mapping rules and the lightest path search into a joint method. We also present a novel method for the fully automated and efficient calculation of atom mapping rules. Different graph types and search strategies were analyzed in a genome-scale study.

2 Calculation of atom mapping rules

Representing the compounds of a chemical reaction as molecular graphs [1], atom mapping rules can be calculated using graph partition and graph isomorphism [Ak04]. The underlying idea is that normally in chemical reactions only very few bonds are broken in order to transform the educts into the products. Hence, we can find the mapping rules by removing a limited number of edges in the molecular graphs of the compounds and searching for graph isomorphisms between the remaining connected components. A valid atom mapping contains an isomorphic component of the product side for each connected component of the educt side and vice versa. However, the result of such a search, as presented in a previous work [Ak04], is not necessarily unique and may contain biochemically meaningless mappings alongside the correct one. We can solve this problem by introducing the EC clustering approach. Now, it is possible to filter out the irrelevant mappings by clustering them together with all the mappings of the enzymatic reactions which have the first three digits of their EC number in common. The underlying idea is that only the first three digits describe the reaction mechanism, and the last digit only enumerates the different chemical structures. This allows us to select the atom mapping rule which describes best the reaction mechanism of the EC cluster.

The next paragraph briefly describes the theoretical framework of the approach as introduced earlier [Ak04] followed by the details of our practical algorithm for mapping calculation. The second paragraph presents the EC clustering approach for filtering out irrelevant mappings.

[1] with nodes as atoms (ignoring hydrogen atoms) and undirected edges as chemical bonds

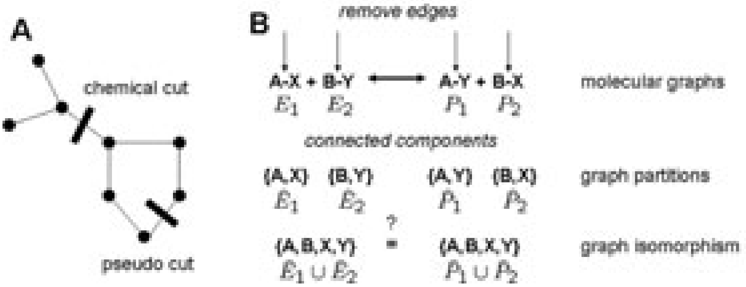

Figure 1: (A) Schematic illustration of chemical cuts and pseudo cuts. (B) The general mapping problem. The example shows a reaction with two educt (E_1, E_2) and two product compounds (P_1, P_2), and a cut-size C=1. Graph partitions ($\hat{E}_1$,$\hat{E}_2$,$\hat{P}_1$,$\hat{P}_2$) were created by removing at most one edge in the molecular graph for each compound. A mapping is found if the multisets $\hat{E}_1 \cup \hat{E}_2$ and $\hat{P}_1 \cup \hat{P}_2$ are equal.

2.1 Problem definition and practical algorithm

Def.: A chemical cut [Ak04] of size C is a partition of a graph G into connected components which are obtained by removing at most C edges whereas the nodes of each removed edge have to belong to different connected components after the removal.

In order to handle reactions modifying ring structures, we must extend the definition of a cut. A pseudo cut removes edges of a graph G which do not disconnect G. The total number of removed edges per compound has still to be not larger than C. An example describing both types of cuts is shown in Fig. 1A.

Def.: Given the chemical reaction equation $E_1 + ... + E_e \leftrightarrow P_1 + ... + P_p$. $E_1, ..., E_e$ and $P_1, ..., P_p$ are molecular graphs representing educt and product compounds. The *mapping problem* is now to find a chemical cut of size C for each $E_1, ..., E_e$ and $P_1, ..., P_p$ such that the resulting multiset of connected components $\hat{E}_1 \cup ... \cup \hat{E}_e$ is equal to the multiset of connected components $\hat{P}_1 \cup ... \cup \hat{P}_p$. Elements of the multisets are equal if they are isomorphic.

For a simple example Fig. 1B illustrates the mapping problem. For fixed values of p, q and C, the problem can be solved in polynomial time, since the number of combinations $(E_1, ..., E_e, P_1, ..., P_p)$ is $\mathcal{O}(n^{C(e+p)})$, where n is the maximum size of a compound in the reaction [Ak04]. Practical algorithms solving the problem for the special case of $C = 1$ and $e = p = 2$ were presented earlier [Ak04]. Here, we introduce a procedure for solving the general problem.

We distinguish two types of mapping rules. Given a chemical reaction, a fragment mapping rule (FMR) defines which connected component (called fragment) of an educt molecular graph is isomorphic to which connected component of a product molecular graph. Such a rule consists of a list of isomorphic fragment pairs. An atom mapping rule (AMR) defines which atom of an educt compound is transferred to which atom of a product compound. A rule of this type consists of a list of atom pairs. From the fragment mapping rules, we can deduce atom mapping rules using the canonical graph representations created by

Morgan's algorithm [Mo65]. [2] We use unique SMILES [WWW89] to detect isomorphic components. The advantage is that this permits the simple incorporation of stereochemical information and reduces the number of inferred irrelevant atom mapping rules. Furthermore, we define two functions which are necessary for the mapping calculation. The first function, $CSF(X)$, transforms the multiset X, which contains connected components as elements, to the multiset Y where the elements of X are replaced by their chemical formulas. Accordingly, the second function, $SMILES(X)$ replaces the elements of X by their unique SMILES.

Practical algorithm: All valid atom mapping rules corresponding to a minimal cut size C can be computed as follows:

1. $C \leftarrow 0$
2. For the molecular graphs of the educts E_1,...,E_e and products P_1,...,P_p create all possible partitions $\hat{E}_{1_i}$,...,$\hat{E}_{e_j}$ and $\hat{P}_{1_k}$,...,$\hat{P}_{p_l}$ using cut size C.
3. Create all possible multisets of connected components $\tilde{E}_s = \hat{E}_{1_i} \cup ... \cup \hat{E}_{e_j}$ and $\tilde{P}_r = \hat{P}_{1_k} \cup ... \cup \hat{P}_{p_l}$.
4. Select all pairs ($\tilde{E}_s$,$\tilde{P}_r$) with $CSF(\tilde{E}_s) = CSF(\tilde{P}_r)$.
5. From all pairs calculated in Step 3 select all pairs ($\tilde{E}_s$,$\tilde{P}_r$) with $SMILES(\tilde{E}_s) = SMILES(\tilde{P}_r)$ and a minimum number of removed edges producing pseudocuts accumulated for all educts and products. Each pair represents an FMR.
6. If no FMR is found in Step 4: $C \leftarrow C + 1$, repeat from Step 1.
7. Extract the final AMRs from the FMRs using the canonical graph representation calculated by Morgan's algorithm.

The third step is introduced to improve the calculation time significantly. It is not necessary to compute the unique SMILES for all partitions. In the first iteration of the algorithm we simply compute the chemical formulas of the connected components and use them to collect a set of candidate partitions for the molecular graphs. Step 4 insures that the mappings found are based on a minimum number of removed edges. If we would search for all mappings allowing the maximum possible cut size C as well as the maximum number of edges producing pseudo cuts, the number of irrelevant mappings per reaction would be much higher. Note that a mapping found by the cut size $C = 0$ typically represents isomerization or oxidoreductive reactions.

2.2 EC clustering

For a significant number of reactions (approximately 40%, data not shown), there is more than one possible mapping rule. An example is shown in Fig. 2A. Using cut size $C =$

[2] The algorithm assigns an unique integer label to each node in a molecular graph, based on the node degree and the degrees of its neighbours. Topologically equivalent nodes in isomorphic graphs get the same labels.

1, there are three possible mapping rules for the reaction catalyzed by serine-pyruvate transaminase (EC 2.6.1.51). But only the first mapping rule describes the underlying reaction mechanism which exchanges the amino group of L-serine with a keto group of pyruvate. To filter out biochemically irrelevant mappings, we introduce the EC clustering approach. The idea is that the mechanism of many chemical reactions consists of shifting or exchanging small functional groups like amino, keto, methyl, phosphate or carboxyl groups. All reactions which have the first three digits of their EC number in common also share the reaction mechanism. The last digit only enumerates the different chemical structures operating as substrates. Typical examples are reactions transferring a phosphate (EC 2.7.1.-) or a methyl group (EC 2.1.1.-) from one molecule to another.

At first, we define an EC cluster (ECC) as a set of enzymatic reactions which have the first three digits of their EC number in common. Given an EC cluster, a reaction mechanism rule (RMR) generally describes, for the reactions in the cluster, how the educts are transformed into the products. Our aim is now to automatically infer the RMR by identifying the relevant functional groups or parts of the substrates. The next step is to select the FMR and appendant AMR which correspond to the inferred reaction mechanism rule and to discard all the other FMRs.

Reaction mechanism rules are represented as strings and constructed from FMRs. The following syntax is used to describe them. The two sides of a reaction are separated by '$=$'. The fragments of each compound are separated by ',' and enclosed by '$<$' and '$>$'. The first fragment representing a non-relevant structure, is designated with '$X1$', the second with '$X2$' and so on. Relevant fragments like the mentioned functional groups are represented using their SMILES (e.g. `N`, `O`, `C`, `OP(O)O`, `C(O)O`). [3] An empty fragment is represented by '$' and is used in graph partitions for compounds in which no edge is removed. The strings representing both the fragments and the whole reaction sides are alphabetically ordered to ensure uniqueness in the comparison with RMRs from different reactions. Fig. 2B shows an RMR for each FMR shown in Fig. 2A.

Note that there is no predefined list of relevant fragments. We generate all combinatorial possible RMRs from the FMRs of a given reaction by allowing each fragment to be relevant or not. Given an EC cluster and a reaction mechanism rule, the occurrence frequency of this rule accumulated over all reactions in the cluster is called the EC cluster score (ECCS). An RMR occurs in a reaction if it can be constructed from at least one FMR of the reaction. From all generated RMRs we select that to be relevant which has the highest score. The EC clustering procedure performs the following steps:

1. For each given FMR containing n educt as well as product fragments, construct for all $\binom{n}{k}$ combinations with $k = 0, ..., n-1$, RMRs in which k fragments are marked as non-relevant (represented as '$X1$', '$X2$', and so on).

2. For all RMRs deduced from an FMR of a reaction in an EC cluster, calculate the EC cluster scores.

3. Assign each FMR of a reaction in a EC cluster the maximum ECCS of the RMRs

[3] Note that the SMILES shown lack double bonds since bond types (parallel edges) are ignored in our molecular graphs for simplicity.

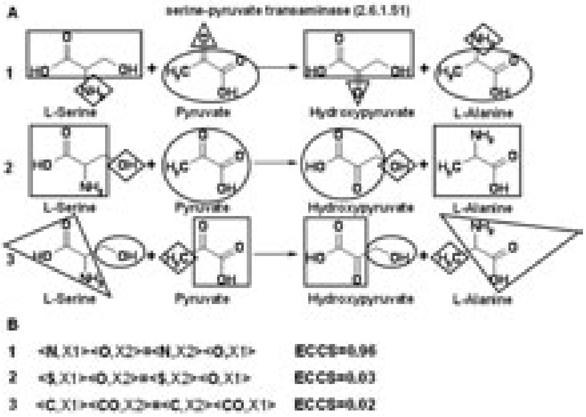
A
serine-pyruvate transaminase (2.6.1.51)
1
2
3
L-Serine
Pyruvate
Hydroxypyruvate
L-Alanine
B
1 <N,X1>+<O,X2>=<N,X2>+<O,X1> ECCS=0.96
2 <S,X1>+<O,X2>=<S,X2>+<O,X1> ECCS=0.03
3 <C,X1>+<CO,X2>=<C,X2>+<CO,X1> ECCS=0.02

Figure 2: (A) A reaction with multiple mapping rules. The atom transfer between both sides of the reaction is represented by equal geometric shapes. The different shapes within a compound also represent the connected components in the corresponding molecular graphs. Only the first rule is biochemically relevant. (B) Each mapping rule is assigned the maximum score (ECCS) of all reaction mechanism rules (RMRs) which were derived from the mapping. The mapping with the highest score is detected as the relevant mapping. For each mapping rule, the best RMR with corresponding score is shown.

which were constructed from the FMR.

4. For each reaction select the FMR (and its corresponding AMR) with the highest score as the relevant mapping.

Considering the example shown in Figure 2, it becomes possible to detect the first mapping rule as biochemically relevant, since the assigned score is significantly larger than the scores of the other two mapping rules. The score of 0.96 for the first RMR indicates that for 96% of the reactions in the EC cluster 2.6.1.- (overall 90 reactions using data from KEGG), the mechanism can be described as exchange of an amino group with a keto group. If there is more than one fragment mapping with the highest score or there is a reaction with no EC number, then we select the mapping as relevant with the minimum number of transferred atoms (the number of atoms of the relevant chemical groups).

3 Finding feasible biotransformation routes

3.1 Degree-weighted metabolic networks

In the degree-weighted metabolic networks approach [CCWH06], the metabolic network of an organism is mapped on a bipartite graph including all compounds and reactions as nodes. Directed edges connect the compound nodes (educts and products) with the reaction nodes. Both directions of a reaction were represented by two independent nodes per reaction. The key idea of a weighted metabolic network is to assign each compound node a weight equal to its degree (e.g. the number of in- and outgoing edges) and each reaction node the weight 1 by default. The weight of a pathway in the graph is then defined

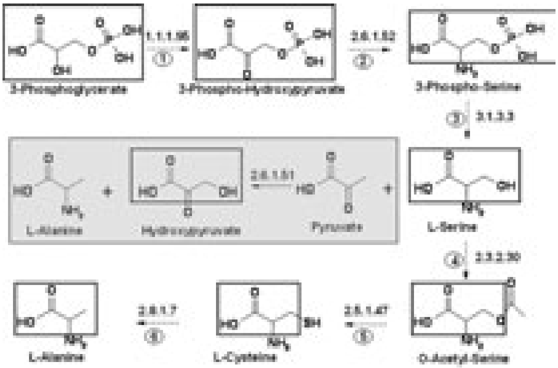

Figure 3: Transformation of 3-phosphoglycerate into L-alanine. The dashed arrows represent a valid pathway which conserves the structural moiety (shown using rectangles). No atom is transferred from 3-phosphoglycerate to L-alanine via L-serine in the reaction shown in the light grey box.

as the sum of the weights of its nodes. This implies that the overall weight of a path is much larger if it contains highly connected compounds like typical pool metabolites or co-factors (e.g. NADP, ATP, water, etc.). Searching for paths with the lowest weight reduces the probability of finding unfeasible biotransformation routes which contains pool metabolites as intermediates between two successive reactions.

A fundamental problem of the lightest path search is its inability to handle important biotransformation routes involving the biosynthesis of pool metabolites (e.g. the purine biosynthesis in which AMP, ADP, GMP and GDP are intermediates). The method fails to reconstruct these routes because pool metabolites participate in many reactions of other transformation processes and, therefore, are assigned very large node weights.

3.2 Combining atom mapping rules with lowest weight paths

The above mentioned problem can be overcome by combining the lightest path search with atom mapping rules. The key idea is to use atom mapping rules to identify biochemically irrelevant paths of low weight. To this end all relevant paths must satisfy the structural moiety constraint which can be defined as follows. A path and its corresponding biotransformation route can only be feasible if at least one atom of the source compound is transferred, via the intermediates, to the target compound. In many cases, this helps to filter out biochemically irrelevant lowest weight paths. We also show that the combination of atom mapping rules with lowest weight paths performs better than searching for the shortest path in the unweighted atom mapping graph. The example shown in Figure 3 illustrates the concept of using the structural moiety constraint for path validation. 3-phosphoglycerate, also known as an intermediate in the degradation of glycose in glycolysis, is used as source metabolite and the amino acid L-alanine as target. The dashed arrows describe a path which consists of six enzymatic steps for transforming 3-phosphoglycerate into L-alanine. Five intermediates are required. The rectangles mark the conserved substructures. In this example 3-phosphoglycerate serves as carbon source for L-alanine. The sequential application of atom mapping rules, linking the educts and products in each reaction, enables a tracing of the conserved structure. Now, it is clear that this path fulfills the structural moiety constraint. The light grey box in the middle complements L-serine to the enzymatic reaction with EC number 2.6.1.51. A path using this reaction as final conversion step would require only four steps in total. However, the application of the atom mapping rule of the reaction implies that no atom could be transferred from 3-phosphoglycerate to L-alanine via L-serine. Hence, this path fails the structural moiety constraint.

We performed several experiments for testing our approach. For this purpose, a bipartite graph was constructed from the EcoCyc database. All reactions from the small molecule metabolism were included representing the *E. coli* metabolic network at genome-scale. We investigated the search performances based on four different network types. The *blind search graph (bsg)* contains only the connectivity information extracted from the metabolic network. Using the *atom mapping graph (amg)*, atom mapping rules are available via the educt-reaction-product node relations. In the *weighted graph (wg)* each edge representing a compound-reaction relation is assigned a weight equal to the degree of the compound in the whole metabolic network. Finally, the *weighted atom mapping graph (wamg)* contains all of the available information as described for the other three network types.

We always search for feasible biotransformation routes between two given nodes (source and target). Using the blind search graph, feasible routes are found by searching for the shortest path. In the atom mapping graph, we search for the shortest path which fulfills the structural moiety constraint. The lightest path is searched in the weighted graph and in the weighted atom mapping graph, we search for the lightest path fulfilling the structural moiety constraint. Paths between two given nodes were calculated using Eppstein's k-shortest path algorithm [Ep98] which efficiently computes the first k-shortest or lightest paths in a directed graph. [4] The algorithm was adopted to incorporate the atom mapping rules. For this purpose, we used an analogous approach for path validation which was proposed by Arita [Ar03]. Each extracted path is validated by an sequential application of the atom mapping rules. In the beginning, all atoms of the source metabolite are available for the mapping. After this, for each step, only those atoms of an educt are available for mapping to the next compound which could be reached by a mapping in the step before. If no atom reaches the target metabolite, the path is rejected as not valid. Atom mapping rules are available for only 63% of the reactions (explained in Section 4.1). This fact is considered in the procedure. If such a reaction is reached, the validation process is restarted with the next reaction which has an atom mapping rule. Hence, both the atom mapping graph and the weighted atom mapping graph can also find paths violating the structural moiety constraint. It should also be mentioned that oxygen and hydrogen atoms are ignored in the process. [5]

[4] The algorithm creates an implicit representation of the k-lightest paths in a directed graph with n vertices and m edges in $\mathcal{O}(m + n \log n + kn)$ which can be traversed using breadth-first-search.

[5] Hydrogen atoms are implicitly represented in the molecular graphs and not considered in the mapping calculation. Although oxygen atoms were considered in the mapping calculation, these atoms are ignored in the path validation process. The problem is that the water molecule is the most frequent pool metabolite and it is often impossible to detect a correct and unique mapping.

Table 1: The results of the atom mapping calculation using the EcoCyc and KEGG data sets.

		EcoCyc	KEGG
reactions	overall	1348	6811
	selected	850 (63.1%)	4621 (67.9%)
	successful	833 (98.0%)	4516 (97.7%)
mappings	overall	1236	5913
	per reaction	1.51	1.31
cut size	$C = 0$	197 (24.0%)	807 (17.8%)
	$C = 1$	553 (67.4%)	3272 (72.5%)
	$C = 2$	71 (8.6%)	437 (9.7%)

4 Results

4.1 Inferring atom mapping rules

Atom mapping rules were inferred from chemical reactions extracted from the KEGG and the EcoCyc databases. The maximum cut-size was restricted to $C = 2$ and the maximum number of compounds permitted per reaction was set to 10. This ensures an efficient calculation. Reactions containing compounds for which the structural information was incomplete or non-existent, and reactions with an unbalanced reaction equation were not considered. This reduces the number of reactions from 6811 to 4621 for KEGG, and from 1348 to 850 for EcoCyc. Tab. 1 summarizes the results of the calculation. For 833 (98%) of the reactions selected from EcoCyc and 4516 (97.7%) from KEGG, at least one atom mapping rule could be found. The overall number of mappings per reaction was 1.51 (EcoCyc) as well as 1.31 (KEGG). The number of reactions with mapping rules using cut size $C = 0$ was 197 (23.6%) for EcoCyc and 807 (17.8%) for KEGG. These are typically stereoisomerization or oxidoreductive reactions in which the transfer of substructures between molecules is not necessary (e.g. EC 1.1.1.-). The majority of the reactions - 563 (67.6%) for EcoCyc and 3272 (72.5%) for KEGG - require atom mapping rules with cut size $C = 1$. Typical representatives are reactions transferring phosphate or methyl groups (e.g. EC 2.7.1.- or EC 2.1.1.-). Seventy-three (8.8%) of the EcoCyc and 437 (9.7%) of the KEGG reactions require atom mapping rules with the cut size $C = 2$. Examples are reactions belonging to EC 1.13.11.- in which two oxygen atoms, originating from molecular oxygen, are transferred. We manually inspected 17 reactions (2%) from EcoCyc and 105 (2.3%) from KEGG for which no atom mapping rule could be inferred. These reactions require mapping rules with a cut size greater than $C = 2$. The hydrolysis of allophanate resulting in two carbon dioxide molecules and two ammonia molecules (EC 3.5.1.54) is an example of a reaction requiring cut size $C = 3$. Another example is the uroporphyrinogen carboxy-lyase reaction (EC 4.1.1.37), in which four molecules of carbon dioxide are cleaved off from uroporphyrinogen ($C = 4$).

4.2 Inferring relevant biotransformation routes

The search performances of the presented network types and search strategies were evaluated by trying to find experimental verified biotransformation routes in the metabolic network of *E. coli*. For this purpose, all annotated biotransformation routes of the small molecule metabolism with at least three reactions were extracted from EcoCyc (137 overall). Given the main source and target metabolites of the annotated routes as start and end nodes, we calculated the shortest (lightest) path constrained to use the first as well as the last reaction of the annotated route. If n annotated routes share the same main source as well as target metabolites and start as well as end reaction, we computed the n shortest (lightest) paths. The quality of the routes found was measured by comparing the intermediate compounds and reactions with the annotated routes, and is expressed using sensitivity, specificity and relevancy score, which are defined as follows:

$$sensitivity = \frac{tp}{tp+fn}\ ;\ specificity = \frac{tp}{tp+fp}\ ;$$
$$relevancy = \frac{sensitivity+specificity}{2} * smc$$

where:

- tp (true positives): The number of compounds and reactions of the route found which are also present in the annotated route. The first and last compounds and reactions are not considered.
- fp (false positives): The number of compounds and reactions of the route found which are not present in the annotated route.
- fn (false negatives): The number of compounds and reactions of the annotated route which are not present in the route found.
- smc (structural moiety constraint): This value is set to 1 if the route found satisfies the structural moiety constraint, and set to 0 otherwise.

If an extracted route was not identical to an annotated one and contains reactions without atom mapping rules, we manually checked the structural moiety constraint.

The search results are shown in Tab. 2. Searching for the shortest path in the blind search graph delivers poor search results. The average relevancy score is only 0.31. Incorporating atom mapping rules for about two-thirds of the reactions in the graph doubles the search performance up to an relevancy score of 0.61. A further improvement can be achieved by searching for the lightest path in the weighted graph. This approach produces significantly more relevant routes. The relevancy score is 0.77. But only 80% of the routes found fulfill the structural moiety constraint which is clearly better in the atom mapping graph (+ 8%). The search for the lightest path in the weighted atom mapping graph produces the best search performance results. The relevancy score reaches 0.86. Although atom mapping rules are available for only two-thirds of the reactions, 91% of the routes found fulfill the structural moiety constraint, 11% more as for the weighted graph. In the next paragraph, we will demonstrate the search results of the two best approaches using glycolysis as an example. Another example based on tetrahydrofolate biosynthesis, useful for comparing the search results of all approaches, is available in the supplementary material. [6]

[6] http://www-bs2.informatik.uni-tuebingen.de/services/blum/GCB2007/supplementary_material.pdf

Table 2: The search results for 137 experimentally verified biotransformation routes extracted from EcoCyc are shown here. Each row represents one search approach: the blind search graph (bsg), the atom mapping graph (amg), the weighted graph (wg) and the weighted atom mapping graph (wamg). The columns show the average sensitivity (sens), specificity (spec), structural moiety constraint (smc) and relevancy score (rel).

approach	sens	spec	smc	rel
bsg	0.34	0.41	0.47	0.31
amg	0.61	0.66	0.88	0.61
wg	0.82	0.87	0.80	0.77
wamg	0.86	0.87	0.91	0.86

4.2.1 Glycolysis

The biotransformation routes of the glycolysis were searched given D-glucose-6-phosphate as a source and pyruvate as a target as well as EC 5.3.1.9 as a start reaction and EC 2.7.1.40 as an end reaction. Fig. 4A represents the two annotated routes extracted from EcoCyc. The first route (shown as black arrows) contains eight reactions and its weight is 188. Three atoms (ignoring oxygen and hydrogen atoms) are transferred from the source to the target. An additional three atoms, resulting in a second molecule of pyruvate, are transferred via the second route (dark grey arrows). This route contains dihydroxyacetone-phosphate (DHAP) as a further main intermediate which is transformed to D-glyceraldahyd-3-phosphate (GAP). All in all the route contains nine reactions and its overall weight is 203. The first three routes found by the lightest path search in the weighted atom mapping graph are shown in Fig. 4B. The first route (black arrows) requires seven reactions, one resp. two less than the annotated routes. Since the weight is 181, the route will be found before to the annotated routes. Once again three atoms are transferred from D-glucose-6-phosphate to pyruvate. The difference is that the route found needs only one reaction for transforming D-fructose-6-phosphate into GAP. The reaction with the EcoCyc ID RXN0-313 (EC 4.-.-.-) is very interesting since it is not assigned to a pathway in EcoCyc. The enzyme catalyzing this reaction is fructose-6-phosphate aldolase (gene name `fsa`) and was reported as a novel enzyme activity catalyzing an aldol cleavage of D-fructose-6-phosphate [SS01]. The similarity to the standard glycolysis routes is reflected in the relevancy score: 0.84. The second route (dark grey arrows) also detours the annotated transformation of D-fructose-6-phosphate into GAP via fructose-6-phosphate aldolase. The difference is the alternative transformation of 3-phosphoglycerate into 2-phosphoglycerate via glycerate as an additional main intermediate. Three atoms are transferred again but nine reactions are required. The weight of the route is 187 and its relevancy is 0.69. The third route found (light grey arrows) is also shown, because it is equal to the first annotated route. Fig. 4C shows the first two routes found by the lightest path search in the weighted graph. Both routes contain only five reactions and have both the weight 170. However, no atom is transferred to pyruvate. The reaction with the EcoCyc ID 2.7.1.121-RXN is responsible for the failed glycolysis reconstruction. The reaction transferes a phosphate group from GAP to phophoenolpyruvate. In the final reaction (EC 2.7.1.40), the phosphate group is cleaved off so that no atom from GAP could be transferred into pyruvate. The failed reconstruction is reflected by an relevancy score of 0.0.

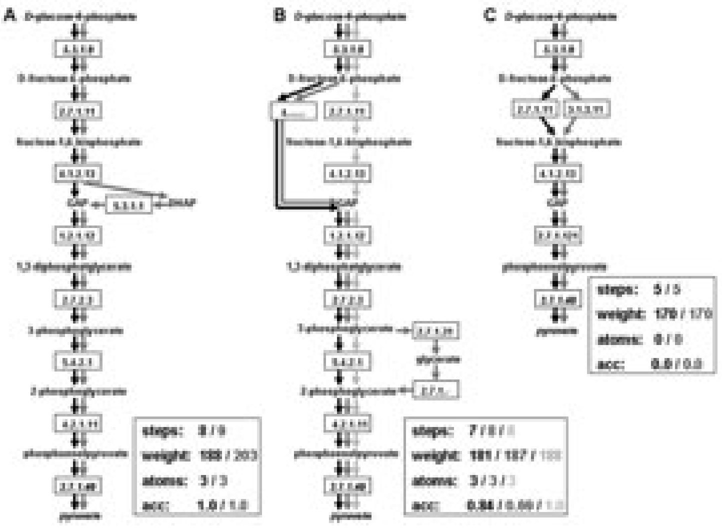

Figure 4: The annotated and the predicted routes for glycolysis, given D-glucose-6-phosphate as a source metabolite, pyruvate as a target and EC 5.3.1.9 as a start reaction and EC 2.7.1.40 as an end reaction. Different colors (black, dark or light grey) represent different routes. For each route, the number of reaction steps, the overall weight, the number of transferred atoms from source to target, and the accuracy (acc) are shown. (A) The annotated routes extracted from EcoCyc. (B) The routes found using the weighted atom mapping graph. (C) The routes found using the weighted graph.

5 Discussion

In this work, we present a novel approach for inferring atom mapping rules from chemical reactions. Fully automated and efficient calculation is the main target and is achieved by introducing pseudo cuts, use of unique SMILES, and EC clustering. Based on atom mapping rules we introduced a novel graph theory-based approach for finding feasible biotransformation routes in metabolic networks. Constraining the lightest path search to those paths where atoms are transferred between source and target nodes yields improved pathway predictions that are more consistent with experimentally verified pathways. Simply by sequentially checking the atom mapping rules of the transforming reactions, problematic routes like those present in the glycolysis or the purine synthesis, can be found quickly. The approach is generally more robust for long pathways or pathways containing typical pool metabolites as intermediates compared to the ordinary lightest path search.

We believe that our approach could be used as a tool that complements existing approaches. It can bridge the gap between the raw genome-scale content stored in pathway databases and well-curated metabolic subnetworks necessary e.g. for applications in metabolic engineering. Fast and intelligent navigation through the network at genome-scale enables a goal-oriented refinement of the search by an iterative addition of constraints. Such constraints contain the identification of side metabolites and the sets of allowed, required or forbidden main intermediates as well as reactions. The search results can help scientists for designing experiments or biotechnologists for defining the constraints necessary for an effecient calculation of stoichiometrically balanced pathways using approaches based on NP-hard convex-analysis [SDF99, SLP00]. Too many constraints at the beginning of the analysis bear the risk of missing relevant pathways. An advanced and iterative graph theory-based search is more robust against the gaps typically present in the networks of newly sequenced organisms [PK02]. It enables the detection of the main routes embedded in network-based pathways even if some balancing side reactions are missing cause of incomplete annotation.

In future work, it should be possible to integrate further relevant information for a better search. Such information could consider the thermodynamic efficiency of the biotransformations or the phylogenetic profiles, expression data as well as subcellular localization of the catalyzing enzymes.

References

[Ak04] Akutsu,T. (2004) Efficient extraction of mapping rules of atoms from enzymatic reaction data., *J. Comput. Biol.*, **11**, 449-62.

[Ar00] Arita,M. (2000) Metabolic reconstruction using shortest paths, *Simulation Practice and Theory*, **8**, 109-125.

[Ar03] Arita,M. (2003) In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism., *Genome Res.*, **13**, 2455-66.

[AS06] Aittokallio,T., Schwikowski,B. (2006) Graph-based methods for analysing networks in cell biology, *Briefings in Bioinformatics*, **7**, 243-55.

[Ca06] Caspi,R., Foerster,H., Carol,A., Fulcher,A., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J., Rhee,S.R., Tissier,C., Zhang,P., Karp,P.D. (2006) MetaCyc: A multiorganism database of metabolic pathways and enzymes., *Nucl. Acids Res.*, **34**, D511-D516

[CCWH06] Croes,D., Couche,F., Wodak,S.J., van Helden,J. (2006) Infering Meaningful Pathways in Weighted Metabolic Networks, *J. Mol. Biol.*, **356**, 222-236.

[Co99] Cordwell,S.J. (1999) Microbial genomes and ”missing” enzymes: redefining biochemical pathways , *Arch. Microbiol*, **172**, 269-279.

[Ep98] Eppstein,D. (1998) Finding the k Shortest Paths, *SIAM Journal on Computing*, **28**, 652-673.

[HWGW02] van Helden,J., Wernisch,L., Gilbert,D., Wodak,S.J. (2002) Graph-based analysis of metabolic networks., *Ernst Schering Res. Found. Workshop*, **38**, 245-74.

[Ka96] Kanehisa,M. (1996) Toward pathway engineering: a new database of genetic and molecular pathways., *Science & Technology Japan*, **59**, 34-38.

[Ke05] Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil, M., Karp,P.D. (2005) EcoCyc: A comprehensive database resource for Escherichia coli, *Nucl. Acids Res.*, **33**, D334-7.

[KS02] Klamt,S., Stelling,J. (2002) Combinatorial complexity of pathway analysis in metabolic networks., *Mol. Biol. Rep.*, **29**, 233-6.

[KSGG03] Klamt,S., Stelling,J., Ginkel,M., Gilles,E.D. (2003) FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps., *Bioinformatics*, **19**, 261-9.

[KZL00] Küffner,R., Zimmer,R., Lengauer,T. (2000) Pathway analysis in metabolic databases via differential metabolic display (DMD), *Bioinformatics*, **16**, 825-836.

[Mo65] Morgan,H.L. (1965) The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service, *J. Chem. Doc.*, **5**, 107-113.

[PK02] Paley,S., Karp,P.D. (2002) Evaluation of computational metabolic-pathway predictions for H. pylori, *Bioinformatics*, **18**, 715-724.

[Ra05] Rahman,S.A., Advani,P., Schunk,R., Schrader,R., Schomburg,D. (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC), *Bioinformatics*, **21**, 1189-93.

[Ro70] Rockafellar,R.T., (1970) Convex analysis, *Princeton Landmarks in Mathematics*, *Princeton University Press*.

[SCS02] Schomburg,I., Chang,A., Schomburg,D. (2002) BRENDA, enzyme data and metabolic information, *Nucl. Acids Res.*, **30**, 47-49.

[SDF99] Schuster,S., Dandekar,T., Fell,D.A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering., *Trends Biotechnol.*, **17**, 53-60.

[SLP00] Schilling,C.H., Letscher,D., Palsson,B. (2000) Theory for the Systematic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective, *J. Theor. Biol.*, **203**, 229-248.

[SS01] Schürmann,M., Sprenger,G. (2001) Fructose-6-phosphate Aldolase is a Novel Class I Aldolase from Escherichia coli and is Related to a Novel Group of Bacterial Transaldolases., *J. Biol. Chem.*, **276**, 11055-11061.

[WWW89] Weininger,D., Weininger,A., Weininger,J.L. (1989) SMILES. 2. Algorithm for Generation of Unique SMILES Notation, *J. Chem. Inf. Comput. Sci.*, **29**, 97-101.