METABOLIC PATHWAY ALIGNMENT (M-Pal) REVEALS DIVERSITY AND ALTERNATIVES IN CONSERVED NETWORKS

YUNLEI LI
DICK DE RIDDER MARCO J. L. DE GROOT MARCEL J. T. REINDERS

Information and Communication Theory Group
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

We introduce a comparative analysis of metabolic reaction networks between different species. Our method systematically investigates full metabolic networks of multiple species at the same time, with the goal of identifying highly similar yet non-identical pathways which execute the same metabolic function, i.e. the transformation of a specific substrate into a certain end product via similar reactions. We present a clear framework for matching metabolic pathways, and propose a scoring scheme which combines enzyme functional similarity with protein sequence similarity. This analysis helps to gain insight in the biological differences between species and provides comprehensive information on diversity in pathways between species and alternative pathways within species, which is useful for pharmaceutical and industrial bioengineering targets. The results also generate hypotheses for improving current metabolic networks or constructing such networks for currently unannotated species.

1 Introduction

The metabolic network of a species represents all known chemical reactions of metabolism within a cell. A single, relatively isolated cascade of such reactions is normally called a metabolic pathway. Most metabolic reactions are catalyzed by specific groups of enzymes. These enzymes are annotated by EC numbers[1], hierarchically organized numbers indicating the type(s) of reaction they catalyze. Studying the metabolic network is a powerful tool to elucidate the cellular machinery. Therefore, it has been an active research field for the last decade[2-13].

Comparing pathways between multiple species provides valuable information to understand evolutionary conservation and variation. Kelly *et al.*[14] align protein interaction networks and predict protein function and interaction using conserved pathways. We extend their alignment concept to the metabolic level, to discover conserved metabolic pathways. Such a pathway transforms a specific substrate into a specific end product via very similar reactions in multiple species. These reactions are similar since they have common substrates and common products. However, they may have different co-substrates or co-products, be catalyzed by different enzymes, need different numbers of reactions to complete the transformation, or reactions may occur in a different order.

Although many comparative analyses at the metabolic level have been performed, little work focuses explicitly on the discrete differences between conserved pathways, and to our knowledge no global search has been carried out yet. For example, Forst *et al.*[5] perform a phylogenetic analysis on four pre-chosen pathways by combining the sequence information of a set of enzymes and gene-coded metabolites in a pathway. Dandekar *et al.*[6] also limit their study, to the glycolysis pathway. As for the similarity measure for matching pathways, Tohsato *et al.*[7] align pathways based on enzyme EC number similarity, discarding information on the involved metabolites. In Clemente *et al.*[8, 9], sets of reactions in multiple pathways are compared, omitting connectivity between the reactions.

Inspired by the PathBLAST algorithm of Kelly *et al.*[14], we propose a novel approach to align metabolic pathways. Our method, Metabolic Pathway ALignment (M-Pal), aligns entire metabolic

networks of different species in order to explore highly conserved pathways. In the resulting aligned pathways, most reactions are identical; the remaining reactions are not identical, yet similar. These conserved pathways are very likely to be essential or efficient pathways. More importantly, our method sheds light on differences between species in the use of non-identical but similar reactions, revealing between-species diversity and within-species alternatives. We introduce *diversity* in a pathway as a term indicating that each species has its own unique mechanism to allow a certain biochemical transformation to take place. If both species share a common reaction, but one of the species has a second, unique reaction to perform the same transformation, then this last transformation forms part of a unique *alternative* pathway.

Diversity and alternatives across species give insight into biological differences between species, provide potential candidate enzymes for bioengineering, and generate hypotheses on missing enzymes or incorrect annotations in current metabolic networks. Moreover, the resulting pathways give more options in pathway engineering and constructing metabolic networks for unannotated species. Finally, this method unites reactions in isolated metabolisms into a large network, relating reactions with upstream substrates and downstream products which might be elusive if we only look at a subset of the network.

We apply M-Pal to *Saccharomyces cerevisiae* and *Escherichia coli*, and find 2518 short conserved pathways. In each conserved pathway, 4−5 reactions from one species are aligned with similar reactions from another species. Among the results, ~1500 pathways are diverse or contain unique alternative enzyme activities. We categorize the differences between pathways and refine the search result by scoring each pathway according to functional and sequence similarity of the enzymes involved. This scoring scheme enables us to focus on highly conserved pathways with similar enzymes. We show that a number of metabolic annotations can be attached to each of the resulting pathways, demonstrating the strength of our systematic search in unearthing novel cross-links in metabolic networks.

We describe M-Pal in detail in Section 2. The results are presented and discussed in Section 3. Section 4 ends with some conclusions and an outlook to further work.

2 Methods

Since we seek to investigate diversity and alternatives in highly conserved metabolic pathways, we align the pathways from two species into a conserved pathway in a rather strict way. That is, we align two pathways only if most of the involved reactions in this two species use similar enzymes to catalyze common substrates into common products, introducing only a limited amount of freedom into the alignment. More specifically, let P_1 and P_2 denote two metabolic pathways in two species containing reactions $[R_{11}, R_{12}\ldots R_{1L}]$ and $[R_{21}, R_{22}\ldots R_{2L}]$, respectively. P_1 and P_2 can be aligned into a conserved pathway only if the individual reactions are aligned in the right order. That is, R_{11} is aligned with R_{21}, R_{12} is aligned with R_{22} *etc.* until R_{1L} is aligned with R_{2L}. We call each pair of matching reactions, e.g. R_{11} and R_{21}, a *building block*.

Given the restrictions mentioned above, we propose an efficient matching mechanism which constructs all building blocks first, and then assembles them into pathways of a desired length, taking reaction directions into account. After the aligned pathways are obtained, we compute an enzyme similarity score for each aligned pathway. In this way, we eventually get a list of conserved pathways, ordered by this score.

This sequential procedure of matching and scoring (see Figure 1) ensures the search for all matching pathways is complete and allows for a flexible scoring function. The exhaustive search results can be pre-computed and, as scoring is performed separately, no potential match will be missed because of prematurely discarding a pathway in the search. Our method is explained in detail in the remainder of this section.

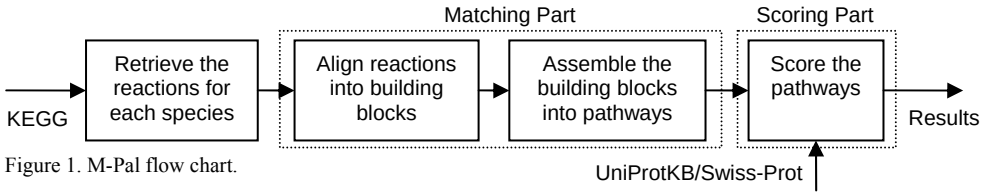

Figure 1. M-Pal flow chart.

2.1 Reaction Retrieval

We obtained the general reaction definitions from Release 42.0 of the KEGG LIGAND composite database[15], updated on May 14, 2007. For each species, we acquired the subset of reactions present in that species, together with the EC numbers and ORF names of the enzymes which catalyze each reaction, from the KEGG/XML and KEGG/PATHWAY databases.

In M-Pal, reactions are represented as a combination of the classic "enzyme-centric" and "compound-centric" representations. Thus, a reaction is represented by all elements involved: metabolites, (a group of) enzyme(s), and its direction. Figure 2a gives an example. To allow us to compare reactions from different species, we plot them next to each other, with the matching substrate or product in the same row. Sometimes, a single reaction and a series of reactions connected in tandem may share common substrates and products. This introduces "*gaps*", indicating that the number of reactions to transform the specific substrates into the specific products differs between species. Figure 2b illustrates this: one reaction from *S. cerevisiae* and two reactions from *E. coli* form a "gap".

2.2 Building Block Alignment

Two reactions R_{1l} and R_{2l} can be aligned to form a building block when they have a common substrate and a common product, and at least one pair of enzymes (one from each species) share functional similarity such that the first two digits of their EC numbers are the same. Note that a reaction can be catalyzed by a group of enzymes, which may have multiple EC numbers. By allowing some variation, we introduce a number of *building block types* (see Figure 3). If R_{1l} and R_{2l} are identical, i.e. the same reaction is present in both species, the resulting building block is called "*identical*" (*i*). If R_{1l} and R_{2l} are different reactions, because of different co-substrates or co-products according to the definition in Section 2.1, they form a "*direct*" building block (*d*). To incorporate alternative pathways, evolutional diversity and annotation errors, we also allow one "mismatch" or one "gap" in a building block. Thus, in an "*enzyme mismatch*" building block (*em*), the first two digits of the EC numbers of the enzymes involved are not the same. The building blocks containing one "gap" are "*direct-gap*" (*dg*) and "*enzyme mismatch-gap*" (*eg*). Furthermore, we include "*enzyme crossover match*" building blocks (*ec*) to accommodate possible variation in the order of the catalyses: there are two reactions in each species sharing common substrates and end products with the EC numbers of the first and second reaction in one species being similar to those of the second and first reaction in the other species, respectively.

To summarize, the reaction alignment method described above results in six types of building blocks, each containing one or two reactions from each species. Note that 26 "current metabolites"[10,13], listed below[a], were excluded from consideration as common substrate or product to avoid finding large numbers of trivial conserved pathways.

[a] ATP, ADP, UTP, UDP, GTP, GDP, AMP, UMP, GMP, NAD, NADH,NADP, NADPH, Acetyl-CoA, CoA, Propanoyl-CoA, L-Glutamine, L-Glutamate, 2-Oxoglutarate, CTP, CDP, CMP, H_2O, CO_2, NH_2, Phosphate.

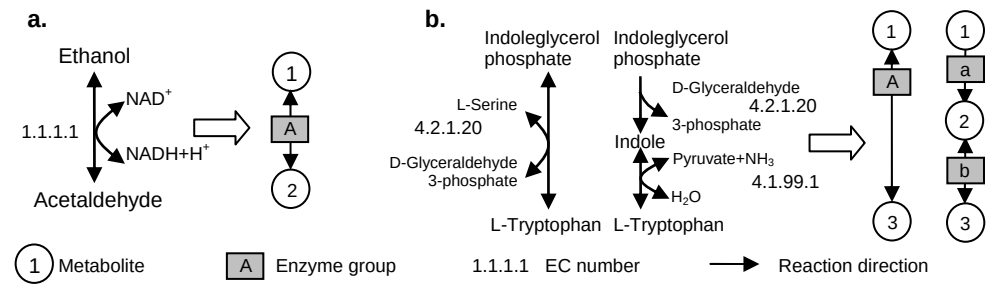

Figure 2. Reaction representation. **a**. Illustration of two representations of reactions in our method. **b**. One reaction from *S. cerevisiae* (on the left) and two reactions from *E. coli* (on the right) share a common substrate (Indoleglycerol phosphate) and product (L-Tryptophan). This situation forms one "gap", i.e. the difference in the number of reactions to transform Indoleglycerol phosphate into L-Tryptophan is one.

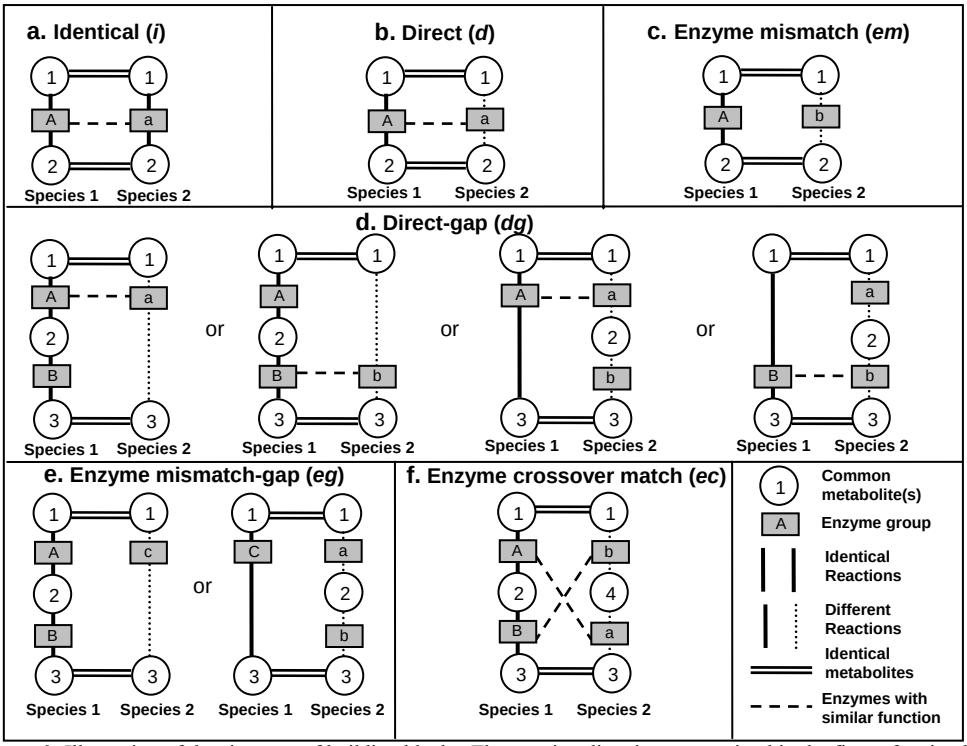

Figure 3. Illustration of the six types of building blocks. The reaction directions are omitted in the figure for simplicity. A dashed link is drawn between two groups of enzymes if they share the same first two digits of their EC numbers.

2.3 Pathway Assembly

Next, we focus on finding conserved short acyclic pathways. We only assemble four building blocks into a pathway, ensuring that one reaction does not appear more than once in a pathway. Moreover, we demand that out of these four building blocks, at least three must be of the type "identical" or "direct", representing the conserved part of the pathway. Only a single building block of type "enzyme mismatch", "direct-gap", "enzyme mismatch-gap" or "enzyme crossover match" is allowed in a pathway. Abbreviations are used to denote the pathway composition of building blocks regardless of the order, e.g. "*i*-*i*-*i*-*d*" indicates a pathway with three reactions of type "identical" and one of type "direct", in any order. In total, there are 21 such compositions possible for pathway alignment. These are used as 21 *pathway categories* in the discussion of our results.

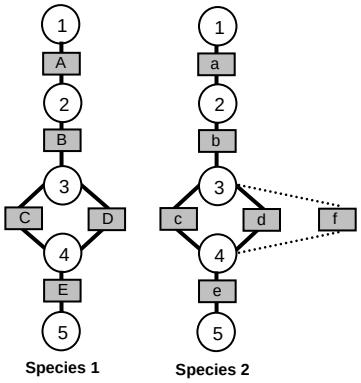

Figure 4. Illustration of the removal of redundant pathways. See Figure 3 for legends. Six possible pathway alignments can be induced in this example (each reaction is represented by the corresponding enzyme groups): ① Reactions A-B-C-E of species 1 with a-b-c-e of species 2, obtaining an *i-i-i-i* alignment. ② Reactions A-B-D-E of species 1 with a-b-d-e of species 2, obtaining an *i-i-i-i* alignment. ③ Reactions A-B-C-E of species 1 with a-b-d-e of species 2, obtaining an *i-i-x-i* alignment, where *x* indicates one of the five non-"identical" building block types. This alignment is redundant with ① and ②. ④ Reactions A-B-D-E of species 1 with a-b-c-e of species 2, which is also redundant with ① and ②. ⑤ Reactions A-B-C-E of species 1 with a-b-f-e of species 2, obtaining an *i-i-x-i* alignment.. This is a novel alternative pathway, since reaction f is unique in species 2, hence *i-i-i-i* alignment is impossible. ⑥ Reactions A-B-D-E of species 1 with a-b-f-e of species 2 also is a novel pathway. In the end, four aligned pathways are obtained: ①, ②, ⑤ and ⑥.

Table 1. Transformation of the total functional similarity $\sum_{b\in P} E_f(b)$ into the score S_f.

$\sum_{b\in P} E_f(b)$	16	15.5	15	14.5	14	13.5	13	12	11	10	8
S_f	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

Table 2. Transformation of the total sequence similarity $\sum_{b\in P} E_s(b)$ into the score S_s.

$\sum_{b\in P} E_s(b)$	$(0, 10^{-80})$	$[10^{-80}, 10^{-60})$	$[10^{-60}, 10^{-40})$	$[10^{-40}, 10^{-20})$	$[10^{-20}, 10^{-10})$	$[10^{-10}, 10^{-6})$
S_s	0	0.1	0.2	0.3	0.4	0.5
$\sum_{b\in P} E_s(b)$	$[10^{-6}, 10^{-2})$	$[10^{-2}, 100)$	[100, 200)	[200, 300)	$[300, \infty)$	
S_s	0.6	0.7	0.8	0.9	1.0	

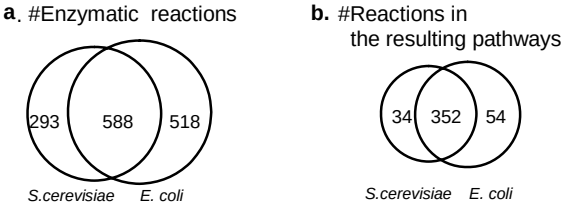

Figure 5. Venn diagrams showing **a.** the total number of enzymatic reactions in the two species and **b.** the number of reactions involved in the results.

Table 3. The number of each of the six types of building blocks.

Type	Identical (*i*)	Direct (*d*)	Direct-gap (*dg*)	Enzyme mismatch (*em*)	Enzyme crossover match (*ec*)	Enzyme mismatch-gap (*eg*)
#Building blocks	516	116	108	27	40	52
#Building blocks in the resulting pathways	352	67	64	11	12	29

To enhance the informativeness of our resulting set of pathways, we remove some redundant pathways. First, building blocks whose substrate and product are identical in one species (after removing current metabolites) will not be selected to construct a pathway. Furthermore, we reduce the redundancy in the result by enforcing uniqueness in choosing the building blocks of the five types

other than "identical", see Figure 4. A non-"identical" building block can be chosen only if it contains at least one reaction absent in one of the species. This is because if all reactions in the building block are present in both species, two building blocks of type "identical" will already be constructed. Consequently, any other combinations of these reactions are redundant. Conversely, a reaction unique to one species provides an interesting novel alternative pathway.

2.4 Scoring Function

Two factors indicate the extent to which an aligned pathway is conserved. One is the pathway category, i.e. the building block composition. For instance, we consider an "*i*-*i*-*i*-*d*" pathway to be more conserved than an "*i*-*i*-*i*-*dg*" pathway. The other factor is enzyme similarity, which we evaluate here based on functional similarity (EC numbers) and sequence similarity. Since they are not fully correlated, we integrate them to introduce a more informative measure of true orthology. In the following, we explain how to calculate functional similarity and sequence similarity of a building block, followed by their integration.

Given a building block containing one reaction from each species, enzyme functional similarity E_f is taken to be the maximum number of digits of EC numbers that the two groups of enzymes share. This is a simple and straightforward manner to measure enzyme functional similarity[12, 17, 18], since EC numbers form a functional hierarchy. Although more complex methods exist[7-9], their validity is still under research. Let the EC numbers in the reaction for species 1 be EC_{11}, EC_{12}, …, EC_{1m}, and for species 2 EC_{21}, EC_{22}, …, EC_{2n}, we count the number of shared digits for each possible pair of EC numbers, and use the maximum as the functional similarity E_f for this building block. For "direct-gap" and "enzyme mismatch-gap" building blocks, for which one group of enzymes should be compared to two groups of enzymes, we compute E_f for both pairs of groups, and choose the larger E_f. For "enzyme crossover match" building blocks, E_f is taken to be the averaged value of the crossover enzyme group comparisons.

For the sequence similarity E_s between two reactions, we take the minimum BLAST *E*-value between all possible enzyme pairs. For "direct-gap" and "enzyme mismatch-gap" building blocks, E_s is computed between the two groups of enzymes which have the larger E_f. For "enzyme crossover match" building blocks, E_s is averaged. BLAST (version 2.2.15) is performed with $e = 100$ on the protein sequences in UniProtKB / Swiss-Prot Release 51.6.

After computing the E_f and E_s scores for all building blocks in a pathway, we sum all E_fs in a pathway and transform the result into a score $S_f \in [0, 1]$; likewise for all E_ss in the pathway to obtain $S_s \in [0, 1]$. Tables 1 and 2 detail these transformations. Since the original values of E_f and E_s have very different ranges, this transformation step actually scales these two measures into the same range in a sensible way, so that they are comparable and easy to combine. The intervals in the transformation tables are chosen to reflect our objective in finding conserved pathways with similar enzymes: high functional similarity values are examined in more detail in the score. For sequence similarity, we focus on the traditional cutoff value 10^{-2} for weak sequence similarity[14], thus the intervals around 10^{-2} are smaller than those for high sequence similarities. We do not restrict ourselves to highly similar sequences because our main interest is to reveal the alternatives and diversities in the pathways. Since the maximum value for E_s is 100 (due to the parameter setting used for BLAST), the intervals for $S_s \geq 0.8$ indicate the number of building blocks with very dissimilar enzyme sequences.

Finally, the two scores are summed so as to combine the functional and sequence similarity:

$$S(P) = S_f\left(\sum_{b\in P} E_f(b)\right) + S_s\left(\sum_{b\in P} E_s(b)\right) \quad (1)$$

in which b denotes a building block and P denotes an aligned pathway. The lower this score, the more similar the enzymes in P are.

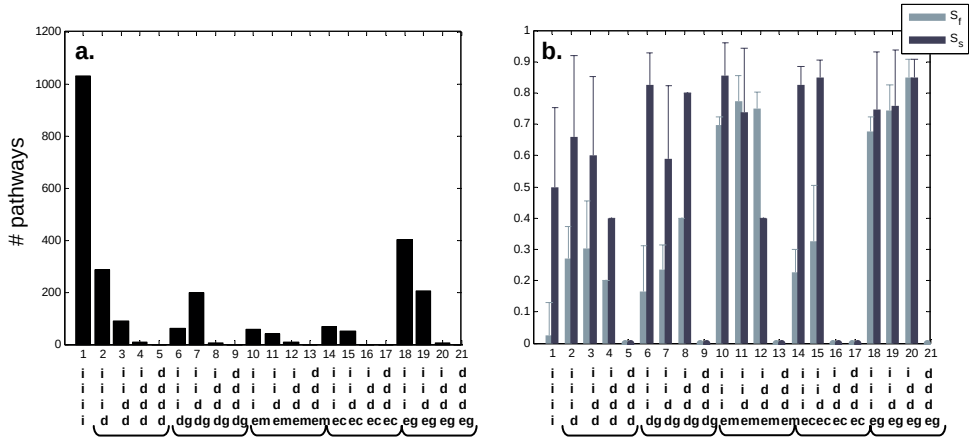

Figure 6. **a.** Total number of pathways in 21 pathway categories. Note that long conserved pathways may result in multiple short overlapping pathways. **b.** The average enzyme functional similarity score and sequence similarity score of each pathway category. Whiskers indicate standard deviations.

3 Results and Discussion

Of 881 enzymatic reactions in *S. cerevisiae* and 1106 in *E. coli*, 588 reactions are present in both species (Figure 5a). Based on the total of 1399 unique reactions, six types of building blocks are assembled into 2518 unique pathways of length 4. Figure 5b shows the number of reaction involved in the resulting pathways. Table 3 summarizes the number of building blocks of each type found. There results indicate that the reactions and building blocks in the resulting pathways reasonably cover all available reactions and building blocks, demonstrating the strength of our systematic search.

For each pathway category containing a specific composition of building blocks, the total number of resulting pathways is shown in Figure 6a and their average functional similarity score S_f and sequence similarity score S_s are shown in Figure 6b. As shown in Figure 6a, ~1000 completely conserved pathways of type "*i-i-i-i*" are found. Not surprisingly, their enzyme sequences are highly similar, with BLAST *E*-values ranging from 10^{-10} to 10^{-6} on average. The pathway with the best score, 0, is depicted in Figure 7a. However, the variance of the sequence similarity score is also large, indicating that some reactions in these pathways do not have enzymes with similar sequences. This might arise because of different specificity, horizontal gene transfer, gene fusions, or the fact that only subunits of the enzymes are the same.

We also found ~1500 highly conserved pathways which contain some diversity between both species or unique alternatives within one species. Each of these pathways has a building block of type "direct", "direct-gap", "enzyme mismatch", "enzyme mismatch-gap", or "enzyme crossover match". Examples are given in Figure 7b-7f. These pathways are of great interest in bioengineering as they manifest the hidden information about pathway diversity and alternatives, which will not be found if we only look at a subset of the metabolic network in one species.

The results are useful in many applications. First, some resulting pathways suggest a more exact EC number annotation of their enzymes is possible and call for detailed comparison of the enzymes. For example, the enzymes in the pathways of type "*i-d-d-em*" in Figure 6b have dissimilar EC numbers, but their sequences are actually very similar (low S_s and high S_f). They might be incorrectly annotated, since they both transform a common substrate into a common product. Another example is given in Figure 7c, in which the enzymes with EC number 4.2.1.20 in *E. coli* (trpA and trpB) could also be annotated as 4.1.2.8, which is the α-subunit of 4.2.1.20. Comparing the enzymes in alternative

pathways in different species can also be beneficial to understand their structural difference and relationship. In Figure 7c for instance, the two enzymes in *E. coli*, 4.2.1.20 and 4.1.99.1, might be different subunits of the enzyme 4.2.1.20 in *S. cerevisiae*. The same can be observed in Figure 6b, where the sequence similarity in the pathways with "*dg*" is generally worse than in those with "*d*" only, implying that the enzymes in "*dg*" are only subunits of the corresponding enzymes in "*d*".

Second, the results can help to understand diversity in metabolism and evolution. Reactions which are unique to one species are highlighted in Figure 7. Investigation of the biological difference between the two species is expected to explain their uniqueness. Further, we can project the knowledge to a new species. For instance, if the new species has the enzymes which catalyze a unique reaction of *S. cerevisiae*, then probably they are very closely related in the phylogenetic tree, and therefore share more common properties. Nevertheless, the revealed diversity might be an artifact of current metabolic network databases. Therefore it is recommended to examine whether the other species also has this unique enzyme, or whether some enzymes (and reactions) are missing in the pathways with "gaps". Another interesting result which might be worthy of further research is shown in Figure 6b, for the group containing enzyme crossover match building blocks (*ec*). Although the crossover enzymes have similar functions, their sequences are very dissimilar. Possible reasons could be that the enzymes have different substrate specificities, or the intermediate substrates are very different. They could also have been isoenzymes in parallel pathways, having become specialized to one species in evolution.

Third, the unique alternative pathways revealed by M-Pal provide potential candidate enzymes for bioengineering. Certain natural enzymes can be removed or changed so that we can choose between different alternative pathways, or enforce the reaction direction to produce the product of our interest. In the pathway shown in Figure 7c, *E. coli* has two alternative pathways to transform Indoleglycerol phosphate into L-tryptophan, one being reversible (catalyzed by 4.2.1.20) and the other one reported to be irreversible (catalyzed by 4.2.1.20 and 4.1.99.1). If the enzymes of 4.2.1.20 in the irreversible pathway are indeed also possibly annotated as 4.1.2.8, we can remove the 4.2.1.20 enzyme activity to enforce the direction towards producing tryptophan, which is an essential amino acid in human nutrition[16].

Finally, our results provide additional opportunities to construct the metabolic networks for currently unannotated species. As discussed above, our method points out possible missing enzymes and suggests related enzymes in well-studied species. The alternative pathways also provide more possibilities for optimizing the network to fit the found enzymes and reactions better.

4 Conclusions

The systematic search of M-Pal associates different parts of metabolic networks with each other and combines information from multiple species to discover diversity and alternatives in highly conserved pathways. The results shed light on the small differences found in the conserved pathways and provide useful information for many applications. Gene knock-out experiments can be performed to test our hypotheses, and the essentiality of the resulting pathways should be examined.

Our research is still at an early stage, and can be refined in a number of ways. Possible extensions include increasing the freedom in the alignment, e.g. allowing for more gaps or mismatches, further separated crossover matches, and longer pathways. This implies the search algorithm will have to become more sophisticated, as exhaustive enumeration will become infeasible. Next, the scoring function can be modified to prefer certain types of alignment. Non-identical metabolites could be included in the matching, implying a need for a compound similarity measure to be added to the scoring function. The enzyme sequence similarity measure could also be refined using protein domain information. The current scoring mechanism assumes functional and sequence similarity is

equally important. Weights could be added to model a trade-off between the two[8]. The scoring function itself could be enhanced by using a probabilistic framework such as in Kelly *et al.*[14], allowing us to look for relatively rather than absolutely conserved pathways and to attach a *p*-value to the pathways found. Other possible enhancements to the score are to take reversibility of reactions and the presence of isoenzymes into account.

Currently, this method is performed on two species only and is expected to give more informative results if applied on species not closely related. An extension could be to apply M-Pal on multiple species, at different evolutionary distances. We expect that larger differences will be found as evolutionary distance increases. The results will give insight to understand evolution and specialization, and will provide new building blocks and alternatives for pathway engineering. Applying this method for prediction of unannotated genes will be of great value. Finally, by relating different sets of enzymes in different species to a common metabolic function, this work provides an infrastructure based on which the regulatory factors can be associated, and functional hypothesis can be generated.

Acknowledgments

The authors would like to thank Rogier J. P. van Berlo, Domenico Bellomo, Wouter van Winden and Peter van Nes for their help and the constructive discussions. This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

References

1. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). http://www.chem.qmul.ac.uk/iubmb/enzyme/.
2. R. Overbeek, N. Larsen, W. Smith, N. Maltsev and E. Selkov, *Gene* **191**, GC1–GC9 (1997).
3. S. Schuster, D.A. Fell and T. Dandekar, *Nat. Biotechnol.* **18**, 326-332 (2000).
4. C. Francke, R.J. Siezen and B. Teusink, *Trends Microbiol*. **13**(11), 550-558 (2005).
5. C.V. Forst and K. Schulten, *J. Mol. Evol.* **52**, 471-489 (2001).
6. T. Dandekar, S. Schuster, B. Snel, M. Huynen and P. Bork, *Biochemical J.* **343**(1), 115-124 (1999).
7. Y. Tohsato, H. Matsuda, and A. Hashimoto, *Proc. of the 8th Inter. Conf. on Intel. Sys. for Mol. Biol.*, 376-383 (2000).
8. J.C. Clemente, K. Satou and G. Valiente, *Bioinformatics* **23**(2): e110-e115 (2006).
9. J.C. Clemente, K. Satou and G. Valiente, *Genome Informatics* **17**(2), 46-56 (2006).
10. D. Zhu and Z.S. Qin, *BMC Bioinformatics* 6-8 (2005).
11. A.G. Malygin, *Biochemistry Moscow* **69**(12), 1379-1385 (2004).
12. M. Heymans and A.K. Singh, *Bioinformatics* **19**, i138-i146 (2003).
13. H.W. Ma and A.P. Zeng, *Bioinformatics* **19**, 270-277 (2003).
14. B.P. Kelley, R. Sharan, R.M. Karp, T. Sittler, D.E. Root, B.R. Stockwell and T. Ideker, *Proc. Natl. Acad. Sci. U S A* **100**, 11394–11399 (2003).
15. S. Goto, T. Nishioka, and M. Kanehisa, *Bioinformatics* **14**, 591-599 (1998).
16. R.J. Wurtman, W.J. Shoemaker and F. Larin, *Proc. Natl. Acad. Sci. U S A* **59**(3), 800-807 (1968).
17. K. Pawlowski, L. Jaroszewski, L. Rychlewski and A. Godzik, *Pac. Symp. Biocomput.* 42-53 (2000).
18. Z. Li, S. Zhang, Y. Wang, X. Zhang and L. Chen, *Bioinformatics* **23**(13), 1631-1639 (2007).

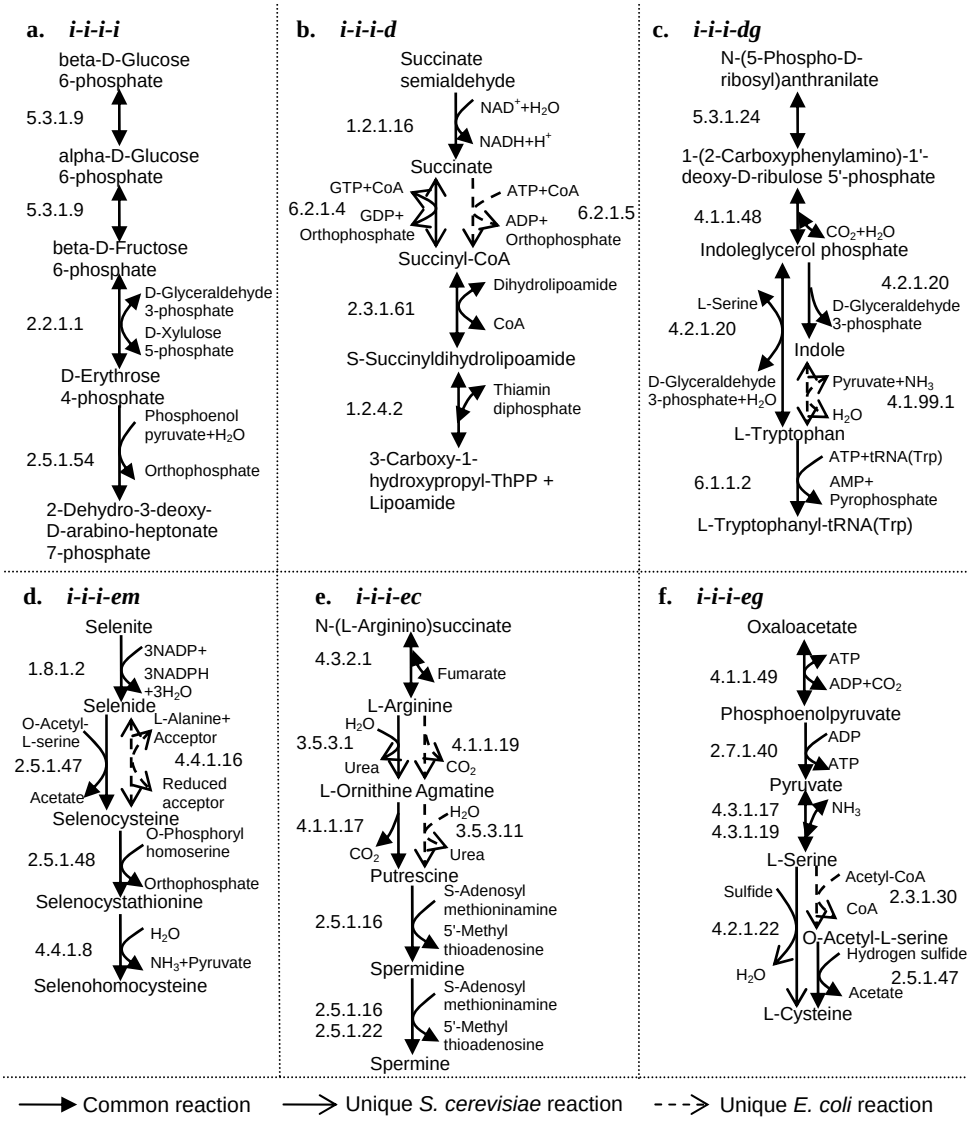

Figure 7. The pathways with the best scores in categories 1, 2, 6, 10, 14, and 18 of Figure 6. If two reactions share a common metabolite, this common metabolite is drawn only once for conciseness. **a**. The pathway with the best score ($S = 0$) in the results. It has an "*i-i-i-i*" alignment. Involved metabolic annotations include: glycolysis/gluconeogenesis; pentose phosphate pathway; starch and sucrose metabolism; and phenylalanine, tyrosine and tryptophan biosynthesis. **b.** One of the pathways with the best score ($S_f = 0.2$, $S_s = 0.1$) within category "*i-i-i-d*". Involved metabolic annotations include: glutamate metabolism; tyrosine metabolism; butanoate metabolism; citric acid cycle (TCA cycle); propanoate metabolism; reductive carboxylate cycle (CO_2 fixation). **c.** The pathways with the best score ($S_f = 0$, $S_s = 0.5$) within category "*i-i-i-dg*". Involved metabolic annotations include: phenylalanine, tyrosine and tryptophan biosynthesis; tryptophan metabolism; aminoacyl-tRNA biosynthesis; benzoxazinone biosynthesis; and nitrogen metabolism. **d**. One of the pathways with the best score ($S_f = 0.7$, $S_s = 0.6$) within category "*i-i-i-em*". Involved metabolic annotations include: selenoamino acid metabolism. **e**. The pathways with the best score ($S_f = 0.2$, $S_s = 0.7$) within category "*i-i-i-ec*". Involved metabolic annotations include: urea cycle and metabolism of amino groups; alanine and aspartate metabolism; arginine and proline metabolism; and beta-alanine metabolism. **f**. One of the pathways with the best score ($S_f = 0.7$, $S_s = 0.3$) within category "*i-i-i-eg*". Involved metabolic annotations include: TCA cycle; pyruvate metabolism; carbon fixation; glycolysis/gluconeogenesis; purine metabolism; glycine, serine and threonine metabolism; cysteine metabolism; and sulfur metabolism.