

Finding Conserved and Non-Conserved Reactions Using a Metabolic Pathway Alignment Algorithm

José C. Clemente¹

clemente@jaist.ac.jp

Kenji Satou¹

ken@jaist.ac.jp

Gabriel Valiente²

valiente@lsi.upc.edu

¹ School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST), 1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan

² Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain

Abstract

Using a metabolic pathway alignment method we developed, we studied highly conserved reactions in different groups of organisms and found out that biological functions vital for each of the groups are effectively expressed in the set of conserved reactions. We also studied the metabolic alignment of different strains of three bacteria and found out several non-conserved reactions. We suggest that these reactions could be either misannotations or reactions with a relevant but yet to be specified biological role, and should therefore be further investigated.

Keywords: metabolic pathway, alignment, conserved reaction, biological annotation

1 Introduction

With the increasing amount of information on metabolic pathways stored in repositories such as KEGG [14], it is of greater importance to develop methods that allow us to understand properties of metabolism in different species. In our previous work [5, 6] we developed an algorithm for metabolic pathway alignment based on the similarity of reactions, metabolites and enzymes.

We have previously utilized our algorithm to obtain metabolic similarity values among organisms. Using those similarity values, we can reconstruct phylogenetic trees depicting the common metabolic history of a group of organisms, or find suitable model organisms in the study of diseases related to certain metabolic conditions [6]. In this work, we will further expand the range of possible applications of our method by investigating the significance behind the pathway alignments that we produce.

Let us first consider how our approach produces an alignment between different reactions in the metabolism of two organisms. Borrowing terminology from sequence alignment, we will define the alignment of reactions as a *perfect match*, a *substitution* or a *gap*. In a perfect match, both reactions are composed of the same set of metabolites and enzymes, and our algorithm would score their similarity as 1. Figure 1 has several examples of perfect matches between reactions in the TCA cycle of *A. fulgidus* and *L. innocua* marked in blue/boldface (00268, 00344, 00351, 00412, 01082 and 01899)¹. Substitutions occur when the reactions are not exactly the same, but share some compounds or enzymes. In such case, the similarity value will be greater than 0 but less than 1. Figure 1 contains three such substitutions: reactions 00342, 00405 and 01197 in *A. fulgidus* are respectively substituted by 01082, 00412 and 00268 in *L. innocua* (alignment marked in green/dashed lines). Finally, gaps occur when a reaction in one of the organisms cannot be mapped with similarity greater than 0 in the other organism. Figure 2 presents an example of a gap: reaction 01698 in *L. innocua* has similarity 0 to any reaction in *A. fulgidus*, and therefore creates a gap in the alignment. Biologically, a gap

¹We will use KEGG notation to identify reactions (00009: $2H_2O_2 \Leftrightarrow O_2 + 2H_2O$), organisms (*hsa*: *Homo sapiens*) and pathways (00020: TCA cycle)

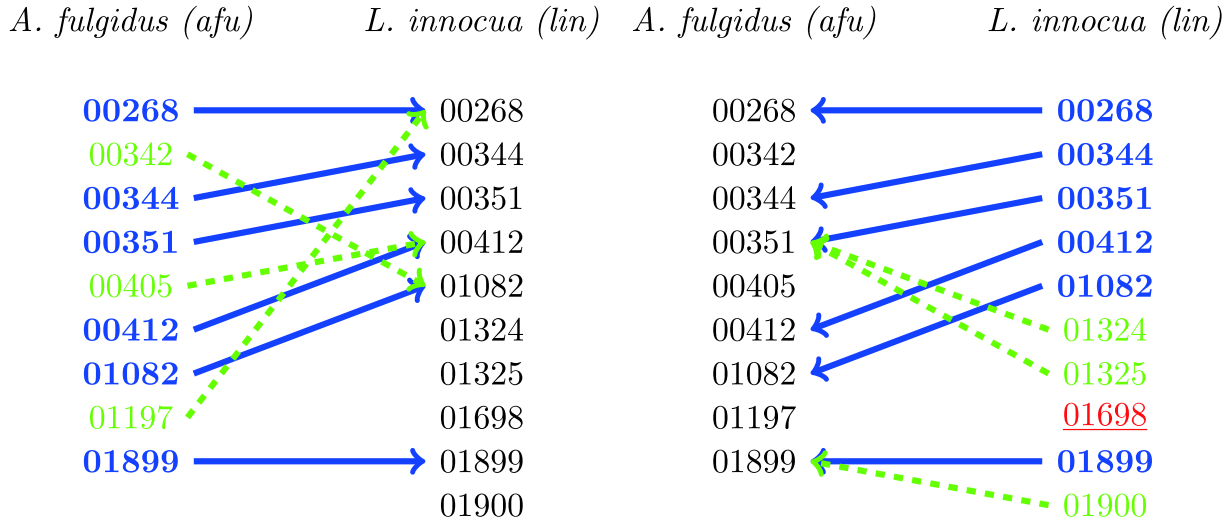


Figure 1: Alignment of TCA cycle from *A. fulgidus* with *L. innocua*. Reactions 00342, 00405 and 01197 in *afu* (green/dashed) are substituted in *lin* for 01082, 00412 and 00268 with similarity < 1; rest of reactions are perfectly aligned, similarity = 1 (blue/bold).

Figure 2: From *L. innocua* to *A. fulgidus*: 01324, 01325 and 01900 (green/dashed) are substituted for 00351, 00351 and 01899 with similarity < 1. 01698 (red/underlined) cannot be mapped with similarity > 0. Remaining reactions are perfectly aligned (blue/bold).

between species *A* and *B* represents a reaction which, in the course of evolution, has been gained by *A* (“insertion” event) or lost by *B* (“deletion”). Generalizing for a set of organisms $\{A, B, C, \dots, Z\}$, a reaction is said to be a gap when it is present in one of the organisms and cannot be aligned with any of the remaining organisms with similarity greater than 0.

Notice that because we are aligning unordered sets of reactions instead of ordered sequences of nucleotides or amino acids, our algorithm produces a directed alignment since the mapping of reactions is not symmetric. Reaction 00342 in *A. fulgidus* is aligned with reaction 01082 in *L. innocua* (Figure 1), but the opposite is not true since reaction 01082 in *L. innocua* gets aligned with 01082 in *A. fulgidus* (Figure 2). Although for simplicity we have described an alignment between two organisms only, the same method can be applied to a set of species by calculating all their respective alignments.

In this work, we will focus on perfect matches and gaps. Perfect matches are interesting since they represent highly conserved reactions in the metabolism of different organisms. Vital biological processes in a group of related species (taxa such as bacteria or archaea) should be conserved and expressed by a significant number of reactions in all the organisms of the group. We will validate this hypothesis by studying perfect matches among bacteria, archaea and eukarya in Section 3.1.

Gap reactions, on the other hand, are interesting since they imply the complete absence of a certain group of reactions in one of the organisms being compared. More specifically, if the organisms being compared are known to be similar we would not expect to find many gaps in the alignment of their metabolism. We will test this hypothesis by studying the alignment of a set of strains (genetic variants of an organism) in Section 3.2.

2 Materials and Methods

All data used in our experiments was obtained from KEGG release 39.0 (July 2006). For each experiment, we selected a set of organisms and retrieved all shared pathways which contained at least one reaction. Given two pathways, $\mathbf{P} = (\mathbf{R})$ and $\mathbf{Q} = (\mathbf{S})$, where \mathbf{R}, \mathbf{S} are sets of enzymatic reactions, our algorithm proceeds by aligning each reaction in one pathway to the most similar reaction in the

other pathway. In order to do so, we calculate three subsets: $\mathbf{R} \cap \mathbf{S}$, $\mathbf{R} \setminus \mathbf{S}$ and $\mathbf{S} \setminus \mathbf{R}$. By definition, $\forall r \in \mathbf{R} : r \in \mathbf{R} \cap \mathbf{S} \Rightarrow \exists s \in \mathbf{S} : s \in \mathbf{R} \cap \mathbf{S}$. All pairs (r, s) have similarity 1 and are therefore aligned by our algorithm. For each reaction $t \in \mathbf{R} \setminus \mathbf{S}$, our algorithm looks for the most similar reaction $u \in \mathbf{S}$, and analogously for each reaction $v \in \mathbf{S} \setminus \mathbf{R}$, we align it with the most similar reaction $w \in \mathbf{R}$. Notice that because our approach is based on set operations, the order in which we process the reactions does not affect the final alignment. Moreover, whenever reactions present in a pathway are missing in other pathways, our method can still perform a significant alignment by looking for reactions which have some degree of similarity. Equation (1) summarizes our metabolic pathway similarity measure.

$$\text{sim}(\mathbf{P}, \mathbf{Q}) = \frac{1}{|\mathbf{R} \cup \mathbf{S}|} \left(|\mathbf{R} \cap \mathbf{S}| + \sum_{R \in \mathbf{R} \setminus \mathbf{S}} \max_{S \in \mathbf{S}} \text{sim}(R, S) + \sum_{S \in \mathbf{S} \setminus \mathbf{R}} \max_{R \in \mathbf{R}} \text{sim}(R, S) \right). \quad (1)$$

Similarity of reactions $R = (\mathbf{C}, \mathbf{E})$ and $S = (\mathbf{D}, \mathbf{F})$ is calculated analyzing the similarity of the respective sets of compounds (\mathbf{C}, \mathbf{D}) and enzymes (\mathbf{E}, \mathbf{F}) , using a parameter α to determine the relative weight of compound and enzyme similarity in the assessment of reaction similarity (see Eq. 2). The similarity of compounds is 1 if they are completely similar, or 0 if they are dissimilar. The similarity of two enzymes is calculated using their enzyme hierarchy similarity measure [18].

$$\begin{aligned} \text{sim}(R, S) = & \frac{1 - \alpha}{|\mathbf{C} \cup \mathbf{D}|} \left(|\mathbf{C} \cap \mathbf{D}| + \sum_{C \in \mathbf{C} \setminus \mathbf{D}} \max_{D \in \mathbf{D}} \text{sim}(C, D) + \sum_{D \in \mathbf{D} \setminus \mathbf{C}} \max_{C \in \mathbf{C}} \text{sim}(C, D) \right) \\ & + \frac{\alpha}{|\mathbf{E} \cup \mathbf{F}|} \left(|\mathbf{E} \cap \mathbf{F}| + \sum_{E \in \mathbf{E} \setminus \mathbf{F}} \max_{F \in \mathbf{F}} \text{sim}(E, F) + \sum_{F \in \mathbf{F} \setminus \mathbf{E}} \max_{E \in \mathbf{E}} \text{sim}(E, F) \right). \end{aligned} \quad (2)$$

Experiments described in Section 3.1 utilized the obtained alignment to select all reactions that were aligned with similarity 1 (perfect alignments) for all organisms in a domain: bacteria, archaea, eukarya, mammals, and plants. With this set of highly conserved reactions, we then calculated the percentage of each pathway that was conserved as the number of conserved reactions appearing in the pathway divided by the total number of reactions of the pathway.

Experiments in Section 3.2 used results from the pathway alignment to obtain gap reactions in any strain, i.e. reactions that could not be mapped with similarity greater than 0 to any of the remaining strains, thus inducing a gap in the alignment. Finally, we calculated the number and type of enzymes involved in such reactions by counting the number of reactions in which each enzyme appeared.

3 Results

3.1 Highly Conserved Reactions

Given a set of organisms, we expect reactions conserved with high similarity to perform functions related to biologically relevant processes in the set. We performed experiments in a group of bacteria, archaea and eukarya and their shared metabolic pathways (Supplementary Material, Table 6. Ideally, reactions conserved with high similarity only in one of the kingdoms (i.e. in all species in that kingdom) and not in the others should reflect some property exclusive to that specific kingdom.

As seen in Table 1(a), Bacteria were very clearly characterized. Most of the reactions conserved with high similarity in bacteria belong to the fatty acid biosynthesis pathway (00061), with 64% of the total number of reactions in the pathway being conserved in all bacteria. The peptidoglycan biosynthesis pathway (00550) also has over 37% of its reactions highly conserved in bacteria. Remnant pathways were conserved in lower proportions.

Archaea had a significant number of conserved reactions in the phenylalanine pathway (00400), with 80% of the total number of reactions in the pathway appearing in all archaea. Pyrimidine metabolism (00860) was partially conserved, with 27% of its reactions present (Table 1(b)).

Results in eukaryota, as seen in Table 1(c), show how oxidative phosphorylation (00190), carbon fixation (00710), and glyoxylate and dicarboxylate metabolism (00630) pathways have over 60% of

3.2.1 *Streptococcus pyogenes*

S. pyogenes is a Gram-positive bacterium associated with different diseases through the release of toxins, as in scarlet fever and toxic shock syndrome. We used 11 strains from *S. pyogenes* currently stored in KEGG, and aligned 64 of their shared pathways which contained at least one reaction.

As it can be seen in Table 3, strains recently introduced in KEGG tend to have a larger number of reactions that cannot be aligned with reactions in any other strain. For instance, the serotype M3 strains *sph*, *spi*, *spj* and *spk* [3], introduced in KEGG in 2006, contain many reactions that cannot be aligned and therefore appear as gaps in other strains. These four strains can have all their reactions aligned with no gaps among them, which indicates a high degree of similarity.

Strains *spg* and *spz* are also serotype M3, but they have some gaps in their alignment to previous M3 strains. Serotype M18 strain *spm*, associated with acute rheumatic fever outbreaks, shares the gaps with both *spg* and *spz*. Strains *spy* and *spz* are serotype M1. The more recent strain *spz* (2005) has no gaps, while *spy* (2001) contains several ones. Serotype M28 strain *spb* (2005) also presents no gaps.

We also investigated which enzymes catalyze the gap reactions (Supplementary material, Table 10). Most of these enzymes are still not fully characterized, as indicated by the “-” symbol in their EC identifier.

Table 3: Gap reactions: *S. pyogenes*.

Strain	Year	Serotype	Gap reactions
<i>sph</i>	2006	M3	(none)
<i>spi</i>	2006	M3	(none)
<i>spj</i>	2006	M3	(none)
<i>spk</i>	2006	M3	(none)
<i>spz</i>	2005	M1	(none)
<i>spb</i>	2005	M28	(none)
<i>spa</i>	2004	M6	00118 01906 03540 03674 03730 03937 04008 04172 04594 04732 04885 04906 05202 05601 05602 06369 06906 06925
<i>spz</i>	2003	M3	00118 00148 00501 01906 01966 02518 02520 03113 03234 03540 03674 03730 03937 04008 04142 04172 04360 04594 04732 04885 04906 04937 04938 05202 05601 05602 06021 06097 06192 06198 06369 06404 06729 06906 06925
<i>spm</i>	2002	M18	00118 00148 00501 01906 01966 02518 02520 03113 03234 03540 03674 03730 03937 04008 04142 04172 04360 04594 04732 04885 04906 04937 04938 05202 05601 05602 06021 06097 06192 06198 06369 06404 06729 06906 06925
<i>spg</i>	2002	M3	00118 00148 00501 01906 01966 02518 02520 03113 03234 03540 03674 03730 03937 04008 04142 04172 04360 04594 04732 04885 04906 04937 04938 05202 05601 05602 06021 06097 06192 06198 06369 06404 06729 06906 06925
<i>spy</i>	2001	M1	00118 00148 00501 01906 02518 02520 03113 03234 03540 03544 03545 03674 03730 03937 04008 04142 04172 04360 04594 04732 04885 04906 05202 05601 05602 06021 06097 06192 06198 06369 06404 06729 06906 06925

3.2.2 *Escherichia coli*

E. coli, a bacterium present in the lower intestine of mammals, is a common bacterial model organism responsible for different kinds of infections, as well as for food-poisoning in contaminated meat. We used 7 strains stored in KEGG, and aligned 82 shared pathways containing at least one reaction.

As in the experiment with *S. pyogenes* (Section 3.2.1), strains of *E. coli* recently incorporated in KEGG seem to have far less gap reactions than older ones: *eci* has only 11 gaps, and *ecp* 20. The number of gaps does not strictly correspond with the year of publication, though: strain *ecc* was introduced after *ecj*, *ece* or *eco*, but has a larger number of gaps.

Strains representing the toxigenic *E. coli* O157:H7 (*ece* and *ecs*) share their set of gap reactions. K-12 type strains *eco* (MG1655) and *ecj* (W3110), on the other hand, do not fully share their gap reaction sets. These two strains were published respectively in 1997 and 2006 (the web entry in KEGG's organism list incorrectly marks 2001), with the second publication correcting some of the entries of the first one [11]. Inspection of enzymes involved in gap reactions in *E. coli* shows again a large predominance of not fully-characterized enzymes, as it can be seen in Table 11 (Supplementary material).

Table 4: Gap reactions: *E. coli*.

Strain	Year	Gap reactions
<i>eci</i>	2006	02000 02002 04986 06405 06782 06783 06784 06785 06786 06787 06920
<i>ecp</i>	2006	02000 02002 04910 04986 05049 05615 05617 05625 05644 05645 05646 05647 06397 06782 06783 06784 06785 06786 06787 06858
<i>ecc</i>	2002	00069 01452 01453 01719 01966 02000 02002 02383 02518 02912 03674 03730 03811 03937 03955 04008 04131 04172 04306 04732 04826 04857 04885 04895 04906 04910 04916 04917 05001 05448 05449 05504 05505 05602 05623 06367 06369 06396 06397 06398 06400 06405 06406 06407 06735 06736 06782 06783 06784 06785 06786 06787 06853 06854 06905 06906 06907 06914 06920 06925 06935
<i>ecj</i>	2001	00069 01452 01453 01719 02518 02912 03317 03674 03730 03811 03937 03955 04008 04131 04142 04172 04306 04313 04360 04375 04732 04809 04813 04826 04857 04885 04895 04910 04916 04917 05001 05448 05449 05504 05505 05602 05623 06369 06396 06397 06398 06400 06405 06406 06407 06735 06736 06853 06854 06905 06906 06907 06914 06920 06925 06935
<i>ece</i>	2001	00069 01452 01453 01719 01966 02383 02518 02912 03674 03730 03811 03937 03955 04008 04131 04142 04172 04306 04360 04375 04515 04732 04784 04809 04813 04826 04857 04885 04895 04910 04916 04917 05001 05448 05449 05504 05505 05602 05623 06369 06396 06397 06398 06400 06405 06406 06407 06735 06736 06853 06854 06905 06906 06907 06920 06925
<i>ecs</i>	2001	00069 01452 01453 01719 01966 02383 02518 02912 03674 03730 03811 03937 03955 04008 04131 04142 04172 04306 04360 04375 04515 04732 04784 04809 04813 04826 04857 04885 04895 04910 04916 04917 05001 05448 05449 05504 05505 05602 05623 06369 06396 06397 06398 06400 06405 06406 06407 06735 06736 06853 06854 06905 06906 06907 06920 06925
<i>eco</i>	1997	00069 01452 01453 01966 02518 02912 03674 03730 03811 03937 03955 04131 04172 04306 04732 04826 04857 04885 04895 04906 04910 05001 05448 05449 05504 05505 05602 05623 06367 06369 06396 06397 06398 06400 06405 06406 06853 06854 06906 06920 06925

3.2.3 *Staphylococcus aureus*

S. aureus is a Gram-positive bacterium that can cause a wide range of diseases, such as pneumonia, meningitis, or toxic shock syndrome. The bacterium has become resistant to many antibiotics in the last years, and may be fatal in cases of severe infections. KEGG currently contains data on 9 different strains, and for our experiments we used 68 of their shared pathways containing at least one reaction.

As in Sections 3.2.1 and 3.2.2, recently published strains tend to have less gap reactions. In this case, *saa* (2006) and *sab* (2005) have the least number, closely followed by *sao* (2006). Methicillin-resistant (MRSA) and methicillin-susceptible (MSSA) strains *sar* and *sas* were published together [13] and have an identical set of gap reactions. The three strains for which the primary repository is the NITE/Juntendo database (*sau*, *sav* and *sam*) have the same gap set as well. A detailed revision of the papers associated with the first two strains [16] and the last one [1] reveals that the group that published the three strains is in fact the same. Finally, the strain *sac* has a significant number of gaps despite its recent publication. Most of the enzymes present in gap reactions in *S. aureus* were again not fully-characterized (Supplementary material, Table 12).

Table 5: Gap reactions: *S. aureus*.

Strain	Year	Gap reactions
<i>saa</i>	2006	01966 03113 03234 03540 03544 03545 04142 04360 04937 04938 06405 06920
<i>sao</i>	2006	00118 01452 01453 02383 03811 04254 04306 04313 04594 04826 05001 05118 05623 06398 06400 06405 06406 06413 06916 06920 06926
<i>sac</i>	2005	00118 01452 01453 01966 02383 03730 03811 04131 04306 04594 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
<i>sab</i>	2005	00118 01452 01453 02383 03811 04306 04594 04826 05001 06372 06398 06400 06406
<i>sar</i>	2004	00118 01452 01453 01719 01966 02383 02528 03113 03234 03730 03811 04131 04142 04306 04360 04594 04809 04813 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
<i>sas</i>	2004	00118 01452 01453 01719 01966 02383 02528 03113 03234 03730 03811 04131 04142 04306 04360 04594 04809 04813 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
<i>sam</i>	2002	00118 01452 01453 01719 01966 02383 03730 03811 04131 04306 04594 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
<i>sau</i>	2001	00118 01452 01453 01719 01966 02383 03730 03811 04131 04306 04594 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
<i>sav</i>	2001	00118 01452 01453 01719 01966 02383 03730 03811 04131 04306 04594 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935

4 Discussion

Understanding which metabolic processes are fundamental for a group of organisms can be of great utility [2]. In this paper, we have presented an approach based on a metabolic pathway alignment algorithm [5, 6] to detect which pathways have a significant number of reactions conserved with high similarity. Results show how can we establish which are the most relevant metabolic pathways for taxa such as bacteria, archaea, mammals or plants.

We also investigated how an alignment among strains of three different bacteria reveals a significant number of reactions that belong exclusively to some of the strains, and therefore produce a gap in the alignment with other strains. We found out such gap reactions to be much more common in species recently introduced in KEGG. Also, most gap reactions seem to be catalyzed by enzymes which are not yet fully determined. These evidences seem to imply that these reactions are, in fact, misannotations.

KEGG is probably the most complete resource on metabolic information publicly available and, despite its high quality standards, it is to be expected that misannotations are introduced due to the automated nature of the annotation process. Methods to detect such errors are well-known in the case of sequence data, and also for metabolic data [8, 9]. The work presented in this paper provides another step towards detecting such misannotations.

4.1 Conserved Reactions in a Group of Organisms

By studying conserved reactions in a group of organisms, we expected to find biological processes which are relevant for some of the organisms. The fatty acid biosynthesis pathway was significantly conserved in bacteria when compared to the other two kingdoms (Table 1(a)). It is known that enzymes of fatty acid biosynthesis represent excellent targets for drugs against bacteria [4, 17], therefore the relevance of this pathway and its high number of conserved reactions in our experiments.

Our method was also able to detect the relevance of the peptidoglycan biosynthesis pathway (Table 1(a)). These polymers enable bacteria to withstand high osmotic pressures. Highly conserved in bacteria, peptidoglycans have no parallel in eukaryota, and disruption of the peptidoglycan pathway can be lethal for bacteria [19].

We also found out the phenylalanine pathway was significantly conserved in Archaea (Table 1(b)). Phenylalanine is an essential alpha amino acid that cannot be synthesized by animals, which have to obtain it from their diet. It is produced from prephenate, an intermediate on the shikimate pathway. Interestingly, although this pathway is present in archaea, bacteria, fungi, and plants, there are two kinds of shikimate kinases: archaeal and non-archaeal. Archaeal shikimate kinases are, by sequence similarity, distantly related to homoserine kinases (GHMP kinase domain superfamily) [7], while all non-archaeal shikimate kinases (the typical form) belong to the (structurally unrelated) NMP kinase domain superfamily [15]. We found 6 entries for enzymes related to shikimate in KEGG: 1.1.1.25, 1.1.1.282, 1.14.13.36, 2.3.1.133, 2.5.1.19 and 2.7.1.71. Except for 1.14.13.36 and 2.3.1.133, which are not annotated to any pathway, all enzymes belong to the phenylalanine pathway.

Eukaryotes were more complex to analyze. The first experiment (see Table 1(c)) shows a series of conserved pathways which cannot be directly linked to fundamental metabolic processes in all eukarya: carbon fixation (00710), for instance, is vital only for plants. It is not clear either how the glyoxylate and dicarboxylate metabolism (00630) are of relevance to the selected eukaryotes. After detailed inspection of results we found out reactions conserved in these pathways are actually annotated to more than one metabolic pathway. Additionally, the definition of a metabolic pathway in KEGG does not necessarily correspond to the traditional understanding expressed in the literature, and KEGG pathways are known to overlap. Several reactions in the Calvin cycle are also present in the pentose phosphate pathway, which might explain the high value for conserved reactions in the carbon fixation pathway among all eukaryotes.

The most conserved pathway in eukaryota was oxidative phosphorylation, which is the final path-

way of cellular respiration after glycolysis and the cytric acid cycle. Its basic function is to transfer electrons from NADH or FADH₂ to molecular oxygen through protein complexes located in the mitochondria. Arguably, the fact that mitochondria are not found in bacteria or archaea [12] could explain that metabolic pathways related to activity in the mitochondria would be relevant to eukaryotic organisms only. Tables 2(a) and 2(b) show the relevance of this pathway both in mammals and plants, which is consistent with this hypothesis. Additionally, Table 2(b) also describes how carbon fixation in photosynthetic organisms is effectively conserved for plants, as would be expected.

It should be noticed that genome sequences for nearly all bacteria and archaea present in KEGG are fully determined, which only happens for a few of the eukarya. Given that metabolic information in KEGG is obtained by analyzing sequence data from GenBank, this might explain the poorer quality of results for eukaryota.

4.2 Non-Conserved Reactions in a Group of Strains

Non-conserved reactions (gaps) are those reactions which are contained in one strain but cannot be aligned with similarity greater than 0 with any other reaction in different strains. This means that the set of enzymes and compounds contained in such reactions do not appear in any other reaction. Since strains are variants of one single organism, gap reactions are worth studying because they represent a set of enzymes and compounds unique to a certain strain. We argue that such reactions are most probably an annotation mistake.

Sections 3.2.1, 3.2.2 and 3.2.3 describe how strains recently included in KEGG contain a significant number of reactions that appear as gaps in older strains. Specifically, strains introduced in years 2006 and 2005 in *S. pyogenes* (*sph*, *spi*, *spj*, *spk*, *spz*, *spb*; Table 3), *E. coli* (*eci*, *ecp*; Table 4), and *S. aureus* (*sca*, *sco*, *scc*, *scb*; Table 5) contained the smallest number of gaps and the largest number of reactions that are gaps in older strains. This suggests that such strains contain misannotated reactions due to the automated nature of the annotation process. Although our experiments are limited to alignment among strains of an organism, it is reasonable to expect similar results for any organism in general. Recent results confirm that KEGG contains a certain number of misannotations [8, 9] and therefore, data about recently included species should be considered as tentative until further validation.

Additionally, most gap reactions are catalyzed by enzymes not fully determined (Supplementary material, Tables 10, 11 and 12). Analysis of the references associated with each of the strains in KEGG did not explain why these enzymes should be annotated to those strains, which further implies that these reactions might indeed be misannotations. Even for gap reactions with fully determined enzymes, such as 3.1.3.73 in *S. pyogenes*, 1.18.1.4 in *E. coli*, and 3.1.3.73 in *S. aureus*, we found out that such enzymes do not have an associated gene annotated to them. EC numbers are not always supported by experimental validation, and they can introduce errors in metabolic pathway repositories as those presented in this work (see [10] for further details on errors associated with EC numbers).

Acknowledgments

The authors would like to thank Profs. Susumu Goto and Minoru Kanehisa (KEGG), Profs. Takashi Gojobori and Kazuho Ikeo (NIG), and Prof. Satoru Miyano (University of Tokyo) for their valuable comments. We also appreciate the suggestions of three anonymous reviewers, which greatly improved the quality this manuscript.

References

- [1] Baba, T., Takeuchi, F., Kuroda, M., Yuzawa, H., Aoki, K., Oguchi, A., Nagai, Y., Iwama, N., Asano, K., Naimi, T., Kuroda, H., Cui, L., Yamamoto, K., and Hiramatsu, K., Genome and vir-

- ulence determinants of high virulence community-acquired MRSA, *The Lancet*, 359(9320):1819–1827, 2002.
- [2] Balmer, Y., Vensel, W.H., Cai, N., Manieri, W., Schurmann, P., Hurkman, W.J., and Buchanan, B.B., A complete ferredoxin/thioredoxin system regulates fundamental processes in amyloplasts, *Proc. Natl. Acad. Sci. USA*, 103(8):2988–2993, 2006.
 - [3] Beres, S.B., Richter, E.W., Nagiec, M.J., Sumby, P., Porcella, S.F., DeLeo, F.R., and Musser, J.M., Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group *A. streptococcus*, *Proc. Natl. Acad. Sci. USA*, 103(18):7059–7064, 2006.
 - [4] Campbell, J.W. and Cronan, J.E., Bacterial fatty acid biosynthesis: Targets for antibacterial drug discovery, *Ann. Rev. Microbiol.*, 55(1):305–332, 2001.
 - [5] Clemente, J., Satou, K., and Valiente, G., Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology, *Genome Informatics*, 16(2):45–55, 2005.
 - [6] Clemente, J., Satou, K., and Valiente, G., Phylogenetic reconstruction from non-genomic data, *Bioinformatics*, in press.
 - [7] Daugherty, M., Vonstein, V., Overbeek, R., and Osterman, A., Archaeal shikimate kinase, a new member of the GHMP-kinase family, *J. Bacteriology*, 183(1):292–300, 2001.
 - [8] Félix, L. and Valiente, G., Efficient validation of metabolic pathway databases, *Proc. 6th Int. Symp. Computational Biology and Genome Informatics*, 1209–1212, 2005.
 - [9] Félix, L. and Valiente, G., Validation of metabolic pathway databases based on chemical substructure search, *Biomolecular Engineering*, in press.
 - [10] Green, M.L. and Karp, P.D., Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers, *Nucleic Acids Res.*, 33(13):4035–4039, 2005.
 - [11] Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B.L., Mori, H., and Horiuchi, T., Highly accurate genome sequences of *E. coli* K-12 strains MG1655 and W3110, *Mol. Syst. Biol.*, 2(2), 2006.
 - [12] Henze, K. and Martin, W., Evolutionary biology: Essence of mitochondria, *Nature*, 426(6963):127–128, 2003.
 - [13] Holden, M.T.G., *et al.*, Complete genomes of two clinical *Staphylococcus aureus* strains: Evidence for the rapid evolution of virulence and drug resistance, *Proc. Natl. Acad. Sci. USA*, 101(26):9786–9791, 2004.
 - [14] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, 32:D277–D280, 2004.
 - [15] Krell, T., Coggins, J.R., and Laphorn, A.J., The three-dimensional structure of shikimate kinase, *J. Mol. Biol.*, 278(5):983–997, 1998.
 - [16] Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., Cui, L., Oguchi, A., Aoki, K., and Nagai, Y., Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*, *The Lancet*, 357(9264):1225–1240, 2001.
 - [17] Muñoz-Elías, E.J. and McKinney, J.D., *Mycobacterium tuberculosis* isocitrate lyases 1 and 2 are jointly required for *in vivo* growth and virulence, *Nature Medicine*, 11(6):638–644, 2005.

- [18] Tohsato, Y., Matsuda, H., and Hashimoto, A., A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy, *Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology*, 376–383, 2000.
- [19] Walsh, C., *Antibiotics: Actions, Origins, Resistance*, ASM Press, 2003.