# Reconstruction of Phylogenetic Relationships from Metabolic Pathways Based on the Enzyme Hierarchy and the Gene Ontology

**José C. Clemente**[1] **Kenji Satou**[1] **Gabriel Valiente**[2]
clemente@jaist.ac.jp ken@jaist.ac.jp valiente@lsi.upc.edu

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[2] Department of Software, Technical University of Catalonia, E-08034 Barcelona, Spain

**Abstract**

There has been much interest in the structural comparison and alignment of metabolic pathways. Several techniques have been conceived to assess the similarity of metabolic pathways of different organisms. In this paper, we show that the combination of a new heuristic algorithm for the comparison of metabolic pathways together with any of three enzyme similarity measures (hierarchical, information content, and gene ontology) can be used to derive a metabolic pathway similarity measure that is suitable for reconstructing phylogenetic relationships from metabolic pathways. Experimental results on the Glycolysis pathway of 73 organisms representing the three domains of life show that our method outperforms previous techniques.

**Keywords:** metabolic pathway, similarity measure, hierarchical similarity, information content, gene ontology, clustering, phylogenetic reconstruction

## 1 Background

The understanding of evolutionary relationships among species has recently shifted from more conventional studies that exploit polymorphism information in DNA or protein sequence to assess the phylogenetic relationship among species, to new studies aimed at assessing the evolution of complete biological processes. There has been much interest in the structural comparison and alignment of metabolic pathways, and several techniques have been conceived to assess the similarity of such pathways for different organisms. In the comparative analysis of metabolic pathways, pathways from different genomes are aligned upon similar enzymes, substrates, and products [2, 12, 13].

There has also been much interest in the phylogenetic analysis of metabolic pathways, and several techniques have been conceived to extend the similarity assessment of these pathways into phylogenies for different organisms. Previous phylogenetic analyses have been based on the number of common enzymes between two organisms [4, 5], on profiles of the presence and absence of the various metabolic pathways [9], and on the topology of the underlying enzyme-enzyme relational graphs [6]. The produced phylogenies have often been evaluated by comparing them against the NCBI taxonomy [15], which is based on Ribosomal RNA 16S sequences, and the best results so far have been obtained by Heymans and Singh [6]. The phylogenetic analysis of metabolic pathways has also lead to the identification of conserved pathway modules in different organisms [16].

The assessment of structural similarity of metabolic pathways of different organisms involves both a graph representation of a metabolic pathway and a similarity measure between individual reactions, enzymes, and compounds present in the pathway. Metabolic pathways are represented as directed hypergraphs, with the compounds and enzymes as nodes and the reactions activated by the enzymes

as hyperarcs [3]. For instance, the directed hypergraph for the Citric Acid Cycle pathway in the bacterium *E.coli* consists of 35 nodes and 18 hyperarcs.

A more abstract representation, called the enzyme-enzyme relational graph, was used in [6, 11], where nodes represent enzymes and arcs represent compounds shared between successive reactions. For instance, the enzyme-enzyme relational graph for the Citric Acid Cycle pathway in the bacterium *E.coli* consists of only 14 nodes and 23 arcs.

Similarity of reactions can be assessed by measuring the similarity of the enzymes activating them. Hierarchical similarity [13] and information content similarity [12, 13] are two commonly used enzyme similarity measures based on the enzyme hierarchy [14].

## 2 Results

We have used a new heuristic algorithm for the comparison of metabolic pathways with three enzyme similarity measures (hierarchical, information content, and gene ontology) to assess the structural similarity of metabolic pathways of different organisms. Gene ontology similarity is a new enzyme similarity measure we developed which is based on the shortest path between enzymes in a mapping of the enzyme hierarchy to the Gene Ontology. We have determined the structural similarity of the Glycolysis pathway across the 73 organisms studied in [6], namely: those organisms in KEGG [8] which have at least three enzymes present in the Glycolysis pathway (see Table 1).

We have clustered these organisms based on metabolic pathway similarity, using average-link hierarchical clustering [7]. In order to evaluate the effectiveness of our method, we have compared the produced phylogenies with both the NCBI taxonomy (restricted to the 73 organisms in Table 1) and the phylogeny for the same organisms in [6]. One of the produced phylogenies can be contrasted in Figure 1 with the NCBI taxonomy [15] and the phylogenetic tree in [6, Fig. 2]. Besides, phylogenies obtained using hierarchical, information content, and gene ontology enzyme similarity can be contrasted in Figure 2.

The new measure of metabolic pathway similarity is parameterized by the relative weight of compound similarity to enzyme similarity in the assessment of reaction similarity, and we have also studied the influence of this weight parameter on metabolic pathway similarity on the Glycolysis pathway across the aforementioned 73 organisms. The pattern of variation changes with the underlying enzyme similarity measure: while hierarchical enzyme similarity yields the highest metabolic pathway similarity for a weight of about 30%, gene ontology enzyme similarity yields the highest metabolic pathway similarity for a weight between 45% and 65%, and information content enzyme similarity yields the highest metabolic pathway similarity for a weight of about 40% (see Figure 3).

Further, in order to facilitate comparison of results with previous work [6], we have also used the *cousins* tool [17] to compute similarity measures between phylogenies. We were unable to reproduce the 0.19 similarity claimed in [6, Table 2] for any parameter of the cousins tool, though, and have adopted the one that gives the closest result (similarity up to second cousins of the trees) for our experiments.

The similarity measures of our technique and Heymans and Singh's technique [6], obtained by using the NCBI taxonomy as a standard, for the Glycolysis pathway across 73 organisms are shown in Table 2. Our method outperforms the best previous technique by a significant margin.

## 3 Discussion

Glycolysis is a metabolic pathway that serves, among other functions, to generate high-energy ATP molecules. This pathway has been thoroughly studied in the literature, being highly conserved in the genetic code and occurring in most species. Because of these characteristics, similarity among different organisms can be studied by analyzing the similarity of their respective Glycolysis pathways.

Table 1: Organisms studied, classified by domain (E: Eukaryota, B: Bacteria, A: Archaea), together with their identifier in the NCBI taxonomy.

| Code | Organism | Domain | NCBI |
|---|---|---|---|
| ATH | *A.thaliana* | E | 3702 |
| CEL | *C.elegans* | E | 6239 |
| DME | *D.melanogaster* | E | 7227 |
| HSA | *H.sapiens* | E | 9606 |
| MMU | *M.musculus* | E | 10090 |
| RNO | *R.norvegicus* | E | 10116 |
| SCE | *S.cerevisiae* | E | 4932 |
| SPO | *S.pombe* | E | 4896 |
| AAE | *A.aeolicus* | B | 224324 |
| ANA | *Anabaena* | B | 103690 |
| ATC | *A.tumefaciens_C* | B | 176299 |
| ATU | *A.tumefaciens* | B | 176299 |
| BHA | *B.halodurans* | B | 86665 |
| BME | *B.melitensis* | B | 224914 |
| BSU | *B.subtilis* | B | 224308 |
| CAC | *C.acetobutylicum* | B | 272562 |
| CCR | *C.crescentus* | B | 190650 |
| CJE | *C.jejuni* | B | 192222 |
| CMU | *C.muridarum* | B | 243161 |
| CPA | *C.pneumoniae_AR39* | B | 115711 |
| CPJ | *C.pneumoniae_J138* | B | 138677 |
| CPN | *C.pneumoniae* | B | 115713 |
| CTR | *C.trachomatis* | B | 272561 |
| DRA | *D.radiodurans* | B | 243230 |
| ECE | *E.coli_O157* | B | 155864 |
| ECJ | *E.coli_J* | B | 83333 |
| ECO | *E.coli* | B | 83333 |
| ECS | *E.coli_O157J* | B | 83334 |
| FNU | *F.nucleatum* | B | 190304 |
| HIN | *H.influenzae* | B | 71421 |
| HPJ | *H.pylori_J99* | B | 85963 |
| HPY | *H.pylori* | B | 85962 |
| LIN | *L.innocua* | B | 272626 |
| LLA | *L.lactis* | B | 272623 |
| LMO | *L.monocytogenes* | B | 169963 |
| MLE | *M.leprae* | B | 272631 |
| MLO | *M.loti* | B | 266835 |

| Code | Organism | Domain | NCBI |
|---|---|---|---|
| MTC | *M.tuberculosis_CDC1551* | B | 83331 |
| MTU | *M.tuberculosis* | B | 83332 |
| NMA | *N.meningitidis_A* | B | 122587 |
| NME | *N.meningitidis* | B | 122586 |
| PAE | *P.aeruginosa* | B | 208964 |
| PMU | *P.multocida* | B | 272843 |
| RPR | *R.prowazekii* | B | 272947 |
| RSO | *R.solanacearum* | B | 267608 |
| SAU | *S.aureus_N315* | B | 158879 |
| SAV | *S.aureus_Mu50* | B | 158878 |
| SCO | *S.coelicolor* | B | 100226 |
| SME | *S.meliloti* | B | 266834 |
| SPN | *S.pneumoniae* | B | 170187 |
| STM | *S.typhimurium* | B | 99287 |
| STY | *S.typhi* | B | 220341 |
| SYN | *Synechocystis* | B | 1148 |
| TMA | *T.maritima* | B | 243274 |
| TTE | *T.tengcongensis* | B | 273068 |
| VCH | *V.cholerae* | B | 243277 |
| XCC | *X.campestris* | B | 190485 |
| XFA | *X.fastidiosa* | B | 160492 |
| YPE | *Y.pestis* | B | 214092 |
| AFU | *A.fulgidus* | A | 224325 |
| APE | *A.pernix* | A | 56636 |
| HAL | *Halobacterium* | A | 64091 |
| MAC | *M.acetivorans* | A | 188937 |
| MJA | *M.jannaschii* | A | 243232 |
| MMA | *M.mazei* | A | 192952 |
| MTH | *M.thermoautotrophicum* | A | 187420 |
| PAB | *P.abyssi* | A | 272844 |
| PAI | *P.aerophilum* | A | 13773 |
| PFU | *P.furiosus* | A | 186497 |
| SSO | *S.solfataricus* | A | 273057 |
| STO | *S.tokodaii* | A | 273063 |
| TAC | *T.acidophilum* | A | 273075 |
| TVO | *T.volcanium* | A | 50339 |

The phylogenetic trees obtained in this work include several biologically relevant clusters. In Figure 1 (right), we can appreciate how archaea organisms are clustered in two groups: MTH, MJA, PFU and PAB in the first cluster (with the thermococci PFU and PAB forming a subcluster), and APE, PAI, TVO, TAC, SSO, HAL, AFU, STO, MMA, and MAC in the second (with methanosarcinales MMA and MAC forming a subcluster). Regarding bacteria, the Chlamedya CPN, CPJ, CPA, CTR, and CMU are also clustered together, as well as the proteobacteria gamma STY, STM, YPE, ECJ, ECS, ECO, and ECE (with the Escherichia in one subcluster and the Salmonella STY and STM in another one). Firmicutes bacillus appear in two main clusters: one for the Bacillales LIN, LMO, BHA, SAU, and SAV, and another one for the Lactobacillales and Clostridia TTE, SPN, and LLA. The proteobacteria delta are also clustered in one group, HPJ, HPY, and CJE.

Despite the goodness of our approach to find relevant clusters, detailed inspection shows we are still far from a fully significative taxonomy. Heymans and Singh's method (Figure 1, center), is capable of clustering together all the proteobacteria alpha but one (MLO). The eukaryota also appear grouped into two clusters: mammals (HSA, RNO, MMU), and the remaining eukaryota (DME, SCE, CEL, SPO, and ATH).

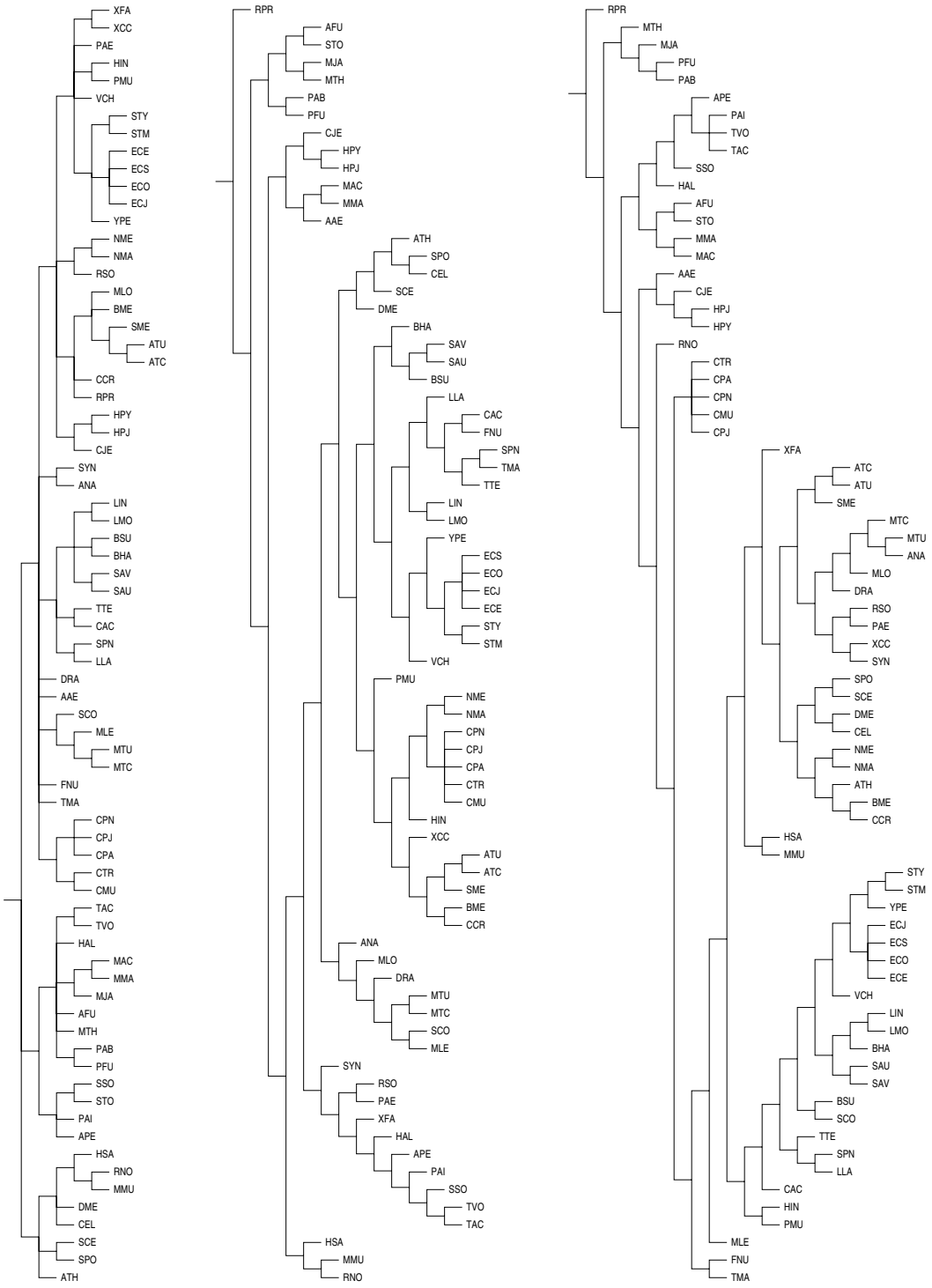


Figure 1: Phylogenetic trees obtained from the Glycolysis pathway for 73 organisms: NCBI (left), Heymans and Singh (middle), and gene ontology enzyme similarity (right, average-link hierarchical clustering, $\alpha = 40\%$).

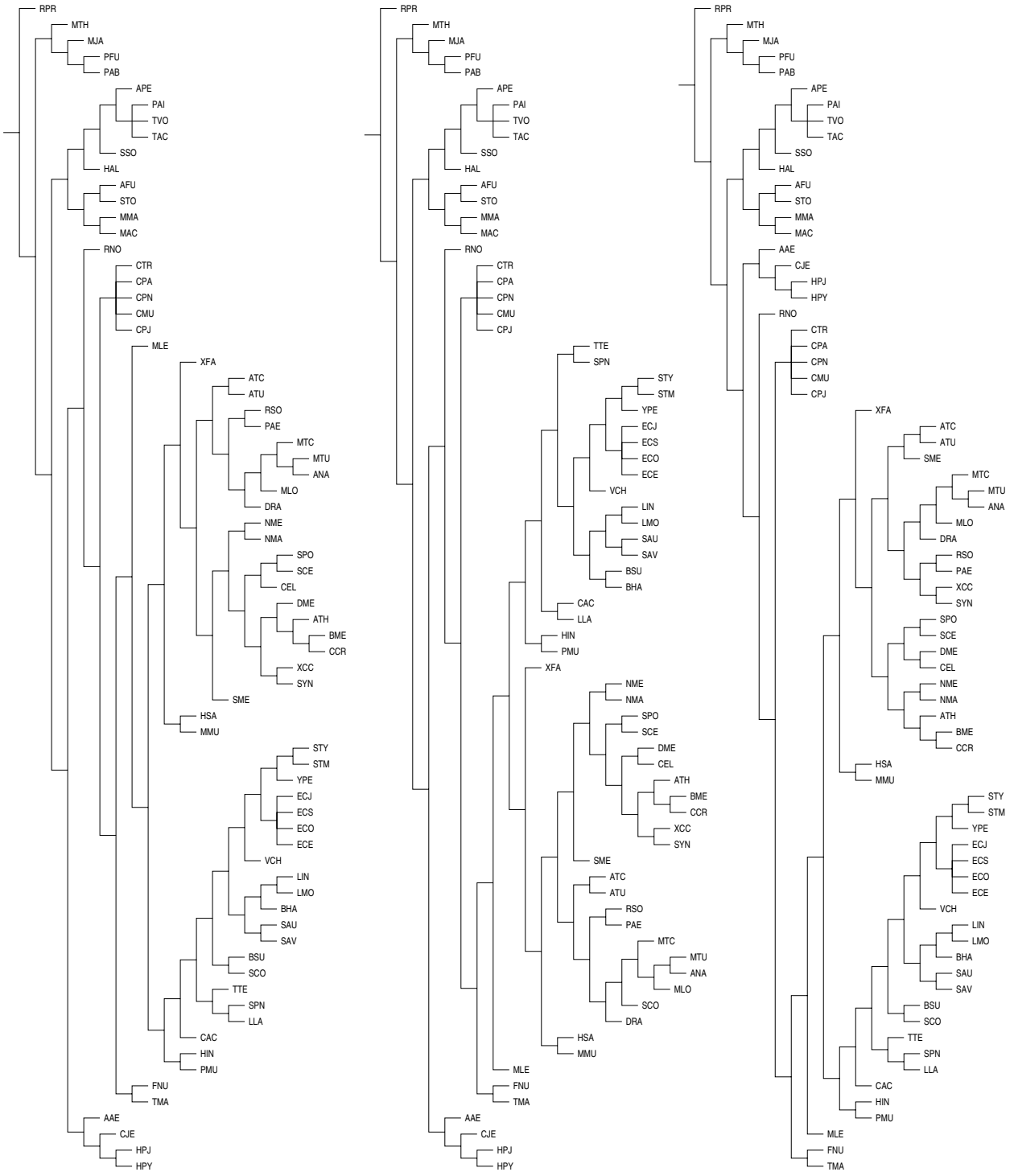


Figure 2: Phylogenetic trees obtained from the Glycolysis pathway for 73 organisms using average-link hierarchical clustering: hierarchical enzyme similarity (left, $\alpha = 30\%$), information content enzyme similarity (middle, $\alpha = 50\%$), and gene ontology enzyme similarity (right, $\alpha = 40\%$).

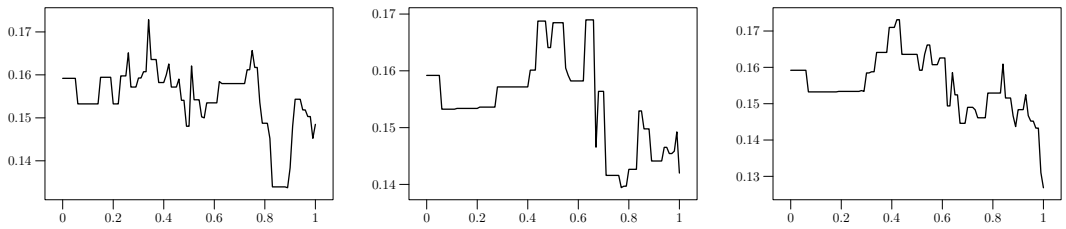

Figure 3: Influence of the $\alpha$ weight parameter on metabolic pathway similarity: hierarchical (left), information content (middle), and gene ontology (right) enzyme similarity.

Table 2: Similarity measures based on the NCBI taxonomy for the Glycolysis pathway across 73 organisms.

| Technique | Similarity |
|---|---|
| Our technique | 0.1709924 |
| Heymans and Singh's technique | 0.1346749 |

Although qualitative analyses of this kind are useful for presentation purposes, it is difficult to evaluate them in practice. Therefore, we measured the similarity of the obtained phylogenies to the NCBI taxonomy as discussed in the previous section, with our approach outperforming previous work by a significant margin.

Figure 2 presents the best trees obtained for each of the three enzyme similarity measures (hierarchical, information content, and gene ontology). Any of these trees is more similar to the NCBI taxonomy than the tree obtained by Heymans and Singh, and the three of them have a similar clustering of related organisms, showing the robustness of our approach.

## 4 Conclusion

We have presented a biologically significative measure of similarity between the compounds and enzymatic reactions in metabolic pathways, and have applied the metabolic pathway similarity measure to the reconstruction of phylogenetic relationships from metabolic pathways across organisms. On the basis of experimental results on the Glycolysis pathway across 73 organisms representing the three domains of life, we have shown that produced phylogenies are robust to changes in the underlying enzyme similarity measure and are also more similar to the NCBI taxonomy than phylogenies produced with previous techniques.

We have used hierarchical enzyme similarity and information content enzyme similarity in these experimental studies, along with a new enzyme similarity measure that we developed based on the gene ontology. While the reconstruction of phylogenetic relationships from metabolic pathways using any of the three enzyme similarity measures produced a similar clustering of related organisms, we have only used a rather simple measure to assess the similarity of compounds, and replacing it by a measure of compound similarity based on a compound ontology [19] is a line of future work. More research is also needed to better understand the influence of the $\alpha$ parameter (that establishes the relative weight of compound similarity to enzyme similarity in the assessment of reaction similarity) on metabolic pathway similarity.

## 5 Methods

We have adopted the usual representation of metabolic pathways as directed hypergraphs, with the compounds and enzymes as nodes and the reactions activated by the enzymes as hyperarcs [3]. As in

previous studies [10, 18], we have discarded so-called *current metabolites*, which function as cofactors in many reactions, namely: $H_2O$, ATP, $NAD^+$, NADH, NADPH, $NADP^+$, $O_2$, ADP, Orthophosphate, CoA, $CO_2$, Pyrophosphate, $NH_3$, and UDP.

In order to assess the similarity of compounds, we have just taken a similarity of 1 for identical compounds and 0 for distinct compounds. We plan to replace it by a measure of compound similarity based on a compound ontology, which is still under development.

### Hierarchical enzyme similarity

In order to assess the similarity of enzymes, we have studied three different enzyme similarity measures: hierarchical, information content, and gene ontology. These measures are based on the enzyme hierarchy, an accepted system for naming and classification of enzymes developed by the Enzyme Commission [14] of the International Union of Biochemistry and Molecular Biology.

In the enzyme hierarchy, enzymes are divided into six main classes on the basis of the reaction activated by the enzyme. Each enzyme is assigned a code, the EC number, which is a string of four digits, separated by dots. The first digit shows the main class which the enzyme belongs to. The second and third digits in the EC number further describe the kind of reaction being activated, and their meanings are defined separately for each of the main classes. The fourth digit distinguishes between enzymes activating very similar but non-identical reactions, by defining the actual substrate.

Consider, for instance, EC code 3.2.1.108. This code corresponds to the lactase enzyme, which activates the hydrolysis of the disaccharide lactose to its component monosaccharides glucose and galactose. The first digit corresponds to class 3.-.-.-, the hydrolases. For the hydrolases, the second digit identifies the type of bond hydrolyzed and the third digit further describes the type of bond hydrolyzed. In the case of lactase, 3.2.-.- are the glycosylases, which have a glycosidic bond (linking carbohydrate units) and 3.2.1.- are the glycosidases (enzymes hydrolysing O-glycosyl and S-glycosyl compounds). Lactase actually activates hydrolysis of the O-glycosyl bond. The fourth and last digit identifies the particular reaction. In the case of lactase, 3.2.1.108 identifies the actual lactose being hydrolized.

The *hierarchical similarity* of two enzymes [13] is the number of common most significant EC digits of the enzymes over 4. The five possible values of hierarchical similarity are thus: 0, for two dissimilar enzymes (with their first digit different); 0.25, if the first digit is identical and the second digit is different; 0.5, if the first two digits are identical but the third digit is different; 0.75, if the first three digits are identical but the last digit is different; and 1, for two identical enzymes (with all four digits identical). The intuition behind hierarchical similarity is to measure how close two enzymes are to each other in the enzyme hierarchy, with higher similarity values for closer enzymes. As a matter of fact, the hierarchical similarity between two enzymes is related to the shortest path between the enzymes in the enzyme hierarchy.

For instance, the hierarchical similarity between the enzymes lactase (3.2.1.108) and glycosylceramidase (3.2.1.62) is 0.75, because they share the first three digits, while the hierarchical similarity between lactase and adenosine nucleosidase (3.2.2.7) is 0.5, and the hierarchical similarity between lactase and phloretin hydrolase (3.7.1.4) is 0.25.

### Information content enzyme similarity

The similarity of two enzymes can also be taken to be the information content of their least common ancestor in the enzyme hierarchy. The *information content similarity* of two enzymes [12, 13] is minus the logarithm of the size of the enzyme hierarchy subtree rooted at the least common ancestor of the enzymes. Similarity values based on information content range from a smallest value of 0, for two identical enzymes, to a largest negative value of about $-12$, for two dissimilar enzymes, and they can be easily normalized by dividing over the size of the whole enzyme hierarchy. The intuition behind information content similarity is also to measure how close two enzymes are to each other in

the enzyme hierarchy, with higher similarity values for closer enzymes. Unlike hierarchical similarity, though, the basis of the similarity measure is not the shortest path between the enzymes in the enzyme hierarchy but the whole subtree rooted at their least common ancestor.

For instance, the information content similarity between lactase (3.2.1.108) and glycosylceramidase (3.2.1.62) is −7.24, because class 3.2.1.- has 151 enzymes, while the information content similarity between lactase and adenosine nucleosidase (3.2.2.7) is −7.50, because class 3.2.-.- has 176 enzymes, and the information content similarity between lactase and phloretin hydrolase (3.7.1.4) is −10.29, because class 3.-.-.- has 1,252 enzymes.

### Gene ontology enzyme similarity

The third method we studied to assess the similarity of two enzymes is based on the gene ontology. Gene Ontology (GO) is a widely accepted standard for describing genes and gene products [1]. GO is composed of *concepts*, each of them described by an unique identifier and one or more strings to name the concept. GO concepts are related to each other by *is-a* and/or *part-of* relations, arranged as a directed acyclic graph.

GO includes three different ontologies: *molecular function*, to describe activities at the molecular level; *biological process*, which deals with series of events accomplished by molecular functions; and *cellular component*, describing different parts of the cell. The molecular function ontology contains concepts representing most of the enzymes present in the Enzyme Commission (EC) database. In this work, we introduce a new enzyme similarity measure based on the shortest distance in the GO hierarchy (not considering direction or type of relation) between the concepts representing any pair of enzymes. Enzymes that have no associated GO entry are substituted by the concept corresponding to the closest sibling enzyme. Dijkstra's algorithm is used to calculate the minimum distance between GO concepts.

The gene ontology similarity measure is conceptually similar to the hierarchical one, since both are based on the shortest path between enzymes, but using a different representation of the enzyme taxonomy, namely, the corresponding associated subgraph of the Gene Ontology.

Using the examples presented in the previous section, the gene ontology distance between lactase (mapped to GO:0000016, "lactase activity") and glycosylceramidase (mapped to GO:0017042, "glycosylceramidase activity") is 2, since the shortest path is [GO:0000016, GO:0004553, GO:0017042]. Lactase and adenosine nucleosidase (GO:0047622, "adenosine nucleosidase activity") are at distance 4 through the shortest path [GO:0000016, GO:0004553, GO:0016798, GO:0016799, GO:0047622]. The most dissimilar examples under hierarchical similarity are also the most distant ones under gene ontology similarity: lactase and phloretin hydrolase (GO:0050180, "phloretin hydrolase activity") are at distance 6 in the path [GO:0000016, GO:0004553, GO:0016798, GO:0016787,GO:0016822, GO:0016823, GO:0050180]. We have obtained normalized gene ontology similarity values from these distances by dividing over the maximum distance (among all enzymes in the metabolic pathway) and subtracting from 1.

### Heuristic algorithm for the comparison of metabolic pathways

We have developed a new heuristic algorithm for the comparison of metabolic pathways, which can be used together with any of the three previous enzyme similarity measures (hierarchical, information content, and gene ontology) to derive a metabolic pathway similarity measure. The algorithm takes all compounds, enzymes, and reactions present in the metabolic pathways into account, and it is based on the idea of computing the intersection and the symmetric difference of the sets of compounds, enzymes, and reactions. Each non-common compound, enzyme, and reaction in one metabolic pathway is then mapped to the most similar compound, enzyme, and reaction, respectively, in the other metabolic pathway.

Let $sim(C, D)$ be the similarity of two compounds $C$ and $D$, and let $sim(E, F)$ be the similarity of two enzymes $E$ and $F$ (either hierarchical, information content, or gene ontology similarity). The similarity of two enzymatic reactions $R = (\mathbf{C}, \mathbf{E})$ and $S = (\mathbf{D}, \mathbf{F})$, where $\mathbf{C}, \mathbf{D}$ are sets of compounds and $\mathbf{E}, \mathbf{F}$ are sets of enzymes, is

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

where $\alpha$ is a weight parameter with $0 \leqslant \alpha \leqslant 1$. The $\alpha$ parameter establishes the relative weight of compound similarity to enzyme similarity in the assessment of enzymatic reaction similarity.

Equation (1) can be simplified when compound and enzyme similarity are normalized metrics, as in the case of hierarchical, information content, and gene ontology similarity, as follows:

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

The similarity of two metabolic pathways $\mathbf{P} = (\mathbf{C}, \mathbf{R})$ and $\mathbf{Q} = (\mathbf{D}, \mathbf{S})$, where $\mathbf{C}, \mathbf{D}$ are sets of compounds and $\mathbf{R}, \mathbf{S}$ are sets of enzymatic reactions, is

$$sim(\mathbf{P}, \mathbf{Q}) = \frac{1}{|\mathbf{R} \cup \mathbf{S}|} \left( \sum_{R \in \mathbf{R} \cap \mathbf{S}} \max_{S \in \mathbf{R} \cap \mathbf{S}} sim(R, S) + \sum_{R \in \mathbf{R} \setminus \mathbf{S}} \max_{S \in \mathbf{S}} sim(R, S) + \sum_{S \in \mathbf{S} \setminus \mathbf{R}} \max_{R \in \mathbf{R}} sim(R, S) \right) \quad (3)$$

Equation (3) can also be simplified when enzymatic reaction similarity is a normalized metric, as follows:

$$sim(\mathbf{P}, \mathbf{Q}) = \frac{1}{|\mathbf{R} \cup \mathbf{S}|} \left( |\mathbf{R} \cap \mathbf{S}| + \sum_{R \in \mathbf{R} \setminus \mathbf{S}} \max_{S \in \mathbf{S}} sim(R, S) + \sum_{S \in \mathbf{S} \setminus \mathbf{R}} \max_{R \in \mathbf{R}} sim(R, S) \right) \quad (4)$$

The metabolic pathway similarity measure is a normalized metric, and it can be computed in time quadratic in the number of compounds, enzymes, and reactions in the metabolic pathways.

## Acknowledgments

The research described in this paper was partially supported by the Spanish CICYT, project GRAMMARS (TIN2004-07925-C03-01), by the Japan Society for the Promotion of Science through Long-term Invitation Fellowship L05511 for visiting JAIST (Japan Advanced Institute of Science and Technology), and by the Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Japan.

## References

[1] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G., Gene Ontology: tool for the unification of biology, *Nat. Genet.*, 25(1):25–29, 2000.

[2] Dandekar, T., Schuster, S., Snel, B., Huynen, M., and Bork, P., Pathway alignment: Application to the comparative alignment of glycolytic enzymes, *Biochem. J.*, 343(1):115–124, 1999.

[3] Deville, Y., Gilbert, D., van Helden, J., and Wodak, S. J., An overview of data models for the analysis of biochemical pathways, *Brief. Bioinform.*, 4(3):246–259, 2003.

[4] Forst, C. V. and Schulten, K., Evolution of metabolisms: A new method for the comparison of metabolic pathways using genomic information, *J. Comp. Biol.*, 6(3–4):343–360, 1999.

[5] Forst, C. V. and Schulten, K., Phylogenetic analysis of metabolic pathways, *J. Mol. Evol.*, 52(1):471–489, 2001.

[6] Heymans, M. and Singh, A. K., Deriving phylogenetic trees from the similarity analysis of metabolic pathways, *Bioinformatics*, 19(Suppl. 1):i138–i146, 2003.

[7] Jain, A. K., Murty, M. N., and Flyn, P. J., Data clustering: A review, *ACM Comput. Surv.*, 31(3):264–323, 1999.

[8] Kanehisa, M. and Goto, S., KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, 28(1):27–30, 2000.

[9] Liao, L., Kim, S., and Tomb, J.-F., Genome comparisons based on profiles of metabolic pathways, In *Proc. 6th Int. Conf. Knowledge-Based Intelligent Information and Engineering Systems*, 469–476, 2002.

[10] Ma, H. and Zeng, A.-P., Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms, *Bioinformatics*, 19(2):270–277, 2003.

[11] Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M., A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucleic Acids Res.*, 28(20):4021–4028, 2000.

[12] Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E., and Ziv-Ukelson, M., Alignment of metabolic pathways, *Bioinformatics*, 21(16):3401–3408, 2005.

[13] Tohsato, Y., Matsuda, H., and Hashimoto, A., A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy, In *Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology*, 376–383, 2000.

[14] Webb, E. C., editor, *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*, Academic Press, 1993.

[15] Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A., Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.*, 28(1):10–14, 2000.

[16] Yamada, T., Goto, S., and Kanehisa, M., Extraction of phylogenetic network modules from prokaryote metabolic pathways, *Genome Informatics*, 15(1):249–258, 2004.

[17] Zhang, K., Wang, J. T.-L., and Shasha, D., On the editing distance between undirected acyclic graphs, *Int. J. Foundations Comput. Sci.*, 7(1):43–57, 1996.

[18] Zhu, D. and Qin, Z. S., Structural comparison of metabolic pathways in selected single cell organisms, *BMC Bioinformatics*, 6:8, 2005.

[19] European Bioinformatics Institute, Chemical entities of biological interest, Database of small molecular entities available at `http://www.ebi.ac.uk/chebi/`.