Fuzzy Clustering improves Phylogenetic Relationships Reconstruction from Metabolic Pathways

Jaume Casasnovas
Dept. Math. and Comp. Sc.
Univ. Balearic Islands
E-07122 Palma, Spain
jaume.casasnovas@uib.es

José C. Clemente
School of Knowledge Sc.
JAIST
Ishikawa 923-1292, Japan
clemente@jaist.ac.jp

Joe Miró-Julià
Dept. Math. and Comp. Sc.
Univ. Balearic Islands
E-07122 Palma, Spain
joe.miro@uib.es

Francesc Rosselló
Dept. Math. and Comp. Sc.
Univ. Balearic Islands
E-07122 Palma, Spain
cesc.rossello@uib.es

Kenji Satou
School of Knowledge Sc.
JAIST
Ishikawa 923-1292, Japan
ken@jaist.ac.jp

Gabriel Valiente
Dept. of Software
Tech. Univ. of Catalonia
E-08034 Barcelona, Spain
valiente@lsi.upc.edu

Abstract

The interest in reconstructing phylogenetic relationships from data on structural similarity of metabolic pathways is growing. The similarity notions and the techniques involved in this reconstruction are assessed by building phylogenetic relationships for model sets of organisms from the similarity measures of the same metabolic pathway for all of them, and then the phylogenetic trees obtained are compared to the NCBI taxonomy. The best technique proposed so far is due to some of the authors of this paper [2], using a new similarity relation for metabolic pathways and average-link hierarchical clustering to compute the phylogenetic tree. In this paper we prove that using a fuzzy clustering method to compute the phylogenetic relationships from this similarity relation the resulting trees are usually closer to the NCBI taxonomy.[1]

Keywords: metabolic pathway, similarity, hierarchical similarity, information content, gene ontology, phylogenetic reconstruction, fuzzy clustering.

[1] All data and programs are available from the corresponding author (F. Rosselló) upon request.

1 Introduction

Evolutionary relationships among species have been mainly understood through the "molecular approach," which exploits polymorphism information in DNA or protein sequences to assess the phylogenetic relationship among species. Nevertheless, the choice of sequences for comparison greatly affects the final result, since non-local mutations like the transfer, duplication, deletion, and functional replacement of genes can prevent the obtention of significant information from the analysis of the evolution of short sequences [16]. This has led to an increasing interest in phylogenetic reconstruction based on non-genomic information, as for instance information on global biological processes in the organisms, specially the use of metabolic pathways. Recent advances in metabolomics experimental techniques also suggest that most metabolic network data will soon be independent from genomic information, which will turn metabolic pathways into a true alternative to biomolecular sequences in phylogeny reconstruction [6].

Structural similarity of metabolic pathways entails 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 usually represented as directed hypergraphs, with the compounds and enzymes as nodes and the reactions activated by the enzymes as hyperarcs [4]. A

more abstract representation, called the *enzyme graph* [10, 18], uses nodes to represent enzymes and arcs to represent compounds shared between successive reactions.

Reaction similarity can be assessed by measuring the similarity of the enzymes activating the reactions and the compounds involved in them. *Hierarchical similarity* [23] and *information content similarity* [20, 23] are two commonly used enzyme similarity measures based on the enzyme hierarchy [24]. In this paper we shall use them together with a third one, the *gene ontology similarity*, introduced in [2] and based on the Gene Ontology [1]. Concerning similarity of compounds, we shall only use the simplest one, the equality, but more complex similarities, for instance based on a compound ontology [5], could also be considered.

Metabolic pathways from different genomes have been aligned upon similar enzymes, substrates, and products [3, 20, 23]. Since the alignment of metabolic pathways is computationally hard, previous phylogenetic analyses of metabolic pathways have been based on the number of common enzymes between two organisms [7, 8], on the presence and absence of metabolic pathways [13], on the topology of enzyme graphs [10], and on the presence and absence of reactions in whole metabolic networks [6]. The produced phylogenies are assessed by comparing them against the NCBI taxonomy [25], which is based on Ribosomal RNA 16S sequences, using some standard phylogenetic tree comparison tool. For instance, [2, 10] used the `cousins` software package [21, 22] to compare their results to previous work, and we have also used it to facilitate the comparison of our results with theirs. This tool compares unordered trees to labelled leaves, like phylogenetic trees, by comparing in a specific way, and up to a certain cousin distance, the sets of *cousin pairs*, triples consisting of a pair of leaves and their *cousin distance*: 0 if they are siblings (they share the same parent), 0.5 if the parent of one of them is the grandparent of the other, 1 if they are cousins (they share the same grandparent but not the same parent), 1.5 if their first common ancestor is the grandparent of one of them and the great-grandparent of the other one, 2 if they are second cousins (they share the same great-grandparent but not the same grandparent) and so on.

In [2] we introduced a new heuristic algorithm to compare metabolic pathways to be used together with any enzyme similarity measure. This algorithm computes the intersection and symmetric difference of the sets of compounds, enzymes, and reactions present in the metabolic pathways. Each non-common compound, enzyme, and reaction in one metabolic pathway is mapped to the most similar one in the other pathway. We showed that the metabolic pathway similarity measure obtained, together with an average-link hierarchical (ALH) clustering method [11], produced better phylogenetic trees than any previous method.

In this paper we show that replacing the ALH clustering method by a fuzzy equivalence relations-based (FER) hierarchical clustering method [17, §4.2], the resulting phylogenetic trees are in most occasions significantly better. Fuzzy clustering has been successfully used in bioinformatics, mostly through variants of the *fuzzy c-means* (FCM) clustering method. For instance, [19] introduced a method for DNA-based phylogenetic tree reconstruction based on FCM and Markov models. But FCM-based hierarchical clustering methods have the disadvantage of requiring the desired number of clusters be given *a priori* in each step. Alternatively, all possible number of clusters must be tried and then the optimal number chosen according to some "least fuzzy partitions" criterion, although this method is extremely time consuming. The FER clustering method overcomes these drawbacks: it is faster, logically simpler, and naturally hierarchical [26]. Although it has found several applications in health sciences (see [17, Ch. 4] and the references therein), to our knowledge, it has only been used in one occasion to produce phylogenetic trees [14].

In the FER clustering method, we determine a fuzzy similarity relation S (reflexive and sym-

metric) on the set of objects and compute the fuzzy equivalence relation E generated by this similarity as the max-min transitive closure of the matrix of S. Then, for each t appearing in E's matrix the *t-cut* crisp equivalence relation obtained by replacing every entry in E's matrix smaller than t by 0 and every entry greater than or equal to t by 1, induces a crisp partition of the set of objects: each element of the partition is a maximal subset of objects that have "E-equivalence value" $\geqslant t$ with each other. These partitions, together with the hierarchy induced by the increasing order of the values t, yields a classification tree for the objects.

For instance, consider the similarity matrix

	MGE	HIN	MTU	MJA	ECO	AFU
MGE	1.00	0.33	0.07	0.02	0.17	0.22
HIN	0.33	1.00	0.33	0.32	0.34	0.27
MTU	0.07	0.33	1.00	0.09	0.20	0.20
MJA	0.02	0.32	0.09	1.00	0.18	0.24
ECO	0.17	0.34	0.20	0.18	1.00	0.32
AFU	0.22	0.27	0.20	0.24	0.32	1.00

on the set of organisms

$$\{\text{MGE}, \text{HIN}, \text{MTU}, \text{MJA}, \text{ECO}, \text{AFU}\}$$

(see Table 1). The fuzzy equivalence relation generated by this similarity is given by the matrix

$$\begin{pmatrix} 1.00 & 0.33 & 0.33 & 0.32 & 0.33 & 0.32 \\ 0.33 & 1.00 & 0.33 & 0.32 & 0.34 & 0.32 \\ 0.33 & 0.33 & 1.00 & 0.32 & 0.33 & 0.32 \\ 0.32 & 0.32 & 0.32 & 1.00 & 0.32 & 0.32 \\ 0.33 & 0.34 & 0.33 & 0.32 & 1.00 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 & 1.00 \end{pmatrix}$$

The hierarchy of partitions of the set of organisms defined by the t-cuts of this fuzzy equivalence relation is:

t	Partition corresponding to the t-cut
1.00	{MGE} {HIN} {MTU} {MJA} {ECO} {AFU}
0.34	{HIN, ECO} {MGE} {MTU} {MJA} {AFU}
0.33	{MGE, HIN, MTU, ECO} {MJA} {AFU}
0.32	{MGE, HIN, MTU, MJA, ECO, AFU}

This hierarchical clustering yields a classification tree, depicted as a dendogram in Fig. 1, and which is very close to the NCBI taxonomy tree for these six organisms, the only difference being that in the latter the archaea MJA and AFU are also clustered (that is, they should be at cousin distance 1 from MGE and MTU, instead of 0.5).

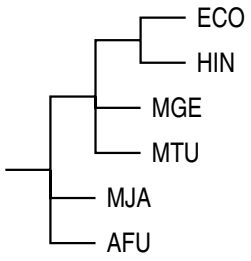

Figure 1: A dendogram for MGE, HIN, MTU, MJA, ECO, and AFU similar to their NCBI taxonomy.

To compare the performance of FER clustering and ALH clustering, we have computed the similarities defined in [2] of the Glycolysis pathways of a model set of 16 organisms. The Glycolysis pathway, which mainly serves to generate ATP molecules, has been thoroughly studied in the literature, being highly conserved in the genetic code and occurring in most species. Similarity among different organisms can therefore be studied by analyzing the similarity of their respective Glycolysis pathways. We have computed the phylogenetic trees generated by these similarities by using both the ALH clustering, as presented in [2], and the FER hierarchical clustering method explained above. Finally, we have computed the similarity of the trees obtained in this way to the NCBI taxonomy of the 16 organisms using the `cousins` tool up to cousin distance 2. The values obtained through FER are often significantly better than those obtained through ALH clustering.

2 Materials and methods

The metabolic pathways for the organisms have been downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) server [12], a repository of metabolic pathways for organisms with completely sequenced genomes that provides information about the enzymes present in the pathways and their classification. We only consider the Glycolysis pathway for the 16 organisms listed

in Table 1, which have already been studied by other authors [10, 13].

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

AFU	*A.fulgidus*	A	224325
MJA	*M.jannaschii*	A	243232
CPN	*C.pneumoniae*	B	115713
MGE	*M.genitalum*	B	243273
MPN	*M.pneumoniae*	B	272634
HIN	*H.influenzae*	B	71421
SYN	*Synechocystis*	B	1148
DRA	*D.radiodurans*	B	243230
MTU	*M.tuberculosis*	B	83332
TPA	*T.pallidum*	B	243276
BSU	*B.subtilis*	B	224308
AAE	*A.aeolicus*	B	224324
TMA	*T.maritima*	B	243274
ECO	*E.coli*	B	83333
HPY	*H.pylori*	B	85962
SCE	*S.cerevisiae*	E	4932

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 [4]. This representation can be extracted automatically from the KEGG files. As in previous studies [15, 27], we have discarded the *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.

To assess the similarity of enzymes, we have considered three different enzyme similarity measures: *hierarchical*, *information content*, and *gene ontology*. The first two are based on the *enzyme hierarchy* [24], an accepted system for naming and classifying enzymes developed by the Enzyme Commission of the International Union of Biochemistry and Molecular Biology. In the enzyme hierarchy, each enzyme is assigned a code, the EC number: a string of four digits, separated by dots. The first digit shows the main class (on the basis of the reaction it activates) 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. The *hierarchical similarity* of two enzymes [23] is the number of common most significant EC digits of the enzymes divided by 4, and the *information content similarity* of two enzymes [20, 23] is minus the logarithm of the size of the enzyme hierarchy subtree rooted at the least common ancestor of the enzymes.

The third method we use to assess the similarity of two enzymes, introduced in [2], is based on the *Gene Ontology* (GO), a widely accepted standard for describing genes and gene products [1] that is composed of *concepts*, each of them identified by a unique index and one or more strings to name the concept. GO concepts are related to each other by *is-a* or *part-of* relations, arranged as a directed acyclic graph. The *gene ontology similarity* of two enzymes is then the shortest distance in the GO graph (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. The minimum distance between GO concepts is computed using Dijkstra's algorithm.

We represent the enzymatic reactions by a pair $(\mathbf{C}, \mathbf{E})$, where $\mathbf{C}$ is the set of compounds and $\mathbf{E}$ the set of enzymes. The *similarity* of two enzymatic reactions $R = (\mathbf{C}, \mathbf{E})$ and $S = (\mathbf{D}, \mathbf{F})$ is given by

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

where *sim* for compounds is 1 for identical compounds and 0 for distinct compounds, and

sim for enzymes stands for either the hierarchical (*hier*), information content (*info*), or gene ontology (*go*) similarity measures. The weight parameter $\alpha \in [0,1]$ establishes the relative weight of compound similarity to enzyme similarity in the assessment of enzymatic reaction similarity.

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 then

$$sim_\alpha(\mathbf{P}, \mathbf{Q}) = \frac{1}{|\mathbf{R} \cup \mathbf{S}|}\Big(|\mathbf{R} \cap \mathbf{S}| + \sum_{R \in \mathbf{R} \setminus \mathbf{S}} \max_{S \in \mathbf{S}} sim_\alpha(R, S) + \sum_{S \in \mathbf{S} \setminus \mathbf{R}} \max_{R \in \mathbf{R}} sim_\alpha(R, S)\Big)$$

We have computed the similarity of the Glycolysis pathway for each pair of our model set of 16 organisms, for each of the three enzyme similarity measures, and for each $\alpha = 0, 0.1, 0.2, \ldots, 0.9, 1$. This has produced 33 16×16 matrices with entries in $[0,1]$. These matrices are symmetrical and all entries in their main diagonal are 1. For instance, the sample similarity matrix used in the Introduction to explain the FER method is the restriction of $go_{0.8}$ to the corresponding set of 6 organisms. We have then clustered our 16 organisms based on these similarities, using ALH clustering [11].

To use the FER hierarchical clustering on these matrices we have computed their max-min transitive closure using the algorithm derived from [17, Thm. 4.2.1]. In this way we have obtained the matrix of the fuzzy equivalence generated by each one of the 33 similarity matrices on the set of the 16 organisms. We have then computed the classification tree given by each one of these fuzzy equivalences, as explained in the introduction, and we have considered them as the organisms' phylogenetic trees.

3 Results and discussion

To evaluate the effectiveness of our method we have compared the phylogenetic trees obtained in the previous section to the NCBI taxonomy restricted to the 16 organisms considered in this study. We have used the `cousins` software package to compute similarity measures between phylogenies.

We were unable to reproduce the 0.27 similarity claimed in [9, Table 5] for any parameter of the `cousins` tool, though, and we have adopted the parameter setting that provides the closest result (similarity up to second cousins, that is, up to cousin distance 2) for our experiments, as was already done in [2], which yields a similarity of 0.19355 between NCBI's and Heymans-Singh's trees.

Table 2: Similarity values for both clustering methods and all similarity measures.

	ALH	*FER*
$go_{0.0}$	0.23864	0.26042
$go_{0.1}$	0.23864	0.27369
$go_{0.2}$	0.22222	0.31959
$go_{0.3}$	0.22222	0.27368
$go_{0.4}$	0.22222	0.26596
$go_{0.5}$	0.22222	0.26596
$go_{0.6}$	0.22222	0.23077
$go_{0.7}$	0.22222	0.21277
$go_{0.8}$	0.22222	0.29474
$go_{0.9}$	0.22222	0.26000
$go_{1.0}$	0.25275	0.20430
$hier_{0.0}$	0.23864	0.26042
$hier_{0.1}$	0.22222	0.27369
$hier_{0.2}$	0.22222	0.30208
$hier_{0.3}$	0.22222	0.29032
$hier_{0.4}$	0.22222	0.25532
$hier_{0.5}$	0.22222	0.28421
$hier_{0.6}$	0.22222	0.23404
$hier_{0.7}$	0.22222	0.21978
$hier_{0.8}$	0.25275	0.23404
$hier_{0.9}$	0.25275	0.19355
$hier_{1.0}$	0.25275	0.19565
$info_{0.0}$	0.23864	0.26042
$info_{0.1}$	0.23864	0.31250
$info_{0.2}$	0.22222	0.25263
$info_{0.3}$	0.22222	0.30208
$info_{0.4}$	0.22222	0.22340
$info_{0.5}$	0.23864	0.21053
$info_{0.6}$	0.23864	0.18280
$info_{0.7}$	0.23864	0.14130
$info_{0.8}$	0.27778	0.15054
$info_{0.9}$	0.27778	0.16304
$info_{1.0}$	0.27778	0.18681

Table 2 shows the similarity values (rounded

to 5 decimal digits) to the NCBI taxonomy tree of the phylogenetic trees obtained through ALH clustering (column *ALH*) and through FER hierarchical clustering (column *FER*) for each of the similarity measures go_α, $hier_\alpha$ and $info_\alpha$, $\alpha = 0, 0.1, 0.2, \ldots, 0.9, 1$. Recall that this parameter α captures the relative weight of compound similarity to enzyme similarity in the assessment of reaction similarity: the smaller this parameter, the smaller the relative weight of enzyme similarity.

It is clearly shown in Table 2 that the gene ontology similarity yields better results when using the FER clustering method. For all values of α except $\alpha = 1$ (that is, except when compound similarity is not taken into account) and $\alpha = 0.7$ the FER tree is closer to the NCBI taxonomy than the ALH tree. The greatest similarity is $go_{0.2}$ with FER (almost 0.32), while the maximum with ALH is $go_{1.0}$ (slightly under 0.253). The average similarity of the FER trees to the NCBI taxonomy is 0.260, while the average similarity of the ALH trees is 0.228.

FER clustering also generates better trees than ALH for the hierarchical similarity. The FER tree is closer to the NCBI taxonomy than the ALH tree for all values of α except, in this case, four: all $\alpha \geqslant 0.7$. The greatest similarity is reached again for $hier_{0.2}$ using FER (slightly over 0.3), while the maximum with ALH is reached for $hier_\alpha$ with $\alpha \geqslant 0.8$ (slightly under 0.253). The average similarity of the FER trees to the NCBI taxonomy is in this case 0.249, while the average similarity of the ALH trees is 0.232.

Interestingly, FER behaves worse than ALH for the information content similarity: the FER tree is more similar to the NCBI taxonomy than the ALH tree for all $\alpha \leqslant 0.4$, while the ALH tree is better when $\alpha \geqslant 0.5$. The greatest similarity is reached in this case for $info_{0.1}$ with FER (slightly over 0.31), while the maximum with ALH is obtained for $info_\alpha$ with $\alpha \geqslant 0.8$ (slightly above 0.277).

The better performance of FER when using gene ontology and hierarchical similarity might be explained by the fact that these measures are conceptually similar, since both are based on shortest path distance among enzymes (in the GO graph and the EC tree, respectively). On the other hand, information content similarity (where FER obtained worse results) is based on EC subtree size, which results in a more fine-grained measure than gene ontology or hierarchical similarity.

It is also evident from Table 2 that for all three types of enzyme similarity the best results are obtained using FER and low values of α. Indeed, if we only take into account the values $\alpha = 0, \ldots, 0.4$, the average similarity of the phylogenetic trees to the NCBI taxonomy is 0.27867 for *go*, 0.27637 for *hier* and 0.27021 for *info*.

The best phylogenetic trees obtained with FER clustering for each one of the three enzyme similarity measures are shown in Fig. 2.

4 Conclusion

We have recalled from [2] a new measure of similarity between metabolic pathways and applied it to the reconstruction of phylogenetic relationships from metabolic pathways across organisms using the FER hierarchical clustering algorithm. We have used a set of 16 organisms representing the three domains of life. We have restricted our experiments in this work to a few organisms to provide a simple empirical proof of the advantages of our method on a well-studied dataset. Results for larger sets of organisms, as those considered in [2, 9, 10] will appear elsewhere. Our results on the Glycolysis pathway for the chosen 16 organisms show that the produced phylogenies are more similar to the NCBI taxonomy than phylogenies produced with previous techniques.

We have used hierarchical, information content, and gene ontology enzyme similarity, and our metabolic pathways similarity measure involves a parameter α that can shift weight from the enzyme to the compound similarity, which in this paper was taken as the 1-0 compound equality. We have shown that using FER together with gene ontology and

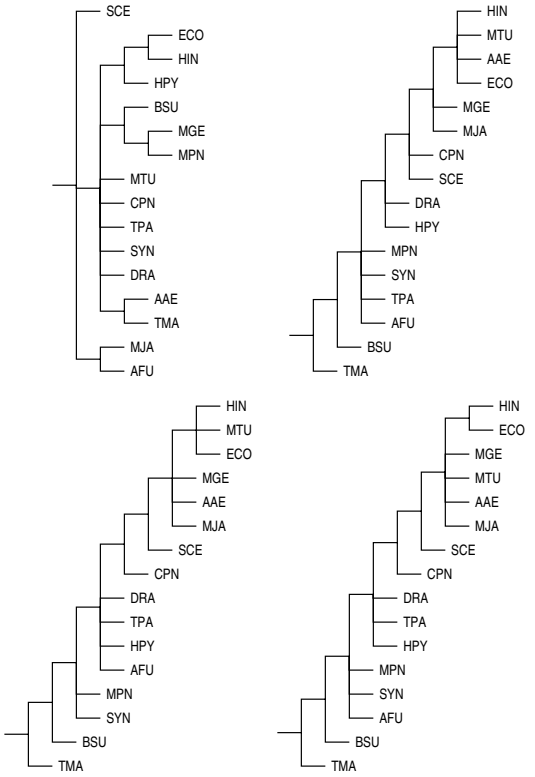

Figure 2: Phylogenetic tree for the set of 16 organisms (NCBI taxonomy, top left) and best trees obtained with FER clustering from the similarity of their Glycolysis pathways, using gene ontology (top right), hierarchical (bottom left), and information content enzyme similarity (bottom right).

small values of α, and thus giving more weight to the compound equality, yield better results. But, as it can be seen in Fig. 2, and despite the improvements achieved by using a fuzzy clustering algorithm, we are still far from obtaining a fully correct taxonomy. More details on the analysis of our results will also appear elsewhere.

Acknowledgements

The research described in this paper was partially supported by the Spanish CICYT, projects GRAMMARS (TIN2004-07925-C03-01) and ALBIOM (BFM2003-0071), 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] M. Ashburner, C. A. Ball, and al. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[2] J. C. Clemente, K. Satou, and G. Valiente. Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology. *Genome Informatics*, 16(2):45–55, 2005.

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

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

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

[6] C. V. Forst, C. Flamm, I. L. Hofacker, and P. F. Stadler. Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics*, 2005. Submitted.

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

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

[9] M. Heymans and A. K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways.

Technical Report 2002-33, available at http://cs.ucsb.edu/research/tech_reports/reports/2002-33.pdf.

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

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

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

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

[14] L. Luo, F. Ji, and H. Li. Fuzzy classification of nucleotide sequences and bacterial evolution. *Bulletin of Mathematical Biology*, 57(4):527–537, 1995.

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

[16] W. Martin. Mosaic bacterial chromosomes: a challenge en route to a tree of genome. *Bioessays*, 21:99–104, 1999.

[17] J. M. Mordeson, D. S. Malik, and S.-C. Cheng. *Fuzzy Mathematics in Medicine*, volume 55. Physica-Verlag, 2000.

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

[19] T. D. Pham, D. Beck, and D. I. Crane. Fuzzy clustering of stochastic models for molecular phylogenetics. *WSEAS Trans. Mathematics and Computers in Biology and Biomedicine*, 1(2):87–92, 2005.

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

[21] D. Shasha. Unordered tree comparison based on cousin distance. http://www.cs.nyu.edu/cs/faculty/shasha/papers/cousins.html.

[22] D. Shasha, J. T.-L. Wang, and S. Zhang. Unordered tree mining with applications to phylogeny. In *Proc. 20th International Conference on Data Engineering, ICDE 2004*, pages 708–719. IEEE Computer Society, 2004.

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

[24] E. C. Webb, 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.

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

[26] L. A. Zadeh. Similarity relations and fuzzy orderings. *Information Sciences*, 3(1):177–206, 1971.

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