Genome Comparisons Based on Profiles of Metabolic Pathways

Li Liao[1,*], Sun Kim[2], Jean-Francois Tomb[3]

1. Department of Computer and Information Sciences
University of Delaware, Newark, DE 19716, USA
2. School of Informatics, Indiana University, Bloomington, IN 47405, USA
3. DuPont Central Research and Development
Experimental Station, Wilmington, DE 19880, USA

Abstract. A computational method to compare organisms based on genome-wide metabolic pathway analysis was developed. Using the WIT database, a metabolic pathway profile for each completed genome is generated. These profiles are records of the presence and absence of the various metabolic pathways, and constitute the basis for a comparison of organisms. A scoring scheme and an algorithm were developed to evaluate generic profiles, which are based on attributes that bear hierarchical relationships. This metabolic pathway profile-based (MPP-based) classification, as applied to analyzing fully sequenced genomes, can shed light on evolution of metabolic pathways.

1. Introduction

As more genomes are sequenced and the metabolic pathways of organisms reconstructed, it becomes possible to perform organism comparisons from a biochemical-physiological perspective. Such comparisons may yield novel insights into the evolution of metabolic pathways and may be relevant to metabolic engineering of industrial microbes. Studies in this direction focusing on individual pathways have been attempted [1,4].

In this work, we propose a novel approach for comparing genomes and classifying organisms. It is based on comparing metabolic pathway profiles (i.e. presence and absence of metabolic pathways), representing the entire metabolic repertoire of an organism. Thus, this approach enables a very rich phenotypic comparison that was not possible before (i.e. morphology, pigmentation, and substrate use). A scoring scheme and algorithm were developed for evaluating generic profiles, which are based on attributes that bear hierarchical relationships. The relationships among pathways are represented as a Master

* Correspondence should be addressed to L.L. Work was accomplished while employed at DuPont.

tree (See section 2.2), and the pathways are represented as leaves. An overall similarity score for two profiles is obtained by averaging scores of matches and mismatches at leaves within a tree branch and propagating average scores bottom-up to the root of the master tree. The developed methodology is applicable to organisms for which we have a complete or nearly complete genomic sequence. In the context of this MPP-based genome comparison method, grouping organisms according to metabolic pathway lineages may shed light on studying evolution and reconstruction of metabolic pathways.

2. *Methods*

In this section, we describe how to utilize genome-wide information about the presence and absence of metabolic pathways to generate profiles, and how to use these profiles for comparing genomes in a hierarchical manner. Section 2.1 shows that the information about the presence and absence of metabolic pathways in an organism can be represented as a binary profile -- a string of zeros and ones. To compare two profiles, a weighted scoring method is proposed in section 2.2 to take into account correlation among pathways. Section 2.3 discusses hierarchical clustering of the profiles using distances obtained through pairwise comparison of profiles.

2.1 Pathway profiling

As a first step, all metabolic pathways are treated independently. Since *n* pathways are independent of one another, there is no inherent order to numbering them. An arbitrary order is chosen and kept constant for all organisms when profiling metabolic pathways. If organism *i* has pathway $P_{j,}$ put 1 into the corresponding element, otherwise put 0. Thus, an organism is represented as a string of zeros and ones, as shown in Table 1.

Table 1. Profiles of presence and absence of metabolic pathways, $P_{1,}$ $P_{2,}$ …, P_n in organisms, $Org_{1,}$ $Org_{2,\ ...,}$ Org_m.

	P_1	P_2	.	.	.	P_n
Org_1	1	0	.	.	.	1
Org_2	0	1	.	.	.	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Org_m	1	0	.	.	.	1

The rule to determine whether a pathway is present in a given organism is strict: all enzymes at every step of the pathway should be present in the organism. However, the assertion of whether an enzyme is present or absent in an organism is based on sequence similarity and also on when to consider two genes as an orthologous pair. Sequence similarity invokes a threshold (*e.g.*, p-score equal to or less than 1.0 e^{-5}), and two genes, from two organisms, are considered orthologues when they satisfy the relationship of bi-directional best hit (For finding orthologues, see the WIT database [6] and COGs [8]).

2.2 Pairwise comparison of pathway profiles

The task of comparing genomes in terms of their pathway profiles (rows in Table 1) is now reduced to a comparison of strings (composed of 1s and 0s). This comparison needs a scoring scheme for matches and mismatches at individual positions. One scoring scheme can be: 1 for matches, -1 for mismatches. To obtain the scored similarity between two profiles, a simple way is to sum up the scores at each position (bit).

$$S = \sum_i^n s(i) \qquad (1)$$

This unweighted score *S* can be normalized by the length n of the profile. Length *n*, in this case, is equal to the total number of pathways. When two profiles have the same length, the score *S* is equal to the normalized Hamming distance [5]. However, Equation (1) becomes inadequate when correlation exists among different positions. As known**,** metabolic pathways are related to one another in terms of physiological functions. Therefore, in contrast to the simple summation over all positions, as shown in Equation (1), a formula that captures correlation among pairs, triples, quadruples pathways, and so forth, is needed.

$$S = \sum_{i\neq j}^{n} s(i)c(i,j)s(j) + \sum_{i\neq j\neq k}^{n} s(i)s(j)s(k)c(i,j,k) + \ldots \qquad (2)$$

However, the coefficients c(i,j), c(i,j,k), *…*, that embody the correlation, are not known *a priori* and can at best be fitted from a set of training data. Despite this difficulty, we postulate that the correlation among pathways may be structured as a hierarchy. The WIT database presents a *pragmatic* classification of the metabolic pathways, which we use as a surrogate for the "real" relationship among pathways. Though the "correctness" of the WIT classification may be arguable, it is reasonable to believe that a hierarchy is a sensible way to structure the relationship among pathways. Consequently, when we score raw profiles of pathway presence and absence, the hierarchical relationships among pathways is considered.

In this work, we propose a heuristic way to include the hierarchical relationships of pathways. To test our methodology, the classification of all known metabolic pathways in the WIT database is adopted. The relationships among pathways are represented as a Master tree, and the pathways are represented as leaves.

Definitions

Master Tree is a tree that represents the hierarchy of all pathways on which the profiles are constructed.

p-Tree is a tree that is derived from the master tree by dropping off leaves whose corresponding pathways are absent from the organism.

In this representation, a profile is no longer a simple string of zeros and ones, where each bit is treated equally and independently. Instead, it is mapped into a p-tree so that the

hierarchical relationship among bits is conserved. Scoring these profiles is equivalent to answering the question: how different are these p-trees? How to compare trees is itself an interesting topic with wide applications, and has been the subject of numerous studies [9]. The task here is narrowly focused and relatively simple: to compare trees (p-tree) that are derived from the same tree (master tree).

The comparison of two p-trees evaluates the difference between the two profiles they represent, with the following correlation being accounted for:

Correlation 1: *pi* and *pj* are sibling, versus *pi* and *pj* are remotely related
Correlation 2: *pi* and *pj* are at different levels of the hierarchy

In generating an overall similarity score for two profiles, a score scheme ought to weight (mis)matches according to their positions in the tree (i.e., take into account hierarchy). To achieve this, we propagate matches and mismatches scores bottom-up to the root of the master tree in four steps. 1) overlay two trees; 2) score mismatches and matches between two trees and label scores at the corresponding leaves on the master tree; 3) average scores from siblings (weight breadth) and assign the score to the parent node; 4) iterate step 3 until the root is reached. The main routines of this algorithm are described as pseudo code in List 1.

```
procedure Main
  input: two profiles, pf1 and pf2; a tree ROOT
  output:f  a score S
        n := length of pf1
        for k := 1 to n do
                id := pf1(k)
                if pf2(k) = id then
                tree.setScoreAtNode(id, 1)
                else
                tree.setScoreAtNode(id, -1)
        endfor
        return S = tree.getScore
endprocedure Main

procedure getScore
        input: a tree node R
        output: a score S
        max := # of children of tree node R
        if max > 0  then
                for j:=1  to max  do
                        node = R.getChildNode(j)
                        S = S + node.getScore
                Endfor
        S =  S / max
        return S
endprocedure getScore
```

List 1. Algorithm for pair-wise comparison of pathway profiles

The path of traversing the master tree is in post-order. A score is obtained at the root, and this score is used to evaluate how ‘‘close” two p-trees are.

2.3 Hierarchical clustering of profiles

In the last section, we discussed a method for comparing two organisms based on metabolic pathway profiles. The comparison result is a similarity score between two profiles, which is a real number equal to or smaller than 1. Score 1 is achieved when two profiles are identical. This similarity score can be interpreted as "distance" by the following formula: *distance = 1 - score*. The distance given by this formula between two identical profiles is zero. It is straightforward to apply pair-wise comparison to a group of N organisms, which results in an N by N distance matrix. Once this distance matrix is obtained, a hierarchical clustering can be accomplished using any distance-based method (See[2]).

3. Results and Discussion

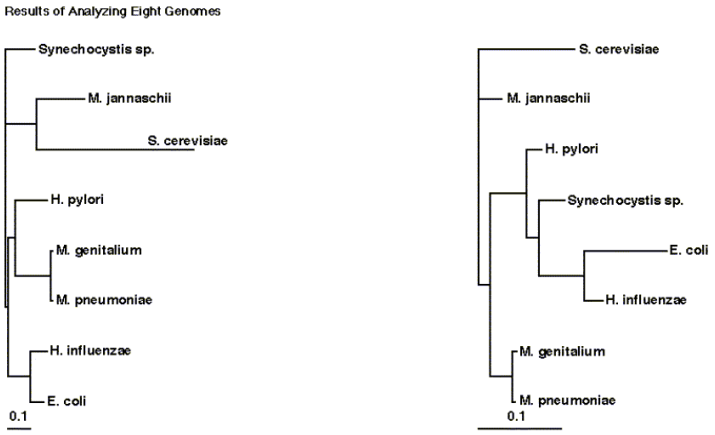
We applied the MPP-based genome comparison method on a group of eight completely sequenced organisms, and then tested how our approach scaled up when the number of organisms was increased by 2 fold and by 4 fold (Figures 1, 2, and 3). We also compared the results to phylogenetic trees based on 16S rRNA analysis (16S rRNA sequences were retrieved from Genbank). In each set, representatives from the three domains of life (Bacteria, Archaea, Eukarya) were included. Trees in Figures 1 to 3 were constructed using *neighbor* and *retree* available in the *Phylip* package [3], and rendered as graphics using *TreeView* [7].

3.1 General observations

With few exceptions, scaling up from 8 to 31 organisms did not affect the relative position of the different organisms on the distance trees generated by the MPP-based approach. This result suggests the adequacy of a hierarchical clustering and robustness of our scoring and algorithm.

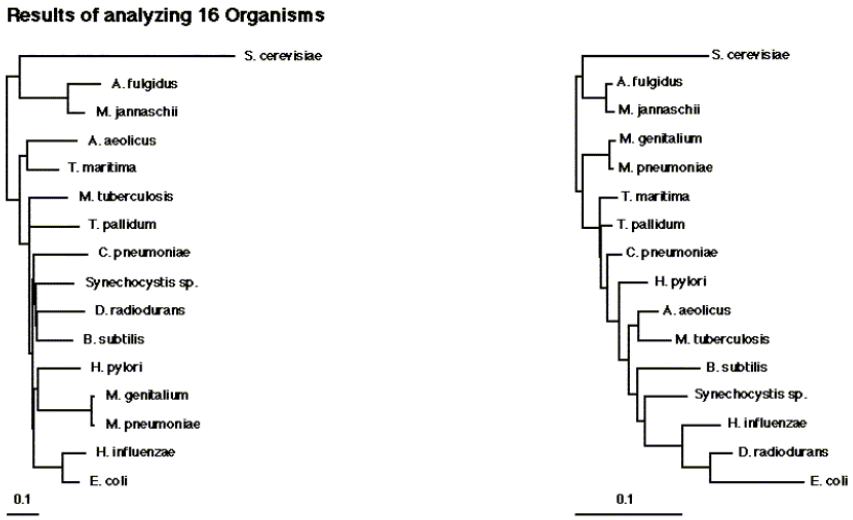
3.2 Similarities between 16S rRNA-based trees and metabolic pathway profile-based trees

Clustering of organisms using the MPP-based approach results in distance trees that are congruent with rRNA-based trees at several levels. The MPP-based approach correctly clusters organisms according to the three domains of life (Figures 1 to 3). Organisms belonging to the archaeal domain are subdivided according to the kingdom lineage they belong to, namely Euryarchaeota (*M. thermoautotrophicum, M. jannaschii, A. fulgidus),* and Crenarchaeota (*A. pernix*) shown in Figures 1 and 3. Furthermore, organisms belonging to the same genus, (*e.g. C. trachomatis,* and *C. pneumoniae; P. abysii,* and *P. horikoshii; M. genitalium,* and *M. pneumoniae*) branch off from the same respective common ancestors. Similarly, the two spirochetes *T. pallidum,* and *B. burgdoferi,* and the gram-positive bacteria (e.g. *B. subtilis, C. acetobutilicum; M. tuberculosis, and M. leprae*) are all closely clustered on the MPP-based trees (Figure 2 and 3).



16S rRNA MPP-based

Figure 1. 16S rRNA and MPP-based trees for eight organisms: Trees were constructed using *neighbor*, a program in the Phylip package [3]. For the 16S rRNA tree, the distance matrix was obtained using *dnadist*, from the Phylip package. For MPP-based trees, the distance matrix was obtained from metabolic pathway profiles using methods presented in this work. Note that the distance scales (0.1), on the 16S rRNA tree and on the MPP-based trees, are not directly comparable, they are derived from sequence difference (or similarity) and from differences in the number of common pathways, respectively



16S rRNA MPP-based

Figure 2. 16S rRNA and MPP-based trees for 16 organisms.

3.3 Discrepancies between the 16S rRNA-based trees and the MPP-based trees

While the above observations are consistent with results from 16S rRNA phylogenetic trees, the relative positions of these clusters on the MPP-based trees are different from the

ones observed on the rRNA-based trees. Also, the position of several organisms on the MPP-based trees is inconsistent with the molecular-based trees. Particularly, the extremely radiation resistant organism, *D. radiodurans*, is positioned in the *E. coli* metabolic lineage; the deep branching organism, *A. aeolicus*, clusters close to the high GC gram-positive Mycobacteria; the cyanobacterium, Synechocystis, branches off from within the proteobacteria cluster. Also, in contrast to the 16S rRNA classification, the proteobacterium *R. prowazekii*, an obligate intracellular parasite, is clustered with the chlamydia group (also obligate intracellular parasites).

The validity of our results is based on the assumptions that the functional role assignments in the WIT database are internally consistent, and the majority of the genes with unknown functions do not represent unknown metabolic pathways. While our first assumption is valid (in the WIT database, the metabolic profiles of related species are very similar or identical) it is difficult to estimate the number of metabolic pathways that remain undiscovered. Nonetheless, the noted deviations from the classical 16S rRNA phylogeny may suggest metabolic pathways undergo evolutions that transcend the boundaries of species and genera. Clustering of metabolic pathway data using different tree alignment distances is underway to study evolutionary of microbial metabolism [10].

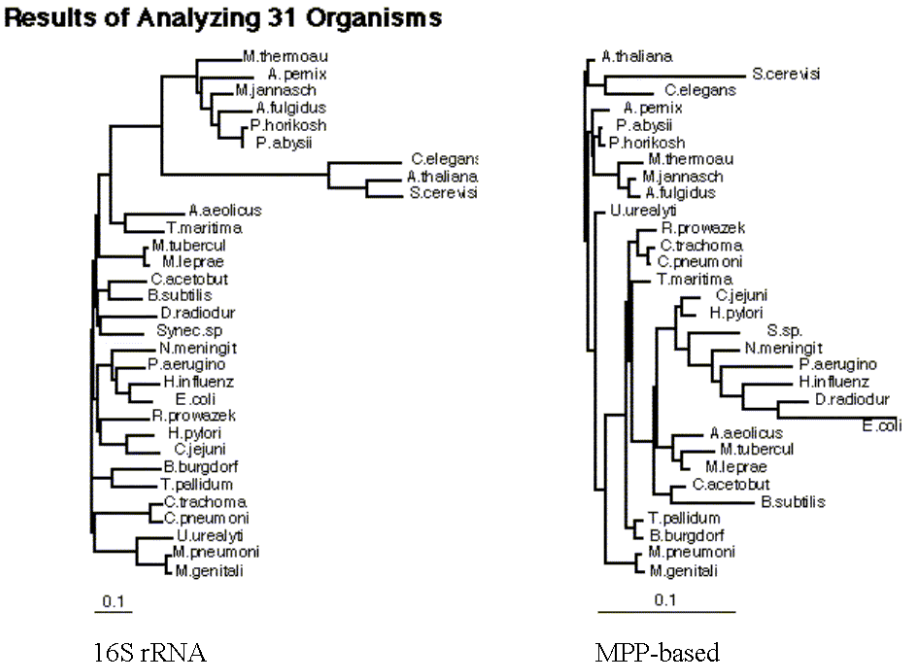


Figure 3. 16S rRNA and MPP-based trees for 31 organisms.

3.4 Conclusion

In summary, we presented a novel approach to utilizing whole genome metabolic pathway information for comparing genomes. A scoring scheme and algorithm were developed for evaluating generic profiles that are based on attributes, which bear hierarchical

relationships. Based on this methodology, organisms are grouped according to metabolic pathway profiles. The results provide a perspective on the relationship among the studied organisms different from the 16S rRNA-based trees (e.g., the branching of *D. radiodurans* and *A. Aquifex* with the *E. coli* and the Gram-positive metabolic lineage respectively). Interpretation of these results would, however, be enhanced if considered in the context of a comprehensive comparative genomics framework. In such a framework, metabolic profiles would be considered as one parameter of likeness to be examined in conjunction with genetic, regulatory, and physiological networks. Our genome wide-based functional comparison of organisms is also relevant to the development of a rational strategy for choosing a new production platform (*e.g., E. coli* would be the preferred recipient of the *D. radiodurans* radiation resistance genes).

Acknowledgements

We thank J. Martin Odom and Jonathan A. Eisen for critical review of the manuscript and constructive comments. Also, we thank the anonymous referees for their valuable comments.

References

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

[2] Durbin, R., Eddy, S., Krough, A., and Mitchison, G., *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.

[3] Felsenstein, J. 1993. PHYLIP (Phylogenetic Inference Package), http://evolution.genetics.washington.edu/phylip/software.html

[4] Forst, Christian V., and Schulten, Klaus, Evolution of Metabolisms: A New Method for the Comparison of Metabolic Pathways Using Genomics Information. *Journal of Computational Biology* **6**(1999)343.

[5] Lin, Jimmy and Gerstein, Mark, Whole-genome Trees Based on the Occurrence of Folds and Orthologs: Implications for comparing Genomes on Different Levels *Genome Research* **10**(2000)808-818.

[6] Overbeek, R., Niels, L., et al, WIT:integrated systems for high throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research* **28**(2000)123-125.

[7] Page, R.D. TreeView: An application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci* **12**(1996)307-331.

[8] Tatusov, R.L., Koonin, E.V., and Lipman, D.J.. A genomic perspective on protein families. *Science* **278**(1997)631-7; Tatusov, R.L., *et al* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. **29**(2001)22-8

[9] Wang, Jason T.L., and Zhang, Kaizhong, Finding similar consensus between trees: An algorithm and a distance hierarchy. *Pattern Recognition* **34**(2001)35-45.

[10] Zhang, S., Liao., L., Tomb, J-F., and Wang, Jason T.L., Clustering of metabolic pathway data using different tree alignment distances. Submitted to *BioKDD* 2002.