BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(7)

# Constructing phylogenetic trees using interacting pathways

**Peng Wan & Dongsheng Che***

Department of Computer Science, East Stroudsburg University, East Stroudsburg, PA 18301; Dongsheng Che – Email: dche@esu.edu; Phone: 1-570-422-2731; *Corresponding author

Received March 24, 2013; Accepted March 25, 2013; Published April 13, 2013

**Abstract:**
Phylogenetic trees are used to represent evolutionary relationships among biological species or organisms. The construction of phylogenetic trees is based on the similarities or differences of their physical or genetic features. Traditional approaches of constructing phylogenetic trees mainly focus on physical features. The recent advancement of high-throughput technologies has led to accumulation of huge amounts of biological data, which in turn changed the way of biological studies in various aspects. In this paper, we report our approach of building phylogenetic trees using the information of interacting pathways. We have applied hierarchical clustering on two domains of organisms—eukaryotes and prokaryotes. Our preliminary results have shown the effectiveness of using the interacting pathways in revealing evolutionary relationships.

**Keywords:** Phylogenetic trees, Metabolome, Hierarchical clustering, Interacting pathways

**Background:**
A phylogenetic tree is a graphic representation of the evolutionary relationships of species, and the phylogenetic distances among the species reflect the closeness of evolutionary relationships. Traditional construction of phylogenetic trees was mainly based on physical similarities and differences. However, the way of the measurement has been changed due to the generation of huge amounts of biological data. For instance, high-throughput sequencing technologies have generated genome sequences in several thousand organisms. A genomic sequence is basically a string of four different kinds of nucleotides (A, C, G and T), with the length from hundreds of thousands to millions. It has been widely accepted that the genomic sequences are highly similar for evolutionary closed organisms, but not similar for evolutionary distant organisms. Therefore, genomic sequences have been widely used for building phylogenetic trees **[1-3].**

The construction of phylogenetic trees using genomic sequences does have some issues. The genomic sequences are usually long, thus comparing genomic sequences across species for building phylogenetic trees is computationally expensive. On the other hand, living organisms in a small niche frequently exchange their genetic materials each other, also known as horizontal gene transfer, making it harder to determine evolutionary relationships based on genomic sequences only. Furthermore, current genomic sequence similarity measurement cannot truly reveal evolutionary relationships across the species. Thus, it is necessary to use other data and methods to reveal true relationships **[4].**

In parallel to the high-throughput genome sequencing technologies, high-throughput metabolic data have also been generated in the past decade. The study of using of metabolic data for biological studies is also known as Metabolomics. The metabolic data from organisms are very informative since they can reveal internal metabolism mechanisms. Theoretically, evolutionary distant species should have different metabolic activities and patterns, while closely related species should have similar metabolic patterns. Therefore, it is desirable to use metabolic data for phylogenetic exploration, or complement the gene-based phylogenetic exploration to some degree.

Metabolic data have been mined and annotated, and corresponding metabolic related databases have been built for scientific communities. For example, KEGG **[5]** is one of databases hosts the Network of Interacting Pathways (NIPs). NIP is a useful resource for phylogenetic distance analysis. The quantitative annotations of metabolic reaction pathways can facilitate for identifying phylogenetic distances **[6].**

As we know, metabolic reaction pathways can be represented as directed or undirected graphs. The nodes in graphs can either be represented as metabolites that are linked by the enzymes, or be represented as enzymes linked by metabolites. Thus, using the information encoded in the graphs can reveal evolutionary relationships across the species.

We can compute phylogenetic distances using actual node information such as enzymes or metabolites in the graphs. Some issues **[6]** of using this approach include: (a) the existence of so-called ubiquitous metabolites. For instance, water, can involve functionally far metabolites together without real biological meaningful representatives; (b) Structures of such networks are highly delicate to inappropriate citation; (c) The induction of some enzymes in a set of reactions may not be the ones which are really involved in those referenced species.

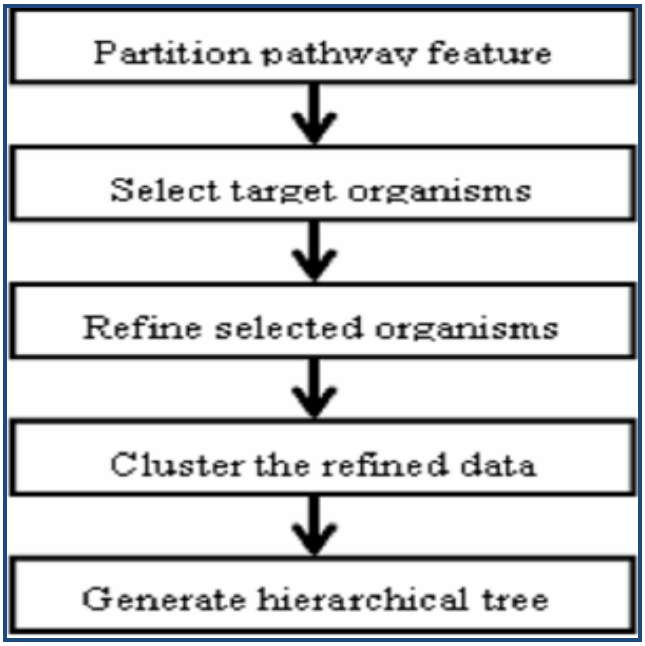


**Figure 1:** Flow Chart of Our Approach.

**Methodology:**

In this paper, we aim to reveal phylogenetic distances across the species using structural information, rather than detailed node information in the graphs. We use the data of metabolic reactions, which are represented as a directed network of interacting pathways (NIPs), and from KEGG Metabolic Relation Network (Directed) Data Set **[7].** In the relation network, enzymes and genes are represented as nodes, while the substrate and product compounds are represented as edges. The related structural information from the graphs was used for computing phylogenetic distances. By utilizing this higher functional approach—metabolic networks, we anticipate revealing phylogenetic relationships across the species.

*Description of data set:*

We collected the processed dataset for KEGG Metabolic Relation Networks from UCI machine learning repository **[7].** This dataset has 53,414 instances in total, each of which represents the derived structural information of a metabolic reaction of any species. The structural information is characterized by 24 features. The features types include integer, real and text. The detailed description of all 24 features for graph structures is listed in **Table 1 (see supplementary material).**

*Data processing*:

The flow chart of our approach is outlined in **(Figure 1)**, and the detailed description of each step is presented as follows:

**(a)** ***Partition pathway feature***: For each instance of the dataset, the first attribute consists of two parts. In order to cluster every species, and calculate the distances based on pathway's variation, the pathway feature must be split into species identifier and pathway identifier. For example, aac00010 is one from the original dataset file, representing the metabolic reaction of Glycolysis Gluconeogenesis (00010) for the species of Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446 (aac). We broke it into aac and 00010.

**(b)** ***Select target species***: The original dataset covered 788 species (including eukaryotes and prokaryotes), and 117 pathways **Table 2 (see supplementary material)** shows the first ten pathway IDs and names. Due to the fact that not all species have all 117 pathways, we want to select those species that have at least minimum number of pathways for our study. To do so, we constructed a zero-one matrix which is based on species and pathways. Based on the zero-one matrix, we set a threshold of pathways' number so that we pick the species with large number of related pathways.

**(c)** ***Refine the target instances***: Even if we pick those species with high number of pathways, we cannot guarantee that any two species have the same pathways. In most cases, the pathways in these species could be overlapped. For instance, species #1 may contain pathways #10, #20 and #30, while species #2 may contain pathways #10, #20 and #40. Most of selected species will miss some pathways out of 117 pathways, which is problematic when such data are applied in distance calculation in next step. To resolve this problem, we filled those missing pathways using the average values of existing pathways.

**(d)** ***Calculate Euclidean distances***: Generally, the input for the clustering algorithms is the similarity or distance matrix between entities in the data. There are various distance (or similarity) metrics, and they all produce broadly similar results **[8].** In this study, we have chosen the most commonly used distance metric, Euclidean distance **(For formula see supplementary material).**

*Clustering algorithm*:

Clustering is an unsupervised learning algorithm that finds the hidden structure in the unlabeled data. In this study, we used

the hierarchical clustering, which is based on the core idea of objects being more related to nearby objects than to objects farther away. We choose Cluster 3.0 **[9]** to cluster our processed data. The results were visualized by Java TreeView **[10].**

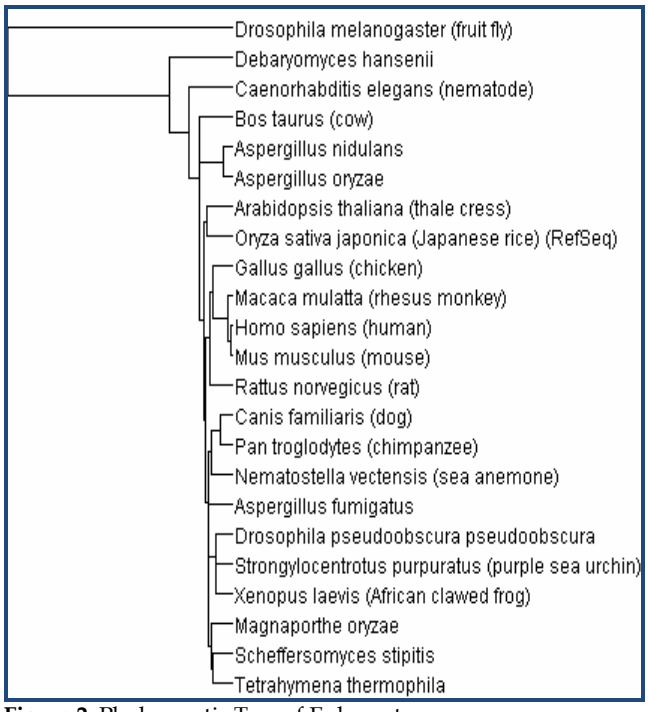


**Figure 2:** Phylogenetic Tree of Eukaryotes.

**Discussion:**

In order to test the effectiveness of our approach, we separated the original dataset into subgroups, one group for the eukaryotic species, and other group for the prokaryotic species. Thus, we produced two distance matrices, one for eukaryotic organisms, and the other for the prokaryotic organisms. For the eukaryotic organisms, we set the threshold of 80 for the minimum number of pathways for each species, and we selected 23 species, and a distance matrix of 23 × 23 was produced. For the prokaryotic organisms, we set the threshold of 90 for the minimum number of pathways for each species, and we selected 30 species, and a distance matrix of 30 × 30 was produced.

We have applied Cluster 3.0 on these two groups of datasets, and used Java TreeView to generate the dendrograms (or phylogenetic trees) for each of the dataset. **Figure 2 & 3** show the dendrograms of eukaryotic species and prokaryotic species respectively, with the lengths of the branches reflecting the distances between species. Therefore, the shorter the branches, the evolutionarily closer the species are, and the longer the branches, the evolutionarily more distant the species are.

From the phylogenetic tree in **(Figure 2)**, we can find some interesting results. For instance, we can see the closest species with human beings is *Mus musculus*, so-called house mouse. To understand why these two species stay close, we did literature search about the closeness of these two species. We found that almost all genes in the mouse were also present in humans. Actually, researchers have reported that approximately 99% of mouse genes have counterparts in humans **[11]**. Therefore, it is not such a surprise that the phylogenetic distance between mice and humans, In **Figure 2**, we can also see that the two plant species, *Arabidoposis thaliana*, and *Oryza sativa japonica* (also known as Japanese rice) stay close in the dendrogram, which is consistent with our expectations. **Figure 3** shows the dendrogram of 30 prokaryotic organisms. Like the tree for the eukaryotes, most of phylogenetically close species stay very close. For instance, two *Escherichia coli* (O1 39:H28 E24377A, and O9 HS) are very close in the tree, two *Bradyrhizobium* (sp. BTAi1 and sp. ORS 278) are in the same branch. The dendrogram result for the prokaryotes also strongly indicates the effectiveness of our approach.

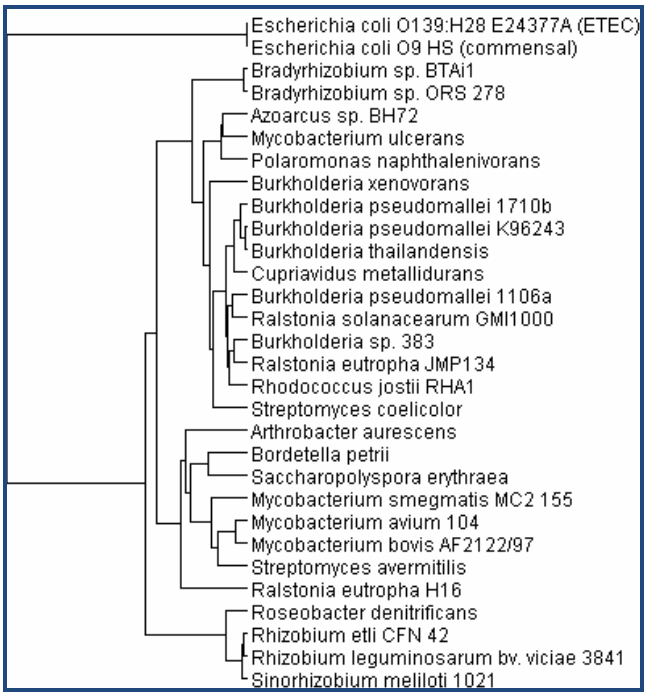


**Figure 3:** Phylogenetic Tree of Prokaryotes.

**Conclusion:**

In this paper, we have reported our approach that uses the information of metabolic reactions to unveil the relation of evolution among species. The main contribution of this research is to demonstrate that with the usage of clustering, the phylogeny of species can be constructed by a higher level functional component—metabolomics. While we only relied on the structural information from these metabolic reactions, our experimental results have shown that our approach is pretty accurate in most of cases, strongly indicating that effectiveness of our approach.

We do realize that the evolutionary distances for some species were not accurately characterized in our study. The inaccurateness could be caused by the followings: **i)** The missing data of pathways within some species, where we used the average values to fill them up; **ii)** for each pathway, we treated them the same weights in computing the distance of phylogeny. We hope more metabolic data will be available to fill the gap, and a sophisticated weight schemes for different pathway may help improve our approach.

**Acknowledgement:**
This research was partially supported by President Research Fund, FDR major grant, and FDR mini grant at East Stroudsburg University, USA.

**References:**

**[1]** Snel B *et al. Nat Genet*. 1999 **21**: 108 [PMID: 9916801]
**[2]** Li Ming *et al. Bioinformatics*. 2001 **17**: 149 [PMID: 11238070]
**[3]** Herniou EA *et al. Journal of virology*. 2001 **75**: 8117 [PMID: 11483757]
**[4]** Ma HW & Zeng AP, *Bioinformatics*. 2003 **19**: 270 [PMID: 12538249]
**[5]** http://www.genome.jp/kegg/
**[6]** Mazurie A *et al. Bioinformatics*. 2008 **24:** 2579 [PMID: 18820265]
**[7]** http://archive.ics.uci.edu/ml/datasets/KEGG+Metabolic+Relation+Network+(Directed)
**[8]** Taylor J *et al. Bioinformatics*. 2002 **18 Suppl 2**: 241 [PMID: 12386008]
**[9]** de Hoon MJ *et al. Bioinformatics*. 2004 **20**: 1453 [PMID: 14871861]
**[10]** Saldanha AJ, *Bioinformatics*. 2004 **20:** 3246 [PMID: 15180930]
**[11]** http://www.ebi.ac.uk/2can/genomes/eukaryotes/Mus_musculus.html

**Edited by P Kangueane**
**Citation**: **Wan & Che,** Bioinformation 9(7): 363-367 (2013)

BIOINFORMATION open access

# Supplementary material:

**Methodology:**

***Calculate Euclidean distances*:**

Generally, the input for the clustering algorithms is the similarity or distance matrix between entities in the data. There are various distance (or similarity) metrics, and they all produce broadly similar results [8]. In this study, we have chosen the most commonly used distance metric, Euclidean distance. The formula is as follows.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

where *p* and *q* are vectors with *n* number of feature values. In our study, we computed the distance between two species using 117 pathways, with each pathway containing 23 features (See Table 1). Thus, the total number of features in our study is 117 × 23 = 2,691. We calculated Euclidean distance for selected eukaryotic and prokaryotic species. Table 3 shows the partial 2-dimensional distance matrix for eukaryotes.

**Table 1:** Feature Description for Our Dataset.

| No. | Feature Info. | Value Type | Range |
|---|---|---|---|
| 1 | Pathway | text | NULL |
| 2 | Nodes | integer | [2, 116] |
| 3 | Edges | integer | [1, 606] |
| 4 | Connected components | integer | [1, 13] |
| 5 | Network Diameter | integer | [1, 30] |
| 6 | Network Radius | integer | [1, 2] |
| 7 | Shortest Path | integer | [1, 3277] |
| 8 | Characteristic Path Length | real | [1, +∞) |
| 9 | Avg.numNeighbours | real | [1, +∞) |
| 10 | Isolated Nodes | integer | [0, 1] |
| 11 | Number of Self Loops | integer | [0, 0] |
| 12 | Multi-edge Node Pair | integer | [0, 57] |
| 13 | Neighborhood Connectivity | real | [1, +∞) |
| 14 | Outdegree | real | [0.5, +∞) |
| 15 | Stress | real | [0, +∞) |
| 16 | SelfLoops | integer | [0, 0] |
| 17 | PartnerOfMultiEdgedNodePairs | real | [0, +∞) |
| 18 | EdgeCount | real | [1, +∞) |
| 19 | BetweennessCentrality | real | [0, +∞) |
| 20 | Indegree | real | [0.5, +∞) |
| 21 | Eccentricity | real | any |
| 22 | ClosenessCentrality | real | (0, 1] |
| 23 | AverageShortestPathLength | real | any |
| 24 | ClusteringCoefficient | real | [0, +∞) |

**Table 2:** A List of Ten Pathway Names.

| Identifier | Name |
|---|---|
| 00010 | Glycolysis Gluconeogenesis |
| 00020 | Citrate cycle (TCA cycle) |
| 00030 | Pentose phosphate pathway |
| 00031 | Undocumented |
| 00040 | Pentose and glucuronate interconversions |
| 00051 | Fructose and mannose metabolism |
| 00052 | Galactose metabolism |
| 00053 | Ascorbate and aldarate metabolism |
| 00061 | Fatty acid biosynthesis |
| 00062 | Fatty acid elongation |

**Table 3:** Euclidean Distance Matrix for 10 Eukaryotes Species. The 10 eukaryotic species in this table are *afm*, *ani*, *aor*, *ath*, *bta*, *cel*, *cfa*, *dha*, and *dme*.

| | S#1 | S#2 | S#3 | S#4 | S#5 | S#6 | S#7 | S#8 | S#9 | S#10 |
|---|---|---|---|---|---|---|---|---|---|---|
| S#1 | 0 | | | | | | | | | |
| S#2 | 770.277 | 0 | | | | | | | | |
| S#3 | 774.409 | 242.721 | 0 | | | | | | | |
| S#4 | 1070.79 | 592.15 | 643.246 | 0 | | | | | | |
| S#5 | 1609.22 | 1433.93 | 1456.4 | 1353.45 | 0 | | | | | |
| S#6 | 1111.33 | 767.811 | 777.271 | 801.645 | 1025.74 | 0 | | | | |
| S#7 | 1559.44 | 1785.64 | 1855.92 | 2030.91 | 1984.29 | 1869.18 | 0 | | | |
| S#8 | 1023.76 | 885.934 | 926.689 | 842.742 | 1015.88 | 832.66 | 1857.57 | 0 | | |
| S#9 | 8554.11 | 8581.38 | 8581.43 | 8604.97 | 8147.58 | 8368.75 | 8582.12 | 8381.47 | 0 | |
| S#10 | 901.819 | 763.328 | 853.568 | 982.096 | 1523.79 | 868.906 | 1324.21 | 1067.3 | 8535.55 | 0 |