Computational Phylogeneticity of Biological Pathways: A developmentalstudy of TCA cycle over a set of organisms

Chapter · January 2011

CITATIONS
0

READS
303

3 authors:

Losiana Nayak
Indian Statistical Institute
25 PUBLICATIONS **125** CITATIONS
SEE PROFILE

Dr. Namrata Tomar
Medical College of Wisconsin
54 PUBLICATIONS **404** CITATIONS
SEE PROFILE

Rajat Kumar De
Indian Statistical Institute
158 PUBLICATIONS **1,515** CITATIONS
SEE PROFILE

Chapter 14

Computational Phylogenicity of Biological Pathways : A Study of Some TCA Cycles

LOSIANA NAYAK, NAMRATA TOMAR AND RAJAT K. DE
Machine Intelligence Unit, Indian Statistical Institute
203 B.T. Road, Kolkata-700108, West Bengal, India

ABSTRACT

Biological Pathways represent system level sophistication of organisms. An exclusive peek into the extent of sophistication of an organism may provide proof of its past evolutionary struggle and position in the developmental trend from pathway point of view. This knowledge may not align with the traditional universally acknowledged standard 16s rRNA trend. Analyzing the reasons behind such discrepancy can constitute an important research work. To map sophistication of one organism with respect to other organisms, phylogeny is required. Here, we are mapping the level of sophistication in terms of enzymes among a set of TCA cycles belonging to different species, arbitrarily chosen from the KEGG/PATHWAY database. For this purpose, we converted the metabolic pathway information into simplified enzyme-enzyme relational (*E-E-R*) graphs. Similarity scores obtained from inter-genus, intra-genus as well as inter-species comparison of these graphs are used for phylogenetic tree construction. These phylogenetic trees throw light on the trend of development of metabolic pathways among different species. But peculiarly, these trees have less similarity with the conventional evolutionary phylogenetic trees constructed from NCBI taxonomy data. In this paper, we have tried to find some justifications for this dissimilarity among phylogenetic trees, by considering their habitats.

Phylogenetics

Phylogenetics is the study of evolutionary relatedness among various groups of organisms. The term phylogenetics is of Greek origin. It is created from the terms phyle/phylon, meaning "tribe, race," and genetikos, meaning "relative to birth" from genesis ("birth"). Evolution is regarded as a branching process, whereby populations are altered over time and may speciate into separate branches, hybridize together, or terminate by extinction. This may be visualized in a phylogenetic tree.

A phylogenetic tree or evolutionary tree is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics. Early representations of branching phylogenetic trees include a "palaeontological chart" showing the geological relationships among plants and animals in the book Elementary Geology, by Edward Hitchcock [1]. Charles Darwin (1859) also produced one of the first illustrations of an evolutionary tree in his book The Origin of Species [2]. Some of the earlier research papers mentioning the notion of phylogenetic tree are

By H. Marshall Ward (1884) Quoted "Apart from their interests more directly affecting mankind, the Fungi have seemed to present problems of life in some respects simpler than other forms, and have thus in a manner promised a solution of phylogenetic and physiological questions more nearly approaching the ideal of the evolutionist." [3]

By F.O. Bower (1889) Paper titled "The comparative examination of the meristems of ferns, as a phylogenetic study." [4]

By J.S. Kingsley (1894) Paper titled "The classification of the Arthropoda." [5]

By J.E.S. Moore (1898) Quoted "The characters of the nervous system of Typhobia show thus in a manner which does not appear to be capable of serious disputation, that this Gastropod has no relation to, nor indeed any but the most remote phylogenetic connection with, the hitherto recognized fresh-water forms." [6]

By F.F. Blackman (1900) Quoted "As a result we seem now at last to be in a position to grasp something of the phylogenetic relations of what one seemed a chaos of forms, and to correct by the evidence of these vestiges of the early stages of evolution of the vegetable kingdom our conception of plant nature and plant possibilities drawn previously only from the study of the higher types." [7]

A phylogenetic tree can be represented in various forms. Rooted trees are directed trees with the notion of a common ancestor for the nodes at the leaf level. Unrooted trees illustrate the relatedness of the leaf nodes without making assumptions about ancestry. Dendrogram is a broad term for the diagrammatic representation of a phylogenetic tree. Cladograms are formed using cladistic methods. They only represent branching pattern not evolutionary time. In a phylogram, branches reflect the number of changes that have taken place in a particular DNA sequence in that lineage through their branch lengths. A Chronogram represents evolutionary time through its branch lengths.

General Methods to infer Phylogenetic Trees

A Phylogenetic tree can be created by various methods, depending on the type of data available and the type of information it should contain. In this section we tried to describe briefly some of the popularly used methods to create phylogenetic trees with their positive and negative aspects. Some methods are particularly useful for certain branches of science. Some are used to create trees of a set of closely related species or reproductively isolated populations of a single species.

Cladistics or phylogenetic systematics

Cladistics is about how to use the past to understand the present. It is an approach to classify the living things in which organisms are defined and grouped by the possession of one or more shared characteristics (called characters) that are derived from a common ancestor. It is a method of reconstructing evolutionary relationships that emphasizes the importance of descent and common ancestry rather than chronology. The word 'clade' was coined by Julian Huxley in 1957 for referring characteristics that can be used as units for setting limits to classes and establishing hierarchies which may or may not be associated with evolutionary sequences [8]. Cladistics places species in a group, or clade, based on a shared character. Within a clade, species that share other characters unique to them are grouped together, and so on, until a cladogram is assembled. In living creatures, genetic characters or behaviors as well as more obvious anatomical features might be considered in assembling a cladogram. Cladistics is especially significant in paleontology, as it points out gaps in the fossil evidence. Cladistic analytical methods have application in genetic epidemiology, conservation biology, basic evolutionary biology and species inference [9]. Now-a-days people have started looking beyond cladistics to extend evolutionary classifications into deeper time levels [10].

Maximum Parsimony

Parsimony is a non-parametric statistical method for estimating phylogenies that require least evolutionary change to explain some observed

data. The word derives from Middle English parcimony, from Latin parsimonia, from parsus, past participle of parcere, whose meaning is "to spare". Parsimony is part of a class of character-based tree estimation methods which use a matrix of discrete phylogenetic characters to infer one or more optimal phylogenetic trees for a set of taxa, commonly a set of species or reproductively isolated populations of a single species. The tree with the most favorable score is taken as the best estimate of the phylogenetic relationships of the included taxa.

Maximum parsimony is a cladistic "optimality criterion" method based on the principle of parsimony. The method can be used with most kinds of phylogenetic data. It is often criticized as being a statistically unsound method that fails to make an explicit underlying "model" of evolution [11]. The branch bound algorithm [12, 13], The Sankoff-Morel-Cedergren algorithm [14, 16] and MALIGN [15] and POY [17] are some of the algorithms that operate following this methodology.

Maximum-likelihood method

The method uses standard statistical techniques for inferring probability distributions to assign probabilities to possible phylogenetic trees. It is broadly similar to the maximum-parsimony method, but allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites. It assumes that "evolution at different sites and along different lineages must be statistically independent". New algorithms and methods to estimate maximum-likelihood phylogenies are getting devised with increasing speed [18]. Some of them are dedicated for making huge alignments possible [19]. One of the strengths of the maximum likelihood method of phylogenetic estimation is the ease with which the hypotheses can be formulated and tested [20]. It is well suited to the analysis of distantly related sequences, but because it requires search of all possible combinations of tree topology and branch length, it is computationally expensive to perform on more than a few sequences. With Maximum Likelihood Method a doubt always persists that how correct is the choice of underlying parameters/distributions [11].

Bayesian Inference1

Bayesian methods assume a prior probability distribution of the possible trees, which may simply be the probability of any one tree among all the possible trees that could be generated from the data, or may be a more sophisticated estimate derived from the assumption that divergence events such as speciation occur as stochastic processes. It can be used to produce phylogenetic trees in a manner closely related to the maximum likelihood methods. Choice of prior distribution is a point of contention among the users of Bayesian-inference based phylogenetics methods [16].

Implementations of Bayesian methods generally use Markov chain Monte Carlo sampling algorithms, although the choice of move set varies. Selections used in Bayesian phylogenetics include circularly permuting leaf nodes of a proposed tree at each step and swapping descendant subtrees of a random internal node between two related trees [21]. Improvisations have been done in the form of Bayesian Markov chain Monte Carlo method for the multi-species coalescent[1*] [22]. Certain models with new features like preliminary data smoothing process and no hypotheses on the rate matrix of the Markov process are also present [23]. The use of Bayesian methods in phylogenetics has been controversial, largely due to incomplete specification of the choice of move set, acceptance criterion, and prior distribution in published work [16].

Phenetics

Phenetics (taximetrics) is an attempt to classify organisms based on overall similarity like morphology or other observable traits regardless of their evolutionary relation. It is closely related to numerical taxonomy which is concerned with the use of numerical methods for taxonomic classification. Phenetics has largely been superseded by cladistics for research into evolutionary relationships among species. However, certain phenetic methods (neighbor-joining) have found their way into cladistics as a reasonable approximation of phylogeny when more advanced methods (Example: Bayesian inference) are too computationally expensive. Phenetic techniques include various forms of clustering and ordination. These are sophisticated ways of reducing the variation displayed by organisms to a manageable level. The process involves measuring of dozens of variables and their presentation as two or three dimensional graphs. Most of the technical challenges in phenetics revolve around balancing the loss of information in such a reduction against the ease of interpreting the resulting graphs.

Distance matrix methods

These methods rely on a measure of "genetic distance" between the sequences and require a multiple sequence alignment as an input. Distance is often defined as the fraction of mismatches at aligned positions with gaps either ignored or counted as mismatches [24]. Distance methods attempt to construct an all-to-all matrix from the sequence query set describing the distance between each sequence pair. A phylogenetic tree created from this

[1] Coalescent theory is a metrospective model of population genetics. It employs a sample of individuals from a population to trace all alleles of a gene shared by all members of the population to a single ancestral copy, known as the Most Recent Common Aneestor (MRCA).

matrix places closely related sequences under the same interior node whose branch lengths closely reproduce the observed distances between sequences. These methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them. They are frequently used as the basis for progressive and iterative types of multiple sequence alignments. The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions that appear across multiple subtrees [16].

Neighbor-joining methods: The simple neighbor-joining method produces unrooted trees, but it does not assume a constant rate of evolution. On the other hand, UPGMA (Unweighted Pair Group Method with Arithmetic Mean) produces rooted trees and requires a constant-rate assumption.

The Fitch-Margoliash method: It uses a weighted least squares method for clustering based on genetic distance [25]. Closely related sequences are given more weight in the tree construction process to correct for the increased inaccuracy in measuring distances between distantly related sequences. The distances used as input to the algorithm must be normalized to prevent large artifacts in computing relationships between closely related and distantly related groups. The distances calculated by this method must be linear. The least-squares criterion applied to these distances is more accurate but less efficient than the neighbor-joining methods.

Using outgroups: Independent information about the relationship between sequences or groups can be used to help reduce the tree search space and root unrooted trees. Standard usage of distance-matrix methods involves the inclusion of at least one outgroup sequence known to be only distantly related to the sequences of interest in the query set [24]. An appropriately chosen outgroup will have a much greater genetic distance and a longer branch length than any other sequence and it will appear near the root of a rooted tree. Choosing an appropriate outgroup requires the selection of a sequence that is moderately related to the sequences of interest. Too close a relationship defeats the purpose of the outgroup and too distant adds noise to the analysis [24]. Care should also be taken to avoid situations in which the species from which the sequences were taken are distantly related but the gene encoded by the sequences is highly conserved across lineages. Horizontal gene transfer, especially between otherwise divergent bacteria, can also confound outgroup usage.

A series of simulation studies exploring the relative performance of several phylogenetic network approaches (statistical parsimony, split decomposition, union of maximum parsimony trees, neighbor-net, simulated history

recombination upper bound, median-joining, reduced median joining and minimum spanning network) compared to standard tree approaches, (neighbor-joining and maximum parsimony) in the presence and absence of recombination was done by [26].

Their findings suggested that in the absence of recombination, all methods recovered the correct topology and branch lengths nearly all of the time when the substitution rate was low, except for minimum spanning networks, which did considerably worse. At a higher substitution rate, maximum parsimony and union of maximum parsimony trees were the most accurate. With recombination, the ability to infer the correct topology was halved for all methods and no method could accurately estimate branch lengths. Hence, there is need for more accurate phylogenetic network methods by improvements of existing methods and development of novel algorithms.

Bioinformatics tools and softwares dedicated to phylogeny

These days many bioinformatics tools, softwares and databases are available for creating phylogenetic trees. Some of them are stand-alone applications, while others are web-based. Some are individual applications, while others allow different levels of sophistication in building pipelines of applications. In this section, we describe a few recent tools and servers along with a database of phylogenetic trees. All of them can be accessed from different websites listed in Table 1.

Table 1 : Tools, Software and Databases for Phylogeny

Names	URLs	Remarks
AlignX	http://tools.invitrogen.com/content.cfm?pageid = 10192	Vector NTI module that performs MSA and displays multi-color graphics
GeConT 2	http://bioinfo.ibt.unam.mx/gecont	Does comparative genome analysis of related genes to search for potential conserved regulatory motifs
DIVEIN	http://indra.mullins.microbiol.washington.edu/DIVEIN	Estimates evolutionary parameters and phylogenetic trees
MAVID	http://bio.math.berkeley.edu/mavid/download/	MSA program suitable for many large genomic regions

Contd...

Phylogeny.fr	http://www.phylogeny.fr/	MSA, phylogenetic reconstruction and graphical representation of trees
Tcoffee@igs	http://igs-server.cnrs-mrs.fr/Tcoffee/	Aligns protein, RNA or DNA sequences and processes datasets of up to 100 sequences
M-Coffee	www.tcoffee.org	Assembles MSA by combining the output of several individual methods into one single MSA
PhyloView	http://www.ogic.ca/projects/phyloview/	Tool for coloring phylogenetic trees upon arbitrary taxonomic properties
PHYML	http://atgc.lirmm.fr/phyml	Estimates maximum likelihood phylogenies from DNA and protein sequences
SATCHMO-JS	http://makana.berkeley.edu/q/satchmo/	Employs Multiple alignment using Fast Fourier Transform (MAFFT) iterative MSA method to align closely related sequences
Signature	http://www.cmbi.ru.nl/signature	Identifies and phylogenetically characterizes the signature genes in a set of query sequences
SoRT[2]	http://genome.cs.nthu.edu.tw/SORT2/[1]	For genome rearrangement analysis and inferring phylogenetic trees
SplitsTree[4]	http://www-ab.informatik.uni-tuebingen.de/software/splitstree4/welcome.html	Framework for tree and network oriented phylogenetic analysis
TreeDomViewer	http://www.bioinformatics.nl/tools/treedom/	Dedicated to alignment of structural domains
CVTree	http://tlife.fudan.edu.cn/cvtree	New whole genome-based, alignment-free composition vector (CV) method for phylogenetic analysis
iTOL	http://itol.embl.de	Manipulates and creates phylogenetic trees in 'New Hampshire' or Newick format
POWER	http://power.nhri.org.tw/power/home.htm	Uses ClustalW and PHYLIP and generates a high-quality tree topology

Contd...

MEGA4	http://www.megasoftware.net/	Maximum Composite Likelihood (MCL) method to estimate evolutionary distances between all pairs of sequences simultaneously with or without incorporating rate variation among sites and substitution pattern heterogeneities among lineages
TreeFam	http://www.treefam.org, http://treefam.genomics.org.cn	Contains curated trees for 690 families and automatically generated trees for another 11,646 families

AlignX [27] is the Vector NTI module that performs multiple sequence alignments and displays them with easily interpretable multi-color graphics. Based on the popular ClustalW algorithm, AlignX incorporates features including profile alignment, secondary structure consideration, automatic consensus calculation, graphic display of a phylogenetic tree, dot matrix comparison and some alignment editing capabilities. But it doesn't have flexibility of ClustalX to align only selected regions or molecules directly in the existing alignment. AlignX's profile alignment is limited and only one sequence or existing alignment can be used as the profile. It supports only MSF (Multiple Sequence File) format as alignment format. FASTA, ClustalW, PHYLIP and NEXUS[2] formats need to be supported in order to gain better compatibilities with other types of alignments and phylogeny programs.

The Gene Context Tool [28] (GeConT) allows users to visualize the genomic context of a gene or a group of genes and their orthologous relationships within fully sequenced bacterial genomes. An updated version of GeConT is also available as GeConT 2 [29]. This updated version has increased query options. An important feature of GeConT 2 is its potential to do comparative genome analysis of related genes to look for potential conserved regulatory motifs.

DIVEIN [30] estimates evolutionary parameters and phylogenetic trees while allowing the user to choose from a variety of evolutionary models. It reconstructs the consensus (CON), most recent common ancestor (MRCA), and center of tree (COT) sequences. It also provides tools for condensing sequence alignments (to show only informative sites or private mutations), computing phylogenetic or pairwise divergence from any user-specified

[2] Modular format with a file consisting of separate blocks each containing one particular kind of information and consisting of standardized commands. Public blocks contain information about taxa, morphological and molecular characters, distances, genetic codes, assumptions, sets, trees, etc., while private blocks contain information of relevance to single programs.

sequence (CON, MRCA, COT or existing sequence from the alignment), computing and outputting all genetic distances in column format, calculating summary statistics of diversity and divergence from pairwise distances and finally, graphically representing the inferred tree and plots of divergence, diversity, and distance distribution histograms. It accepts aligned nucleotide or amino acid sequences in NEXUS, PHYLIP, or FASTA format.

MAVID [31] is a multiple alignment program suitable for many large genomic regions and it is fast being capable of aligning hundreds of kilobases in less than a minute. It is suitable for the alignment of mitochondrial sequences, viral genomes and other data sets. The output is organized in such a way that conserved regions between subsets of sequences can be quickly identified for further investigation.

Phylogeny.fr [32] platform offers a variety of programs to automatically perform leading methods for multiple sequence alignment, phylogenetic reconstruction and graphical representation of trees, and chains these methods into a pipeline that can be executed in three modes. The 'One Click' mode is suitable for non-specialists and provides a ready-to use pipeline of programs with recognized accuracy and speed like MUSCLE [33] for multiple alignment, PhyML [34, 35] for tree building, and TreeDyn [36] for tree rendering. The 'Advanced' mode uses the same pipeline but allows the parameters of each program to be customized by users. The 'A la Carte' mode offers more flexibility and sophistication, as users can build their own pipeline by selecting and setting up the required steps from a large choice of tools to suit their specific needs.

Tcoffee@igs [37] uses the latest version of the T-Coffee package. Given a set of unaligned sequences, the server returns an evaluated multiple sequence alignment and the associated phylogenetic tree. Tcoffee@igs can be used for aligning protein, RNA or DNA sequences and it can process datasets of up to 100 sequences (2000 residues long).

M-Coffee [38, 39] is an extension of T-Coffee. It is an open source package distributed under a GPL license. It is available in both stand-alone and web-server from. It is a meta-method for assembling multiple sequence alignments (MSAs) by combining the output of several individual methods into one single MSA. The combination procedure is a rather robust process able to cope with a significant amount of noise. It uses consistency to estimate a consensus alignment by assuming that incorrect alignments are less likely to be consistent than correct ones. Best results can be obtained if the right combination of methods can be selected carefully. The main issue with such a selection is that it may be hard to automate the whole process and will always require expert knowledge. M-Coffee is very similar to the standard T-Coffee, but it does not require the estimation of the pairwise library.

PhyloView [40] is a web based tool for coloring phylogenetic trees upon arbitrary taxonomic properties of the species represented in a protein sequence phylogenetic tree.

PHYML (PHYlogenetic inferences using Maximum Likelihood) estimates maximum likelihood phylogenies from DNA and protein sequences [34,35]. It provides the user with a number of options like nonparametric bootstrap and estimation of various evolutionary parameters, in order to perform comprehensive phylogenetic analyzes on large datasets in reasonable computing time.

The SATCHMO-JS webserver [41], an extension of the SATCHMO algorithm (jump-start SATCHMO) estimates protein multiple sequence alignments (MSAs) and phylogenetic trees simultaneously. The server takes a set of sequences in FASTA format as input and outputs a MSA based phylogenetic tree. It employs a divide-and-conquer strategy to jump-start SATCHMO at a higher point in the phylogenetic tree, reducing the computational complexity of the progressive all versus all HMM-HMM[3] scoring and alignment. It employs computationally efficient Multiple Alignment using Fast Fourier Transform[4] (MAFFT) iterative MSA method to align closely related sequences and saving the use of computationally expensive HMM-HMM alignment for only those subgroups that are more distantly related.

Signature genes are unique to a taxonomic clade and are common within it. They have an enormous amount of information about clade-specific processes and contain a strong evolutionary signal that can be used phylogenetically to characterize a set of sequences, such as a metagenomic sample. As signature genes are based on gene content, they provide a means to assess the taxonomic origin of a sequence sample that is complementary to sequence-based analyzes. Signature [42] is a web server that identifies the signature genes in a set of query sequences and phylogenetically characterizes them. The server produces a list of taxonomic clades that share signature genes with the set of query sequences, along with an insightful image of the tree of life in which the clades are color coded based on the number of signature genes present.

SoRT[2] [43] is a web server that allows the user to perform genome rearrangement analysis involving reversals, generalized transpositions and

---

[3] The Hidden Marker Model is a finite set of states, each of which is associated with a probability distribution. In a particular state an observation can be generated according to the associated probability distribution. It is only the outcome, not the state visible to an enternal observer. The states are "hidden", hence the name Hidden Markov Model.

4 An algorithm for computing the fourier transform of a set of discrete data values.

translocations (including fusions and fissions) and infer phylogenetic trees of genomes based on their pairwise genome rearrangement distances.

The SplitsTree program [44, 45] is based on so-called splits and phylogenetic networks. It is aimed at providing a general framework for both tree- and network-oriented phylogenetic analysis. Fundamental data types supported by the program include unaligned and aligned sequences, distances, splits, trees, networks and quartets. The freely available package provides commonly used distance-based algorithms. The main features of the program are visualization of phylogenetic trees and networks, interactive exploration of the visualization, bootstrapping and transformations of molecular sequences to distances.

TreeDomViewer [46] is a visualization tool available as a web-based interface that combines phylogenetic tree description, multiple sequence alignment and InterProScan [47] data of sequences. It generates a phylogenetic tree projecting the corresponding protein domain information onto the multiple sequence alignment. It adopts an evolutionary perspective on how domain structure of two or more sequences can be aligned and compared to subsequently infer the function of an unknown homolog. One feature of major importance in TreeDomViewer is the alignment of structural domains. This allows for quick checking of the alignment quality, easy inference of homology even when the sequence residue similarity is very low and support for the phylogeny based on functional characteristics evidences.

The CVTree web server [48] is a new implementation of the whole genome-based, alignment-free composition vector (CV) method for phylogenetic analysis. It is an alignment and parameter free phylogenetic tool using composition vectors (CVs) inferred from whole genome data. It can find the phylogenetic position of the user's-specific genome data from the server's in-built database.

Interactive Tree Of Life (iTOL) [49] is a web-based tool for the display, manipulation and annotation of phylogenetic trees. It provides an easy way to manipulate and create customized graphical representations of phylogenetic trees in the standard 'New Hampshire' or Newick format. Various types of data such as genome sizes or protein domain repertoires can be mapped onto the tree. It can automatically determine taxonomic classes of all internal nodes and assign proper scientific names to the leaves. It supports the export of several bitmap and vector graphics formats.

POWER, the PhylOgenetic WEb Repeater [50], is a web based service designed to perform user-friendly pipeline phylogenetic analysis. It uses an open-source LAMP structure [(Linux operating system), Apache (web server), MySQL (relational database) and PHP (html-embedded scripting language)].

It infers genetic distances and phylogenetic relationships using well established algorithms (ClustalW and PHYLIP). It incorporates a novel tree builder based on the graphical display library (.png format) to generate a high-quality tree topology according to the calculated result. Tracking changes in a tree that result from repeated parameter tuning and addition or deletion of sequences is difficult. POWER overcomes this problem by providing detailed logs of user-defined parameters, all output files and process history.

MEGA4 [51, 52], a fourth version of MEGA software, expands on the existing facilities for editing DNA sequence data from autosequencers, mining web-databases, performing automatic and manual sequence alignment, analyzing sequence alignments to estimate evolutionary distances, inferring phylogenetic trees, and testing evolutionary hypotheses. The new feature added in MEGA4 is the Maximum Composite Likelihood (MCL) method for estimating evolutionary distances between all pairs of sequences simultaneously, with or without incorporating rate variation among sites and substitution pattern heterogeneities among lineages. MCL method can also be used to estimate transition/transversion bias and nucleotide substitution pattern without knowledge of the phylogenetic tree. An updated version of this software is now available as MEGA5.

TreeFam [53] is a database of phylogenetic trees of gene families found in animals. Release 1.1 of TreeFam contains curated trees for 690 families and automatically generated trees for another 11,646 families. In TreeFam, orthologs and paralogs are inferred from the phylogenetic tree of a gene family. In this way, ortholog inference in TreeFam is different from that used by most other ortholog databases like in HomoloGene [54].

Role of Phylogenetic Trees in inferring Biological Pathways

Phylogenetic analysis on existing pathways can throw light on their evolutionary developmental trends. Information can also be inferred by phylogeny with detection of new interactions and their conservancy over different species. Phylogeny and Pathway Biology are often found hand in hand in attempts for discovery of new biology.

Advantage of the phylogenetic analysis of biological pathways resides in the combined analysis of more than one functional role. The analysis is understood as an extension of the classic phylogenetic analysis of individual sequences towards a higher level of description. Pathway phylogenies classify relationships between genes but also between pathways and multi-enzyme systems, gene regulatory systems, protein-protein interaction systems and signal transduction systems among others. A phylogenetic tree can be inferred from a set of biological pathways by various means like multiple nucleotide sequence alignment [55], multiple protein sequence alignment, multiple orthologous

sequence alignments of batches of 'substrates and products' [56] or 'enzymes' or 'substrates and products and enzymes' or 'substrates and products along with energy currencies'. Phylogenetic trees can also be inferred by comparing whole pathways [57–61] or a few purpose specific (disease-specific, function-specific) interacting pathways or partial [62] or whole genomes, metabolomes [63], proteomes, interactomes, phenomes, transcriptomes and protein-protein interaction networks [64–68] among others. This is possible at the advent of many -omics wise studies.

Computational phylogenetics can be used to infer individual protein interactions and protein interaction networks. Some phylogenetic trees are created from multiple sequence alignments of nucleotide sequence based phylogenetic distance matrices for inferring protein interactions [64]. Other trees, based on multiple sequence alignments of orthologous sequences, help in assessing protein co-evolution. They assist in exploiting co-evolution of interacting proteins to discover interaction specificity by using known protein sequences [65] and in prediction of the interactome [66, 68]. They try to indicate that interacting domain pairs exhibit higher level of co-evolution than the non-interacting domain pairs [67].

Phylogenetic trees can also be inferred from nucleotide sequence of members participating in signal transduction pathways [69]. The trees indicate at new cascades of interacting proteins, those arise through gene duplications, evolving towards different specificity [70]. Despite the bewildering number of cell types and patterns found in the animal kingdom, only a few signaling pathways are required to generate them. Most cell-cell interactions during embryonic development involve the Hedgehog, Wnt, TGF-β, RTK, Notch, JAK/STAT and nuclear hormone pathways. Looking at how these pathways evolved might provide insights into how a few signaling pathways can generate so much cellular and morphological diversity during the development of individual organisms and the evolution of animal body plans [71].

Inference of phylogenetic trees from metabolic pathways

A metabolic pathway is a series of biochemical reactions within a biological cell, catalyzed by enzymes, which either result in the removal of a molecule from the environment to be used or stored by the cell (metabolic sink), or the initiation of another metabolic pathway (also called flux generating step). In such pathways, enzymes, substrates, and reactions are grouped conceptually into networks. A metabolic pathway can be represented as a network that in turn, can be represented as a binary square adjacency matrix having the intervening metabolites as rows and columns and taking '1' or '0' values depending on the presence or absence of an arc between the corresponding elements. In the case of metabolic networks, an arc corresponds to the presence

of one (or more) enzyme catalyzing a chemical reaction transforming one metabolite into another. The irreversibility of some reactions makes the adjacency matrix asymmetrical [61].

An organism's metabolism is usually depicted in a graph based representation commonly referred to as a metabolic network. Metabolic network information is of interest among researchers for various purposes. Although all elements of the network are closely connected, large functionally independent parts of it are considered as metabolic pathways [58]. Comparative analysis of metabolic pathways in different genomes can help understanding the evolutionary and organizational relationships among species. This type of analysis allows one to measure the evolution of complete processes (with different functional roles) rather than the individual elements of a conventional analysis [57]. Data obtained from metabolomic experiments can be organized into networks, based on their pair wise correlations and the key challenge is to deduce unknown pathways based on the observed correlations. The data generated networks reflect the structure of the underlying biochemical pathways and help us to develop a systematic relationship between observed correlation networks and the underlying biochemical pathways [60].

Metabolic pathways in an organism can also be investigated using information about its functionally characterized proteins. Existing pathway information can be reconstructed to create novel pathways for engineering new metabolic capabilities of interest to microbiologists and metabolic engineers [72]. Reconstructing and analyzing the metabolic map of organisms is an important challenge in bioinformatics [73]. If genome is represented as a graph with genes as nodes and metabolic pathway as another graph with gene products or enzymes or Enzyme Commission (EC) numbers as nodes, graph comparison can be used to identify local similarities, termed correlated clusters, between two graphs, which allow gaps and mismatches of nodes and edges. This is especially suitable for detecting biological features [74]. A number of metabolic databases being available electronically, some with features for querying and visualizing metabolic pathways and regulatory networks, information from each data source can be extracted and compiled into a PETRI net. A petri net (also known as place / transition net on P/T net) is one of several mathematical modeling languages for the description of distributed systems. PETRI nets of the allow investigate the content in metabolic pathway databases, to map and integrate genomic information and functional annotations. Such an approach helps to define, generate and search paths and pathways in biological networks. Differential Metabolic Displays (DMDs) are useful for function prediction, especially in the context of the interpretation of gene expression data [75]. The comparison and phylogenetic analysis of

metabolic pathways may be useful for gene-diagnostics and gene-therapy that are currently based on comparative genomics. By comparing metabolic pathways, complex relationships between genes can be detected and more sophisticated directions for the cure of complex diseases may become feasible [56].

A Phylogenetic tree can be constructed from metabolic pathways based on the nucleotide sequence information of enzymes and substrates [56], nucleotide sequence information of the metabolome to find the conservativeness of the known metabolic complement of *E. coli* at the enzyme level [63], homology character code of partial metabolomes of universal biochemical compounds like amino acids, fatty acids and monosaccharides etc. to do universal phylogenetic analysis [62], distance matrix based on Relative Description Length (RDL) of metabolic networks to find conservation and evolution [76], the enzyme hierarchy, information content and gene ontology of metabolic pathways [77, 78], multiple sequence alignments of orthologous sequences to infer biological networks with output kernel trees [68], portions of metabolomes (Network of Interacting Proteins (NIPs) responsible for metabolism) to find phylogenicity among them [79]. Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks has potential in the macroscopic analysis of phylogenetic relationships among organisms in relation to horizontal gene transfer [59].

Application: developmental study of TCA cycle over a set of organisms via phylogeny

In this section, we construct enzyme graphs from Tri-Carboxylic Acid (TCA) cycle of different species. An enzyme graph removes the information on metabolites and substrates from a pathway and considers only the order of different enzymes present in the pathway [57]. Adjacency matrices representing these graphs are used for pathway comparison among different species. Distance matrix obtained from these adjacency matrix comparisons are used to construct phylogenetic tree that throw light on the development of TCA cycle among organisms.

Data

Metabolic pathway information was taken from PATHWAY Database of Kyoto Encyclopedia of Genes and Genomes (KEGG) [80–83]. The PATHWAY database is a collection of manually drawn diagrams called the KEGG reference pathway diagrams (maps), each corresponding to a known network of functional significance. Each reference pathway can be viewed as a network of enzymes or a network of EC numbers. EC numbers in the metabolic pathways play roles as identifiers of the nodes (enzymes). From the

manually drawn reference pathways, many organism specific pathways are generated by superimposing (coloring) genes in given organisms. Metabolic pathways are networks of direct and indirect protein-protein interactions, which are in turn, can be viewed as networks of enzyme-enzyme relations [84]. We took down the active part of the reference pathway, highlighted by thick black lines as shown in Fig. 1, *i.e.*, the original species-specific metabolic pathway. The rest genes are ignored, as they are not identified in the concerned species [57].

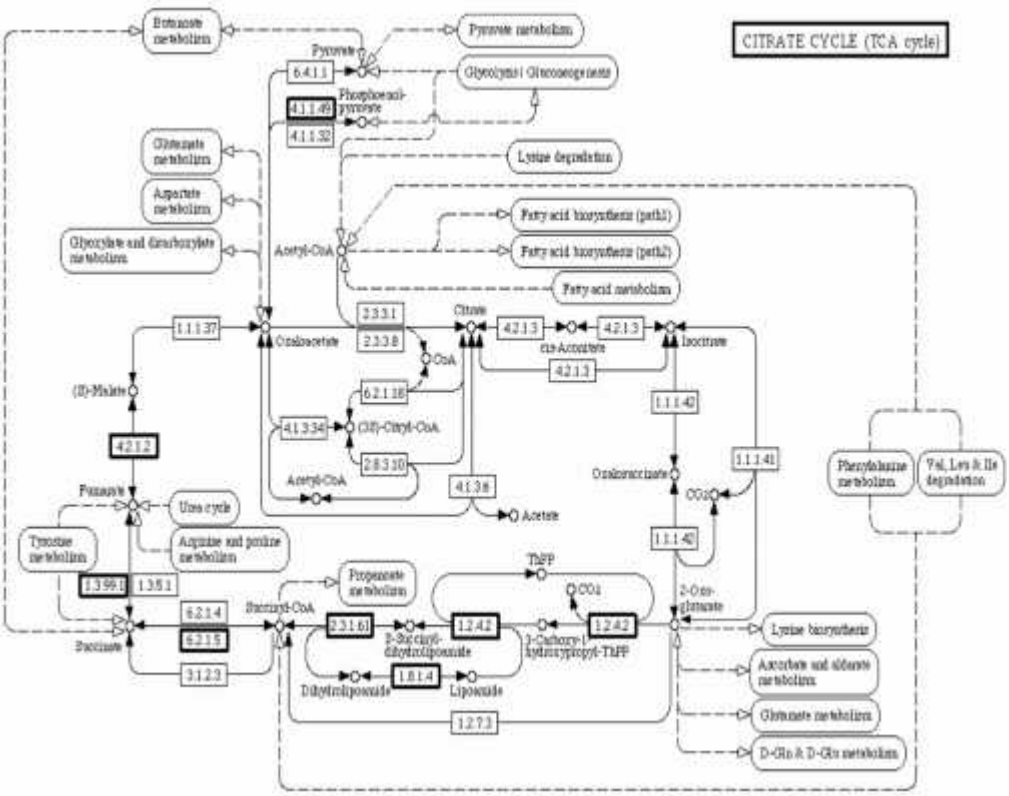


Fig. 1 : TCA Cycle of *W. glossinidia*

We divide our task of creating data from a species-specific metabolic pathway into four steps. 1) Every reaction that is a component of the metabolic pathway is studied according to its reversibility or irreversibility, 2) The Enzyme-Enzyme Relational (*E-E-R*) graph is created from the reaction components, 3) The *E-E-R* graph is simplified with introduction of self nodes (*sn*s) and double self-nodes (*dsn*s) and 4) Adjacency matrix of the simplified *E-E-R* graph is prepared. With the adjacency matrices as input data we can step into pathway comparison.

Assumptions

For our convenience, we used some assumptions that make the data generation steps easier.

1) The generalized reference pathway of TCA cycle as given in KEGG contains all total 21 different types of biomolecules (substrates and products). Here we denoted them as capital alphabets: pyruvate → A, phosphoenolpyruvate → B, acetylcoenzymeA → C, coenzymeA → D, oxaloacetate → E, citrate → F, (3S) citrylcoenzymeA → G, acetat → I, cis-aconitate → J, isocitrate → K, oxaloacetate → L, 2-oxo-glutarate → M, 3-carboxy -1-hydroxypropyl-ThPP → N, S-succinyl-dihydrolipoamide → O, ThPP → P, succinylcoenzymeA → Q,dihydrolipoamide → R, lipoamide → S, succinate → T, fumarate → U, (S) –malate → V.

2) In addition the EC number set contains 23 elements that we denote as small alphabets: 6.4.1.1 → a, 4.1.1.4a → b, 4. [comment : colons must be replaced by dots] [point 2] 1:1:32 → c, 2:3:3:1 → d, 2:3:3:8 → e, 4:1:3:34 → f, 2:8:3:10 → g, 6:2:1:18 → h, 4:1:3:6 → i, 4:2:1:3 → j, 1:1:1:41 → k, 1:1:1:42 → l, 1:2:4:2 → m, 2:3:1:61 → n, 1:8:1:4 → o, 1:2:7:3 → p, 6:2:1:4 → q, 6:2:1:5 → r, 3:1:2:3 → s, 1:3:99:1 → t, 1:3:5:1 → u, 4:2:1:2 → v, 1:1:1:37 → w.

A Member Reaction (MR) is an individual reaction of the metabolic pathway featuring its respective substrate, enzyme involved (in case of KEGG represented by EC number) and strictly only one of the products. A reversible reaction is considered as two MRs (forward reaction and backward reaction) and a reaction with two products is also considered as two MRs (each reaction involved in creating one product) as shown in Fig. 2.

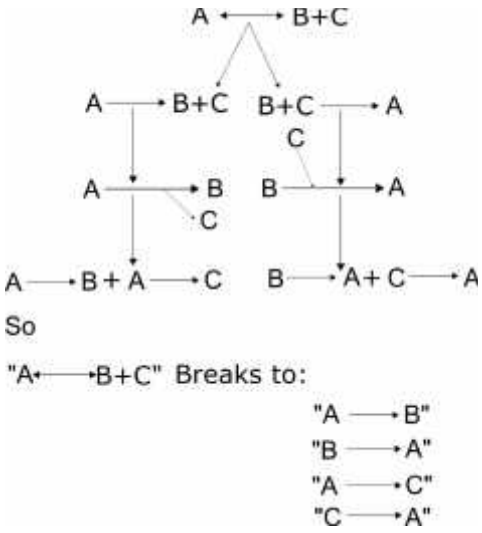
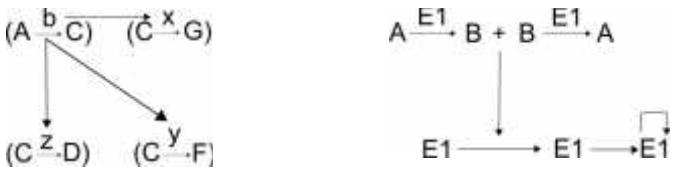


Fig. 2 : Decomposition of a reversible bi-product reaction

Construction of Enzyme-Enzyme Relational (E-E-R) graph

Each created enzyme graph can be represented as $G = (V, E)$ where V is the set of vertices (EC numbers) and $E \subseteq V \times V$ is the set of edges representing the relation between successive reactions. For graph construction each reaction component, represented by its EC number is considered as a node. If initially a single reaction component is picked randomly, i.e., b is a node having substrate A and product C as given in Fig. 3(a), in next step we have to search for reaction components having C as substrate. Let x, y, z are three reaction components starting with substrate C, then we have to link b with x, y and z by separate directed edges (Fig. 3(a)). This process is repeated till all the elements of the MR set get incorporated in the E-E-R graph.

If we have two reaction components with same EC number so that their union set is a reversible reaction, then the corresponding node is defined as a "system" and represented in the graph as '*sn*' as seen in Fig. 3(b).



(a) Graph construction (b) Representation of a self-node (*sn*)

Fig. 3 : Steps for simplification of enzyme graph

The process of simplification

A simplified version of the enzyme graph can be obtained by introducing the concept of *dsn*. If two *sn*s are reversibly connected in an E-E-R graph, then they can be represented as a *dsn*. If in the enzyme graph a node is present more than once, then only its connectivity with other nodes will be taken into account after representing it once in the simplified version.

$$E_{1(sn)} \leftrightarrow E_{1(sn)} \Rightarrow E_{1(dsn)} \quad (1)$$

Representation of graph by adjacency matrix

If G be a directed graph with n number of vertices, containing no parallel edges, then the adjacency matrix $X = [X_{ij}]$ of the digraph G is an $n \times n$ matrix whose element $x_{ij} = 1$, if there is an edge directed from i^{th} vertex to j^{th} vertex otherwise $x_{ij} = 0$. According to the definition, all the *sn* and *dsn* will have a score of 1 as they have self-loops. So, finally the E-E-R graph will be represented as an adjacency matrix.

The similarity computation algorithm

Here, we have used the algorithm proposed by Heymans and Singh [57] for computation of similarities among different graphs.

Nodal similarity and dissimilarity: We first define a similarity score *Sim* between every pair of objects represented by the nodes of the two graphs. We have taken the similarity between EC numbers as *Sim* values. The basic intuition behind the approach is that two nodes are similar if they reference and are referenced by similar nodes.

The similarity scores between nodes, $S(a,b)$ are initialized with $Sim(a,b)$, and then updated simultaneously according to the following mutually recursive rule: two nodes are similar if they link to similar nodes, are referenced by similar nodes, have both missing ingoing (outgoing) edges from (to) similar nodes and have mismatches between edges from (to) dissimilar nodes. The similarity between two nodes (a,b) is computed by summing their similarities and subtracting their dissimilarities. The former consists of four similarity terms, A_{11} - A_{14}, and the latter consists of four dissimilarity terms, D_{11} - D_{14}.

Term $A_{11}(a,b)$ represents the average similarity between the in-neighbors of a and the in-neighbors of b. We first obtain the sum of similarities of the pair of nodes (a_2,b_2) (while $a_2 \in G_1$ and $b_2 \in G_2$) from which a and b have incoming edges. We normalize the sum by dividing it by the total number of in-neighbor pairs, $deg_{in}(a).deg_{in}(b)$ ($deg_{in}(a)$ denotes the number of incoming edges to node a). A slight technicality here is that either a and/or b may not have any in-neighbors. If both a and b have an in-degree of 0, then the term A_{11} is defined as the sum of similarities of every pair (a_2,b_2) (with $a_2 \in G_1$ and $b_2 \in G_2$) normalized by the total number of such pairs, $n_1 \times n_2$. If only one of them has an in-degree of 0, then A_{11} is set to 0.

Term $A_{12}(a,b)$ represents the average similarity between the out-neighbors of a (nodes to which a has outgoing edges) and the out-neighbors of b. It is computed over the pair of nodes (a_2,b_2) (with $a_2 \in G_1$ and $b_2 \in G_2$) to which the nodes a and b have outgoing edges. It is defined analogously to A_{12}.

The next two terms are motivated by the fact that the absence of edges to similar nodes may be as meaningful as the presence of edges to similar nodes. Term $A_{13}(a,b)$ is similar to $A_{11}(a,b)$ except that it works on the complement of the input graphs. It represents the average similarity between the non-in-neighbors of a (nodes from which a has no incoming edges) and the non-in-neighbors of b. We first obtain the sum of similarities of the pair of nodes (a_2,b_2) (with $a_2 \in G_1$ and $b_2 \in G_2$) from which the nodes a and b have no incoming edges. The sum is normalized by dividing by the total number of non-in-neighbor pairs, $(n_1 - deg_{in}(a)).(n_2 - deg_{in}(b))$.

Term $A_{14}(a,b)$ represents the average similarity between the non-out-neighbors of a (nodes to which a has no outgoing edges) and the non-out-neighbors of b. It is computed over the pair of nodes (a_2,b_2) (with $a_2 \in G_1$ and $b_2 \in G_2$) to which the nodes a and b have no outgoing edges. It is defined analogously to A_{13}.

Term $D_{11}(a,b)$ represents the dissimilarity between nodes a and b on account of the in coming edges. It is computed over the pair of nodes (a_2,b_2) (with $a_2 \in G1$ *and* $b_2 \in G2$) from which node a has an incoming edges ($a_2 \rightarrow a$) but b does not (b2 $\not\rightarrow$ b).

Term $D_{12}(a,b)$ is the analogue of $D_{11}(a,b)$. It considers the similarity of nodes from which a has no incoming edges but b does. Term $D_{13}(a,b)$ considers the similarity of nodes to which a has an outgoing edge but b does not. Term $D_{14}(a,b)$ is the analogue of D_{13}. It considers the similarity of nodes to which a has no outgoing edges but b does.

The similarity scores $S(a,b)$ are computed by iteration to a fixed point [57]. We initialize the scores $S_0(a,b)$ to $Sim(a,b)$. The scores $S_{(k+1)}(a,b)$ are then recursively computed based on S_k. Since we are only interested in the relative scores, the scores are normalized after each iteration. Here is the outline of the iterative process [57].

Initialization:

$$S^0(a,b) = Sim(a,b) \tag{2}$$

Iterative step:

$$S^{k+1}(a,b) = [\frac{A^k{}_{11}(a,b) + A^k{}_{12}(a,b) + A^k{}_{13}(a,b) + A^k{}_{14}(a,b)}{4} - \frac{D^k{}_{11}(a,b) + D^k{}_{12}(a,b) + D^k{}_{13}(a,b) + D^k{}_{14}(a,b)}{4}] \times sim(a,b) \tag{3}$$

Normalization:

$$S \leftarrow \frac{S}{\|S\|_2} \tag{4}$$

In equation (2), the similarity scores $S(a,b)$ are multiplied by $Sim(a,b)$ in order to combine the neighborhood similarity with the similarity of the objects represented by the nodes. Since each of the four terms A_{11} to A_{14}, and each of the four terms D_{11} to D_{14} have a range between -1 and 1, $S(a,b)$ is also divided by 4 in order to have a range between -1 and 1. For normalization we are using Frobenius norm. The similarity scores are symmetric, i.e., $S(a,b) = S(b,a)$. Threshold value considered for convergence of the above iterative process is 0.000001 [85].

Graph matching

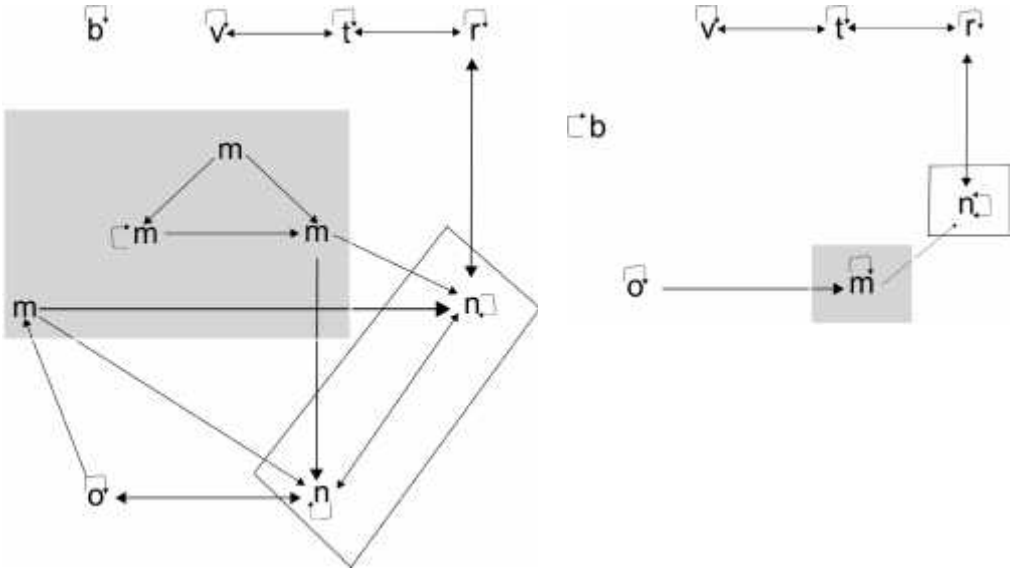
At the end of the first phase, we obtain a matrix *S* of the similarity values. These are the weighs that show extent of similarity between the considered pair of nodes where each of them belong to a different graph. If the graphs will be considered in ascending order of nodes and in each column of the similarity matrix (generated from a pair of graphs) the best weight will be chosen then we will get an optimal matched set of nodes that will be equivalent to bipartite graph matching. With the best matching so obtained, we define an n_1 x n_2 boolean matrix *M* whose entry *M(a,b)* is set to 1 if nodes *a* and *b* have been matched.

Computation of similarity of the selected pairs of nodes

After we find the best correspondence between graphs G_1 and G_2, we need to obtain the similarity score for this correspondence. As in the first phase, we combine the structural similarity with the node similarity to compute this score. We perform one iteration of a system of equations similar to A_{11}-A_{14} and D_{11}-D_{14}. The new set of equations A_{21}-A_{24} and D_{21}-D_{24} is similar to the previous one except that we use *M(a, b)* instead of *Sim(a, b)*. We also use a new normalization that is square root of the previous one. This is necessary since the maximum size of a matching is the smaller of the input graph sizes; specifically, if a graph is compared to itself then *M(a, b)* is given by the identity mapping: the similarity terms A_{21}-A_{24} reduce to 1 and the dissimilarity terms D_{21}-D_{24} reduce to 0.

Terms A_{21}-A_{24} and D_{21}-D_{24} incorporate the similarity and the dissimilarity of the best match between graphs G_1 and G_2. We combine these terms and multiply by the similarity of the nodes to obtain the final value of *S(a,b)* [57].

$$S(a,b) = [\frac{A_{21}(a,b) + A_{22}(a,b) + A_{23}(a,b) + A_{24}(a,b)}{4} - \frac{D_{21}(a,b) + D_{22}(a,b) + D_{23}(a,b) + D_{24}(a,b)}{4}] \times sim(a,b) \quad 5$$



(a) Enzyme graph of *W. glossinidia* (b) Simplified enzyme graph of *W. glossinidia*

Fig. 4 : Simplification of enzyme graph of *W. glossinidia*

The final similarity score calculation between two graphs

Finally, to obtain the similarity score $S_{G1,G2}$ between the graphs G_1 and G_2 [57], we sum the similarity scores computed in the previous phase over the pair of matched nodes, and normalize the sum by the square root of the product of the number of nodes of G_1 and G_2, in order to have a similarity score between -1 and 1. When G_1=G_2, the similarity score will be equal to 1.

$$S_{G1,G2} = \frac{\sum_{a \in G_1, b \in G_2, M(a,b)=1} S(a,b)}{\sqrt{n_1 \times n_2}} \quad (6)$$

Data Processing with an example

For example here construction of adjacency matrix from TCA Cycle of *W. glossinidia* (Fig. 1) is demonstrated. The species specific pathway contains all total 7 nodes and 19 *MR*s. The *E-E-R* graph is comparatively a simple one than that of other species like *H. sapiens* or *E. coli*. In the graph, node *n* is present twice (enclosed region in Fig. 4(a)); both of them are *sn*s and reversibly connected. So instead of these two *sn*s we can introduce a *dsn*. The connectivities are adjusted accordingly. Node *m* is present four times (striped region in Fig. 4(a)), one of them being a *sn*. So in the simplified graph, node *m* is represented as a *sn* which takes care of the connectivities of the four nodes

with each other as well as with other nodes. Following the above criteria the simplified *E-E-R* graph of *W. glossinidia* is created as shown in Fig. 4(b). These graphs can be represented as an adjacency matrix, those used for distance calculation between TCA cycles of two species. A distance matrix containing the distances of a set of species among themselves is used for creating phylogenetic tree(s).

Results and Discussion

We have chosen 27 species (the list is given in Table 2) randomly for phylogenetic tree construction from KEGG organisms list. The tree is drawn with the distance matrix calculated from TCA cycle using two tools (Fitch and Drawgram) of the Phylip 3.65 package [86]. The created tree is used to get a view of TCA cycle development among the taken species (Fig. 5). In other words, the TCA cycle tree represents closeness/distance among these species from metabolic point of view. The tree roughly shows a trend that indicates positive evolution in TCA cycle with increase in number of nodes. But directional connectivity among the nodes also plays certain role in defining the trend. That is why; *P. furiosus* is the nearest neighbor of *M. Mazei* and *C. acetobutylicum* rather than *P. horikoshii*. This tree can also be used to compare TCA cycle development with the conventional evolutionary tree obtained from NCBI taxonomy data (Fig. 6).

The NCBI taxonomy tree gives an approximation of closeness or distance among species based on evolution. But exact evolutionary distance among the species is not obtainable. Here the considered species are divided into three groups as seen in Fig. 6. Four species belong to the group Eukaryota, eight species to the group Archaea and fifteen species to the group Bacteria. These three groups are further divided into many sub-groups and sub-sub-groups in accordance with their closeness in classification. The four eukaryotes (*C. elegans, H. sapiens, M. musculus and R. norvegicus*) are positioned closer to each other with respect to other species. Group Archaea is divided into sub-groups Crenarchaeota (a major group of Archaea, containing many extremely thermophilic organisms) and Euryarchaeota (a major group of Archaea that include the methanogens, which produce methane and are often found in intestines, the halobacteria, which survive extreme concentrations of salt, and some extremely thermophilic aerobes and anaerobes). Members of group Bacteria are divided into seven sub-groups, three of which are further divided into sub-sub-groups and so on.

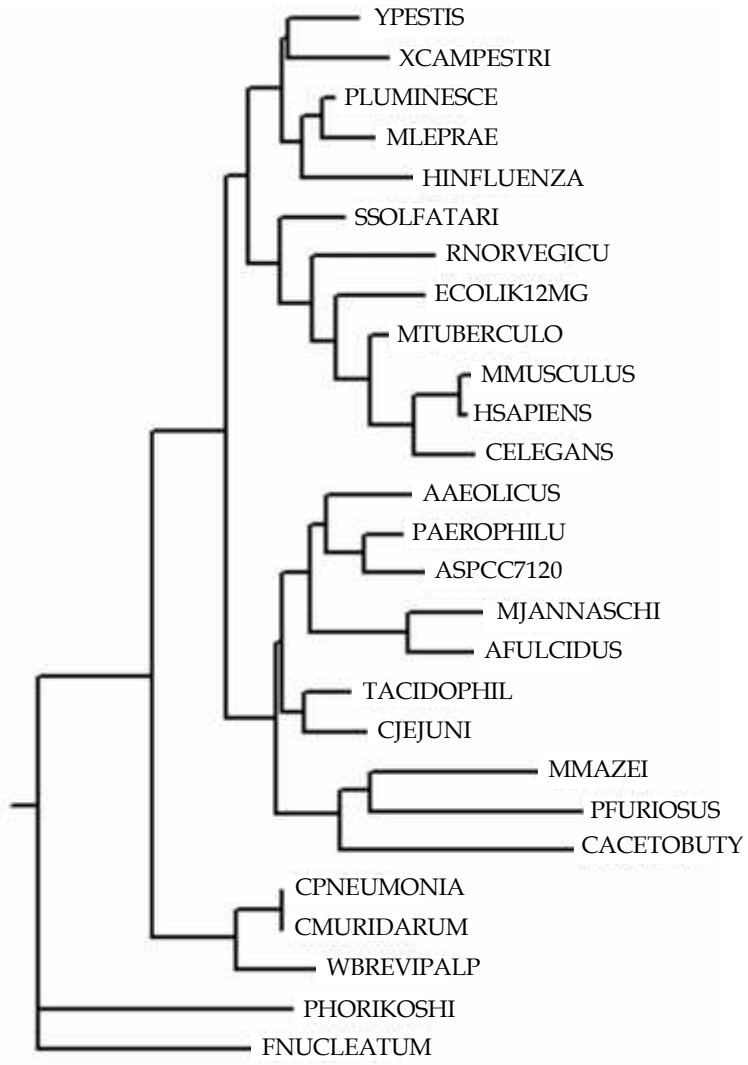


Fig. 5 : TCA Cycle tree: by Heymans's method

But the TCA cycle tree (Fig. 5) does not follow this trend. *E. coli K12 MG1655*, being a member of the group Bacteria is closer to members of group Eukaryota (*M. musculus and H. sapiens*) in comparison to *R. norvegicus*. The species *S. solfataricus* and *P. aerophilum* are positioned quite at distance irrespective of the fact that they belong to the same class Thermoprotei. Similarly, two members of the genus Pyrococcus are placed away from each other in the TCA cycle tree. All these observations confirm the fact that TCA cycle tree is not in accordance with phylogenetic tree of life and from evolution point of view closest species may have different kind of metabolism.

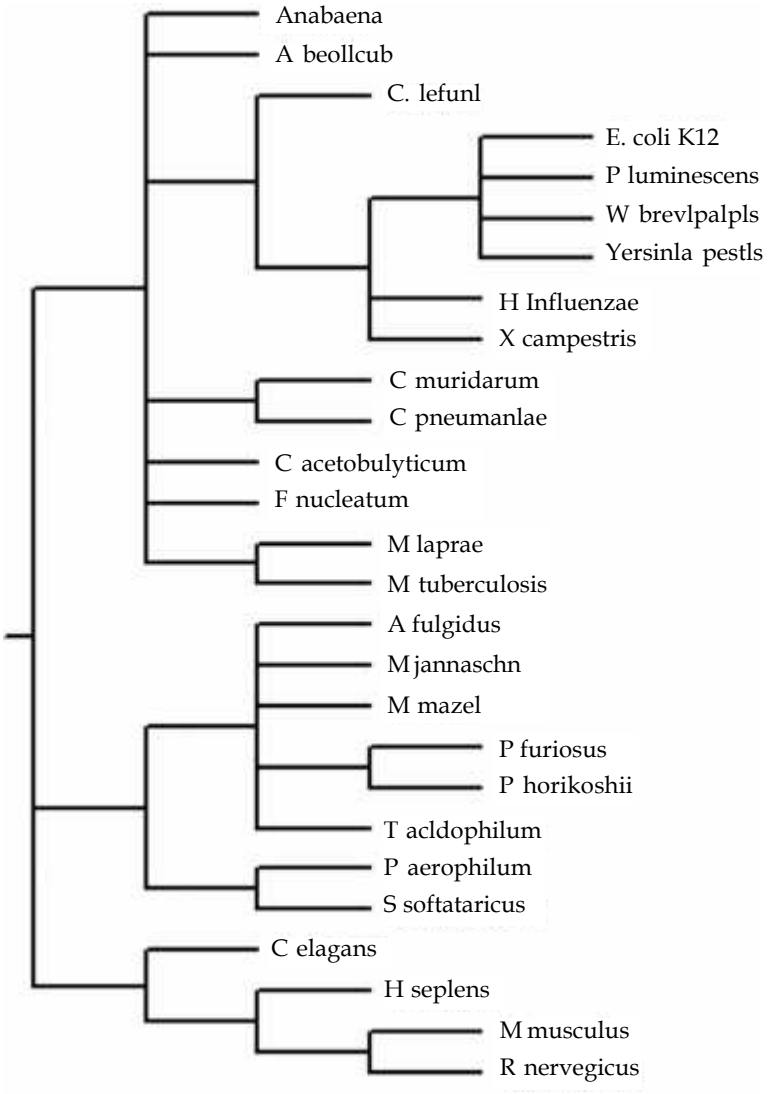


Fig. 6 : NCBI Taxonomy tree

Then naturally many questions arise about TCA cycle evolution. Some of them are: how the TCA cycle evolution occurred, what are the crucial limiting factors that regulated it and how did it shaped up in different organisms among others. We have tried to find answer for some of these questions by studying habitats of the 27 species considered here for tree construction (Table 2). In general, it is the habitat (living environment) and the type of abundant available energy source (food) that shape up an organism's metabolism. So, we did a habitat study of the considered organisms. Habitat study gave us some satisfactory results. Habitat of all the considered species

are given in Table 2. According to the table, all the pathogens are placed closely except a few. It is understandable that once inside the host body they must have faced the same kind of adjustment problems like lack of oxygen, refusal of the host acceptance etc. The hyperthermophiles show a tendency to stick together. Terrestrial mammals are closely placed with an exception of *E. coli*. But there is some justification of placing *E. coli* among the mammals as it is a habitant of mammalian intestine and mostly adapted to its habitat.

Table 2 : Habitat details of the set of considered species

Organism name	**Habitat**
Y. pestis	Pathogen
X. campestris	Plant Pathogen
P. luminescens	Insect Pathogen
M. leprae	Human Pathogen
H. influenzae	Human Pathogen
S. solfataricus	Thermophile of hot springs
R. norvegicus	Terrestrial
E. coli K12 MG1655	Gut flora (lives in mammalian intestine)
M. tuberculosis	Human Pathogen
M. musculus	Terrestrial
H. sapiens	Terrestrial
C. elegans	Soil nematode found in temperate regions
A. aeolicus	Most thermophile bacteria of hot springs
P. aerophilum	Hyperthermophile found in alkaline boiling water
Anabaena sp.PCC7120	Aquatic cyanobacteria capable of nitrogen fixation
M. jannaschii	Methane producing hyperthermophile
A. fulgidus	Sulphur metabolizing hyperthermophile
T. acidophilum	Found in coal refuse piles
C. jejuni	Pathogen
M. mazei	Terrestrial, methane producing organism
P. furiosus	Hyperthermophile, enzymes contain tungsten
C. acetobotulicum	Soil living
C. pneumoniae	Obligate intracellular parasite of human
C. muridarum	Intracellular infecting agent of family Muridae
W. glossinidia	Symbiont, lives inside gut of blood sucking tsetse fly
P. horikoshii	Hyperthermophile
F. nucleatum	Found in normal flora of mouth

Conclusions

In this paper, we tried to find some justifications for the pattern of development of TCA cycle among some species. The results indicated that habitat of an organism does play crucial role in designing an organism's metabolism. Yet there are other things to be considered and clarified like, the mixing up of soil nematodes with thermophiles, the reason behind the sulphur metabolizing and methane producing thermophiles having similar TCA cycles, and many more found in the TCA cycle tree. We hope that study of TCA cycles of different species along with other metabolic pathways in large scale will bring the complete metabolic evolution into picture.

Acknowledgement

Losiana Nayak is grateful to Council of Scientific and Industrial Research, India, for providing her a Senior Research Fellowship [No. 9/93(102)08].

URLs

1. http://en.wikipedia.org/
2. http://www.ncbi.nlm.nih.gov/Taxonomy/
3. http://www.answers.com/topic/cladistics
4. http://www.genome.jp/kegg/pathway.html
5. http://quizlet.com/10202/ap-bio-chapter-25-vocabulary-flash-cards/

References

1. Hitchcock, E. (1844). Elementary Geology. Mark H. Newman School Book Publishers and Book Sellers, New York.
2. Darwin, C. (1859). On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray Publishers, London.
3. Ward, H. (1884). Memoirs: On the Sexuality of the Fungi. *Journal of Cell Science* 2(94) 262–310.
4. Bower, F.O. (1889). The comparative examination of the meristems of Ferns, as a Phylogenetic Study. Annals of Botany 3(11) 305–392.
5. Kingsley, J.S. (1894). The Classification of the Arthropoda. *American Naturalist* 28(326) 118–135.
6. Moore, J.E.S. (1898). Memoirs: The Molluscs of the Great African Lakes.-II. The Anatomy of the Typhobias, with a Description of the New Genus (Batanalia). *Journal of Cell Science* 2(161) 181–204.
7. Blackman, F.F. (1900). The Primitive Algae and the Flagellata. An account of modern Work bearing on the Evolution of the Algae. *Annals of Botany os*-14(4) 647–688.

8. Patterson, C. (1980). Cladistics. The Biologist 27: 234–240.
9. Templeton, A.R. (2010). The Diverse Applications of Cladistic Analysis of Molecular Evolution, with Special Reference to Nested Clade Analysis. *Int. J. Mol. Sci.* 11: 124–139.
10. Horandl, E., Links, S.B. (2010). Beyond cladistics: Extending evolutionary classifications into deeper time levels. *Taxon* 59(2): 345–350.
11. Steel, M., Penny, D. (2000). Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. *Molecular Biology and Evolution* 17(6): 839–850.
12. Hendy, M.D., Penny, D. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* 59(2): 277–290.
13. Ratner, V.A., Zharkikh, A.A., Kolchanov, N., Rodin, S., Solovyov, S., Antonov, A.S. (1995). Molecular Evolution Biomathematics Series Vol 24. Springer-Verlag, New York, NY.
14. Sankoff, D., Morel, C., Cedergren, R.J. (1973). Evolution of 5S RNA and the non-randomness of base replacement. *Nature: New Biology* 245(147): 232–234.
15. Wheeler,W.C., Gladstein, D.S. (1994). MALIGN: A Multiple Sequence Alignment Program. *Journal of Heredity* 85(5): 417–418.
16. Felsenstein, J. (2004). Inferring Phylogenies. Sinauer Associates, Sunderland, MA.
17. Varon, A., Vinh, L.S., Wheeler, W.C. (2010). POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics* 26(1): 72–85.
18. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol*. 59(3): 307–321.
19. Price, M.N., Dehal, P.S., Arkin, A.P. (2010). FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* 5(3): e9490.
20. Huelsenbeck, J.P., Crandall, K.A. (1997). Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood. *Annual Review of Ecology and Systematics* 28(1): 437–466.
21. Yang, Z., Rannala, B. (1997). Bayesian Phylogenetic Inference Using DNA Sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol* 46: 409–418.
22. Heled, J., Drummond, A.J. (2010). Bayesian Inference of Species Trees from Multilocus Data. *Molecular biology and evolution* 27(3): 570–580.
23. Bohl, E., Lancaster, P. (2006). Implementation of a Markov model for phylogenetic trees. *Journal of Theoretical Biology* 239(3): 324–333.
24. Mount, D.W. (2004). Bioinformatics: Sequence and Genome Analysis. CSHL press, USA.
25. Fitch, W.M., Margoliash, E. (1967). Construction of Phylogenetic Trees. *Science* 155: 279–284.

26. Woolley, S.M., Posada, D., Crandall, K.A. (2008). A Comparison of Phylogenetic Network Methods using Computer Simulation. *PLoS ONE* 3(4): e1913.

27. Lu, G., Moriyama, E.N. (2004). Vector NTI, a balanced all-in-one sequence analysis suite. *Briefings in Bioinformatics* 5(4): 378–388.

28. Ciria, R., Goodger, C.A., Morett, E., Merino, E. (2004). GeConT: gene context analysis. *Bioinformatics* 20(14): 2307–2308.

29. Martinez-Guerrero, C.E., Ciria, R., Goodger, C.A., Hagelsieb, G.M., Merino, E. (2008). GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic Acids Research* 36: W176–W180.

30. Deng, W., Nickle, B.S.M.D.C., Learn, G.H., Liu, Y., Heath, L., Pond, S.L.K., Mullins, J.I. (2010). DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. *BioTechniques* 48(5): 405–408.

31. Bray, N., Pachter, L. (2003). MAVID multiple alignment server. *Nucleic Acids Research* 31(13): 3525–3526.

32. Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M., Gascuel, O. (2008) . Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* 36: W465–W469

33. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(113).

34. Guindon, S., Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate larges phylogenies by maximum likelihood. *Syst. Biol*. 52(5): 696–704.

35. Guindon, S., Lethiec, F., Duroux, P., Gascuel, O. (2005). PHYML Online-a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research* 33: W557–W559.

36. Chevenet, F., Brun, C., Banuls, A.L., Jacq, B., Christen, R. (2006). TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics 7(439).

37. Poirot, O., OToole, E., Notredame, C. (2003). Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Research* 31(13): 3503–3506.

38. Moretti, S., Armougom, F., Wallace, I.M., Higgins, D.G., Jongeneel, C.V., Notredame, C. (2007). The M-Coffee web server: a meta method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Research* 35: W645–W648.

39. Wallace, I.M., OSullivan, O., Higgins, D.G., Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research* 34(6): 1692–1699.

40. Palidwor, G., Reynaud, E.G., Andrade-Navarro, M.A. (2006). Taxonomic colouring of phylogenetic trees of protein sequences. *BMC Bioinformatics* 7(79).

41. Hagopian, R., Davidson, J.R., Datta, R.S., Samad, B., Jarvis, G.R., lander, K.S. (2010). SATCHMO-JS: a webserver for simultaneous protein multiple sequence alignment and phylogenetic tree construction. *Nucleic Acids Research* 38: W29–W34.

42. Dutilh, B.E., He, Y., Hekkelman, M.L., Huynen, M.A. (2008). Signature, a web server for taxonomic characterization of sequence samples using signature genes. *Nucleic Acids Research* 36: W470–W474.

43. Huang, Y.L., Huang, C.C., Tang, C.Y., Lu, C.L. (2010). SoRT^2: a tool for sorting genomes and reconstructing phylogenetic trees by reversals, generalized transpositions and translocations. *Nucleic Acids Research* 38: W221–W227.

44. Huson, D.H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformtics* 14(1): 68–73.

45. Huson, D.H., Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23(2): 254-267.

46. Alako, B.T.F., Rainey, D., Nijveen, H., Leunissen, J. A. M. (2006). TreeDomViewer: a tool for the visualization of phylogeny and protein domain structure. *Nucleic Acids Research* 34: W104–W109.

47. Zdobnov, E.M., Apweiler, R. (2001). InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9): 847–848.

48. Xu, Z., Hao, B. (2009). CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Research* 37: W174–W178.

49. Letunic, I., Bork, P. (2007). Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1): 127–128.

50. Lin, C., Lin, F.K., Lin, C.H., Lai, L.W., Hsu, H.J., Chen, S.H., Hsiung, C.A. (2005). POWER: PhylOgenetic WEb Repeater-an integrated and user-optimized framework for biomolecular phylogenetic analysis. *Nucleic Acids Research* 33: W553–W556.

51. Tamura, K., Dudley, J., Nei, M., Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol. Biol. Evol*. 24(8): 1596–1599.

52. Kumar, S., Nei, M., Dudley, J., Tamura, K. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics* 9(4): 299–306.

53. Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., Wong, G.K.S., Zheng, W., Dehal, P., Wang, J., Durbin, R. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research* 34: D572–D580.

54. Wheeler, D.L., Barrett, T., on, D.A.B., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., W.H. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 33: D39–D45.

55. Mittenhuber, G. (2001). Phylogenetic Analyses and Comparative Genomics of Vitamin B6 (Pyridoxine) and Pyridoxal Phosphate Biosynthesis Pathways. *Journal of Molecular Microbiology and Biotechnology* 3(1): 1–20.

56. Forst, C.V., Schulten, K. (2001). Phylogenetic Analysis of Metabolic Pathways. *Journal of Molecular Evolution* 52(6): 471–489.

57. Heymans, M., Singh, A.K. (2003). Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics* 19(suppl. 1): 138–146.

58. Gagneur, J., Jackson, D.B., Casari, G. (2003). Hierarchical analysis of dependency in metabolic networks. *Bioinformatics* 19(8): 1027–1034.

59. Oh, S., Joung, J.G., Chang, J.H., Zhang, B.T. (2006). Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *BMC Bioinformatics* 7(284).

60. Steuer, R., Kuths, J., Fiehn, O., Weckwerth, W. (2003). Observing and interpreting correlations in metablic networks. *Bioinformatics* 19(8): 1019–1026

61. Tun, K., Dhar, P.K., Palumbo, M.C., Giuliani, A. (2006). Metabolic pathways variability and sequence/networks comparisons. *BMC Bioinformatics* 7(24).

62. Cunchillos, C., Lecointre, G. (2005). Integrating the Universal Metabolism into a Phylogenetic Analysis. *Molecular Biology and Evolution* 22(1): 1-11.

63. Peregrin-Alvarez, J.M., Tsoka, S., Ouzounis, C.A. (2003). The Phylogenetic Extent of Metabolic Enzymes and Pathways. *Genome research* 13(3): 422–427.

64. Gertz, J., Elfond, G., Shustrova, A., Weisinger, M., Pellegrini, M., Cokus, S., Rothschild, B. (2003). Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* 19(16): 2039–2045.

65. Ramani, A.K., Marcotte, E.M. (2003). Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity. *Journal of Molecular Biology* 327(1): 273–284.

66. Pazos, F., Ranea, J.A.G., Juan, D., Sternberg, M.J.E. (2005). Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome. *Journal of Molecular Biology* 352(4): 1002–1015.

67. Jothi, R., Cherukuri, P.F., Tasneem, A., Przytycka, T.M. (2006). Co-evolutionary Analysis of Domains in Interacting Proteins Reveals Insights into Domain-Domain Interactions Mediating Protein-Protein Interactions. *Journal of Molecular Biology* 362(4): 861–875.

68. Geurts, P., Touleimat, N., Dutreix, M., d'Alch Buc, F. (2007). Inferring biological networks with output kernel trees. *BMC Bioinformatics* 8(Suppl 2) (S4).

69. Sun, J., Daniel, R., Wagner-Dobler, I., Zeng, A.P. (2004). Is autoinducer-2 a universal signal for interspecies communication: a comparative genomic and phylogenetic analysis of the synthesis and signal transduction pathways. *BMC Evolutionary Biology* 4(36).

70. Caffrey, D.R., O'Neill, L.A.J., Shields, D.C. (1999). The Evolution of the MAP Kinase Pathways: Coduplication of Interacting Proteins Leads to New Signaling Cascades. *Journal of Molecular Evolution* 49(5): 567–582.

71. Pires-daSilva, A., Sommer, R.J. (2003). The evolution of signalling pathways in animal development. *Nature Reviews Genetics* 4(1): 39–49.

72. McShan, D., Rao, S., Shah, I. (2003). PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics* 19(13): 1692–1698.

73. Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I., Dandeker, T. (2002). Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics* 18(2): 351–361.

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

75. Kuffner, R., Zimmer, R., Lengauer, T. (2000). Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics* 16(9): 825–836.

76. Chor, B., Tuller, T. (2007). Biological Networks: Comparison, Conservation, and Evolution via Relative Description Length. *Journal of Computational Biology* 14(6): 817–838.

77. Clemente, J.C., Satou, K., Valiente, G. (2005). Reconstruction of Phylogenetic Relationships from Metabolic Pathways Based on the Enzyme Hierarchy and the Gene Ontology. *Genome Informatics* 16(2): 45–55.

78. Clemente, J.C., Satou, K., Valiente, G. (2007). Phylogenetic reconstruction from non-genomic data. *Bioinformatics* 23(2): e110–e115.

79. Mazurie, A., Bonchev, D., Schwikowski, B., Buck, G.A. (2008). Phylogenetic distances are encoded in networks of interacting pathways. *Bioinformatics* 24(22): 2579–2585.

80. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research* (database issue) 32: D277–D280.

81. Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A. (2002). The KEGG database at GenomeNet. *Nucleic Acids Research* 30(1): 42-46.

82. Kanehisa, M. (1997). A database for post-genome analysis. *Trends in Genetics* 13(9): 375–376.

83. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* 34(9): D354–D357.

84. Kanehisa, M., Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28(1): 27–30.

85. Heymans, M., Singh, A.K. (2002). Deriving phylogenetic trees from the similarity analysis of metabolic pathways (Technical Report). 1–30.

86. Felsenstein, J. (1989). PHYLIP- Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164–166.

❐❐❐