# Biological Networks: Comparison, Conservation, and Evolutionary Trees
## (Extended Abstract)

Benny Chor and Tamir Tuller*

School of Computer Science, Tel Aviv University
{bchor, tamirtul}@post.tau.ac.il

**Abstract.** We describe a new approach for comparing cellular-biological networks, and finding conserved regions in two or more such networks. We use the length of describing one network, given the description of the other one, as a distance measure. We employ these distances as inputs for generating phylogenetic trees. Our algorithms are fast enough for generating phylogenetic tree of more than two hundreds metabolic networks that appear in KEGG. Using KEGG's metabolic networks as our starting point, we got trees that are not perfect, but are surprisingly good. We also found conserved regions among more than a dozen metabolic networks, and among two protein interaction networks. These conserved regions seem biologically relevant, proving the viability of our approach.

**Keywords:** Biological networks, tree reconstruction, relative description length, compression, metabolic networks, Conserved regions, networks' comparison, network evolution.

## 1 Introduction

With the advent of bio technologies, huge amounts of genomic data have accumulated. This is true not only for biological sequences, but also with respect to biological *networks*. Prominent examples are metabolic networks, protein-protein interaction networks, and regulatory networks. Such networks are typically fairly large, and are known for a number of species. On the negative side, they are error prone, and are often partial. For example, in the KEGG database [17] there are over 250 metabolic networks of different species, at very different levels of details. Furthermore, some networks are directly based on experiments, while others are mostly "synthesized" manually.

The goal in this study is to devise a quantitative and efficient method for local and global comparisons of such networks, and to examine their evolutionary signals. Our method of comparing two networks is based on the notion of *relative description length*. Given two labeled network $A$ and $B$, we argue that the more similar they are, the fewer bits are required to describe $A$ given $B$ (and vice

* Corresponding author.

A. Apostolico et al. (Eds.): RECOMB 2006, LNBI 3909, pp. 30–44, 2006.

versa). Mathematically, this can give rise to Kolmogorov complexity-like measures, which are incomputable and inapproximable. Other approaches, based on labeled graph alignment, subgraph isomorphism, and subgraph homeomorphism, are computationally intractable [12].

By way of contrast, our algorithm is efficient: Comparing the man-mouse metabolic networks takes 10 seconds on a 3 years old PC (996 MHZ, 128 MB RAM, Pentium 3 ), and all $(240 \times 239)/2$ pairwise comparisons of the KEGG database took less than three days on the same machine. We extend the relative description length approach to *local* comparison of two or multiple networks. For every label of the nodes (describing a metabolic substrate), we identify if that label exists in the various networks, and build local neighborhoods of equal radius around these labels. Neighborhoods with high similarity, according to our criteria, are likely to be conserved. We seek a method that is efficient not only for one pair of networks, but for all $\binom{n}{2}$ pairs. Our global comparison produces a matrix for expressing the pairwise distances between networks. To test its quality we have built an evolutionary tree, based on the distance matrix constructed from KEGG's metabolic networks. To the best of our knowledge, this is the first time evolutionary trees are constructed based on biological networks. The results are surprisingly good. For example, the tree for 20 taxa with large networks (more than 3000) in the KEGG database perfectly clusters the taxa to Eukaryotes, Prokaryots and Archea, and clusters almost perfectly sub-partitions within each type. Neither the 20 taxa tree nor another KEGG based tree for 194 taxa are perfect, but this is hardly surprising given the huge disparity in detail between KEGG's metabolic networks, where some have more than 3000 nodes (metabolites) while as many as 10% of species have metabolic networks with fewer than 10 nodes. Bio networks are still at a state where the available data is much more fragmented and less accessible than biological sequences data. But network information certainly goes beyond sequence information, and our work makes some preliminary steps at the fascinating questions of network comparison and evolution.

Relative description length proved to be a useful parameter for measuring the disparity between biological sequences, such as genomes. In [21], Li *et al.* describe a distance based on compression [4] that was used for generating phylogenetic trees. In [3] Burstain *et. al* present a simple method based on string algorithms (average common substring) for generating phylogenetic trees. The main innovation in the present work is the use of the paradigm of relative description length in the domain of biological networks, which is very different than the one dimensional domain of biological sequences. The different domain necessitates a different approach. Our is based on the reasonable assumption that homologous nodes in close taxa will share more similar neighborhood, as compared to remote taxa.

To the best of our knowledge, this is the first time relative description length is used for comparing networks and constructing evolutional signals (trees). Ogata *et al.* [25] developed a heuristic for finding similar regions in two metabolic pathways. Their method is based on comparing the distances between pairs of

nodes in two metabolic pathways. Schreiber [28] developed a tool for visualization of similar subgraphs in two metabolic pathways. Tohsato *et al.* [31] deals with alignment of metabolic pathways, where the topology of the pathways is restricted to chains. Kelley *et al.* [18] data-mine chains in protein-protein networks by searching paths with high likelihood in a global alignments graph, where each node represents a pair of proteins (one from each network). This last work was generalized to identify conserved paths and clusters of protein-protein interaction networks inmultiple organisms, by Sharan *at al.* [29]. They build a graph with a node for each set of homologue proteins (one protein for each organism). Two nodes are connected by an edge if all the pairs of proteins interact in each organism respectively. The second step is searching paths and clusters in this graph. Koyuturk *et al.* [19] used a bottom up algorithm for finding frequent subgraphs in biological networks. Pinter *et. al* [26] suggested an $O(n^3/\log(n))$ algorithm for the alignment of two trees. While related, this does not solve our problem as it is restricted to trees, and is not efficient enough for multiple species. Another problem with the alignment approach is to define the costs of deletion, mismatches. This problem is true for both sequences and graphs' alignment. Chung, and Matula [5, 23] suggest algorithms for a similar problem of subgraph isomorphism on trees.

The rest of the paper is organized as follows: In section 2 we discuss the general problem of comparing directed labelled graphs. Then we describe our approach, the relative description length (RDL) method. In section 3 we describe the properties of our measure. In section 4 we describe a method based on the relative description measure for finding conserved regions in network. In section 5 we demonstrate the method, where the inputs are metabolic networks from KEGG. Section 6 contain concluding remarks and suggestions for further research.

## 2 Distances and Phylogeny from Biological Networks

In this section we discuss the problem of comparing labeled, directed graphs. We then describe our RDL method for computing distances between networks. The "design criteria" is to find measures that accurately reflects biological disparity, while concurrently be efficiently computable. The networks in this paper are directed graphs with uniquely labeled nodes. Specifically we used the format of Jeong *et al.* [16] for representing a metabolic networks, only the nodes have labels, the edges have no labels. But our algorithms apply, *mutatis mutandis*, to other types of networks with such representation. All metabolic substrates are represented by graph nodes, and the reaction links in the pathway, associated with enzymes, are represented by directed graph edges.

The basic measure we are interested in is the amount of bits needed to describe a network $G_2$, given the network $G_1$. The natural measure to consider here is *Kolmogorove complexity* defined as follows $k(x) = k_U(x)$ is the length of a shortest string, $z$ that when given as an input to $U$, an *Universal Turing Machine* (TM) [30], $U$ emits $x$ and halts, namely $U(z) = x$ [22]. One may

consider *relative* Kolmogorov complexity. Given two strings $x$ and $y$, $k(x|y)$ is defined as the length of the shortest string $z$ one need to add to the the string $y$ as an input to a universal TM, $U$, such that $U(z, y) = x$. A variant of this measure is known to be a metric [22], *i. e.* it is symmetric and it satisfies the triangle inequality. Unfortunately, it is well known that Kolmogorov complexity, in its unconditional and conditional forms, is incomputable. Furthermore, there is a non constant function $f(x)$, a function that increases with $x$, such that even an $f(x)$ approximation of $k(x)$, and thus of $k(x|y)$ is incomputable. We now turn to the definition of the relative description length measure. Let $pa_i$ denote the set of nodes that are parents of $i$ in the network. A directed graph or network, $G$, with $n$ labelled nodes can be encoded by using $\log(n)$ bits to denote the number of parents of each node, and $\log \binom{n}{|pa_i|}$ bits to name $x_i$'s parents (for *sparse* networks, this is more succinct than the $n$ bits per node of the naive description). Let $DL(G)$ denote the description length of $G$. Then for an $n$ node network $DL(G) = \sum_{i=1}^{n} \left( \log(n) + \log \binom{n}{|pa_i|} \right)$. Suppose now we have a collection $\{G_i\}$ of labelled directed graphs, and let $n_i$ denote the number of nodes in $G_i$. Let $n_{i,j}$ denote the number of labelled nodes that appear *both* in $G_i$ and $G_j$. Let $pa_v(G)$ denote the number of parents of node $v$ in the graph $G$. For encoding a subset $T$ of a known set $S$, one needs $\log(|T|) + \log \binom{|S|}{|T|}$ bits. The first expression describes the size of the group $T$, and the second is for describing the subset out of $\binom{|S|}{|T|}$ possible subsets. We denote the number of bits encoding sub-set $T$ of a known set $S$ by $Enc(T|S)$. Two assumptions underly our procedure for describing one graph given the other:

1. The distance among corresponding pairs of nodes in networks of closely related species are similar.
2. It is possible that two nodes, corresponding to different species, have the same role even if their labeling is not identical.

The procedure for describing the graph $G_2$, given the graph $G_1$ was defined as follows:

**DL($G_2|G_1$)**

1. There are $n_1 - n_{1,2}$ nodes that appear in $G_1$ and do not appear in $G_2$. Given $G_1$, they can be encoded using $Enc(n_1 - n_{1,2}|n_1)$ bits.
2. For each node $v$ common to $G_1$ and $G_2$:
   (a) The node $v$ has $|pa_v(G_1) \cap pa_v(G_2)|$ parents, which appear both in $G_1$ and $G_2$. We encode these nodes by $Enc(pa_v(G_1) \cap pa_v(G_2)|n_1)$ bits.
   (b) The node $v$ has $|pa_v(2) \setminus (pa_v(G_1) \cap pa_v(G_2))|$ parents which appear in $G_2$ but not in $G_1$. We encode these nodes by $Enc\left(|pa_v(G_2) \setminus (pa_v(G_1) \cap pa_v(G_2))| \, |n_2 - n_{1,2}\right)$ bits.

(c) The rest of the parents of the node $v$ in $G_2$ appear in both $G_1$ and $G_2$, but are not parents of $v$ in $G_1$. Denote the size of this set by $n_v$. Let $d$ denote the minimal bidirectional radius of a ball around the node $v$ in $G_1$ that contains all these parents. Let $n^{v,d}$ denote the number of nodes in this ball. We encode these parents using $\log(d) + \log(n_v) + \log \binom{n^{v,d}}{n_v}$ bits.

3. For each node $v$ that appears in $G_2$ and not in $G_1$: Let $c_v$ denote the number of bits need to describe the parents of node $v$ by other node that appear both $G_1$ and $G_2$ using steps 1, 2. We encode the parents of the node by $1 + min(\log(n_1) + c_v, \log(n_2) + \log \binom{n_2}{|pa_v|})$ bits.

**Definition 1.** *Given two labelled, directed networks $G_i$ and $G_j$, we define their relative description length "distance", $RDL(G_i, G_j)$, as follows: $RDL(G_i, G_j) = DL(G_i|G_j)/DL(G_i) + DL(G_j|G_i)/DL(G_j)$.*

The first term in this expression is the ratio of the number of bits needed to describe $G_i$ when $G_j$ is given and the number of bits needed to describe $G_i$ without additional information. The second term is the dual. In general, $D(G_1, G_2)$ is larger when the two networks are more dissimilar, and $0 \leq D(G_1, G_2) \leq 2$. The extreme cases are $G_1 = G_2$, where $D(G_1, G_2)$ is $O(1/|V_1| + 1/|V_2|)$, and when $G_1$, $G_2$ have no nodes in common, where $D(G_1, G_2) = 2$.

In the preprocessing stage we first calculate the distances between all pairs of nodes in $G_1$ and $G_2$ by Dijkstra algorithm [6] or Johnson algorithm [6], we ignore directionality. The running time of these algorithms is $O(|E| \cdot |V| + |V|^2 \log(|V|))$. In all metabolic networks, the input degree of each node is bounded (in all the network in KEGG no one have more than 40 parents, usually it was much less, between 1 and 3 parents), thus $E = \Theta(V)$, and the time complexity is $O(|V|^2 \log(|V|))$ for all pairs. Note that there are algorithms of time complexity $O(|V|^{2.575})$ for finding distances between all pairs of nodes without any assumptions on the graphs structure [32] . We now sort the distance vector of each node in $O(|V| \log(|V|)$ time, so the total time is $O(|V|^2 \log(|V|))$. In the next stage, we sort the node names in each net in lexicographic order in $O(|V| \log(|V|)$ time. Then we sort each parent list in lexicographic order, this is done in $O(|V| \log(|V|)$ time.

Stage 1. in the procedure $DL(G_2|G_1)$ is done in linear time given a lexicographic ordering of the nodes in the two networks. The total of stages 2.($a$) for all the nodes is done in linear time given a lexicographically ordered list of all the parent list. The total of stages 2.($b$) for all the nodes is done in $O(|V| \log(|V|)$ time given a lexicographic sort of the nodes in $G_1$. The total of stages 2.($c$) for all the nodes is done in $O(|V| \log(|V|)$ time given the sorted distances matrix of the network. Stage 3 done in total time of $O(|V|^2 \log(|V|)$ for all the nodes.

Thus the total time complexity of the pairwise network comparison algorithm is $O(|V|^2 \log(|V|)$.

We used neighbor joining algorithm [27] for generating a tree from the distance matrix. Recent variants of NJ run in in $O(N^2)$, where $N$ is the number of taxa [9]. Thus the total time complexity of our method for generating a phylogenetic tree for $N$ networks of up to $|V|$ nodes each, is $O(N^2 \cdot |V|^2 \log |V|)$. We discovered empirically that by skipping stage 3., the precision decreases by a few percentage points, while the time complexity becomes close to linear. Such shortcut may be suitable for larger inputs.

## 3 Properties of the RDL Networks Comparison Measure

It is easy to see that the measure $D(G_i, G_j)$ is symmetric. While $D(G, G) > 0$, it is small for large graphs. In general, our measure does not satisfy the triangle inequality. For example the distance of the following three networks in KEGG do not satisfy the triangle inequality. The networks are the bacteria Aquifex aeolicus (*aae*), the archea Archaeoglobus fulgidus (*afu*), and the bacteria Bacteroides fragilis YCH46 (*bfr*). The distance between *aee* and *bfr* is 4.7, while the distance between *aae* and *afu* is 0.7 and the distance between *afu* and *bfr* is 3.92. However, by empirically checking all the triplets in a distance matrix generated for all the 240 networks in KEGG we found that only a very small fraction of all triplets do not satisfy the triangle inequality - 363 triplets out of $2,257,280$ possible triplets. Usually these triplets involve very partial nets. For example the *bfr* network mentioned above includes only four nodes. After removing all the networks with less than 100 nodes, we got 194 networks left. For this set of species, all the triplets satisfy the triangle inequality.

We performed preliminary empirical studies, showing that our measure increases linearily as a function of the "evolutionary time". We used the following simple minded model: At each time period there is a probability $p_1$ of adding a new node to a net, probability $p_2$ of removing a node from a net (all nodes have the same probability to be removed), probability $p_3$ of adding a directed edge between any two vertices, probability $p_4$ of removing a directed existing edge between any two vertices (all edges have the same probability to be removed). We chose $p_1 = p_2$ in order to maintain the expected number of nodes in the graph, and choose $p_3 = p_4$ in order to maintain the average number of edges in the graph.

In the resultant graphs the growth was close to linear, suggesting that for networks with similar sizes, our method for generating phylogenetic trees using distances based methods, such as neighbor jointing, is justified. Furthermore, our method can also be used to estimate branch lengths of phylogenetic trees. These consequences do not necessarily apply to networks of different sizes. Of course, the preliminary simulation used a very simplistic model. More sophisticated ones, including unequal grows and elimination rates, may give a better indication for more realistic instances.

## 4 Finding Conserved Regions in Networks

In this section we describe our method for finding conserved regions in two or more networks, and the rationale behind it. The method is based on the RDL measure described in section 2. Consider a ball of bidirectional distance at most $d$ from node $v$ in the directed graph $G$. The $d$ conservations score of the node $v$ in two is $\infty$ if it is not appear in the two networks, if it appear in the two networks it defined as follows:

**Definition 2.** *A $(d, c)$ conserved node:*
*Let $v$ be a shared node among $G_1$ and $G_2$. Let $B_1$ and $B_2$ be the balls of bidirectional radius $d$ around $v$ in $G_1$ and $G_2$, respectively. We say that $v$ is $(d, c)$ conserved in $G_1$, $G_2$ if $D(B_1, B_2) \leq c$.*

The $(d, c)$–conservated region of the two network $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is defined as the intersection of the two subgraphs of $G_1$, $G_2$ induced by the $(d, c)$ conserved nodes with respect to $G_1$, $G_2$. Algorithmically, we get it as follows

**Find the $(d, c)$ conserved region of $G_2$, $G_1$:**

1. For each node common to $G_1$ and $G_2$, compute its $d$-conservations score.
2. Generate a graph $G_1' = (V_1', E_1')$ where $V_1'$ includes the nodes in $G_1$ that are $(d, c)$ conserved with respect to $G_1$, $G_2$. The edge $e$ is an directed edge in $G_1'$ if its two endpoints are in $V_1'$, and it is a directed edge in $E$.
3. The graph $G_2' = (V_2', E_2')$ is defined analogously.

The parameters $d$ (radius) and $c$ (RDL score), determine the two conserved regions $G_1'$, $G_2'$. It is easy to see that decreasing $c$ decreases the sizes of $G_1'$, $G_2'$. Increasing $d$ may either increase or decrease the sizes of the conserved graphs.

In a similar way we now define a conservation score for a node with respect to more than two network.

**Definition 3.** *$(d, c, k)$ conservation node:*
*Let $k$ satisfy $1 \leq k \leq \binom{N}{2}$. A node $v$ is $(d, c, k)$ conserved with respect to the $N$ networks, $G_1$, $G_2$,.., $G_N$, if $v$ is $(d, c)$ conserved in at least $k$ out of the $\binom{N}{2}$ networks pairs.*

We adjusted the parameters $d$, $c$, $k$ to our input graphs, by choosing parameters such that a random node is picked as conserved with probability smaller than $p$, where $p$ is a pre-defined threshold (usually $p = 0.05$). The rational behind our approach is that the probability of mutations in "more important" parts of the network is smaller (just like for sequences). We filter noise by finding subgraphs that are conserved for sufficiently many pairs ($k$) of networks. Since every node in the network is a part of a process (*e.g.* a metabolic pathway, or a protein

signaling pathway in a protein interaction network), we expect an "important" node to share "important" pathways and thus have a conserved neighborhood, which our definition is supposed to capture.

## 5 Experimental Results

In this section we describe the results of running our algorithms on the metabolic networks in the KEGG database. First, we describe the phylogenetic trees our method generated (for two different subsets of species), and discuss the similarity of these trees to the common taxonomy [8]. Then, we describe the results of applying our method for finding conserved regions and discuss the biological relevance of the results.

### 5.1 Phylogenetic Trees

We started with a relatively small subset, containing 19 taxa: 9 eukaryotes, 5 prokaryotes, and 5 archea. We chose species whose networks in KEGG have more than 900 nodes. We generated a distance matrix based on RDL, and finally constructed a tree, using the Phylip [11] implementation of NJ algorithm [27]. The tree with the true edges' length is depicted in figure 1. The resulting tree is reasonably close to the common accepted taxonomy of these species [8]. The five archea, the five prokaryotes, and the nine eukaryotes form a clade each. Within the eukaryotes, the three mammals (rat, mouse, and human) are clustered together. The fruit fly and the worm C. elegance, both from the *Bilateria* super family, are clustered together. The three yeasts (S. Scerevisiae, A. Gossyppi, and S. Pombe) are clustered together. One example of inaccuracy in our tree is the split inside the mammals, putting the human and mouse together and the rat as an outgroup. One possible explanation is that mouse is a much more popular model animal than rat (it indeed have about 30% more nodes in KEGG), consequently its investigated pathways are closer to human and this is reflected in KEGG. The length of the branches are reasonable, compared to analog methods for phylogeny that are based on sequences' compression [3, 21].

In the next step we generated a tree for all the 194 networks having more than 100 nodes in KEGG (KEGG has additional 56 species with smaller metabolic networks). The resulting tree is depicted in figure 2. Of the 194 taxa in the tree 13 are eukaryotes, 17 archea, and 164 are prokaryotes. This subset includes about 50 species with networks of a few hundreds nodes, and about 80 species with thousands nodes, the largest network (for example human or the bacteria Bardyrhizobium Japonicum - a gram negative bacteria that develops a symbiosis with the soybean plant) has more than 3000 nodes. The names of the taxa are their code name in KEGG. We colored eukaryotes blue, archea grin, and prokaryotes red.

All the archea formed a clade and so did the prokaryotes. All the eukaryotes but one, *plasmodium falciparum* ($pfa$). Plasmodium is placed among the bacteria. One possible explanation is the loss of genes and mtabolic pathways that plasmodium, the malaria parasite, went through [13, 20]. The dataset we used has two super-families of archea. The first is *Euryarchaeota*, which contains

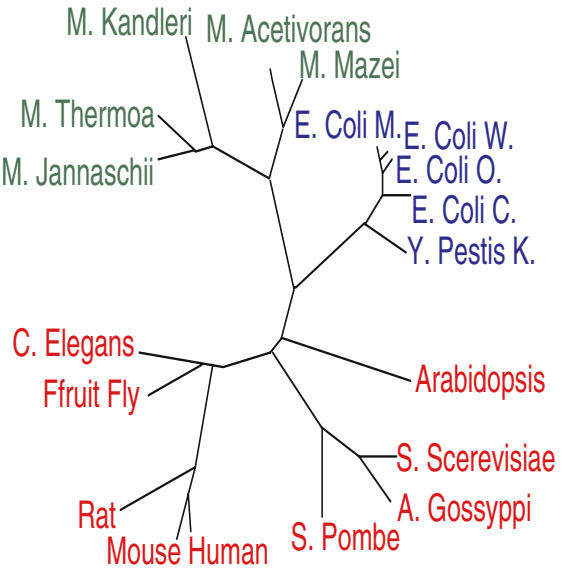


**Fig. 1.** A small phylogenetic tree, built upon distances of metabolic networks as computed using our method (tree topology and edges' length from NJ algorithm)

the species *pab*, *pho*, *hal*, *mja*, *afu*, *hma*, *pto*, *mth*, *tac*, *tyo*, and *mma*. The other is *Crenachaeota*, containing the species *pai*, *sto*, *sso*, *ape*. The only archea that "jumps family" from the second super family to the first is *Pyrobaculum aerophilum* (*pai*), which an extremely thermoacidophilic anaerobic taxa [1]. The partitioning within the eukaryotes kingdom is similar to its partition in the tree for the small dataset (figure 1). Most of the prokaryotes families are clustered together: For example the gamma proteobacteria *vvu*, *vvy*, *vpa*, *ppr*, *vch*, *son* form a clade. Most of the alpha bacteria are clustered together: *Mlo*, *Sme*, *Atu*, *Atc*, *Bme*, *Bms*, *Bja*, *Rpa*, and *Sil*. With the exception of *Ehrlichnia ruminantium Welgevonden* (*Eru*) that joined to the malaria parasite *pfa*, and of Caulobacter Crescentus (*ccr*) that is close (few splits away) but not in the same main cluster alpha bacteria. The two *Bartonella Bhe* and *Bqu* are clustered together, *Zmo* and *gox* are clustered close together but not in the main cluster of alpha bacteria. Considering the large variability in the sizes of the networks and the noisy inputs, we view the results as very good.

### 5.2 Conserved Regions in Metabolic Networks

In this section we describe the results of our algorithm for finding conserved regions on few dataset. The first contains two species:A bacteria and human, the second contains nine eukaryotes, and the last dataset has ten species, including four eukaryotes, three prokaryotes, and three archea. We also discuss another dataset of three species (Human, E. Coli and yeast) whose their pathways in KEGG are known to be constructed independently. For a lack of space we

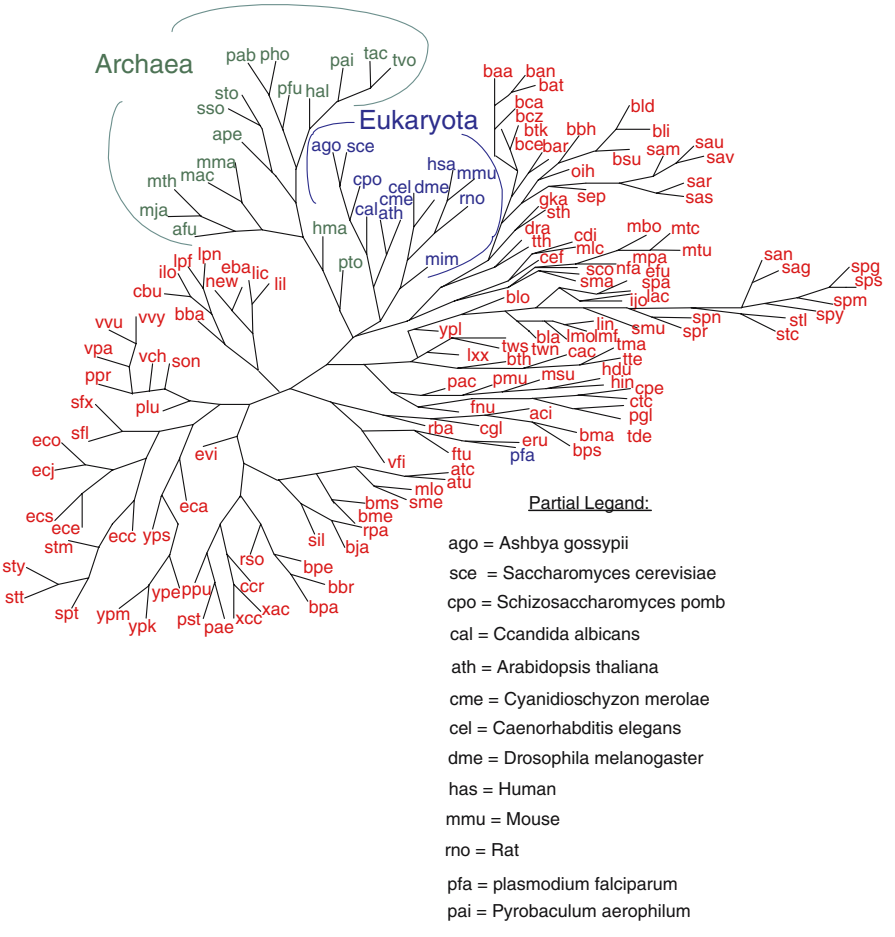


**Fig. 2.** Phylogenetic tree for 194 based on metabolic networks, all with more than 100 nodes in KEGG

describe here only the stoichiometric formulas of the compounds and very small fraction of the graphs we got, full details of the compounds can be found in KEGG. We describe here few of the subgraphs we found in the results conserved subgraphs. Note that even the relatively short subgraphs described here have since by our definition they are surrounded by a relatively conserved neighborhood.

Our first set contains two very far species: Human and the Gamma Enterobacteria - Yersinia Pestis KIM. Since these two species were split billions of years ago, we expect that the conserved regions found are common to many other taxa. The thresholds to our algorithm was diameter $d$ of 20 nodes, and relative description score 0.9. From our experience, a threshold of 0.9 or lower is fairly strict.

KEGG's metabolic network of human includes more than 3000 nodes, while the metabolic network of the bacteria includes more than 2000 nodes. The

resulting conserved networks includes 160 nodes, that are common to the two species. We describe here few of many other results we found: One of long simple paths in the conserved graph represents the metabolic pathway $C_{10}H_{14}N_2O_5$ ($C00214$) $\rightarrow$ $C_{10}H_{15}N_2O_8P$ ($C00364$) $\leftrightarrow$ $C_{10}H_{16}N_2O_{11}P_2$ ($C00363$) $\leftrightarrow$ $C_{10}H_{17}N_2O_{14}P_3$ ($C00459$), which is a part of the pyrimidine metabolism [14]. It includes the last four nodes at the end of the pathway Pyrimidine synthesis. Pyrimidine are the nucleotides $T$ and $C$, which are building blocks of DNA.

Another simple path of length four represent the sub metabolic pathway $C_5H_{11}O_8P$ ($C00620$) $\leftrightarrow$ $C_5H_{11}O_8P$ ($C00117$) $\rightarrow$ $C_5H_{13}O_{14}P_3$ ($C00119$) $\leftarrow$ $C_{27}H_{46}O_3$ ($C01151$). This is a part of the the pentose phosphate pathway [24]. One of the functions of this anabolic pathway is to utilizes the 6 carbons of glucose to generate 5 carbon sugars, necessary for the synthesis of nucleotides and nucleic acids. This pathway is also part of purine synthesis metabolism, again - one of the building blocks of DNA.

In the next stage we checked for conserved regions in nine Eukaryotes. We chose Eukaryotes with networks larger than 2000 nodes in KEGG. We generated the $(20, 0.7, 6)$ conserved graph for this set of species.

The resulting nine conserved metabolic networks includes between 84 to 106 nodes, while each of the input networks has more than 2000 nodes. We describe here few of the results we found, some ultra conserved regions: The first subgraph $C_6H_9NO_2S_2R_2$ ($C00342$) $\leftrightarrow$ $C_6H_7NO_2S_2R_2$ ($C00343$) is shared by all nine subnetworks. It is part of the pyrimidine synthesis metabolism.

The second pathway is part of the Riboflavin (the left node in the pathway) synthesis metabolism: $C_{27}H_{33}N_9O_{15}P_2$ ($C00016$) $\leftrightarrow$ $C_{17}H_{21}N_4O_9P$ ($C00061$) $\leftrightarrow$ $C_{17}H_{20}N_4O_6$ ($C00255$) Riboflavin is a vitamin that supports energy meta bolism and biosynthesis of a number of essential compounds in eukaryotes, such as human, mouse, fruit fly, rat, S. Cerevisiae, and more [17]. The following ultra conserved subgraph is part of the Cysteine synthesis metabolism:

$C_6H_{12}N_2O_4S_2$ ($C00491$) $\leftrightarrow$ $C_3H_7NO_2S_2$ ($C01962$). Cysteine (the right node in the pathway above) is an amino acid with many important physiological functions in eukaryotes. It is part of Glutathione and is a precursor in its synthesis, which is found in almost all the eukaryotes tissues and has many functions such as activating certain enzymes, and degrading toxic compounds and chemical that contain oxygen.

The last dataset we includs four eukaryotes, three archea, and three bacteria. From each class, we chose species with a large number of nodes in KEGG, the input networks include between 1500 and 3000 nodes. We generated the $(20, 0.7, 6)$ conserved graph for this set of species. The resulting ten conserved metabolic networks include between 58 to 93 nodes. We describe here few of the interesting results. We found a ultra conserved sub-networks, related to nucleotides metabolism, this is the same part of the pyrimidine synthesis metabolism described above. Another path is part of the Bile acid biosynthesis metabolism: $C_{27}H_{48}N_2O_3$ ($C05444$) $\leftrightarrow$ $C_{27}H_{46}O_3$ ($C05445$). Bile acid is essential for fat digestion, and for eliminating wastes from the body. It is also generated by bacteria in the intestine [15].

An unexpected ultra conserved path, the subnetwork $C_2Cl(4)$ ($C06789$) $\rightarrow$ $C_2HCl_3$ ($C06790$) $\rightarrow$($C_2H_2Cl_2$ ($C06791$), $C_2H_2Cl_2$ ($C06792$)) $\rightarrow$ $C_2H_3Cl$ ($C06793$) is the first part of the Tetrachloroethene degradation pathway. Tetrachloroethene is a toxin (also known as PCE). Different organisms have developed different processes for degrading PCE [2, 10, 7]. However, the part of this pathway we find here is shared by to many species (and in nine out of ten species in our dataset).

There are few species whose pathway in KEGG were reconstructed independently. Three such species are Human, E. Coli, S. cerevisiae (yeast). We implemented our method for finding conserved regions on these three species which have between 2000 to 3000 nodes in KEGG. We generated the $(20, 0.9, 3)$ conserved graph for this set of species. The conserved graphs of the Human, E. Coli, S. cerevisiae respectively included 79, 79, and 101 nodes respectively. Major fraction of the pathways found for other sets of species are also found here. One such example is the sub-graph of Pyrimidine synthesis.

In all the above results we noticed that conserved node, *i. e.* nodes that are part of the plotted resulting graphs, tend to be with a relative high in- and out-degrees, *i. e.* at least four, in the original networks. Note that in our graph representation of metabolic networks the edges (enzymes names) were unlabelled. However, in the case of the conserved sub-graphs described here the edges were also conserved.

### 5.3 Conserved Regions in Protein Interaction Networks

In addition to the metabolic networks, we have preliminary results on finding conserved regions in two protein interaction networks. In this subsection we report an initial study of finding conserved regions in the protein interaction networks of yeast and drosophila (7164 and 4737 nodes, respectively). We emphasis that these are preliminary results, which mainly establish the application of our approach to networks whose characteristics differ from metabolic networks. In contrast to the metabolic networks, protein interaction networks do not have labels that are shared across species. To identify corresponding nodes, we used Blast results. Two protein were declared identical if the drosophila's protein have the best blast score for the yeast protein, and the score were $< e^{-10}$. We now ran our algorithm. The two nodes with the highest conservation score the first node is the protein $YML064C$ in yeast and his homolog in drosophila (the protein $CG2108$). This protein catalyzed the basic reaction $GTP + H_2O \rightarrow GDP + phosphate$, and as such it is expected a-priori to be conserve. The second protein is $YLR447C$ in yeast ( the protein $CG2934$ in drosophila) also involve in "basic" activities such as hydrogen-exporting ATPeas activity, catalyzing the reaction: $ATP + H_2O + H^+(in) \rightarrow ADP + phosphate + H^+(out)$.

## 6 Concluding Remarks and Further Research

We presented a novel method for comparing cellular-biological networks and finding conserved regions in two or more such networks. We implemented our

method, and produced a number of preliminary biological results. It is clear that networks contains information, which is different than sequence information, and also differ from information in gene content. This work opens up a number of algorithmic and biological questions. The various networks in KEGG were not built independently. This biases the results, especially those of conserved regions. Interestingly, despite this fact, the our results seem surprisingly good.

The experimental work here concentrated mainly on metabolic networks taken from the KEGG database. Of course, there is no reason to consider only KEGG, and only metabolic networks. More importantly, we plan to examine our methods on more protein interaction networks, regulatory networks, and possibly a mixture thereof.

Our representation of the networks followed that of Jeong *et al.* [16] and ignored the edge labels (enzyme names). As shown in the conserved regions, identical node labels (substrates) seem to determine the enzymes involved. Yet, it is desireable to include edge labels explicitly. Indeed, the RDL approach allows such modification at relative ease. A more meaningful extension is to consider labels not just as equal or unequal. A continuous scale of similarity, as implied for example from the chemical description of substrates, can be used. Different representations of the directed graph (*e.g.* children instead of parents) are also possible. Other algorithms, based on variants of labeled subgraph isomorphism, can be considered as well. However, their efficiency should be carefully analyzed.

When dealing with biological networks, we should always keep in mind that they are still in their infancy. They are noisy due to experimental conditions, and they are partial, due to budgetary limitations and biases of the researchers. Thus the precision of the results is likely to evolve and improve, as more reliable data are gathered.

Finally, it will be of interest to combine different sources of data, for example sequence data (proteins and genes) and network data, to construct trees and find conserved regions. Of special interest are regions where the signals from the various sources are either coherent or incoherent. Of course, this work is only a first step, and calls for possible improvements.

## Acknowledgements

We would like to thank Nadir Ashkenazi, Nathan Nelson, Eytan Ruppin, Roded Sharan, and Tomer Shlomi for helpful discussions.

## References

1. S. Afshar, E. Johnson, S. Viries, and I. Schroder. Properties of a thermostable nitrate reductase from the hyperthermophilic archaeon pyrobaculum aerophilum. *Journal of Bacteriology*, pages 5491–5495, 2001.
2. D. M. Bagly and J. M. Gossett. Tetrachloroethene transformation to trichloroethene and cis-1,2-dichloroethene by sulfate-reducing enrichment cultures. *Appl. Environ. Microbio*, 56(8), 1990.

3. D. Burstein, I. Ulitsky, T. Tuller, and B. Chor. Information theoretic approaches to whole genome phylogenies. *RECOMB05*, pages 296–310, 2005.
4. X. Chen, S. Kwong, and M. Li. A compression algorithm for dna sequences and its applications in genome comparison. *RECOMB*, pages 107–117, 2000.
5. M. J. Chung. $o(n^{2.5})$ time algorithms for subgraph homeomorphisem problem on trees. *J. Algorithms*, 8:106–112, 1987.
6. T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. The MIT Press, Cambridge, 1990.
7. J. Damborsky. Ttetrachloroethene–dehalogenating bacteria. *Folia. Microbiol*, 44(3), 1999.
8. NCBI Taxonomy Database. http://www.ncbi.nlm.nih.gov/entrez/linkout/tutorial/taxtour.html.
9. I. Elias and J. Lagergren. Fast neighbor joining. In *Proc. of the 32nd International Colloquium on Automata, Languages and Programming (ICALP'05)*, volume 3580 of *Lecture Notes in Computer Science*, pages 1263–1274. Springer-Verlag, July 2005.
10. B. Z. Fathepure and S. A. Boyd. Dependence of tetrachloroethylene dechlorination on methanogenic substrate consumption by methanosarcina sp. strain dcm. *Appl. Environ. Microbio*, 54(12), 1988.
11. J. Felsenstein. Phylip (phylogeny inference package) version 3.5c. *Distributed by the author. Department of Genetics, University of Washington, Seattle*, 1993.
12. M. R. Garey and D. S. Johnson. *Computers and Intractability*. Bell Telephone Laboratories, 1979.
13. R. Hernandez-Rivas, D. Mattei, Y. Sterkers, D. S. Peterson, T. E. Wellems, and A. Acherf. Expressed var genes are found in plasmodium falciparum subtelomeric regions. *Mol. Cell. Biol*, pages 604–611, 1997.
14. P. A. Hoffee and M. E. Jones. *Purin and pyrimidine nucleotide metabolism*. Academic Press, 1978.
15. A. F. Hofmann. Bile acids: The good, the bad, and the ugly. *News Physiol. Sci.*, 14, 1999.
16. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000.
17. M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28:27–30, 2000.
18. B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stock-well, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS*, 100:11394–11399, 2003.
19. M. Koyuturk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, 20:i200–i207, 2004.
20. D. M. Krylov, Y. I. Wolf, I. B. Rogozin, and E. V. Koonim. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, 13:2229–2235, 2003.
21. M. Li, J. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 2001.
22. M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 2005.
23. D. W. Matula. Subtree isomorphisem in $o(n^{5/2})$ time. *Ann. Discrete Math*, 2:91–106, 1978.
24. G. Michal. *Biochemical pathways*. John Wiley and Sons. Inc, 1999.

25. H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, 28:4021–4028, 2000.
26. R. Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 2005.
27. N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.
28. F. Schreiber. Comparison of metabolic pathways using constraint graph drawing. *Proceedings of the Asia-Pacific Bioinformatics Conference (APBC'03), Conferences in Research and Practice in Information Technology*, 19:105–110, 2003.
29. R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA 102*, pages 1974 – 1979, 2005.
30. M. Sipser. *Introduction to the Theory of Computation*. PSW Publishing Company, 1997.
31. Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proc. 8th International Conference on Inelligent Systems for Molecular Biology (ISMB2000)*, pages 376–383, 2000.
32. U. Zwick. All pairs shortest paths using bridging sets and rectangular matrix multiplication, 2000.