Biological Networks: Comparison, Conservation, and Evolution via Relative Description Length

By: Tamir Tuller & Benny Chor

Presented by:
Noga Grebla
Presenting the goals of the research

Reviewing basic terms required for understanding the article

Defining the problem discussed in the research

Comparing biological networks & calculating the measure

Building evolutionary trees

Finding conserved regions

Experimental running and its results

Main conclusions and future directions
Goals of the study

Goals

Devising quantitative and efficient methods for:

- Comparison of two or more cellular-biological networks

- Examination of their evolutionary signals through:
  - Building phylogenetic trees
  - Finding conserved regions

Rational

These methods can be used to examine evolutionary signals implied by conserved regions and reflected in phylogenetic trees.
Goals of the study

Innovations

- New measure for the degree of similarity among the networks which is quantitative and efficient
- New method for comparing biological networks and finding conserved regions
- Building phylogenetic trees based on using biological networks comparison.
Basic terms

Presenting the goals of the research

Reviewing basic terms required for understanding the article

Defining the problem discussed in the research

Comparing biological networks & calculating the measure

Building evolutionary trees

Finding conserved regions

Experimental running and its results

Main conclusions and future directions
Basic terms

**List of basic terms in bioinformatics**

* Biological network
* Conserved regions
* Phylogenetic trees
* Homologue nodes
* Pathways
* Taxa
Basic terms – biological networks

- A biological network describes relations between biological components
- The components are represented by nodes
- The relations are represented by edges
- A cellular biological network, describes relations between biological components within the cell, such as DNA, RNA, proteins, etc’.
- Typically fairly large
Basic terms – biological networks

Biological cellular network in this article

- Metabolic networks
  Describe the biochemical reactions between chemical compounds which convert one compound into another

- Protein interaction networks
  Describe interactions between different types of proteins
Basic terms – biological networks

Available data

- known for a number of species

- KEGG database contains over 250 metabolic networks of different species, at very different levels of details

- Some networks are directly based on experiments, while others are synthesized

- The data are often fragmented, incomplete, and error prone

Yet: network information goes far beyond sequence information.
Conserved regions
Conserved regions of biological networks are parts which remained unchanged through evolution of many species, since such mutations could not survive.

Therefore they provide very significant evolutionary signals.
phylogenetic tree

A phylogenetic tree or evolutionary tree is a branching diagram or a 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.
Homologue nodes
Homologues nodes in biological networks represent homologues function/s.

Homologue = the same in certain aspect.
Basic terms - Pathways

**Biological pathway**
- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell.
- Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move.
Basic terms - Taxa

Taxa = hierarchical categories

- In the science of biological classification (taxonomy) biological taxa are arranged in a hierarchy (tree).

- Taxa which are joined together in a common branch of the tree are implied to have descended from a common ancestor.

- In the classification of prokaryotes, plants, and animals, the taxonomic categories in descending order are: kingdom, phylum (in plants, division), class, order, family, genus, species, and race.
The research problem

- Presenting the goals of the research
- Reviewing basic terms required for understanding the article
- Defining the problem discussed in the research
- Comparing biological networks & calculating the measure
- Building evolutionary trees
- Finding conserved regions
- Experimental running and its results
- Main conclusions and future directions
The main goal
Achieving efficient quantitative measure for the degree of similarity between biological networks.

Basic argument
Given two labeled networks A and B, the more similar they are, the fewer bits are required to describe B given A (and vice versa).

The desired measure
The minimal number of bits needed to describe one network given the description of the other.

* minimal number of bits \(\equiv\) shortest distance *
The research problem

**Mathematic model**
- $G_1, G_2 = $ the networks
- $D = $ The similarity measure (shortest distance between $G_1, G_2$)

=> $F(G_2 | G_1) = D$.

This implies **Kolmogorov complexity of string X**

In its relative form "$K(X|Y) = D$" :

$X, Y$ are strings

$D$ is the length of the shortest string $Z$, that when appended to the string $Y$ as an input to a universal Turing machine $U$, it yields $X$.

=> $U(Z, Y) = X$

**Kolmogorov complexity is incomputable**
The research problem

The problem

Finding the “shortest distance” is incomputable

Therefore:
A weaker notion should be used for the measure, by giving up the requirement for absolute minimum.

The problem of the research is to find such measure which:
- represents reliably the above "shortest distance"
- can be calculated efficiently
Networks comparison

- Presenting the goals of the research
- Reviewing basic terms required for understanding the article
- Defining the problem discussed in the research
- Comparing biological networks & calculating the measure
- Building evolutionary trees
- Finding conserved regions
- Experimental running and its results
- Main conclusions and future directions
Networks comparison – Intuitive direction

Intuitive direction

- The number of bits required to describe one network (B) given another network (A), reflects their differences.

- Therefore: it may be used as the measure for the degree of similarity between these networks.

* Describing $\equiv$ Encoding *
1. **Similar distances**
   The “distances” among corresponding pairs of nodes in networks of closely related species, are similar.

2. **Similar neighborhoods**
   Two homologue nodes may appear in networks of different species, with different labels. Even in such cases, their “neighborhood” in these networks will often be similar.

The distance  $\Rightarrow$  The length of the shortest paths, if directions are ignored.

Homologue  $\Rightarrow$  Node that have the same role.

Neighborhood  $\Rightarrow$  Defined in the next slide.
Basic definitions

- \( G(V, E) \) = Network \( G \) with a set of nodes \( V \), and a set of edges \( E \).

- \( (u, v) \) = Edge between the nodes \( u \) and \( v \). (nodes \( u \) and \( v \) are neighbors in \( G \).)

- A node \( u \) is a parent of node \( v \) if:
  - There is a directed edge \( (u, v) \) from \( u \) to \( v \)
  - In the case of undirected edge, any neighbor of \( v \) is also its parent.

- \( pa_v(G) \) = The set of the parents of node \( v \) in network \( G \).
Networks comparison – Network encoding

**Network encoding**

Each node $v$ in a directed network $G(V, E)$ with $n$ nodes, can be encoded using:

- $\log(n)$ bits, for the number of its parents
- $\log(|pa_v(G)|)$ bits, for the names of its parents

The entire network can be encoded using:

$$\sum_{v=1}^{n} \left( \log n + \log \left( \binom{n}{|pa_v(G)|} \right) \right) \text{ bits}.$$ 

**Description Length (DL) - the trivial encoding of (G)**

$DL(G) = \text{The number of bits required to encode the network } G(V, E) \text{ with its } n \text{ nodes}$

$$= \sum_{v=1}^{n} \left( \log n + \log \left( \binom{n}{|pa_v(G)|} \right) \right) \text{ bits}.$$
Networks comparison – Network subset encoding

Network subset encoding

A subset $T$ of a known set $S$ can be encoded by:

$$\log(|T|) + \log\left(\frac{|S|}{|T|}\right)$$

bits, denoted as: $Enc(T|S)$

- $\log(|T|)$ bits, for describing $T$
- $\log\left(\frac{|S|}{|T|}\right)$ bits, for describing a specific subset $S$

out of $\frac{|S|}{|T|}$ possible subsets.

This formula can be used for describing one network based on the other, namely in this study: 
"describe($G_2|G_1$)" - by certain subsets of them.
Main subsets in the algorithm

The algorithm is based on looking at the following subsets of nodes in $G_2$:

1. $V_1 \cap V_2 = \text{The group of nodes which appear in both } G_1 \text{ and } G_2 \text{, denoted as } V_{1,2}$

2. The parents of node $v$ where $v$ is in $V_1 \cap V_2$

3. The parents of node $v$ where $v$ is in $V_2 \setminus V_1$
Stage (1) - encoding the nodes in $V_{1,2}$

$V_{1,2} = V_1 \cap V_2$. (all the purple nodes 🌟 in the figure below.)

Using the formula:

$$Enc(T|S) = \log(|T|) + \log\left(\frac{|S|}{|T|}\right)$$

The group $V_{1,2}$ can be encoded by:

$Enc(V_{1,2}|V_1)$ bits.

Where:

$T = V_{1,2}$
$S = V_1$
Networks comparison – Describing $G_2$ given $G_1$

Stage(2) - encoding the parents of each node $v$ in $V_{1,2}$

Three cases to consider:

2(a) - Parents of node $v$ in both $G_1$ and $G_2$

2(b) - Parents of node $v$ in $G_2$ but not in $G_1$

2(c) - Parents of node $v$ which appear in $G_1$ but are not parents of $v$ in $G_1$
Stage 2(a) - encoding the parents of node $v$ which appear in both $G_1$ and $G_2$

This group: $A$, is consisted of all the nodes within:
$$pa_v(G_1) \cap pa_v(G_2)$$

Using again the formula for $Enc(T|S)$
This group can be encoded using:
$$Enc((pa_v(G_1) \cap pa_v(G_2))|pa_v(G_1))$$

Where:
$$T = pa_v(G_1) \cap pa_v(G_2)$$
$$S = pa_v(G_1)$$
Networks comparison – Describing $G_2$ given $G_1$

Stage 2(b) - encoding the parents of node $v$ which appear in $G_2$ but not in $G_1$

This group: B, is consisted of all the nodes within: $pa_v(G_2) \setminus V_1$.

Using again the formula for $Enc(T|S)$
This group can be encoded using:

$$Enc\left((pa_v(G_2) \setminus V_1) | V_2 \setminus V_{1,2}\right)$$

Where:

- $T = pa_v(G_2) \setminus V_1$
- $S = V_2 \setminus V_{1,2}$
Stage 2(c)- encoding the parents of node \( v \)

which appear in both \( G_1 \) and \( G_2 \)

but are not parents of \( v \) in \( G_1 \)

Using “\( Enc(T|S) \)”, this group: C, can be encoded by:

\[
\log(d) + \log(n_v) + \log\left( \frac{n_v^{v,d}}{n_v} \right) \text{ bits.}
\]

Where:

- \( n_v \) = the size of group C
- \( d \) = the minimal bidirectional radius around \( v \) in \( G_1 \)
- \( n_v^{v,d} \) = the number of all the nodes in this ball

* C is a subset of the nodes in the black ball
Networks comparison – Describing $G_2$ given $G_1$

Stage 2(c) - encoding the parents of node $v$

which appear in both $G_1$ and $G_2$
but are not parents of $v$ in $G_1$

Based on assumption 2 ("similar neighborhood") : usually $d$ is small.
Thus, this encoding of the set $C$ is economical.

If no such minimal radius $d$ exists, then $C$ is encoded by:

$$\log(n_v) + \log \left( \frac{V_{1,2}}{n_v} \right)$$

bits.

Where:

$n_v = \text{the size of group } C$
$V_{1,2} = \text{is used instead of the ball, for the group that contains } C$
Stage(3) - encoding the parents of each node $\nu$

where $\nu$ is in $V_2 \setminus V_1$

This group can be encoded by either:
- The trivial network encoding
- The neighbors of its homologue node in $G_1$.

(Based on assumption 2: “similar neighborhoods”, describing the parents of $\nu$ in $G_2$ by the parents of its homologue node in $G_1$, is economical).
Networks comparison – Describing $G_2$ given $G_1$

Stage(3) - encoding the parents of each node $v$

where $v$ is in $V_2 \setminus V_1$

Encoding by neighbors of homologue node:

$log |V_1| + c_v$ bits

Where:

$c_v = \text{the number of bits needed to describe the parents of node } v \text{ by the parents of its homologue node (using steps 1,2)}$

Trivial network encoding:

$log |V_2| + \log \left( \left| \text{pa}_v(G_2) \right| \right)$ bits.

In this stage the minimum between them is chosen.
Goals

Basic terms

The problem

Comparison

Evolution

Conservation

Experiment

discussion

Networks comparison – Describing $G_2$ given $G_1$

Stage(4) – Selecting minimum value for the DL

After completing stages 1-3, the algorithm should choose the minimum between:

- The value accumulated from those stages

- The trivial encoding of network $G_2$:
  \[
  \sum_{k=1}^{n} \log(n) + \log\left(|pa_k(G_2)|\right) \text{ bits}
  \]

This final result is the:

Description Length of $G_2$ given $G_1$, namely: $DL(G_2|G_1)$ - the intuitive measure for the distance between $G_1$ and $G_2$
Networks comparison – $DL(G_2|G_1)$

$DL(G_2|G_1)$ properties

This distance measure should be:
- Non negative ✓
- Symmetric ✗
- Satisfying the triangle inequality

DL is not symmetric because it does not reflect nodes which appear in $G_1$ and not in $G_2$
Networks comparison – $RDL(G_1, G_2)$

Relative Description Length (RDL) - definition

$RDL(G_1, G_2)$ is defined in order to have a symmetric measure:

$$RDL(G_1, G_2) = \frac{DL(G_1|G_2)}{DL(G_1)} + \frac{DL(G_2|G_1)}{DL(G_2)}$$

The first term is the ratio:

(The number of bits needed to describe $G_1$ given $G_2$)

(The number of bits in the trivial description of $G_1$)

The second term is symmetric accordingly.
Networks comparison – $RDL(G_1, G_2)$

Relative Description Length (RDL) - properties

$$RDL(G_1, G_2) = DL(G_1 | G_2) / DL(G_1) + DL(G_2 | G_1) / DL(G_2)$$

1. Non negative $\quad 0 \leq RDL(G_1, G_2) \leq 2$
2. Symmetric $\quad \checkmark$
3. Time complexity
   $$O(|E| |V| + |V|^2 \log|V|) \Rightarrow \text{computable and efficient}$$

* The next expected properties should be examined through experimental running. The results will be described in that context. *

4. RDL should grow as function of the evolutionary time
5. RDL should satisfy the triangle inequality
Summary

The measure $RDL(G_1, G_2)$ for comparing two networks is calculated in the following steps:

1. Calculating $DL(G_1|G_2)$
2. Calculating $DL(G_2|G_1)$
3. Calculating $RDL(G_1, G_2) = \frac{DL(G_1|G_2)}{DL(G_1)} + \frac{DL(G_2|G_1)}{DL(G_2)}$

Time complexity of the calculations is:

$O(|E||V| + |V|^2 \log |V|)$
Goals

Presenting the goals of the research

Basic terms

Reviewing basic terms required for understanding the article

The problem

Defining the problem discussed in the research

Comparison

Comparing biological networks & calculating the measure

Evolution

Building evolutionary trees

Conservation

Finding conserved regions

Experiment

Experimental running and its results

discussion

Main conclusions and future directions
Evolution – phylogenetic trees.

The method for building phylogenetic trees

1. Calculating $RDL(G_1, G_2)$ for each pair of networks

2. Neighbor joining algorithm
Evolution – phylogenetic trees.

**The time complexity for building phylogenetic trees**

- Neighbor joining algorithm – for $N$ networks: $O(N^3)$

- Calculating $RDL(G_1, G_2)$ with up to $V$ nodes each:
  $O(|E||V| + V^2 \log|V|)$

- Building phylogenetic trees
  $O(N^3 + N^2 (|E||V| + V^2 \log|V|))$
Finding conserved regions

- Presenting the goals of the research
- Reviewing basic terms required for understanding the article
- Defining the problem discussed in the research
- Comparing biological networks & calculating the measure
- Building evolutionary trees
- Finding conserved regions
- Experimental running and its results
- Main conclusions and future directions
**Finding conserved regions - Definitions**

**Auxiliary definitions**

\( G_1, G_2 = \text{two networks} \)

\( v = \text{a node shared to} \ G_1, G_2 \)

\( B_1, B_2 = \text{two balls with bidirectional radius} \ d \ \text{centered at} \ v \ \text{in} \ G_1 \ \text{and} \ G_2, \ \text{respectively.} \)

\( (d \ \text{is an integer}) \)

**d-conservation score**

The “d-conservation score” of the node \( v \) with respect to \( G_1, G_2 \) and \( d \) as defined above is:

\[ RDL(B_1, B_2) \]

If \( v \) does not appear in the two graphs, then: the “d-conservations score” of the node \( v \) with respect to the two graphs is defined to be \( \infty \)
Based on the previous auxiliary definitions:

"\((d, c)\)-conserved" node

A node \(v\) is "\((d, c)\)-conserved" in \(G_1\) and \(G_2\) if 
\[ RDL(B_1, B_2) \leq c \]

\(c\) is a real number

"\((d, c)\)-conserved" regions

The "\((d, c)\)-conserved" region of two networks \(G_1\) and \(G_2\) is the intersection of the two sub graphs of \(G_1\) and \(G_2\) induced by "\((d, c)\)-conserved" nodes with respect to \(G_1\) and \(G_2\).

(In each network the conserved region contains all the conserved nodes and all the original edges between them.)
Finding Conserved regions - In two networks

Finding the "\((d, c)\)-conserved" region of \(G_1\) and \(G_2\)

Given the networks \(G_1\), \(G_2\), radius \(d\) and a number \(c\)

1. For each node common to \(G_1\) and \(G_2\), compute its "d-conservation score": \(RDL(B_1, B_2)\)

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 \(e \in E'_1\) is an edge in \(G'_1\) if:
     its two endpoints are in \(V'_1\)
     and it is an edge in \(E_1\)

3. The graph \(G'_2 = (V'_2, E'_2)\) is defined analogously.

\(G'_1\) and \(G'_2\) are the "\((d, c)\)-conserved" regions of \(G_1\) and \(G_2\) accordingly.
Generalization for N networks

\[ G_1, G_2, \ldots, G_N = \text{N labeled graphs} \]

The natural generalization for N networks is the following requirement:

Finding nodes that are “\((d, c)\)-conserved” in all the \( \binom{N}{2} \) networks pairs of \( G_1, G_2, \ldots, G_N \)

This requirement is too strict and should be relaxed!
Finding Conserved regions - Generalization

**Generalization - requirement relaxation**

\[ k = \text{an integer in the range } 1 \leq k \leq \binom{N}{2} \]

**Definition:**

"\((d, c, k)\)-conserved" node

A node \(v\) is \"\((d, c, k)\)-conserved\" with respect to the \(N\) networks \(G_1, G_2, \ldots, G_N\) if:

\[ v \text{ is } (d, c)\text{-conserved in at least } k \text{ out of the } \left(\binom{N}{2}\right) \text{ pairs} \]

**Properties:** for larger \(k\) (a stricter requirement), the sizes of the conserved graphs decrease.
Experimental results

Presenting the goals of the research

Reviewing basic terms required for understanding the article

Defining the problem discussed in the research

Comparing biological networks & calculating the measure

Building evolutionary trees

Finding conserved regions

Experimental running and its results

Main conclusions and future directions
Experimental results

Experiment contents

- Building phylogenetic trees
- Finding conserved regions
- Further tests of the RDL

The experiment was performed by:

- Running the algorithms described in this study on various sets of metabolic networks taken from KEGG database
- Running the algorithm for finding conserved regions on two protein interaction networks
- Reviewing the results
**Building phylogenetic trees**

- The first running: small subset of 19 taxa: 9 eukaryotes, 5 prokaryotes, 5 arches. Each network with more than 900 nodes.
- The resulting tree is reasonably close to the conventional taxonomy of these species.

**An exception:**

This result may suggest that mouse is more similar to human than rat, when complete metabolic networks are compared.
Building phylogenetic trees

- The second running: all 194 networks with more than 100 nodes.

- In the resulting tree:
  - Many taxa appeared in the suitable clusters
  - The exceptions were explained by reasons such as:
    - Imperfect data
    - Specific changes in metabolic pathways which are identified by the algorithm as opposed to the conventional taxa.

=> Most of the results validated the algorithm
Building phylogenetic trees

- The results of the second running:
Experimental results - Finding conserved regions

Finding conserved regions in metabolic networks

- 4 different dataset were used as input

- The parameters: $c$ & $k$ were adjusted specifically to each of these datasets, so the algorithm choose less than 10% of the nodes, for the conserved regions.

- The "conserved graphs" generated from running the algorithm, were sorted lexicographically by their labels, in order to show their similarity
Examples (1)
- Dataset: Human and Gamma Enterobacteria
- The conserved regions are expected to be common to many taxa, because human and bacteria diverged billions years ago
- Resulting conserved networks had 94 nodes which were common to both species (out of 900 in human and 1010 in the bacteria).
- Example of pathways included in these conserved networks: pathways which are parts of different building blocks of the DNA => indeed common to many taxa
Experimental results - Finding conserved regions

Examples(1)
- The conserved graphs of: Human and Gamma Enterobacteria
Experimental results - Finding conserved regions

Examples (2)

- Dataset: nine Eukaryotes with more than 900 nodes in KEGG database.

- The parameters for $d, c, k$ were: 20, 0.7, 6 accordingly.

- The resulting 9 conserved networks include 84-95 nodes

- Example of conserved sub-graphs which were found: parts of the pathways of
  - Pyrimidine synthesis metabolism
  - Riboflavin synthesis metabolism
  - Cysteine synthesis metabolism
  - ...

⇒ The algorithm found many conserved regions which are typical to Eukaryotes (among others).
Experimental results - Finding conserved regions

Examples (2)
- The conserved graph of the nine Eukaryotes

- Human
- Mouse
- Rat
- Fruit Fly
- C. elegans
- Arabidopsis
- S. cerevisiae
- S. pombe
- A. Gossypii
Experimental results - Finding conserved regions

Finding conserved regions in PPI networks

Current limitations and difficulties:

- Protein interaction networks do not have unique labels that are shared across species

- The method used in the research for deciding if two proteins are identical is simplistic, and sometimes may miss the true homologous proteins

- Smaller database of known networks: only for 8 species.

- The existing protein interaction predictions are not too reliable (false positive rate can be > 20%)

- Complicated network topology: many cycles and many number of parents
Example (3)
- Dataset: PPI networks of yeast and drosophila

- Dashed edges belong to drosophila
- Continuous edges belong to yeast
Example (3)
- Dataset: PPI networks of yeast and drosophila

Despite all the problems mentioned above:
From checking the functionalities of few dozens of the conserved nodes, the results seemed good.
Further tests of the RDL (completion to slide of “RDL properties”)

4. **RDL should grow as function of the evolutionary time**

- A simulation of evolution was performed on initial network, by steps of random adding or removing a node or an edge.

- RDL was calculated for the initial network and the new network after each step of the simulation.
The simulation results:

The graph shows that RDL growth is close to linear as function of the evolutionary time.
Further tests of the RDL (completion to slide of “RDL properties”)

5. RDL should satisfying the triangle inequality

- Not all triplets in the distance matrix of all 240 networks of KEGG satisfied the triangle inequality

- However:
  by empirically checking all these triplets it was found that only 363 out of 2,257,280 (0.016%) violated this rule, usually due to partial data.

- After removing all network with less than 100 nodes, the triplets of all the remaining 194 networks, satisfied the triangle inequality
Conclusions from the experiment results

- For metabolic networks:
The results were good and validated the RDL measure and the algorithms proposed in the study.

- For protein interaction network:
The method for deciding that two proteins are identical, is too simplistic and adjustments are needed.

Despite all the problems the findings supported the feasibility of using RDL measure for analyzing these networks.

- RDL grows with evolutionary time and for large enough networks also satisfies the triangle inequality.
Discussion

Presenting the goals of the research

Reviewing basic terms required for understanding the article

Defining the problem discussed in the research

Comparing biological networks & calculating the measure

Building evolutionary trees

Finding conserved regions

Experimental running and its results

Main conclusions and future directions
General insights

- The procedure described in this study is biologically reasonable, and gives meaningful results.

- The algorithms may be applied on additional types of biological networks, and from other databases (not only KEGG).

- Since the database of biological networks, are still in their infancy, the precision of the results is likely to evolve and improve, as more reliable data is gathered.
Discussion - future direction

Main procedure enhancements and generalization

- With certain modifications, the procedure may improve speed or accuracy. (Still, the best variant of the procedure, and its exact parameters for finding conserved regions, may depend on the type of the input).

- It is desirable to include explicit edge labels in the representation of the networks, and the RDL approach allows such modification at relative ease.

- Different representations of the directed graph are also possible, but their efficiency should be carefully analyzed.
Biological Networks: Comparison, Conservation, and Evolution via Relative Description Length

By: Tamir Tuller & Benny Chor