Optimal Network Alignment with Graphlet Degree Vectors

Tijana Milenković, Weng Leong Ng, Wayne Hayes, Nataša Pržulj

ZHENG Yufei

Nov. 25, 2015
Background

• What is **A Network Alignment**? (General idea)
  – A mapping between the nodes of networks being compared

• Categories
  – Local
    • Subnetwork to subnetwork(s)
  – Global
    • Node to node
    • Overall similarity
• What does network alignment do?
  – To do comparative analysis of networks

• What can we know from comparative analysis of biological networks?
  – Infer phylogenetic relationships
  – Guide the transfer of knowledge
  – Give insights into evolutionary conservation of protein

• Why network alignment specifically?
  – Exact network comparisons are infeasible
• What makes this method (H-GRAAL) special?
  – Find optimal alignments relative to any fixed, deterministic cost function
  – Based solely on network topology

Details of Hungarian algorithm can be found in:

Methods

• Signature similarities

Notations: G – Network, V – the set of all nodes in G

**An induced subgraph** on a node set \( S \subseteq V \) of G is obtained by taking S and *all* edges of G having both end-nodes in S.

**Graphlets** – small connected non-isomorphic induced subgraphs of a large network
Small-world nature...

2-node graphlet

3-node graphlets

4-node graphlets

5-node graphlets

G₀  G₁  G₂  G₃  G₄  G₅  G₆  G₇  G₈

G₉  G₁₀  G₁₁  G₁₂  G₁₃  G₁₄  G₁₅  G₁₆  G₁₇  G₁₈  G₁₉

G₂₀  G₂₁  G₂₂  G₂₃  G₂₄  G₂₅  G₂₆  G₂₇  G₂₈  G₂₉
Methods

• Signature similarities

Notations: G – Network, V – the set of all nodes in G

An induced subgraph on a node set $S \subseteq V$ of $G$ is obtained by taking $S$ and all edges of $G$ having both end-nodes in $S$.

Graphlets – small connected non-isomorphic induced subgraphs of a large network

“symmetries” between nodes of a graphlet

Orbits
**Graphlet degrees** - the number of graphlets that the node touches

The **signature** of a node *(GDV)* - The full vector of 73 coordinates
The computation of *signature similarity* between two nodes:
For a node $u \in G$, $u_i$ denotes the $i^{th}$ coordinate of its signature vector.

The *distance* $D_i(u,v)$ between the $i^{th}$ orbits of nodes $u$ and $v$:

$$D_i(u,v) = w_i \times \frac{\log(u_i + 1) - \log(v_i + 1)}{\log(\max\{u_i, v_i\} + 2)}$$

The *total distance* $D(u,v)$ between nodes $u$ and $v$:

$$D(u,v) = \frac{\sum_{i=0}^{72} D_i}{\sum_{i=0}^{72} w_i}$$

*Signature similarity* $S(u,v)$ between nodes $u$ and $v$

$$S(u,v) = 1 - D(u,v)$$

How to determine the *weight* of each orbit?
Intuitions about weight...

2-node graphlet

3-node graphlets

4-node graphlets

5-node graphlets

G_0

G_1

G_2

G_3

G_4

G_5

G_6

G_7

G_8

G_9

G_10

G_11

G_12

G_13

G_14

G_15

G_16

G_17

G_18

G_19

G_20

G_21

G_22

G_23

G_24

G_25

G_26

G_27

G_28

G_29
$O_i$ - (orbit dependency count) the number of orbits that affect orbit $i$.

$O_{44} = 5$

$O_{15} = 4$

$w_i = 1 - \frac{\log(o_i)}{\log(73)}$

A few definitions...

\[ V(G) \triangleq \text{the set of nodes of network } G \]

Let \( G_1 \) and \( G_2 \) be networks and assume without loss of generality that \( |V(G_1)| \leq |V(G_2)| \)

An **alignment** of \( G_1 \) to \( G_2 \) is a set of ordered pairs \((u,v)\), \( u \in V(G_1) \) and \( v \in V(G_2) \), such that no two ordered pairs share the same \( G_1 \)-node or the same \( G_2 \)-node.

A **maximum alignment** of \( G_1 \) to \( G_2 \) is an alignment where every \( G_1 \)-node is in some aligned pair of the alignment.

\[ \text{Alignment} \triangleq \text{maximum alignment} \]
• **H-GRAAL Algorithm**
  – Cost function

\[ \alpha \in [0,1] \triangleq \text{A parameter that weighs the contribution of signature similarity} \]

\[ 1 - \alpha \in [0,1] \triangleq \text{A parameter that weights the contribution of degrees of the nodes } u \text{ and } v. \]

\[ \deg(u) \triangleq \text{The degree of a node } u \]

\[ \max_{\text{deg}}(G) \triangleq \text{The maximum degree of node in network } G \]

The degree of the most ‘popular’ node.
The cost of aligning nodes $u$ and $v$ $C(u,v)$:

$$C(u, v) = 2 - ((1 - \alpha) \times T(u, v) + \alpha \times S(u, v))$$

A complete bipartite graph with $V(G_1)$ and $V(G_2)$ as the bipartition and each edge $(u,v)$ from $V(G_1)$ to $V(G_2)$ is labeled with the node alignment cost between $u$ and $v$.
– H-GRAAL

- Find an alignment by minimizing the cost summed over all aligned pairs
- Ways to generate other optimal alignments
  - A removal of \((u,v)\)
  - Dynamic Hungarian algorithm

- Core alignment

```
Procedure Find-All-Optimizing-Pairs \((U, V, C_0)\)
1: \(A_0, M_0 \leftarrow \text{H-GRAAL} \,(U, V, C_0)\)
2: \(C_{\text{min}} \leftarrow \text{Alignment-Cost} \,(A_0)\)
3: \(S \leftarrow \emptyset\)
4: for all \(u \in U\) do
5: \(A \leftarrow A_0, M \leftarrow M_0, C \leftarrow C_0\)
6: while \(\text{Alignment-Cost} \,(A) = c_{\text{min}}\) do
7: \(v \leftarrow A[u]\)
8: \(S \leftarrow S \cup \{v\}\)
9: \(A, M \leftarrow \text{Dynamic-H-GRAAL} \,(U, V, C, A, M, \{((u, v), +\infty)\})\)
10: \(C[u][v] \leftarrow +\infty\)
11: end while
12: end for
```
Measures of Quality

• Theoretically
  – The edge correctness (EC)
  – The node correctness (NC)
  – The interaction correctness (IC)

• In real applications
  – The edge correctness (EC)
    percentage of edges in one graph that are aligned to edges in the other graph
  – Common connected subgraphs (CCSs)
    connected subgraphs that appears in both networks

Both high EC and large and dense CCSs are desired.
Results and Discussion

H-GRAAL vs. GRAAL

- Method validation

Align the largest connected component of the high-confidence yeast *S. cerevisiae* PPI network by Collins et al.

- High-confidence data
- 20% lower-confidence data
- 15% lower-confidence data
- 10% lower-confidence data
- 5% lower-confidence data
- 25% lower-confidence data

Interactions from the lower-confidence data set described by Collins et al.
Comparison of H-GRAAL with GRAAL with respect to **NC, EC** and **IC**

- NC
- EC
- IC
- Non-zero $\alpha$
- All noise levels
• Pairwise alignment of PPI networks of yeast and human
  – Why these 2 networks?
    • Yeast PPI network by Collins et al - lower levels of noise
    • Human PPI network by Radivojac et al - give valuable insights into complex diseases
    • These two networks have already been analyzed with GRAAL
  – Data
    • the highest EC of 10.92% over all $\alpha$ between 0 and 1 (slightly lower than GRAAL)
    • More contiguous and denser

$16127 \times 10.92\% = 1761$
The largest **common connected subgraph** resulting from the alignment of the yeast and human PPI networks, consisting of 1,290 interactions amongst 317 proteins.
– Statistical significance

• Compared to a random alignment of these two particular networks

Given a random alignment of the yeast and human PPI networks, compute the probability of obtaining EC of 10.92% or better.

Given \( G_1(V_1, E_1), G_2(V_2, E_2) \), define our null model of random alignment as a random mapping: \( g : E_1 \rightarrow V_2 \times V_2 \)

\[
\begin{align*}
n_1 &= |V_1|, n_2 = |V_2|, m_1 = |E_1|, m_2 = |E_2|, EC = x\%, p &= \frac{n_2(n_2-1)}{2} \\
k &= [m_1 \times x\%] \\
\text{the probability } P \text{ of successfully aligning } k \text{ or more edges: } \quad P = \sum_{i=k}^{m_2} \frac{C_{m_2}^i C_{m_1-i}^{m_1-i}}{C_{p-m_2}^{m_1}}
\end{align*}
\]
• The amount of topological similarity

Estimate how much similarity one would expect to find with H-GRAAL between two random networks

Align with H-GRAAL networks drawn from several different random graph models that have the same number of nodes and edges as the yeast and human networks

Apply the Vysochanskij—Petunin inequality:

\[ P(|X - \mu| \geq \lambda \mu) \leq \frac{4}{9\lambda^2} \]

Chebyshev’s Inequality

\[ P(|X - \mu| \geq k \mu) \leq \frac{1}{k^2} \]
Biological significance

- quantify the biological significance
- validate the biological quality
– Application to protein prediction

Annotated proteins

Unannotated proteins

MF
BP
CC

literature search & text mining

Validation hit-rate

Molecular function (MF)
Biological process (BP)
Cellular component (CC)
Validation hit-rates to given biological characteristic for human:

- MF
- BP
- CC

Validation hit-rates to given biological characteristic for yeast:

- MF
- BP
- CC

Human Biological characteristics:

- H-GRAAL
- GRAAL

Yeast Biological characteristics:

- H-GRAAL
- GRAAL
• Reconstruction of phylogenetic trees by aligning metabolic pathways across species
  – Related attempts
– H-GRAAL vs. Existing attempts

information source is fundamentally different
• Results

A

Alveolates

TPV
TAN
CHO
CPV
PFA

Entamoeba

EHI

Cellular slime mold

DDI

Sequence-based tree

H-GRAAL’s tree

GRAAL’s tree

B

Ascomycetes

SCE
AGO
CAL
SPO

Microsporidian

ECU

Basidiomycetes

CNE

Sequence-based tree

H-GRAAL’s tree

GRAAL’s tree
About the signature similarity...

- In this paper

$$D_i(u,v) = w_i \times \frac{|\log(u_i + 1) - \log(v_i + 1)|}{\log(\max\{u_i, v_i\} + 2)}$$

$$D(u,v) = \frac{\sum_{i=0}^{72} D_i}{\sum_{i=0}^{72} w_i}$$

$$S(u,v) = 1 - D(u,v)$$

- In functional analysis

$H$ - Hilbert Space

$$\{m_1, m_2, \ldots m_k\}$$

- An orthonormal set

$$\forall h \in H,$$

$$h = \sum_{i=1}^{k} \langle x, m_i \rangle m_i$$

Each coordinate

S = \frac{\langle h_u, h_v \rangle}{\|h_u\| \|h_v\|}$$

GDV

73 unit vectors
Thanks for listening! מודה!