CENG 734
Advanced Topics in Bioinformatics

Week 13
Bioimage Informatics

Fall 2010-2011
Quiz #10

1. In the paper discussed last week, how is “network alignment” used to compute an edge score for an edge in a PPI network? You do not need to give any formula, verbal explanation is sufficient.
From last week: Paper #2

- **Global Network Alignment** by Kuchaiev and Przulj, Nature Precedings (not yet published).
Overview

• Main aim is NOT to find conserved modules between two/multiple species
• Main aim is to provide a global (i.e., for all proteins of the query network) matching
• Combines sequence similarity with network topology similarity to define a matching confidence matrix between proteins
Formal definition

• Provide a matching between two graphs two maximize the edge correctness score

\[
EC = \frac{\left| \{(u, v) \in E_1 \land (f(u), f(v)) \in E_2\} \right|}{|E_1|} \times 100\%
\]

• NP-complete if want to find and EC of 100
• Therefore maximizing this is equally difficult \(\rightarrow\) NP-Hard
• A heuristic solution is provided
Measures used in the “confidence matrix”

1. Graphlet degree signature distance ($SD$) (Milenkovic and Pržulj, 2008)
2. Relative degree difference ($DD$)
3. Relative clustering coefficient difference ($CD$)
4. Relative eccentricity difference ($ED$)
5. BLAST $E$-value for protein sequence similarity ($SeqD$)

Combined confidence:

$$C(i, j) = conf_{SD}(i, j) + conf_{DD}(i, j) + conf_{CD}(i, j) +$$
$$+ conf_{ED}(i, j) + conf_{SeqD}(i, j)$$
The algorithm

Algorithm 1 $M$-GRAAL($G_1, G_2$)

Construct, or read in the cost matrices and build the matrix of confidence scores, $C$, as well as the priority queue of node pairs ordered by their confidence scores.

Initialize alignment $A$ to an empty set.

while there are unaligned nodes in $G_1$ do

Use the priority queue to find a seed pair of nodes, $(u, v)$, $u \in G_1, v \in G_2$, i.e., the pair of nodes that can be aligned with the highest confidence, $C(u, v)$. Break ties randomly.

Add $(u, v)$ to alignment $A$.

for all $k \in \{1, ..., \min\{eccen(u), eccen(v)\}\}$ do

Construct the $k^{th}$ neighborhood of $u$ in $G_1$, $N_{G_1}^k(u)$, and the $k^{th}$ neighborhood of $v$ in $G_2$, $N_{G_2}^k(v)$.

$align\_neighborhoods(N_{G_1}^k(u), N_{G_2}^k(v), C, A)$

end for

If there are still unaligned nodes in $G_1$, raise both graphs to the next power (up to the 3rd power).

end while

return alignment $A$. 
Algorithm 2 align_neighborhods\( (N^k_{G_1}(u), N^k_{G_2}(v), C, A) \)

1. Construct a bipartite graph \( BP(N^k_{G_1}(u), N^k_{G_2}(v), E) \) with node partitions being \( N^k_{G_1}(u) \) and \( N^k_{G_2}(v) \) as follows:

   - Check the current alignment \( A \) and add an edge \((u', v')\) to \( E \), \( u' \in N^k_{G_1}(u) \), \( v' \in N^k_{G_2}(v) \), if and only if nodes \( u' \) and \( v' \) have aligned neighbors. Hence, aligning them will increase \( EC \) by at least 1.

   - To each edge \((n, m)\) in \( E \), assign the weight \( C(n, m) \), the confidence with which we can align \( n \) and \( m \).

2. Solve the Maximum Weight Bipartite Matching Problem for bipartite graph \( BP \) constructed above.

3. Add the optimal matching found in Step 2 above to the current alignment \( A \).
Results

• Yeast and human networks are aligned
• Yeast network: 16,127 interactions amongst 2,390 proteins
• Human network: 41,456 interactions amongst 9,141 proteins.
• All combinations of the five measures have been tried
Results

• The highest edge correctness is achieved by the graphlet signatures only

• However, the most stable alignment is obtained by using four of the measures (all except eccentricity)
## Results

Table 1. Fraction of protein pairs in the alignment of yeast and human that share GO terms.

<table>
<thead>
<tr>
<th></th>
<th>Alignment 1</th>
<th>Alignment 2</th>
<th>Alignment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 1$</td>
<td>46.67% ($10^{-9}$)</td>
<td>50.58% ($3.6 \times 10^{-8}$)</td>
<td>47.84% ($10^{-9}$)</td>
</tr>
<tr>
<td>$\geq 2$</td>
<td>14% ($3.5 \times 10^{-4}$)</td>
<td>20.52% ($4 \times 10^{-8}$)</td>
<td>16.67% ($10^{-9}$)</td>
</tr>
<tr>
<td>$\geq 3$</td>
<td>3.58% ($8.4 \times 10^{-2}$)</td>
<td>8.19% ($10^{-9}$)</td>
<td>6.08% ($10^{-9}$)</td>
</tr>
<tr>
<td>$\geq 4$</td>
<td>1.01% (0.36)</td>
<td>4.10% ($5 \times 10^{-8}$)</td>
<td>2.81% ($10^{-9}$)</td>
</tr>
<tr>
<td>$\geq 5$</td>
<td>0.32% (0.49)</td>
<td>1.89% ($1.8 \times 10^{-8}$)</td>
<td>1.61% ($10^{-9}$)</td>
</tr>
<tr>
<td>$\geq 6$</td>
<td>0.05% (0.36)</td>
<td>0.97% ($1.4 \times 10^{-8}$)</td>
<td>0.97% ($10^{-9}$)</td>
</tr>
</tbody>
</table>

*Alignment 1* is purely topological. *Alignment 2* is obtained when only sequence information is used to score node pairs. *Alignment 3* is obtained when signatures, degrees, clustering coefficients and BLAST E-values are used together to score node pairs. Numbers in brackets represent p-values.
Using network comparison for phylogenetic analysis

Fig. 1. Phylogeny of the five investigated herpesviruses. (A) The “gold standard” tree (McGeoch and Gatherer, 2005; McGeoch et al., 2006); (B) Unrooted phylogenetic tree reconstructed from edge correctness scores of topological alignments produced by M-GRAAL.
Bioimage Informatics

• Automated analysis of time-lapse fluorescence microscopy images: from live cell images to intracellular foci

• By Dzyubachyk et al. in Bioinformatics 26(19), 2010.
Overview

• Provide a complete system with no parameters to tune (i.e., no training) to analyze live cell images both at the cellular and subcellular level.

• Process multilayer time-lapse microscopy images (4D)

• Provide comparable results to human experts

• Foci: plural of “focus”.
  – May correspond to a certain subcellular region that is of interest.
Tasks

• Cell segmentation
• Cell tracking
• Foci segmentation
  – Foci counting
• Foci pattern analysis
  – Cell phase identification
• Consider fluorescent foci
Cell segmentation and tracking

• Level-set based technique proposed earlier is able to handle topology changes in robust under different intensities

• Extend the level-set based technique with motion correction.

  – Register each cell to a global coordinate system to differentiate between global cell movement from intracellular component movement
Validation experiments

• In the first experiment, they investigate the time course of nuclear foci formation and disappearance upon treatment with ionizing radiation of the 53BP1 DNA repair protein.
  – Foci counting

• In the second experiment, they test the system for identifying cell phases in time-lapse images of proliferating cell nuclear antigen (PCNA)–green fluorescent protein (GFP)-stained cells.
Results

Fig. 1. Example of motion correction using the proposed approach. The two top rows show the motion of one cell extracted from a time-lapse fluorescence microscopy image dataset (outlined in white). One slice ($z = 1$) is shown for time steps 1, 11, 21, 31, 41, 51, 61, 71, 81 and 84. The third row shows (magnified) the result of cell motion correction after segmentation and tracking. In this case, only the global motion of the nucleus is subtracted.
**Fig. 2.** Example of foci segmentation using our algorithm: (a) images of the same nucleus in five different time steps (1, 9, 46, 65, 71), each representing one of the phases of the cell cycle (G1, early-S, middle-S, late-S, G2); (b) results of applying patch-based reconstruction to each image; (c) initially detected foci markers (dots in different shades of gray); (d) results of the graph-cut-based segmentation algorithm; and (e) final results after foci selection. All images are the first slice ($z=1$) of the corresponding 3D image stack.
Fig. 3. Comparison between manual (light-gray) and automated (dark-gray) 53BP1 foci counting for normal ES cells (IB-10) in terms of (a) the percentage of the positive cells and (b) the average number of foci per cell at various time points. For each of the measures the corresponding values and the obtained polynomial trend lines are shown.
Results

Fig. 6. Comparison between manual and automated detection of phase transition moments in PCNA-stained cells. The four plots correspond to the four possible phase transitions: (a) G1 to early-S (21 cases); (b) early-S to middle-S (29 cases); (c) middle-S to late-S (26 cases); and (d) late-S to G2 (22 cases). In each case, the difference in detection times between the automated method and each of the two observers is plotted. A missing point on one of the curves in (c) means that the corresponding phase transition was not detected by the corresponding observer.
Next Week

• Project presentations
• 12 projects total
• 6 presentations per week
• Max 25 min per presentation
• Here are the lucky(!) presenters for next week
  – Hilal and Yigit
  – Serdar
  – Erhan
  – Sefa
  – Ayse Gul
  – Burak