Network problems

- Network inference
  - Given raw experimental data
  - Infer network structure
- Motif finding
  - Identify common subgraph topologies
- Module detection
  - Identify subgraphs that perform same function
- Conserved modules
  - Identify modules that are shared in networks of multiple species
Network motifs

- Problem: Find subgraph topologies that are statistically more frequent than expected
- Brute force approach
  - Count all topologies of subgraphs of size m
  - Randomize graph (retain degree distribution) and count again
  - Output topologies that are over/under represented

*Feed-forward loop*: over-represented in regulatory networks

not very common
Network modules

• Modules: dense (highly-connected) subgraphs (e.g., large cliques or partially incomplete cliques)

• Problem: Identify the component modules of a network

• Difficulty: definition of module is not precise
  • Hierarchical networks have modules at multiple scales
  • At what scale to define modules?
Conserved modules

- Identify *modules* in multiple species that have “conserved” topology

- Use sequence alignment to identify homologous proteins and establish correspondence between networks

- Using correspondence, output subsets of nodes with similar topology
Comparative network analysis

• Compare networks from different...
  • interaction detection methods
    • yeast 2-hybrid, mass spectrometry, etc.
  • conditions
    • heat, media, other stresses
  • time points
    • development, cell cycle
  • species
Comparative tasks

- Integration
  - Combine networks derived from different methods (e.g. experimental data types)

- Alignment
  - Identify nodes, edges, modules common to two networks (e.g., from different species)

- Database query
  - Identify subnetworks similar to query in database of networks
Conserved interactions

- Network comparison between species also requires sequence comparison
- Protein sets compared to identify orthologs
- Common technique: highest scoring BLAST hits used for establishing correspondences
Conserved modules

- Conserved module: orthologous subnetwork with significantly similar edge presence/absence
Network alignment graph

• Analogous to pairwise sequence alignment
**Conserved module detection**

<table>
<thead>
<tr>
<th>Biological networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 1 (Condition/type 1)</td>
</tr>
<tr>
<td>Species 2 (Condition/type 2)</td>
</tr>
</tbody>
</table>

- **Matched proteins**
  - Match protein pairs that are sequence-similar
  - Protein sequences: PKSDIDVDLCSELMAKACSE-GV
  - PKS +D+DLSEL+ KAC++ + PKSSLIDDLCSELIIKACTDCKI

- **Network alignment**
  - Conserved interactions
  - Matched protein pairs

- **High-scoring conserved subnetworks**

- **Search algorithm**

(Sharan & Ideker, 2006)
Real module example

Module for RNA metabolism (Sharan et al., 2005)

- Yeast
- Worm
- Fly

- Note: a protein may have more than one ortholog in another network
Basic alignment strategy

• Define scoring function on subnetworks
  • high score \(\Rightarrow\) conserved module

• Use BLAST to infer orthologous proteins

• Identify “seeds” around each protein: small conserved subnetworks centered around the protein

• Grow seeds by adding proteins that increase alignment score
Subnetwork modeling

• We wish to calculate the likelihood of a certain subnetwork $U$ under different models

• Subnetwork model ($M_s$)
  • Connectivity of $U$ given by target graph $H$, each edge in $H$ appearing in $U$ with probability $\beta$ (large)

• Null model ($M_n$)
  • Each edge appears with probability according to random graph distribution (but with degree distribution fixed)

(Sharan et al., 2005)
Noisy observations

• Typically weight edges in graph according to confidence in interaction (expressed as a probability)

• Let

  • $T_{uv}$: event that proteins $u$, $v$ interact
  • $F_{uv}$: event that proteins $u$, $v$ do not interact
  • $O_{uv}$: observations of possible interactions between proteins $u$ and $v$
Subnetwork model probability

- Assume (for explanatory purposes) that subnetwork model is a clique:

\[
Pr(O_U|M_s) = \prod_{(u,v) \in U \times U} Pr(O_{uv}|M_s)
\]

\[
= \prod_{(u,v) \in U \times U} \left[ Pr(O_{uv}|T_{uv}, M_s)Pr(T_{uv}|M_s) + Pr(O_{uv}|F_{uv}, M_s)Pr(F_{uv}|M_s) \right]
\]

\[
= \prod_{(u,v) \in U \times U} \left[ \beta Pr(O_{uv}|T_{uv}) + (1 - \beta) Pr(O_{uv}|F_{uv}) \right]
\]
Null model probability

• Given values for $p_{uv}$: probability of edge $(u,v)$ in random graph with same degrees

$$Pr(O_U|M_n) = \prod_{(u,v)\in U\times U} [p_{uv}Pr(O_{uv}|T_{uv}) + (1 - p_{uv})Pr(O_{uv}|F_{uv})]$$

• How to get random graph if we don’t know true degree distribution? Estimate them:

$$d_i = \sum_j Pr(T_{ij}|O_{ij})$$

$$Pr(T_{uv}|O_{uv}) = \frac{Pr(O_{uv}|T_{uv})Pr(T_{uv})}{Pr(O_{uv}|T_{uv})Pr(T_{uv}) + Pr(O_{uv}|F_{uv})(1 - Pr(T_{uv}))}$$
Likelihood ratio

• Score subnetwork with (log) ratio of likelihoods under the two models

\[ L(U) = \log \frac{Pr(O_U|M_s)}{Pr(O_U|M_n)} \]

\[ = \sum_{(u,v) \in U \times U} \log \frac{\beta Pr(O_{uv}|T_{uv}) + (1 - \beta) Pr(O_{uv}|F_{uv})}{p_{uv} Pr(O_{uv}|T_{uv}) + (1 - p_{uv}) Pr(O_{uv}|F_{uv})} \]

• Note the decomposition into sum of scores for each edge
Seed construction

• Finding “heavy induced subgraphs” is NP-hard (Sharan et al, 2004)

• Heuristic:
  • Find high-scoring subgraph “seeds”
  • Grow seeds greedily

• Seed techniques: for each node $v$:
  • Find heavy subgraph of size 4 including $v$
  • Find highest-scoring length 4 path with $v$
Randomizing graphs

- For statistical tests, need to keep degree distribution the same
- Shuffle step:
  - Choose two edges \((a, b), (c, d)\) in the current graph
  - Remove those edges
  - Add edges \((a, d), (c, b)\)
Predictions from alignments

- Conserved modules of proteins enriched for certain functions often indicate shared function of other proteins

- Use to predict function of unannotated proteins

- Sharan et al., 2005: annotated 4,645 proteins with estimated accuracy of 58-63%

- Predict missing interactions

- Sharan et al., 2005: 2,609 predicted interactions in fly, 40–52% accurate
Parallels to sequence analysis

(Biological sequence comparison)

- 1960: First protein sequences by Sanger, others
- 1970: Dayhoff, Jukes/Cantor, Needleman/Wunsch
- 1980: PAM, BLOSUM, Smith/Waterman, Swiss-Prot, GenBank, EMBL-Bank, Stormo
- 1990: Haussler, Borodovsky, Church, Taylor, Lipman, others, BLAST

(A new type of data becomes routinely available)

- Mathematical models of evolution
- Scoring via transition probabilities
- Public genome-scale databases
- Mining for motifs and domains
- Hidden Markov models
- Database queries are staple of molecular biology

(Biological network comparison)

- 1990: Interaction detection with two-hybrid mass spec.
- 2001: Interologs: evolutionary models
- 2002: MaWish
- 2003: BIND, DIP, MINT, GRID
- 2004: Path BLAST
- 2005: Scale-free property; robustness
- 2010?: Sharan/Karp/Ideker

(Sharan & Ideker, 2006)