Using Probabilistic Clustering to help identify multiple protein instances in static PPI datasets

Colin Mclean

The University of Edinburgh, School of Informatics.
Overview

- Introduction:
  - SYNSYS data: www.synsys.eu/
  - Pre-synaptic architecture.
  - Active Zone/docking PPI networks.

- Clustering Algorithms:
  - Stochastic Poisson Block Model.

- Multiple Instance Proteins:
  - Degree-correlated-Bridgeness.
SYNSYS data

- SYNSYS consortium we’re interest in:
  - modeling PPI data of mammalian synapse.
  - identify structure/function of multi-protein complexes in pre-synapse.
  - improve predictions of disease related genes.

- Data:
  - IP/Pull-down dataset from mouse experiments.

- 93 Bait proteins covering the pre-synapse.
  - pulls out ~2000 proteins.
  - ~9000 significant PPI.
Distribution of Bait Proteins


Key events occurring at an excitatory synapse

Synaptic vesicles are formed and recycled at the presynaptic side; vesicles in this process are numbered in sequential order. Synaptic vesicles containing glutamate as a neurotransmitter (NT) are recruited to specialized release sites known as active zones. Upon stimulation, the vesicles exocytose to release NT, which diffuses across the synaptic cleft and binds to postsynaptic NT receptors. Activation of these receptors results in an influx of Ca²⁺, which triggers postsynaptic signal transduction cascades. Multispanning proteins that make up a region known as the post-synaptic density mediate clustering of receptors and cell-adhesion molecules, and orchestrate the coupling of diverse signaling components.
Active Zone - Baits

- MALS/CASK/Mint-1/liprin-a complex linking presynaptic machinery to Active zone
- Multiple Instances of proteins occur across Active zone: Docking, Priming, Vesicle fusion
Building Networks

Baits → Baits/Preys → Baits/Preys/db Int.
Active Zone Network

<table>
<thead>
<tr>
<th>Data / Databases</th>
<th>Proteins</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Zone network (1)</td>
<td>479</td>
<td>858</td>
</tr>
<tr>
<td>EUROSPIN -Y2H</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Intact &amp; Ensembl*</td>
<td></td>
<td>58</td>
</tr>
<tr>
<td>hmr-dbs**</td>
<td></td>
<td>394</td>
</tr>
<tr>
<td>HIPPIE</td>
<td></td>
<td>401</td>
</tr>
</tbody>
</table>

- **hmr-dbs includes: BioGrid, CCSB, Intact, MDC
- Network covers 27 (of 93) Baits
- 7% of 1st order interactions directly from pull-down data.
Docking subprocess

- Evidence of 13 (of 93) baits in Docking process.
- Use to build network of the docking process.
- AP2A2, BSN, CASK, PCLO, RAB3A, RIMSI SNAP25, STXBPI...
Docking Network

<table>
<thead>
<tr>
<th>Docking network (2)</th>
<th>Proteins</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>164</td>
<td>563</td>
</tr>
</tbody>
</table>

- Proteins based on manually curate annotation list for docking (EUROSPIN).
- Interactions based on same databases: hmr-dbs, HIPPIE.
Overview

- Introduction
- Clustering Algorithms
- Gene Multi-Instance
Clustering - Probabilistic

- Node belonging to one community: In densely connected networks, subnetworks may often overlap to such an extent that this becomes too restrictive.

- Assigning probabilities to edges and nodes, allowing nodes to belong to more than one community.

\[
P(A|B) = \frac{P(B|A) \ P(A)}{P(B)}
\]

P. K. Gopalan & D. M. Blei, PNAS (2013)

Clustering - Probabilistic

• Goal to determine the value of edge parameters, $\theta_{iz}\theta_{jz}$, that best fits the network and from these determine the overlapping communities.

• Assign probability $\theta_{iz}$ to i’th node belonging to the z’th community.

$\theta_{i1}\theta_{j1} = 0.75$
$\theta_{i1}\theta_{j2} = 0.05$
$\theta_{i2}\theta_{j1} = 0.05$
$\theta_{i2}\theta_{j2} = 0.1$
Stochastic Block Model


- The model:
  - Generates networks of nodes (N) and edges (M) into a given no: (K) of communities.
  - Parameterised by \( \theta_{iz} \)'s. Two nodes, i and j, with high theta values, have a high probability being connected by edge in community z.

- The probability of generating a network with given community structure modeled by Poisson distributions.

\[
P(G|\Theta) = \prod_{i<j} \frac{(\Theta)^{A_{ij}} \exp(-\Theta)}{A_{ij}!}
\]

\[
\Theta = \sum_{z} \theta_{iz} \theta_{jz}
\]

Graph treated as a multigraph, i.e. multiedges. In practice \( \theta_{iz} \)'s small, so expected no: of edges small (hence error induced small)
Stochastic Block Model

- Fit model to observed network by maximising \( P(G|\Theta) \) with respect to \( \theta_{iz} \)

\[
\log(P(G|\Theta)) \geq \sum_{ijz} [A_{ij}q_{ij}(z)\log \frac{\theta_{iz}\theta_{jz}}{q_{ij}(z)} - \theta_{iz}\theta_{jz}]
\]

Jensen's inequality

\[
\sum_{z} q_{ij}(z) = 1
\]

Expectation: \( q_{ij}(z) = \frac{\theta_{iz}\theta_{jz}}{\sum_{z} \theta_{iz}\theta_{jz}} \)

\( \Delta \log(P(G|\Theta)) < 10^{-5} \)

Maximisation: \( \theta_{iz} = \frac{\sum_{j} A_{ij}q_{ij}(z)}{\sqrt{\sum_{ij} A_{ij}q_{ij}(z)}} \)

\( \theta_{iz} \) Edges Probs.

\( \sum_{z} \theta_{iz} \) Node Probs.
Partition Density function - $D_k$

- Model requires no: communities ($K$) as an input.
- Investigated using Partition Density function ($D_k$) to find optimal $K$ for the network.

$$m_K = \sum_{(i,j)=1}^{M} \theta_{iK} \theta_{jK}$$

$$D_K = \frac{m_K - (n_K - 1)}{n_K(n_K-1)/2 - (n_K-1)}$$

$$n_K = \sum_{i=1}^{N} \theta_{iK}$$


Active Zone (27 Baits) $N=479$, $E=858$

$K$ Min : 24
Predicting community


- Bayesian approach:
  - posterior, conditional, distribution over hidden variables:

  \[
P(\theta, z | G) = \frac{p(\theta, z, G)}{p(G)}
\]

- approx. posterior dist. using Mean-Field Variational Inference, in combination with Stochastic Optimisation.

- K = 35 for Docking network.

Performance of stochastic inference (SVI) and Poisson (POI) community models, for synthetic networks of 10k, 100k, 1M nodes.

©2013 by National Academy of Sciences

Institute for Adaptive and Neural Computation
Solution Space

- Run Poisson Model 5000 times looking at distribution of solution space for different K
Overview

• Introduction  • Clustering Algorithms  • Gene Multi-Instance
Bridgeness


- Bridgeness:
  - extent to which a vertices is shared among different K:

\[ b_i = 1 - \sqrt{\frac{K}{K-1} \sum_{j=1}^{K} (u_{ji} - \frac{1}{K})^2} \]

- \( b_i [0,1] \):
  - 0 implies vertices belong to one community.
  - 1 implies vertices belong to all communities with the same strength, i.e. outline in graph (belongs to none of the communities).

- Distinguish outliners and ‘real’ bridges by also using centrality of vertices:
  - Degree, Betweenness.
  - High centrality too supports assumption vertex is a bridge.
Degree-correlated-Bridgenessness

- Test if the Bridge vertices are Multiple Instance Protein (MIP) candidates.
  - e.g. CASK, RIMS1, RAB’s

Candidate proteins:
- Degree > 10 &&
  Bridgeness > 0.1
MIP Candidates (1) - K35

Tuesday, 12 November 13
Docking network (2)

<table>
<thead>
<tr>
<th>Docking network (2)</th>
<th>Proteins</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>164</td>
<td>563</td>
</tr>
</tbody>
</table>

- $K=22$ for Poisson approach.
- $K=19$ for Bayesian approach:

P. K. Gopalan & D. M. Blei, PNAS (2013)
Summary

- Built presynaptic PPI networks based on experimental data.
- Applied Block Probabilistic Clustering algorithms to networks.
- Degree-correlated-Bridgeness to identify multiple instance proteins.
Reference Material

- Papers:
  - Molecular Profiling of Synaptic Vesicle Docking Sites Reveals Novel Proteins but few Differences between Glutamatergic and GABAergic Synapses, Neuron Article, (2013).
Active Zone Annotations

- Exp. data as ‘guide’ to identify proteins (Baits) involved in the docking process.
  J. Boyken et. al., Neuron Article (2013).

- 493 proteins identified and assigned ratio:
  \[
  \frac{\text{iTRAQ frac. of protein identified as SV Docking}}{\text{iTRAQ frac. of protein identified as free}}
  \]

- Annotation list, to predict annotation of unknown proteins in network.
Partition Density function - \( D_k \)

- Fit Function \( f(K) = \exp + \text{poly}(1) \)

**Computer Generated**

- \( N = 100, E = 272, K = 14 \)

**Y2H dataset**

- \( N = 656, E = 2311, K = 48 \) (Geodesic Edge Betweenness)

**MASC Complex**

- \( K = 13 \)  
  - Min: 12  
  - TP: 12

- \( K = 24 \)  
  - Min: 23  
  - TP: 23 [16-28]

- \( K = 48 \)  
  - Min: 29  
  - TP: 39 [28-48]
Multi-Protein Instances

- Know genes play key role in multiple processes. Interest using probabilistic clustering to find genes in overlapping communities in our data:
Probabilistic

- Interest in finding gene, X, that has 50/50 split between two communities (C1 & C3):

```
<table>
<thead>
<tr>
<th></th>
<th>g1</th>
<th>g2</th>
<th>g3</th>
<th>g4</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Edges

- C1
- C2
- C3
- C4

- X → g1
- Xa → g1
- Xb → g4
- g2 → C2
- g3 → C3
- g4 → C4

Tuesday, 12 November 13
MIP Candidates (2) - K22

Bridging proteins for SYNSYS Docking Network

Tuesday, 12 November 13