COMP598: Advanced Computational Biology Methods & Research

Protein-Protein Interaction Networks

Jérôme Waldispühl
School of Computer Science, McGill
(includes slides from M.Lavallée-Adam & B. Berger)
What is system biology?

“The science of integrating genetic, genomic, biochemical, cellular, physiological and clinical data to create a system network that can be used to predictively model a biological event(s).”

Morel, N.M., et al. Mayo Clin Proc, 2004. 79(5): p. 651-8.
Reductionism vs. Systems Biology
In the 20th century...

Norbert Wiener (1948)
Cybernetics
“The science of communications and automatic control systems in both machine and living things.”

Hudgkin & Huxley (1952)
Math model explaining the action potential propagating along the axon of a neuronal cell.

Denis Noble (1960)
Math model of cardiac cells.

Mihajlo Mesarovic (1968)
Organized the 1st “System theory and Biology” symposium. Launch of a new scientific discipline!

“-omics revolution”!

(1990s)

Ludwig von Bertalanffy (1928)
General Systems Theory
“general science of wholeness”
"-omics revolution"

Genomics → Proteomics → Metabolomics

Transcriptomics ↔ Functional proteomics/genomics

SYSTEMS BIOLOGY

Morel, N.M., et al. Mayo Clin Proc, 2004. 79(5): p. 651-8.
Back to the car analogy

How would use a system approach to understand how a car functions?

1. Preliminary understanding -> formulate a simple model
2. Define all the components: mechanical, electrical, and control.
3. Perturb the car and compare to normal car
4. Integrate data and compare to your model
5. Discrepancies? ->new hypothesis -> repeat step 3-5.
A test system: galactose utilization in *S. cerevisiae*

- **9 elements:**
  - **4 enzymes** catalyze conversion of galactose (*gal*) to glucose-6-P
  - **1 transporter** molecule
    - Sets the state of the system
  - **4 transcription factors (TFs)**
    - Turn system on/off depending on galactose presence/absence

---

*Image of yeast cells*

---

Ideker, T., et al., *Science*, 2001. **292**(5518): p. 929-34.
Perturb the system and compare

- Yeast strain used:
  - 9 knock-out (KO)
  - 1 wild-type (WT)

DNA Microarray

GAL1
GAL2
GAL3
GAL4
GAL5
GAL6
GAL7
GAL10
GAL80

- gal Δ vs. wt
- gal Δ vs. wt

OBSERVED CHANGE IN mRNA LEVEL (log_{10} ratio)

Ideker, T., et al., Science, 2001. 292(5518): p. 929-34.
Decontaminator

Modeling contaminants in AP-MS/MS experiments

(Lavallee-Adam et al., J. Proteome Research, 2010)
Protein interactions obtained by Tandem Affinity Purification
Background

Cell Culture

2D-LC

Database search

SDS-PAGE

TAP

First affinity step

Bead

Calmodulin

Second affinity step

Bead

IgG

Protein A

Protease cleavage

MS/MS

Target

CBP
Cell Culture

2D-LC

SDS-PAGE

TAP

Mistaken Purification

Non-specificity of TAG antibody

False positive sources

Gel contamination

Over-expression

Faulty Purification

Misidentification

Carry-over
Experimental Improvements

In-cell normal expression → Cell Culture

Additional purifications → TAP

Robot gel band cutting LC column washing → SDS-PAGE

SDS-PAGE 2D-LC

MS/MS Database search

False positive sources

Over-expression → Non-specificity of TAG antibody → Faulty Purification → Gel Contamination → Carry-over

Misidentification
Computational Filtering

Cell Culture

TAP

SDS-PAGE

2D-LC

MS/MS

Database search
Computational Filtering

Cell Culture

TAP

SDS-PAGE

2D-LC

DeContaminator
(Lavallee-Adam et al., JPR)

Peptide/Protein
Prophet (Keller et al., Nesvizhskii et al.)
Percolator (Kall et al.)

Database search

Krogan et al., 2006
Chua et al., 2006
Ewing et al., 2007
Cloutier et al., 2009
Collins et al., 2007
Simple contaminant detection

Manually label all contaminants and systematically remove all interactions with these proteins.

Limitations:

- A contaminant for one bait might be a true interaction for another.
- Could not detect sporadic contamination.
Alternate contaminant detection method

Two experiments for a given bait $b$
- Induced experiment: expression of the bait vector is induced.
- Control experiment: expression of the bait vector is not induced.

$MRatio$ method:

For a prey $p$:
    If $MS\_Score(b_{\text{induced}}, p) < 5 * MS\_Score(b_{\text{control}}, p)$
        $p$ is a contaminant
    Else
        $p$ is truly interacting with $b$

[Jeronimo et al., 2007]

Limitations:
- Expensive both in terms of time and resources
- One-to-one comparisons of noisy low abundance prey MS/MS results
- Control might show leaky expression of the bait
DeContaminator
(Lavallee-Adam et al., Journal of Proteome Research)

• **Goal:** Use a limited number of controls for the proper identification of contaminants in TAP-MS/MS PPI data.

• **Advantages:**
  • No one-to-one comparisons of MS scores have to be performed.
  
  • Accurate modeling with limited resource usage.
  
  • Using a limited number of high-quality controls avoids expression leakiness issues.
Limited number of matched control/induced experiments

Control
  \( b_1 \)
  \( b_2 \)
  \( \ldots \)
  \( b_{14} \)

Induced
  \( b_1 \)
  \( b_2 \)
  \( \ldots \)
  \( b_{14} \)

Control experiments noise model

Induced experiments noise model

Mascot score for prey \( p \) upon induction of bait \( b \)

P-value calculation

\[ p\text{-value}(M_{b,p}^f) = \Pr[M_p^N \geq M_{b,p}^f | M_{b,p}^N, M_{b,p}^I, M_{b,p}^{N_1}, \ldots, M_{b,p}^{N_{14}}] \]

False Discovery Rate (FDR) calculation

\[ \text{FDR}(b,p) \sim \text{Confidence level that } p \text{ is a contaminant for bait } b \]

(Lavallee-Adam et al., JPR)
**Modeling Contaminants**

**Objective:** Compute the posterior distribution of $MCM_p$ given all MS score observations $CM_{b,p}$ $\forall b \in B$

![Diagram showing the relationship between control and induced MS score mean distributions and the calculation of p-value](image)

\[ pvalue(IM_{b,p}) = Pr[MCM_p \geq MIM_{b,p} | IM_{b,p}, CM_{b1,p}, CM_{b2,p}, \ldots, CM_{b14,p}] \]
Modeling Contaminants

Non-induced results from the set of baits $B$ are pooled.

Using a weighted $k$-nearest neighbours smoothing of the frequency of each $MCM_p$ value, conditional on $CM_{b,p}$ values $\forall b \in B$, we obtain an estimate of:

$$\Pr[MCM_p = mcm|CM_{b,p} = cm]$$

After Bayes rule:

$$\Pr[CM_{b,p} = cm|MCM_p = mcm]$$

The posterior distribution of $MCM_p$ scores is then:

$$\Pr[MCM_p|CM_{b_1,p}, \ldots, CM_{b_{14},p}] = \Pr[MCM_p] \prod_{i=1}^{14} \Pr[CM_{b_i,p}|MCM_p]/\zeta$$

The posterior distribution of $MIM_{b,p}$ is computed in a similar fashion:

$$\Pr[MIM_{b,p}|IM_{b,p}]$$
Contaminants Assessment

p-value that prey \( p \) is a contaminant for bait \( b \)

\[
p_{value}(IM_{b,p}) = \Pr[MCM_p \geq MIM_{b,p}|IM_{b,p}, CM_{b_1,p}, CM_{b_2,p}, \ldots, CM_{b_{14},p}]\
\]

False Discovery Rate (FDR) for an interaction with a given \( p \)-value:

\[
FDR(p-value) = \frac{\sum_{b \in B} |\{np \in NP_b| np \leq p\text{-value}\}|}{|NP_b|} \\
\frac{\sum_{b \in B} |\{ip \in IP_b| ip \leq p\text{-value}\}|}{|IP_b|}\
\]

\( NP_b \) and \( IP_b \) are the sets of non-induced and induced interaction \( p \)-values.
Protein-Protein Interaction Network

- 89 baits and 11894 interactions [Fortier, Lacombe et al., 2010]
- Human cell line: HEK293
- Proteins in the network are mainly involved in transcription and RNA processing.
- 14 representative baits out of the 89 have been selected for control experiments.
False Discovery Rates

Number of predicted interactions

FDR 1%: DeContaminator: 2430 interactions
Z-score approach: 1011 interactions

FDR

Z-score
DeContaminator
IsoRank

Comparison of PPI networks
Comparative Genomics

Look at the same kind of data across species with the hope that areas of high correlation correspond to functional parts or modules of the genome.
Why understanding function-level differences is important

- Increased complexity (function) is not explained simply by variations in gene (or protein) count

Estimated Number of Genes:
- 6600
- 21000
- 14000
- 24500
- 23000

Estimated Number of Proteins:
- 6600
- 27000
- 19000
- 32000
- 49000

Numbers from http://www.ensembl.org
Protein-Protein Interactions (PPIs)

- Often, proteins interact with other proteins to perform their functions
- Many cellular activities are a result of protein interactions

MAPK Signaling Cascade

Image from:
http://focosi.altervista.org/mapkmap2.html
Modeling PPIs

- Traditional perspective: low-throughput, structural
- New perspective: high-throughput, network-based
Protein-Protein Interaction (PPI) Network

Yeast 2-Hybrid method

X + Y = ?

Cusick et al. Hum Med Gen, 05

http://internal.binf.ku.dk
Motivation behind Network Comparison

• Compare PPI networks at the species level

• Transfer annotation from one species to another
  – More feasible, cheaper and easier than in humans
  – Error detection

• Compute functional orthologs
  – Functional orthologs: proteins which perform the same function across species
The Problem

Given two protein-protein interaction networks, find for a piece of one network, something that has a comparative structure in the other network

Our approach: match neighborhood topologies
Algorithm: IsoRank

Sequence similarity

|     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|
| a5  | b7  | 2.1 |
| a5  | b9  | 1.5 |
| a3  | b2  | 3.4 |

Functional similarity for each possible node pairing

|     |     |     |
|-----|-----|-----|
| a5  | b7  | 1e-2 |
| a5  | b1  | 2e-8 |
| a5  | b3  | 1e-7 |
| a5  | b9  | 1e-4 |
| a3  | b1  | 5e-4 |
| a3  | b6  | 3e-9 |
|     |     | ... |
Functional Similarity Score: Intuition

• Compute pairwise scores $R_{ij}$:

Goal: “high $R_{ij}$” $\Rightarrow$ “i and j are a good match”

Intuition: i and j are a good match if their sequences align and their neighbors are a good match

$R_{a5,b1} = ?$
Computing $R_{ij}$

- Combine both sequence and network data

$$R_{ij} = (1 - \alpha)E_{ij} + \alpha N_{ij}$$

- Functional similarity
- Sequence similarity
- Network similarity
Simple Case: $\alpha=1$ (no $E_{ij}$)

- $R_{ij}=N_{ij}$. $R_{ij}$ depends on neighborhoods of $i$ and $j$

$$R_{ij} = N_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{|N(u)||N(v)|} R_{uv}$$

- $N(a)$ is the set of neighbors of $a$

$$R_{a1,b4} = \frac{1}{2 \times 3} R_{a2,b3}$$
Simple case: $\alpha=1$ (no $E_{ij}$)

- $R_{ij}=N_{ij}$. $R_{ij}$ depends on neighborhoods of $i$ and $j$

$$R_{ij} = N_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{\|N(u)\| \|N(v)\|} R_{uv}$$

- $N(a)$ is the set of neighbors of $a$

$$R_{a_2, b_2} = \frac{1}{1 \times 1} R_{a_1, b_1} + \frac{1}{1 \times 3} R_{a_1, b_3} + \frac{1}{3 \times 1} R_{a_3, b_1} + \frac{1}{3 \times 3} R_{a_3, b_3}$$
Example: Computed $R_{ij}$ values

|     | b1   | b2   | b3   | b4   | b5   |
|-----|------|------|------|------|------|
| a1  | 0.0312 |      | 0.0937 |      |      |
| a2  |      | 0.1250 |      | 0.0625 | 0.0625 |
| a3  | 0.0937 |      | 0.2813 |      |      |
| a4  |      | 0.0625 |      | 0.0312 | 0.0312 |
| a5  |      | 0.0625 |      | 0.0312 | 0.0312 |

Empty cell indicates $R_{ij} = 0$
Example: Computed $R_{ij}$ values

$$
\begin{array}{ccccc}
 & b1 & b2 & b3 & b4 & b5 \\
\hline
a1 & 0.0312 & & 0.0937 & & \\
a2 & & 0.1250 & & 0.0625 & 0.0625 \\
a3 & 0.0937 & & 0.2813 & & \\
a4 & & 0.0625 & & 0.0312 & 0.0312 \\
a5 & & 0.0625 & & 0.0312 & 0.0312 \\
\end{array}
$$

Empty cell indicates $R_{ij} = 0$
Example: Computed $R_{ij}$ values

$$R$$

|     | b1    | b2    | b3    | b4    | b5    |
|-----|-------|-------|-------|-------|-------|
| a1  | 0.0312|       | 0.0937|       |       |
| a2  |       | 0.1250|       | 0.0625| 0.0625|
| a3  | 0.0937|       | 0.2813|       |       |
| a4  | 0.0625|       |       | 0.0312| 0.0312|
| a5  | 0.0625|       |       | 0.0312| 0.0312|

Empty cell indicates $R_{ij} = 0$
Capturing non-local effects?

- The algorithm can resolve between \( p-r \) vs. \( p-q \)

\[
R_{pr} = 8.12 \times 10^{-3} \\
R_{pq} = 8.64 \times 10^{-3}
\]
Computing $R$: an eigenvalue problem

- The equations for $R$ describe an eigenvalue problem

\[
R = AR
\]

\[
A[ij][uv] = \frac{1}{N(u)|N(v)|}
\]

\[
\text{size}(A) = N_1 N_2 \times N_1 N_2
\]

$R$ is the principal eigenvector of $A$

- $A$ is about $10^8 \times 10^8$ when aligning yeast and fly networks
  - However, both $A$ and $R$ are very sparse
  - We use the Power method to efficiently compute $R$

- Extension to weighted edges is straightforward

$N_1 = \# \text{ nodes in Graph 1}$

$N_2 = \# \text{ nodes in Graph 2}$
General Case: $0 \leq \alpha \leq 1$

- Let $B_{ij}$ = sequence similarity score between $i$ (from graph #1) and $j$ from (graph #2)
- $E_{ij} = B_{ij} / |B|_1$

\[ R = (1 - \alpha)E + \alpha AR \]

$0 \leq \alpha \leq 1$
Results: Yeast-Fly Global Alignment

- # of edges in the common subgraph: 1420
  - Implies about 5% overlap! Why so low?
  - PPI data currently is noisy and low-coverage

- # of edges in the largest component: 35

- The value of $\alpha$ used: 0.6
  - Provided best overall agreement with previous gene correspondence predictions
Various Topologies Are Found

Existing local alignment methods (PathBlast; Kelley et al.) often find only specific topologies.
Role of $\alpha$: why the dip?
Robustness to Error in PPI data
Robustness to Error in PPI data

![Graph showing the fraction of nodes mapped correctly against the percent of randomized edges. The graph has two curves, one for $\alpha=1$ and another for $\alpha=10^6$. The red curve indicates a higher robustness compared to the blue curve. The true curve is somewhere around the red curve.](image)
Functional Orthologs
• Genes that perform similar functions
  – “functional orthologs” vs “plain old orthologs”
  – distinguish between orthologs and paralogs
• Bandyopadhyay et al. [Genome Res. ’06]
  – Use local network alignment results
  – Then use a MRF to partially resolve ambiguities
• We compared our results with theirs
## Functional Orthologs: IsoRank Pairwise Alignment Predictions

| Protein   | IsoRank | Bandyopadhyay et al. |
|-----------|---------|----------------------|
| Gid8      | CG6617  | CG6617               |
|           |         | CG18467              |
| Gpa1      | Goα47a  | Goα47a               |
|           |         | Giα65a               |
| Kap104    | Trn     | Trn                  |
|           |         | CG8219               |
| CG18617   | Vph1    | Vph1                 |
|           |         | Stv1                 |
| Egd1      | Bic     | Bcd                  |