Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms

GIW2016
Joint work with Limsoon Wong

Wilson Wen Bin Goh
Some background

The traditional network utilisations
- DNA
- Perturbation
- Validation
- Correlating phenotype to network (static projection)

The new network utilisations
- RNA
- Protein
- DNA
- RNA
- Protein

Complexes work much better than predicted clusters from reference networks

The problem

- No formalization of the classes of methods for complex-based analysis
- A comprehensive means of evaluation/benchmarking is not available
Network-Paired approach

ESSNet

- Let $g_i$ be a protein in a given protein complex
- Let $p_j$ be a patient
- Let $q_k$ be a normal
- Let $\Delta_{i,j,k} = \text{Expr}(g_i, p_j) - \text{Expr}(g_i, q_k)$
- Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

- Newest addition to complex-based methods
- Null hypothesis is “Complex C is irrelevant to the difference between patients and normals, and the proteins in C behave similarly in patients and normals”
- No need to restrict to most abundant proteins
  $\Rightarrow$ Potential to reliably detect low-abundance but differential proteins

Lim et al. *A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small.* *JBCB*, 13(4):1550018, 2015
Five methods to compare with

- Network-based methods
  - Over-Representation Analysis (Hypergeometric enrichment, HE)
  - Direct group (GSEA)
  - Hit-Rate (qPSP) \cite{goh2015biology}
  - Rank-Based Network Analysis (PFSNET), \cite{goh2016jbc}

- Standard t-test on individual proteins (SP)
Simulated data

- Simulated datasets from Langley and Mayr
  - D.1.2 is from study of proteomic changes resulting from addition of exogenous matrix metallopeptidase (3 control, 3 test)
  - D2.2 is from a study of hibernating arctic squirrels (4 control, 4 test)

- Both D1.2 and D2.2 have 100 simulated datasets, each with 20% significant features
  - Effect sizes of these differential features are sampled from one out of five possibilities (20%, 50%, 80%, 100% and 200%), increased in one class and not in the other

- Significant artificial complexes are constructed with various level of purity (i.e. proportion of significant proteins in the complex)
  - Equal # of non-significant complexes are constructed as well
**Precision, Recall and the F-score**

Elements = features

relevant elements

- **false negatives**
- **true negatives**
- **true positives**
- **false positives**

selected elements

- **How many relevant items are relevant?**
- **How many relevant items are selected?**

**Precision**

Of the selected feature, how many are correct?

**Recall**

Of the selected feature, what is the proportion of all the correct ones we got?

Precision and recall can be combined as:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$
SP shows poor performance on simulated data. Can network-based methods do better?

Supplementary Figure 1 Single protein (SP) precision-recall performance on D1.2. The f-score (pink), precision (blue) and recall (green) shows that SP performs abysmally on simulated data. HE is shown next to SP as a reference.
ESSNET shows excellent recall/precision on simulated data
Renal cancer control data (RCC)

• 12 runs originating from a human kidney tissue digested in quadruplicates and analyzed in triplicates

• Excellent for evaluating false-positive rates of feature-selection methods
  – Randomly split the 12 runs into two groups. Report of any significant features between the groups must be false positives

All methods control false positives well.

Dash line corresponds to expected # of false positives at alpha 0.05 (~30 complexes)
Renal cancer data (RC)

- 12 samples are run twice so that we have technical replicates over 6 normal and 6 cancer tissues

- Excellent opportunity for testing reproducibility of feature-selection methods
  - A good method should report similar feature sets between replicates

- Can also test feature-selection stability
  - Apply feature-selection method on subsamples and see whether the same features get selected
ESSNET & PFSNET show excellent cross-replicate reproducibility

This table is computed on by applying the methods on the full RC dataset.
Feature-selection stability

THE BINARY MATRIX is USEFUL FOR COMPARING STABILITY AND CONSISTENCY OF SIGNIFICANT FEATURES PRODUCED BY SOME FEATURE-SELECTION METHOD

THE ROWS REPRESENT EACH SIMULATION
THE COLUMNS ARE A NOMINAL FEATURE VECTOR. RED REPRESENTS FEATURES REPORTED AS SIGNIFICANT WHILE PINK ARE NON-SIGNIFICANT.
THE ROW SUMS PROVIDES INFORMATION ON THE NUMBER OF SIGNIFICANT FEATURES WHILE THE COLUMN SUMS PROVIDE INFORMATION ON THE RELATIVE STABILITY OF EACH FEATURE (I.E., OUT OF N SIMULATIONS, HOW MANY TIMES IS THE FEATURE REPORTED AS SIGNIFICANT)

ESSNET & PFSNET show excellent feature-selection stability
ESSNET & PFSNET show excellent stability

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE</td>
<td>0.022</td>
<td>0.016</td>
<td>0.047</td>
<td>0.030</td>
</tr>
<tr>
<td>DG</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>ESSNET</td>
<td>0.714</td>
<td>0.941</td>
<td>1.000</td>
<td>0.885</td>
</tr>
<tr>
<td>QPSP</td>
<td>0.149</td>
<td>0.282</td>
<td>0.991</td>
<td>0.470</td>
</tr>
<tr>
<td>PFSNET</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
ESSNET can assay low-abundance complexes that qPSP cannot

A: QPSP-ESSNET significant-complex overlaps

B: P-value distribution for overlapping and non-overlapping QPSP complexes.

C: Sampling abundance distribution. The left panel is a zoom-in of the right. The y-axis is the protein abundance while the four categories are the distribution of abundances of complexes found in QPSP, ESSNET, ESSNET unique (complement), and all proteins in RC.
ESSNET can assay low-abundance complexes that PFSNET cannot

Of the 5 ESSNET-unique complexes, PFSNET can detect 4; the missed complex consists entirely of low-abundance proteins.

If p-value threshold is adjusted by Benjamini-Hochberg 5% FDR, PFSNET can detect only 3 of the 5 ESSNET-unique complexes while ESSNET continues to detect them all.
What have we learnt?

• We’ve seen how five statistical methods can be used in conjunction with complex-based analysis
• ESSNET, adapted for proteomics is a powerful approach that can sensitively detect low-abundance complexes
References

- Goh & Wong. *Integrating networks and proteomics: Moving forward*. Trends in Biotechnology, in press
Acknowledgements

Professor Limsoon Wong
National University of Singapore