Advancing clinical proteomics using protein complexes as a contextualization framework

Limsoon Wong

(Based on joint work with Wilson Goh, Kevin Lim)
Challenges in proteomic profile analysis

- Poor reproducibility of measurements
- Sparse # of features measured
- Uncertainty in mapping peptides to proteins
- Small sample size

⇒ Much more challenging than gene expression profile analysis
Some exciting ideas in gene expression profile analysis can be useful in improving proteomic profile analysis...

A DETOUR TO GENE EXPRESSION ANALYSIS
Why small sample size?

• **Biological constraint**
  – Comparing cell lines
  – Comparing mutants vs wildtype

• **Rare-sample constraint**

• **Population-size constraint**
  – Singapore is small, we often wait a long time for enough patients presenting the desired phenotype

• **Cost & technological constraints**
Outline

• Ideals of a perfect method for gene selection in gene expression profile analysis

• Failure of commonly-used methods

• Reproducible precise & sensitive selection of genes, even when sample size is extremely small

• Reliable accurate cross-batch classification, even when batch effect is severe and sample size is small
THE IDEAL
A perfect method for identifying causal factors of a disease

• A perfect method should …
  – Completeness: Report all causal factors in a dataset
  – Soundness: Not report any non-factor

⇒ When applied to two representative datasets of the disease, the two sets of identified factors should be the same
⇒ Factors identified from a subset of a dataset should be subset of factors identified from the whole dataset
⇒ Factors identified from one dataset should do well when used for classifying new datasets
THE REALITY
Percentage of overlapping genes

- Low % of overlapping genes from diff expt in general
  - Prostate cancer
    - Lapointe et al, 2004
    - Singh et al, 2002
  - Lung cancer
    - Garber et al, 2001
    - Bhattacharjee et al, 2001
  - DMD
    - Haslett et al, 2002
    - Pescatori et al, 2007

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEG</th>
<th>POG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prostate Cancer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td><strong>Lung Cancer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td><strong>DMD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

“Most random gene expression signatures are significantly associated with breast cancer outcome”

Gene regulatory circuits

- Each disease has some underlying cause

- There is some unifying biological theme for genes that are truly associated with a disease
Overlap analysis: ORA

ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (we use all genes whose t-statistic meets the 5% significance threshold), and checking the significance of the size of the intersection using the hypergeometric test.

Disappointing performance

- DMD gene expression data
  - Pescatori et al., 2007
  - Haslett et al., 2002

- Pathway data
  - PathwayAPI, Soh et al., 2010
THE REASONS
Issue #1 with ORA

- Its null hypothesis basically says “Genes in the given pathway behaves no differently from randomly chosen gene sets of the same size”

- This null hypothesis is obviously false
  \[ \implies \text{Lots of false positives} \]

- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behaviour of genes in a pathway is more coordinated than random ones
Issue #2 with ORA

- It relies on a pre-determined list of DE genes
- This list is sensitive to the test statistic used and to the significance threshold used
- This list is unstable regardless of the threshold used when sample size is small
Issue #3 with ORA

- It tests whether the entire pathway is significantly differentially expressed

- If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch
GSEA in gene-permutation mode

- **Issue #2 & #3 solved to different degrees**
  - Does not need pre-determined list of DE genes, but gene ranking (based on t-test p-value) is still unstable for small sample size
  - Irrelevant genes in pathway have only small effect on the ES(S) peak

- **Issues #1 (when sample size is small) is unsolved**
Better performance, but not great

![Graph showing subnetwork agreement vs sample size (N) for GSEA and ORA]

- **GSEA**
- **ORA**
PFSNet: Exploiting subnetworks

- Induce subnetworks from pathways by considering only genes highly expressed in majority of patients in any class
  
  $\beta_1^*(g_i) = \sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|}$
  
  $\beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{fs(e_{g_i,p_j})}{|\neg D|}$

  $Score_{1}^{pk}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) \times \beta_1^*(g_i)$
  
  $Score_{2}^{pk}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) \times \beta_2^*(g_i)$

- For an irrelevant subnetwork $S$, the two scores above for each patient $P_k$ should be roughly equal, regardless of class
  - Interestingly, expression of the *same gene* is not compared between patients!

- Do a paired t-test to decide whether $S$ is relevant
  - Get null distribution by permuting class labels

- All 3 issues solved, but not when sample size is small
  - $\beta$ weights become unstable
  - Cannot generate null distribution
Much better performance but still not great
THE QUANTUM LEAP
EVEN WHEN SAMPLE SIZE IS EXTREMELY SMALL
ORA-Paired: Paired test and new null hypothesis

- Let $g_i$ be genes in a given pathway $P$
- 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

- **Issue #1 is solved**
  - Null hypothesis is “Pathway $P$ is irrelevant to the difference between patients and normals, and the genes in $P$ behave similarly in patients and normals”

- **Issue #2 is solved**
  - No longer need a pre-determined list of DE genes

- **Issue #3 is unsolved**

- **Is sample size now larger?**
  - $|\text{patients}| \times |\text{normals}| \times |\text{genes in P}|$
Testing the null hypothesis

“Pathway P is irrelevant to the difference between patients and normals and so, the genes in P behave similarly in patients and normals”

• **Method #1**
  – T-test w/ a conservative degree of freedom
    • E.g., # normals + # patients

• **Method #2**
  – By the null hypothesis, a dataset and any of its class-label permutations are exchangeable
  ⇒ Get null distribution by class-label permutations
    • Only for large-size sample

• **Method #3**
  – Modified null hypothesis
    • “Pathway P induces gene-gene correlations, and genes in P behave according to these gene-gene correlations;
    • P is irrelevant to the diff betw patients and normals and so, genes in P behave similarly in patients and normals”
  ⇒ Get null distribution using datasets that conserve gene-gene correlations in the original dataset
    • E.g., array rotation
Similar to PFSNet, good but not great.
NEA-Paired: Paired test on subnetworks

- Given a pathway $P$
- Let each node and its immediate neighbourhood in $P$ be a subnetwork
- Apply ORA-Paired on each subnetwork individually

- Issues #1 & #2 are solved as per ORA-Paired

- Issue #3 is partly solved
  - Testing subnetworks instead of whole pathways
  - But subnetworks derived in a simple-minded way
Much better performance

Upregulated in DMD

Sample size (N)

Subnetwork agreement

NEA-Paired
ORA-Paired
PFSNet
GSEA
ORA
ESSNet: Larger subnetworks

- Compute the average rank of a gene based on its expression level in patients in any class
- Use the top $\alpha\%$ to extract large connected components in pathways
- Test each component using ORA-Paired

- Gene rank is very stable
- Issues #1 - #3 solved
Fantastic performance

upregulated in DMD

- ESSNet
- NEA-Paired
- ORA-Paired
- PFSNet
- GSEA
- ORA
More datasets tested
ESSNet is unlikely to report junk

Table 4.2: Average number of subnetworks predicted by ESSNet over the sample sizes (N); the first number denotes the number of subnetworks in the numerator of the subnetwork-level agreement and the second number denotes the number of subnetworks in the denominator of the subnetwork-level agreement; cf. equation 4.5.

<table>
<thead>
<tr>
<th>sample size (N)</th>
<th>DMD</th>
<th>ALL</th>
<th>BCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8.2/13.4</td>
<td>7.0/11.9</td>
<td>4.8/12.6</td>
</tr>
<tr>
<td>3</td>
<td>11.1/15.9</td>
<td>11.3/17.9</td>
<td>5.0/11.7</td>
</tr>
<tr>
<td>4</td>
<td>13.18/16.5</td>
<td>11.9/15.9</td>
<td>6.2/10.4</td>
</tr>
<tr>
<td>5</td>
<td>14.2/16.7</td>
<td>14.6/18.3</td>
<td>7.9/12.7</td>
</tr>
<tr>
<td>6</td>
<td>15.14/17.6</td>
<td>14.9/18.0</td>
<td>11.0/15.7</td>
</tr>
<tr>
<td>7</td>
<td>15.2/17.4</td>
<td>16.1/19.2</td>
<td>12.9/17.5</td>
</tr>
<tr>
<td>8</td>
<td>15.4/17.5</td>
<td>16.2/19.0</td>
<td>15.3/20.4</td>
</tr>
<tr>
<td>9</td>
<td>16.6/18.8</td>
<td>17.0/19.8</td>
<td>15.8/20.8</td>
</tr>
<tr>
<td>10</td>
<td>17.6/19.7</td>
<td>17.3/19.7</td>
<td>16.2/20.8</td>
</tr>
</tbody>
</table>
A negative-control experiment showing that ESSNet does not report junk
ESSNet also dominates when sample size is large

Table 4.3: Number of subnetworks predicted by the various methods on a full dataset where the null distribution is computed using array rotation (rot), class-label swapping (cperm) and gene swapping (gswap); the first number denotes the number of subnetworks in the numerator of the subnetwork-level agreement and the second number denotes the number of subnetworks in the denominator of the subnetwork-level agreement; cf. equation 4.5.

<table>
<thead>
<tr>
<th>Method</th>
<th>DMD</th>
<th></th>
<th>ALL</th>
<th></th>
<th>BCR</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rot</td>
<td>cperm</td>
<td>rot</td>
<td>cperm</td>
<td>rot</td>
<td>cperm</td>
</tr>
<tr>
<td>ESSNet</td>
<td>20/23</td>
<td>13/15</td>
<td>22/24</td>
<td>25/27</td>
<td>24/29</td>
<td>30/32</td>
</tr>
<tr>
<td>NEA-paired</td>
<td>77/98</td>
<td>91/115</td>
<td>140/163</td>
<td>109/119</td>
<td>176/192</td>
<td>37/43</td>
</tr>
<tr>
<td>ORA-paired</td>
<td>30/62</td>
<td>30/62</td>
<td>34/74</td>
<td>34/74</td>
<td>53/99</td>
<td>53/99</td>
</tr>
<tr>
<td>ORA-hypergeo</td>
<td>20/46</td>
<td>41/141</td>
<td>24/60</td>
<td>48/73</td>
<td>4/14</td>
<td>32/166</td>
</tr>
<tr>
<td>GSEA</td>
<td></td>
<td></td>
<td>23/64</td>
<td>24/69</td>
<td>8/52</td>
<td>17/48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7/57</td>
<td>5/46</td>
</tr>
</tbody>
</table>
Do ESSNet results agree on small datasets vs big datasets?

- Use ESSNet’s results on entire datasets as the benchmark to evaluate ESSNet’s results on small subsets of the datasets.
- The precision (i.e., agreement) is superb, though some subnetworks are missed when smaller datasets are analysed.

| Sample size (N) | Precision | | Recall | |
|----------------|-----------|-----------------|--------|
|                | DMD       | ALL             | BCR    | DMD       | ALL             | BCR    |
|                | D | −D | D | −D | D | −D | D | −D | D | −D | D | −D |
| 2              | 0.96 | 0.88 | 0.87 | 0.95 | 0.93 | 0.91 | 0.45 | 0.31 | 0.34 | 0.25 | 0.19 | 0.17 |
| 3              | 0.93 | 0.86 | 0.99 | 0.89 | 0.90 | 0.87 | 0.56 | 0.45 | 0.56 | 0.41 | 0.21 | 0.16 |
| 4              | 0.88 | 0.88 | 0.97 | 0.92 | 0.91 | 0.87 | 0.67 | 0.50 | 0.51 | 0.53 | 0.35 | 0.48 |
| 5              | 0.89 | 0.88 | 0.94 | 0.90 | 0.89 | 0.90 | 0.73 | 0.52 | 0.74 | 0.55 | 0.36 | 0.38 |
| 6              | 0.82 | 0.88 | 0.93 | 0.92 | 0.89 | 0.91 | 0.78 | 0.62 | 0.74 | 0.62 | 0.44 | 0.438 |
| 7              | 0.85 | 0.86 | 0.95 | 0.93 | 0.90 | 0.87 | 0.75 | 0.59 | 0.66 | 0.64 | 0.55 | 0.53 |
| 8              | 0.84 | 0.89 | 0.97 | 0.94 | 0.90 | 0.92 | 0.81 | 0.69 | 0.74 | 0.66 | 0.61 | 0.66 |
| 9              | 0.88 | 0.90 | 0.94 | 0.92 | 0.89 | 0.89 | 0.90 | 0.67 | 0.76 | 0.74 | 0.65 | 0.67 |
| 10             | 0.88 | 0.93 | 0.97 | 0.92 | 0.90 | 0.90 | 0.86 | 0.84 | 0.89 | 0.74 | 0.66 | 0.73 |
Leukemias: IL-4 signaling in ALL

For the Leukemia dataset (in which patients are either classified to have acute lymphoblastic leukemia or acute myeloid leukemia), one of the significant subnetworks that is biologically relevant is part of the Interleukin-4 signaling pathway, see figure 6b (supplementary material). The binding of Interleukin-4 to its receptor (Cardoso et al., 2008) causes a cascade of protein activation involving JAK1 and STAT6 phosphorylation. STAT6 dimerizes upon activation and is transported to the nucleus and interacts with the RELA/NFkB1 transcription factors, known to promote the proliferation of T-cells (Rayet and Gelineas, 1999). In contrast, acute myeloid leukemia does not have genes in this subnetwork up-regulated and are known to be unrelated to lymphocytes.
Remarks

• Consistent successful gene expression profile analysis needs deep integration of background knowledge

• Most gene expression profile analysis methods fail to give reproducible results when sample size is small (and some even fail when sample size is quite large)

• Logical analysis to identify key issues and simple logical solution to the issues can give fantastic results
DIFFICULTY OF CROSS-BATCH CLASSIFICATION
Batch effects

- Batch effects are common
- Batch effects cannot always be removed using common normalization methods
Gene-feature-based classifiers do badly when there are batch effects, even after normalization.

Predictive accuracy of gene-feature-based classifiers with and w/o rank normalization in the ALL/AML dataset.

Gene selection by t-test, SAM, or rank product. Classifier by naïve Bayes.
SUCCESSFUL CROSS-BATCH CLASSIFICATION

WHEN SAMPLE SIZE IS LARGE
PFSNet-based features

- **PFSNet**
  - Induce subnetworks from pathways by considering only genes highly expressed in majority of patients in any class
  - For each subnetwork $S$ and each patient $P_k$, compute a pair of scores:
    \[
    \beta_1^*(g_i) = \sum_{p_j \in D} \frac{f_s(e_{g_i,p_j})}{|D|} \quad \beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{f_s(e_{g_i,p_j})}{|\neg D|}
    \]
    \[
    Score_{p_k}^1(S) = \sum_{g_i \in S} f_s(e_{g_i,p_k}) \cdot \beta_1^*(g_i) \quad Score_{p_k}^2(S) = \sum_{g_i \in S} f_s(e_{g_i,p_k}) \cdot \beta_2^*(g_i)
    \]

- Straightforward to use these scores as features
Successfully reducing batch effects

**Figure 5.6**: A figure showing that the batch effects are reduced by PFSNet subnetwork features. The colors red and blue represent different batches.
Successful cross-batch classification

**Figure 5.7:** A figure showing that data points are separated by class labels instead of batch when PFSNet features are used. The colors green and orange represent different classes.
SUCCESSFUL CROSS-BATCH CLASSIFICATION EVEN WHEN SAMPLE SIZE IS SMALL
ESSNet scores subnetworks but not patients.

How to produce feature vectors for patients?

ESSNet

- Induce subnetworks using genes highly expressed in majority of samples in any class
  - Let $g_i$ be genes in a given subnetwork S
  - Let $p_j$ be patients
  - Let $q_k$ be normals
  - 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
ESSNet-based features

• The idea is to see whether the pairwise differences of genes with a subnetwork between a given subject $p_x$ and the two separate classes ($D$ and $\neg D$) have a distribution around 0

$$\Delta_{(D)}(S, p_x) = \{e_{g_i, p_x} - e_{g_i, p'} \mid g_i \in S \text{ and } p' \in D\}$$

$$\Delta_{(\neg D)}(S, p_x) = \{e_{g_i, p_x} - e_{g_i, p'} \mid g_i \in S \text{ and } p' \in \neg D\}$$

• We expect $\Delta(D)(S, P_x)$ and $\Delta(\neg D)(S, P_x)$ to have +ve or –ve median for patients in one of the classes iff subnetwork $S$ is useful for classification
  – The median and $\pm$2 std dev of $\Delta(D)(S, P_x)$ and $\Delta(\neg D)(S, P_x)$ give 6 features for $P_x$
ESSNet-based features

- We also obtain pairwise differences of genes within a subnetwork among all possible pairs of patients in D and \(-D\)

\[
\Delta_{(D \rightarrow \neg D)}(S) = \{ e_{g_i, p'} - e_{g_i, p''} \mid g_i \in S \text{ and } p' \in D \text{ and } p'' \in \neg D \}
\]

Similarly for \(\Delta_{(-D \rightarrow \neg D)}(S), \Delta_{(-D \cdot D)}(S), \Delta_{(D \cdot D)}(S)\)

- This gives 4 more features

\[
\begin{align*}
\text{ESSNet\_feature}_7^{px, S} &= T\_statistic(\Delta_{(-D)}(S, p_x), \Delta_{(D \rightarrow \neg D)}(S)) \\
\text{ESSNet\_feature}_8^{px, S} &= T\_statistic(\Delta_{(-D)}(S, p_x), \Delta_{(-D \cdot \neg D)}(S)) \\
\text{ESSNet\_feature}_9^{px, S} &= T\_statistic(\Delta_{(D)}(S, p_x), \Delta_{(D \cdot D)}(S)) \\
\text{ESSNet\_feature}_{10}^{px, S} &= T\_statistic(\Delta_{(D)}(S, p_x), \Delta_{(-D \cdot \neg D)}(S))
\end{align*}
\]
ESSNet-based features lead to high cross-batch classification accuracy
ESSNet-based features retain high cross-batch classification accuracy even when training-sample size is small.
Remarks

• Traditional methods of classifying gene expression profiles often have difficulty predicting outcome of new batches of patients
  – Normalization does not always help

• ESSNet-based features are much more robust even when training-sample size is small
  – Subnetworks found by ESSNet are reproducible and gave high cross-batch classification accuracy

⇒ ESSNet is successful in isolating disease-relevant subnetworks from pathways
BACK TO PROTEOMICS
Not so easy to use the ESSNet idea in proteomics

\[ \Delta_{i,j,k} = \text{Expr}(g_i, p_j) - \text{Expr}(g_i, q_k) \]

in ESSNet compares expression of gene \( g_i \) in subjects \( p_j \) and \( q_k \)

Proteomic profiling is “semi random”
- A protein/peptides may get measured in \( p_j \) but may not get measured in \( q_k \)

PFSNet, interestingly, does not need to compare the expression of the same genes in two subjects

So use the PFSNet idea for proteomic profile analysis
Analyzing proteomic profiles in context of protein complexes

SNET, FSNET, PFSNET
SNet

1/ Identify DE complexes, rather than DE proteins

2/ Only highest-abundance proteins get to vote

- Given a protein \( g_i \) and a class of tissues \( C_j \), let

\[
\beta(g_i, C_j) = \sum_{p_k \in C_j} \frac{f_s(g_i, p_k)}{|C_j|}
\]

- where \( f_s(g_i, p_k) = 1 \), if the protein \( g_i \) is among the top \( \alpha\% \) most abundant proteins in the tissue \( p_k \), and \( = 0 \) otherwise

- Let the score of a protein complex \( S \) and a tissue \( p_k \) wrt to a class \( C_j \) be defined as:

\[
\text{score}(S, p_k, C_j) = \sum_{g_i \in S} [f_s(g_i, p_k) \times \beta(g_i, C_j)]
\]

- The test statistic is defined as:

\[
f_{SNet}(S, X, Y, C_j) = \frac{\text{mean}(S, X, C_j) - \text{mean}(S, Y, C_j)}{\sqrt{\frac{\text{var}(S, X, C_j)}{|X|} + \frac{\text{var}(S, Y, C_j)}{|Y|}}}
\]

- where \( \text{mean}(S, #, C_j) \) and \( \text{var}(S, #, C_j) \) are respectively the mean and variance of the list of scores \{ score(S, p_k, C_j) \mid p_k \text{ is a tissue in } # \}.

- Given two classes \( C_1 \) and \( C_2 \), the set of significant complexes returned by SNet is the union of \{ \text{S} \mid f_{SNet}(S, C_1, C_2, C_1) \text{ is significant} \} and \{ \text{S} \mid f_{SNet}(S, C_2, C_1, C_2) \text{ is significant} \}. 
FSNet

1/ Identify DE complexes, rather than DE proteins

2/ Only highest-abundance proteins get to vote

3/ Give other high-abundance proteins partial vote

Given a protein $gi$ and a class of tissues $Cj$, let

$$\beta(gi, Cj) = \sum_{pk \in Cj} \frac{fs(gi, pk)}{|Cj|}$$

and $fs(gi, pk) = 1.0, 0.8, 0.6, 0.4, 0.2, 0.0$ depending on how abundant $gi$ is in $pk$

Let the score of a protein complex $S$ and a tissue $pk$ wrt to a class $Cj$ be defined as:

$$score(S, pk, Cj) = \sum_{gi \in S} [fs(gi, pk) \ast \beta(gi, Cj)]$$

The test statistic is defined as:

$$f_{FSNET}^{(S, X, Y, Cj)} = \frac{\text{mean}(S, X, Cj) - \text{mean}(S, Y, Cj)}{\sqrt{\frac{\text{var}(S, X, Cj)}{|X|} + \frac{\text{var}(S, Y, Cj)}{|Y|}}}$$

where $\text{mean}(S, #, Cj)$ and $\text{var}(S, #, Cj)$ are respectively the mean and variance of the list of scores $\{\text{score}(S, pk, Cj) | pk \text{ is a tissue in } #\}$.

Given classes $C1$ and $C2$, the set of FSNet-significant complexes is the union of $\{S \mid f_{FSNET}(S, C1, C2, C1) \text{ is significant}\}$ and $\{S \mid f_{FSNET}(S, C2, C1, C2) \text{ is significant}\}$. 

PFSNet

1/ Identify DE complexes, rather than DE proteins

2/ Only highest-abundance proteins get to vote

3/ Give other high-abundance proteins partial vote

4/ Let the votes be weighted by their abundance in both phenotypes

- Let $\delta(S,pk,X,Y) = score(S,pk,X) - score(S,pk,Y)$, where $score(S,pk,#)$ is as in FSNet

- If complex $S$ is irrelevant, $E(\delta(S,pk,X,Y)) = \sim 0$. So define a one-sample t-statistic:

$$f_{PFSNet}(S,X,Y,Z) = \frac{\text{mean}(S,X,Y,Z)}{\text{se}(S,X,Y,Z)}$$

- where $\text{mean}(S, X, Y, Z)$ and $\text{se}(S, X, Y, Z)$ are respectively mean and s.e. of the list $\{ \delta(S, pk, X, Y) \mid pk \text{ is a tissue in } Z \}$

- Given two classes $C_1$ and $C_2$, the set of PFSNet-significant complexes is union of $\{ S \mid f_{PFSNet}(S, C_1, C_2, Z) \text{ is significant} \}$ and $\{ S \mid f_{PFSNet}(S,C_2,C_1, Z) \text{ is significant} \}$, where $Z = C_1 \cup C_2$
The SWATH dataset from (Guo et al. 2015) was used in this and later slides. It contains 24 SWATH runs from 6 pairs of non-tumorous and tumorous clear-cell renal carcinoma tissues, which have been swathed in duplicates (12 normal, 12 cancer).
Agreement of significant complexes betw different subsamples
Stability of significant complexes
Precision & recall wrt complexes identified using the whole dataset

Ave F = 0.67

Ave F = 0.71

Ave F = 0.92
Cross-validation performance

<table>
<thead>
<tr>
<th>Group</th>
<th>SNET</th>
<th>FSNet</th>
<th>PFSNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. significant features (0.05)</td>
<td>self_validatio n</td>
<td>cross_validati on</td>
</tr>
<tr>
<td>1.00</td>
<td>21.00</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>2.00</td>
<td>23.00</td>
<td>1.00</td>
<td>0.92</td>
</tr>
<tr>
<td>3.00</td>
<td>20.00</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>4.00</td>
<td>17.00</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>5.00</td>
<td>26.00</td>
<td>0.92</td>
<td>0.83</td>
</tr>
<tr>
<td>6.00</td>
<td>18.00</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>7.00</td>
<td>15.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>8.00</td>
<td>19.00</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>9.00</td>
<td>18.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>10.00</td>
<td>23.00</td>
<td>0.92</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SNET</th>
<th>FSNet</th>
<th>PFSNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>20.00</td>
<td>0.96</td>
<td>0.86</td>
</tr>
<tr>
<td>s.d.</td>
<td>3.30</td>
<td>0.06</td>
<td>0.16</td>
</tr>
<tr>
<td>COV</td>
<td>0.16</td>
<td>0.06</td>
<td>0.19</td>
</tr>
</tbody>
</table>

- Naïve Bayes training using score(S,pk,#), delta(S,pk,X,Y) and paired(S,pk,X,Y) for SNet/FSNet, PFSNet
- Good performance despite small # of features used
Closing remarks

• SNet/FSNet/PFSNet are based on ranks, not actual abundance level
• They also do not rely on comparing abundance level of the same proteins in different tissues

⇒ Potentially more robust in future data batches
⇒ Extend utility of proteomic analysis, and increase the likelihood of identifying stable, consistent and generalizable biomarkers
Acknowledgements

• **My students**
  – Donny Soh
  – Dong Difeng
  – Kevin Lim
  – Wilson Goh

• **& collaborators**
  – Choi Kwok Pui
  – Li Zhenhua
  – Tian Nan Guo
  – Ruedi Aebersold

• **Singapore Ministry of Education**

• Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Finding Consistent Disease Subnetworks Across Microarray Datasets.** *BMC Genomics*, 12(Suppl. 13):S15, November 2011

• Kevin Lim, Limsoon Wong. **Finding consistent disease subnetworks using PFSNet.** *Bioinformatics*, 30(2):189--196, January 2014

• Kevin Lim, Zhenhua Li, Kwok Pui Choi, Limsoon Wong. **A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small.** *JBCB*, in press.