CS4220: Knowledge Discovery Methods for Bioinformatics
Unit 4: Proteomic Profile Analysis

Wong Limsoon
Delivering more powerful proteomic profile analysis

- Basic proteomic profile analysis
- Common issues in proteomic profile analysis
- Improving consistency
  - PSP, PDS
- Improving coverage
  - CEA, PEP, Max Link
Basic Proteomic Profile Analysis
Typical Proteomic MS Experiment

See also http://www.slideshare.net/joachimjacob/bits-introduction-to-mass-spec-data-generation

Diagnosis Using Proteomics

Technology-dependent
a) peptide and protein identification from PSMs
b) feature detection, quantification, annotation, and alignment
c) peptide significance analysis
d) protein significance analysis

Technology-independent
e) class discovery
f) class prediction
g) data integration
h) pathway analysis

A rather nice set of proteomic profiles of leukemia patients.

Protein Identification by Mass Spec

Step 1: MS/MS instrument

Database search
- Sequest, Mascot, InSpect
- de Novo interpretation
- Lutefisk, Peaks, PepNovo

Source: Leong Hon Wai
Tandem Mass-Spectrometry

Source: Leong Hon Wai
Breaking Protein into Peptides, and Peptides into Fragment Ions

- Proteases, e.g. trypsin, break protein into peptides
- A Tandem Mass Spectrometer further breaks the peptides down into fragment ions and measures the mass of each piece
- Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones
- Mass Spectrometer measures mass/charge ratio of an ion

Source: Leong Hon Wai
Peptide Identification by Mass Spec

MS/MS instrument

Database search
• Sequest, Mascot, InSpect
de Novo interpretation
• Lutefisk, Peaks, PepNovo

Step 2: Understanding an MS/MS Spectrum

Source: Leong Hon Wai
Peptide Fragmentation

Collision Induced Dissociation

$$\text{H}...-\text{HN-CH-CO} \quad \ldots \quad \text{NH-CH-CO-NH-CH-CO-}\ldots\text{OH}$$

 Prefix Fragment

Suffix Fragment

- Peptides tend to fragment along the backbone
- Fragments can also loose neutral chemical groups like NH$_3$ and H$_2$O

Source: Leong Hon Wai

Peptide Fragmentation

Figure 1: (a) The structure of an amino-acid. (b) An ionized peptide. (c) $y_{n-1}^+$ ion
... and fragments due to neutral losses

Source: Leong Hon Wai
• The peaks in the mass spectrum:
  – Prefix and Suffix Fragments
  – Fragments with neutral losses (\(-\text{H}_2\text{O}, -\text{NH}_3\))
  – Noise and missing peaks

Source: Leong Hon Wai

Example MS/MS Spectrum

![MS/MS Spectrum](image)

**Figure 2:** MS/MS spectrum for peptide SGFLEEDK.
Protein Identification with MS/MS

Source: Leong Hon Wai
Peptide Identification by Mass

**Step 3: Computational Methods**

- Database search: Sequest, Mascot
- de Novo interpretation: Lutefisk, Peaks, PepNovo

Source: Leong Hon Wai
Database Search Algorithms

• Database search
  – Used for spectrum from known peptides
  – Rely on completeness of database

• General Approach
  – Match given spectrum with known peptide
  – Enhanced with advanced statistical analysis and complex scoring functions

• Methods
  – SEQUEST, MASCOT, InsPecT, Paragon
Theoretical Spectrum for a Peptide

- Given this peptide
  
  \[ G \quad V \quad D \quad L \quad K \]

- Its theoretical spectrum is

  ![Peptide Mass Spectrum](image)

- Theoretical spectrum is dependent on
  - Set of ion-types considered
  - Larger if multi-charge ions are considered

Source: Leong Hon Wai
Database Search Algorithm

**Database Search**

**Database of known peptides**

<table>
<thead>
<tr>
<th>Peptide Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIIMNVRT, HEWAILF, GHNWAMNAC, GVFGSVLRA, EKLNKAATYIN..</td>
</tr>
</tbody>
</table>

**Match**

**Matching Score for this peptide**

**Repeat for all the peptides in the Database**

Source: Leong Hon Wai
De Novo Sequencing Algorithms

• Given a spectrum
  – Build a spectrum graph
  – Peptides are paths in this graph
  – Find the best path

Source: Leong Hon Wai
Spectrum Graph for a Peptide

- Connect peaks together
  - If their mass difference = mass of an amino acid
- Theoretical spectrum is dependent on
  - Set of ion-types considered
  - Larger if multi-charge ions are considered

Source: Leong Hon Wai
Building a Graph from a Spectrum

Source: Leong Hon Wai

De Novo Sequencing Algorithms

Find longest directed acyclic path

AVGELTK
De Novo vs. Database Search

Database Search

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIMMNVT, AVGELTK, HEWAILF, GHNLWAMNAC, GVFGSLRA, EKLNAATYIN..

De Novo

Database of all peptides $20^n$

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIMMNVT, AVGELTK, HEWAILF, GHNLWAMNAC, GVFGSLRA, EKLNAATYIN...

AVGELTK

Source: Leong Hon Wai
De Novo vs. Database Search: A Paradox

- The database of all peptides is huge ≈ $O(20^n)$
- The database of all known peptides is much smaller ≈ $O(10^8)$

- However, de novo algorithms can be much faster, even though their search space is much larger!
  - A database search scans all peptides in the search space to find best one
  - De novo eliminates the need to scan all peptides by modeling the problem as a graph search

Source: Leong Hon Wai
Protein Identification

- After all the peptides have been identified, they are grouped into protein identifications.
- Peptide scores are added up to yield protein scores.
- Confidence of a particular peptide identification increases if other peptides identify the same protein and decreases if no other peptides do so.
- Protein identifications based on single peptides should only be allowed in exceptional cases.

Cf. Gene Expression Profile Analysis

- Once the proteins are identified, the proteomic profile of a sample can be constructed
  - I.e., which protein is found in the sample and how abundant it is

- Similar to gene expression profile. So gene expression profile analysis techs can be applied

- Some key differences
  - Proteomic profile has much fewer features
  - Proteomic profiling study has much fewer samples
Common Issues in Proteomic Profile Analysis
Peptide & protein identification by MS is still far from perfect

• “... peptides with low scores are, nevertheless, often correct, so manual validation of such hits can often ‘rescue’ the identification of important proteins.”

Typical frequency distribution of proteins detected in proteomic profiles.

Only 25 out of 800+ proteins are common to all 5 mod-stage HCC patients!

Image credit: Wilson Goh
Issues in Proteomic Profiling

• Coverage
• Consistency

⇒ Thresholding
  – Somewhat arbitrary
  – Potentially wasteful

• By raising threshold, some info disappears

Detected protein
Present but undetected protein

Image credit: Wilson Goh
Improving Consistency in Proteomic Profile Analysis
An inspiration from gene expression profile analysis

Gene Regulatory Circuits

- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype
- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype

Contextualization!

Taming false positives by considering pathways instead of all possible groups

Group of Genes

- Suppose
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
  - What is the chance of a group of 5 genes being perfectly correlated to these samples?
- Probi(group of genes correlated) = $\binom{1}{2}^5$
  - Good, << $1/2^5$
- # of groups = $100000 \times C_5^5$
- E(# of groups of genes correlated) = $100000 \times C_5^5 \times (1/2)^5 = 2.6 \times 10^{-7}$

$\Rightarrow$ Even more false positives?
- Perhaps no need to consider every group

Copyright 2011 © Limsoon Wong
Intuitive Example

- Suppose the failure to form a protein complex causes a disease
  - If any component protein is missing, the complex can’t form

⇒ Diff patients suffering from the disease can have a diff protein component missing
  - Construct a profile based on complexes?
We try an adaptation of SNet on proteomics profiles…

“Proteomic Signature Profiling” (PSP)
“Threshold-free” Principle of PSP

Applying PSP to a HCC Dataset

Consistency: Samples segregate by their classes with high confidence

Cluster dendrogram with AU/BP values (%)

Distance: euclidean
Cluster method: ward
Feature Selection


Feature Selection

\[
\text{t-score} = \frac{\bar{HA} - \bar{HB}}{S_{HA,HB} \sqrt{\frac{1}{n} + \frac{1}{m}}}
\]

where

\[
S_{HA,HB} = \sqrt{\frac{(m - 1)S_{HA}^2 + (n - 1)S_{HB}^2}{m + n - 2}}
\]
## Top-Ranked Complexes

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>p_val</th>
<th>mod_score</th>
<th>poor_score</th>
<th>cluster_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>5179</td>
<td>0.000300541</td>
<td>0.513951977</td>
<td>3.159758312</td>
<td>NCOA6-DNA-PK-Ku-PARP1 complex</td>
</tr>
<tr>
<td>5235</td>
<td>0.000300541</td>
<td>0.513951977</td>
<td>3.159758312</td>
<td>WRN-Ku70-Ku80-PARP1 complex</td>
</tr>
<tr>
<td>1193</td>
<td>0.000300541</td>
<td>0.513951977</td>
<td>3.159758312</td>
<td>Rap1 complex</td>
</tr>
<tr>
<td>159</td>
<td>0</td>
<td>0</td>
<td>2.810927655</td>
<td>Condensin I-PARP-1-XRCC1 complex</td>
</tr>
<tr>
<td>2657</td>
<td>0.008815869</td>
<td>0</td>
<td>2.55616281</td>
<td>ESR1-CDK7-CCNH-MNAT1-MTA1-HDAC2 complex</td>
</tr>
<tr>
<td>3067</td>
<td>0.00911641</td>
<td>0</td>
<td>2.55616281</td>
<td>RNA polymerase II complex, incomplete (CDK8 complex), chromatin structure modifying</td>
</tr>
<tr>
<td>1226</td>
<td>0.013323983</td>
<td>0.715352108</td>
<td>2.420592827</td>
<td>H2AX complex I</td>
</tr>
<tr>
<td>5176</td>
<td>0</td>
<td>0.513951977</td>
<td>2.339059313</td>
<td>MGC1-DNA-PKcs-Ku complex</td>
</tr>
<tr>
<td>1189</td>
<td>0</td>
<td>0.513951977</td>
<td>2.339059313</td>
<td>DNA double-strand break end-joining complex</td>
</tr>
<tr>
<td>5251</td>
<td>0</td>
<td>0.513951977</td>
<td>2.339059313</td>
<td>Ku-ORC complex</td>
</tr>
<tr>
<td>2766</td>
<td>0</td>
<td>0.513951977</td>
<td>2.339059313</td>
<td>TERF2-RAP1 complex</td>
</tr>
</tbody>
</table>

## Top-Ranked GO Terms

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Description</th>
<th>No. of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0016032</td>
<td>viral reproduction</td>
<td>36</td>
</tr>
<tr>
<td>GO:0000398</td>
<td>nuclear mRNA splicing, via spliceosome</td>
<td>34</td>
</tr>
<tr>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>28</td>
</tr>
<tr>
<td>GO:0000084</td>
<td>S phase of mitotic cell cycle</td>
<td>28</td>
</tr>
<tr>
<td>GO:0006366</td>
<td>transcription from RNA polymerase II promoter</td>
<td>26</td>
</tr>
<tr>
<td>GO:0006283</td>
<td>transcription-coupled nucleotide-excision repair</td>
<td>22</td>
</tr>
<tr>
<td>GO:0006369</td>
<td>termination of RNA polymerase II transcription</td>
<td>22</td>
</tr>
<tr>
<td>GO:0006284</td>
<td>base-excision repair</td>
<td>21</td>
</tr>
<tr>
<td>GO:0000086</td>
<td>G2/M transition of mitotic cell cycle</td>
<td>21</td>
</tr>
<tr>
<td>GO:0000079</td>
<td>regulation of cyclin-dependent protein kinase activity</td>
<td>20</td>
</tr>
<tr>
<td>GO:0010833</td>
<td>telomere maintenance via telomere lengthening</td>
<td>20</td>
</tr>
<tr>
<td>GO:0033044</td>
<td>regulation of chromosome organization</td>
<td>19</td>
</tr>
<tr>
<td>GO:0006200</td>
<td>ATP catabolic process</td>
<td>18</td>
</tr>
<tr>
<td>GO:0042475</td>
<td>odontogenesis of dentine-containing tooth</td>
<td>18</td>
</tr>
<tr>
<td>GO:0034138</td>
<td>toll-like receptor 3 signaling pathway</td>
<td>17</td>
</tr>
<tr>
<td>GO:0006915</td>
<td>apoptosis</td>
<td>17</td>
</tr>
<tr>
<td>GO:0006271</td>
<td>DNA strand elongation involved in DNA replication</td>
<td>17</td>
</tr>
</tbody>
</table>
False Positive Rate Analysis

• Divide 7 poor patients into 2 groups
  – Significant complexes produced by PSP here are false positives

• Repeat many times to get dull distribution
  – Median = 40, mode = 6

• Cf. 523 complexes in CORUM (size ≥4) used in PSP. At p ≤ 5%, 523 * 5% ≈ 27 false positives expected
A Shortcoming of PSP

• Protein complex databases are still relatively small & incomplete…

⇒ Augment the set of protein complexes by protein clusters predicted from PPI networks!

• Many protein complex prediction methods
  – CFA, Habibi et al. BMC Systems Biology, 4:129, 2010
  – …
Another Shortcoming of PSP

• Protein complexes provided a biologically-rich feature set for PSP
  – But it is only one aspect of biological function

• The other aspect is biological pathways
  – But coverage issue of proteomic profiles create lots of “holes”

• Can we extract and use subnets from pathways?
Another adaptation of SNet on proteomics profiles…

“Pathway-Derived Subnets” (PDS)
Pathway-Derived Subnets (PDS)

• Identify the set $S_i$ of proteins detected in more than 50% of samples having phenotype $P_i$
  – Do this for each phenotype $P_1, \ldots, P_k$

• Overlay $\cup_i S_i$ to pathways

• Remove nodes not covered by $\cup_i S_i$
  ⇒ This fragments pathways into subnets

• Use these subnets to form “proteomic signature profiles”
  – The rest of the steps is same as PSP

Source: Wilson Goh

**PDS consistently segregates mod vs poor patients**

Source: Wilson Goh
What have we learned?

• Contextualization (into complexes and pathways) can deal with consistency issues in proteomics

• GO term analysis also indicates that context-based methods (PSP, PDS) select clusters that play integral roles in cancer

• Context-based methods (PSP, PDS) reveal many potential clusters and are not constrained by any prior arbitrary filtering which is a common first step in conventional analytical approaches
Improving Coverage in Proteomic Profile Analysis
Typical proteomic profiling misses many proteins

Need to improve coverage!

Image credit: Wilson Goh
FCS

• Rescue undetected proteins from high-scoring protein complexes

• Why?

Let A, B, C, D and E be the 5 proteins that function as a complex and thus are normally correlated in their expression. Suppose only A is not detected and all of B–E are detected. Suppose the screen has 50% reliability. Then, A’s chance of being false negative is 50%, & the chance of B–E all being false positives is \((50\%)^4=6\%\). Hence, it is almost 10x more likely that A is false negative than B–E all being false positives.

• Shortcoming: Databases of known complexes are still small
CEA

- Generate cliques from PPIN
- Rescue undetected proteins from cliques containing many high-confidence proteins

- Reason: Cliques in a PPIN often correspond to proteins at the core of complexes

- Shortcoming: Cliques are too strict
  ⇒ Use more power complex prediction methods

- Map high-confidence proteins to PPIN
- Extract immediate neighbourhood & predict protein complexes using CFinder
- Rescue undetected proteins from high-ranking predicted complexes

- Reason: Exploit powerful protein complex prediction methods

- Shortcoming: Hard to predict protein complexes
  - Do we need to know all the proteins a complex?
MaxLink

- Map high-confidence proteins ("seeds") to PPIN
- Identify proteins that talk to many seeds but few non-seeds
- Rescue these proteins

- Reason: Proteins interacting with many seeds are likely to be part of the same complex as these seeds

- Shortcoming: Likely to have more false-positives
“Validation” of Rescued Proteins

• Direct validation
  – Use the original mass spectra to verify the quality of the corresponding y- and b-ion assignments
  – Immunological assay, etc.

• Indirect validation
  – Check whether recovered proteins have GO terms that are enriched in the list of seeds
  – Check whether recovered proteins show a pattern of differential expression between disease vs normal samples that is similar to that shown by the seeds
An example using the PEP approach to recover undetected proteins …
Background

• **HCC (Hepatocellular carcinoma)**
  – Classified into 3 phases: differentiated, moderately differentiated and poorly differentiated

• **Mass Spectrometry**
  – iTRAQ (Isobaric Tag for Relative and Absolute Quantitation)
  – Coupled with 2D LC MS/MS
  – Popular because of ability to run 8 concurrent samples in one go
Poor and mod proteins are widely interspersed

• In the subnet of reported proteins in mod and poor, poor and mod genes are well mixed

- Mod and Poor
- Poor only

Image credit: Wilson Goh
Identify the “seeds”
Ratio < 0.8 and > 1.25 for Mod (min 3 patients)
Ratio < 0.8 and > 1.25 for Poor (min 4 patients)

PEP Workflow

Expansion to include neighbors greatly improves coverage.

Returning to Mass Spectra

- Test set: Several proteins (ACTR2, CDC42, GNB2L1, KIF5B, PPP2R1A, PKACA and TOP1) from top 34 clusters not detected by Paragon

- The test: Examine their GPS and Mascot search results and their MS/MS-to-peptide assignments

- Assessment of MS/MS spectra of their top ranked peptides revealed accurate y- and b-ion assignments and were of good quality (p < 0.05)

⇒ In silico expansion verified

Successful Verification

**ACTR2**

*Goh et al. Journal of Proteome Research, 10(5):2261--2272, 2011*

**CDC42**

*Goh et al. Journal of Proteome Research, 10(5):2261--2272, 2011*
Another Experiment: Comparison

- **Valproic acid (VPA)-treated mice vs control**
  - VPA or vehicle injected every 12 hours into postnatal day-56 adult mice for 2 days
  - Role of VPA in epigenetic remodeling

- **MS was scanned against IPI rat db in round #1**
  - 396 proteins identified

- **MS was scanned against UniProtkb in round #2**
  - 393 additional proteins identified

- **All recovery methods ran on round #1 data and the recovered proteins checked against round #2**
Moderate level of agreement of reported proteins between various recovery methods
Performance Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Novel Suggested Proteins</th>
<th>Recovered proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEP</td>
<td>375</td>
<td>158</td>
</tr>
<tr>
<td>Maxlink</td>
<td>910</td>
<td>226</td>
</tr>
<tr>
<td>FCS (predicted)</td>
<td>678</td>
<td>224</td>
</tr>
<tr>
<td>FCS (complexes)</td>
<td>789</td>
<td>775</td>
</tr>
</tbody>
</table>

- Looks like running FCS on real complexes is able to recover more proteins and more accurately.
Precision vs recall of running FCS on real complexes
Must Read

• Steen & Mann. The ABC’s and XYZ’s of peptide sequencing. Nature Reviews Molecular Cell Biology, 5:699-711, 2004


Good to Read


From proteomics to metabolomics & lipidomics: Can the same network-based approach be applied?
Acknowledgements

- The slides on peptide identification were adapted from those given to me by A/P Leong Hon Wai

- A lot of the slides on PSP, PDS, and PEP came from the work of Wilson Goh