Introduction to Proteomics & Bioinformatics

Simon Hubbard
Faculty of Life Sciences,
The University of Manchester
Overview

• Proteomics
• Lecture 1
  – 30,000 ft view of experimental proteomics
  – Generalised proteomics workflow
  – Bioinformatics problems in
    • Peptide Mass Fingerprinting
    • Tandem Mass Spectroscopy
• Lecture 2
  – Quantitative proteomics
  – Genome annotation
  – Protein-protein interactions
    • Mass-spec based
Why study the proteome?

- Proteins are the cells functional entities
  - Dynamic
  - State-dependent

- One gene does not equal one protein
  - Isoforms
  - Post-translational modifications

- Transcriptome is not perfectly correlated
  - Bigger picture needed for systems biology

- Interactions, molecular machines, systems biology
  - Much of this happens at the protein level
So what can proteomics tell us about protein function

- **Qualitative**
  - Protein exists!
    - Gene structure in genome (validation)
    - Isoform exists
  - Post-translational modifications
  - Protein-protein interactions

- **Quantitative**
  - Relative expression
    - ICAT, iTRAQ
    - SILAC
    - Counting peptides
    - Peak intensities
  - Absolute levels
    - AQUA, QconCAT
A generalised proteomics experiment

- Various routes through this map
- Complexity reduction

Separating by size or charge in most cases

Identify peptides as a proxy for proteins, comparing theoretical to experimental spectra
Key analytical technique is Mass Spectrometry

- Ionisation methods
- 2 Nobel Prizes
- Key advance

Proteins and peptides don’t want to be in the gas phase!
**General proteomics experiment**

- **genome**
  - Knowledge + prediction
  - Transcription & regulation
  - Post-translational modifications

- **real proteome**
  - Separation methods
  - 2D-gels, functional separations, n-dimensional chromatography

- **simple mixtures & single proteins**
  - Digest

- **“virtual” proteome**
  - Knowledge + prediction
  - Peptide mass and fragment database
  - Bioinformatics identification

- **Peptide mass map fingerprint**

- **Fragment ion spectra**
Proteomic database searching

- 2 principle types of database search
  
  - Peptide Mass Fingerprinting (PMF)
    - Submit the m/z values of hydrolytic peptides
    - PROBLEM: Identify the most likely protein which gave rise to the observed peptide m/z values
  
  - Tandem MS approaches
    - Submit the precursor ion and product ion m/z values
    - PROBLEM: Identify the most likely peptide which has the same precursor ion m/s and which gave rise to a similar fragment ion spectra.
      - Database search
      - De Novo sequencing
Parameters to consider

- **Mass spectrometric**
  - peptide masses
    - mass type (monoisotopic, average)
    - Number (more is not always good! Mixtures ?)
  - peptide charge states
  - mass accuracy (instrument resolution)
  - **missed cleavages**
  - partial (or complete) sequence information
- **Proteome protein**
  - Protein mass
  - Protein isoelectric point (pI)
  - Database (type, size, quality, etc)
- **Modifications**
  - fixed or variable (cysteines, oxidation of met S, N-termini, etc.)
Tryptic digestion - Missed Cleavages

- We have mass spectra, and we already know:
  - Masses of amino acids (to 4 dp)
  - Sequences of entire genomes
  - Rules of tryptic cleavage
Parameter example - Missed cleavages

- Trypsin specificity is not 100%

<table>
<thead>
<tr>
<th>Position relative to cleavage site</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4 P3 P2 P1 P1' P2 P3</td>
<td></td>
</tr>
<tr>
<td>[WYF] [RK] [*RK]</td>
<td>Monigatti and Berndt, 2005</td>
</tr>
<tr>
<td>[*RK] [RK] [WYF]</td>
<td>Monigatti and Berndt, 2005</td>
</tr>
<tr>
<td>[DE] [RK] [*RK]</td>
<td>Monigatti and Berndt, 2005; Yen et al., 2006</td>
</tr>
<tr>
<td>[*RK] [RK] [DE]</td>
<td>Monigatti and Berndt, 2005; Thiede et al., 2000; Yen et al., 2006</td>
</tr>
<tr>
<td>[RK] [*RK] [DE] [DE]</td>
<td>Yen et al., 2006</td>
</tr>
<tr>
<td>[DE] [DE] [*RK] [RK]</td>
<td>Yen et al., 2006</td>
</tr>
<tr>
<td>[DE] [*RK] [RK] [*RK] [DE]</td>
<td>Yen et al., 2006</td>
</tr>
<tr>
<td>[*RK] [RK] [RK] [DE]</td>
<td>Monigatti and Berndt, 2005; Thiede et al., 2000</td>
</tr>
<tr>
<td>[*RK] [RK] [H]</td>
<td>Monigatti and Berndt, 2005</td>
</tr>
<tr>
<td>[RK] [P]</td>
<td>Monigatti and Berndt, 2005; Thiede et al., 2000; Yen et al., 2006</td>
</tr>
</tbody>
</table>

Lots of acidic residues local to scissile peptide bond
PMF Scoring systems

- The problem

- Search list of peptides
  - Test if database protein matches search list
  - Score for each protein
  - Simplest score = # matching peptides

Protein database
PMF Scoring systems

- Scores based on:
  - Number of matching expt peptides
  - Total number of theoretical matched peptides
  - Mass defect (error)
  - Protein size (bias against small proteins)

- Significance assessed by chance of match occurring randomly
  - MASCOT score: probabilistic basis, score > 70 is supposedly significant (at 5% level)
  - Profound/KNEXUS: gives a true probability
    - hence 1 = certain hit, and 0.95 is 95% certain.
MASCOT Example

Mascot Search Results

User: C. Alucard
Email: alucard@bats.ac.tz
Search title: Unknown red protein
Database: MSDB 20010401 (634857 sequences; 196694506 residues)
Timestamp: 24 Apr 2001 at 08:16:01 GMT
Top Score: 103 for 1A6K, myoglobin - sperm whale

Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event. Protein scores greater than 71 are significant (p<0.05).

Look for this threshold
Other features are important

- Recent paper from Brown lab (Aberdeen)
  - 3 simple features out-perform Mascot scoring
    - Hit-rate (no. masses matched/no. masses submitted)
    - Coverage (fraction seq covered * protein size)
    - Excess of limit peptides (over missed ones)
Tandem MS: Basic principles

Peptides delivered via LC

Select 1 peptide ion

Compare with library of theoretical peptide spectra

SEQUEST, MASCOT, X!TANDEM, PHENYX, OMSSA
Peptide ion fragmentation

- Different series of ions are produced
- Dependent on ion instrument type, energy etc.
SEQUEST Model Spectrum
Example – Matching Spectra to Theoretical Spectra

(protonated mass 1410.6)

<table>
<thead>
<tr>
<th>mass</th>
<th>b-ions</th>
<th>y-ions</th>
<th>mass'</th>
</tr>
</thead>
<tbody>
<tr>
<td>88.1</td>
<td>S</td>
<td>FAFDSINAMELK</td>
<td>1323.6</td>
</tr>
<tr>
<td>181.2</td>
<td>SF</td>
<td>AFDSINAMELK</td>
<td>1226.4</td>
</tr>
<tr>
<td>256.3</td>
<td>SFA</td>
<td>FDSINAMELK</td>
<td>1155.4</td>
</tr>
<tr>
<td>403.5</td>
<td>SPAF</td>
<td>DSIAMELK</td>
<td>1008.2</td>
</tr>
<tr>
<td>518.5</td>
<td>SPAFD</td>
<td>SINAMELK</td>
<td>893.1</td>
</tr>
<tr>
<td>605.6</td>
<td>SPAFDS</td>
<td>INAMELK</td>
<td>806.0</td>
</tr>
<tr>
<td>718.8</td>
<td>SPAFDSI</td>
<td>HAMELK</td>
<td>692.3</td>
</tr>
<tr>
<td>830.0</td>
<td>SPAFDSIM</td>
<td>AEMLK</td>
<td>561.7</td>
</tr>
<tr>
<td>921.1</td>
<td>SFAFDSIMA</td>
<td>DKELK</td>
<td>490.6</td>
</tr>
<tr>
<td>1050.2</td>
<td>SPAFDSIMAE</td>
<td>TELK</td>
<td>361.5</td>
</tr>
<tr>
<td>1151.3</td>
<td>SPAFDSIMAE</td>
<td>LEK</td>
<td>260.4</td>
</tr>
<tr>
<td>1264.4</td>
<td>SPAFDSIMATEL</td>
<td>K</td>
<td>147.2</td>
</tr>
</tbody>
</table>
Tandem MS & database searching

- **Sequest (Yates)**
  - Can find hits & provide sequence from uninterpreted data
  - 2 step search
    - 1) Find set of top-scoring 500 peptides (can ignore enzyme cleavage rules) which match to parent ion
    - 2) Generate detailed theoretical spectra of all fragment ions
Sequest

• Stage 1 – look at likely candidates within +/- 2Da

\[ S_p = \left( \sum i_m \right) n_i (1 + \beta)(1 + \rho) / n_t \]

• \( n_t \) = total no of predicted sequence ions
• \( n_i \) = total no of predicted ions that match expt
• \( i_m \) = ion intensity (abundance) of matching ions

• This filters out many possible precursors
• Takes top 500
Tandem MS & database searching

- Sequest (Yates)
  - 2) Generate detailed theoretical spectra of all fragment ions and cross correlate to experimental one (Xcorr)
How are spectra matched to peptides?

• **SEQUEST**
  
  – Experimental spectra post processing
  
  – “normalise” experimental & theoretical
    
    • b and y-ions normalised to 50
    • 25.0 for neutral losses, 10 for a-ions
  
  – Calculate $X_{\text{corr}}$
  
  – Normalise top $X_{\text{corr}}$ score to 1
  
  – (sometimes referred to as $C_n$)
  
  – Calculate $\Delta C_n$
    
    • Calculate difference between first and second $C_n$ values
  
  – $X_{\text{corr}} > 2.0$ and $\Delta C_n > 0.1$
    worth investigating further
MASCOT also provides MS-MS searches

- Ion scores
- 2 cutoffs
  - “identity”
  - “homology”

Protein score
Duplicates

Ion scores
- Mass: 71221
- Score: 305
- Queries matched: 9

Proteins matching the same set of peptides:
- AAF17824
- APS42068: NID: - Bos taurus
- AAAS1411
- BOVALBUMIN: NID: - Bos taurus
How do I know my score is correct?

- Find out “random” score distribution
  - Random sequences
  - Shuffled sequences
  - Reversed sequences
- Statistical model, or FDRs
Significance: Reverse database searching

- Concatenate forward DB and reverse DB
- Hits to “reverse” entries must be “false”
- If there are \( n_{\text{for}} \) hits to the forward database, \( n_{\text{rev}} \) to the reverse

\[
\text{FDR} = 2 \times \frac{n_{\text{rev}}}{n_{\text{rev}} + n_{\text{for}}}
\]

- Solution to multiple testing problem

Threshold chosen to give a fixed FDR (eg 1%)
Quantitative proteomics

• Why is it difficult?
  – Ion abundance != amount of peptide (proxy for protein)

• Standards are needed
  – Internal or external

• Relative or absolute
  – iTRAQ
  – SILAC
  – AQUA
  – QconCAT
Quantitative proteomics

**SI LAC**
Metabolic stable-isotope labelling

**ICAT/ITRAQ**
Isotope tagging by chemical reaction

**Heavy water**
Stable-isotope incorporation via enzyme reaction

Protein labelling

Digest

Data collection

Mass spectrometry

Data analysis

Common feature
Proteomic quantitative methods

- **iTRAQ™ (Applied Biosystems)**

  **Isobaric Tag**
  (Total mass = 145)

  **Reporter**
  (Mass = 114 thru 117)

  **Peptide Reactive Group**

  **Balance**
  (Mass = 31 thru 28)

  **PBG**

= MS/MS Fragmentation Site
Proteomic quantitative methods

- iTRAQ™ (Applied Biosystems)
SI LAC

- Stable Isotope LAbelling in Cell culture
- Quantitation achieved from integrated chromatogram
- Area under peak (heavy:light ratio gives quant)

Extracted ion chromatograms

Mass spectra
Recent example using SILAC

- Genome-wide coverage AND quantitation
  - Heavy Lys in one sample
  - SILAC is used here to test whether peaks are being missed in the MS
  - And yes, there are, because the instrument can’t keep up with the data coming in!
  - ~2000 protein ids

~8 Da shift for real peaks
Full quantitation, genome wide?

- Is now a reality, at least for yeast

De Godoy et al., Nature 455, 1251-
Yeast proteome quantified

- 3 expt protocols using SILAC
- Haploid vs Diploid yeast
- Lys-C digestion
- Revealed known/novel proteins expressed in different forms

De Godoy et al, Nature, 455, 1251-
• Many genes in higher abundance in haploid state
Whole genome absolute quantification

- *Leptospira interrogans* – spirochete, human pathogen
- 83% of proteome quantified
- Isotope labelling, label free methods
Proteomic applications

• Genome annotation
  – Why ?
    • Bone fide evidence
    • Better than transcripts
    • Can locate splice variants
    • Post-translational modifications

• Protein-protein interactions
  – High-throughput
  – Genome wide
    • 2 “classic” studies in yeast
Example peptides mapped to TIM

identified peptides

GAFTGEI$\text{SPAMIK}$  DIGAANVILGHSER  HVFGESDELI$\text{G}$QK  AIADNVK  VVLAYEPVWAIGTGK  IIYGGSV$\text{TGG}$NCK

TEVVC$\text{GAPSIYLD}$FARQKLDAKIKGVAAQNCYKV$\text{PK}$GAFTGEI$\text{SPAMIK}$DIGAANVILGH$\text{S}$ER  ERRHVFGESDELI$\text{G}$QKVH$\text{LA}$EAGL$\text{G}$VIA$\text{C}G$EKLDEREAGI$\text{TEK}$VVF$\text{EQ}$TKAIADNVK$\text{D}$  WSKVVLAYEPVWAIGTGK$\text{T}A$TPQQAQEVHEKLRGWLKS$\text{HVS}$DAVAQ$\text{STR}$IIYGGSV$\text{TGG}$NCK  CKELASQH$\text{D}$GFLVGGASLKPEFVDIINAK
TAP tagging

- Recombinant techniques
- TAG is attached to each ORF
Protein-protein interactions

- Gavin et al - TAP tagging approach
  - Tandem-affinity-purification

232 multi-protein complexes, 344 proteins, 231 novel functions
Yeast -> human protein function

- PsiBLAST, SMART-HMM used to define orthology
Comparison of proteomics datasets

- Different methods have different levels, and distinctly different patterns, and different pros/cons

Methods should be “centred” on diagonal

- Correlated mRNA expression
- in silico predicted interactions
- Genetic interactions (synthetic lethality)
- Overlap of high-throughput methods
- Reference set: known complexes (MIPS, YPD)
Comparison of proteomics datasets

- Comparison to MIPS/YPD set
  - 10,907 complexes
- Best results when methods agree
- Bioinformatic methods were quite good
Gavin et al 2006. Socio-Affinity clustering

- TAP-tagging yields proteins with many clusters
- Which proteins are just “common” and which interactions are “real”?

\[ A(i, j) = S_{i\land j = \text{bait}} + S_{i=\text{bait}, j}\land i = \text{bait} + M_{i, j}; \quad S_{i, j \land i = \text{bait}} = \log \left( \frac{n^\text{prey}_{i,j}}{f^\text{prey}_i f^\text{prey}_j \sum_{\text{all baits}} n^\text{prey} (n^\text{prey} - 1)/2} \right) \]

- Fraction of all purifications where \( i \) is the prey
- Fraction of all purifications where \( j \) is the prey
- Total number of baits (=purifications)
- Fraction of all purifications where \( j \) is the prey
- No. of purifications with baits other than \( i \) or \( j \)
- No. times protein \( i \) retrieves \( j \) when \( i \) is tagged
- No. Preys retrieved with protein \( i \) as bait
- Fraction of all purifications where \( i \) is the bait
Generates more informative protein complexes

Find best clustering parameters by comparison with gold standard sets
Summary

- Proteomics
- Bioinformatics tools necessary to make identifications
  - Identifications underpin
    - Qualitative
    - Quantitative
  - Functional information gained on
    - Protein is “expressed”
    - Gene structure
    - Isoforms
    - PTMs
    - Protein-protein interactions
    - Quantitation (level) via labelling