Protein1: Last week's take home lessons

- Protein interaction codes(s)?
- Real world programming
- Pharmacogenomics : SNPs
- Chemical diversity : Nature/Chem/Design
- Target proteins : structural genomics
- Folding, molecular mechanics & docking
- Toxicity animal/clinical : cross-talk
Protein2: Today's story & goals

- Separation of proteins & peptides
- Protein localization & complexes
- Peptide identification (MS/MS)
  - Database searching & sequencing.
- Protein quantitation
  - Absolute & relative
- Protein modifications & crosslinking
- Protein - metabolite quantitation
Why purify?

- Reduce one source of noise (in identification/quantitation)
- Prepare materials for in vitro experiments (sufficient causes)
- Discover biochemical properties
(Protein) Purification Methods

- **Charge**: ion-exchange chromatography, isoelectric focusing
- **Size**: dialysis, gel-filtration chromatography, gel-electrophoresis, sedimentation velocity
- **Solubility**: salting out
- **Hydrophobicity**: Reverse phase chromatography
- **Specific binding**: affinity chromatography
- **Complexes**: Immune precipitation (± crosslinking)
- **Density**: sedimentation equilibrium
Protein Separation by Gel Electrophoresis

• Separated by *mass*: Sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis.
  – Sensitivity: 0.02ug protein with a silver stain.
  – Resolution: 2% mass difference.

• Separated by *isoelectric point (pI)*: polyampholytes pH gradient gel.
  – Resolution: 0.01 pI.
Comparison of predicted with observed protein properties (localization, postsynthetic modifications)

E.coli

See Link et al. 1997 Electrophoresis 18:1259-313 (Pub)
## Computationally checking proteomic data

<table>
<thead>
<tr>
<th>Property</th>
<th>Basis of calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein charge</td>
<td>RKHYCDE (N,C), pKa, pH <a href="http://gcg.nhri.org.tw/isoelectric.html#algorithm">Pub</a></td>
</tr>
<tr>
<td>Protein mass</td>
<td>Calibrate with knowns (complexes)</td>
</tr>
<tr>
<td>Peptide mass</td>
<td>Isotope sum (incl.modifications)</td>
</tr>
<tr>
<td>Peptide LC</td>
<td>aa composition linear regression</td>
</tr>
<tr>
<td>Subcellular</td>
<td>Hydrophobicity, motifs <a href="http://gcg.nhri.org.tw/isoelectric.html#algorithm">Pub</a></td>
</tr>
<tr>
<td>Expression</td>
<td>Codon Adaptation Index (CAI)</td>
</tr>
</tbody>
</table>
Protein2: Today's story & goals

• Separation of proteins & peptides
• Protein localization & complexes
• Peptide identification (MS/MS)
  – Database searching & sequencing.
• Protein quantitation
  – Absolute & relative
• Protein modifications & crosslinking
• Protein - metabolite quantitation
Cell fraction: Periplasm
2D gel: SDS mobility
isoelectric pH

See Link et al. 1997 Electrophoresis 18:1259-313
(Pub)
Cell localization predictions


Using the information from the topology of 70 membrane proteins... correctly identifies 295 transmembrane helical segments in 70 membrane proteins with only two overpredictions.
Isotope calculations

Mass resolution 0.1% vs. 1 ppm

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>H(1)</td>
<td>1.007825</td>
<td>99.99</td>
<td>H(2)</td>
<td>2.014102</td>
<td>0.015</td>
</tr>
<tr>
<td>C(12)</td>
<td>12.000000</td>
<td>98.90</td>
<td>C(13)</td>
<td>13.003355</td>
<td>1.10</td>
</tr>
<tr>
<td>N(14)</td>
<td>14.003074</td>
<td>99.63</td>
<td>N(15)</td>
<td>15.000109</td>
<td>0.37</td>
</tr>
<tr>
<td>O(16)</td>
<td>15.994915</td>
<td>99.76</td>
<td>O(17)</td>
<td>16.999131</td>
<td>0.038</td>
</tr>
<tr>
<td>S(32)</td>
<td>31.972072</td>
<td>95.02</td>
<td>S(33)</td>
<td>32.971459</td>
<td>0.75</td>
</tr>
</tbody>
</table>
## Computationally checking proteomic data

<table>
<thead>
<tr>
<th>Property</th>
<th>Basis of calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein charge</td>
<td>RKHYCDE (N,C), pKa, pH <a href="http://gcg.nhri.org.tw/isoelectric.html#algorithm">Pub</a></td>
</tr>
<tr>
<td>Protein mass</td>
<td>Calibrate with knowns (complexes)</td>
</tr>
<tr>
<td>Peptide mass</td>
<td>Isotope sum (incl.modifications)</td>
</tr>
<tr>
<td>Peptide LC</td>
<td>aa composition linear regression</td>
</tr>
<tr>
<td>Subcellular</td>
<td>Hydrophobicity, motifs <a href="http://gcg.nhri.org.tw/isoelectric.html#algorithm">Pub</a></td>
</tr>
<tr>
<td>Expression</td>
<td>Codon Adaptation Index (CAI)</td>
</tr>
</tbody>
</table>
High Performance Liquid Chromatography
Mobile Phase of HPLC

• The interaction between the mobile phase and sample determine the migration speed.
  – Isocratic elution: constant migration speed in the column.
  – Gradient elution: gradient migration speed in the column.
Stationary Phase of HPLC

• The degree of interaction with samples determines the migration speed.
  – Liquid-Solid: polarity.
  – Liquid-Liquid: polarity.
  – Size-Exclusion: porous beads.
  – Normal Phase: hydrophilicity and lipophilicity.
  – Reverse Phase: hydrophilicity and lipophilicity.
  – Ion Exchange.
  – Affinity: specific affinity.
See


A Map is Like a 2D Peptide Gel

First Dimension:
Reverse Phase Chromatography
Separation By Hydrophobicity

Second Dimension:
Mass Spectrometry
Separation by Mass
What Information Can Be Extracted From A Single Peptide Peak

Isotopic Variants of DAFLGSFLYEYSR

@ 36.418 min

K.Leptos 2001
Directed Analysis of Large Protein Complexes by 2D separation: strong cation exchange and reversed-phased liquid chromatography.

A new 40S subunit protein

Protein2: Today's story & goals

• Separation of proteins & peptides
• Protein localization & complexes
• Peptide identification (MS/MS)
  – Database searching & sequencing.
• Protein quantitation
  – Absolute & relative
• Protein modifications & crosslinking
• Protein - metabolite quantitation
Tandem Mass Spectrometry


Quadrople Q1 scans or selects m/z.
Q2 transmits those ions through collision gas (Ar).
Q3 Analyzes the resulting fragment ions.
Ions

Droplet after reduction of size due to Evaporation

"Coulombic Explosion" (Disintegration)

Ion Evaporation

$M (H^+)_{n}$

Multiply Protonated Analyte
Peptide Fragmentation and Ionization

An example

b-ions

y-ions

b1: (R1)+
y1: (R4)+

b2: (R1-R2)+
y2: (R3-R4)+

b3: (R1-R2-R3)+
y3: (R2-R3-R4)+

b4: (R1-R2-R3-R4)+
y4: (R1-R2-R3-R4)+
Mass Spectrum Interpretation Challenge

- It is unknown whether an ion is a b-ion or an y-ion or else.
- Some ions are missing.
- Each ion has multiple of isotopic forms.
- Other ions (a or z) may appear.
- Some ions may lose a water or an ammonia.
- Noise.
- Amino acid modifications.
A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry

**SEQUEST: Sequence-Spectrum Correlation**

Given a raw tandem mass spectrum and a protein sequence database.

- For every protein in the database,
- For every subsequence of this protein
  - Construct a hypothetical tandem mass spectrum
  - Overlap two spectra and compute the correlation coefficient (CC).

- Report the proteins in the order of CC score.

Protein2: Today's story & goals

- Separation of proteins & peptides
- Protein localization & complexes
- Peptide identification (MS/MS)
  - Database searching & sequencing.
- Protein quantitation
  - Absolute & relative
- Protein modifications & crosslinking
- Protein - metabolite quantitation
# Expression quantitation methods

<table>
<thead>
<tr>
<th>RNA</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes immobilized labeled RNA</td>
<td>Antibody arrays</td>
</tr>
<tr>
<td>RNAs immobilized labeled genes-</td>
<td>Westerns</td>
</tr>
<tr>
<td>Northern gel blot</td>
<td>-none-</td>
</tr>
<tr>
<td>QRT-PCR</td>
<td>same</td>
</tr>
<tr>
<td>Reporter constructs</td>
<td>same (Antibodies)</td>
</tr>
<tr>
<td>Fluorescent In Situ (Hybridization)</td>
<td>-none-</td>
</tr>
<tr>
<td>Tag counting (SAGE)</td>
<td>mass spec</td>
</tr>
</tbody>
</table>
Molecules per cell

E.coli/yeast         Human

Individual mRNAs:
$10^{-1}$ to $10^3$ $10^{-4}$ to $10^5$

Proteins:
$10$ to $10^6$ $10^{-1}$ to $10^8$
MS Protein quantitation $R = .84$

Yeast Protein ESI-MS Quantitation

$y = 0.8754x + 0.1573$
$R^2 = 0.8381$

Link, et al
MS quantitation reproducibility
Sample: Angiotensin, Neurotensin, Bradykinin

Map: 600 – 700 m/z

Coefficients of Variance

\[ CV = \frac{\sigma}{\mu} \]
Correlation between protein and mRNA abundance in yeast

Normality tests

See Weiss 5th ed. Page 920.
Types of non-normality: kurtosis, skewness (www) (log) transformations to normal.
(http://www.marketminer.com/prophet/statguide/n-dist_exam_res.html)

Spearman correlation rank test

\[ r_s = 1 - \frac{6S}{(n^3-n)} \]

Rank (from 1 to n, where n is the number of pairs of data) the numbers in each column. If there are ties within a column, then assign all the measurements that tie the same median rank. Note, avoids ties (which reduce the power of the test) by measuring with as fine a scale as possible. \( S \) = sum of the square differences in rank. (ref) (http://www.wisc.edu/ecology/labs/phenology/spearman.html)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Rx</th>
<th>Ry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

\( n=4 \) 6 4 3 3
Correlation of (phosphorimager $^{35}$S met) protein & mRNA

$r_p = 0.76$ for log(adjusted RNA) to log(protein)

$r_s = .74$ overall; 0.62 for the top 33 proteins & 0.56 (not significantly different) for the bottom 33 proteins
Observed (Phosphorimage) protein levels vs. Codon Adaptation Index (CAI)

Codon Adaptation Index (CAI) Sharp and Li (1987); $f_i$ is the relative frequency of codon $i$ in the coding sequence, and $W_i$ the ratio of the frequency of codon $i$ to the frequency of the major codon for the same amino-acid.

$$\ln(\text{CAI}) = \sum_{i=1,61} f_i \ln (W_i)$$
ICAT Strategy for Quantifying Differential Protein Expression.

X = H or D

Mass Spectrum and Reconstructed Ion Chromatograms.

Protein & mRNA Ratios +/- Galactose

See Ideker et al 2001
(http://arep.med.harvard.edu/pdf/Ideker01.pdf)
Protein2: Today's story & goals

- Separation of proteins & peptides
- Protein localization & complexes
- Peptide identification (MS/MS)
  - Database searching & sequencing.
- Protein quantitation
  - Absolute & relative
- **Protein modifications & crosslinking**
- Protein - metabolite quantitation
Post-synthetic modifications

- Radioisotopic labeling: PO$_4$ S,T,Y,H
- Affinity selection:
  - Cys: ICAT biotin-avidin selection
  - PO$_4$: immobilized metal Ga(III) affinity chromatography (IMAC)

Specific PO$_4$ Antibodies

Lectins for carbohydrates

- Mass spectrometry
32P labeled phosphoproteomics

Low abundance cell cycle proteins not detected above background from abundant proteins

Natural crosslinks

- Disulfides: Cys-Cys
- Collagen: Lys-Lys
- Ubiquitin: C-term-Lys
- Fibrin: Gln-Lys
- Glycation: Glucose-Lys
- Adeno primer proteins: dCMP-Ser
Crosslinked peptide Matrix-assisted laser desorption ionization Post-Source Decay (MALDI-PSD-MS)

tryptic digest of BS3 cross-linked FGF-2. Cross-linked peptides are identified by using the program ASAP and are denoted with an asterisk (9). (B) MALDI-PSD spectrum of cross-linked peptide E45-R60 (M + H+ = m/z 2059.08).
Constraints for homology modeling based on MS crosslinking distances

The 15 nonlocal throughspace distance constraints generated by the chemical cross-links (yellow dashed lines) superimposed on the average NMR structure of FGF-2 (1BLA). The 14 lysines of FGF-2 are shown in red.

See Young et al 2000, PNAS 97: 5802 (Pub) (http://www.pnas.org/cgi/content/full/97/11/5802)
Homology modeling accuracy

Swiss-model  RMSD of the test set in Angstroms

(http://www.expasy.ch/swissmod/SM_LikelyPrecision.html)
Top 20 threading models for FGF ranked by crosslinking constraint error

See Young et al. PNAS | May 23, 2000 | vol. 97 | no. 11 | 5802-5806
http://www.pnas.org/cgi/content/full/97/11/5802
Protein2: Today's story & goals

• Separation of proteins & peptides
• Protein localization & complexes
• Peptide identification (MS/MS)
  – Database searching & sequencing.
• Protein quantitation
  – Absolute & relative
• Protein modifications & crosslinking
• Protein - metabolite quantitation
Challenges for accurately measuring metabolites

• Rapid kinetics
• Rapid changes during isolation
• Idiosyncratic detection methods: enzyme-linked, GC, LC, NMR (albeit fewer molecular types than RNA & protein)
1634 Metabolite Masses
256 amino acids

Y = Frequency

598 have identical mass
e.g. Ile & Leu = 131.17

Ogata et al. (1998) Biosystems 47:119-128 KEGG
Y = RP LC retention time in min. (higher hydrophocity)

X = Mass

ESI-MS (LCQ) data pooled aminoacids
Metabolite fragmentation & stable isotope labeling

Isotopomers

See Klapa et al. Biotechnol Bioeng 1999; 62:375. Metabolite and isotopomer balancing in the analysis of metabolic cycles: I. Theory. (Pub) "accounting for the contribution of all pathways to label distribution is required, especially ... multiple turns of metabolic cycles... 13C (or 14C) labeled substrates."

MetaFoR: Metabolic Flux Ratios

Fractional 13C labeling > Quantitative 2D NMR
Why use amino acids from proteins rather than metabolites directly?

See:


Dauner et al. 2001 Biotec Bioeng 76:144
(http\www.biotech.biol.ethz.ch\sauer\pdf\2001 - Michael Dauner et al B&B.pdf)
A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations

Types of interaction models

Quantum Electrodynamics
Quantum mechanics
Molecular mechanics
Master equations

**Phenomenological rates ODE**
Flux Balance
Thermodynamic models
Steady State
Metabolic Control Analysis
Spatially inhomogenous models

subatomic
electron clouds
spherical atoms (101Pro1)
stochastic single molecules (Net1)

Concentration & time (C,t)
\[ \frac{dC_{ik}}{dt} \text{ optima steady state (Net1)} \]
\[ \frac{dC_{ik}}{dt} = 0 \text{ k reversible reactions} \]
\[ \sum \frac{dC_{ik}}{dt} = 0 \text{ (sum k reactions)} \]
\[ \frac{d(dC_{ik}/dt)}{dC_j} \text{ (i = chem.species)} \]
\[ \frac{dC_i}{dx} \]

Increasing scope, decreasing resolution
How do enzymes & substrates formally differ?

Catalysts increase the rate (&specificity) without being consumed.
Enzyme rate equations with one Substrate & one Product

\[ \frac{dP}{dt} = \frac{V(S/Ks - P/Kp)}{1 + S/Ks + P/Kp} \]

As \( P \) approaches 0:

\[ \frac{dP}{dt} = \frac{V}{1 + Ks/S} \]
Enzyme Kinetic Expressions

Phosphofructokinase

\[ v_{PFK} = \frac{v_{mx}^{PFK}}{N_{PFK}} \left( \frac{F6P}{K_{F6P}^{PFK}} \right) \left( \frac{Mg \cdot ATP}{K_{Mg \cdot ATP}^{PFK}} \right) \]

\[ N_{PFK} = 1 + L_0^{PFK} \left( \frac{1 + ATP_{free}}{K_{ATP}^{PFK}} \right)^4 \left( \frac{1 + Mg}{K_{Mg}^{PFK}} \right)^4 \left( \frac{1 + AMP}{K_{AMP}^{PFK}} \right)^4 \left( \frac{1 + F6P}{K_{F6P}^{PFK}} \right)^4 \]

Allosteric kinetic parameters for AMP, etc.
Human Red Blood Cell ODE model

ODE model
Jamshidi et al.
2000 (Pub)
(http://atlas.med.harvard.edu/gmc/rbc.html)
Red Blood Cell in Mathematica

In[4] := PFK

\frac{251.732 \text{ F6P MGATP}}{
(0.1 + \text{F6P}) \left(1 + (0.001072 (1. + 2.27273 \text{ MG})^4 (1. + 101.919 (\text{ATP-MGATP})^4) / (1. + 30.303 \text{ AMP})^4 (1. + 10. \text{ F6P})^4))
\right) (0.068 + \text{MGATP})}
\]
Protein2: Today's story & goals

• Separation of proteins & peptides
• Protein localization & complexes
• Peptide identification (MS/MS)
  – Database searching & sequencing.
• Protein quantitation
  – Absolute & relative
• Protein modifications & crosslinking
• Protein - metabolite quantitation