Shotgun Proteomic Analysis

Department of Cell Biology
The Scripps Research Institute
Biological/Functional Resolution of Experiments

Organelle

Multiprotein Complex

Cells/Tissues

Expression Analysis

Information Content

Interaction Analysis

Function
“Shotgun Proteomics”

Protein Mixture → proteolysis → Peptide Mixture

**μLC** → **MS** → **MS/MS**

Output Filtering and Re-Assembly

DTASelect → SEQUEST → 80 node Beowulf Computer Cluster
Data Processing Issues with Shotgun Proteomics

1:1 Mixture of Unlabeled/\textsuperscript{15}N-Labeled Yeast Soluble Proteins Analyzed Using a Single 12h Analysis

<table>
<thead>
<tr>
<th></th>
<th>LCQ</th>
<th>LTQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS/MS Spectra</td>
<td>18,970</td>
<td>86,950 (4.5 x)</td>
</tr>
<tr>
<td>Protein ID’s</td>
<td>559</td>
<td>891 (1.6x)</td>
</tr>
<tr>
<td>(*1 peptide confirmed w/ RelEx)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein ID’s</td>
<td>157</td>
<td>304 (1.9x)</td>
</tr>
<tr>
<td>(*2 peptides confirmed w/ RelEx)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*RelEx was used to evaluate the presence of labeled isotopomer
Processing Tandem Mass Spectra

Spectral Quality
Subtractive Analysis
Database Analysis
De Novo: Unusual, Unanticipated Features
Quantification
Filter and Assembly

Filter
LibQUEST DTASelect/Contrast
SEQUEST™ PepProb
GutenTag
RelX
DTASelect ProtProb
Data Issues

- Data quality
- How is a match determined?
  - Protease issues?
  - Validation issues?
- Posttranslational Modifications?
- Quantification?
- Sampling Issues?
### Spectral Filtering with Hand Crafted Features

<table>
<thead>
<tr>
<th>Called Good</th>
<th>Called Bad</th>
<th>%Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 GOOD</td>
<td>671</td>
<td>75</td>
</tr>
<tr>
<td>+1 BAD</td>
<td>5585</td>
<td>11475</td>
</tr>
<tr>
<td>+2/+3 GOOD</td>
<td>3166</td>
<td>348</td>
</tr>
<tr>
<td>+2/+3 BAD</td>
<td>11611</td>
<td>26684</td>
</tr>
<tr>
<td>All GOOD</td>
<td>3837</td>
<td>423</td>
</tr>
<tr>
<td>All BAD</td>
<td>17196</td>
<td>38159</td>
</tr>
</tbody>
</table>

Bern, Goldberg, MacDonald, Yates *Bioinformatics* (in press)
### Mass: 2076.64
### Database: TCA60ug06

<table>
<thead>
<tr>
<th>#</th>
<th>Peptide</th>
<th>d13C</th>
<th>d15N</th>
<th>i01</th>
<th>Reference</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.24</td>
<td>0.00</td>
<td>54.34</td>
<td>96.45</td>
<td>YKClH62W</td>
<td>KATHLDFGPGASGLVILTHRN</td>
</tr>
<tr>
<td>2</td>
<td>0.97</td>
<td>0.00</td>
<td>3.46</td>
<td>1.85</td>
<td>YKClH62W</td>
<td>YVDPYR1LHVHTLDILK1</td>
</tr>
<tr>
<td>3</td>
<td>4.89</td>
<td>0.55</td>
<td>2.45</td>
<td>3.03</td>
<td>YKClH62W</td>
<td>YSDPYR1LHVILTDILK1</td>
</tr>
<tr>
<td>4</td>
<td>5.55</td>
<td>0.00</td>
<td>2.19</td>
<td>2.05</td>
<td>YKClH62W</td>
<td>YDPRP6</td>
</tr>
<tr>
<td>5</td>
<td>2.54</td>
<td>0.61</td>
<td>2.14</td>
<td>7.37</td>
<td>YKClH62W</td>
<td>YVDDAVP2LYNR1TDTLSQ</td>
</tr>
</tbody>
</table>

**Peptide evaluation**

- G: Green
- C: Cyan
- R: Red
- Y: Yellow

**Submit Peptide**

**Locus evaluation**

- G: Green
- C: Cyan
- R: Red
- Y: Yellow

**Submit Locus**
Multi-Enzyme Digestion Procedures

- Identification of PTMs
- Sample is split into 3 aliquots
- Digest using 3 different proteases
- Mix and analyze by LC/LC-MS/MS
- Interpret spectra using SEQUEST

Diagram:
- trypsin
- elastase
- subtilisin
- MudPIT
- SEQUEST
High pH/Proteinase K Method (hpPK Method)

Overlapping Peptide Coverage

**(TM7)** gi|13794265|ref|NP_056312.1| DKFZP564G2022 protein

MAAAAWLQVL PVILLLLGAH PSPLSFFSAG PATVAAADRS
KWHIPIPSGK NYFSFGKILF RNTTIFLKFD BEPDLSLNI
TWYLKSADCY NEIYNKAEE VELYLEKLKE KRGLSNGIQT
SSKLFPQCSE LFKTQTFSGD FMHLPLILGE QKEAKENGTN
LTFIGDKTAM HEPLQTQODA PYFIVHGIG SSSKESSKEN
SLSNLFTMTV EVKGPYEYLT LEDYPLMIFF MVMCIVYVLF
GVWLFAWSAC YWRLILRIFQ WIGAVIFLGM LEKAVFYAEF
QNIRYKGEVS QGALILAEEL SAVKLSLART LVSIVSLGYG
IVKPRLGVTL HKVVVAGALY LFSGMGVL RVTGAQDGA
SLAFPLAFL DTALCWFLFI SLTQTMKLLK LRRNIVKLSL
YRHFTNLI AAVAASIFII WTTMKFRVT COSDWRELWV
DDAIWRLIFS MIFGMLVLS PNSANQLRFQ FSPLSEEEEE
DEQKEPMLKE SFEGMKMRST QCEFPGNSKV NKAQEDDLKW
VEENVPSVVT DVALPALLDS DEER MITHFE RSKME

**(TM6)** gi|14249524|ref|NP_116213.1| hypothetical protein FLJ14681

MVAACRSVAG LLPRRRRCFP ARAPLLRVAL CLLCWTPAAV
RAVPEGLWEL ETVNDSGSGL IFRKTMENST DKLTSVKSFPH
CSGPVKFTIV WHLKYHTCHN EHSNLEELFQ KHKLVSDEVDF
CHYLKNDNCW TTKNENLDNC SDSQFVFSLN NKELEIRNV
SNQERSMDVV ARTQKDGFI PIIVSKRTENT DASWNLNVSL
SMIGPHGYIS ASDWPLMIFY MVMCIVYLY GILWLTWSAC
YWKDILRFQ WIAAVIFLGM LEKAVFYSEY QNISNTGLST
QGLLIFAELI SAIKRTLARL LVIIVSLGYG IVKPRLGTVL
HRVIGGLLLY LIFAAVEGVM RVIGSNHLA VVLDIIALAV
IDSIFVWFIF ISLAQTMKTL RLKNTVKFS LYHRFKNLTVI
FAVLASIFVM GWTTKFRFIA KCDGSDMERW VDDAFWSLF
SLILIYMFIL WRPSANQRNY AFMLPDIDSS DSIEEMFTVS
ENLTEGIKLR ASKSVNQTA KATSENFDE DLK WVEENIP

WVEENVPSVTDVALPALLDS*DEER
VEENVPSVTDVALPALLDS*DEER
EENVPSVTDVALPALLDS*DEER
VENVPSVTDVALPALLDS*DEER
ENVPSSVTDVALPALLDS*DEER
VPSVTDVALPALLDS*DEER
PSSVTDVALPALLDS*DEER
LPALLDS*DEER
PALLDS*DEER

WVEENVPSVTDVALPVLVDS*DEEIMTR
IPSSFTDVALPVLVDS*DEEIMTR
PSSFTDVALPVLVDS*DEEIMTR
SFTDVALPVLVDS*DEEIMTR
TDVALPVLVDS*DEEIMTR
DVALPVLVDS*DEEIMTR
VALPVLVDS*DEEIMTR
ALPVVLDS*DEEIMTR
LPVLVDS*DEEIMTR
PVLVDS*DEEIMTR
gi|27229118|ref|NP_082129| RIKEN cDNA 0610006F02; S-adenosylmethionine-dependent methyltransferase activity [Mus musculus]

MDALVLFLQL LVLLLTLPLH LLALLSCWQP ICKTPPYPM AMLTARSYKK MESKKRELFS QIKDLKGTSG NVALLELGCG

QRVLRPGGLL FFWHRVAEPQ GSRAFLWQRV LEPTWKHIGD GCHLTRETWK DIERAQFSEV QLEWQPPPFR WLPVGPHIM

KAVK
Dimethyl Arginine Containing Peptide

Golgi Peptide

Synthetic Peptide
Shotgun Proteomic Experiments and Sampling Issues

- Based on prior studies in yeast, we know not every protein present is id’d.
- Reproducibility is good for high abundance proteins 70-80%.
- Reproducibility is not as good for low abundance proteins. 20-30%.
- Is this predictable?
Random Sampling Model for Data Dependent Acquisition

\[ K = n_L \times (1 - (1 - L / N)^S) \]

- \( n_L \) = \# of protein species at particular level
- \( L \) = abundance level
- \( N \) = total number of proteins
- \( S \) = experiments

![Graph showing the relationship between the number of MudPIT runs and the number of identified proteins.](image)

- Solid line: Experimental
- Dashed line: Semi-empirical Model
- Dotted line: Random Model
Distribution of Protein Identifications After Repeating Analysis 9 times

![Graph showing the distribution of protein identifications. The x-axis represents the number of MudPIT's, and the y-axis represents the number of proteins. The graph illustrates the number of proteins identified in all 9 experiments and in 1 experiment.](image-url)
RelX software

DTASelect Output
Peptide Sequence
LVNHFIQEFK

1) Predict m/z Range

2) Sum Signal in Range

3) Store Mass Chromatograms

4) Peak Detection

5) Correlation
Systematic Errors are Present in Samples

TSA1 = 1.335±0.015
SSA1 = 1.004±0.018
ADH1 = 0.661±0.045
Spectral Sampling for Relative Quantification

Combined Data for 6 proteins added to Yeast Soluble Cell Lysate at 4 different levels

Linear dynamic range 2-orders
Measuring small changes is not as reliable

$y = 1.0613x + 0.2608$
$R^2 = 0.9997$
Synthesis of Ribosomal Proteins in *Plasmodium falciparum*

Striking trend: almost all ribosomal proteins increase over ring to troph transition.
Standards

1. **Data formats**: McDonald et al. *MS1, MS2, and SQT - three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications*, *RCM* 2004, 18, 2162-8.

Database search standard based on:


MacCoss et al, *Probability Based Validation of Protein Identifications Using a Modified SEQUEST Algorithm*, *Analytical Chemistry* 74, 5593-5599 (2002). Normalized Scores
Standards

- 4. Standards should not prevent innovation
  - Data formats should be practical e.g. storage space

- 7. Data processing tools should be transparent and validated, e.g. published

- 8. Data for publication: information to support biological conclusions - sequences of peptides id’d.

- 9. Archiving data: biological conclusions should be the most important part of the experiment
Probability Distributions

Number of Spectra

Probability Scores

- Charge State +1
- Charge State +2
- Charge State +3
Single Spectral Matches are Problematic: How to tell if they are correct

- Searches determine closeness of fit based on some measure: Compare matches with different programs
- **Probability scoring**: \( P = \text{random match} \) based on frequency of fragment ions in database
- **SEQUEST**: XCorr measures how close the spectrum fits to ideal spectrum
- Manual validation, experimental validation
- *de novo* interpretation
Multi-Enzyme Digest

- Sample is split into 3 aliquots
- Digest using 3 different proteases
Multi-Enzyme Digest

- Sample is split into 3 aliquots
- Digest using 3 different proteases
- Mix and analyze by LC/LC-MS/MS
- Interpret spectra using SEQUEST
Properties of Data Dependent Data Acquisition

Most invariant property is spectral copy number

![Bar chart showing % of Proteins Identified, % of Unique Peptides, and % of Spectrum Copies for different data sets P1, P2, P3, and P4.](chart.png)
**GutenTag: Partial de novo Sequencing of Tandem Mass Spectra**

- Database searching assumes minimal errors in the database and sequence variations between strains, individuals and species.
- Modifications need to be specified in database searches, so unanticipated modifications will be missed.
- Partial *De novo* analysis of tandem mass spectra in large-scale can identify peptides containing sequence variations and unanticipated modifications.
1. Generate sequence tags
GutenTag

- 6170 tandem mass spectra: LC/LC/MS/MS analysis of simple digested protein mixture
- 1987 spectra matched by SEQUEST
- 1328 spectra matched by GutenTag
- 766 partial matches suggesting modifications and sequence variations
- Total matching spectra by GutenTag: 2,094
- Partial de novo will extend identifications
- Software is Automated and Large-Scale
Improved Spectral Quality Effects Peptide Identification
Infusion of 1 pmol/µl Angiotensin I

**LCQ-Classic**
761 of 1000 MS/MS Spectra Matched the Correct Sequence

**LTQ**
970 of 1000 MS/MS Spectra Matched the Correct Sequence
Database Searching with Tandem Mass Spectra

- The goal is to identify peptides using MS/MS spectra and amino acid sequence databases.

- Develop a probabilistic model that establishes a relationship between the database sequences and the spectrum to complement quantitative measures of closeness-of-fit.

- Develop non-empirical probabilistic measures using cross-correlation measurements.
Probability Model

• Null hypothesis: All fragment matches to MS/MS spectrum are by random.

• $N, K$ – number of all fragments and all fragment matches, respectively. $N_1$ is the number of fragments of a particular peptide which has $K_1$ matches.

\[
P_{K,N}(K_1, N_1) = \frac{C^K_{K_1} \cdot C^{N_1-K_1}_{N-K}}{C^{N_1}_N}
\]

• We seek an amino acid sequence that has the smallest probability of being a random match.
Model Hypergeometric Distributions

$N, K$ - number of all fragments and all fragment matches, respectively. $N_1$ is the number of fragments of a particular peptide which has $K_1$ matches.
Significance of Peptide Identification

• Not all identifications are significant: poor quality spectra of peptides, incomplete peptide fragmentation, inaccuracies in database, posttranslational modifications, MS/MS of chemical noise and non-peptide molecules.

• Significance of a match (P_value) is also obtained from the hypergeometric distribution.
Significance of a Match

Significance of $K_1 = 14$

$N = 1000$
$K = 300$
$N_1 = 30$
Yeast Database

Number of Fragment Matches

- Predicted Distribution
- Observed Frequency

N = 5284150
K = 569160
N_1 = 40
Mass/Charge State Dependence

- Scores that use closeness of fit measures can artificially inflate with weight/mass.

- This complicates use of uniform criteria for identification.

- Probabilities generated by hypergeometric distribution are charge/weight independent.
Cross-Correlation Score Distribution

Number of Spectra

Cross-Correlation Scores

- XCorr+1
- XCorr+2
- XCorr+3
Database Dependence

- The probability inferred from the hypergeometric distribution is in principle database dependent.

- However, the dependency is very weak.
Pep_Probe Summary

- Implements 4 scoring schemes: hypergeometric, poisson, maximum likelihood and cross-correlation. Sorts results either by hypergeometric or cross-correlation scores.
- No enzyme specificity is assumed.
- Reports significance of each match.
- Can search for posttranslational modifications to three different amino acids.
- Has been implemented to run on a standalone or compute clusters.
- Runs on heterogeneous cluster of computers, in WINDOWS and LINUX platforms.
Processing Tandem Mass Spectra

1. Spectral Quality
   - Eliminate Poor Quality Spectra, Score Quality of Other spectra

2. Database Analysis
   - Multiple Methods With Different Selectivity's

3. Analyze Unusual or Unanticipated Features
   - De novo or partial De novo analysis

4. Assemble and Annotate Data
   - Biological Significance
Increased Data Production Requires Automated Data Analysis

**LIBQUEST**
- Library Searching
- Comparative Analysis
- Subtractive Analysis


**SEQUEST & Pep_Prob**
- Database Search

**SEQUEST-SNP**
- Variant Search
  - Related Sequence SNP Analysis
  - Mutation Analysis

**GutenTag**
- De Novo Sequencing
- Alternate Splicing
- Unantic. Mod.
- Unknown ORFs

**Probability for Protein Identification**

**Spectral Quality**

Bioinformatics (in press)

**Relative Abundance**

**m/z**

**Existing Sequence**

**PTM**

**Probability for Protein Identification**

**Database Search**

**Existing Sequence**

**PTM**

**Related Sequence SNP Analysis**

**Mutation Analysis**

**Alternate Splicing**

**Unantic. Mod.**

**Unknown ORFs**

**Yates et al., JASMS 5, 976 (1994)**


**Sadygov and Yates, Anal. Chem (in press)**

**Link et al. Nature Biotech. 17, 676-682 (1999)**

**Tabb et. al. J. Proteome Res. 1, 26, (2002)**

Data Considerations

• Different types of experiments
• Different types of data analysis
Integrated Multi-Dimensional Liquid Chromatography

RP  SCX  RP

100 micron FSC

5 μM
100-300 nL/min

Waste

hV
Comprehensive Analysis of Complex Protein Mixtures

• Protein Identification: *What’s there*
• Post Translational Modifications: *Regulation*
• Quantification: *Dynamics*
• Proteomic Data to Knowledge: *Genetics, RNAi, siRNA*

Translation of technology development into biological discovery