Working at the Interface between Proteomics & Informatics at Indiana University

Randy J. Arnold

Proteomics Manager
National Center for Glycomics & Glycoproteomics

1533
March 23, 2006
Overview

• Introduction to Proteomics
• Proteomics Platforms / Approaches
• Peptide Fragmentation – Predictable?
• Biology-based Separations
• Influence of Informatics
What is Proteomics?

Proteomics is the study of protein expression, regulation, modification, and function in living systems for understanding how living systems use proteins. Using a variety of techniques, proteomics can be used to study how proteins interact within a system, or how proteins change due to applied stresses.

Proteomics requires the use of advanced measurement techniques with an emphasis on separations and mass spectrometry.
The Central Dogma of Life

DNA - genetic information

RNA - instructions for proteins

Proteins - structure and function of living cells

Glycans - sugars attached to other biomolecules
The Basics - Proteins

Polymers of amino acids (20 naturally occurring)

R_1 = CH_3  
R_2 = CH_3OH

Alanine  
Serine  
AlanylSerine

Ala  
Ser  
Ala-Ser

A  
S  
AS
The Basics – Protein Structure

primary structure – amino acid sequence

DRLEFI VTALLKPW

N-terminus  C-terminus

tertiary structure – protein folding

secondary structure – local spatial arrangement

β-sheet

α-helix

quaternary structure – multimeric complexes

http://www.path.cam.ac.uk/~mrc7/igs/mikeimages.html

http://mcl1.ncifcrf.gov/integrase/asv_secstr.html

http://www.path.cam.ac.uk/~mrc7/igs/mikeimages.html
Proteomics Approaches

**Bottom-up**
- Lyse cells
- Digest proteins
- Separate proteins
- Peptide MS & MS/MS
- LC resolution
- MS sensitivity
- PTM ID
- Sequence Coverage

**Top-down**
- Lyse cells
- Separate proteins
- Digest proteins
- Separate peptides
- Peptide MS & MS/MS
- Protein MS
- Protein MS/MS

Advantages
- LC resolution
- MS sensitivity
- Sequence Coverage
- PTM ID
Chromatographic Separation

Reverse-phase chromatography

75 µm

5 µm

100 Å pores

hydrophobic stationary phase
Chromatographic Separation

Molecular Interactions

1) 95% aqueous / 5% organic mobile phase
2) 75% aqueous / 25% organic mobile phase
3) 65% aqueous / 35% organic mobile phase
Electrospray Ionization

2002 Nobel Prize in Chemistry awarded to John Fenn for advancing electrospray

to 10-port valve

vent

analytical column
C18-5µm-100Å, 75µm x 15cm

Sample trapping column
C18-5µm-200Å, 75µm x 1.5cm

Pt lead (2.5kV)

Mass Spectrometer

Compliments of Dr. Myeong Hee Moon
Mass Spectrometry

Time-of-Flight MS

1) Ions enter source region, accelerated toward reflectron.

2) Ions separate in space based on their relative mass-to-charge (m/z).

3) Ions reverse path in reflectron.

4) Ions impact detector.

Fligh time

Signal

0 kV 0 kV

+20 kV
Instrumentation

- nanoLC-QIT MS/MS
  - LC Packings / ThermoFinnigan

- capLC-QTOF MS/MS
  - Waters / Micromass

- 2-D gel spot cutter
  - BioRad

- MALDI-cid-TOF/TOF MS/MS
  - Applied Biosystems
nanoLC-MS/MS Data

Base Peak Chromatogram

Mass Spectrum at 48.08 min.

Tandem MS of m/z 540.4

“b” ions  b₂  b₃  b₄  b₅

T A A Y V N A I E K

“y” ions  y₅  y₆  y₇  y₈

b₂ b₃ b₄ b₅
Database Searching
(Informatics)

Database (SwissProt)
Actin
MYTCVPIASEQUENCEMI MEWTPQSDLI RPTVCI MNERCVGGPYI LCMTEND
Amylase
DSLI KRNYTI PMCSQI RECNHI PLMTRCH GYYKWSI ALAI NTQSFGI VRI VAMNLPS SCRTI VGHWEDRI CTMQNCI SPPEKELIA VARGTSP

Results

Proteins found
Hemoglobin, beta chain
<table>
<thead>
<tr>
<th>Pept.</th>
<th>Mass</th>
<th>Score</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>738.84</td>
<td>41</td>
<td>HLDNLK</td>
</tr>
<tr>
<td>2</td>
<td>912.01</td>
<td>61</td>
<td>VHLTDAEK</td>
</tr>
<tr>
<td>3</td>
<td>915.06</td>
<td>56</td>
<td>AAVNGLWGK</td>
</tr>
<tr>
<td>4</td>
<td>1090.24</td>
<td>41</td>
<td>VI NAFNDGLK</td>
</tr>
<tr>
<td>5</td>
<td>1122.33</td>
<td>62</td>
<td>WAGVASALAHK</td>
</tr>
<tr>
<td>6</td>
<td>1218.42</td>
<td>70</td>
<td>LVI NAFNDGLK</td>
</tr>
</tbody>
</table>

Database searching software
MASCOT®
Identification approaches

- **SEQUEST**\(^1\) – cross-correlation
  - \( b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8 b_9 b_{10} b_{11} \)
  - \( y_{11} y_{10} y_9 y_8 y_7 y_6 y_5 y_4 y_3 y_2 y_1 \)

- **Mascot**\(^2\) – probability-based scoring

Intensity Prediction

• Kinetic model$^1$
  – model based on chemical mechanism (CID)
  – universally applicable?

• Decision tree$^2$
  – +2 only
  – b & y ions only

A Machine Learning Approach to Predicting Peptide Fragmentation Spectra

Randy J. Arnold, Narmada Jayasankar, Divya Aggarwal, Haixu Tang, and Predrag Radivojac

Collaborative project between the Department of Chemistry and the School of Informatics

Presented at PSB in Jan. 2006
Peptide charge state

MLQLVEESKDAGIR

+2

+3
Method – classification model

• Precursor sequence $S$
  – Charge $q_S \in \{+2, +3\}$

• Estimate probabilities: $P(l(i) \geq t \mid S, q_S)$
  – where $l(i)$ is peak intensity of any fragment ion:
    – $i \in \{\text{precursor-H}_2\text{O}, b, b\text{-H}_2\text{O}, b\text{-NH}_3, b\text{-H}_2\text{O-NH}_3, y, y\text{-H}_2\text{O}, y\text{-NH}_3, y\text{-H}_2\text{O-NH}_3, b^{2+}, y^{2+}\}$
    – $t = 1\%$ of total intensity of the spectrum
## Datasets

<table>
<thead>
<tr>
<th>Ion</th>
<th>Doubly charged precursors</th>
<th>Triply charged precursors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positives</td>
<td>Negatives</td>
</tr>
<tr>
<td>precursor – H₂O</td>
<td>239</td>
<td>1484</td>
</tr>
<tr>
<td>b</td>
<td>5210</td>
<td>16916</td>
</tr>
<tr>
<td>b – H₂O</td>
<td>1700</td>
<td>20426</td>
</tr>
<tr>
<td>b – NH₃</td>
<td>678</td>
<td>21448</td>
</tr>
<tr>
<td>b – H₂O – NH₃</td>
<td>249</td>
<td>21877</td>
</tr>
<tr>
<td>b²</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>y</td>
<td>9323</td>
<td>12802</td>
</tr>
<tr>
<td>y – H₂O</td>
<td>431</td>
<td>21695</td>
</tr>
<tr>
<td>y – NH₃</td>
<td>286</td>
<td>21840</td>
</tr>
<tr>
<td>y – H₂O – NH₃</td>
<td>145</td>
<td>21981</td>
</tr>
<tr>
<td>y²</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Features – 202 in total

Tryptic peptide, \( n \) residues long, from human p53

- amino acids at positions \( k, k - 1, k + 1, k + 2 \)
- amino acid at position 1
- amino acid compositions for both fragment ions
- length and mass of each fragment ion
- various physical / chemical properties
Model selection & training

- Ensembles of two-layer feed-forward neural networks
- Experimented with network architecture
- Datasets were high-dimensional & class-imbalanced
- Different set of negatives for each network in the ensemble
- Applied feature selection & PCA
- Separate validation set for each individual model (20% of training set)
- Each ensemble contained 30 neural networks
## Performance evaluation (ROC)

<table>
<thead>
<tr>
<th>Ion</th>
<th>Doubly charged precursors</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>sn</td>
<td>sp</td>
<td>acc/AUC</td>
<td>sn</td>
<td>acc/AUC</td>
</tr>
<tr>
<td>precursor – H₂O</td>
<td>72.0</td>
<td>60.8</td>
<td>66.4/70.7</td>
<td>81.3</td>
<td>68.5</td>
<td>74.9/79.7</td>
</tr>
<tr>
<td>b</td>
<td>80.4</td>
<td>75.4</td>
<td>77.9/85.8</td>
<td>80.6</td>
<td>71.9</td>
<td>76.3/84.6</td>
</tr>
<tr>
<td>b – H₂O</td>
<td>76.8</td>
<td>76.3</td>
<td>76.5/84.6</td>
<td>76.2</td>
<td>60.2</td>
<td>68.2/76.8</td>
</tr>
<tr>
<td>b – NH₃</td>
<td>75.8</td>
<td>76.0</td>
<td>75.9/82.8</td>
<td>76.9</td>
<td>65.0</td>
<td>70.9/78.6</td>
</tr>
<tr>
<td>b – H₂O – NH₃</td>
<td>69.1</td>
<td>64.6</td>
<td>66.8/73.1</td>
<td>81.8</td>
<td>51.9</td>
<td>66.9/68.1</td>
</tr>
<tr>
<td>b²⁺</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>88.4</td>
<td>75.8</td>
<td>82.1/88.5</td>
</tr>
<tr>
<td>y</td>
<td>84.7</td>
<td>79.3</td>
<td>82.0/89.5</td>
<td>88.9</td>
<td>79.1</td>
<td>84.0/91.4</td>
</tr>
<tr>
<td>y – H₂O</td>
<td>66.4</td>
<td>66.2</td>
<td>66.3/72.2</td>
<td>82.6</td>
<td>56.5</td>
<td>69.6/73.0</td>
</tr>
<tr>
<td>y – NH₃</td>
<td>70.3</td>
<td>70.8</td>
<td>70.6/79.0</td>
<td>81.2</td>
<td>59.8</td>
<td>70.5/77.8</td>
</tr>
<tr>
<td>y – H₂O – NH₃</td>
<td>60.7</td>
<td>51.1</td>
<td>55.9/56.5</td>
<td>83.2</td>
<td>54.3</td>
<td>68.7/69.6</td>
</tr>
<tr>
<td>y²⁺</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.9</td>
<td>72.6</td>
<td>80.2/86.8</td>
</tr>
</tbody>
</table>
Amino Acid preferences

Figure 2. The amino acid preferences for peptide fragmentation. The frequencies of observing ion types (b-, y- or b2, y2) were plotted in grey scaling from 0 (white) to 1 (black). The rows indicate amino acid on the left-hand side, while the columns indicate amino acids on the right-hand side of the cleavage site.

Similar results to from studies by Smith, Wysocki, and others
+2 peptides w/o Proline

GYSFTTTAER

mobile proton
+2 peptides w/o Proline

VFDKDGNGYISAAELR
+2 peptides w/ Proline

GAAQNIIPASTGAAK

mobile proton
+2 peptides w/ Proline

TYFSHIDVSPGSAQVK
+3 peptides w/o Proline

VFVKGNGYIISAAELR

mobile proton
+3 peptides w/o Proline

HRDTGILDSIGR

Experimental Spectrum

Predicted Spectrum
+3 peptides w/ Proline

GSHSQTPSPGALPLGR

mobile proton
+3 peptides w/ Proline
HVLSGTLGVPEHTYR
# Peptide ID - Scoring

vs. 500 random sequences; 25 in each category

<table>
<thead>
<tr>
<th>Scoring scheme</th>
<th>Doubly charged precursors</th>
<th></th>
<th></th>
<th>Triply charged precursors</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mobile proton</td>
<td>Non-mobile proton</td>
<td>Mobile proton</td>
<td>Non-mobile proton</td>
<td>Mobile proton</td>
<td>Non-mobile proton</td>
</tr>
<tr>
<td>w/o Pro</td>
<td>w Pro</td>
<td>w/o Pro</td>
<td>w Pro</td>
<td>w/o Pro</td>
<td>w Pro</td>
<td>w/o Pro</td>
</tr>
<tr>
<td>New</td>
<td>.32 ± .03</td>
<td>.26 ± .04</td>
<td>.30 ± .03</td>
<td>.24 ± .04</td>
<td>.13 ± .03</td>
<td>.14 ± .04</td>
</tr>
<tr>
<td>Simple</td>
<td>.22 ± .02</td>
<td>.14 ± .03</td>
<td>.23 ± .02</td>
<td>.15 ± .02</td>
<td>-.01 ± .02</td>
<td>-.03 ± .02</td>
</tr>
<tr>
<td>New</td>
<td>1.1 ± 0.1</td>
<td>1.4 ± 0.2</td>
<td>1.1 ± 0.1</td>
<td>1.5 ± 0.2</td>
<td>1.8 ± 0.7</td>
<td>1.5 ± 0.2</td>
</tr>
<tr>
<td>Simple</td>
<td>1.1 ± 0.1</td>
<td>1.4 ± 0.2</td>
<td>1.0 ± 0.1</td>
<td>1.3 ± 0.2</td>
<td>9.0 ± 1.8</td>
<td>19.0 ± 4.5</td>
</tr>
</tbody>
</table>
Fragmentation Prediction
Conclusions

• Peptide MS/MS spectra are predictable using a neural network machine learning approach

• Observation of known mechanisms (enhanced fragmentation N-term to Pro)

• Potential to study subtle effects (mobile vs. no mobile proton)
“Undersampling” Problem
Incomplete proteome coverage

Mass spectrum at 72.31 min.

AMGLPEDLI QK
from Fatty acid-binding protein

VVDLLAPYAK
from ATP synthase beta chain

APAAI GAYSQAVLVDR
from 14.5 kDa translational inhibitor protein

AVQEV LVTHG EDTADRPPVPI FK
from Cytochrome P450 2D5
Organelle Enrichment

Compartmentalization:
organization of eukaryotic cells into organelles

- Nucleus
- Mitochondria
- E.R.
- Golgi
- Vacuoles
- Cytoskeleton
1. Lyse cells
2. Low speed centrifugation
3. Medium speed centrifugation
4. High speed centrifugation
## Organelle Enrichment

**Mouse Brain** tissue - Trypsin digest LC-IT-MS/MS (2x)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Score &gt;28</th>
<th>1+ Peptides</th>
<th>2+ Peptides</th>
<th>3+ Peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole cells</td>
<td>674</td>
<td>227</td>
<td>115</td>
<td>84</td>
</tr>
<tr>
<td>Nuclear</td>
<td>896</td>
<td>266</td>
<td>154</td>
<td>94</td>
</tr>
<tr>
<td>Mitochondrial</td>
<td>871</td>
<td>278</td>
<td>149</td>
<td>101</td>
</tr>
<tr>
<td>Microsomal</td>
<td>439</td>
<td>162</td>
<td>72</td>
<td>42</td>
</tr>
<tr>
<td>Cytosol</td>
<td>815</td>
<td>255</td>
<td>114</td>
<td>80</td>
</tr>
<tr>
<td>Combined</td>
<td><strong>1501</strong></td>
<td><strong>538</strong></td>
<td><strong>244</strong></td>
<td><strong>164</strong></td>
</tr>
</tbody>
</table>
### Protein Results Parser 3.0

http://newweb.chem.indiana.edu/facilities/proteomics/parser/main.htm
“Cytoplasmic” Proteins

- Phosphoglycerate kinase (44.4 kDa) – glycolysis
  - Peptide Count: 0.75 / 0.5 / 0.0 / 10.25 (Nucl / Mito / Micr / Cyto)
- L-lactate dehydrogenase (36.5 kDa) – anaerobic glycolysis
  - Peptide Count: 3.0 / 2.75 / 1.5 / 11.25
- Glyceraldehyde-3-phosphate dehyd (35.7 kDa) – glycolysis
  - Peptide Count: 4.25 / 4.0 / 10.25 / 9.0
- Dihydropyrimidinase related protein – 2 (62.2 kDa) – axon elaboration?
  - Peptide Count: 7.5 / 8.25 / 11.75 / 16.25
  - Subcellular Location: tightly, but noncovalently, associated with membranes

Data averaged from 4 LC-MS/MS analyses of two rat hippocampus tissue samples.
Protein subcellular location noted as “cytoplasmic” in SwissProt/TrEMBL (http://us.expasy.org/)
## Ribosomal Proteins

Peptide counts: \(\text{Nucl.} / \text{Mito.} / \text{Micr.} / \text{Cyto.}\)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>L18 (21.5 kDa)</td>
<td>0.0</td>
<td>0.0</td>
<td>3.0</td>
<td>0.0</td>
</tr>
<tr>
<td>S6 (28.7 kDa)</td>
<td>0.0</td>
<td>0.0</td>
<td>2.75</td>
<td>0.0</td>
</tr>
<tr>
<td>L12 (17.8 kDa)</td>
<td>0.0</td>
<td>0.0</td>
<td>2.5</td>
<td>0.0</td>
</tr>
<tr>
<td>37 other proteins</td>
<td>0.0</td>
<td>0.0</td>
<td>30.5</td>
<td>0.0</td>
</tr>
<tr>
<td>(\text{SA ?} (32.7 \text{ kDa}))</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>S12 (14.4 kDa)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>S28 (7.8 kDa)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.75</td>
</tr>
</tbody>
</table>
“Marker” Proteins

Peptide counts:  **Nucl.** / **Mito.** / **Micr.** / **Cyto.**

- **ATP synthase** – **mitochondrial inner membrane**
  - alpha (58.8 kDa): 13.75 / 13.25 / 10.0 / 0.0
  - beta (56.3 kDa): 11.25 / 14.5 / 6.0 / 0.25

- **Glutamate dehydrogenase** (61.4 kDa)
  - mitochondrial matrix: 2.75 / 4.0 / 3.75 / 3.0

*incomplete separation?*

*damaged mitochondria?*
Organelle Enrichment Conclusions

• Fast, semi-quantitation using peptide counts
• Nuclear / Mitochondrial separation challenging
• Biologically relevant information found for cytoplasmic and ribosomal proteins
Proteomics Needs Informatics for…

• Locating peaks in 2 or more dimensions
• MS/MS spectra interpretation
• Protein/Peptide quantification
• Peptide detectability
• Experimental data → Biological information
  – enzyme or pathway regulation
  – disease susceptibility
  – drug efficacy
# Acknowledgments

## Faculty
- Milos Novotny
- David Clemmer
- James Reilly
- Stephen Jacobson

## Collaborations
- William McBride
- Frank Witzmann
- Meei-Huey Jeng
- Linda Malkas
- Haixu Tang
- Robert Hickey
- John Foley
- Ken Nephew
- Richard DiMarchi
- Martha Oakley
- Pedja Radivojac
- J.-T. Zhang

## Scientists / Post-docs
- Yehia Mechref
- Myeong Hee Moon
- Petra Hrncirova
- Iveta Klouckova
- Steve Valentine
- Weidong Cui
- Dariusz Janecki
- Wendy Strother-Robinson
- Li-Yun Chang

## Graduate students
- Chemistry
  - John Taraszka
  - Matt Thompson
  - Arugadoss Devakumar
  - Rená Sowell
  - Ruwan Kurulugama
  - Zhiyin (Ella) Xun
- Informatics
  - Kiran Annaiah
  - Divya Aggarwal
  - Narmada
  - Jayasankar
  - C.J. Fleck

## Undergraduate students
- Chet Linson
- Amy Ho

## Resources
- Information Technology Group

## $$ Funding $$
- INGEN – Indiana Genomics Initiative
  - State of Indiana 21st Century
  - NIH – National Center for Glycomics & Glycoproteomics