Proteomics and Metabolomics

Advanced Genome Analysis

April 21st, 2015 - UCHSC

Kirk Hansen, University of Colorado
Overview

- Mass Spectrometry
- Metabolomics
- Proteomics
- Qualitative Proteomics
- Data analysis - method to improve proteome coverage
3 main components of mass spectrometer:

- Electron impact
- Fast atom bombardment
- Matrix-assisted laser desorption/ionization
- Electrospray ionization

- Quadrupole
- Time-of-flight
- Ion trap
- Fourier transform ion cyclotron resonance
- Hybrid instruments

- Electron multiplier
- Image current
Electrospray ionization (ESI)
Why does the bottle label say: Mol. Wt.: 46.07?
Stable isotopes of most abundant elements of peptides

<table>
<thead>
<tr>
<th>Element</th>
<th>Mass</th>
<th>Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>1.0078</td>
<td>99.985%</td>
</tr>
<tr>
<td></td>
<td>2.0141</td>
<td>0.015</td>
</tr>
<tr>
<td>C</td>
<td>12.0000</td>
<td>98.89</td>
</tr>
<tr>
<td></td>
<td>13.0034</td>
<td>1.11</td>
</tr>
<tr>
<td>N</td>
<td>14.0031</td>
<td>99.64</td>
</tr>
<tr>
<td></td>
<td>15.0001</td>
<td>0.36</td>
</tr>
<tr>
<td>O</td>
<td>15.9949</td>
<td>99.76</td>
</tr>
<tr>
<td></td>
<td>16.9991</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>17.9992</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Metabolite assignment

Lactate = 89.023

1.0034 Da

$^{13}\text{C} - ^{12}\text{C}$ mass difference

1.1%

Natural abundance of $^{13}\text{C}$
Monoisotopic versus average mass

\[ \text{C}_{48} \text{H}_{82} \text{N}_{16} \text{O}_{17} \]
Mass accuracy and resolving power

\[ \frac{m}{\Delta m} \text{ (PWHH)} = \text{resolving power} \]

\[ \frac{415.98}{0.04} = 10,400 \]

**Accuracy** = \[
\frac{m_{\text{theor}} - m_{\text{exp}}}{m_{\text{theor}}} = \frac{0.01}{415.99} = 0.0024\% \text{ error} = 24 \text{ ppm}
\]
Molecules and mass spectra

Similar structures may or may have not similar mass spectra

Electron impact (70 eV) mass spectra; Source: NIST05; Created using structure similarity search in NIST MS Search program
Molecules and mass spectra

Similar mass spectra may or may have not similar structures

Electron impact (70 eV) mass spectra; Source: NIST05; Created using spectral similarity search in NIST MS Search program
Why small molecules...?

- >95% of all diagnostic clinical assays test for small molecules
- 89% of all known drugs are small molecules
- 50% of all drugs are derived from pre-existing metabolites
- 30% of identified genetic disorders involve diseases of small molecule metabolism
- Small molecules serve as cofactors and signaling molecules to thousands of proteins
The Metabolome is Connected to All Other “Omes”

- Small molecules (i.e. AMP, CMP, GMP, TMP) are the primary constituents of the genome & transcriptome
- Small molecules (i.e. the 20 amino acids) are the primary constituents of the proteome
- Small molecules (i.e. lipids) give cells their shape, form, integrity and structure
- Small molecules (sugars, lipids, AAs, ATP) are the source of all cellular energy
- Small molecules serve as cofactors and signaling molecules for both the proteome and the genome
  - *The genome & proteome largely evolved to catalyze the chemistry of small molecules*
Sensitivity = improved detection of low level compounds
Specificity = assignment of individual metabolites (not just classes)

Metabolomics: from NMR to UPLC-MS
Metabolome
= the total metabolite pool

All low molecular weight (< 1500 Da)

Sugars, Nucleosides
Organic acids, Ketones
Aldehydes, Amines
Amino acids
Small peptides
Lipids, Steroids
Terpenes
Alkaloids
Drugs (xenobiotics)
The complexity of the metabolome

Human Metabolomes

- **3100 (T3DB)**
  - Toxins/Env. Chemicals

- **1000 (DrugBank)**
  - Drug metabolites

- **30000 (FooDB)**
  - Food additives/Phytochemicals

- **1450 (DrugBank)**
  - Drugs

- **8500 (HMDB)**
  - Endogenous metabolites
Molecular cartography of the human skin surface in 3D

Amina Boussimani\textsuperscript{a,1}, Carla Porto\textsuperscript{b,3}, Christopher M. Rath\textsuperscript{a,1}, Mingxun Wang\textsuperscript{b}, Yurong Guo\textsuperscript{a}, Antonio Gonzalez\textsuperscript{d,e,2}, Donna Berg-Lyons\textsuperscript{a,4}, Gail Ackermann\textsuperscript{e,6}, Gitte Julie Moeller Christensen\textsuperscript{a,7}, Teruaki Nakatsuji\textsuperscript{b}, Lingjuan Zhang\textsuperscript{b}, Andrew W. Borkowski\textsuperscript{b}, Michael J. Meehan\textsuperscript{b}, Kathleen Dorrestein\textsuperscript{a}, Richard L. Gallo\textsuperscript{b}, Nuno Bandeira\textsuperscript{a,8,9}, Rob Knight\textsuperscript{e,6,10}, Theodore Alexandrov\textsuperscript{a,4,11,12}, and Pieter C. Dorrestein\textsuperscript{a,4,12}

\textsuperscript{a}Department of Dermatology, San Diego, CA, USA; \textsuperscript{b}Department of Microbiology, San Diego, CA, USA; \textsuperscript{c}Department of Microbiology and Immunology, Stanford University, Stanford, CA, USA; \textsuperscript{d}Department of Microbiology, University of California, San Diego, CA, USA; \textsuperscript{e}Department of Physics, University of California, San Diego, CA, USA; \textsuperscript{f}Department of Chemistry, University of California, San Diego, CA, USA; \textsuperscript{g}Department of Cell Biology and Molecular Genetics, University of California, San Diego, CA, USA; \textsuperscript{h}Department of Medicine, University of California, San Diego, CA, USA; \textsuperscript{i}Department of Biomedical Engineering, University of California, San Diego, CA, USA; \textsuperscript{j}Department of Computer Science, University of California, San Diego, CA, USA; \textsuperscript{k}Department of Bioinformatics, University of California, San Diego, CA, USA.
• Accurate intact mass (< 5 ppm) against KEGG pathway database
• Retention time match (CV < 5 seconds) against standards (>630 standard library)
• Chemical formula determination using isotopic patterns
• MS/MS analysis and transition fingerprints if further disambiguation is needed
• $^{13}\text{C}$ spiked in internal standards for CV calculation
Study Design: RBC Storage in AS-3

**AS-3 (Nutricel)**

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaCl</td>
<td>70 mM</td>
</tr>
<tr>
<td>NaH$_2$PO$_4$</td>
<td>23 mM</td>
</tr>
<tr>
<td>Citric acid</td>
<td>2 mM</td>
</tr>
<tr>
<td>Na-citrate</td>
<td>23 mM</td>
</tr>
<tr>
<td>Adenine</td>
<td>2 mM</td>
</tr>
<tr>
<td>Glucose</td>
<td>55 mM</td>
</tr>
<tr>
<td>pH</td>
<td>5.8</td>
</tr>
</tbody>
</table>

PRE  | POST  | LR  | D1  | D7  | D14 | D21 | D28 | D35 | D42 |

5 donors
11 time points

2 fractions

Supernatant

RBC

+ and – ion mode

220 analytical runs

Relative Abundance vs. Time (min)
Class example: Amino Acids

Robust and Accurate
Compound classes and pathways

M$^{15}$ – HILIC Sample Analysis

- FA
- Sugars
- OA
- Aminoacids
- Phosphate compounds
- Polyamines
Selectivity and Sensitivity

Selective

Sensitive

**D**
Glutamate Calibration curve
- RSQ = 0.998
- LOD = 3 pg
- LOQ = 9 pg (61 fmol)

**E**
Calibration curve (zoom in)
- LOQ = 9 pg; 61 fmol
3k to >10k features detected

Higher in MOLM-13 + Flt3 inhibitor
Node size = abundance

Higher in MOLM-13 Untreated Controls
Collaboration with James DeGregori
Multivariate Analyses

**PLS-DA**

RBC Extracts

- PC1 (31.2%)
- PC2 (8%)

- PRE
- D42
- POST
- D35
- 3h
- D28
- D21
- D14
- D1
- LR

Supernatants

- PC1 (33.5%)
- PC3 (6.3%)

- PRE
- POST
- 3h
- LR
- D1
- D7
- D14
- D21
- D28
- D35
- D42

**Hierarchical Clustering Analysis**
Loading plots inform on the pathways affected by storage lesions.

**RBC Extracts**

- PC1 (31.2%)
- PC2 (8%)

**Supernatants**

- PC1 (33.5%)
- PC3 (6.3%)

### PLS-DA

- **PRE**
- **D42**
- **D35**
- **POST**
- **D3**
- **D1**
- **3h**
- **LR**
- **D21**
- **D14**
- **D28**
- **D7**

### Metabolites

- **Orthophosphate**
- **D-Hexose 6-phosphate**
- **Glutamine**
- **Adenine**
- **Phosphonopyruvate**
- **Cys-Gly**
- **CTP**
- **Pyruvate**
- **Glyceraldehyde 3-phosphate**
- **3-Phospho-D-erythrose**
- **Diphosphate**
- **alpha-D-Glucosamine**
- **Iminoerythrose 4-phosphate**
- **dGMP**
- **AMP**
- **Inosine**
- **Hypoxanthine**
- **Threonate**
- **Glyceraldehyde**
Glutathione Homeostasis: synthesis

- SER
- GLY
- CYS
- GSH
- GLN
- CYS
- GLU
- 5-OXO
- GLN
- LAC
- SARC
- PYR
- ALA
- OXAL
- ASP
- meTHF
- THF

- Pre
- Post
- LR
- 3h
- D1
- D7
- D14
- D21
- D28
- D35
- D42

RBCs
Supernatants
Limitations of metabolomics

• High biological variance in metabolite levels (i.e., the variation between genetically identical plants grown in the same conditions)

• Unlike nucleic acids and proteins, metabolites have a vast range of chemical structures and properties. Their molecular weights span two orders of magnitude (20–2000 Da). Therefore no single extraction or analysis method works for all metabolites.

• The concentrations of various metabolites can vary dramatically from mM to pM concentrations.

• Some metabolites are labile and won’t survive extraction and analysis
Traditional Proteomics

Use antibodies to quantify proteins
Standard Proteomics Workflow

Cells or tissue

Step 1
Extraction

Proteins

Step 2
Digestion

Peptides

Step 3
PTM

UHPLC

Step 4
ESI

Mass spectrometer

Chromatogram

MS1

Intensity vs. m/z

MS2

Intensity vs. m/z

Subcellular fractionation

Protein interaction

Felix Meissner & Matthias Mann, Nature Immunology 2014
Fragmentation of peptides yields amino acid sequence.

C-terminal fragments

\[ Y_{n-1} \]

\[ Y_{n-2} \]

N-terminal fragments

\[ b_1 \]

\[ b_2 \]
Standard DB Search

Sequence database entry

In-silico digestion

- MAILAGGHSVR
  - FGPK
  - AFAEVNGETFYSR
  - VITLESTNMFNEIIK
  - YPNVVIDDENNDK
  ...

Theoretical proteolytic peptides

In-silico fragmentation

- MAILAG
  - MAIL
  - M
  - MAI
  - AIILAG

Theoretical fragmented peptides

Theoretical peptide list

Ions peaklists

340.695086
676.960634
498.8283
545.564
1171.967066
261.107346
342.51458
456.727405
363.268453

Result: ranked list of peptide and protein candidates

340.695086
676.960634
498.8283
545.564
1171.967066
261.107346
342.51458
456.727405
363.268453
Draft Proteomes

Geiger T. et. al, Mol Cel Proteomics, 2013


Chromosomal coverage of the 18,097 proteins identified
Blue bars indicate the density of proteins in a particular chromosomal region.

Gene ontology analysis of the ‘missing’ proteome identifies GPCRs, secreted and keratin-associated proteins as the major protein classes underrepresented in proteomic experiments.

Comparative 2D Gel Electrophoresis

- Mr
- pI

Excision/digestion

LC-MS/MS
Fluorescence 2D Difference Gel Electrophoresis (DIGE)

- Pooled internal standard
  - Label with Cy2
- Protein extract 1
  - Label with Cy3
- Protein extract 2
  - Label with Cy5

Mix labelled extracts → 2-D separation → Image gel with Typhoon Variable Mode Imager → Image analysis and data quantitation with DeCyder Differential Analysis Software
### MS Methods for Quantification

<table>
<thead>
<tr>
<th></th>
<th>Label Free</th>
<th>Stable Isotope Labeling</th>
<th>SIL Internal Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Quantifiable proteins per MS run:</strong></td>
<td>1000s</td>
<td>1000s</td>
<td>10s – 100s</td>
</tr>
<tr>
<td><strong>Accuracy:</strong></td>
<td>+</td>
<td>+ to +++</td>
<td>++++</td>
</tr>
<tr>
<td><strong>Precision:</strong></td>
<td>+</td>
<td>+++</td>
<td>++++</td>
</tr>
<tr>
<td><strong>Complexity of experiment / setup time:</strong></td>
<td>++++</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td><strong>Cost:</strong></td>
<td>+++*</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td><strong>Absolute quantification:</strong></td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>
Stable isotope labeling by amino acids in cell culture (SILAC)

Introduction

SILAC relies on metabolic incorporation of a given ‘light’ or ‘heavy’ form of the amino acid into the proteins. The method relies on the incorporation of amino acids with substituted stable isotopic nuclei (e.g. deuterium, 13C, 15N). Thus in an experiment, two cell populations are grown in culture media that are identical except that one of them contains a ‘light’ and the other a ‘heavy’ form of a particular amino acid (e.g. 12C and 13C labeled l-lysine, respectively). When the labeled analog of an amino acid is supplied to cells in culture instead of the natural amino acid, it is incorporated into all newly synthesized proteins. After a number of cell divisions, each instance of this particular amino acid will be replaced by its isotopelabeled analog. Since there is hardly any chemical difference between the labeled amino acid and the natural amino acid isotopes, the cells behave exactly like the control cell population grown in the presence of normal amino acid. It is efficient and reproducible as the incorporation of the isotope label is 100%. We anticipate that potential applications of SILAC will lead to its use as a routine technique in all areas of cell biology.

http://silac.org/
**Stable Isotope Labeling with Amino acids in Cell culture**

-SILAC-labeling of tissue samples are not possible. However, primary culture of cells harvested from tissue in SILAC is possible.

Incorporation of stable isotopes

(from Ong et. al 2002, MCP)
Relative Quantification

• Intensity-based absolute quantification (iBAQ)

\[
\text{sum of all identified peptide intensities} \quad = \quad \text{iBAQ}
\]

48 accurately quantified human proteins with abundances over 6 orders of magnitude

Schwanhäusser, Nature 2011
Correlation Between mRNA and Protein Levels
ECM Proteins are under-represented in proteomic datasets

7000 Proteins Quantified:
- Only 4 ECM proteins in the top 100 proteins for each tissue

All tissues
#5085 COL1A1

<table>
<thead>
<tr>
<th>Fetal tissue (6 organs)</th>
<th>Adult tissue</th>
<th>Lung</th>
</tr>
</thead>
<tbody>
<tr>
<td>#91 COL6A3</td>
<td>#11 COL6A3</td>
<td>#7</td>
</tr>
<tr>
<td>#317 COL1A1</td>
<td>#15 HSPG2</td>
<td>#15</td>
</tr>
<tr>
<td>#531 COL1A2</td>
<td>#59 COL6A1</td>
<td>#31</td>
</tr>
<tr>
<td></td>
<td>#60 FN1</td>
<td>#47</td>
</tr>
<tr>
<td></td>
<td>#70 COL14A1</td>
<td>#135</td>
</tr>
<tr>
<td></td>
<td>#85 COL6A2</td>
<td>#57</td>
</tr>
<tr>
<td></td>
<td>#167 COL1A1</td>
<td>#66</td>
</tr>
<tr>
<td></td>
<td>#227 COL1A2</td>
<td>#138</td>
</tr>
</tbody>
</table>
Differential ECM extraction of soluble and insoluble ECM
# Methods to Quantify ECM Proteins

<table>
<thead>
<tr>
<th>Method</th>
<th>Quantifiable proteins per MS run</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Complexity of experiment / setup time</th>
<th>Cost</th>
<th>Absolute quantification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label Free</td>
<td>1000s</td>
<td>+</td>
<td>+</td>
<td>+++</td>
<td>+++*</td>
<td>NO</td>
</tr>
<tr>
<td>Stable Isotope Labeling</td>
<td>1000s</td>
<td>+ to +++</td>
<td>++++</td>
<td>++</td>
<td>++</td>
<td>NO</td>
</tr>
<tr>
<td>SIL Internal Standards</td>
<td>10s – 100s</td>
<td>++</td>
<td>++++</td>
<td>-</td>
<td>-</td>
<td>YES</td>
</tr>
</tbody>
</table>

- **Run 1**
- **Run 2**

### Notes
- Accuracy: + to +++
- Precision: + to +++
- Complexity of experiment / setup time: ++++ to -
- Cost: +++* to ++
- Absolute quantification: NO to YES
LC/MS Properties: GPMDB

- Compares peptides to a collection of previously observed results
- Determines how many times the peptide has been observed by others
- Most proteins show very reproducible peptide patterns
Quantitative Peptide Concatemers

- Arg & Lys BL21 Auxotroph: >99% incorporation
- Create 40+ SIL probes in a single Qconcat
- Absolute Quantification:
  - Intra/Inter-laboratory & experiment applications
1. Differential ECM extraction of soluble and insoluble ECM

2. Absolute quantification using internal standard peptides and targeted MS
### Mass Spectrometry Methods

<table>
<thead>
<tr>
<th></th>
<th>data-dependent acquisition</th>
<th>targeted data acquisition</th>
<th>data-independent acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantifiable proteins per MS run:</strong></td>
<td>1000s</td>
<td>10s – 100s</td>
<td>1000s</td>
</tr>
<tr>
<td><strong>Reproducibility:</strong></td>
<td>+</td>
<td>++++</td>
<td>++++</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td>++</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td><strong>Software availability</strong></td>
<td>++++</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td><strong>Cost of equipment:</strong></td>
<td>+</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td><strong>Runs times / sample:</strong></td>
<td>2 to 24 hrs</td>
<td>1 to 2 hrs</td>
<td>2 to 10 hrs</td>
</tr>
</tbody>
</table>
Targeted Assay

Quantification of 214 Peptides in a single 25 min gradient.

Optimized Transitions, Collision Energy, and Declustering Potential for optimal sensitivity and reproducibility.
### Protein Modification DBs

<table>
<thead>
<tr>
<th>Modification</th>
<th>Short name</th>
<th>Monoisotopic Average Composition</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>amidino</td>
<td>amidino</td>
<td>42.021798 42.0400 H(2) C N(2)</td>
<td></td>
</tr>
<tr>
<td>Guanidination</td>
<td>Guanidination</td>
<td>42.021798 42.0400 H(2) C N(2)</td>
<td></td>
</tr>
<tr>
<td>tri-Methylation</td>
<td>tri-Methylation</td>
<td>42.046950 42.0797 H(6) C(3)</td>
<td></td>
</tr>
<tr>
<td>Carbamylation</td>
<td>Carbamyl</td>
<td>43.005814 43.0247 H C N O</td>
<td></td>
</tr>
<tr>
<td>Carboxylation</td>
<td>carboxyl</td>
<td>43.989829 44.0095 C O(2)</td>
<td></td>
</tr>
<tr>
<td>Gamma-carboxylation</td>
<td>Gamma-carboxyl</td>
<td>43.989829 44.0095 C O(2)</td>
<td></td>
</tr>
<tr>
<td>N-carboxylation of Met</td>
<td>N-carboxyl</td>
<td>43.989829 44.0095 C O(2)</td>
<td></td>
</tr>
<tr>
<td>S-Ethylcystine from Serine</td>
<td>S-Eth</td>
<td>44.008456 44.1188 H(4) C(2) O(-1) S</td>
<td></td>
</tr>
<tr>
<td>Ethanolation of Cys</td>
<td>EtOH</td>
<td>44.026215 44.0526 H(4) C(2) O</td>
<td></td>
</tr>
<tr>
<td>Oxidation to nitro</td>
<td>Nitro</td>
<td>44.985078 44.9976 H(-1) N O(2)</td>
<td></td>
</tr>
<tr>
<td>Acetate labeling reagent (N-term &amp; K) (heavy form, +3amu)</td>
<td>Acetyl_heavy</td>
<td>45.029395 45.0552 H(-1) H2(3) C(2) O</td>
<td></td>
</tr>
<tr>
<td>Beta-methylthiolation</td>
<td>b-methylthiol</td>
<td>45.987721 46.0916 H(2) C S</td>
<td></td>
</tr>
<tr>
<td>Methyl methanethiosulfonate</td>
<td>MMTS</td>
<td>45.987721 46.0916 H(2) C S</td>
<td></td>
</tr>
<tr>
<td>Selenium replaces sulphur in cysteine</td>
<td>seCys</td>
<td>47.944449 46.8950 S(-1) Se</td>
<td></td>
</tr>
<tr>
<td>Selenium replaces sulphur in Methionine</td>
<td>SeMet</td>
<td>47.944449 46.8950 S(-1) Se</td>
<td></td>
</tr>
<tr>
<td>cysteine oxidation to cysteic acid</td>
<td>Cysteic_acid</td>
<td>47.984744 47.9982 O(3)</td>
<td></td>
</tr>
</tbody>
</table>

  Unimod, protein modifications DB
- [http://www.ncifcrf.gov/RESID/](http://www.ncifcrf.gov/RESID/) Database of Protein Modifications
Improved Proteome Coverage

![Graph showing Open Mass Modification Search with peaks for various modifications: SNP, -H₂O, Na⁺, Ac, Me, PO₄, Ub, etc. for simple glycans, branched glycans, and xlinked peptides.](image)
Collagen Galactosyl Glycans

~42,000 years old

~120,000 years old