Computational Proteomics
Outline

• Proteomics
• Mass Spectrometer
• Tandem Mass Spectrometry
• Peptide Identification with DB Search
• DB Search Result Validation
• Peptide De novo sequencing
• Quantification
Why Proteomics

The two very different cells have the same genome!
Why Proteomics

- Post Translational Modifications
- Example: Reversible phosphorylation of proteins is an important regulatory mechanism.
- Example: p53 protein has more than 18 different phosphorylation sites.
Proteomics

Genome
-20-25,000 genes

Transcriptome
-100,000 transcripts

Proteome
>1,000,000 proteins

Alternative promoters
Alternative splicing
mRNA editing

Post-translational modifications
HER2-positive breast cancer is a breast cancer that tests positive for a protein called human epidermal growth factor receptor 2 (HER2), which promotes the growth of cancer cells.

"HER2-positive breast cancers tend to be more aggressive than other types of breast cancer. They're also less responsive to hormone treatment. However, treatments that specifically target HER2 are very effective."
• Personalized
• Predictive
• Preventative
• Participatory
Protein

Primary structure is a sequence.
20 frequent amino acids.
Fold into a complex 3D structure.
Amino Acids

• There are 20 amino acids. All have the same basic structure but with different side chains:

\[
\begin{align*}
\text{H} & \quad \text{N} & \quad \text{C} & \quad \text{C} & \quad \text{O} & \quad \text{H} \\
\text{H} & \quad \text{R} & \quad \text{side chain group}
\end{align*}
\]

• Examples:

Glycine, or Gly, or G

Arginine, or Arg, or R
Peptides and Proteins

Glycine, or Gly, or G

Arginine, or Arg, or R

N-terminal

C-terminal

peptide bonds

Atom mass:
C=12.000
H=1.008
N=14.003
O=15.995
Protein Synthesis

- Gene (segments of genome) -> mRNA -> protein
- https://www.youtube.com/watch?v=lkq9AcBcohA
A Protein Sequence

>P02769|ALBU_BOVIN Serum albumin - Bos taurus (Bovine).

MKWVTFISLLLLLFSAYSRGVFRRDHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQCPFDEHVKLVNELTEFAKTCAVESHAGCEKSLHTLFGDELCKVASLRTEYGDMDACCQEKFPERNECFLHSHKDDSPLPDPLKPDNPNTLCDEFKADEKKFWGKYLYEIARRHPYFYAPELILYANKYNGVFOQECQAEKDGACLIPKIETMREKVLASARQRLRCAIFKFGERALKAWSVARLSQKFPKAEFVEVTKLVDLTKVHKECCCHGDLLECADDDARDLAKYICDNQDTISSKLKECCDKPLEKSHCIAEVEKDAIPENLPPLTADFAEDKDVCKNYQEAKDAFLGSFLYEYSRRHPEYAVSVLLRLAKEYEATLCEEACAKDDPHACHYSTVFSDKLKLHVLDEPQNLIKQNCDQFEKLGYGFQNALIVRYTRKVPQVSTPTLVLEVSRSLGKVGTRECTKPESEMRMPCTEDYLSLILNRLCVLHEKTPVSEKVTCCSTESLVNRRPFCSAULTPDETYVPKFDEKLFTFHAICLPDTEKQIKQQTALVELLKHKPKATEEEQLKTVMENFVAFVDKCCAADDEACFAVEGPKLVSTQATA
Protein Sequence Database

• A common way is through gene prediction.
• Also one can use Edmon Degradation for specific protein.
  – Remove an amino acid from the N-terminus of the protein.
  – Separate and determine the removed amino acid.
  – Repeat.
  – https://www.youtube.com/watch?v=7nubm99YOyw
• Mass spectrometry has been used routinely to sequence short peptides, and now started to be used to sequence proteins.
Purpose of Proteomics

• Identify all proteins in biological sample
• Identify PTMs and variations on proteins
• Identify endogenous peptides in sample
• Quantify them
• Study protein-protein interactions
• Find biomarkers
• Today, mostly studied by using mass spectrometer
## Mass (Weights) of Atoms

<table>
<thead>
<tr>
<th>element</th>
<th>nominal mass</th>
<th>exact mass</th>
<th>Percent abundance</th>
<th>average mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>12</td>
<td>12.000000</td>
<td>98.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>13.00335</td>
<td>1.1%</td>
<td>12.0115</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>1.00783</td>
<td>99.98%</td>
<td>1.008665</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.0140</td>
<td>0.02%</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>16</td>
<td>15.99491</td>
<td>99.8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>17.9992</td>
<td>0.2%</td>
<td>15.994</td>
</tr>
<tr>
<td>N</td>
<td>14</td>
<td>14.00307</td>
<td>99.63%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15.00011</td>
<td>0.37%</td>
<td>14.0067</td>
</tr>
<tr>
<td>S</td>
<td>32</td>
<td>31.97207</td>
<td>94.93%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>32.97146</td>
<td>0.76%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>33.96787</td>
<td>4.29%</td>
<td></td>
</tr>
</tbody>
</table>
# Amino Acid Residue Mass Table

<table>
<thead>
<tr>
<th>Name</th>
<th>3-letter code</th>
<th>1-letter code</th>
<th>Monoisotopic Mass</th>
<th>Average Mass</th>
<th>Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>71.03711</td>
<td>71.08</td>
<td>C₅H₅NO</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>156.10111</td>
<td>156.2</td>
<td>C₆H₁₂N₄O</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
<td>114.04293</td>
<td>114.1</td>
<td>C₆H₆N₂O₂</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>Asp</td>
<td>D</td>
<td>115.02694</td>
<td>115.1</td>
<td>C₄H₅NO₃</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td>103.00919</td>
<td>103.1</td>
<td>C₅H₄NOS</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>Glu</td>
<td>E</td>
<td>129.04259</td>
<td>129.1</td>
<td>C₅H₇NO₃</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
<td>128.05858</td>
<td>128.1</td>
<td>C₅H₈N₂O₂</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
<td>57.02146</td>
<td>57.05</td>
<td>C₂H₃NO</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td>137.05891</td>
<td>137.1</td>
<td>C₆H₇N₃O</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td>113.08406</td>
<td>113.2</td>
<td>C₆H₁₁NO</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
<td>113.08406</td>
<td>113.2</td>
<td>C₆H₁₁NO</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
<td>128.09496</td>
<td>128.2</td>
<td>C₆H₁₂N₂O</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
<td>131.04049</td>
<td>131.2</td>
<td>C₅H₉NOS</td>
</tr>
<tr>
<td>Phenyalanine</td>
<td>Phe</td>
<td>F</td>
<td>147.06841</td>
<td>147.2</td>
<td>C₉H₉NO</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
<td>97.05276</td>
<td>97.12</td>
<td>C₅H₇NO</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
<td>87.03203</td>
<td>87.08</td>
<td>C₃H₅NO₂</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
<td>101.04768</td>
<td>101.1</td>
<td>C₃H₇NO₂</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
<td>186.07931</td>
<td>186.2</td>
<td>C₁₁H₁₆N₂O</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
<td>163.06333</td>
<td>163.2</td>
<td>C₉H₉NO₂</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
<td>99.06841</td>
<td>99.13</td>
<td>C₅H₉NO</td>
</tr>
</tbody>
</table>
Fundamental Idea for Peptide ID

\[ b_1 \quad A \text{ NELLLLNVK} \quad y_8 \]
\[ b_2 \quad \text{AN ELLLNVK} \quad y_7 \]
\[ b_3 \quad \text{ANE LLLLNVK} \quad y_6 \]
\[ b_4 \quad \text{ANEL LLNVK} \quad y_5 \]
\[ b_5 \quad \text{ANELL LNVK} \quad y_4 \]
\[ b_6 \quad \text{ANELLL NVK} \quad y_3 \]
\[ b_7 \quad \text{ANELLLLNVK} \quad y_2 \]
\[ b_8 \quad \text{ANELLNNV K} \quad y_1 \]
Not That Simple

• Many difficulties to carry out the idea
  – Noise,
  – Imperfect fragmentation,
  – Ambiguous peak assignments
• How about the protein?
• How about quantification?
• How about PTMs?
MASS SPECTROMETERS
Mass Spectrometer

• Given molecules (small or large), ionize them into ions (molecules with electrical charge), measure the mass to charge ratios (m/z).

• Some ionization techniques let \( z = 1 \). Others can derive \( z \) from data computationally.
A Mass Spectrum

peak list

2789.22  3597.0
2790.22  5018.0
2791.23  4406.0
2792.23  2868.0
2793.23  1234.0

...
How does a mass spectrometer work?

A mass spectrometer measures the mass of individual atoms and molecules, which enables scientists to identify them.

Inside a mass spectrometer, a sample of test material is converted into a high-speed stream of electrically charged particles. The stream is then bent by a magnetic field; heavy, slow moving particles (A) and lighter, fast-moving particles (B) crash to either side, so only particles of a certain mass (C) reach a detector. By changing the magnetic field strength, the instrument can measure particles of varying masses. The results are reported as a graph, in which spikes show the relative abundance in the sample of particles of different masses.

http://www.youtube.com/watch?v=J-wao0O0_qM

This is one of the simplest spectrometers. Today’s instruments are more complex and use different mechanism.
Today’s Instruments Used in Proteomics

- MALDI-R
- Q-Tof Micro
- FT-ICR
- LTQ-Orbitrap
Three Basic Components

• Ionizer
  – MALDI, ESI

• Mass Analyzer
  – Magnetic Sector, Iontrap, TOF, Quadrupole, FT, Orbitrap

• Detector
  – Electron multiplier, FT

• In addition, how to fragment the parent ion is also important for today’s proteomics
Ionization (1) – MALDI

• Matrix-assisted laser desorption/ionization (MALDI) is a soft ionization technique.
History

- 1960s: irradiation of low-mass organic molecules with a high-intensity laser pulse lead to the ion formation.
  - Origins of laser desorption (LD) ionization.
- 1985: Franz Hillenkamp, Michael Karas and their colleagues coined the term.
  - Found that Alanine (D) can be ionized easier if mixed with Tryptophan (W), and irradiated with a pulsed 266 nm laser.
  - Up to 2843 Da peptide could be ionized.
- 1987: Koichi Tanaka of Shimadzu Corp. and his co-workers
  - combined 30 nm cobalt particles in glycerol with a 337 nm nitrogen laser
  - Up to 34,472 Da protein.
  - Shared ¼ of 2002 Nobel prize in Chemistry.
A MALDI plate
Electrospray Ionization: Formation of Charged Droplets

Formation of multiply charged ions
Electrospray
History

• A.B. from Berea College in his new hometown.
• 1940: Ph.D. from Yale University.
• 1962: He joined the Yale University faculty.
• 1987: he reached the mandatory retirement age (70).
• University-mandated move to smaller laboratory space.
• Started to work on ESI.
• 1994: Fenn joined Virginia Commonwealth University.
• 2002: Received Nobel prize.
• The patent rights to ESI became the subject of a legal case between Yale University and Fenn.
• 2005: Yale was awarded over one million dollars and partial patent rights to the technique.

Fenn's first electrospray ionization source (top) coupled to a single quadrupole mass spectrometer
Mass Analyzer (1) – Quadrupole

Quadrupole Mass Analyzer

Mass filter; complete spectrum is obtained by scanning whole range

Mass range 10-4,000 Da
Mass Analyzer (2) – TOF

- Time of Flight.

Time of flight is proportional to $\sqrt{m/z}$
Mass Analyzer (3) – Orbitrap

Moving ions are trapped around an electrode.

By shaping the electrode appropriately, ions also move left and right. Left-right frequency proportional to $\sqrt{m/z}$.

Fourier transform to convert the time-field signal to frequencies.

Notes:
- All-mass detection
- Noise equiv. to 20 ions (1 sec)
Orbitrap History

- Inventor: Alexander Alexeyevich Makarov

- 1989 Moscow Engineering Physics Institute - M.S. Molecular Physics
- 1993 Moscow Engineering Physics Institute - Ph.D. Physics and Mathematics
- 2006 Thermo released Orbitrap.
- 2008 American Society for Mass Spectrometry Distinguished Contribution in Mass Spectrometry Award
Detector (1) – Electron Multiplier

- An electron multiplier is a vacuum-tube structure.
- Secondary emission
  - a single electron can, when bombarded on secondary emissive material, induce emission of roughly 1 to 3 electrons.
- This can be repeated a number of times, resulting in a large shower of electrons all collected by a metal anode, all having been triggered by just one.
Summary

Ionizer
- MALDI
- ESI

Mass Analyzer
- Iontrap
- Quadropole
- TOF
- FT
- Orbitrap

Detector
- Electron Multiplier
- Fourier Transform

Ionize molecules
separate ions with m/z
detect ions

Mixed molecules
intensity
m/z
Isotope Peaks

• Recall that C has two common isotopes: C12 (98.9%) and C13 (1.1%).

• Given enough Carbon atoms in the peptide, the isotope peak will start to show in the spectrum.
In a mass spectrum

Deconvolution adds all the isotopic peaks to the monoisotopic peak. So, the later process does not need to worry about the isotopes.
Isotope & Mass

Check the difference
Isotope & Resolution
Isotope & Charge Deconvolution

- The same molecules may be charged differently, and therefore form a few peaks in the spectrum.

For protein/peptide with positive charges, the charge is obtained from adding protons (which has mass approx. 1 dalton. As a result, a molecule with mass $M$ will have peaks at $(M+Z)/Z$. 

![Graph showing m/z values](image)
Exercise

![Graph with peaks at 395.73, 396.22, and 397.24]
TANDEM MASS SPECTROMETRY
Peptide Identification

• Now we can measure the mass of a peptide.
• That’s not sufficient for identifying it.
  – Many peptides may share the same/similar mass.
  – A peptide may be modified so the mass is not determined by the sequence alone.
• Idea: Fragment the peptide and measure the pieces.
  – So you have more information about the peptide to determine it unambiguously.
Peptide Identification

b₁  A NELL LNVK  y₈
b₂  AN EL LLNVK  y₇
b₃  ANE LLLN VK  y₆
b₄  ANEL LLN VK  y₅
b₅  ANELL LN VK  y₄
b₆  ANELLL NVK  y₃
b₇  ANELLLN VK  y₂
b₈  ANELLLNV K  y₁
Tandem Mass Spectrometry

- Can combine different mass analyzers. E.g. Q-Tof.
Tandem Mass Spectrometry

• One can also couple other mass analyzers to do MS/MS.
• The MS and MS/MS can also be measured with different mass analyzers too.
• Normally a few tandem scans after each survey scan.
• The precursor mass to gate in for each tandem scan is dynamically determined after the survey scan. This is called DDA (Data Dependent Acquisition).
Intact Proteins v.s. Peptides

• An intact protein is usually too large for MS/MS.
  – Although some people are doing this now.
• There are enzymes that digest proteins into short peptides.
Protease

- A **protease** is any **enzyme** that conducts **proteolysis**.
- Or: a protease breaks protein in water.
- Trypsin digests at site [KR] | [^P].

...M-A-L-R-Q-V-...

...M-A-L-R Q-V-...
Liquid Chromatography

• LC or HPLC (High Performance LC) is used to further separate the complex peptide mixture, before they enter MS.

• In proteomics, LC separates peptides according to their hydrophobicity.
  – (from the Attic Greek hydro, meaning water, and phobos, meaning fear)
HPLC

• Read wikipedia or a biochemistry book if you’re interested. But for the purpose of this course:

• Given a mixture of peptides, HPLC will elute different peptides at different time.
  – Elution time or retention time (RT).
  – And hence separate the peptides.

• Separation isn’t perfect.
  – Each peptide elutes at a time window (as an LC peak).
  – Multiple peptides’ LC peaks can overlap.
LC-MS/MS

MS or survey scan

BSA-GluC-1.mzXML: ms=1 RT=31.171 scan=1707 TIC=2.13E8

BSA-GluC-1.mzXML: ms=2 mz=977.1248 z=3 RT=31.1804 scan=1708 TIC=1.14E7
(1) A mixture of protein is digested into peptides with added enzyme (usually trypsin).

(2) The resulting peptides are separated with liquid chromatography (LC). Different peptides elute at different time. The separation may not be perfect.

(3) The spectrometer scans the peptide ions at a particular time and obtains a profile MS scan. Each peak in the MS spectrum supposedly corresponds to a peptide.

(4) The spectrometer selects a peak (a peptide ion) with the first mass analyzer, fragments it and produces an MS/MS scan with the second mass analyzer.
Possible Ways to Interpret MS/MS Data

1. Database search
2. De novo sequencing
3. Inexact homology search

MS/MS Spectra

peptides

protein DB

Inexact protein DB

homology search

homologous peptides
PEPTIDE ID WITH DB SEARCH
Idea

- Idea is very simple:
  - Try every peptide in a sequence database.
  - Evaluate the matching between the peptide’s fragment ions’ mass and the peaks in the MS/MS spectrum.
  - Report the top scoring peptide.
Calculate Fragment Ions

- **b** and **y** the most abundant for CID.
- **c** and **z**, **z+1** the most abundant for ETD.
- Some fragment ions also lose H2O and/or NH3.

### Fragment Ions

<table>
<thead>
<tr>
<th>Ion Type</th>
<th>Mass</th>
<th>Nominal Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>[M]+ 1.0078</td>
<td>[M]+ 1</td>
</tr>
<tr>
<td>c</td>
<td>[M]+ 18.0344</td>
<td>[M]+ 18</td>
</tr>
<tr>
<td>x</td>
<td>[M]+ 44.9977</td>
<td>[M]+ 45</td>
</tr>
<tr>
<td>z</td>
<td>[M]+ 1.9918</td>
<td>[M]+ 2</td>
</tr>
<tr>
<td>z²</td>
<td>[M]+ 2.9997</td>
<td>[M]+ 3</td>
</tr>
</tbody>
</table>

... 

\[
m(b3) = m(A) + m(N) + m(E) + 1.0078 \]
\[
m(b4) = m(A) + m(N) + m(E) + m(L) + 1.0078
\]

... 

\[
m(y2) = m(K) + m(V) + 19.0184
\]
\[
m(y3) = m(K) + m(V) + m(N) + 19.0184
\]

...
DB Search History – The Origin

• 1994: SEQUEST

• This is the first paper for peptide identification with MS/MS and computer software.

• Licensed by Thermo for its LCQ instrument.
John Yates III

- 1980: B.A. University of Maine
- 1983: M.S. University of Maine
- 1987: Ph.D. University of Virginia
- 1994: SEQUEST
- 1996: ASMS Research Award
- 2001: MudPID
- 2004: ASMS Biemann Medal
- Over 350 publications. H-index >110.
SEQUEST

Figure 1. Flow chart that depicts the algorithm for searching protein databases with tandem mass spectrometry data.
The Score Function

• A cross-correlation is used to compare the top 500 candidates.
• First calculate a theoretical spectrum.
Cross Correlation

• For spectrum $x[i]$ and $y[i]$.

\[ R_{\tau} = \sum_{i=0}^{n-1} x[i] y[i + \tau] \]

• $R_{\tau}$ calculated with FFT.

• A good match between $x$ and $y$ should give highest $R_{\tau}$ when $t=0$.

• $R_0$ minus mean for $-75 < t < 75$ is the final score.

• Scores of 500 candidates are normalized to 1.

• Difference between first- and second-ranked candidate greater than 0.1 is a good indication of true match.
Sequest Score

- The score was not arbitrarily chosen.
- The correlation (dot product) was first normalized against the junk spectra (produced by shifting).
- Then it is normalized against the junk peptides (top 500 candidates).
- This makes the PSM (peptide-spectrum match) scores are comparable between different spectra.
Another

• **1999: Mascot**

• **Currently** the most widely used program.

• Get aged.
Yet Another

• 2004: X!TANDEM

• 2006: X!Hunter:

thegpm.org
Multi-round Search

• An interesting idea introduced by X!Tandem was multi-round search.
• First round, high-quality spectra vs. all proteins. Result is a short list of highly confident proteins.
• Second round, all spectra vs. short list of proteins.
• This made Sequest one of the fastest search engine.
More...


Another Worth Mentioning

• 2005 InsPecT:

• These are the more “computational” people

Vineet Bafna, Pavel A. Pevzner: Genome Rearrangements and Sorting by Reversals
FOCS 1993: 148-15
PEAKS DB

- Used in 500 labs.
- One of the best DB search software in the market.
One Difficulty

• Variable PTM.

• A PTM can be on and off. So, a peptide in database becomes exponentially many if multiple variable PTMs are to be searched.
RESULT VALIDATION
Optimality v.s. Reality

I found the optimal solution!

Really? Is it Real?
Peptide Spectrum Matches (PSM)
Result Quality

- Earlier publications reported everything software identified.
- MCP started the guideline in 2004 to ensure accuracy.

The need for these guidelines is driven in part by the fact that a significant but undefined number of the proteins being reported as “identified” in proteomics articles are likely to be false positives (2). These incorrect matches probably result
The Sensitivity and Accuracy Dilemma
Result Quality Control

• Confidence: e.g. PeptideProphet (ISB)

• False Discovery Rate: e.g. decoy database
  – and other publications at the same period.
FDR Estimation

Protein DB

target

decoy

Search Engine

Identified Peptides

# false target hits ≈ # decoy hits

\[ FDR = \frac{\#\text{decoy}}{\#\text{target}} \]

Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry

Joshua E Elias¹ & Steven P Gygi¹,²

FDR Curves
FDR Should Be Used with Caution

• Original target-decoy paper validated SEQUEST’s results.

• Target-decoy was used to validate many search engines in later publications.

• A lot of people do not use it correctly.
Pitfall 1 – Multiple Round Search

Round 1. Fast Search

Round 2. More Sensitive Search

More targets than decoys

# false target hits > # decoy hits

FDR underestimation.

Solution: Decoy Fusion

Decoy sequence append to each target protein.

Fast Search

More Sensitive Search

Equal targets and decoys

# false target hits \( \approx \) # decoy hits

PEAKS DB paper. MCP 2011. 10.1074/mcp. M111.010587.
Idea: Peptides on a multi-hit protein get a bonus on their scores to increase sensitivity.

Pitfall

More multi-hit proteins from target DB ⇒ more false hits are “saved” from target DB ⇒ FDR underestimation.
Solution: Decoy Fusion

Weak false hits are “saved” with approx.
equal probabilities in target and decoy.

Got the sensitivity, but still estimate the FDR correctly.
Percolator

- Retrain the parameters in scoring function to separate target and decoy hits.

Decoy Should Be Used with Caution

Idea: Re-train the coefficients of scoring function for every search after knowing the decoy hits.

Pitfall: Risk of over-fit. Machine learning experts only.

Adjust scoring function to remove decoy hits after search.

Fewer target false hits are removed ⇒ FDR underestimation
Too Many Imperfect Engines

- 2004: Scaffold
  - Become the main product of Proteome Software (California)
Summary

• 1994: Sequest (John Yates)
• 1996: Tag based search (Matthias Mann)
• 1999: Mascot (Matrix Science)
• 2002: PeptideProphet
• 2003-2004:
  – X!Tandem (Ron Beavis)
  – OMMSA (NCBI)
  – Phenyx (GenBio)
• 2004: Scaffold
• 2005-2006:
  – Mascot Error Tolerance Search
  – InsPecT (Vineet Bafna & Pavel Pevzner)
• 2007: False discovery rate.
• 2008: MaxQuant (Matthias Mann)
• Lately: pFind, MSGF-DB, PEAKS DB, ByOnic
PEPTIDE DE NOVO SEQUENCING
PSM Score

• While a human can easily tell good match from bad ones, we need to teach a dumb (but fast) computer to do so.

• A scoring function, $f(S, P)$, measures the “matching” quality between a spectrum $S$ and a peptide $P$. 
PSM Score Important

1. Database search

2. De novo sequencing

De novo sequencing: Given spectrum $S$, construct the peptide $P$ that maximizes $f(S, P)$.

DB Search: Given spectrum $S$, find the peptide $P$ in database that maximizes $f(S, P)$. 

protein DB

peptides

MS/MS Spectra

peptides
Fragment Ions

<table>
<thead>
<tr>
<th>$b_1$</th>
<th>A NELLLNVK</th>
<th>$y_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_2$</td>
<td>AN ELLNVK</td>
<td>$y_7$</td>
</tr>
<tr>
<td>$b_3$</td>
<td>ANEL LNVK</td>
<td>$y_6$</td>
</tr>
<tr>
<td>$b_4$</td>
<td>ANELL NVK</td>
<td>$y_5$</td>
</tr>
<tr>
<td>$b_5$</td>
<td>ANELL NVK</td>
<td>$y_4$</td>
</tr>
<tr>
<td>$b_6$</td>
<td>ANELL NVK</td>
<td>$y_3$</td>
</tr>
<tr>
<td>$b_7$</td>
<td>ANELLLN VK</td>
<td>$y_2$</td>
</tr>
<tr>
<td>$b_8$</td>
<td>ANELLLNV K</td>
<td>$y_1$</td>
</tr>
</tbody>
</table>

Diagram showing mass-to-charge ($m/z$) ratios for the fragment ions.
Likelihood Ratio

• Let \( m \) be the m/z of a y-ion and we observe a peak at \( m \) of the spectrum.

• Two different assumptions:
  – The peptide is the real peptide so peak is caused by the y-ion.
    • \( \Pr(\text{observe a peak at } m | m \text{ is a y-ion m/z of the real peptide}) \)
  – The peptide is a random peptide so the match is purely by chance.
    • \( \Pr(\text{observe a peak at } m | m \text{ is a random mass}) \)
Log Likelihood Ratio

- Learn two probabilities from large training data
  - $p$: Prob(a peak is observed at a y-ion m/z).
  - $q$: Prob(a peak is observed at a random m/z).
  - Usually $p > q$.

- If an expected y-ion is observed, $\log \frac{p}{q}$ is added to score.
  - $\log \frac{p}{q}$ is called the log-likelihood-ratio

- If an expected y-ion is missing, $\log \frac{1-p}{1-q}$, is added to score.

- Thus, matching ion is rewarded and missing ion is penalized.
- Other fragment ion types can be considered similarly.
If the corresponding y and/or b ions are observed for prefix mass \( m \) (and therefore suffix \( M - m \)), then peptide is likely to have a prefix mass \( m \). Let \( f(m) > 0 \). Otherwise, \( f(m) \leq 0 \).

- Note that \( f(m) \) is usually the sum of several related ion types.
- Also \( f(m) \) can be computed without knowing the actual sequence.
Score for a Peptide

- For a sequence $P$ with prefix masses $m_1, m_2, \ldots, m_k$, the peptide score is defined as
  
  $$f(S, P) = f(m_1) + f(m_2) + \cdots + f(m_k)$$

De novo sequencing: Given spectrum $S$, construct the peptide $P$ that maximizes $f(S, P)$. 
Given the score function. Find a path from 0 to M, such that each step length is equal to an amino acid residue mass, and total scores of the covered cells are maximized.
Given the score function. Find a path from 0 to M, such that each step length is equal to an amino acid residue mass, and total scores of the covered cells are maximized.
Dynamic Programming

Let $D[m]$ be the maximum total score of a path from 0 to $m$.

- $D[0] = 0$
- For $m$ from 1 to $M$ by

$$D(m) = f(m) + \max_a D(m - m(a)).$$

The best sequence can be retrieved by a backtracking process by repetitively computing the last amino acid $a$.

Time complexity?
High Resolution Data

• What if the mass values are not nominal?
PTM and De Novo Sequencing

• Variable PTM does not cause major speed slow down for *de novo* sequencing algorithms.
  – Instead of trying 20 regular amino acids in the maximization, the algorithm simply tries all modified amino acids too.
  – The time complexity is increased by a constant factor. (Compare to the exponential growth in database search approach).

• However, since the solution space is larger when many variable PTMs are allowed, the accuracy of the algorithm is reduced.
Calculating $f(m)$ with Other Ions

- Scoring function more accurate if more than one ion types are considered.
- For example, if $b$, $y$, and $y$-$H_2O$ ions are considered, then for prefix mass $m$, the corresponding ion masses are:
  
  - $b = m+1; \ y = M-m+19; \ y$-$H_2O = M-m+1$

- Calculate the log likelihood ratio for each ion type, add them up as $f(m)$. 
Double Count
Solution

• Pretend it does not exist?
  – Rare event in real sequences, anyway.
• No.
  – the algorithm is encouraged to find a peptide that reuses many significant peaks as both y and b ions.
  – And those results found by the algorithm will be wrong.
Idea

• Build a prefix and suffix simultaneously, so that when y and b overlap, we know it and only count the peak once.
• A dynamic programming algorithm can do this.
• Other practical solutions exist.
Mass Segment Error

• Most errors are due to incomplete ion ladders in the spectrum.
  – Thus, a segment of amino acids cannot be determined.
  – However, the total mass of the segment, is fixed.
  – E.g. [242]VLSLLVESK, where 242 = N+Q, N+K, or L+E

• The first two or three residues often have low confidence, because of a lack of fragment ions.

• Most de novo sequencing software uses the precursor mass as a constraint (thus the mass of the derived sequence is usually correct).
APPLICATIONS OF DE NOVO SEQ.
Outline

• Traditionally, de novo seq. was regarded as a method when there is no database available. But people have changed their minds...

• Current de novo performance

• How to utilize de novo results
  1. Improving DB search
  2. De novo + tag search
  3. De novo + homology search
  4. Protein sequencing
De Novo Harder than DB Search

- DB does not require full fragment ion ladder.
De Novo Challenge

• Often de novo sequence partially correct.
  – Mass segment error is the most common.

<table>
<thead>
<tr>
<th>(de novo)</th>
<th>LSCFAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(real)</td>
<td>SLCFAV</td>
</tr>
</tbody>
</table>

• Speed is **NOT** a problem.
  – PEAKS de novo 15 spectra/sec on a $1000 PC.
  – Thanks to computer algorithms.
  – You only care about the optimal peptide, a dynamic programming algorithm will compute it directly, instead of trying each combination exhaustively.
• Most of confident DB results have ≥6 AA correctly identified by de novo.
• Most DB results matching de novo with ≥6 AA are correct.
Application 1: Improving DB Search

1% FDR, 174 peptides

1% FDR, 149 peptides

1% FDR, 183 peptides

1% FDR, 165 peptides
Improving DB Search

• Indeed, this is used in PEAKS DB 5.3.

• Performance on ABRF 2011 iPRG study.
Application 2: Tag Search


[200]CFAV[225]
||||
SL CFAV PK

• To compare: SEQUEST (1994), Mascot (1999).
Theoretical Tag Search Performance

- 1% FDR, 149 peptides
- 0% FDR, 163 peptides

Graph showing the relationship between Mascot score and the number of correct amino acids (AA) for true and false peptides.
Application 3: Tag Homology Search

- What if homologous database?
Listen to Both Parties

de novo

LSCFAK

SLAAFK

bioinformagician

SLCAFK

Homolog
Two Examples

<table>
<thead>
<tr>
<th>(denovo)</th>
<th>(real)</th>
<th>(homolog)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSCFAV</td>
<td>SLCFAV</td>
<td>SLCF−V</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(denovo)</th>
<th>(real)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSCFV</td>
<td>EACFV</td>
</tr>
</tbody>
</table>

$m(\text{LS})=m(\text{EA})=200.1$ Da
Experiment

ALBU_BOVIN

28

PEAKS de novo

13 correct +
15 partially correct

human DB

SPIDER

24 correct +
4 partially correct
Application 4: Complete Protein Sequencing

- Hopper *et al.*, JBC 1989, 106AA
- Martin-Visscher *et al.*, Appl. Env. Microbiology, 2008, 60AA circular

1. Multi-enzyme digestion
2. Sequence each peptide
3. Assemble by overlapping
CHAMPS (Step 1)

DTHKEELHAR
IARHFDDLEFENFQ
QFNLGLVLIA
VEPAAASQYLQ

Step 1. de novo provides partially correct sequences.
CHAMPS (Step 2)

Reference:

DTHKSEIAHRFNDLGEENFQGLVLIAF

DTHKEELHARQFNGLVLIA

IARHFDDLEFENFQVEPAASQYLQ

Step 2. SPIDER finds a reference protein, and the mapping of the de novo peptides to the reference protein.
CHAMPS (Step 3)

Reference: DTHKSEIAHRFNDLGEEENFOGLVLIAFQSYLQ
Target: DTHKEEIAHRFNDLFEENFOGLVLIAAASQYLQ

Step 3. Compute the target protein that minimizes the total de novo errors and mutations.

Listen to many, speak to a few. – William Shakespeare.
CHAMPS Performance

- Testing proteins: ALBU_BOVIN and LYS_CHICK.
- Both removed from Swissprot to form a homologous database.

<table>
<thead>
<tr>
<th>Target Protein</th>
<th>Reference Protein</th>
<th>Reference similarity</th>
<th>Coverage</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBU_BOVIN</td>
<td>ALBU_SHEEP</td>
<td>92.5%</td>
<td>99.6%</td>
<td>100%</td>
</tr>
<tr>
<td>LYS_CHICK</td>
<td>LYS_COTJA</td>
<td>95.3%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
QUANTIFICATION
General Idea of Quantification

• Quantification compares two or more samples, and reveal the peptide quantity *changes*.

• Labeling method
  – Two or more samples labeled with different isotopic reagents; and mixed before LC-MS/MS.
  – ICAT, iTRAQ, SILAC

• Label-free method
  – Two or more samples run LC-MS/MS separately;
The four available tags of identical overall mass vary in their stable isotope compositions such that the reporter group has a mass of 114–117 Da and the balancer of 28–31 Da. The fragmentation site between the balancer and the reporter group is responsible for the generation of the reporter ions in the region of 114–117 m/z.
Peptide 114
Peptide 115
Peptide 116
Peptide 117

Same precursor with four differently labeled peptides.

But they produce different fragment reporter ions.

From four different samples.
**Isotope coded affinity tag (ICAT)**

- Different isotopic tags with different mass (mass change = 8 Da).

Structure of the ICAT reagent. A thiol-reactive group covalently binds to cysteine residues of proteins. A linker can incorporate stable isotope signatures. A biotin tag is used to isolate the ICAT-labeled peptides by avidin affinity chromatography.
Isotope coded affinity tag (ICAT)

- Two samples are labeled with heavy and light labels, respectively.
## Compute Protein Ratios

Table 1. Sequence identification and quantitation of the components of a protein mixture in a single analysis.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Peptide sequence identified</th>
<th>Observed ratio (d0/d8)</th>
<th>Mean ± SD</th>
<th>Expected ratio (d0/d8)</th>
<th>% error</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCA_BOVIN</td>
<td>ALC#SEK</td>
<td>0.94</td>
<td>0.96 ± 0.06</td>
<td>1.00</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>C#EVFR</td>
<td>1.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FLDDLTDIMC#VK</td>
<td>0.92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OVAL_CHICK</td>
<td>ADHPFLFC#IK</td>
<td>1.88</td>
<td>1.92 ± 0.06</td>
<td>2.00</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>YPIPEYLQC#VK</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BGAL_ECOLI</td>
<td>LTAAC#FOR</td>
<td>1.00</td>
<td>0.98 ± 0.07</td>
<td>1.00</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>IGLNC#QLAQVAER</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IIFDGVNSAFHLWC#NGR</td>
<td>1.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LACB_BOVIN</td>
<td>WENEGC#AQK</td>
<td>3.64</td>
<td>3.55 ± 0.13</td>
<td>4.00</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>LSNPFTQLEEQC#HI</td>
<td>3.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3P_RABIT</td>
<td>VPTPNVSVDLTC#R</td>
<td>0.54</td>
<td>0.56 ± 0.02</td>
<td>0.50</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>IVSNASC#TTNC#LAPLAK</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHS2_RABIT</td>
<td>IC#GGWQMEEDWDWLRC</td>
<td>0.32</td>
<td>0.32 ± 0.03</td>
<td>0.33</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>TC#AYTNHTVLPEALER</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WLVLC#NPGLAEIIAER</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Gene names are according to Swiss Prot nomenclature (www.expasy.ch).
b ICAT-labeled cysteinyl residue is denoted by # sign.
c Ratios were calculated for each peptide as shown in Fig. 3.
d Expected ratios were calculated from the known amounts of proteins present in each mixture.
Peptide features are matched, and their signal intensities are used to compute quantity ratio.
Label Free Quantification

• Difficulties
  – Retention time alignment.
  – Feature detection difficult due to noise, low S/N, or overlapping features.
  – Feature matching ambiguity.
  – Signal intensity inaccurate when S/N is low.