Proteomics software at MSI.

Pratik Jagtap
Minnesota Supercomputing Institute

http://www.mass.msi.umn.edu/
Proteomics software at MSI.

- **PROTEOMICS : EMERGING TECHNOLOGY**
- **PROTEOMICS WORKFLOW**
  - SEARCH ALGORITHMS
  - DE NOVO ANALYSIS : Peaks
  - STATISTICAL VALIDATION OF PROTEIN IDENTIFICATION
  - QUANTITATIVE TOOLS
  - TARGETED PROTEOMICS
  - DATA DISSEMINATION : Tranche
Proteomics
Fifteen Years Ago...
Two-Dimensional gel electrophoresis

Proteins are resolved based on their isoelectric point (using isoelectric focusing) and then molecular weight (using SDS-PAGE).

Gels are compared, differentially expressed proteins are excised and identified.
Proteomics
Fifteen Years Ago...

Mass Spectrometry

Data Extraction

Search algorithm

Analysis Software that correlates the protein ID to the excised gel spot.
Genomics: Number of Databases for search and Shotgun analysis.

<table>
<thead>
<tr>
<th>Database Type</th>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Complete Genomes</td>
<td>1058</td>
<td>Contact: Genomesonline</td>
</tr>
<tr>
<td>Search GOLD</td>
<td>4999</td>
<td>Location: <a href="http://www.genomesonline.org">www.genomesonline.org</a></td>
</tr>
<tr>
<td>Metagenomes</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>Archival Ongoing Genomes</td>
<td>96</td>
<td>Last Update: July 14, 2009</td>
</tr>
<tr>
<td>Bacterial Ongoing Genomes</td>
<td>2628</td>
<td></td>
</tr>
<tr>
<td>Eukaryotic Ongoing Genomes</td>
<td>1049</td>
<td></td>
</tr>
</tbody>
</table>
Multi-Dimensional Protein Identification Technology

High-Throughput Proteomics: MudPIT

Plasmodium falciparum
[Sporozoites, Trichozoites, Merozoites, Gametocytes]

Lysis

Proteins

Digestion

Peptide Mixture

Tandem Mass Spectrometer

2D Chromatography

RP

SCX

MS/MS Spectrum

SEQUEST® DTASelect & Contrast

> 1,000 Proteins Identified
mass spectrometry

- gas sample enters here
- ions accelerate towards charged slit
- magnetic field deflects lightest ions most
- filament current ionizes the gas
- ions separated by mass
Mass Spectrometers & data formats

Thermofinnigan  Xcalibur / .raw  Sequest
Life Technologies  Analyst / .wiff ; .t2d  ProteinPilot
Waters  Masslynx / .raw
Bruker  .baf

X! tandem  OMSSA  Mascot
Proteomics workflow

Protein → Peptide → Fragmentation → Mass spectrum

Search against database.
Proteo-Informatics

SEARCH ALGORITHM

Mass Spectrometry

Mass spectral data.

Search algorithm

Protein sequence database

Peptide

QYSSVDDLTNTR

M3MS

Experimental spectrum

Compare

Theoretical spectrum

Ranked list of peptide matches

1. QYSSVDDLTNTR 3.3
2. WEELOPEK1 2.1
3. DLLLOWCQPEEK 1.9
4. EGYV3NTIAEK 1.7
5. GDAVFQEAALNR 1.6
6. SYLFQMEAEG 0.2

Frequency

Database search score

Assign peptide to best match

Peptide

QYSSVDDLTNTR

Score

3.3

Expectation value

$5 \times 10^{-4}$
Mass Spectrometry

Search algorithm

- Sequest
- X!tandem
- OMSSA
- MaxQuant
- ProteinPilot
- Mascot

RDC: sdvlapp32

CPC9; CGL 138.

https://sequest7.msi.umn.edu/mascot
Protip / TINT

Raw Data from Orbitrap → mzxml format → Mgf format
- dta format
- X!TANDEM search
- OMSSA search
- SEQUEST search
- Scaffold Analysis
  - Scaffold Viewer

Powered by Tropix
Performing multiple searches through Protip

HUMAN DATASET

Sequest: 5522, 401
xi tandem: 5137, 370
Mascot: 5486, 411
All Together: 8162, 491
Sequest + Mascot: 6554, 441
Sequest + xi tandem: 6962, 441
xi tandem + Mascot: 7443, 462

Powered by Tropix
PROTEINPILOT

Search algorithm
ProteinPilot accounts for more spectra by screening for large number of modifications.

- Glu->pyro-Glu
- iTRAQ4plex
- Methylthio
- No iTRAQ4plex
- Amino(Y)
- Arg->GluSA(R)
- Cation:Cu[I](D)
- Cation:Cu[I](E)
- Cation:K(D)
- Cation:Na(D)
- Cation:Na(E)
- Deamidated(N)
- Deamidated(Q)
- Dehydrated(D)
- Dehydrated(E)
- Dehydrated(S)
- Dehydrated(T)
- Delta:H(4)C(2)(H)
- Dethiomethyl(M)
- Dioxidation(M)
- Dioxidation(W)
- iTRAQ4plex(H)
- iTRAQ4plex(K)
- iTRAQ4plex(S)
- iTRAQ4plex(T)
- iTRAQ4plex(Y)
- Methylthio(C)

G → A substitution
Dioxidation (W)
Hydroxy-proline
iTRAQ4plex(Ser)
MaxQuant is an integrated suite of algorithms specifically developed for high-resolution, quantitative MS data.

MaxQuant detects peaks, isotope clusters and stable amino acid isotope-labeled (SILAC) peptide pairs as three-dimensional objects in m/z, elution time and signal intensity space.

By integrating multiple mass measurements, mass accuracy in the p.p.b. range is achieved.

MaxQuant quantifies several hundred thousand peptides per SILAC-proteome experiment.

http://www.maxquant.org/
**PEAKS Options**

- **De novo Tools.**
  - No DB
  - Sequence but no ID
  - Protein ID
  - Sequence with ID
  - **inChorus**
    - high confidence Sequence ID
  - **SPIDER**
    - Sequence with putative ID
  - **PTM Finder**
    - PTMs
  - **PEAKS Q**
    - quantity

- Data in different formats
**PEAKS resources at MSI**

- **PEAKS Online**
  - Get a password from Tu.
  - Set up a search using peaklist.
  - Monitor your search status.
  - Links for .anz files that can be used further in PEAKS Client.

- **PEAKS Client v4.5**
  - PEAKS Client available at CGL: CPC7 and CPC10.
  - Use remote access to `cpc7.msi.umn.edu` or `cpc10.msi.umn.edu`
  - Use your .anz file (generated from Online search) for further analysis.
Statistical validation of peptide and protein identifications.
Scaffold

Statistical validation of peptide and protein identifications.

https://www.msi.umn.edu/sw/scaffold-for-pro
Proteomics Quantitation
iTRAQ™: Isobaric Tags for Relative and Absolute Quantification.

Isobaric Tag (Total mass = 145)

Reporter
Charged

Balance
Neutral loss

Peptide Reactive Group

Trypsin digest

Mix
MS

[Reporter-Balance-Peptide]

MS/MS

Mass (m/z)

% Intensity

QGQPILGGEASNDTWI
TTK
MaxQuant quantifies several hundred thousand peptides per SILAC-proteome experiment.
ProteinPilot supports normalization and quantification of iTRAQ datasets.
MRM

Targeted Proteomics

Selectivity, Sensitivity and Dynamic Range...

Proteins of Interest

Quantitative Proteomics Results → Prediction

Proteotypic Peptides

Choose and Optimize Transistions

Quantitative Analysis by MRM
Tranche is a free and open source file sharing tool that enables the storage of large amounts of data. Designed and built with scientists and researchers in mind, Tranche can handle very large data sets, is secure, is scalable, and all data sets are citable in scientific journals.
LAST WORD...

QUESTIONS?

Pratik Jagtap
pratik@msi.umn.edu
http://twitter.com/pratikomics

http://sitemaker.umich.edu/iwsmoi