Tutorial: Proteogenomics

Natalie Castellana

Department of Computer Science and Engineering, UCSD
Center for Computational Mass Spectrometry

https://proteomics.ucsd.edu

https://proteomics.ucsd.edu/LiveSearch

http://proteomics.ucsd.edu
Proteogenomics?

Using the proteome to answer questions about the genome

• Where are the protein coding genes located?
• What splice isoforms are functional for this gene?
• Is this gene’s product post-translationally processed?

http://proteomics.ucsd.edu
Gene Annotation

“The aim of genome annotation efforts is to determine the biochemical and biological function, if any, of each nucleotide in a genome.”

(Brent *Nature* 2008)

http://proteomics.ucsd.edu
Gene Annotation

Prokaryotic Genes

Operon

- Source of Evidence
  - Homology
  - Transcript sequencing
  - Genomic signals
Gene Annotation

Prokaryotic Genes

- Challenges
  - Translation start site (TSS)
  - Programmed frame shift
  - Post-translational processing
Gene Annotation

Eukaryotic Gene

- Challenges
  - Translation start site (TSS)
  - Post-translational processing
  - Splice sites

http://proteomics.ucsd.edu
Gene Annotation

Eukaryotic Gene

• How can the proteome help?
  – Translation start site (TSS)
  – Splicing
  – Translation frame
  – Pseudogenes

• Bonus
  – Post-translational modifications
  – Protein relative abundance

http://proteomics.ucsd.edu
Anatomy of a MS/MS Experiment

Sample of Proteins → Peptides → Mass Spectrometer → Spectra

[Diagram showing the process of converting a sample of proteins into spectra using a mass spectrometer.]

http://proteomics.ucsd.edu
Proteogenomic Outline

• Identify peptides from tandem mass spectra

• Map peptides to genomic locations

• Interpret collections of locations

• Deliver useful information to the community

http://proteomics.ucsd.edu
Peptide Identification – De novo

http://proteomics.ucsd.edu
Peptide Identification – Database search

>AT1G51370.2
MVGKKKTKICDKVSHEEDRISLPEPLISEILFHLSTKD
SVRTSALSTKWRYLWQSVPGLDLPYASNTNTIVSFVE
SFFDSHRDSWIRKLRDLGYYHDKYDLMSWIDAATT

>AT1G50920.1
MVQYNFKRITVPNGKEFVDIILSRTQRTPTVHVKGY
KINRLQFYMVRKVKTQTNFHAKLSAIIDFPRLEQIHPF
YGDLLHVLYNKDHYKLALGQVNTARNLISKIKDYV

>AT1G36960.1
MTRLPPKGGDLGDPFLTFIDLCVQVRIPLYLSELT
VSIAGTLGPILEMEFNQDTSTYVAFIRVKIRLVFIDRLRF
FRREEAAASINTTDQTHMTSSNDISPASIPQ

>AT1G75120.1
MAVRKEKVPFRECGLAVLVGIGFCVCILIPNFVN
FRSSKVASASONPERSPKMFKAEFAISEKNGELRQVSD
DLTEKVRLAEQKEVIAKP

http://proteomics.ucsd.edu
Goal: Discovery of novel coding regions

- Protein Databases containing novel coding sequence

<table>
<thead>
<tr>
<th>Source</th>
<th>Benefits</th>
<th>Shortcomings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ab initio predictions</td>
<td>• Full protein sequences</td>
<td>• Often only 1 transcript per</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Finds ‘predictable’ genes</td>
</tr>
</tbody>
</table>

http://proteomics.ucsd.edu
**Goal: Discovery of novel coding regions**

- **Protein Databases containing novel coding sequence**

<table>
<thead>
<tr>
<th>Source</th>
<th>Benefits</th>
<th>Shortcomings</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ab initio</em> predictions</td>
<td>• Full protein sequences</td>
<td>• Often only 1 transcript per</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Finds ‘predictable’ genes</td>
</tr>
<tr>
<td>Translated ESTs/RNASeq</td>
<td>• Restricted to regions that are transcribed</td>
<td>• Sampling bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Sequencing errors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Limited coverage</td>
</tr>
</tbody>
</table>
Goal: Discovery of novel coding regions

- Protein Databases containing novel coding sequence

<table>
<thead>
<tr>
<th>Source</th>
<th>Benefits</th>
<th>Shortcomings</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ab initio</em> predictions</td>
<td>• Full protein sequences</td>
<td>• Often only 1 transcript per gene • Finds ‘predictable’ genes</td>
</tr>
<tr>
<td>Translated ESTs/RNASeq</td>
<td>• Restricted to regions that are transcribed</td>
<td>• Sampling bias • Sequencing errors • Limited coverage</td>
</tr>
<tr>
<td>Related Proteomes</td>
<td>• Restricted to regions that are translated</td>
<td>• Mutations • Species-specific proteins are missed</td>
</tr>
</tbody>
</table>

http://proteomics.ucsd.edu
Goal: Discovery of novel coding regions

- Protein Databases containing novel coding sequence
  - 6-frame translation

- Shortcomings
  - Very, very large
  - Unable to encode spliced peptides

http://proteomics.ucsd.edu
Exon Splice Graph

- Properties
  - Path = transcript
  - Compact

- New search method required
- Available in InsPect package

http://proteomics.ucsd.edu
Exon Splice Graph

http://proteomics.ucsd.edu
Exon Splice Graph

http://proteomics.ucsd.edu
Exon Splice Graph

http://proteomics.ucsd.edu
Exon Splice Graph

http://proteomics.ucsd.edu
Exon Splice Graph

http://proteomics.ucsd.edu
Exon Splice Graph

http://proteomics.ucsd.edu
Exon Splice Graph

http://proteomics.ucsd.edu
Goal: Discovery of novel coding regions

- Protein Databases containing novel coding sequence
  - Splice graph

Predictions from:
- *ab initio* gene predictions (AUGUSTUS, GeneID, GLIMMER)
- Transcript sequences
- Related proteomes

http://proteomics.ucsd.edu
Goal: Discovery of novel coding regions

- Best coverage is to use a combination of the two

http://proteomics.ucsd.edu
Proteogenomic Outline

• Identify peptides from tandem mass spectra

• Map peptides to genomic locations

• Interpret collections of locations

• Deliver useful information to the community

http://proteomics.ucsd.edu
Genomic Location

- Map peptides to genomic location(s)

http://proteomics.ucsd.edu
Genomic Location

- Map peptides to genomic locations(s)
  - Shared peptides

Same peptide sequence

http://proteomics.ucsd.edu
Genomic Location

• Map peptides to genomic locations(s)
  – Shared peptides
  – Novel locations can only be determined using novel peptides.

http://proteomics.ucsd.edu
What is a ‘novel’ peptide

• **Our working definition:**
  – The peptide sequence cannot be derived from an annotated protein sequence.

• **Disclaimers:**
  – The peptide may appear in an EST/hypothetical protein/theoretical prediction which has not been incorporated into the gene annotation for the target organism.
  – The peptide may exists in a homologous protein in another species.

[http://proteomics.ucsd.edu](http://proteomics.ucsd.edu)
Finding peptide locations

- Identify all locations in the databases

http://proteomics.ucsd.edu
Finding peptide locations

- Map to locations on the genome
  - PepSplice
    - Combined with database search to find spliced peptides.

- Genomic Peptide Finder (GPF)
  - Uses *de novo* generated peptides to map to the genome, allowing for sequencing errors and splicing.

http://proteomics.ucsd.edu
Proteogenomic Outline

• Identify peptides from tandem mass spectra

• Map peptides to genomic locations

• Interpret collections of locations

• Deliver useful information to the community

http://proteomics.ucsd.edu
Novel Peptides and Locations

• Peptide clusters

http://proteomics.ucsd.edu
Novel Peptides and Locations

• Peptide clusters

Intragenic Cluster

http://proteomics.ucsd.edu
Novel Peptides and Locations

- Peptide clusters

http://proteomics.ucsd.edu
Proteogenomic Outline

• Identify peptides from tandem mass spectra
• Map peptides to genomic locations
• Interpret collections of locations
• Deliver useful information to the community

http://proteomics.ucsd.edu
Proteogenomic workflow

- Spectra
- Known Proteome
- Genome

DB Search → Determine Novelty

Confirmed Proteins

Confirmed Peptides

Novel Peptides

Novel Locations

Novel Events

Gene Prediction

Novel and Corrected Gene Models

Plant Proteomes

ESTs

Center for Computational Mass Spectrometry

http://proteomics.ucsd.edu
Creating gene models

Stanke et al. BMC Bioinf. 2006  
http://proteomics.ucsd.edu
Studies on Model Organisms

- Anopheles gambiae (Kalume et al. BMC Gen. 2005)
- Drosophila melanogaster (Brunner et al. Nat. Biot. 2007)
- Homo sapien (Tanner et al. Gen. Res. 2007)
- Arabidopsis thaliana (Castellana et al. PNAS 2008)
- Mycoplasma pneumoniae (Jaffe et al. Prot. 2004)
- Shewanella oneidensis (Gupta et al. Gen. Res. 2007)

http://proteomics.ucsd.edu
Then what?

- Get onboard with the gene annotators
  - TAIR9 used over 300 of our refined and novel models.

- Make accessible to scientists
  - Tracks on Genome Browsers

http://proteomics.ucsd.edu
Challenges to Proteogenomics

• Broad sampling of expressed proteins

• Large databases
  – Zea mays is about 3 Gbp

• Large datasets
  – 10-100 Million spectra

http://proteomics.ucsd.edu
Goal: Proteome Coverage

• Challenge: Limitations of sampling
  – Mass spectrometer can only sample most abundant peptides
  – Small dynamic range (3-4 orders of magnitude versus 6 in reality)
  – Not all peptides from a protein are ‘detectable’

Solution: Diverse sampling and sample preparation

http://proteomics.ucsd.edu
Goal: Identification of novel coding regions

Castellana et al. PNAS 2008

http://proteomics.ucsd.edu
Getting involved with the biologists

• ‘Analysis driven experimentation feedback loop’ (Brunner et al. Nat. Biot. 2007)

• Enrichment techniques
  – Phospho-purification with heavy metals
  – N-terminal labeling (Baudet et al. MCP)
  – Size exclusion
  – Organelles (Ram et al. Science 2005)
  – Basic proteins (Brunner et al.)
Searching Large Databases

- **In Arabidopsis thaliana: (135 Mbp)**
  - Proteome: 13 M AA
  - 6-frame translation: 60 M AA

- **In Zea mays: (3 Gbp)**
  - Splice graph: 68 M AA, 414 K splice junctions
  - 6-frame translation: 2 B AA

Most DB search tools will not work!!
CCMS Computer Cluster

Fully automated pipeline running on 352 cores
Database Filtering

- Scoring is expensive
  - 0.1 sec/PSM = 100,000 sec = 1.5 days for 1,000,000 spectra against a single peptide!

<table>
<thead>
<tr>
<th>Seq1</th>
<th>ARNDDQGGHILKMFKKLILLK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2</td>
<td>MFPVYWTSPNRAARNDCEHLL</td>
</tr>
<tr>
<td>Seq3</td>
<td>KMMMYYVPPSFSFMMILLEHG</td>
</tr>
<tr>
<td>Seq4</td>
<td>QEQGHHIILKKMFPSDDQQGH</td>
</tr>
<tr>
<td>Seq5</td>
<td>HKLMFPSTWYVDRNONNASSCE</td>
</tr>
<tr>
<td>Seq6</td>
<td>FFPFSTWWYVEQGHHDCCNE</td>
</tr>
</tbody>
</table>

http://proteomics.ucsd.edu
Database Filtering

- Scoring is expensive
  - 0.1 sec/PSM = 100,000 sec = 1.5 days for 1,000,000 spectra against a single peptide!

http://proteomics.ucsd.edu
Searching Large Databases

• Database Filtering
  – Parent mass
  – Multi-pass filtering (Craig and Beavis. Bioinf. 2004)

http://proteomics.ucsd.edu
Peptide Sequence Tags

Protein Sequence Tag: Prefix Mass

Tag Generation

- VCD: 71.0
- EED: 434.4
- ECD: 228.1
- VCK: 97.2
- RLK: 850.6
- RKN: 514.2

Filter

- AVCDNDCEQGHL
- NNRADKLMMFPPSS
- HIPSEEDMMFPS
- HDCEQQGKFPSS
- MNNRATSPWYY
- TWWVMFRLK
- QEEQRKNCCGHKMFS
- GGQECDN
- EEHKMMFPTWYVV

http://proteomics.ucsd.edu
Identifying Spectrum using Gapped Peptides

MS-GappedDictionary

Input

Database

... G P E P T I D E G ...

... 103, 57, 129, 57, 101, 113, 115, 129, 57 ...

Spectrum

186

271

244

166

289

141

168

226

103, 57, 129, 57

101, 113, 115, 129, 57

Jeong et al., RECOMB 2010

http://proteomics.ucsd.edu
Identifying Spectrum using Gapped Peptides

MS-GappedDictionary

186  271  244
168  289  244
168  307  226
168  141  166  226

Database
...  G  P  E  P  T  I  D  E  G ...
...  103, 57, 129, 57, 101, 113, 115, 129, 57 ...

No/Match

MS-BPM (Block Pattern Matching)

To be presented in RECOMB 2011
http://proteomics.ucsd.edu
# MS-BPM: Benchmarking

<table>
<thead>
<tr>
<th></th>
<th>De novo</th>
<th>DB search</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-Alignment</td>
<td>-</td>
<td>850,000 secs / mil²</td>
</tr>
<tr>
<td>Sequest</td>
<td>-</td>
<td>180,000 secs / mil²</td>
</tr>
<tr>
<td>InsPect</td>
<td>-</td>
<td>3,000 secs / mil²</td>
</tr>
<tr>
<td>MSGD + BPM (no modification)</td>
<td>0.2 secs / spec</td>
<td>306 secs / mil²</td>
</tr>
<tr>
<td>MSGD + MutBPM (one mutation)</td>
<td>0.2 secs / spec</td>
<td>3,500 secs / mil²</td>
</tr>
<tr>
<td>MSGD + ModBPM (one blind mod)</td>
<td>0.2 secs / spec</td>
<td>4,300 secs / mil²</td>
</tr>
</tbody>
</table>

- DB search time is measured in seconds to process one million spectra per one million characters of the database (secs / mil²).

[http://proteomics.ucsd.edu](http://proteomics.ucsd.edu)
Arabidopsis Score Distribution

~13 Mbp

http://proteomics.ucsd.edu
Maize Score Distribution

~2 Gbp

http://proteomics.ucsd.edu
Searching large databases

- **Decoy strategy**
  - Reduction in sensitivity

http://proteomics.ucsd.edu
Searching large databases

- **Decoy strategy**
  - Reduction in sensitivity

[Graph showing distribution of IDs from Decoy and Standard with Q-axis and arrows indicating Decoy IDs > Q and All IDs > Q.]
Searching large databases

- **Decoy strategy**
  - Reduction in sensitivity

\[
FDR = \frac{\text{Decoy IDs} > Q}{\text{All IDs} > Q}
\]

[http://proteomics.ucsd.edu](http://proteomics.ucsd.edu)
Searching large databases

- Database-independent scoring methods
  - MS-Generating Function

Kim et al. JPR 2008
Searching large databases

- Database-independent scoring methods
  - MS-Generating Function

![Diagram showing score distribution of all peptides with spectral probability formula]

\[ \text{Spectral Probability} = \frac{\text{IDs} > Q}{\text{All IDs}} \]

[http://proteomics.ucsd.edu]
Searching Large Datasets

• Spectrum Filtering
  – Clustering (Frank et al. JPR 2008)
  – Quality filters

http://proteomics.ucsd.edu
Spectral Clustering

http://proteomics.ucsd.edu
Spectral Clustering

Consensus Spectrum

http://proteomics.ucsd.edu
Maize Score Distribution
Clustered Distribution

Score Histogram (355804 True, 355678 False)

http://proteomics.ucsd.edu
Enter for Computational Mass Spectrometry

http://proteomics.ucsd.edu
Challenges of a large experiment

- Large collection of spectra

10% PSM FDR = 10% PID FDR

10% PSM FDR = 33% PID FDR

http://proteomics.ucsd.edu
• Let $C$ be the event that a region of the genome codes for protein.

\[ \Pr(C) = \Pr(P_1 \lor P_2) \]
Confidences

Assume peptide identifications are independent:

$$\Pr(C) = \Pr(P_1 \lor P_2) = 1 - \Pr(\neg P_1)\Pr(\neg P_2)$$

Spectrum identifications are not independent:

$$\Pr(P_1) = \max \left[ \Pr(S_{1,1}), \Pr(S_{1,2}) \right]$$
Searching large databases

- Decoy strategy

\[
FDR = \frac{\text{Decoy IDs} > Q}{\text{All IDs} > Q}
\]

http://proteomics.ucsd.edu
Intragenic Events

- FDR for events
  - Local FDR

1% FDR: 150,000 IDs

http://proteomics.ucsd.edu
Intragenic Events

- FDR for events
  - Local FDR

1% FDR: 150,000 IDs
5% FDR: 180,000 IDs

http://proteomics.ucsd.edu
Intragenic Events

• FDR for events
  – Local FDR

1% FDR: 150,000 IDs
5% FDR: 180,000 IDs
23% local FDR

http://proteomics.ucsd.edu
Intragenic Events

- FDR for events
  - Local FDR

http://proteomics.ucsd.edu
Confidences

- Let $C$ be the event that a region of the genome codes for protein.

\[ \Pr(C) = 1 - \Pr(\neg L_{1,1}) \Pr(\neg L_{2,1}) \]

[Taylor diagram]

[http://proteomics.ucsd.edu]
Confidences

\[ \Pr(C) = 1 - \Pr(\neg L_{1,1}) \Pr(\neg L_{2,1}) \]

Assume locations are equally likely

\[ \Pr(L_{1,1}) = \Pr(L_{1,2}) = \frac{1}{2} \Pr(P_1) \]

Spectrum identifications are not independent:

\[ \Pr(P_1) = \max \left[ \Pr(S_{1,1}), \Pr(S_{1,2}) \right] \]

http://proteomics.ucsd.edu
Novel Peptides and Locations

- Peptide clusters

Intragenic Cluster

http://proteomics.ucsd.edu
Intragenic Events

• Interpretation
  – Is the gene model incorrect, or did we just discover a novel isoform?

• Orthogonal evidence
  – Peptide evidence
  – Homology
  – Ab initio prediction
  – EST support
Novel Peptides and Locations

- Peptide clusters

http://proteomics.ucsd.edu
Intergenic Events

- **Interpretation**
  - Is this a novel gene or an extension of a known gene?

- **Orthogonal evidence**
  - Homology
  - *Ab initio* prediction
  - EST support

http://proteomics.ucsd.edu
Proteogenomic workflow

- DB Search
  - Known Proteome
  - Spectra
  - Genomic

- Gene Prediction
  - Plant Proteomes
  - ESTs

- Determine Novelty
  - Confirmed Peptides
  - Novel Peptides
  - Novel Locations

- Novel and Corrected Gene Models
  - Novel Events

http://proteomics.ucsd.edu
Proteome Annotation

- Wealth of proteomic information
  - Protein quantification
  - Post-translational modifications
  - Signal peptides
  - Cleaved proteins

http://proteomics.ucsd.edu
Using the *genome* to answer questions about the *proteome*.

- What is the sequence of the protein(s) in my sample, if I don’t have a genome or a proteome??
Comparative/Template Proteogenomics

Target Protein

Database Protein

http://proteomics.ucsd.edu
Comparative/Template Proteogenomics

Target Protein

Genome

http://proteomics.ucsd.edu
http://proteomics.ucsd.edu
Comparative Proteogenomics

• What if the genome is not available?

Do you have the **proteome** of a related species?  

Yes

- Modi (Na et al. MCP 2008)
- TagRecon (Dasari et al. JPR 2010)
- Champs (Liu et al. Bionf. 2009)
- MS-Align+ (Liu et al. in prep)

GenoMS (Castellana et al. MCP 2010)

CSPS (Bandeira et al. Nat. Biot. 2009)

http://proteomics.ucsd.edu
Comparative Proteogenomics

- What if the genome is not available?
  
  Do you have the **proteome** of a related species?

  Yes

  Modi (Na et al. MCP 2008)
  TagRecon (Dasari et al. JPR 2010)
  Champs (Liu et al. Bionf. 2009)
  MS-Align+ (Liu et al. in prep)

  No

  GenoMS (Castellana et al. MCP 2010)

  CSPS (Bandeira et al. Nat. Biot. 2009)

http://proteomics.ucsd.edu
Comparative Proteogenomics

- What if the genome is not available?

  Do you have the **proteome** of a related species?

  Yes

  - Modi (Na et al. MCP 2008)
  - TagRecon (Dasari et al. JPR 2010)
  - Champs (Liu et al. Bionf. 2009)
  - MS-Align+ (Liu et al. in prep)

  No

  - GenoMS (Castellana et al. MCP 2010)

  - CSPS (Bandeira et al. Nat. Biot. 2009)

http://proteomics.ucsd.edu
Comparative Proteogenomics

- What if the genome is not available?

Do you have the proteome of a related species?

Yes

- Modi (Na et al. MCP 2008)
- TagRecon (Dasari et al. JPR 2010)
- Champs (Liu et al. Bioinf. 2009)
- **MS-Align+ (Liu et al. in prep)**

No

- GenoMS (Castellana et al. MCP 2010)

- CSPS (Bandeira et al. Nat. Biot. 2009)

http://proteomics.ucsd.edu
Comparative Proteogenomics

- What if the genome is not available?

**Do you have the **proteome** of a related species?**

- **Yes**
  - Modi (Na et al. MCP 2008)
  - TagRecon (Dasari et al. JPR 2010)
  - Champs (Liu et al. Bionf. 2009)
  - MS-Align+ (Liu et al. in prep)

- **No**

**Do you have the **genome** of a related species?**

- **Yes**
  - GenoMS (Castellana et al. MCP 2010)

- **No**
  - CSPS (Bandeira et al. Nat. Biot. 2009)

http://proteomics.ucsd.edu
Comparative Proteogenomics

- What if the genome is not available?

  Do you have the **proteome** of a related species?
    - Modi (Na et al. MCP 2008)
    - TagRecon (Dasari et al. JPR 2010)
    - Champs (Liu et al. Bionf. 2009)
    - MS-Align+ (Liu et al. in prep)
  - No: GenoMS (Castellana et al. MCP 2010)

  Do you have the **genome** of a related species?
  - Yes: CSPS (Bandeira et al. Nat. Biot. 2009)
  - No: CSPS (Bandeira et al. Nat. Biot. 2009)

http://proteomics.ucsd.edu
Comparative Proteogenomics

- What if the genome is not available?

Database Methods

- Modi (Na et al. MCP 2008)
- TagRecon (Dasari et al. JPR 2010)
- Champs (Liu et al. Bionf. 2009)
- MS-Align+ (Liu et al. in prep)

De Novo Methods

- GenoMS (Castellana et al. MCP 2010)
- CSPS (Bandeira et al. Nat. Biot. 2009)

http://proteomics.ucsd.edu
Resources

• Splice graph construction and searching
  – proteomics.ucsd.edu/Inspect

• Proteogenomics workflow on CCMS website
  – Zea mays (Beta Testing)
  – Arabidopsis thaliana
  – Human

http://proteomics.ucsd.edu