Proteomics &
Integrated Omics

Parag Mallick
(paragm@stanford.edu)


CANARY CENTER
AT STANFORD
Topics

• A History Lesson
  – (why you should care about more than the transcripts)

• How does Proteomics Work...
  – Chromatography
  – Mass Spectrometry
  – Identification
  – Quantification
What is a gene?

• 1865 - Mendel Gene as a discrete unit of heredity
  (Versuche über Pflanzenhybriden)

  – If you cross some peas…

The characters which were selected for experiment relate:

1. To the *difference in the form of the ripe seeds.*
2. To the *difference in the color of the seed albumen* (endosperm). The albumen of the ripe seeds is either pale yellow, bright yellow and orange colored, or it possesses a more or less intense green tint. This difference of color is easily seen in the seeds as their coats are transparent.
3. To the *difference in the color of the seed–coat.*
4. To the *difference in the form of the ripe pods.*
5. To the *difference in the color of the unripe pods.* They are either light to dark green, or vividly yellow, in which coloring the stalks, leaf–veins, and calyx participate.*
6. To the *difference in the position of the flowers.*
7. To the *difference in the length of the stem.*

*Variations in traits were caused by variations in inheritable factors (or, in today’s terminology, phenotype is caused by genotype).*

*Genome Res. 2007 17: 669-681*
What is a gene

- Elemente der exakten erblichkeitslehre. Deutsche wesentlich erweiterte ausgabe in fünfundzwanzig vorlesungen (1909)
  Wilhelm Johannsen

1) the GENE, which is an abbreviation of Darwin’s and De Vries’ pangene from Greek gennao, to breed;
2) the PHENOTYPE, from Greekphain-omai, to appear;
3) the GENOTYPE, from gennao, to breed, and typos, a type or imprint.
What is a gene?

- 1909 Archibald Garrod - genetic defects cause many inherited diseases

- He suggested (correctly) that alkaptonuria results from a single recessive gene, which causes a deficiency in the enzyme that normally breaks down alkapton.
What is a gene?

1940s: Gene as a blueprint for a protein

- Beadle and Tatum (1941), who studied *Neurospora* metabolism, discovered that mutations in genes could cause defects in steps in metabolic pathways. This was stated as the “one gene, one enzyme”
Enter the misunderstood central dogma
How do cells work?

- Duh - according to the central dogma.

- Genes make transcripts which make proteins.
  - Proteins go do work and cells do stuff.
A simple model for how cells work

Cell Motility is controlled by ‘Gene X’

Transcript Of gene X → Abundance of Protein X → Motility
Hmmm...but that doesn’t work all the time...

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Description</th>
<th>Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene location and structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intronic genes</td>
<td>A gene exists within an intron of another (Henikoff et al. 1986)</td>
<td>Two genes in the same locus</td>
</tr>
<tr>
<td>Genes with overlapping reading frames</td>
<td>A DNA region may code for two different protein products in different reading frames (Contreras et al. 1977)</td>
<td>No one-to-one correspondence between DNA and protein sequence</td>
</tr>
<tr>
<td>Enhancers, silencers</td>
<td>Distant regulatory elements (Spilanakis et al. 2005)</td>
<td>DNA sequences determining expression can be widely separated from one another in genome. Many-to-many relationship between genes and their enhancers.</td>
</tr>
<tr>
<td>Structural variation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile elements</td>
<td>Genetic element appears in new locations over generations (McClinstock 1948)</td>
<td>A genetic element may be not constant in its location</td>
</tr>
<tr>
<td>Gene rearrangements/structural variants</td>
<td>DNA rearrangement or splicing in somatic cells results in many alternative gene products (Early et al. 1980)</td>
<td>Gene structure is not hereditary, or structure may differ across individuals or cells/tissues</td>
</tr>
<tr>
<td>Copy-number variants</td>
<td>Copy number of genes/regulatory elements may differ between individuals (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005)</td>
<td>Genetic elements may differ in their number</td>
</tr>
<tr>
<td>Epigenetics and chromosome structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epigenetic modifications, imprinting</td>
<td>Inherited information may not be DNA-sequence based (e.g., Dobrovic et al. 1988); a gene’s expression depends on whether it is of paternal or maternal origin (Sager and Kitchin 1975)</td>
<td>Phenotype is not determined strictly by genotype</td>
</tr>
<tr>
<td>Effect of chromatin structure</td>
<td>Chromatin structure, which does influence gene expression, only loosely associated with particular DNA sequences (Paul 1972)</td>
<td>Gene expression depends on packing of DNA. DNA sequence is not enough to predict gene product.</td>
</tr>
<tr>
<td>Post-transcriptional events</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternative splicing of RNA</td>
<td>One transcript can generate multiple mRNAs, resulting in different protein products (Berget et al. 1977; Gelasins and Roberts 1977)</td>
<td>Multiple products from one genetic locus; information in DNA not linearly related to that on protein</td>
</tr>
<tr>
<td>Alternatively spliced products with alternate reading frames</td>
<td>Alternative reading frames of the INK4a tumor suppressor gene encodes two unrelated proteins (Quelle et al. 1995)</td>
<td>Two alternative splicing products of a pre-mRNA produce protein products with no sequence in common</td>
</tr>
<tr>
<td>RNA trans-splicing, homotypic trans-splicing</td>
<td>Distant DNA sequences can code for transcripts ligated in various combinations (Borst 1986). Two identical transcripts of a gene can trans-splice to generate an mRNA where the same exon sequence is repeated (Takahara et al. 2000)</td>
<td>A protein can result from the combined information encoded in multiple transcripts</td>
</tr>
<tr>
<td>RNA editing</td>
<td>RNA is enzymatically modified (Eisen 1988)</td>
<td>The information on the DNA is not encoded directly into RNA sequence</td>
</tr>
<tr>
<td>Post-translational events</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein splicing, viral polyproteins</td>
<td>Protein product self-cleaves and can generate multiple functional products (Villa-Komaroff et al. 1975)</td>
<td>Start and end sites of protein not determined by genetic code</td>
</tr>
<tr>
<td>Protein trans-splicing</td>
<td>Distinct proteins can be spliced together in the absence of a trans-spliced transcript (Handa et al. 1996)</td>
<td>Start and end sites of protein not determined by genetic code</td>
</tr>
<tr>
<td>Protein modification</td>
<td>Protein is modified to alter structure and function of the final product (Wold 1981)</td>
<td>The information on the DNA is not encoded directly into protein sequence</td>
</tr>
<tr>
<td>Pseudogenes and retrogenes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retrogenes</td>
<td>A retrogene is formed from reverse transcription of its parent gene’s mRNA (Varin et al. 1980) and by insertion of the DNA product into a genome</td>
<td>RNA-to-DNA flow of information</td>
</tr>
<tr>
<td>Transcribed pseudogenes</td>
<td>A pseudogene is transcribed (Zheng et al. 2005, 2007)</td>
<td>Biochemical activity of supposedly dead elements</td>
</tr>
</tbody>
</table>
Flow of Biological Information

DNA → RNA → Protein → Functional Protein

Transcriptional Control
- mRNA secondary structure
- mRNA stability
- uRNA

Translational Control
- Protein processing
- Protein modification

Post translational Control
- Protein stability/turnover
- Protein complex formation

DNA Structure

Chromatin Organization
Simultaneous mRNA & Proteins Analysis at the Single-Cell Level

Transcription → Translation

Immunostaining

Single Cell RT-PCR

mRNA

Protein

Cell Loading

Protein Quantification

mRNA Quantification
cMET Overexpression-based resistance

- **cMET mRNA Expression**
  - HCC827 parental
  - HCC827 overexpressed

- **Protein Abundance**
  - HCC827
  - HCC827 MET

- **Graph**
  - Erlotinib (µM)
  - % of control cell growth rate

- **Comparison**
  - mRNA Expression
    - cMET
      - HCC827 parental: 28.68% (34.38%)
      - HCC827 overexpressed: 23.65% (74.60%)

- **Legend**
  - HCC827 Parental
  - HCC827 Overexpressed
Integrated Omics

• At a fundamental level we do not understand how information is transferred and processed in biological system.

• Through mysterious processes, cells are able to take signals from their environment, process those signals and then act.

• Unfolding this mystery of information transfer in biological systems is a critical challenge to modern biology.

• Integrated omics asserts that one must look at lots of different stuff to get a complete picture of how cells are regulated
Time Scales of Cellular Regulation

Characteristic Space [m]

10^{-9} Nanometer
10^{-6} Micrometer
10^{-3} Millimeter
10^{0} Meter

System

Protein Modification
Protein Signaling
Gene Expression
Cell-Cell Communication
Organ Growth
Gene Mutations

Process

10^{0} Seconds
10^{2} Hours
10^{4} Days
10^{6} Months
10^{8} Years

Characteristic Time [s]
Proteomics Basics

The diverse properties of proteins are biologically informative.

Proteomics attempts to (quantitatively) measure the diverse properties of proteins.

Through these quantitative proteomics measurements, we hope to better characterize cancer systems to understand oncogenesis, transformation, therapeutic action and identify protein biomarkers.
What is a “proteome”

The "proteome" is the PROTEin complement of a genOME.

The concept of “the proteome” is fundamentally different to that of “the genome.”

The genome is virtually static and can be well-defined for an organism, the proteome continually changes in response to external and internal events, disease, drugs, mood.

Example Proteomes:

Human Proteome: All human proteins encoded by the genome (in all their modification states, e.g. phosphorylated).

Cell Surface Proteome

Plasma Proteome: All the plasma proteins.
The Biochemical Approach

Protein mixture

Separate and assay

Purify product to homogeneity

Assay structure and function

Sequential cycles of purification and assay
Technologies for Global (quantitative) Analysis

- Genomic DNA → mRNA → Protein products → Functional protein
- Activity profiling
- Protein phosphorylation analysis: inferred activity
- Arrays methods
- Sequencing
- Quantitative protein profiling
- Protein linkage maps (dynamic)
- Protein linkage maps (catalogue)
Observations in Systems Studies

Integrated Genomic & Proteomic Analyses of a Systematically Perturbed Metabolic Network
Ideker, Aebersold, Hood

FUNCTIONAL CATEGORIES
- Galactose-induced metabolic genes
- Respiratory genes
- Ribosomal protein genes
mRNA and Protein Levels & Half-Lives

What is Proteomics?

The "proteome" is the set of PROTEins composing a system. Proteomics is the systematic (quantitative) analysis of the proteins expressed in a “system” at a time.

Why Use a Proteomic Approach?

**Pros**
- Flexibility - multiple approaches
- Multiple types of information on proteins states – amino acid sequence, PTMs, protein-protein and protein-cofactor interactions and complexes, subcellular localization
- Global - system-wide

**Challenges**
- Proteome complexity (e.g. PTMs) requires targeted approaches - focus on informative sub-proteomes
- Detection limitations - typically detect more abundant proteins
- Expensive and requires high level of expertise / collaboration
# The Range of Proteomics

<table>
<thead>
<tr>
<th>Technical expertise</th>
<th>Protein analysis</th>
<th>Protein complexes</th>
<th>Protein networks</th>
<th>Cell culture models</th>
<th>Translational studies</th>
<th>Population proteomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purity and identity</td>
<td><em>In vitro</em> biochemistry</td>
<td><em>In vitro</em> biology</td>
<td><em>In vitro</em> biology</td>
<td><em>In vitro</em> target or marker discovery</td>
<td><em>In vivo</em> marker discovery</td>
<td><em>In vivo</em> marker discovery or verification</td>
</tr>
<tr>
<td>Single PTM</td>
<td>Purity and identity</td>
<td>Single PTM</td>
<td>Quantitative PTM</td>
<td>Multiple PTM</td>
<td>PTM stoichiometry</td>
<td>Splicing, SNPs</td>
</tr>
<tr>
<td>Complex composition</td>
<td>Complex composition</td>
<td>Complex dynamics</td>
<td>Pathway crosstalk</td>
<td>Network dynamics</td>
<td>Complex stoichiometry</td>
<td>Spatial organization</td>
</tr>
<tr>
<td>Network composition</td>
<td>Network composition</td>
<td>Expression profiling</td>
<td>Organelle composition</td>
<td>Cross model analysis</td>
<td>Comprehensive PTM</td>
<td>MALDI imaging</td>
</tr>
<tr>
<td>PTM analysis</td>
<td>PTM analysis</td>
<td>Pathway crosstalk</td>
<td>Organelle composition</td>
<td>Perturbation analysis</td>
<td>Protein activity</td>
<td>Biofluid composition</td>
</tr>
<tr>
<td>Complex dynamics</td>
<td>Complex dynamics</td>
<td>Pathway crosstalk</td>
<td>Organelle composition</td>
<td>Perturbation analysis</td>
<td>Protein activity</td>
<td>Biofluid composition</td>
</tr>
<tr>
<td>Network dynamics</td>
<td>Network dynamics</td>
<td>Expression profiling</td>
<td>Organelle composition</td>
<td>Perturbation analysis</td>
<td>Protein activity</td>
<td>Biofluid composition</td>
</tr>
<tr>
<td>PTM stoichiometry</td>
<td>Complex stoichiometry</td>
<td>Comprehensive PTM</td>
<td>Protein activity</td>
<td>Cross model analysis</td>
<td>Protein activity</td>
<td>Biofluid composition</td>
</tr>
<tr>
<td>Splicing, SNPs</td>
<td>Spatial organization</td>
<td>Comprehensive PTM</td>
<td>Protein activity</td>
<td>Cross model analysis</td>
<td>Protein activity</td>
<td>Biofluid composition</td>
</tr>
</tbody>
</table>
Lots of Options in the Pipeline

- Metabolic labeling (SILAC, $^{15}$N)
- Chemical protein labeling (ICPL)
- Chemical peptide labeling (iCAT, cICAT, iTRAQ, TMT, methylation, esterification)
- Enzymatic peptide labeling ($^{18}$O)
- Absolute quantification (AQ2A, QconCAT)
- Label-free quantitation (spectrum counting, omPAI, APEX, XICs, expression)
- Single/multiple reaction monitoring (SRM, MRM)
- Express, Pepper, MSQuant, MaxQuant, iTracker, TPR, CPAS, TOPP, ProteoWizard

- Biopsy
- Biofluid
- Laser-capture microdissection
- Cell sorting (FACS)
- Primary cell culture
- Stable cell line culture
- Free-flow electrophoresis
- Gradient centrifugation

- Database searching
- De novo sequencing
- Peptide mass fingerprinting (PMF)
- Accurate mass and time tag (AMT)
- Mascot, Sequest, X!Tandem
- OMSSA, Phenyx, Spectrum Mill
- PEAKS, PepNovo, InsPecT, PTM Score, A-Score, ModifiComb

- Protein quantification
- Sample extraction
- Protein identification
- Mass spectrometry
- Protein fractionation
- Peptide fractionation
- 1D and 2D gel electrophoresis
- Isoelectric focusing
- Capillary electrophoresis
- Column chromatography
- Immunoprecipitation
- Pull-downs with tagged proteins
- Cell surface labeling
- Active site labeling
- Affinity depletion
- Phosphoflow
- Glyco-capture

- Electro spray ionization (ESI)
- Matrix-assisted laser desorption/ionization (MALDI)
- Time-of-flight MS (TOF)
- Ion trap MS
- Quadrupole MS
- Orbitrap MS
- Fourier-transform ion cyclotron MS (FT-ICR)
- Liquid chromatography MS (LC-MS)
- Imaging MS
- Ion mobility MS
- Tandem mass spectrometry (MSn)
- Collision-induced dissociation (CID)
- Electron-transfer dissociation (ETD)
- Electron-capture dissociation (ECD)
- Post-source decay (PSD)
Proteomic Pipeline

Sample Prep
- Protein mixture

MS Analysis
- Trypsin
- Peptides
- 1D, 2D, 3D peptide separation
- Tandem mass spectrum

Interpretation
- Correlative sequence database searching
- Protein identification
- Theoretical
- Acquired
Sample Preparation

- Isolate your sub-proteome of interest
  - Cell Disruption
  - Removal of NAs., lipids, salt
  - Protein Solubility
  - Do you require native proteins?

- Prevent modification of your sample.
  - Degradation
  - PTMs

- Fractionate (intact)
- Digest (trypsin protease)
- Fractionate (peptides)
Liquid chromatography (LC)
Liquid chromatography

1. Starting Conditions
2. Load (adsorption)
3. Elute (desorption)
Liquid chromatography

Hydrophilic

Hydrophobic

Time

% eluent

Polyimide Coating

Polished Edge

Frit

Bare Fused Silica

Packed Bed

Fused-silica Tubing, 50 cm
360 μm ID
Proteomic Pipeline

Sample Prep
Protein mixture

MS Analysis

Interpretation

Theoretical
Protein mixture

Peptides
1D, 2D, 3D peptide separation

Q1
Q2 Collision Cell
Q3

Correlative sequence database searching

Protein identification

Theoretical

Acquired

Tandem mass spectrum

m/z
Time (min)
Topics

- Proteomics
  - overview & basics
  - mass spectrometry
  - identification
  - quantification
  - cancer proteomics
  - phospho-proteomics
Why Study Proteins by Mass Spectrometry?

Mass Spectrometers are fancy bathroom scales. They tell you an analytes Molecular Weight (MW) from which you can infer elemental composition. Can also give you chemical structural information (e.g. amino acid sequence).

MS can unambiguously identify proteins

- Gel separated proteins
- Proteins in mixtures
- Protein:protein associations

MS can identify the precise site of a post-translational modification

- Phosphorylation
- N- or C-terminal modification
- Glycosylation
- Many more

MS can quantify the abundance of proteins
How does it work?

A set of molecules is ionized, separated according to mass/charge (m/z) ratio and detected.
MALDI is a method that allows for vaporization and ionization of non-volatile biological samples from a solid-state phase directly into the gas phase.
Chemical Surfaces

- (Hydrophobic)
- (Anionic)
- (Cationic)
- (Metal Ion)
- (Normal Phase)

Biological Surfaces

- (PS10 or PS20)
- (Antibody - Antigen)
- (Receptor - Ligand)
- (DNA - Protein)
Electrospray Ionization (ESI)
Electrospray versus MALDI

**LC-ESI**
- Liquid phase ionization
- Deeper Ionization
- Time-limited – entire sample is consumed during analysis
- Usually used with traps, quadropole analyzers

**MALDI**
- Solid phase ionization
- Sample is consumed slowly, resampling is possible
- Usually used with TOF analyzers
How does it work?

A set of molecules is ionized, separated according to mass/charge (m/z) ratio and detected.
How does it work?

Create gas phase ions
Mass Spectrometry: Time of Flight

Turn on power...

...and start clock
Mass Spectrometry: Time of Flight
Mass Spectrometry: Time of Flight
Mass Spectrometry: Time of Flight
Mass Spectrometry: Time of Flight

Note Time
How does it work?

A set of molecules is ionized, separated according to mass/charge (m/z) ratio and detected.

Note: Intensity of signal is *related* to abundance. That relationship is consistent for a particular peptide. That relationship will vary from peptide to peptide.
Introducing Multidimensional Separation

A set of molecules is separated by hydrophobicity, ionized, separated according to mass/charge (m/z) ratio and detected.
A Standard LC-MS Experiment

I

Protein mixture

Proteolysis → Peptides → Ion Exchange/PAGE → RPLC → 1D, 2D, 3D peptide separation

II

Time 1 (11 min) → Time 2 (12 min)

Q1

Mass spectrum

200 400 600 800 1000 1200

m/z
3 Dimensions of Data

Mass vs. Intensity vs. Time

MS scans
Topics

• Proteomics
  – overview & basics
  – mass spectrometry
  – identification
  – quantification
  – cancer proteomics
  – phospho-proteomics
What kind of info can mass spec give you?

Molecular weight (MW)

Elemental composition

Chemical structural information (e.g. amino acid sequence)
Identification by Mass: A Practical Example

5.000g  5.670g
Tandem Mass Spectrometry

- Digestion
- Ionization
- Isolation
- Fragmentation
- Mass Analysis

Protein → Digestion → Peptides

Ionization:
- Q2 Collision Cell
- Q3 Mass Analysis

Fragmentation:
- peptide fragments
- m/z 200, 400, 600, 800, 1000, 1200

Peptide Digestion → Peptide Fragments
Tandem Mass Spectrometry

Protein Digestion

Peptides

Ionization

Isolation

Fragmentation

Mass Analysis

Mass Spectrum

m/z
Tandem Mass Spectrometry

LLICK 444.2275
ALLICK 557.3116
MALLICK 670.3956
GMALLICK 773.4048
AGMALLICK 919.5104
...

Tandem mass spectrum
Mainly b-ions and y-ions
(breakage at the peptide (weakest) bond),
but other fragments too
MS/MS sequencing

CDK2  Y15\(^\text{P}\)

IGEGTY\(^\text{P}\)GVVYK

<table>
<thead>
<tr>
<th>m/z, amu</th>
<th>Intensity, counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>171.1214</td>
<td>11</td>
</tr>
<tr>
<td>147.1234</td>
<td></td>
</tr>
<tr>
<td>187.0807</td>
<td></td>
</tr>
<tr>
<td>310.1878</td>
<td></td>
</tr>
<tr>
<td>409.2501</td>
<td></td>
</tr>
<tr>
<td>147.1234</td>
<td></td>
</tr>
<tr>
<td>187.0807</td>
<td></td>
</tr>
<tr>
<td>310.1878</td>
<td></td>
</tr>
<tr>
<td>409.2501</td>
<td></td>
</tr>
</tbody>
</table>

b ions: IG, E, G, T

y ions: K, Y, V, V, G, pY, T, G, E, G

MS/MS sequencing

Identification Break

Mapping from a spectrum to a peptide!

– Take your spectrum (find_amino_acids.pdf) and try to determine an amino acid in your spectrum
  (hint – your AA mass table may help)
Making Sense of Our Spectrum

<table>
<thead>
<tr>
<th>b</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 K 9</td>
<td>182.0812</td>
</tr>
<tr>
<td>2 A 8</td>
<td>838.4305</td>
</tr>
<tr>
<td>3 S 7</td>
<td>767.3934</td>
</tr>
<tr>
<td>4 A 6</td>
<td>680.3614</td>
</tr>
<tr>
<td>5 D 5</td>
<td>609.3243</td>
</tr>
<tr>
<td>6 L 4</td>
<td>494.2973</td>
</tr>
<tr>
<td>7 A 3</td>
<td>381.2132</td>
</tr>
<tr>
<td>8 K 2</td>
<td>310.1761</td>
</tr>
<tr>
<td>9 Y 1</td>
<td>182.0812</td>
</tr>
</tbody>
</table>

Theoretical Peak Table:
- b-H: 948.5149
- b-NH: 949.4989
- MH: 966.5255
- Low Mass:
  - 84.0808
  - 101.1073
  - 129.1022
  - 126.0913
- N-terminal:
  - b-NH: 183.1128
  - b-H: 200.1394
- C-terminal:
  - y-NH: 821.4040
  - y-H: 820.4199

Graph showing m/z values and peaks at various m/z ratios.
Model Spectrum

The diagram shows a model spectrum with various peaks labeled as b1, b2, b3, b4, b5, b6, b7, b8, b9, b10, and b11. The peaks are associated with specific charges and masses, with the mass-to-charge (M/z) values ranging from 0 to 1000. The relative intensity is indicated on the y-axis, ranging from 0% to 100%. The spectrum includes labeled peptides with amino acid sequences KEVLSI and ELVSLI, with a mass of 1228.73 AMU + 2 H (Parent Error: 0 ppm).
Model Spectrum

1228.73 AMU +2 H (Parent Error: 0 ppm)

Relative Intensity

M/z

0% 250 500 750 1000

0% 50% 100%

M/z

0% 250 500 750 1000

0% 50% 100%
Mapping from a spectrum to a peptide!

1) Digest a protein.
   1) Go to: [http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msdigest](http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msdigest)
   2) Put this sequence in:

   ```
   >IPI:IPI00791964.1  Tax_Id=9606  ALCAM protein
   MESKGASSCRLLCLLISATVFRPGLGWYTVNSAYGDITIIIPCRLDPQNLMEFKYK
   DGSPVFIAFRSSTKSVQYDDVPEYKDRNLSNLSINARISDEKRFVCMVTE
   VFEAPTIVKVF
   ```
   3) Press the ‘Perform Digest’ Button

2) In a new tab go to: [http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct](http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct)
   1) Enter ‘KASADALAKY’
   2) Click on the top non-n-terminal peptide from your previous digest ((K)DGSPVFIAFR(S))

3) Take your next mystery spectrum and try to determine which of your possible theoretical spectra match.

4) Go to: [http://www.matrixscience.com/search_form_select.html](http://www.matrixscience.com/search_form_select.html)
   1) Click MS/MS Ion Search
   2) Under ‘Data file’ click ‘Choose File’ and upload ‘msms.dta’
   3) Data Format, Sequest (dta)
   4) Click ‘start search’
   5) Under the first hit, Under ‘query’ - Click on ‘1’
Identification Break

You identified a peptide – which protein was in your experiment?

1) Look in Sequences.txt

2) If you identified the peptide ‘PEPTIDEK’ which protein was in your sample?

   GTPase NRas precursor, GTPase HRas precursor, PD-ECGF/TP protein, acheron isoform 2, hypoxia-inducible factor 1, alpha, Beta-nerve growth factor precursor
   
   Hmmm….so NOT MAX, ALCAM

3) We have digested those sequences and made the file Sequences.digestEdit.txt

4) If you identified the peptide ‘DSDDVPMVLVGNK’ which protein was in your sample?

   GTPase NRas precursor, GTPase HRas precursor

5) If you identified the peptides ‘DSDDVPMVLVGNK’ and ‘QAQDLAR’ which protein was in your sample?

   DSDDVPMVLVGNK: GTPase NRas precursor, GTPase HRas precursor
   QAQDLAR: Beta-nerve growth factor, GTPase HRas precursor

6) If you identified the peptide ‘VNIQVEK’ which protein was in your sample?

   hypoxia-inducible factor 1, alpha
Ionization - proteotypic peptides (good flyers and bad flyers)

- **MS-compatible peptides**
  - **MS platform dependent** (e.g. MALDI TOF vs. ESI-Ion Trap)

- **Observed peptides**

- **Proteotypic peptides**

- **Frequency analysis**

- **Prediction**

- 494 amino acid–associated physico-chemical property scales

---

P. Mallick et al.  
25:125

Kuster et al.  
Nat Rev Mol Cell Bio  
6:577 (2005)
Topics

• Proteomics
  – overview & basics
  – mass spectrometry
  – identification
  – Quantification
Quantitation

Differential isotope labeling (light & heavy)
- ICAT
- SILAC
- iTRAQ

Label-free quantitation
(Absolute quantitation using synthetic peptides, AQUA)
Stable Isotope Labeling Strategies

**PROTEIN LABELING**
- Metabolic stable isotope labeling
- Isotope tagging by chemical reaction

**DATA COLLECTION**
- Mass spectrometry

**DATA ANALYSIS**
- Intensity
- m/z

**Light LYS**
- H₂N
- COOH

**Heavy LYS**
- H₂N
- COOH

**Stable Isotope Labeling Strategies**

Bioinorganic mass encoded tether
Functional group
ICAT reagent X = H (D) or D (H)
Label-free quantitation

$p = 4 \times 10^{-8}$
What is Proteomics?

The "proteome" is the set of PROTEins composing a system.

Proteomics is the systematic (quantitative) analysis of the proteins expressed in a “system” at a time.

<table>
<thead>
<tr>
<th>Objective</th>
<th>Proteomics as</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery of new entities</td>
<td>Biol. or Clin. assay:</td>
<td>Comprehensive</td>
</tr>
<tr>
<td>Assay for known entities</td>
<td>Proteome analyzed multiple (infinite) times</td>
<td></td>
</tr>
<tr>
<td>Restricted or targeted</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proteome as database:
- Proteome analyzed once

Example
- HUPO PPP
- Expression profiling
- PTM discovery
- Protein complexes
- Biomarker candidate verification

Approach
- Peptide sequencing, database searching
- Peptide identification and quantification
- SRM or MRM Peptide quantification
• In SRM/MRM assays the first (Q1) and last (Q3) mass analyzers of a triple quadrupole mass spectrometer are used as mass filters to isolate a peptide ion and a corresponding fragment ion.

• The signal of the fragment ion is then monitored over the chromatographic elution time.

• The selectivity resulting from the two filtering stages, combined with the high duty cycle, results in increased sensitivity and quantitative analyses.

• The specific pairs of m/z values associated to the precursor and fragment ions selected are referred to as "transitions" and effectively constitute mass spectrometric assays that allow to identify and quantify a specific peptide and, by inference, the corresponding protein in a complex protein digest.
The process of establishing a SRM/MRM assay for a protein

- Overall, this is a lengthy and iterative process, but, once an MRM assay for a protein is established, it becomes universally useful.

- The tedious assay development process needs to be performed only once, for a given type of mass spectrometer and fragmentation mechanism.

- There exist software and a website that can make this process easier and help with validation.
  - Skyline (MacCoss Lab)
  - proteowizard
  - MRMAtlas
Other omics...

- Metabolomics
- Glycomics
- Lipidomics
- ...

Key to interpretability is having a model for how the different entities relate to each other