

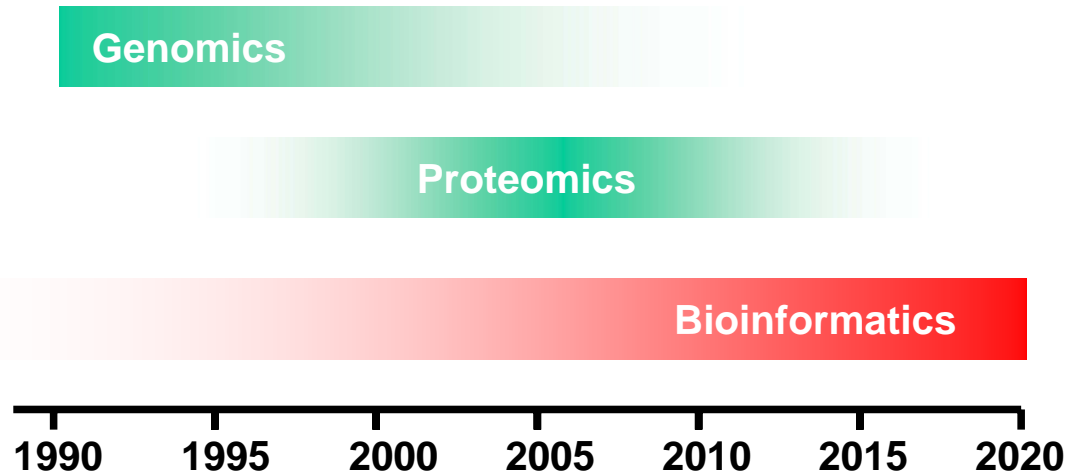
# Proteomics & Bioinformatics Part I

David Wishart  
University of Alberta

## What is Proteomics?

- **Proteomics** - *A newly emerging field of life science research that uses High Throughput (HT) technologies to display, identify and/or characterize all the proteins in a given cell, tissue or organism (i.e. the proteome).*

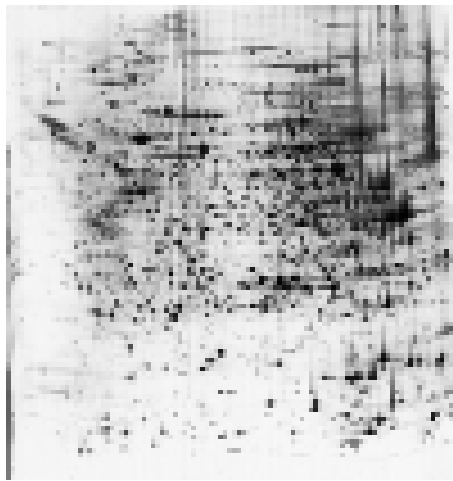
# Proteomics & Bioinformatics



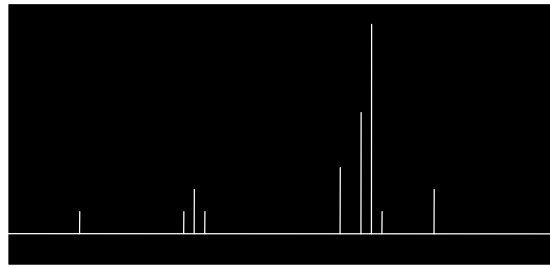
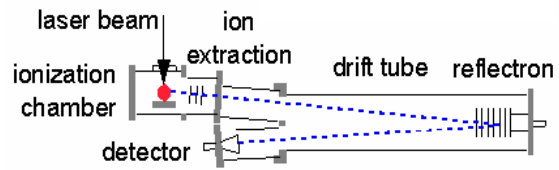
## 3 Kinds of Proteomics

- **Structural Proteomics**
  - High throughput X-ray Crystallography/Modelling
  - High throughput NMR Spectroscopy/Modelling
- **Expressional or Analytical Proteomics**
  - Electrophoresis, Protein Chips, DNA Chips, 2D-HPLC
  - Mass Spectrometry, Microsequencing
- **Functional or Interaction Proteomics**
  - HT Functional Assays, Ligand Chips
  - Yeast 2-hybrid, Deletion Analysis, Motif Analysis

# Expressional Proteomics

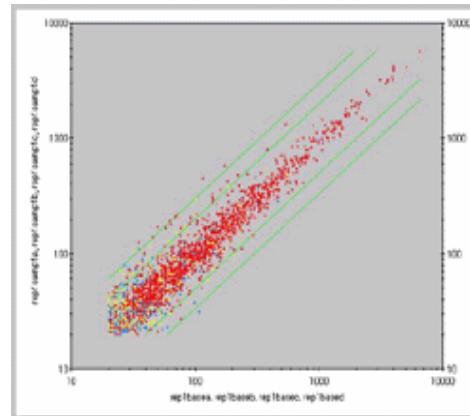
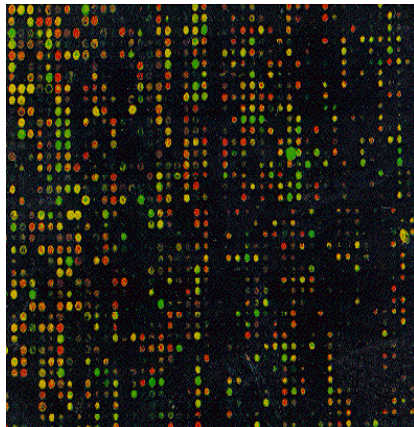


2-D Gel



QTOF Mass Spectrometry

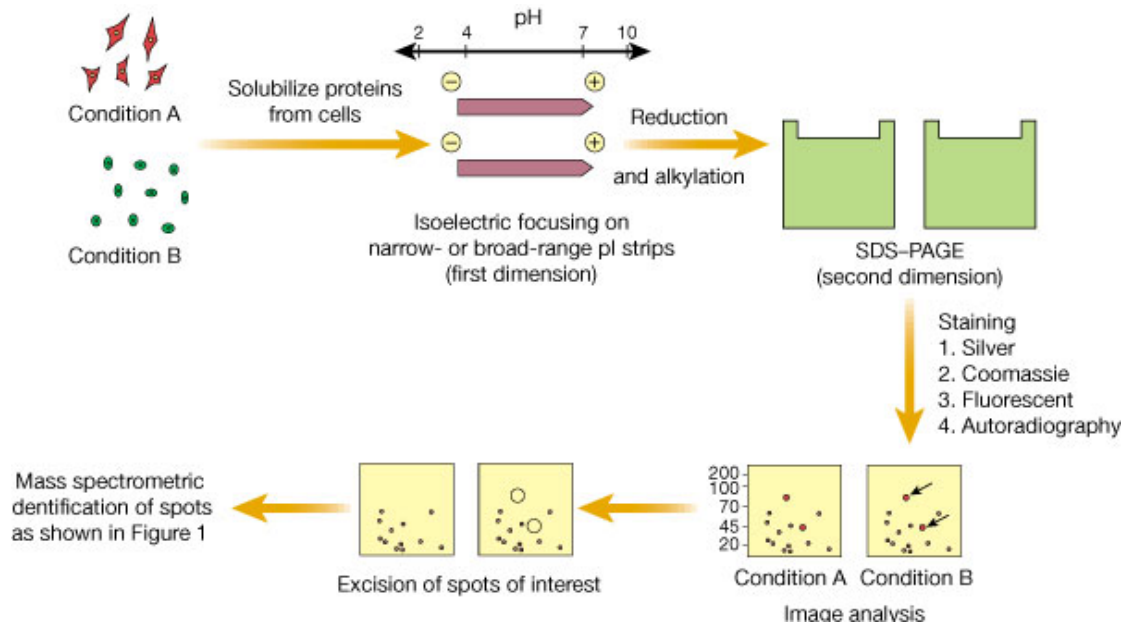
# Expressional Proteomics



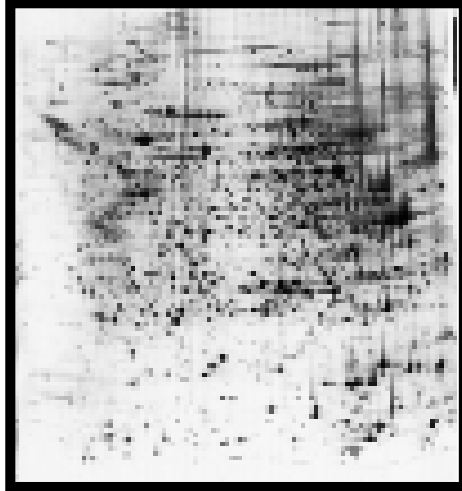
# Expressional Proteomics

- To separate, identify and quantify protein expression levels using high throughput technologies
- Expectation of 100's to 1000's of proteins to be analyzed
- Requires advanced technologies and plenty of bioinformatics support

## Electrophoresis & Proteomics



# 2D Gel Electrophoresis



- **Simultaneous separation and detection of ~2000 proteins on a 20x25 cm gel**
- **Up to 10,000 proteins can be seen using optimized protocols**

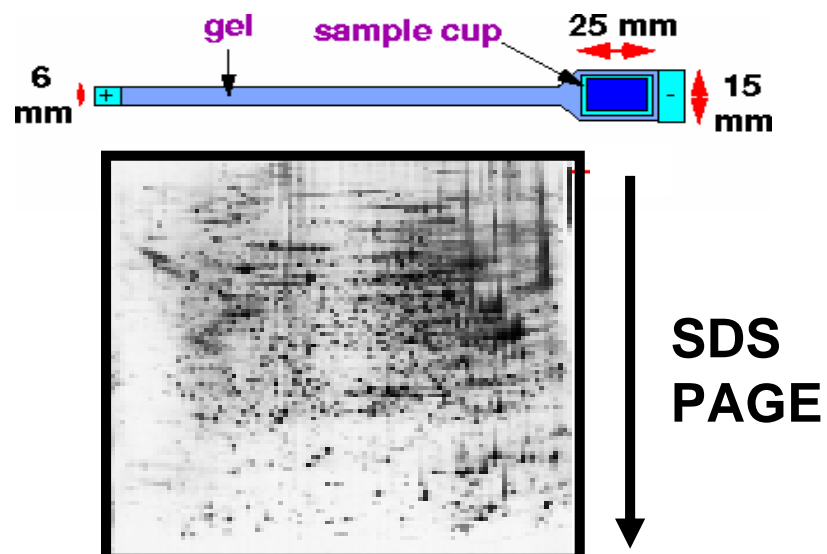
## Why 2D GE?

- **Oldest method for large scale protein separation (since 1975)**
- **Still most popular method for protein display and quantification**
- **Permits simultaneous detection, display, purification, identification, quantification**
- **Robust, increasingly reproducible, simple, cost effective, scalable & parallelizable**
- **Provides pI, MW, quantity**

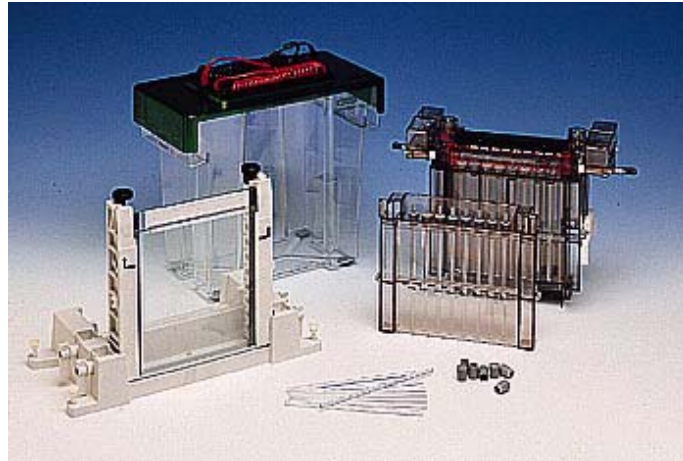
# Steps in 2D GE & Peptide ID

- Sample preparation
- **Isoelectric focusing (first dimension)**
- **SDS-PAGE (second dimension)**
- Visualization of proteins spots
- Identification of protein spots
- Annotation & spot evaluation

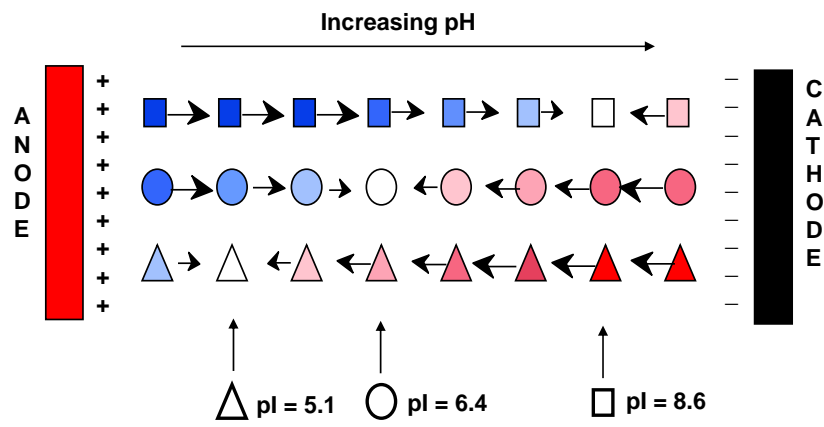
## 2D Gel Principles



# Isoelectric Focusing (IEF)



## IEF Principles



# Isoelectric Focusing

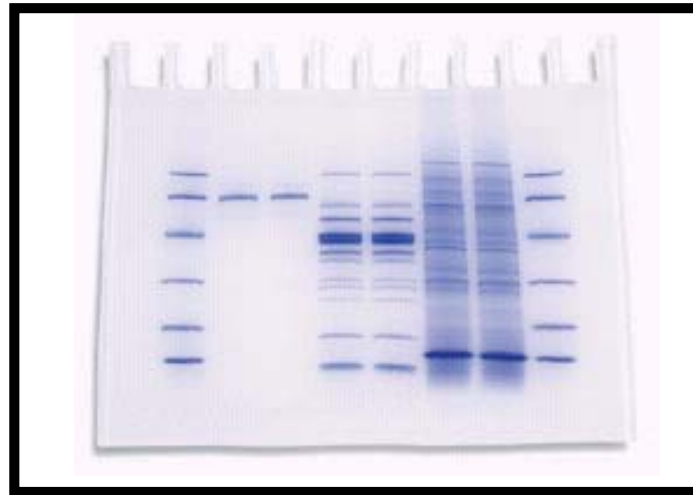
- Separation of basis of pI, not Mw
- Requires very high voltages (5000V)
- Requires a long period of time (10h)
- Presence of a pH gradient is critical
- Degree of resolution determined by slope of pH gradient and electric field strength
- Uses ampholytes to establish pH gradient
- Can be done in “slab” gels or in strips (IPG strips for 2D gel electrophoresis)

## Steps in 2D GE & Peptide ID

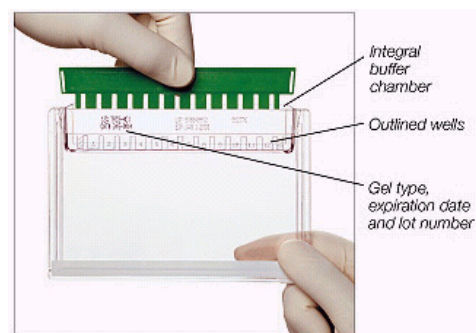
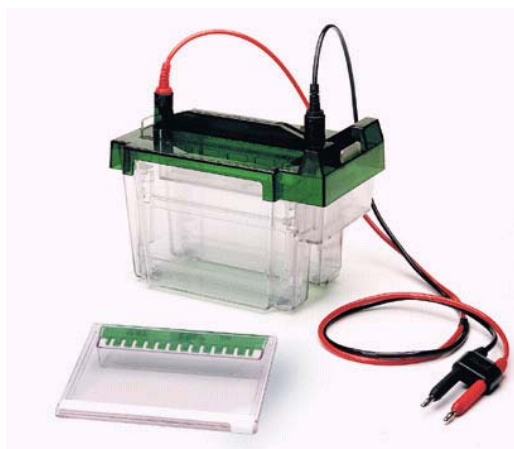
- Sample preparation
- Isoelectric focusing (first dimension)
- **SDS-PAGE (second dimension)**
- Visualization of proteins spots
- Identification of protein spots
- Annotation & spot evaluation



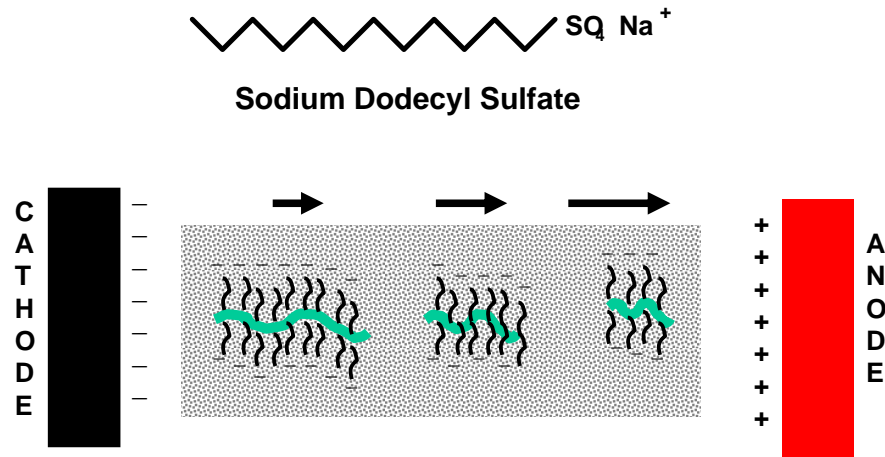
# SDS PAGE



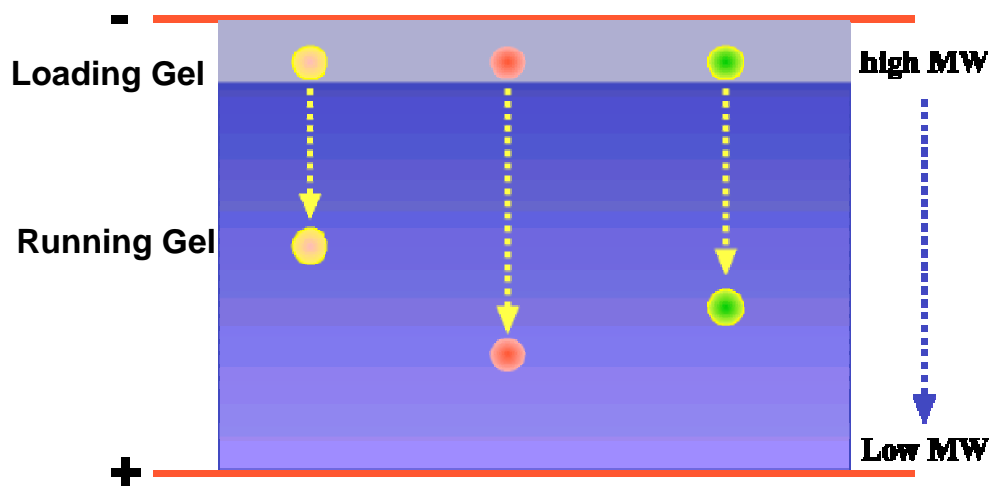
## SDS PAGE Tools



# SDS PAGE Principles



# SDS-PAGE Principles



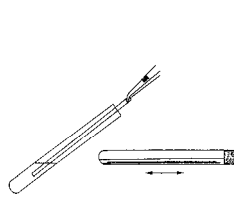
# SDS-PAGE

- Separation of basis of MW, not pI
- Requires modest voltages (200V)
- Requires a shorter period of time (2h)
- Presence of SDS is critical to disrupting structure and making mobility  $\sim 1/\text{MW}$
- Degree of resolution determined by %acrylamide & electric field strength

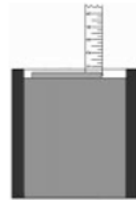
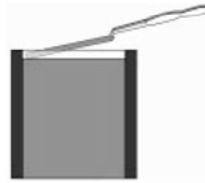
## SDS-PAGE for 2D GE

- After IEF, the IPG strip is soaked in an equilibration buffer (50 mM Tris, pH 8.8, 2% SDS, 6M Urea, 30% glycerol, DTT, tracking dye)
- IPG strip is then placed on top of pre-cast SDS-PAGE gel and electric current applied
- This is equivalent to pipetting samples into SDS-PAGE wells (an infinite #)

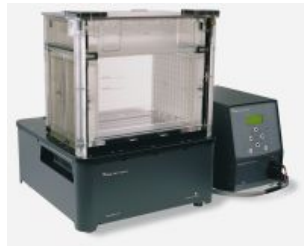
# SDS-PAGE for 2D GE



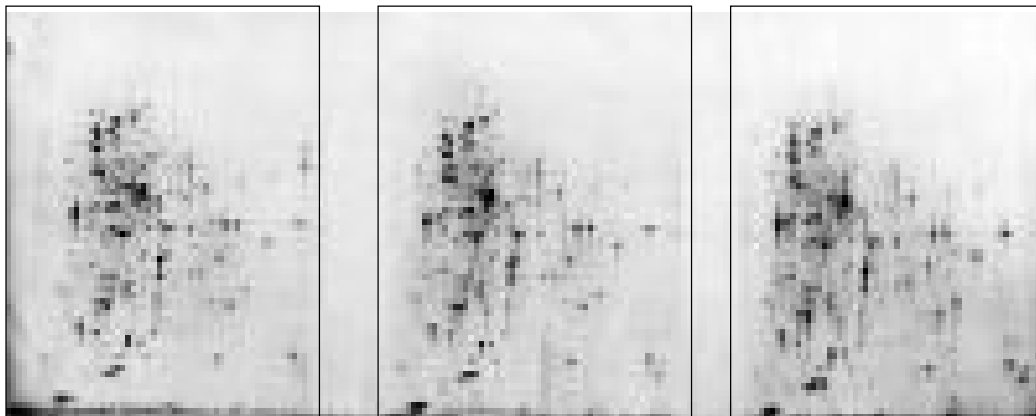
equilibration



SDS-PAGE



# 2D Gel Reproducibility

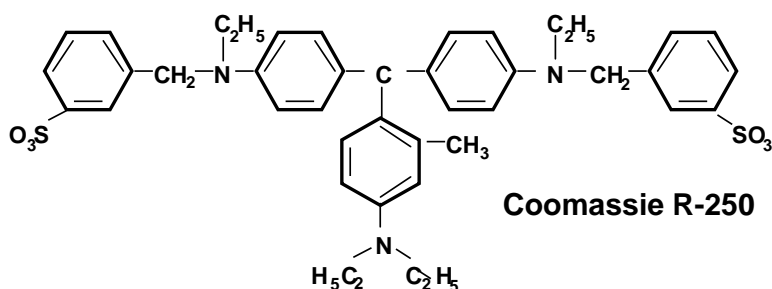


# Advantages and Disadvantages of 2D GE

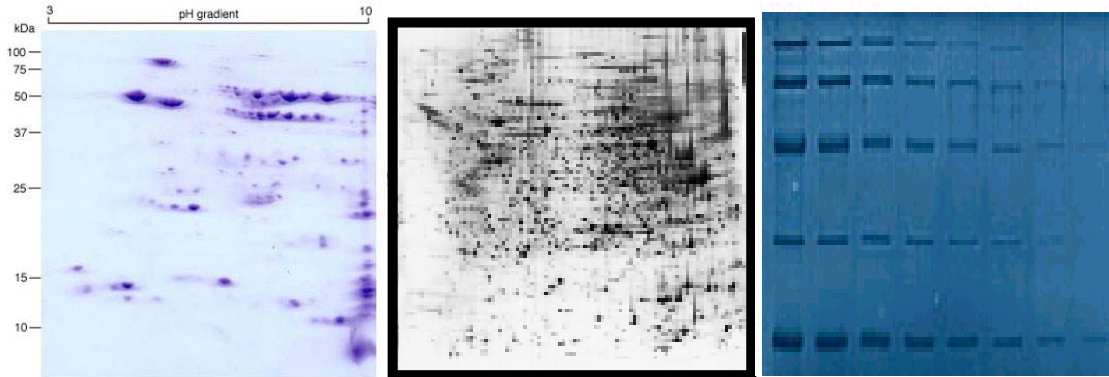
- Provides a hard-copy record of separation
- Allows facile quantitation
- Separation of up to 9000 different proteins
- Highly reproducible
- Gives info on Mw, pI and post-trans modifications
- Inexpensive
- Limited pI range (4-8)
- Proteins >150 kD not seen in 2D gels
- Difficult to see membrane proteins (>30% of all proteins)
- Only detects high abundance proteins (top 30% typically)
- Time consuming

## Protein Detection

- Coomassie Stain (100 ng to 10  $\mu$ g protein)
- Silver Stain (1 ng to 1  $\mu$ g protein)
- Fluorescent (Sypro Ruby) Stain (1 ng & up)



# Stain Examples



Coomassie

Silver Stain

Copper Stain

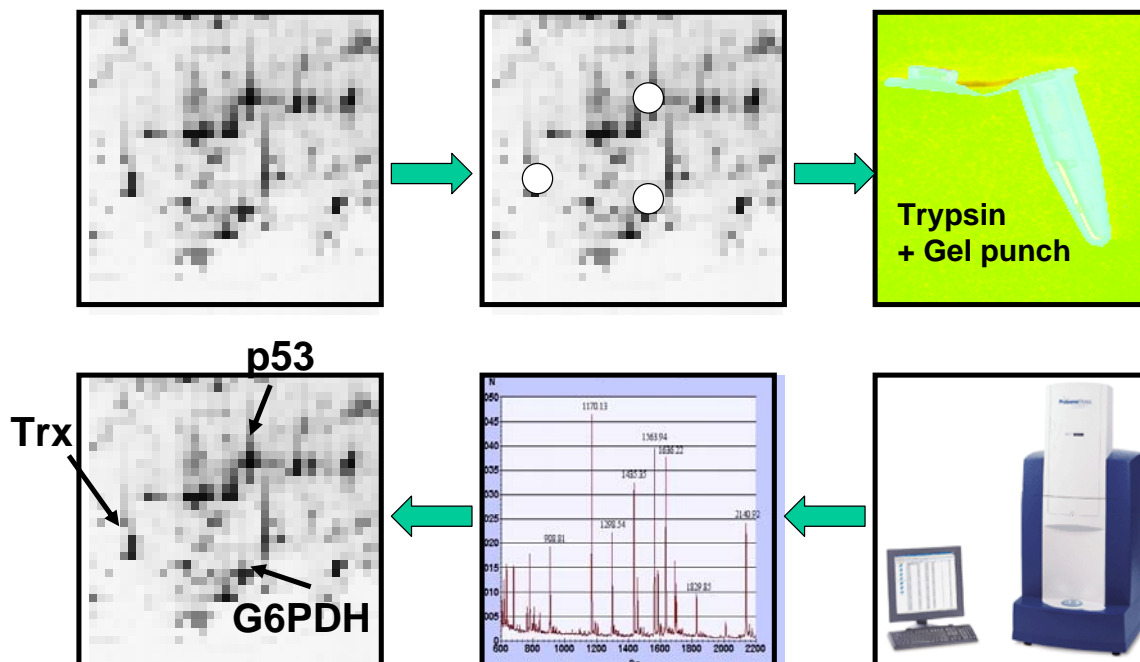
## Steps in 2D GE & Peptide ID

- Sample preparation
- Isoelectric focusing (first dimension)
- SDS-PAGE (second dimension)
- Visualization of proteins spots
- **Identification of protein spots**
- Annotation & spot evaluation

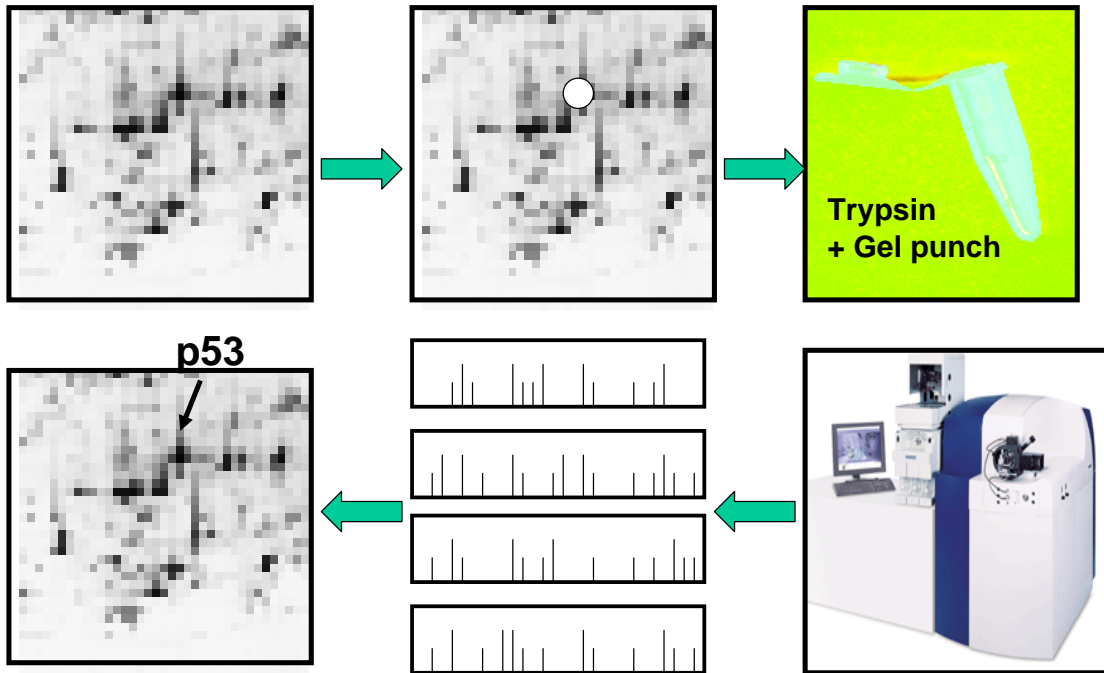
# Protein Identification

- **2D-GE + MALDI-MS**
  - Peptide Mass Fingerprinting (PMF)
- **2D-GE + MS-MS**
  - MS Peptide Sequencing/Fragment Ion Searching
- **Multidimensional LC + MS-MS**
  - ICAT Methods (isotope labelling)
  - MudPIT (Multidimensional Protein Ident. Tech.)
- **1D-GE + LC + MS-MS**
- **De Novo Peptide Sequencing**

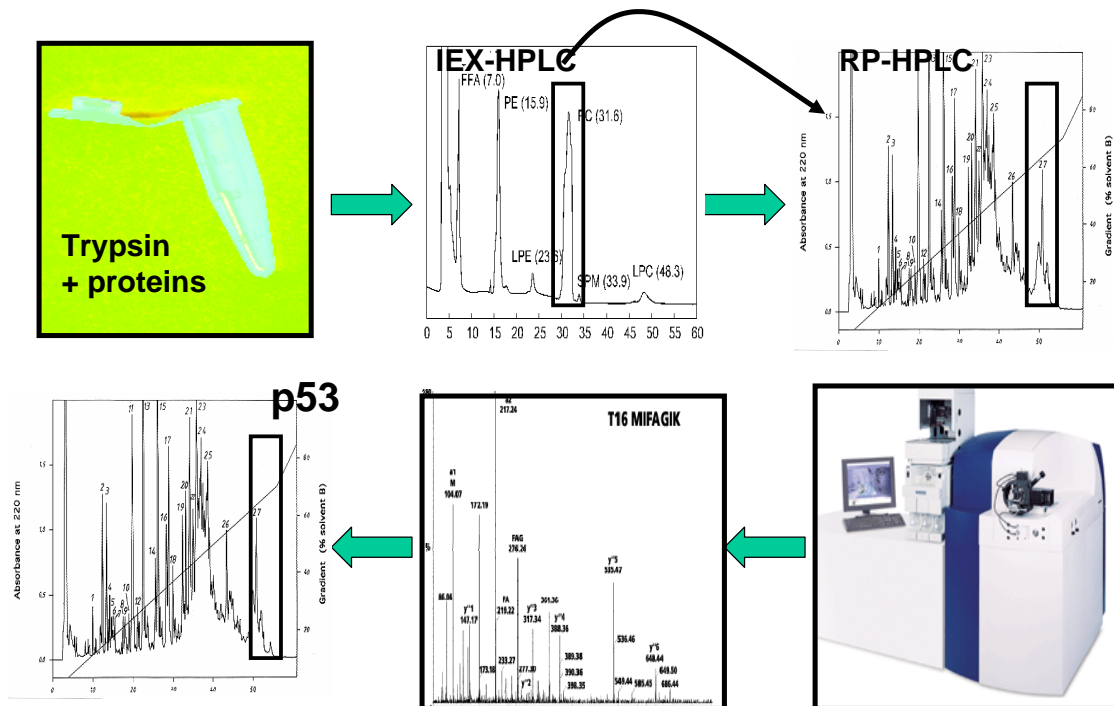
## 2D-GE + MALDI (PMF)



# 2D-GE + MS-MS

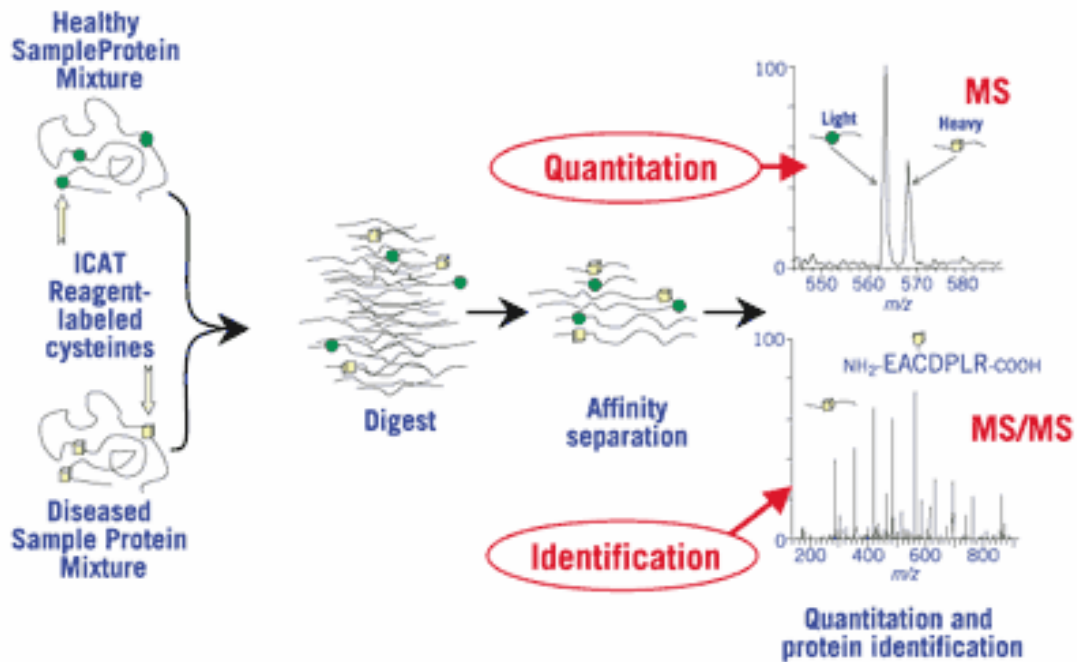


# MudPIT



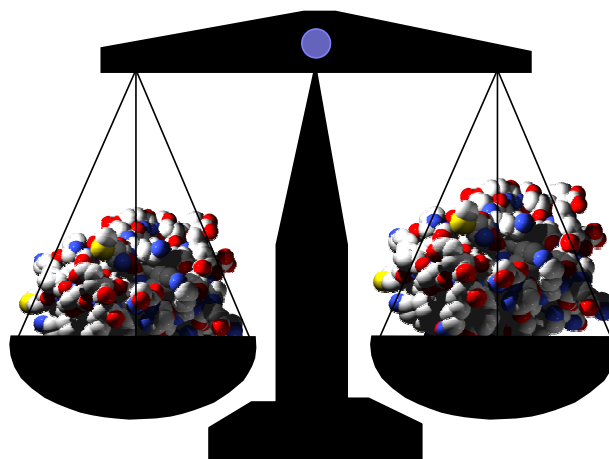


# ICAT (Isotope Coded Affinity Tag)



## Mass Spectrometry

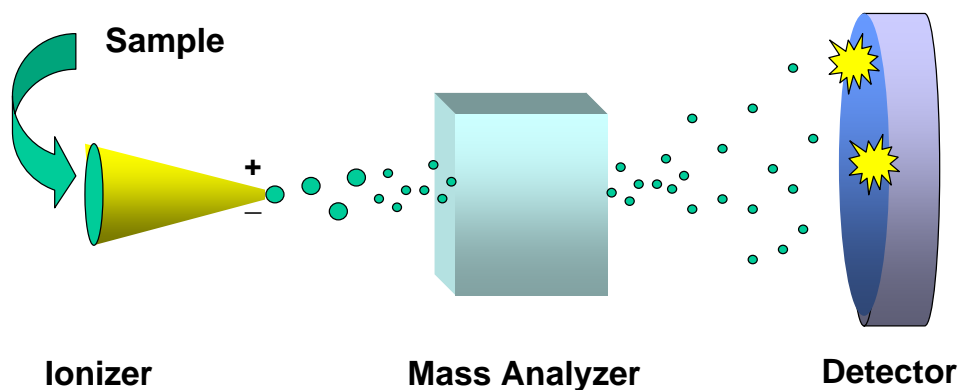
- Analytical method to measure the molecular or atomic weight of samples



# MS Principles

- Find a way to “charge” an atom or molecule (ionization)
- Place charged atom or molecule in a magnetic field or subject it to an electric field and measure its speed or radius of curvature relative to its mass-to-charge ratio (mass analyzer)
- Detect ions using microchannel plate or photomultiplier tube

## Mass Spec Principles

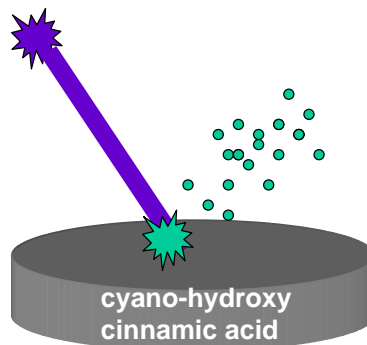


# Typical Mass Spectrometer



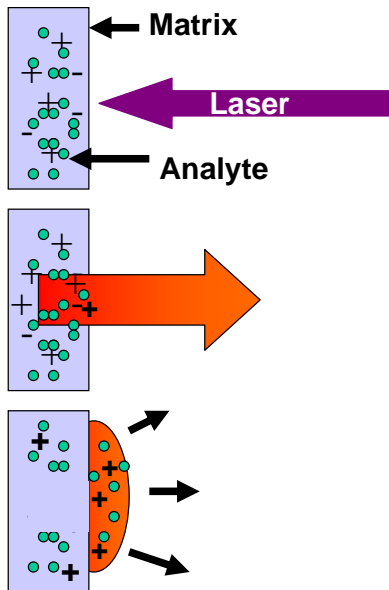
## Matrix-Assisted Laser Desorption Ionization

337 nm UV laser



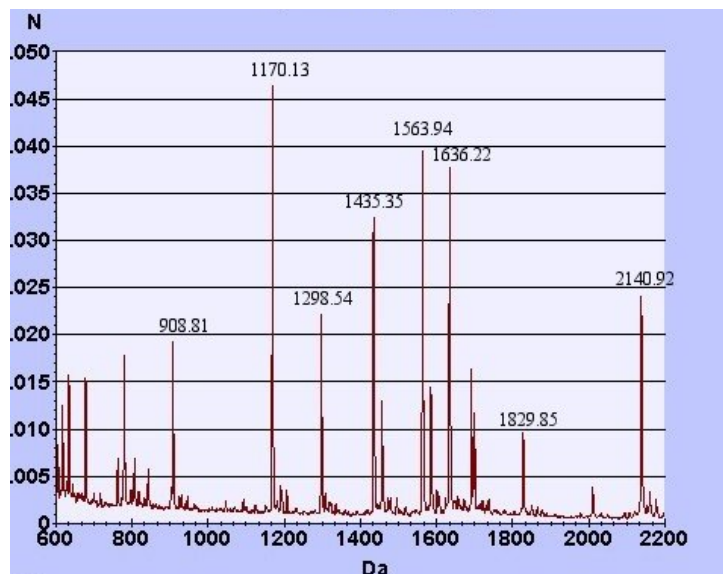
**MALDI**

# MALDI Ionization

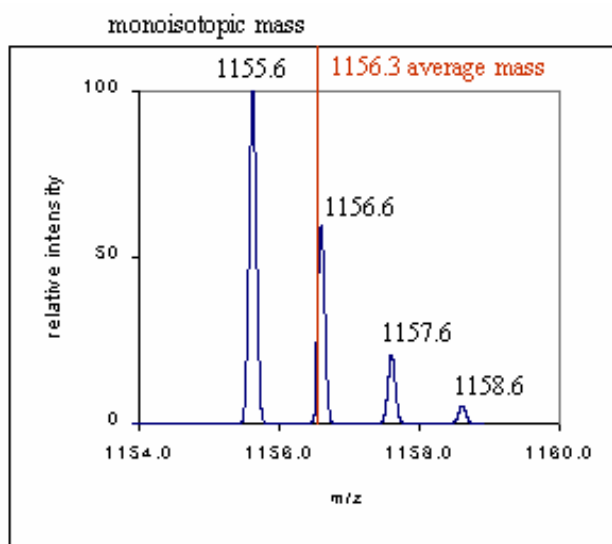


- Absorption of UV radiation by chromophoric matrix and ionization of matrix
- Dissociation of matrix, phase change to super-compressed gas, charge transfer to analyte molecule
- Expansion of matrix at supersonic velocity, analyte trapped in expanding matrix plume (explosion/"popping")

# MALDI Spectra (Mass Fingerprint)



# Masses in MS



- **Monoisotopic mass** is the mass determined using the masses of the most abundant isotopes
- **Average mass** is the abundance weighted mass of all isotopic components

## Amino Acid Residue Masses

### Monoisotopic Mass

Glycine	57.02147	Aspartic acid	115.02695
Alanine	71.03712	Glutamine	128.05858
Serine	87.03203	Lysine	128.09497
Proline	97.05277	Glutamic acid	129.04264
Valine	99.06842	Methionine	131.04049
Threonine	101.04768	Histidine	137.05891
Cysteine	103.00919	Phenylalanine	147.06842
Isoleucine	113.08407	Arginine	156.10112
Leucine	113.08407	Tyrosine	163.06333
Asparagine	114.04293	Tryptophan	186.07932

# Amino Acid Residue Masses

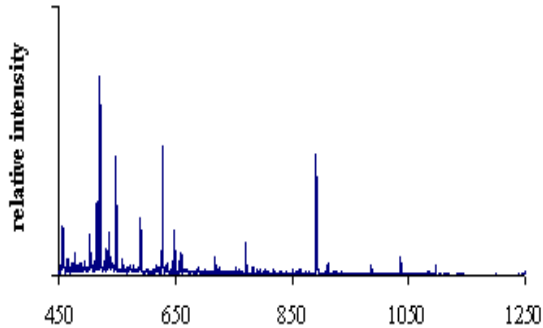
## Average Mass

Glycine	57.0520	Aspartic acid	115.0886
Alanine	71.0788	Glutamine	128.1308
Serine	87.0782	Lysine	128.1742
Proline	97.1167	Glutamic acid	129.1155
Valine	99.1326	Methionine	131.1986
Threonine	101.1051	Histidine	137.1412
Cysteine	103.1448	Phenylalanine	147.1766
Isoleucine	113.1595	Arginine	156.1876
Leucine	113.1595	Tyrosine	163.1760
Asparagine	114.1039	Tryptophan	186.2133

## Calculating Peptide Masses

- Sum the monoisotopic residue masses
- Add mass of H<sub>2</sub>O (18.01056)
- Add mass of H<sup>+</sup> (1.00785 to get M+H)
- If Met is oxidized add 15.99491
- If Cys has acrylamide adduct add 71.0371
- If Cys is iodoacetylated add 58.0071
- Other modifications are listed at
  - <http://prowl.rockefeller.edu/aainfo/deltamassv2.html>
- Only consider peptides with masses > 400

# Peptide Mass Fingerprinting (PMF)



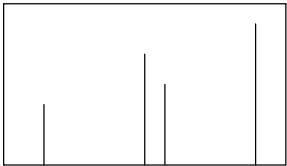
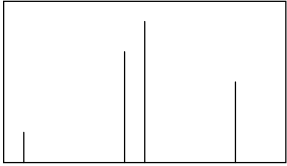
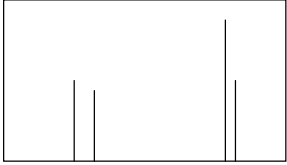
## Peptide Mass Fingerprinting

- Used to identify protein spots on gels or protein peaks from an HPLC run
- Depends on the fact that if a peptide is cut up or fragmented in a known way, the resulting fragments (and resulting masses) are unique enough to identify the protein
- Requires a database of known sequences
- Uses software to compare observed masses with masses calculated from database

# Principles of Fingerprinting

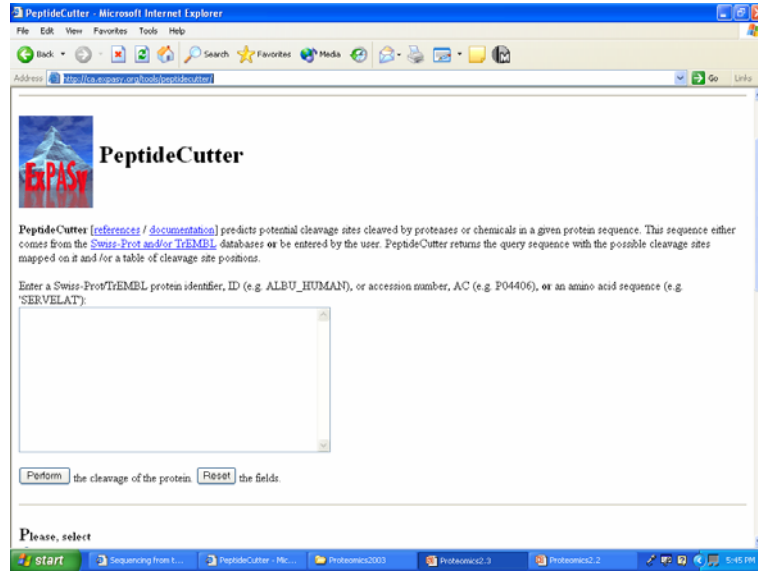
<u>Sequence</u>	<u>Mass (M+H)</u>	<u>Tryptic Fragments</u>
>Protein 1 acedfhsakdfqea sdfpkivtmeeewe ndadnfekqwfe	4842.05	acedfhsak dfgeasdfpk ivtmeeewendadnfek gwfe
>Protein 2 acekdfhsadfqea sdfpkivtmeeewe nkdadnfekqwfe	4842.05	acek dfhsadfgeasdfpk ivtmeeewenk dadnfekqwfe
>Protein 3 acedfhsadfqeka sdfpkivtmeeewe nda kdnf ekqwfe	4842.05	acedfhsadfgek asdfpk ivtmeeewendak dnf ekqwfe

# Principles of Fingerprinting

<u>Sequence</u>	<u>Mass (M+H)</u>	<u>Mass Spectrum</u>
>Protein 1 acedfhsakdfqea sdfpkivtmeeewe ndadnfekqwfe	4842.05	
>Protein 2 acekdfhsadfqea sdfpkivtmeeewe nkdadnfekqwfe	4842.05	
>Protein 3 acedfhsadfqeka sdfpkivtmeeewe nda kdnf ekqwfe	4842.05	



# Predicting Peptide Cleavages



<http://ca.expasy.org/tools/peptidecutter/>

[http://ca.expasy.org/tools/peptidecutter/peptidecutter\\_enzymes.html#Tryps](http://ca.expasy.org/tools/peptidecutter/peptidecutter_enzymes.html#Tryps)



## PeptideCutter

### The cleavage specificities of selected enzymes and chemicals:

A general model of enzymatic cleavage:

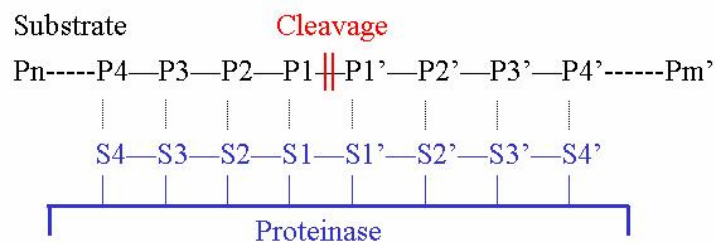


Fig.1 Schematic representation of enzyme-substrate complex with eight binding sites. Positions P<sub>n</sub> to P<sub>m</sub>' in the substrate are counted from the bond between P<sub>1</sub> and P<sub>1</sub>', where the cleavage occurs.

# Protease Cleavage Rules



Trypsin	XXX[KR]--[!P]XXX
Chymotrypsin	XX[FYW]--[!P]XXX
Lys C	XXXXXK-- XXXXX
Asp N endo	XXXXXD-- XXXXX
CNBr	XXXXXM--XXXXX

## Why Trypsin?

- Robust, stable enzyme
- Works over a range of pH values & Temp.
- Quite specific and consistent in cleavage
- Cuts frequently to produce “ideal” MW peptides
- Inexpensive, easily available/purified
- Does produce “autolysis” peaks (which can be used in MS calibrations)
  - 1045.56, 1106.03, 1126.03, 1940.94, 2211.10, 2225.12, 2283.18, 2299.18

# Preparing a Peptide Mass Fingerprint Database

- Take a protein sequence database (Swiss-Prot or nr-GenBank)
- Determine cleavage sites and identify resulting peptides for each protein entry
- Calculate the mass (M+H) for each peptide
- Sort the masses from lowest to highest
- Have a pointer for each calculated mass to each protein accession number in databank

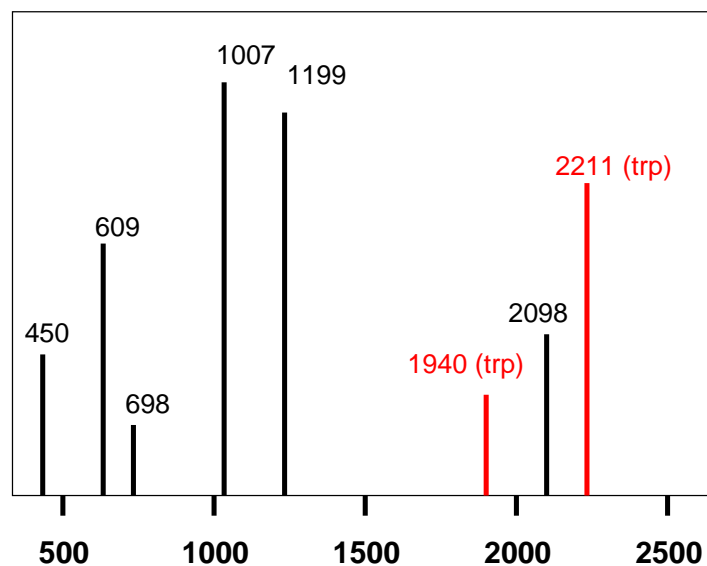
## Building A PMF Database

<u>Sequence DB</u>	<u>Calc. Tryptic Frags</u>	<u>Mass List</u>
>P12345	acedfhsak	450.2017 (P21234)
acedfhsakdfqea	dfgeasdfpk	609.2667 (P12345)
sdfpkivtmeeewe	ivtmeeewendadnfek	664.3300 (P89212)
ndadnfekqwfe	gwfe	1007.4251 (P12345)
		1114.4416 (P89212)
>P21234	acek	1183.5266 (P12345)
acekdfhsadfqea	dfhsadfgeasdfpk	1300.5116 (P21234)
sdfpkivtmeeewe	ivtmeeewenk	1407.6462 (P21234)
nkdadnfeqwfe	dadnfeqwfe	1526.6211 (P89212)
		1593.7101 (P89212)
>P89212	acedfhsadfgek	1740.7501 (P21234)
acedfhsadfqeka	asdfpk	2098.8909 (P12345)
sdfpkivtmeeewe	ivtmeeewendak	
ndakdnfeqwfe	dnfegwfe	

# The Fingerprint (PMF) Algorithm

- Take a mass spectrum of a trypsin-cleaved protein (from gel or HPLC peak)
- Identify as many masses as possible in spectrum (avoid autolysis peaks)
- Compare query masses with database masses and calculate # of matches or matching score (based on length and mass difference)
- Rank hits and return top scoring entry – this is the protein of interest

## Query (MALDI) Spectrum



# Query vs. Database

<u>Query Masses</u>	<u>Database Mass List</u>	<u>Results</u>
450.2201	450.2017 (P21234)	2 Unknown masses 1 hit on P21234 3 hits on P12345
609.3667	609.2667 (P12345)	
698.3100	664.3300 (P89212)	
1007.5391	1007.4251 (P12345)	Conclude the query protein is P12345
1199.4916	1114.4416 (P89212)	
2098.9909	1183.5266 (P12345)	
	1300.5116 (P21234)	
	1407.6462 (P21234)	
	1526.6211 (P89212)	
	1593.7101 (P89212)	
	1740.7501 (P21234)	
	2098.8909 (P12345)	

## What You Need To Do PMF

- A list of query masses (as many as possible)
- Protease(s) used or cleavage reagents
- Databases to search (SWProt, Organism)
- Estimated mass and pI of protein spot (opt)
- Cysteine (or other) modifications
- Minimum number of hits for significance
- Mass tolerance (100 ppm =  $1000.0 \pm 0.1$  Da)
- **A PMF website (Prowl, ProFound, Mascot, etc.)**

# PMF on the Web

- **ProFound**
  - [http://129.85.19.192/profound\\_bin/WebProFound.exe](http://129.85.19.192/profound_bin/WebProFound.exe)
- **MOWSE**
  - <http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse>
- **PeptideSearch**
  - <http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html>
- **Mascot**
  - [www.matrixscience.com](http://www.matrixscience.com)
- **Peptident**
  - <http://us.expasy.org/tools/peptident.html>

## ProFound

**ProFound** - Peptide Mapping [Short Form] Version 4.10.5  
The Rockefeller University Edition

<b>General</b>	<b>Digestion</b>
Sample ID <input type="text"/>	Allow maximum <input type="text" value="1"/> missed cleavages
Database <input type="text" value="NCBI/nr (2002/11/27)"/>	Enzyme <input type="text" value="Trypsin"/>
Taxonomic Category <input type="text" value="All taxa"/>	For user-defined cleavage, please click <a href="#">here</a> .
Search for <input type="text" value="single protein only"/>	<b>Modifications</b>
Protein Mass <input type="text" value="0"/> - <input type="text" value="3000"/> kDa	Complete Modification(s) <input type="text" value="Unmodified"/>
Protein pI <input type="text" value="0"/> - <input type="text" value="14"/>	4-vinyl-pyridine (Cys)
Report Top <input type="text" value="10"/> Candidates	Acrylamide (Cys)
Questions? Please write to <a href="#">ProFound</a>	Iodoacetamide (Cys)
What's new <a href="#">about ProFound?</a>	Iodoacetic acid (Cys)
	Partial Modification <input type="checkbox"/> Methionine oxidation
	For more partial modifications, please click <a href="#">here</a> .
<b>Masses</b>	
Average Masses:	Monoisotopic Masses:
<input type="text"/>	<input type="text"/>
Mass tolerance for average data: +/- <input type="text" value="1"/>	Mass tolerance for monoisotopic data: +/- <input type="text" value="0.1"/>
Tolerance unit: <input checked="" type="radio"/> Da <input type="radio"/> % <input type="radio"/> ppm	Charge state: <input checked="" type="radio"/> M <input type="radio"/> MH+
<input type="button" value="Identify Protein"/>	<input type="button" value="Extra Settings"/>
<input type="button" value="Example"/>	<input type="button" value="Reset Form"/>

# ProFound (PMF)

*PROWL (ProFound)*

The screenshot shows the ProFound web interface with several red annotations:

- Sample ID:** "You can give your sample a designation (eg. spot number) to keep track of mass lists."
- Database:** "Select Database: eg. NCBI, SWISS-PROT"
- Search for:** "Select taxonomy: You should get higher search scores with narrower search parameters. Start with the closest related taxa, and broaden the search as needed."
- Protein Mass:** "With broad pH range IPG strips (eg. 3-10), co-migration of two or more proteins is possible. You can search for multiple proteins here, or resubmit those unmatched peptides after this search."
- Report top:** "Enter protein mass range and pI. \*use gel information if available\*" and "Enter how many 'hits' you would like the search engine to report."
- Enzyme:** "We often see 1 missed cleavage with in-gel trypsin digestion." and "Enter the enzyme that was used to digest the protein."
- Modifications:** "Enter modifications that occur on every instance of the residue in the protein. eg. iodoacetamide" and "We occasionally see methionine oxidation with in-gel trypsinization."
- Masses:** "We give you a monoisotopic mass list." and "For MALDI data, the peptides in the mass list are protonated (+1)."
- Charge state:** "With internal calibration, the peptide mass accuracy is within 0.1 Da (50 ppm)."

## What Are Missed Cleavages?

### Sequence

```
>Protein 1
acedfhsakdfqea
sdfpkivtmeeewe
ndadnfekqgwe
```

### Tryptic Fragments (no missed cleavage)

```
acedfhsak (1007.4251)
dfgeasdfpk (1183.5266)
ivtmeeewendadnfek (2098.8909)
gwfe (609.2667)
```

### Tryptic Fragments (1 missed cleavage)

```
acedfhsak (1007.4251)
dfgeasdfpk (1183.5266)
ivtmeeewendadnfek 2098.8909)
gwfe (609.2667)
acedfhsakdfgeasdfpk (2171.9338)
ivtmeeewendadnfekgwfe (2689.1398)
dfgeasdfpkivtmeeewendadnfek (3263.2997)
```

# ProFound Results

## ProFound - Search Result Summary

Version 4.10.5  
The Rockefeller University Edition

Protein Candidates for search B9403AFB-07C0-76BF87E5 [1209637 sequences searched]							
Rank	Probability	Est'd Z	Protein Information and Sequence Analyse Tools (T)	%	pI	kDa	Ⓢ
1	2.2e-001	0.12	T <a href="#">gi 15222204 ref NP_172776.1</a> putative oxysterol-binding protein, protein id: At1g13170.1 [Arabidopsis thaliana]	8	6.1	92.31	Ⓢ
2	2.2e-001	0.12	T <a href="#">gi 17547403 ref NP_520805.1</a> PROBABLE OXIDOREDUCTASE PYRROLINE-5-CARBOXYLATE REDUCTASE SIGNAL PEPTIDE PROTEIN [Ralstonia solanacearum]	11	5.8	28.10	Ⓢ
3	7.6e-002	-	T <a href="#">gi 23054472 gb ZF_00080629.1</a> hypothetical protein [Geobacter metallireducens]	11	6.1	51.76	Ⓢ
4	7.6e-002	-	T <a href="#">gi 19920902 ref NP_609168.1</a> CG7228-PA [Drosophila melanogaster]	7	8.6	66.18	Ⓢ
5	2.6e-002	-	T <a href="#">gi 19572314 emb CAD19081.1</a> potassium channel beta chain [Stigmatella aurantiaca]	10	9.6	41.10	Ⓢ
+6	2.5e-002	-	T <a href="#">gi 2133779 pir S63985</a> collagen alpha 2 chain precursor - sea urchin (Strongylocentrotus purpuratus) (fragment)	3	4.4	200.03	Ⓢ
7	2.3e-002	-	T <a href="#">gi 15450423 gb AAK96505.1</a> AT4g20760/F21C20_110 [Arabidopsis thaliana]	13	9.8	32.46	Ⓢ
+8	2.0e-002	-	T <a href="#">gi 7495844 pir T25534</a> hypothetical protein C10H11.6 - Caenorhabditis elegans	8	6.7	58.38	Ⓢ
9	1.9e-002	-	T <a href="#">gi 21293583 gb EAA05728.1</a> agCP10259 [Anopheles gambiae str. PEST]	4	6.3	66.10	Ⓢ
10	1.6e-002	-	T <a href="#">gi 16121031 ref NP_404344.1</a> sigma-54 transcriptional regulatory protein [Yersinia pestis]	10	6.1	37.74	Ⓢ

# MOWSE

MOWSE at the HGMP-RC - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse>

**Bioinformatics Applications:**  
**MOWSE at the HGMP-RC**

You can use this page to submit a [MOWSE](#) database search.

MOWSE will search the [owl protein sequence database](#) with protein fragment information, and return the protein(s) which most likely correspond to your peptide-data.

### MOWSE Parameters

You must supply various parameters in order to run MOWSE successfully. We have provided default values for some of these parameters; information to help you select appropriate alternatives can be found by following the links associated with the various parameter fields.

Sample name:

Select the reagent used:

Tolerance:

Whole sequence molecular weight:

Molecular weight filter:

Pfactor:

### Peptide mass data

start Sequencing from the ... MOWSE at the HGMP-... My Documents Proteomics2.3 10:39 PM



# PeptIdent

PeptIdent - Microsoft Internet Explorer

Address: <http://us.expasy.org/tools/peptident.html>

## Peptide Mass Fingerprinting

Name of the unknown protein:  pI:  within pI range:

Database:  Mw:  (in Dalton, not kDa) within Mw range (in percent):

Note: Peptides with masses >6000 Da have not been indexed.

Species to be searched:

Enter a list of peptide masses (separated by spaces or newlines) that correspond to the unknown protein:

Or upload a file in one of the supported formats from your computer. The peptide masses will be extracted automatically from this file.

All peptide masses are

[M+H]<sup>+</sup> or  [M] or  [M-H]<sup>-</sup>, and

monoisotopic or  average.

The peptide masses are

with cysteines treated with:

with acrylamide adducts on cysteines

with methionines oxidized.

Enzyme:

Allow for  missed cleavage sites (MC).

Report only proteins with at least  peptide hits.

Display a maximum of  matching proteins.

Mass tolerance:  Dalton

Print information about sequence portion covered by the matching peptides.

Send the result by e-mail

With this option, you will receive the result (in form of an html table) by e-mail. This is recommended and helps avoid the otherwise frequent Document

start Sequencing from the ... PeptIdent - Microsoft ... My Documents Proteomics2.3 10:42 PM

# MASCOT

## Mascot: Peptide Mass Fingerprint

Your name  Email

Search title

Database

Taxonomy

Enzyme  Allow up to  missed cleavages

Fixed modifications   
AB\_oldest\_ICATd8 (C)  
Acetyl (K)  
Acetyl (N-term)  
Amide (C-term)

Variable modifications   
AB\_oldest\_ICATd8 (C)  
Acetyl (K)  
Acetyl (N-term)  
Amide (C-term)

Protein mass  kDa Peptide tol.  Da

Mass values  MH<sup>+</sup>  M<sub>r</sub> Monoisotopic  Average

Data file

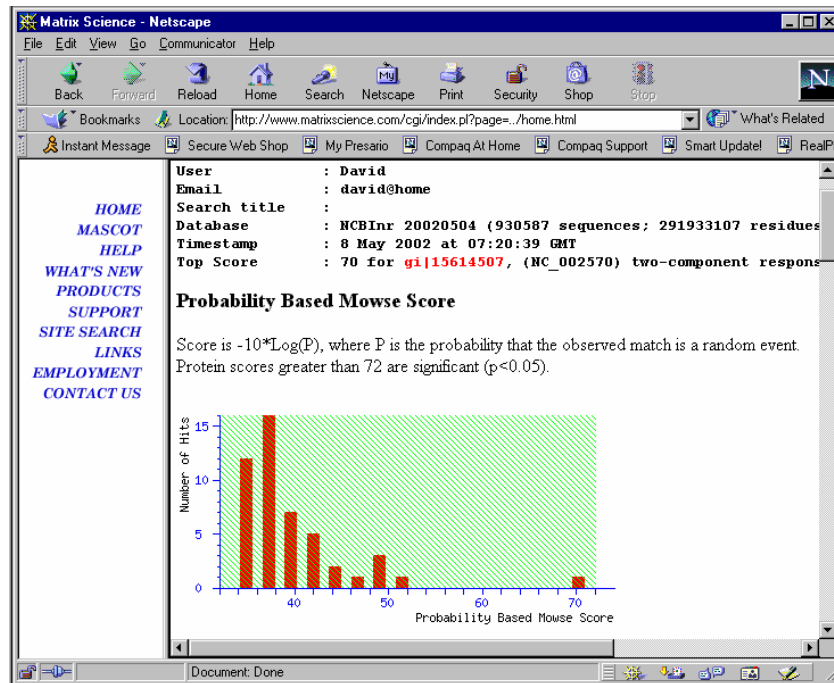
Query

NB Contents of this field are ignored if a data file is specified.

Overview  Report top  hits

Copyright © 2000 Matrix Science Ltd. All Rights Reserved. Last Updated 06/29/2003 01:29:18

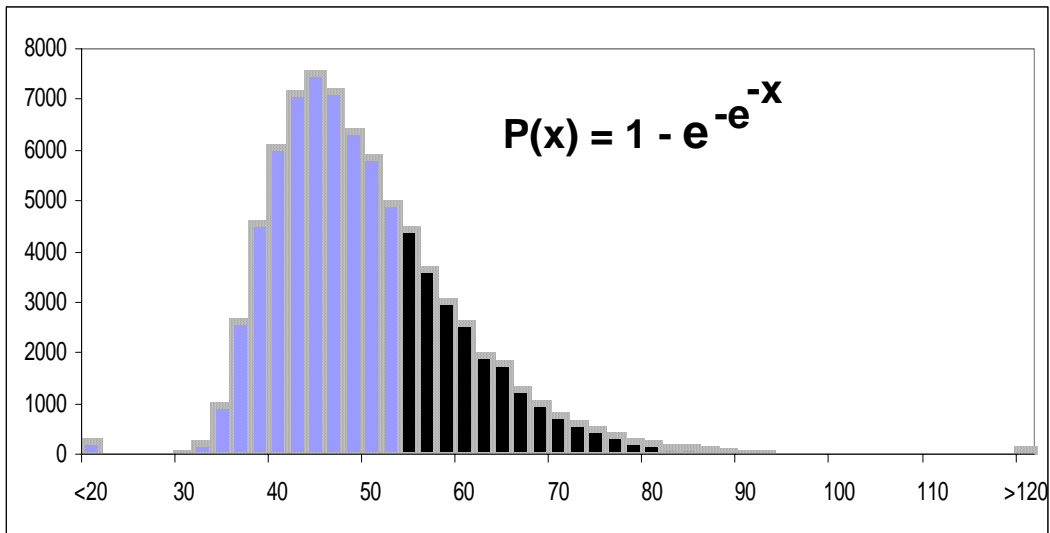
# MASCOT



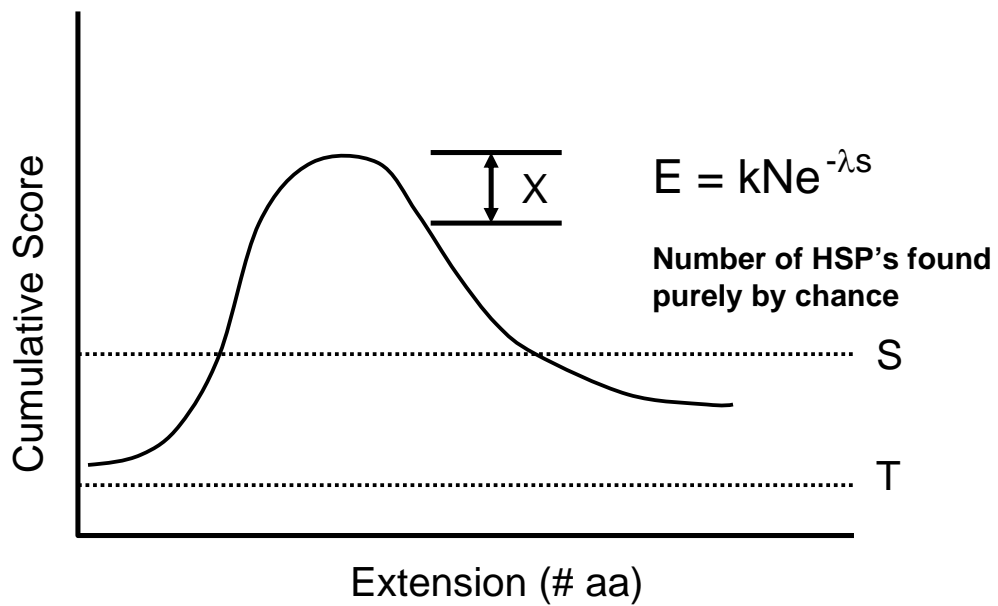
## Mascot Scoring

- The statistics of peptide fragment matching in MS (or PMF) is very similar to the statistics used in BLAST
- The scoring probability follows an extreme value distribution
- High scoring segment pairs (in BLAST) are analogous to high scoring mass matches in Mascot
- Mascot scoring is much more robust than arbitrary match cutoffs (like % ID)

# Extreme Value Distribution



## Extending HSP's



## Mascot/Mowse Scoring

- The Mascot Score is given as  $S = -10 \cdot \log(P)$ , where  $P$  is the probability that the observed match is a random event
- Try to aim for probabilities where  $P < 0.05$  (less than a 5% chance the peptide mass match is random)
- Mascot scores greater than 72 are significant ( $p < 0.05$ ).

## Advantages of PMF

- Uses a “robust” & inexpensive form of MS (MALDI)
- Doesn't require too much sample optimization
- Can be done by a moderately skilled operator (don't need to be an MS expert)
- Widely supported by web servers
- Improves as DB's get larger & instrumentation gets better
- *Very amenable to high throughput robotics (up to 500 samples a day)*

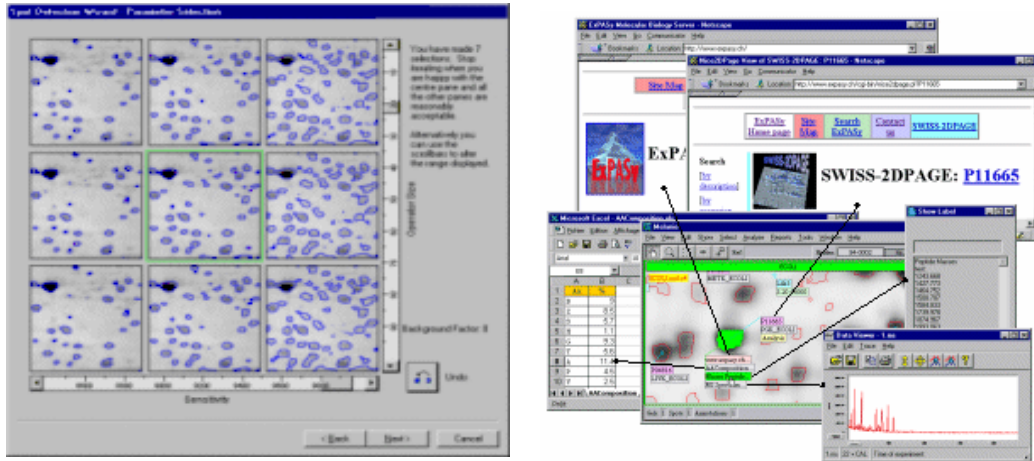
## Limitations With PMF

- Requires that the protein of interest already be in a sequence database
- Spurious or missing critical mass peaks always lead to problems
- Mass resolution/accuracy is critical, best to have <20 ppm mass resolution
- Generally found to only be about 40% effective in positively identifying gel spots

## Steps in 2D GE & Peptide ID

- Sample preparation
- Isoelectric focusing (first dimension)
- SDS-PAGE (second dimension)
- Visualization of proteins spots
- Identification of protein spots
- **Annotation & spot evaluation**

# 2D Gel Software



## Commercial Software

- **Melanie 3 (GeneBio - Windows only)**
  - <http://ca.expasy.org/melanie>
- **ImageMaster 2D Elite (Amersham)**
  - <http://www.imsupport.com/>
- **Phoretix 2D Advanced**
  - <http://www.phoretix.com/>
- **PDQuest 6.1 (BioRad - Windows only)**
  - <http://www.proteomeworks.bio-rad.com/html/pdquest.html>

# Common Software Features

- Image contrast and coloring
- Gel annotation (spot selection & marking)
- Automated peak picking
- Spot area determination (Integration)
- Matching/Morphing/Landmarking 2 gels
- Stacking/Aligning/Comparing gels
- Annotation copying between 2 gels

## GelScape – Gel Annotation on the Web

- Web-enabled gel viewing and annotation tool
- Allows users to post, share and compare gels in a free, platform independent manner
- A Java Applet with extensive Perl and HTML
- Tested and operable on most platforms (UNIX, Linux, Windows, MacOS) using most browsers (IE and Netscape > 4.0)
- Conceptually aligned with web mail
- Developed by Nelson Young & Casper Chang

## **GelScape Supports...**

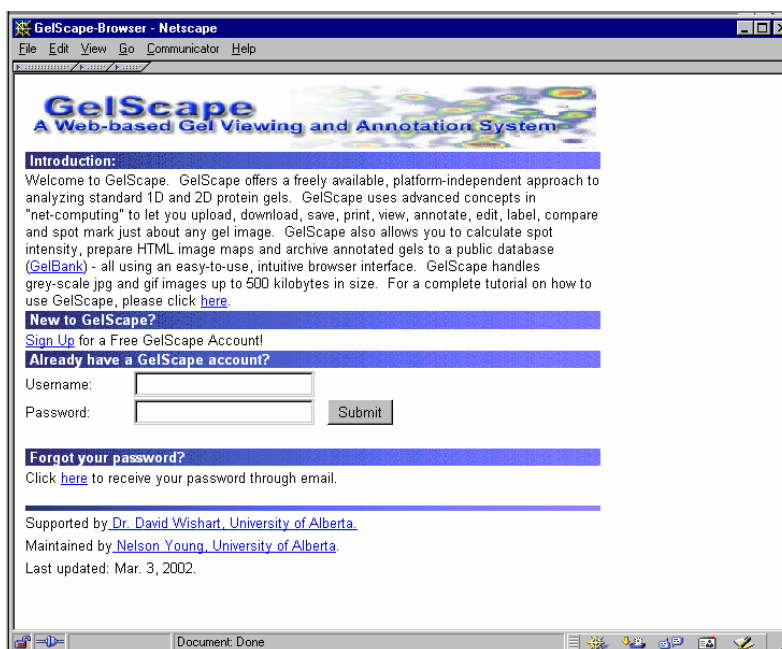
- **1D and 2D gel image uploading (gif and jpg) from local machine**
- **Non-local (server-side) storage of annotated gels**
- **Image resizing (zooming?)**
- **Spot marking and unmarking**
- **Spot annotation (via Swiss Prot ID, mass fingerprint, hand annotation)**

## **GelScape Supports...**

- **MW and pH grid drawing and dragging**
- **Spot edge detection and spot integration**
- **Interactive, image map spot annotation display**
- **Gel comparison (overlying)**
- **Gel legend display**
- **Image saving, image uploading (to GelBank), image printing (preview)**



<http://www.gelscape.org>



## Expressional Proteomics

- **Sample preparation**
- **2D electrophoresis or 2D HPLC separation**
- **Visualization of proteins spots/peaks**
- **Identification of protein spots/peaks**
- **Annotation & spot evaluation**

# 3 Kinds of Proteomics

- **Structural Proteomics**
  - High throughput X-ray Crystallography/Modelling
  - High throughput NMR Spectroscopy/Modelling
- **Expressional or Analytical Proteomics**
  - Electrophoresis, Protein Chips, DNA Chips, 2D-HPLC
  - Mass Spectrometry, Microsequencing
- **Functional or Interaction Proteomics**
  - HT Functional Assays, Protein Chips, Ligand Chips
  - Yeast 2-hybrid, Deletion Analysis, Motif Analysis