Proteomics & Bioinformatics Part I

David Wishart

3-41 Athabasca Hall

david.wishart@ualberta.ca

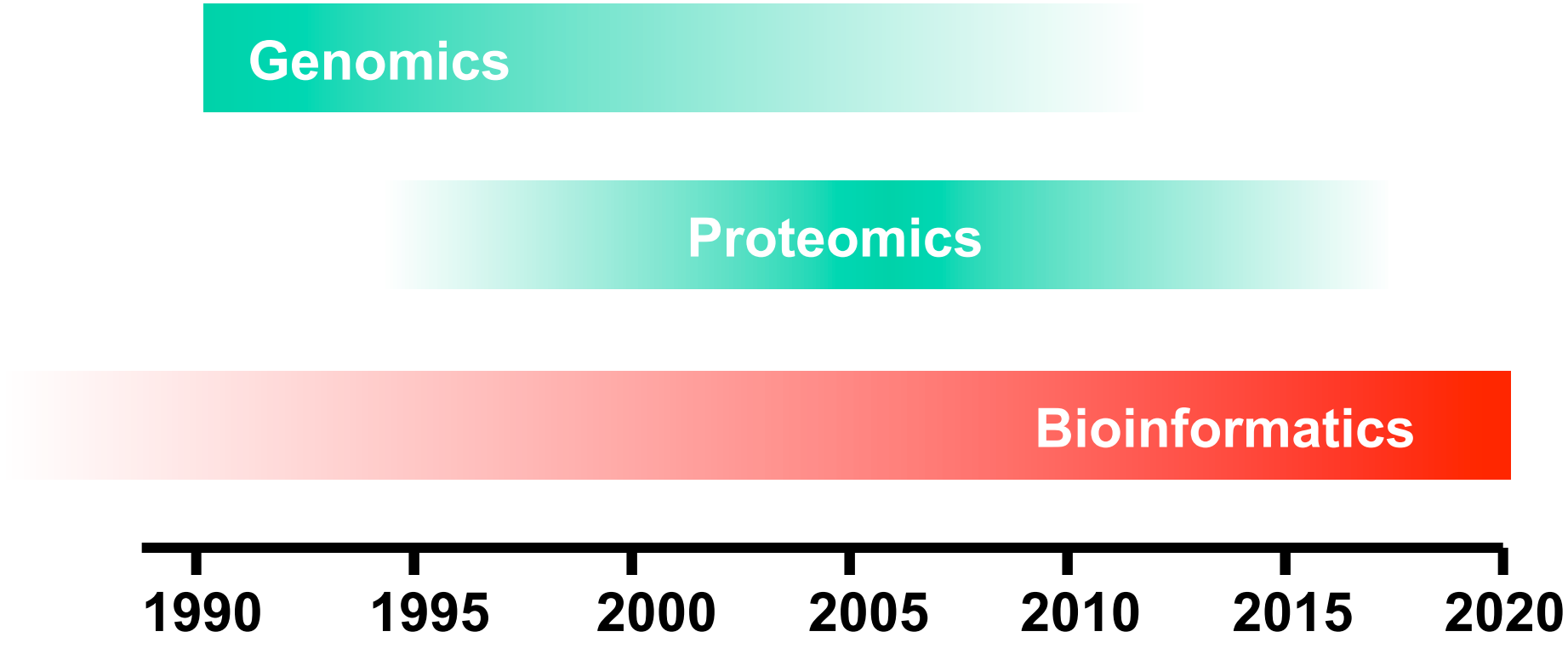
Objectives

- Learn about the 3 different types of proteomics
- Become familiar with expression-based proteomics techniques
- Become familiar with mass spectrometry for protein or peptide ID
- Become familiar with some of the software tools and algorithms for peptide/protein ID

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).***

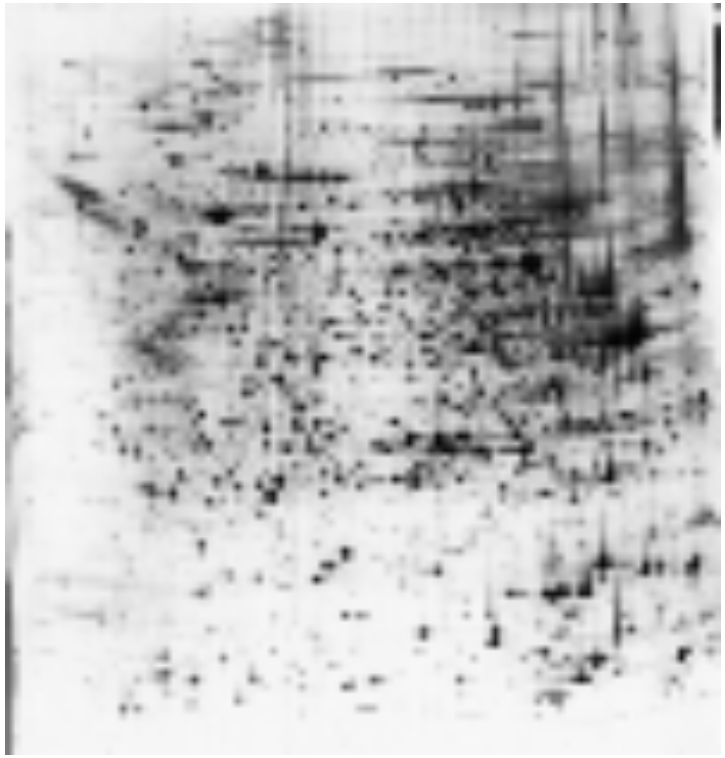
Proteomics & Bioinformatics



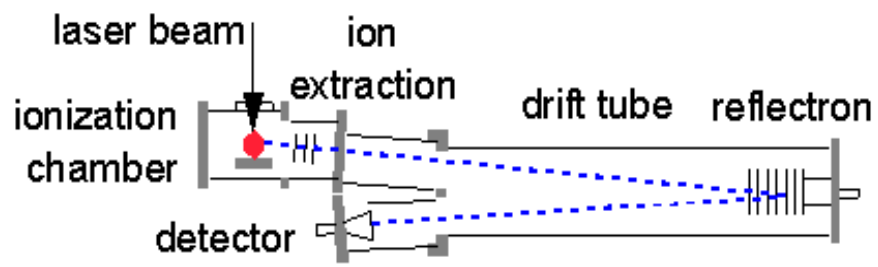
3 Kinds of Proteomics*

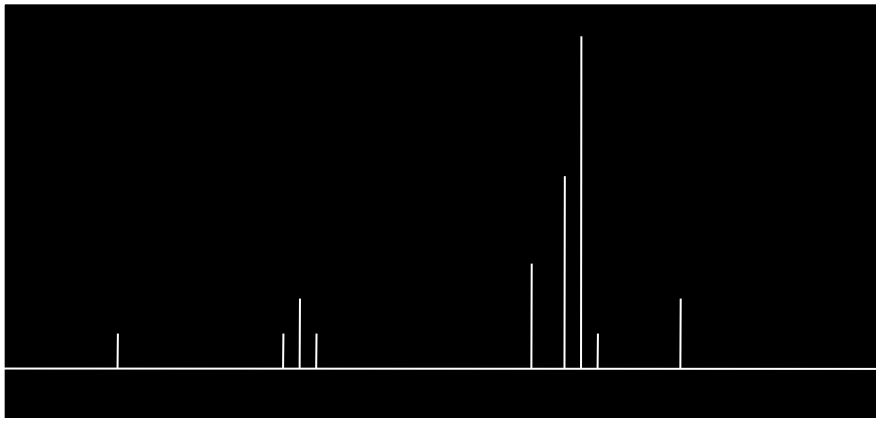
- **Structural Proteomics**
  - High throughput X-ray Crystallography/Modelling
  - High throughput NMR Spectroscopy/Modelling
- **Expressional or Analytical Proteomics**
  - Electrophoresis, Protein Chips, DNA Chips, 2D-HPLC
  - Mass Spectrometry, Microsequencing
- **Functional or Interaction Proteomics**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

Expressional Proteomics



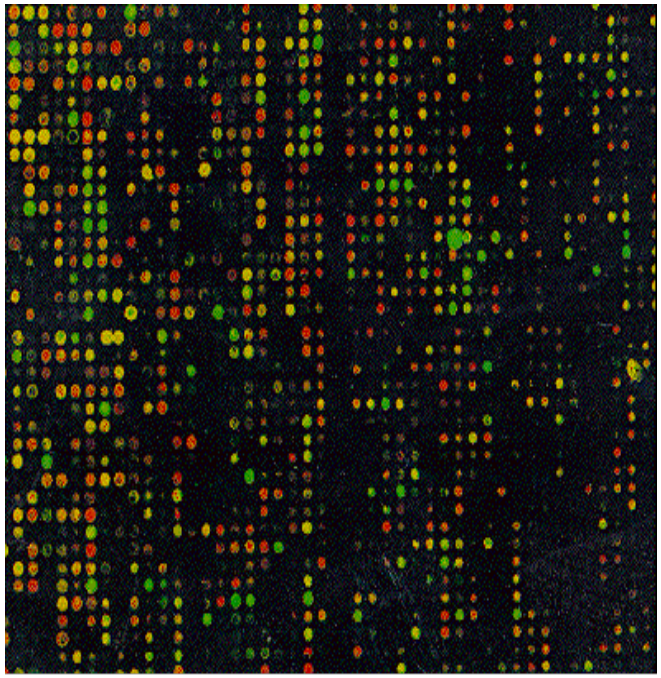
2-D Gel

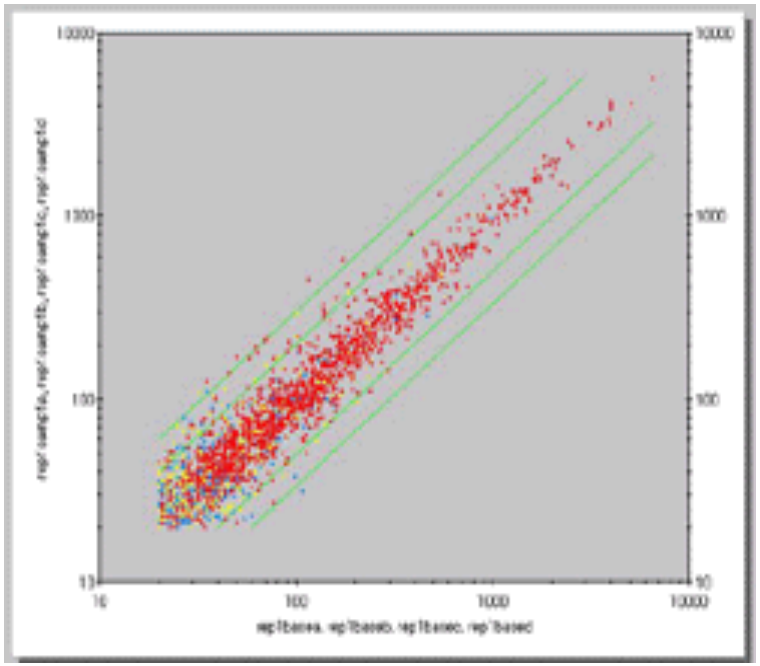




QTOF Mass Spectrometry

Expressional Proteomics

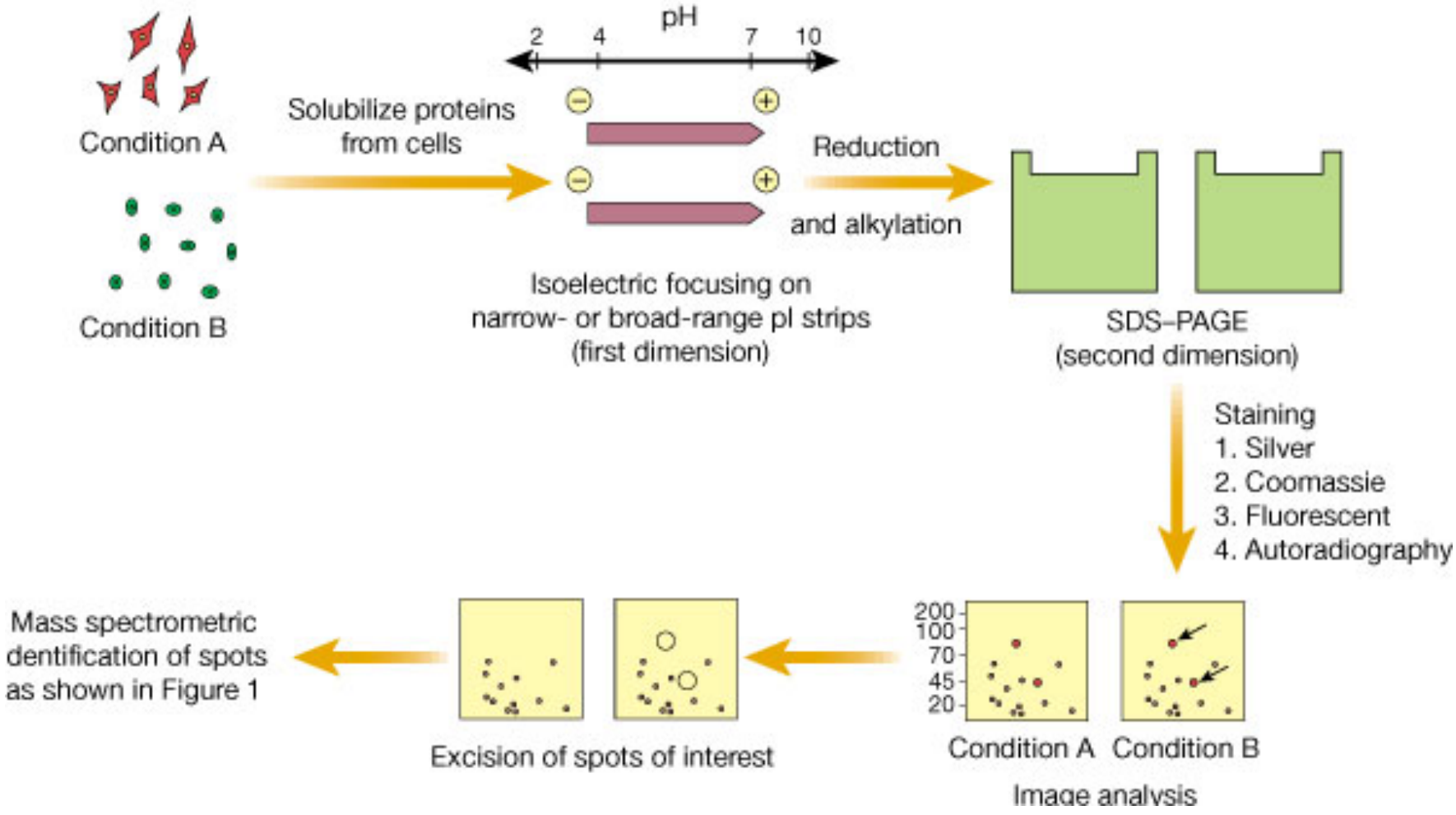




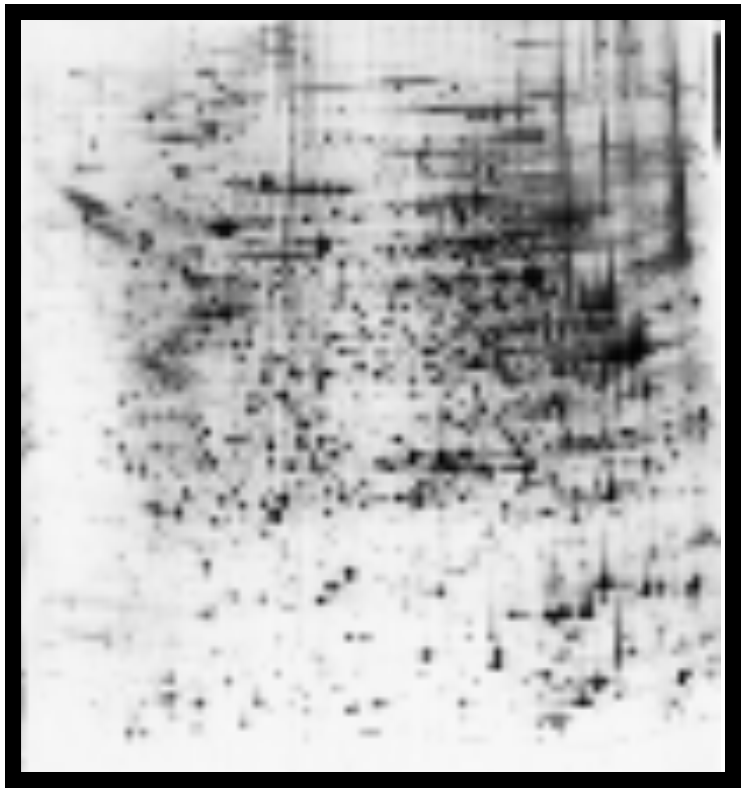
Expressional Proteomics*

- 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**

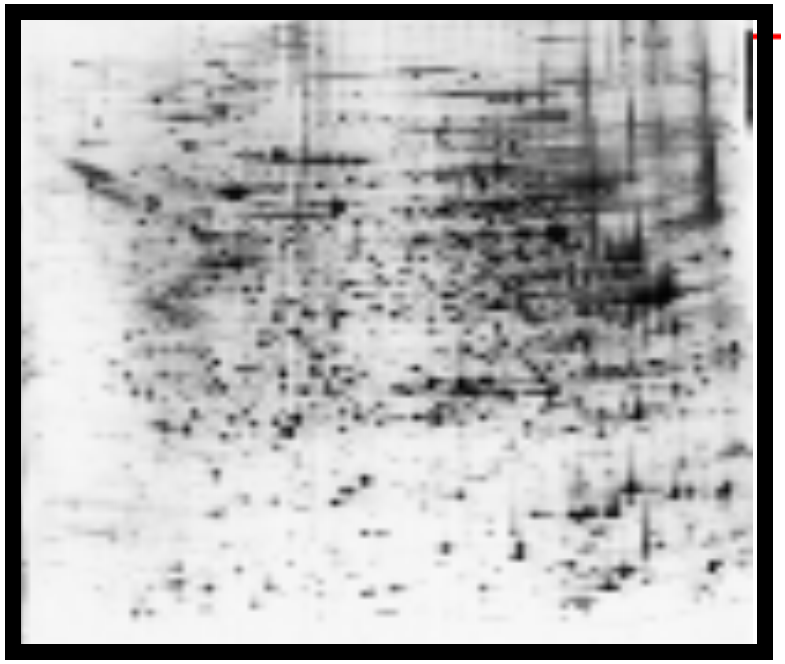
Why 2D GE?*

- 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

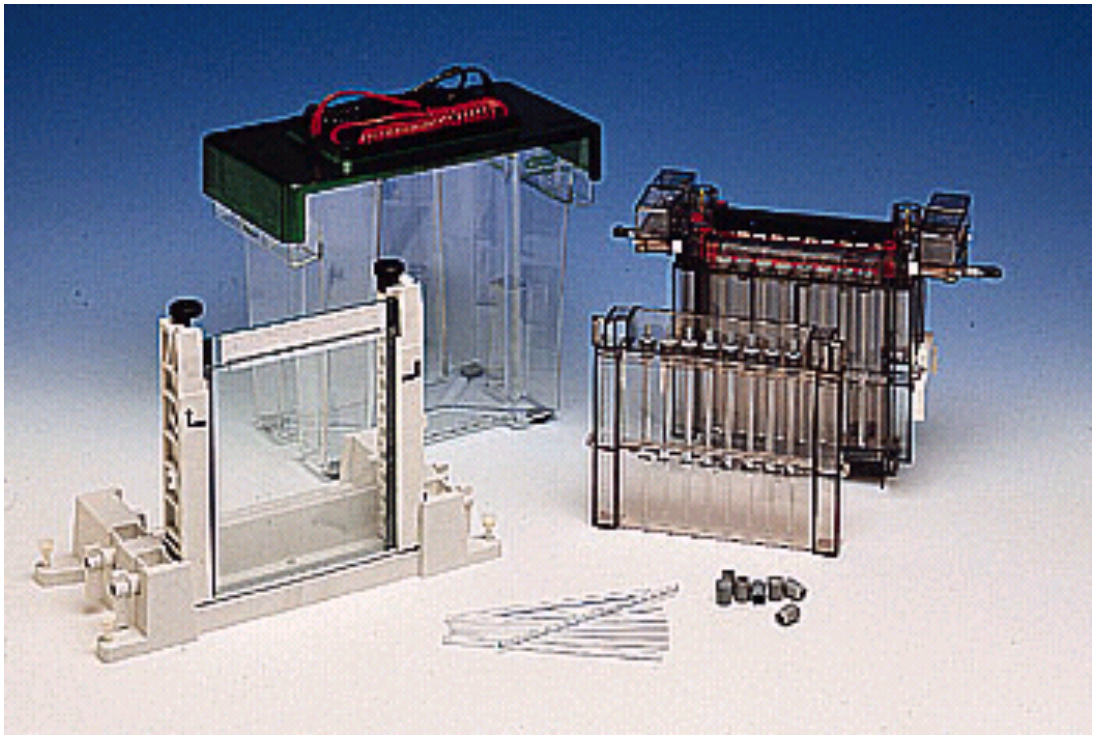
- Sample preparation
- Isoelectric focusing (first dimension)
- SDS-PAGE (second dimension)
- Visualization of proteins spots
- Identification of protein spots
- Annotation & spot evaluation

2D Gel Principles*

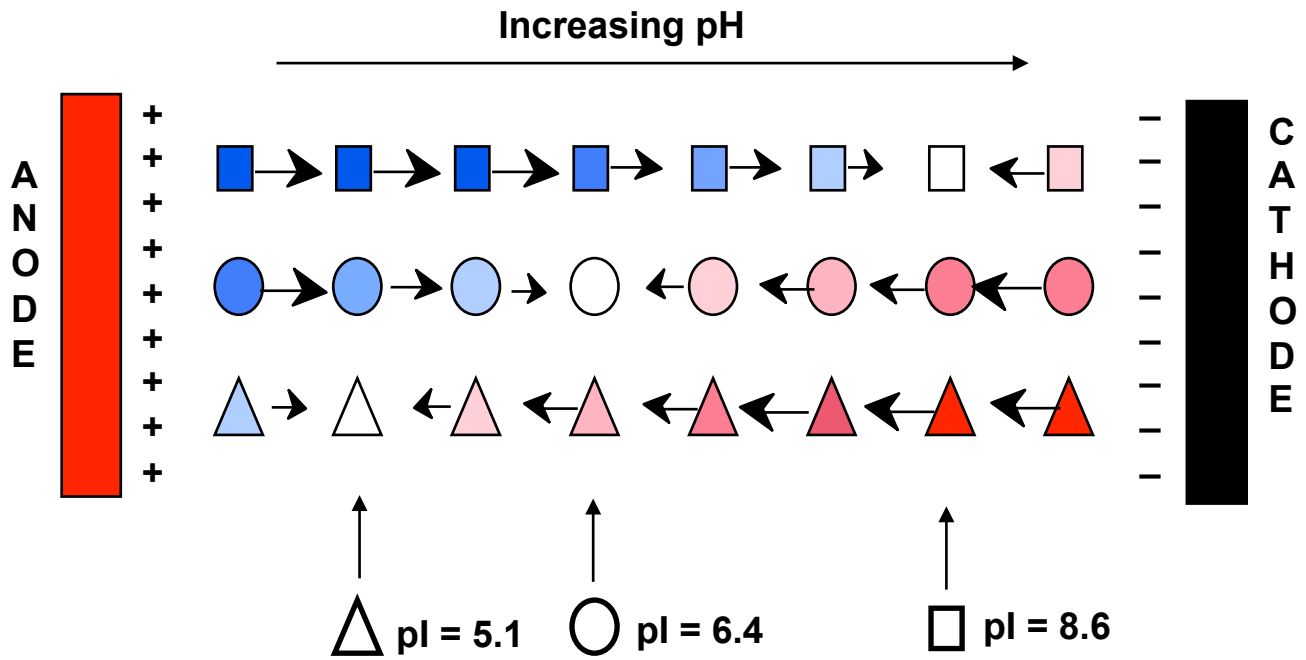


SDS
PAGE

Isoelectric Focusing (IEF)



IEF Principles*



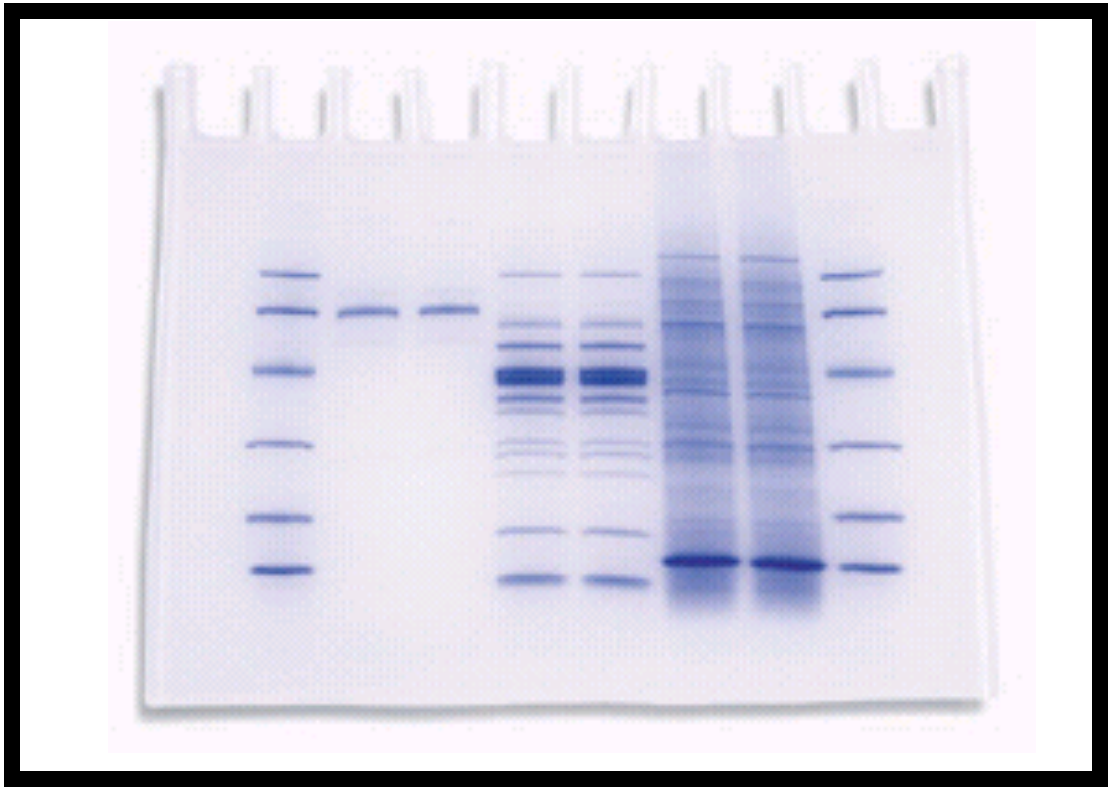
Isoelectric Focusing*

- 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)

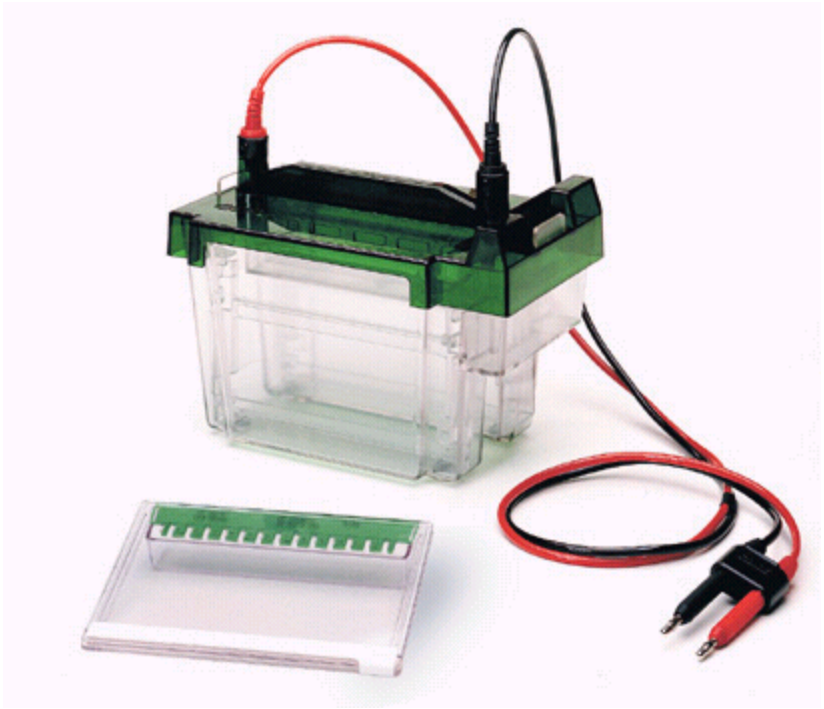
Steps in 2D GE & Peptide ID

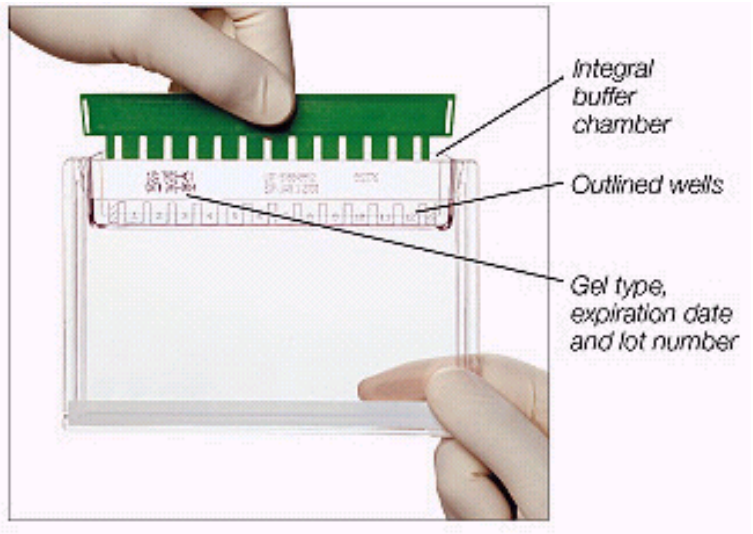
- Sample preparation
- Isoelectric focusing (first dimension)
- SDS-PAGE (second dimension)
- Visualization of proteins spots
- Identification of protein spots
- Annotation & spot evaluation

SDS PAGE



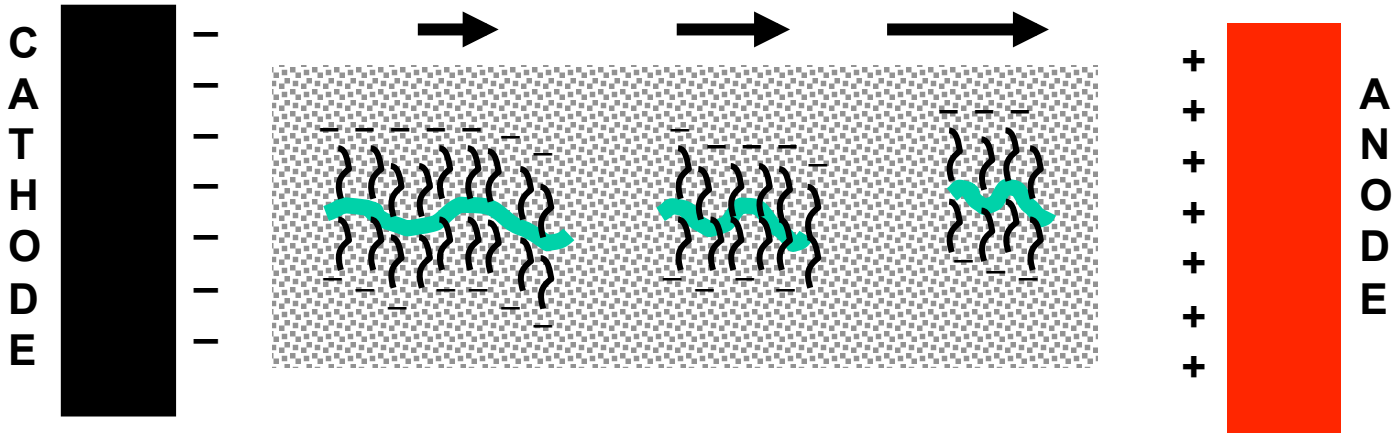
SDS PAGE Tools



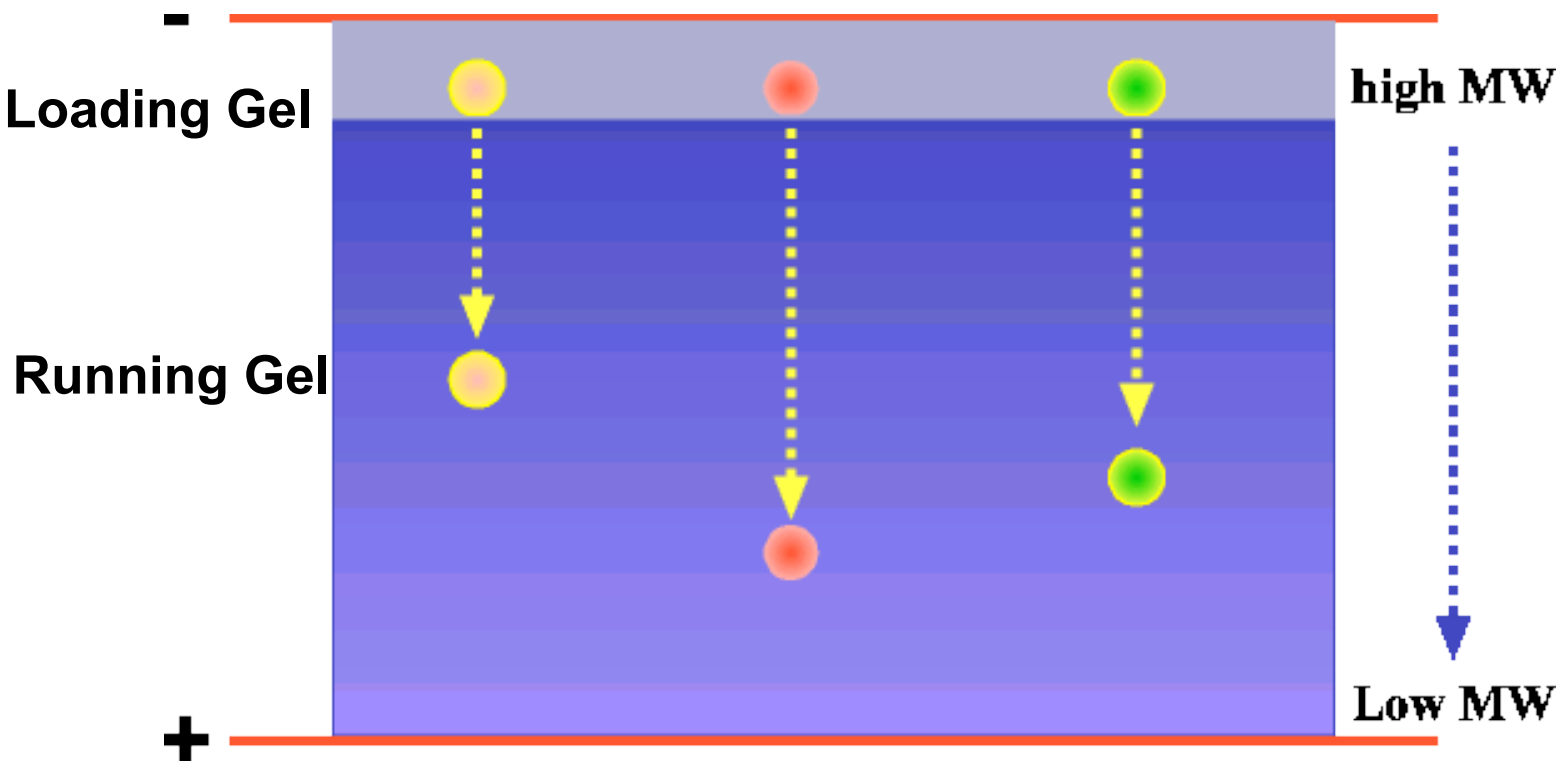


SDS PAGE Principles*

Sodium Dodecyl Sulfate



SDS-PAGE Principles*



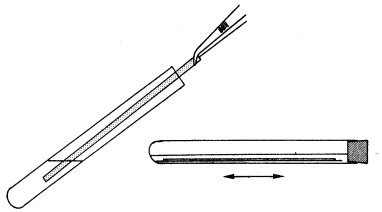
SDS-PAGE

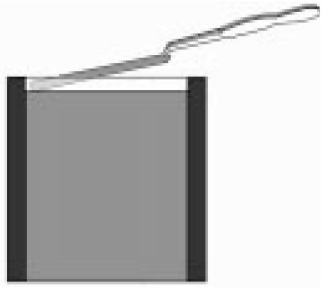
- Separation of basis of MW, not pI
- Requires modest voltages (200V)
- Requires a shorter period of time (2h)
- Presence of SDS is critical to disrupting structure and making mobility ~ 1/MW
- Degree of resolution determined by %acrylamide & electric field strength

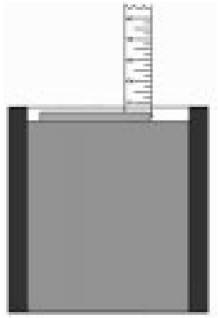
SDS-PAGE for 2D GE

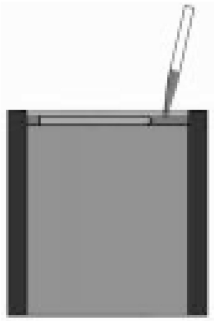
- **After IEF, the IPG strip is soaked in an equilibration buffer (50 mM Tris, pH 8.8, 2% SDS, 6M Urea, 30% glycerol, DTT, tracking dye)**
- **IPG strip is then placed on top of pre-cast SDS-PAGE gel and electric current applied**
- **This is equivalent to pipetting samples into SDS-PAGE wells (an infinite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

SDS-PAGE for 2D GE





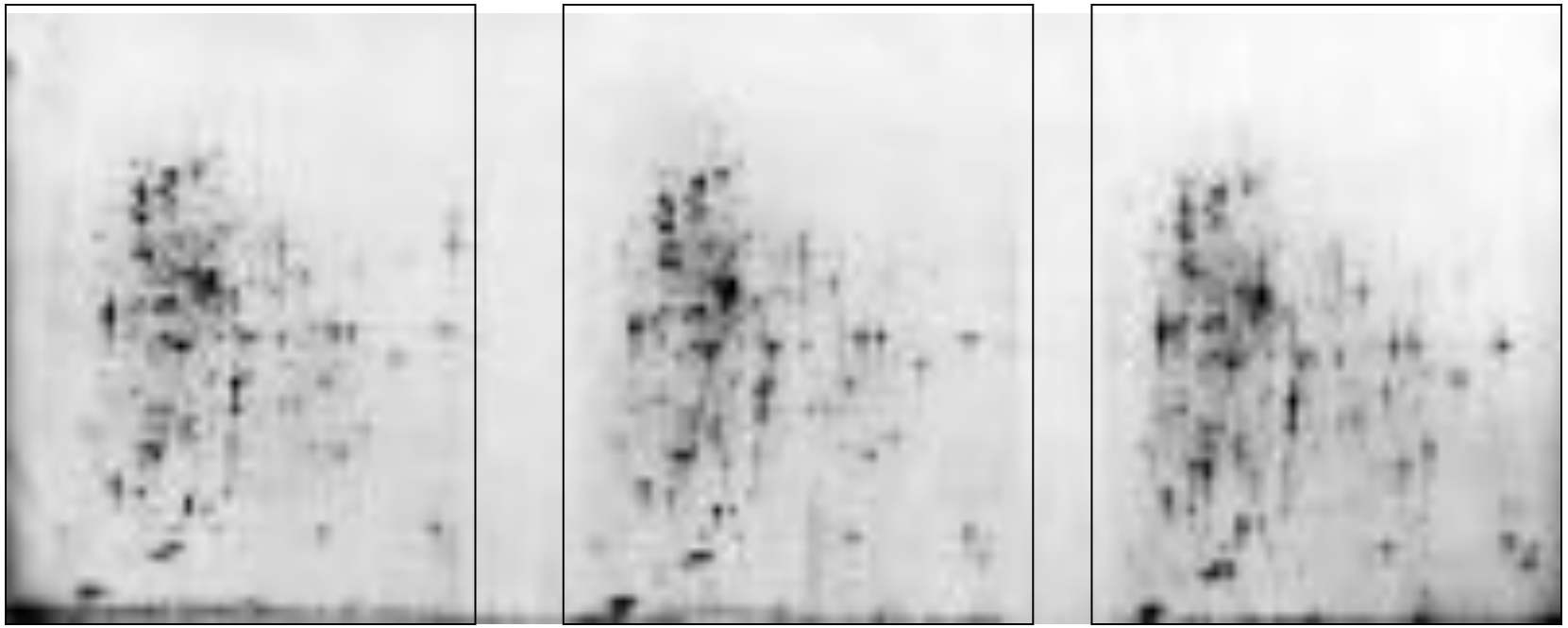




equilibration

SDS-PAGE

2D Gel Reproducibility



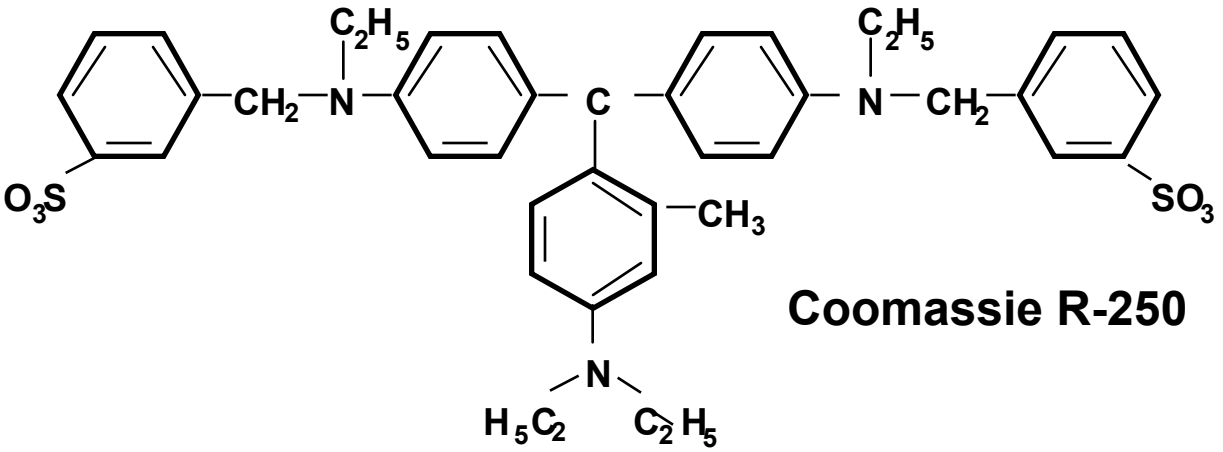
Advantages and Disadvantages of 2D GE*

- 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

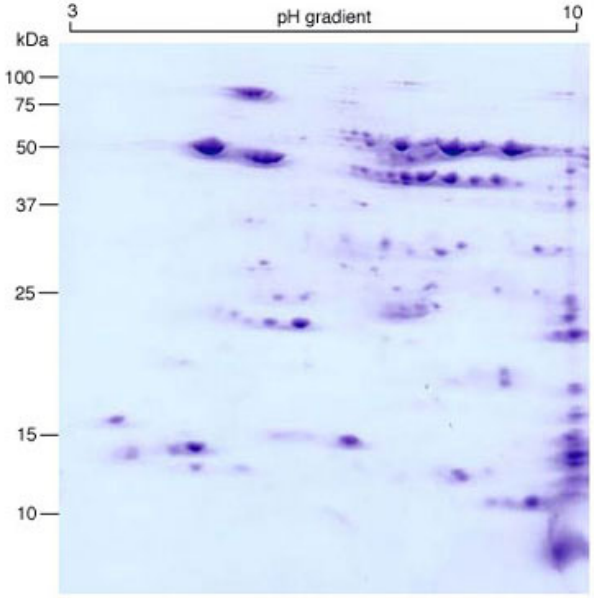
Protein Detection*

- Coomassie Stain (100 ng to 10 μg protein)
- Silver Stain (1 ng to 1 μg protein)
- Fluorescent (Sypro Ruby) Stain (1 ng & up)

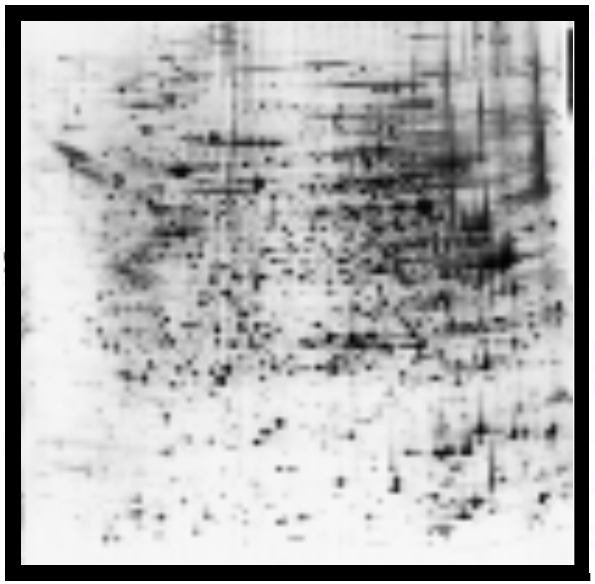


Coomassie R-250

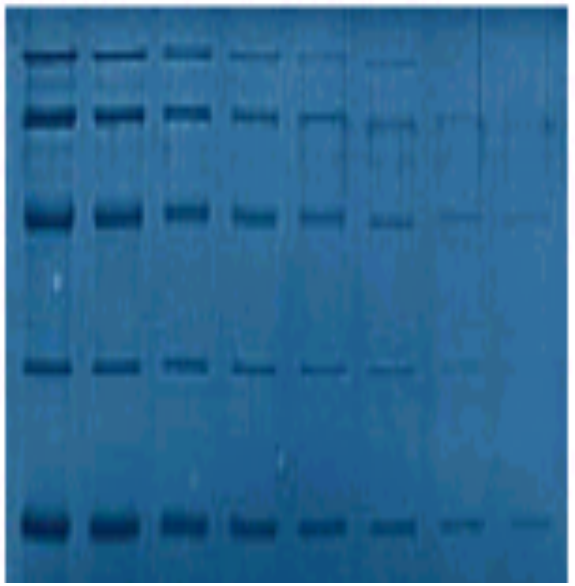
Stain Examples



Coomassie

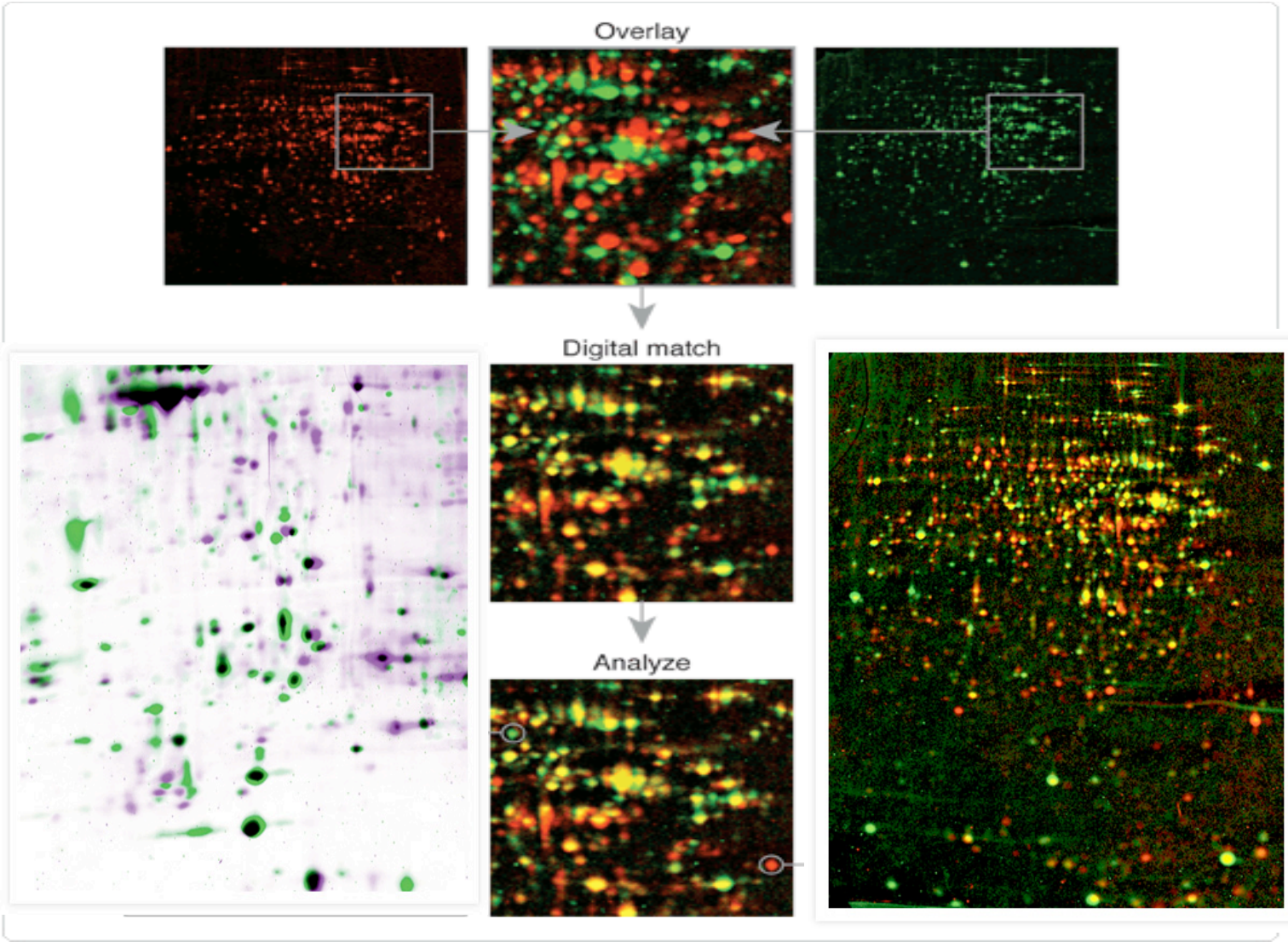


Silver Stain



Copper Stain

Multicolor Staining with Sypro fluorescent stains



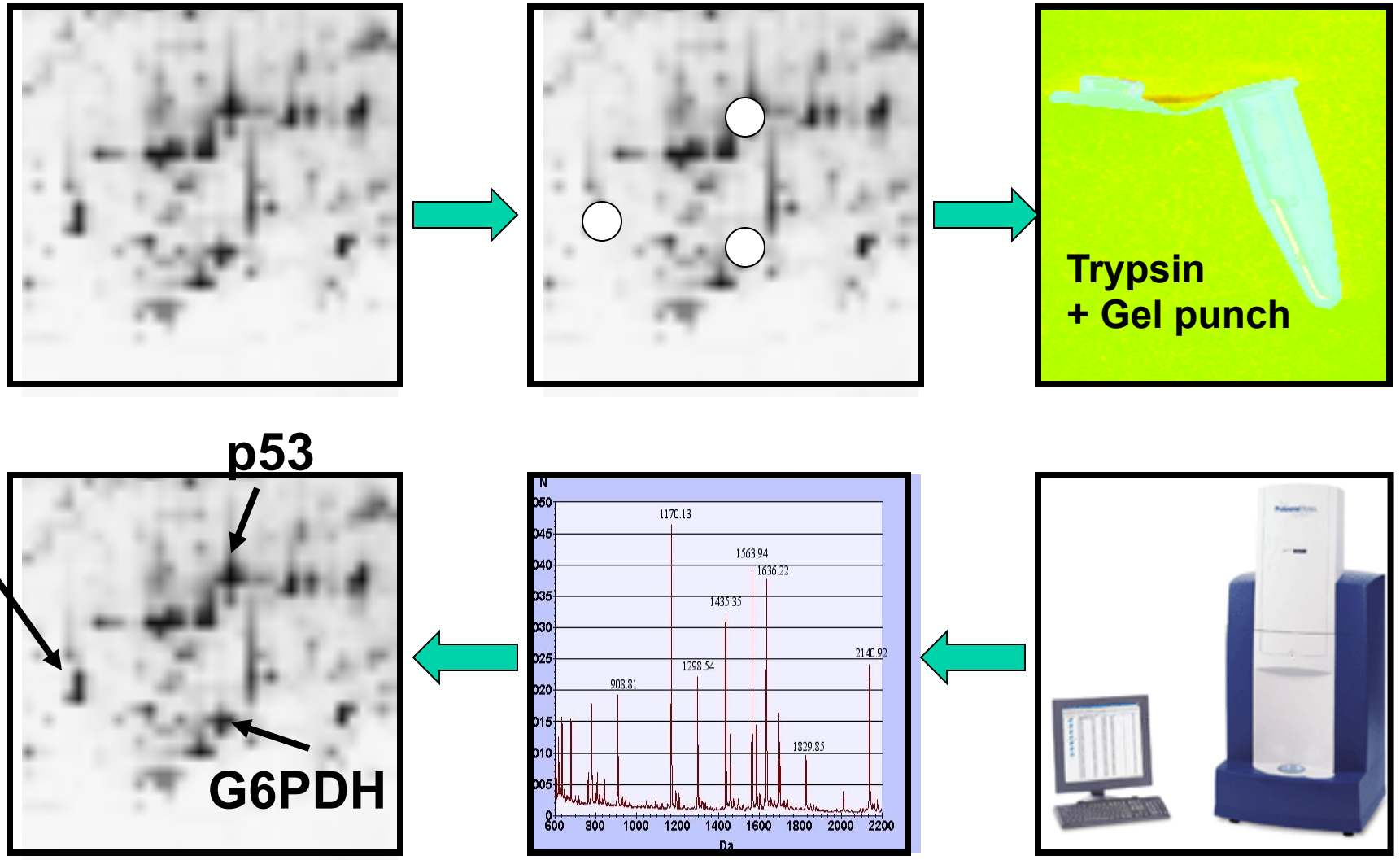
Steps in 2D GE & Peptide ID

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

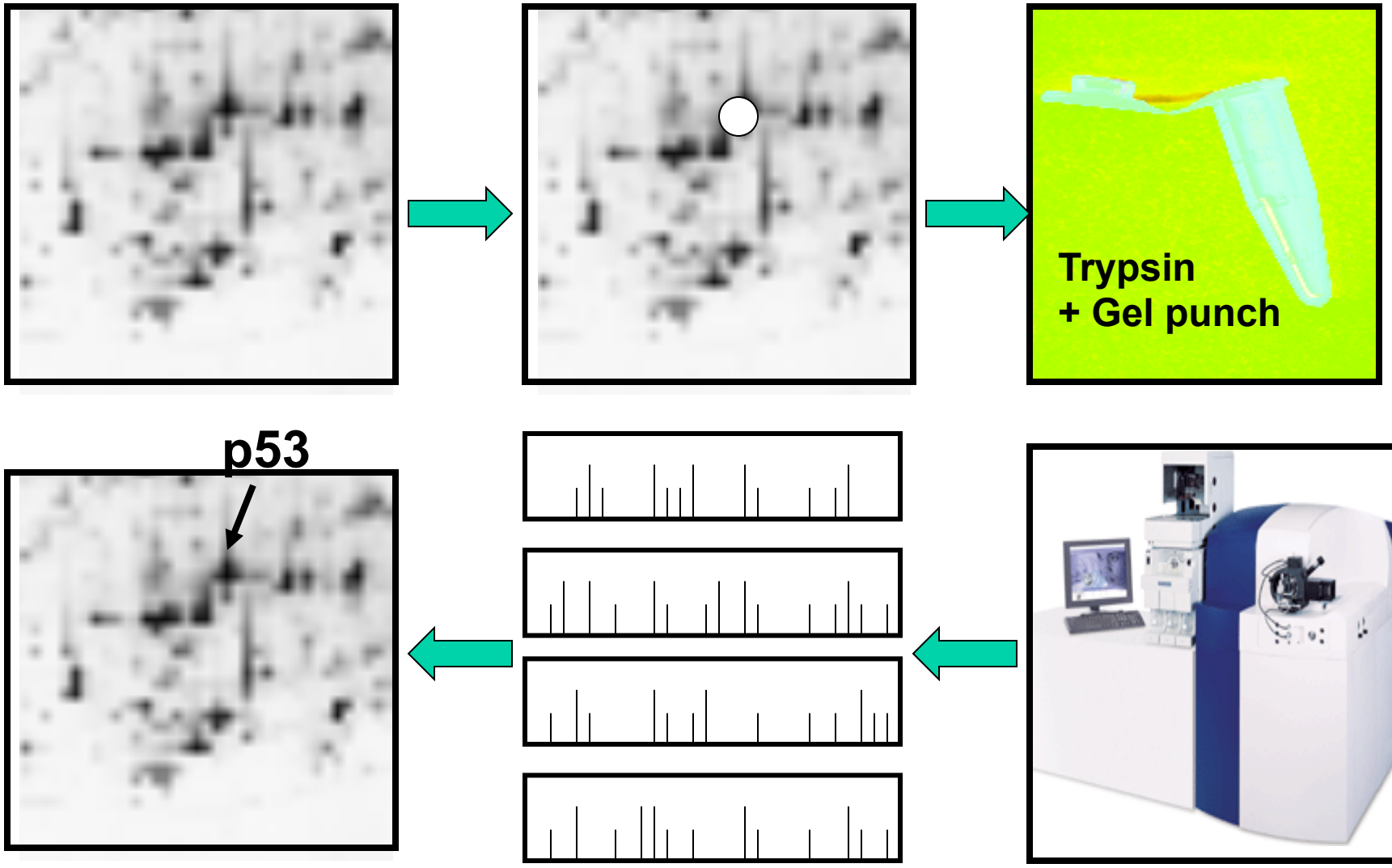
Protein Identification*

- 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

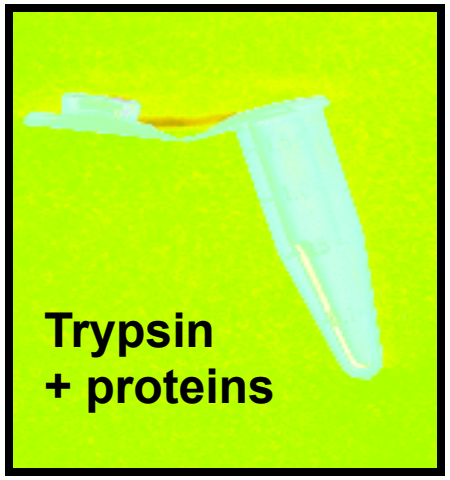
2D-GE + MALDI (PMF)*

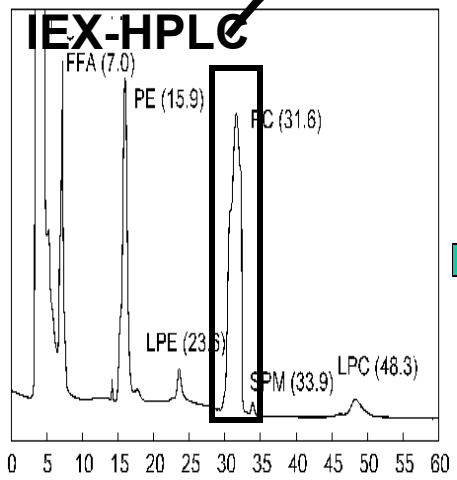


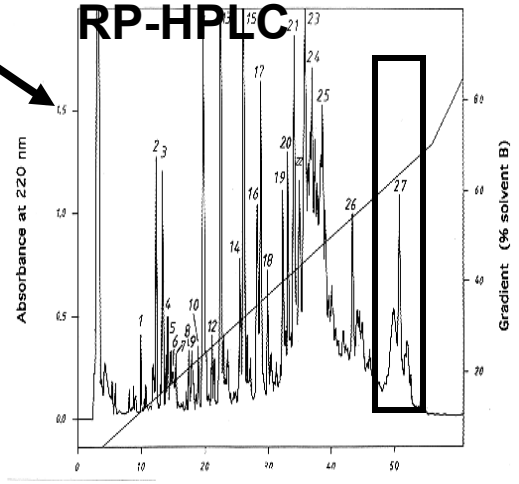
2D-GE + MS-MS

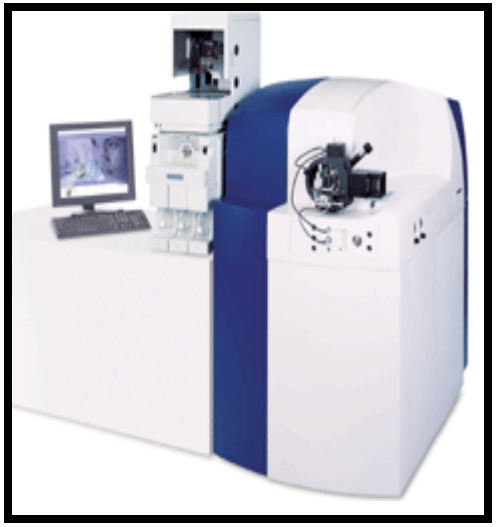


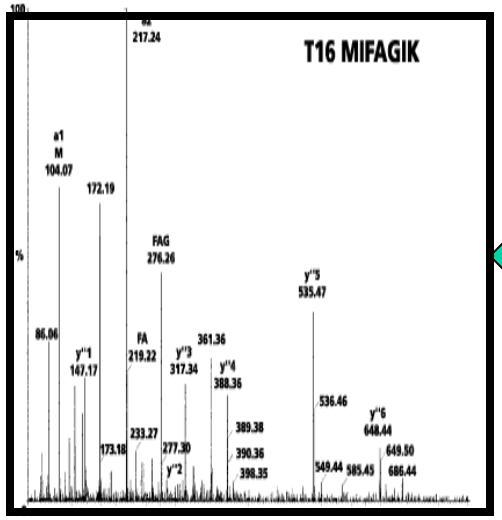
MudPIT

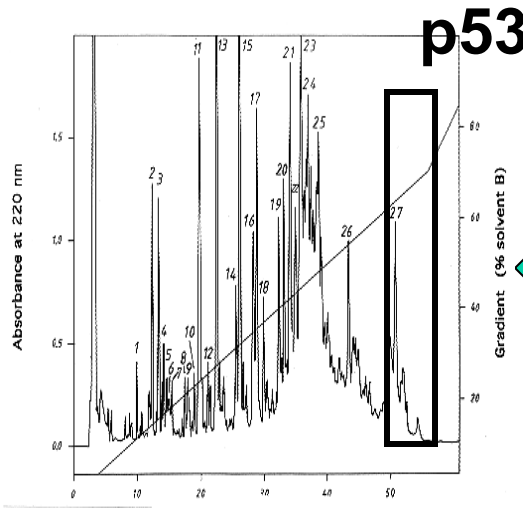

Trypsin
+ proteins


IEX-HPLC
FFA (7.0)
PE (15.9)
PC (31.6)
LPE (23.6)
SPM (33.9)
LPC (46.3)
0 5 10 15 20 25 30 35 40 45 50 55 60

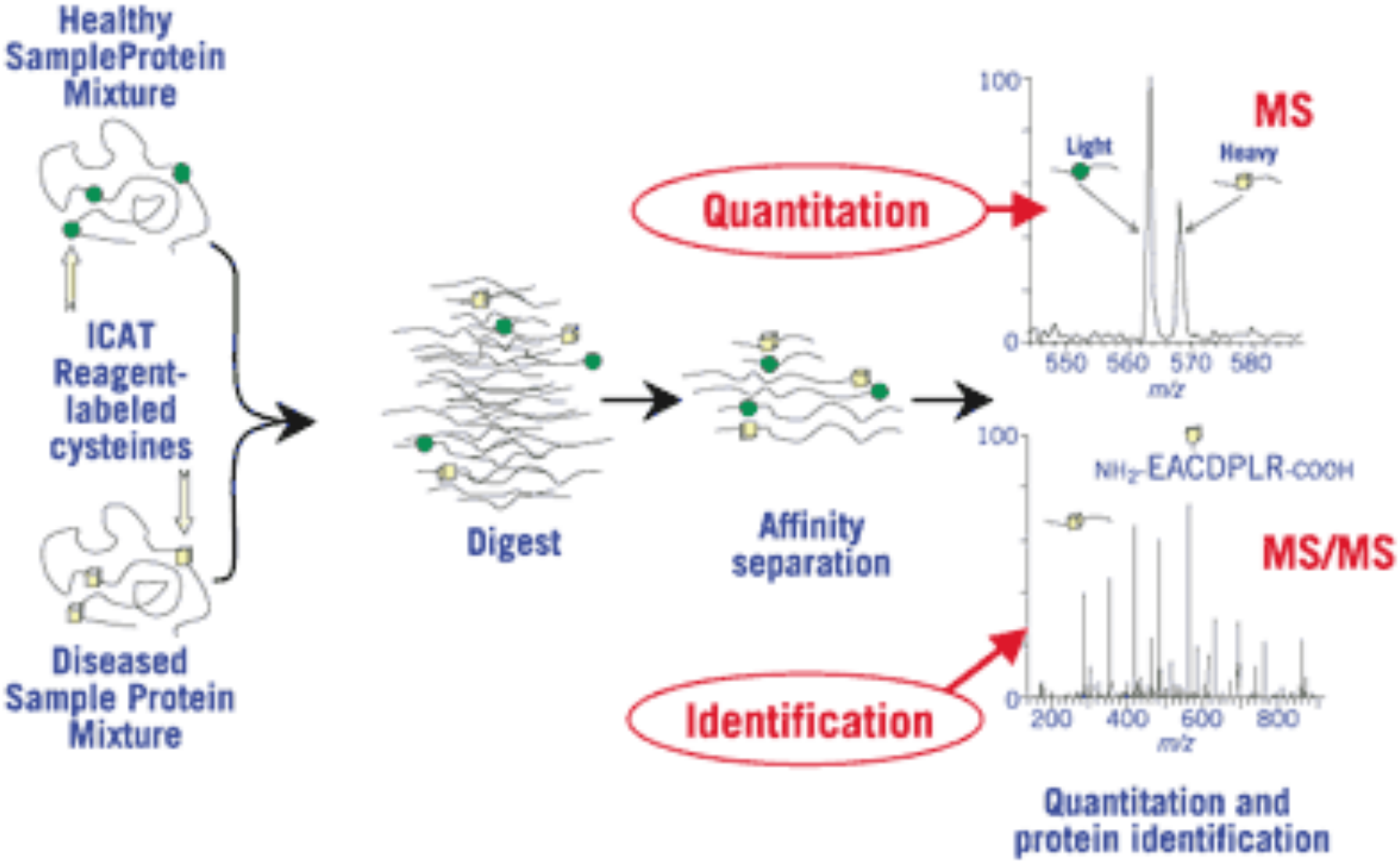

RP-HPLC
Absorbance at 220 nm
Gradient (% solvent B)




T16 MIFAGIK
217.24
a1
M
104.07
172.19
FAG
276.26
FA
219.22
y''1
147.17
y''3
317.34
361.36
y''4
388.36
y''5
535.47
536.46
y''6
648.44
649.50
686.44
549.44
585.45
389.38
390.36
398.35
277.30
233.27
173.18
86.06

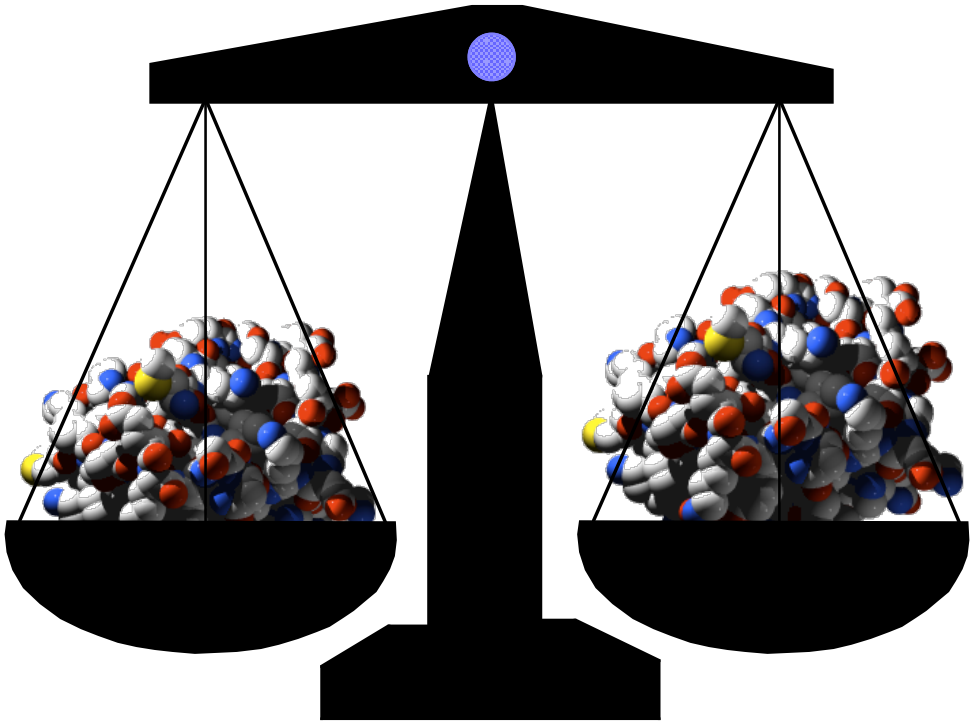

p53
Absorbance at 220 nm
Gradient (% solvent B)

ICAT (Isotope Coded Affinity Tag)*



Mass Spectrometry

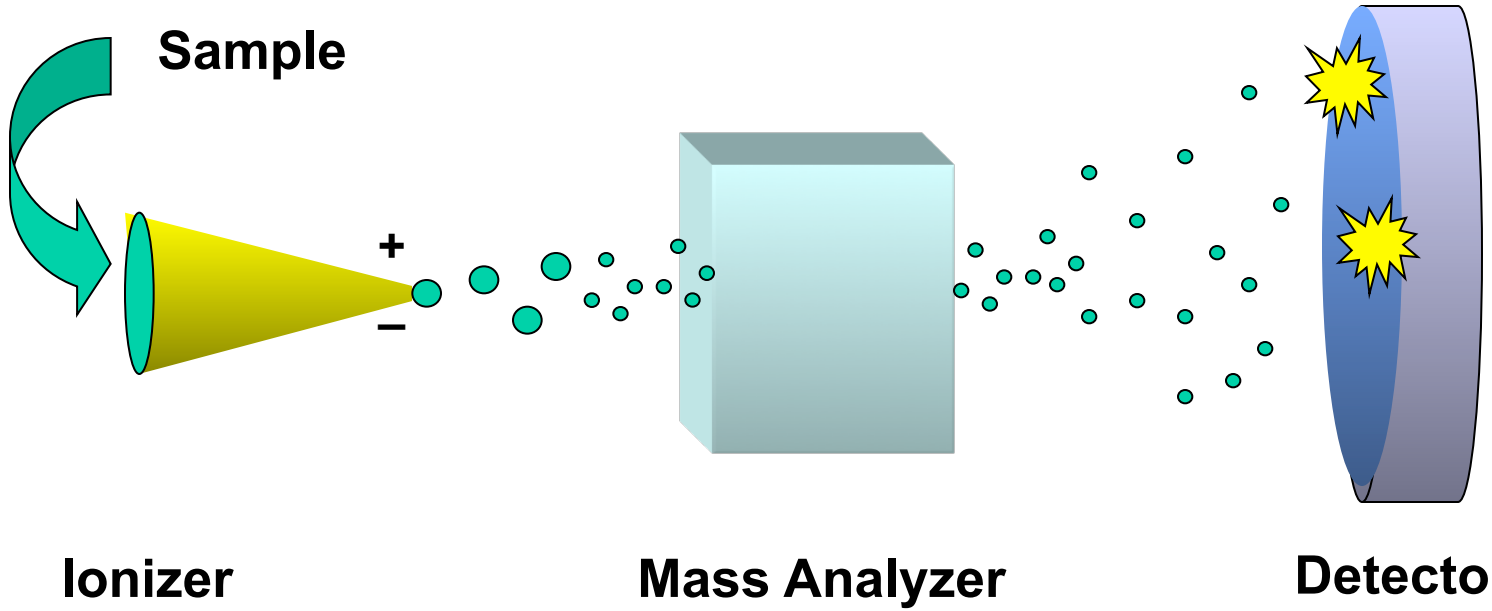
- **Analytical method to measure the molecular or atomic weight of samples**



MS Principles*

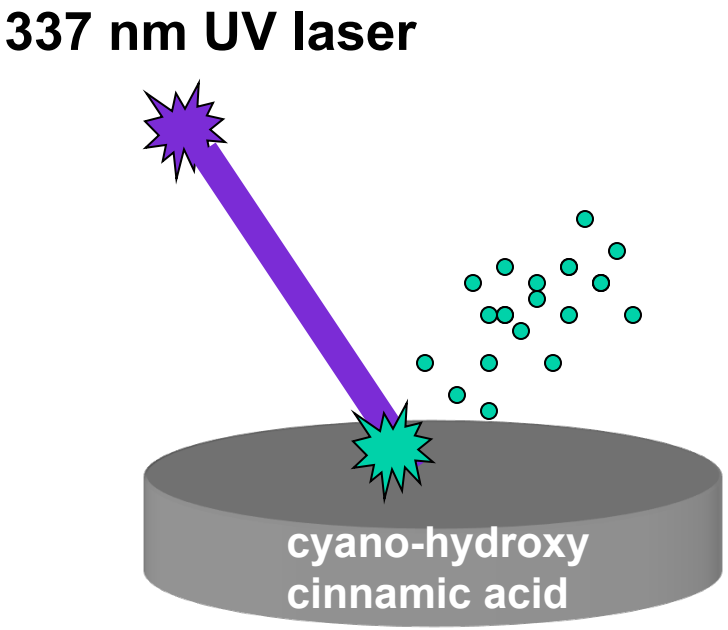
- Find a way to "charge" an atom or molecule (ionization)
- Place charged atom or molecule in a magnetic field or subject it to an electric field and measure its speed or radius of curvature relative to its mass-to-charge ratio (mass analyzer)
- Detect ions using microchannel plate or photomultiplier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

Mass Spec Principles*



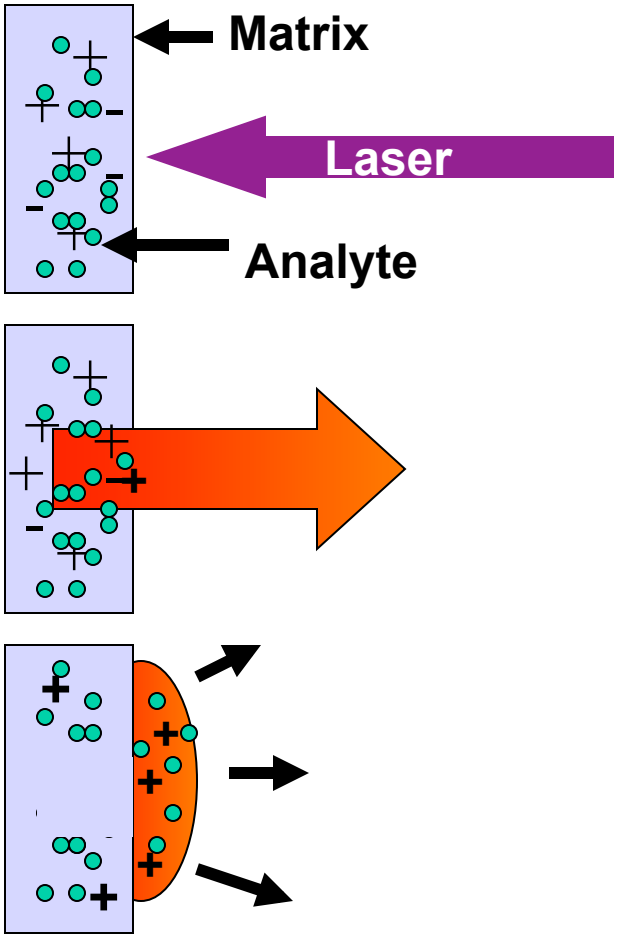
Typical Mass Spectrometer

Matrix-Assisted Laser Desorption Ionization



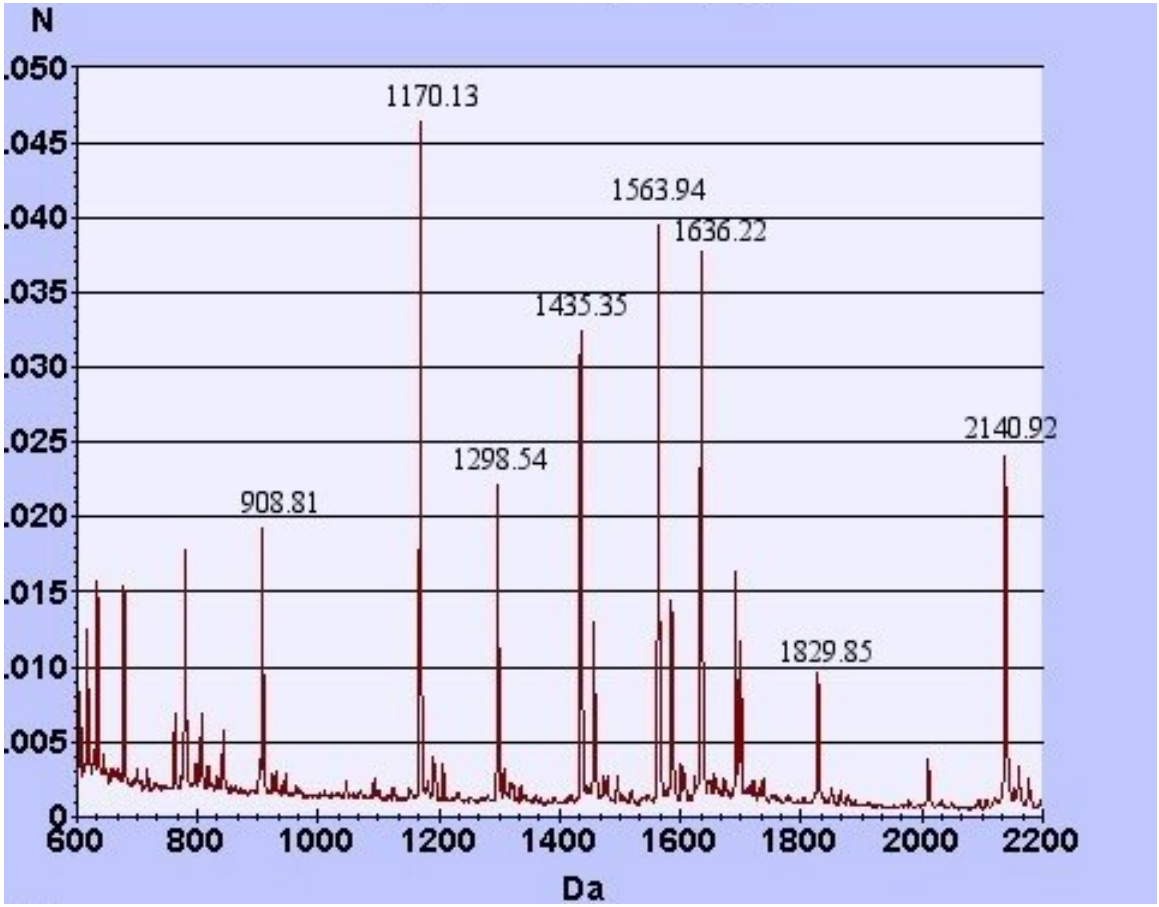
MALDI

MALDI Ionization*

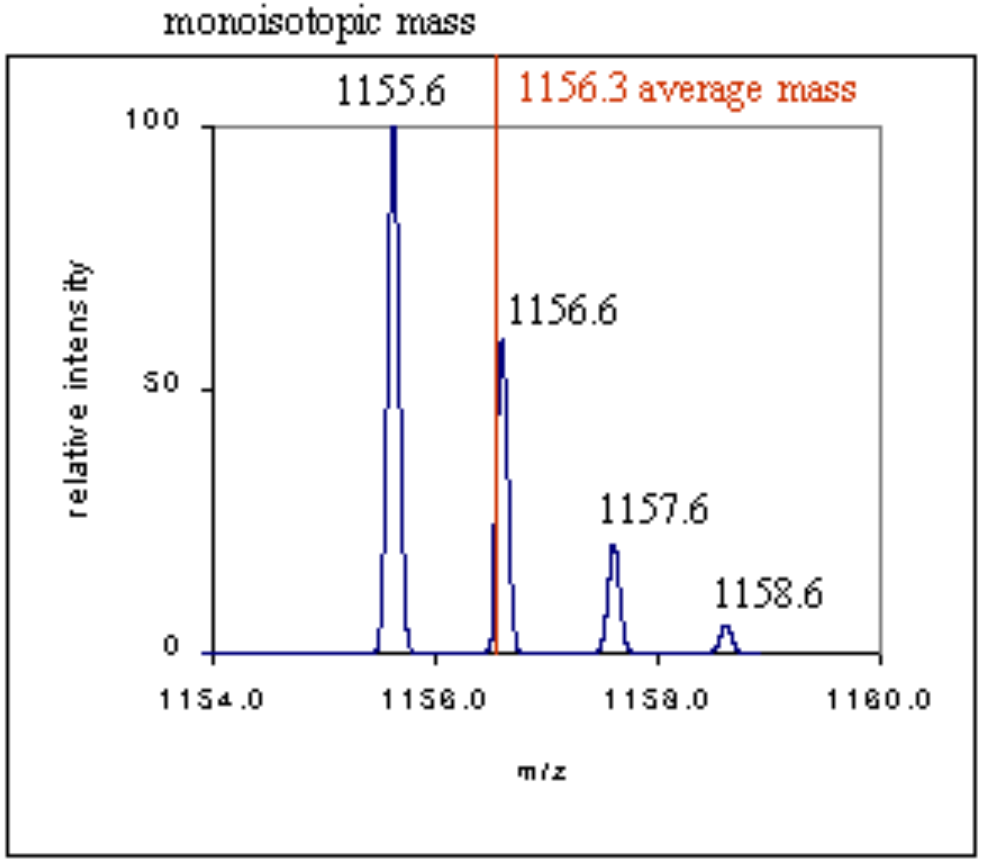


- Absorption of UV radiation by chromophoric matrix and ionization of matrix
- Dissociation of matrix, phase change to super-compressed gas, charge transfer to analyte molecule
- Expansion of matrix at supersonic velocity, analyte trapped in expanding matrix plume (explosion/"popping")

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

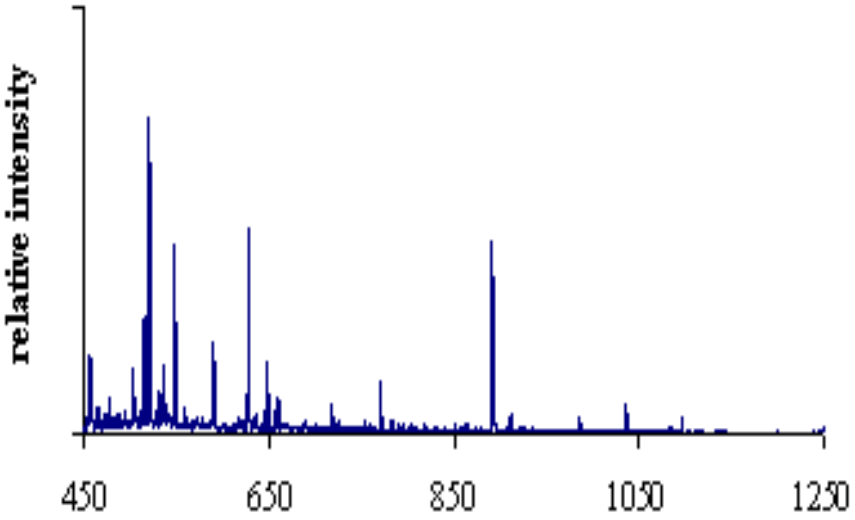
Average Mass

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_2O (18.01056)
- Add mass of H^+ (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)



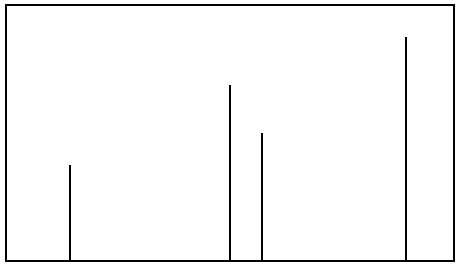
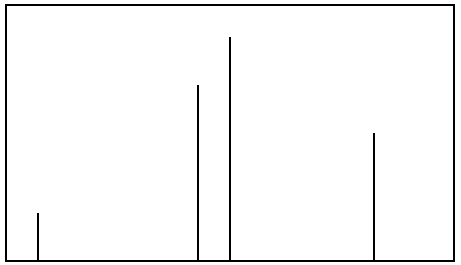
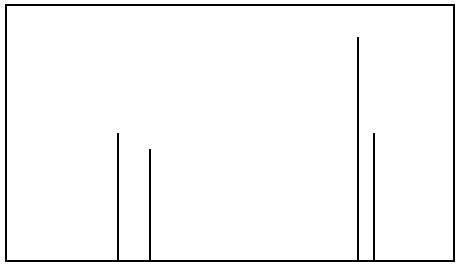
Peptide Mass Fingerprinting*

- Used to identify protein spots on gels or protein peaks from an HPLC run
- Depends of 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

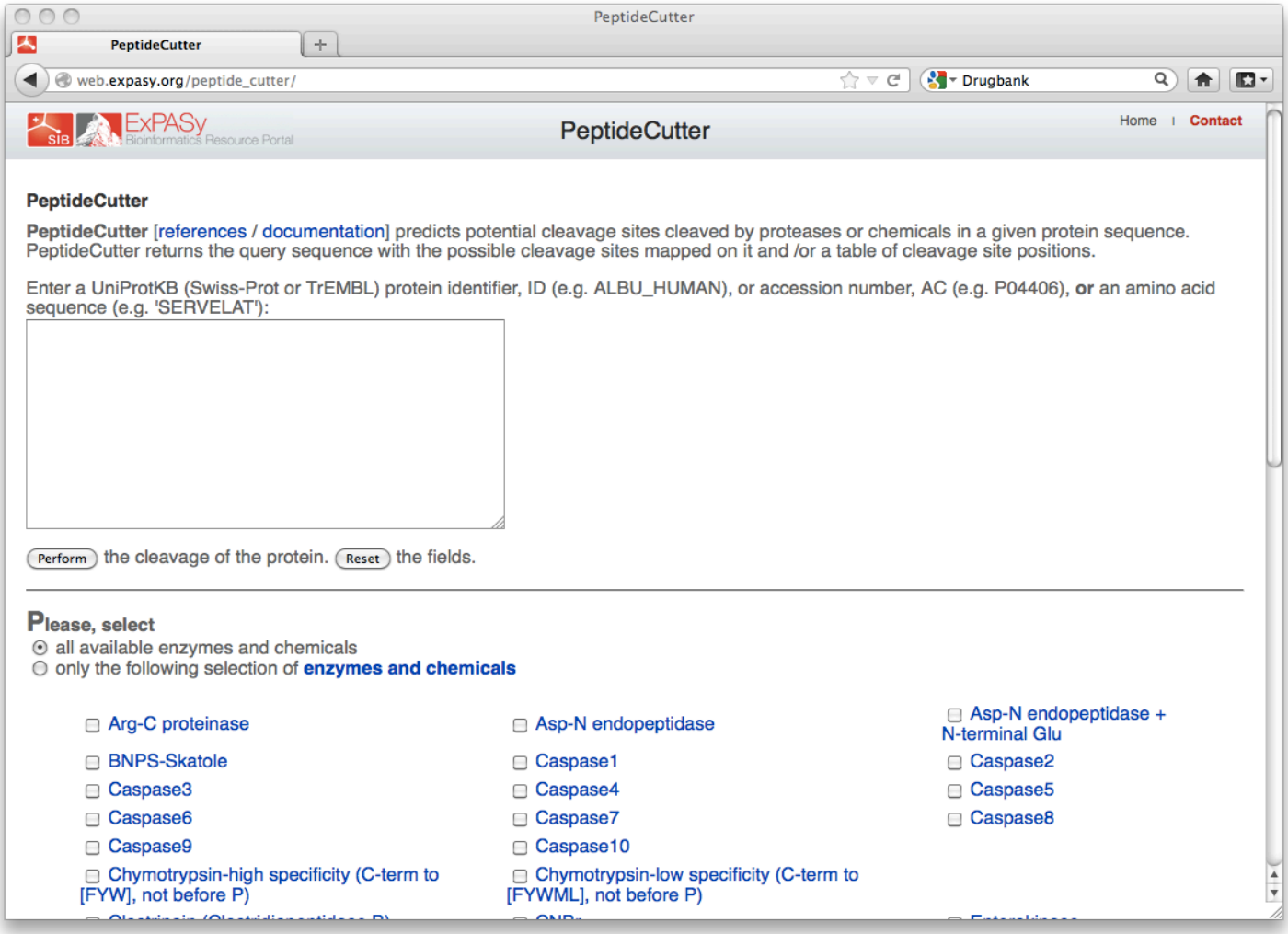
Principles of Fingerprinting*

Sequence	Mass (M+H)	Tryptic Fragments
>Protein 1 acedfhsakdfqea sdfpkivtmeeewe ndadnfekqwfe	4842.05	acedfhsak dfgeasdfpk ivtmeeewendadnfek gwfe
>Protein 2 acekdfhsadfqea sdfpkivtmeeewe nkdadnfeqwfe	4842.05	acek dfhsadfgeasdfpk ivtmeeewenk dadnfeqwfe
>Protein 3 acedfhsadfqeka sdfpkivtmeeewe ndakdnfeqwfe	4842.05	acedfhsadfgek asdfpk ivtmeeewendak dnfegwfe

Principles of Fingerprinting*

Sequence	Mass (M+H)	Mass Spectrum
>Protein 1 acedfhsakdfqea sdfpkivtmeeewe ndadnfekqwfe	4842.05	
>Protein 2 acekdfhsadfqea sdfpkivtmeeewe nkdadnfeqwfe	4842.05	
>Protein 3 acedfhsadfqeka sdfpkivtmeeewe ndakdnfeqwfe	4842.05	

Predicting Peptide Cleavages



http://web.expasy.org/peptide_cutter/

http://web.expasy.org/peptide_cutter/peptidecutter_enzymes.html

PeptideCutter

The cleavage specificities of selected enzymes and chemicals:

A general model of enzymatic cleavage:

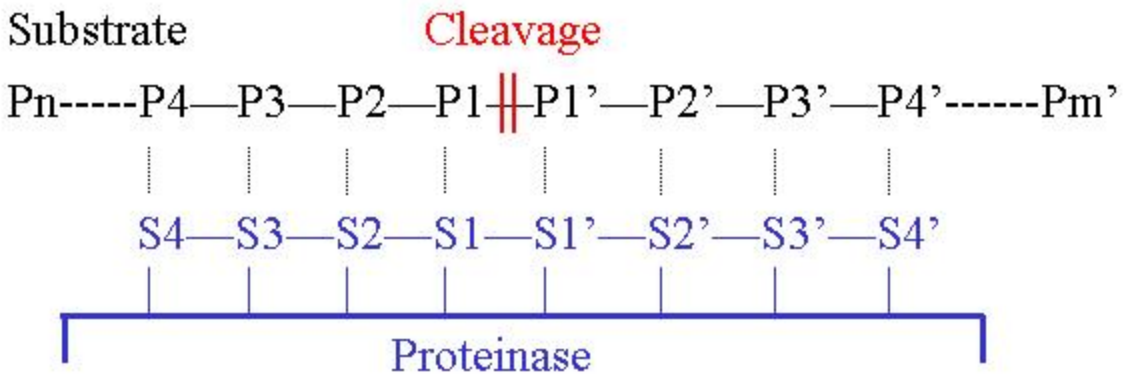
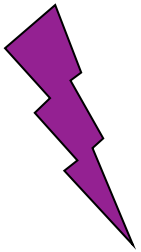


Fig.1 Schematic representation of enzyme-substrate complex with eight binding sites. Positions Pn to Pm' in the substrate are counted from the bond between P1 and P1', where the cleavage occurs.

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

Sequence DB

```
>P12345
acedfhsakdfqea
sdfpkivtmeeewe
ndadnfekqwfe

>P21234
acekdfhsadfqea
sdfpkivtmeeewe
nkdadnfeqwfe

>P89212
acedfhsadfqeka
sdfpkivtmeeewe
ndakdnfeqwfe
```

Calc. Tryptic Frags

```
acedfhsak
dfgeasdfpk
ivtmeeewendadnfek
gwfe

acek
dfhsadfgeasdfpk
ivtmeeewenk
dadnfeqwfe

acedfhsadfgek
asdfpk
ivtmeeewendak
dnfegwfe
```

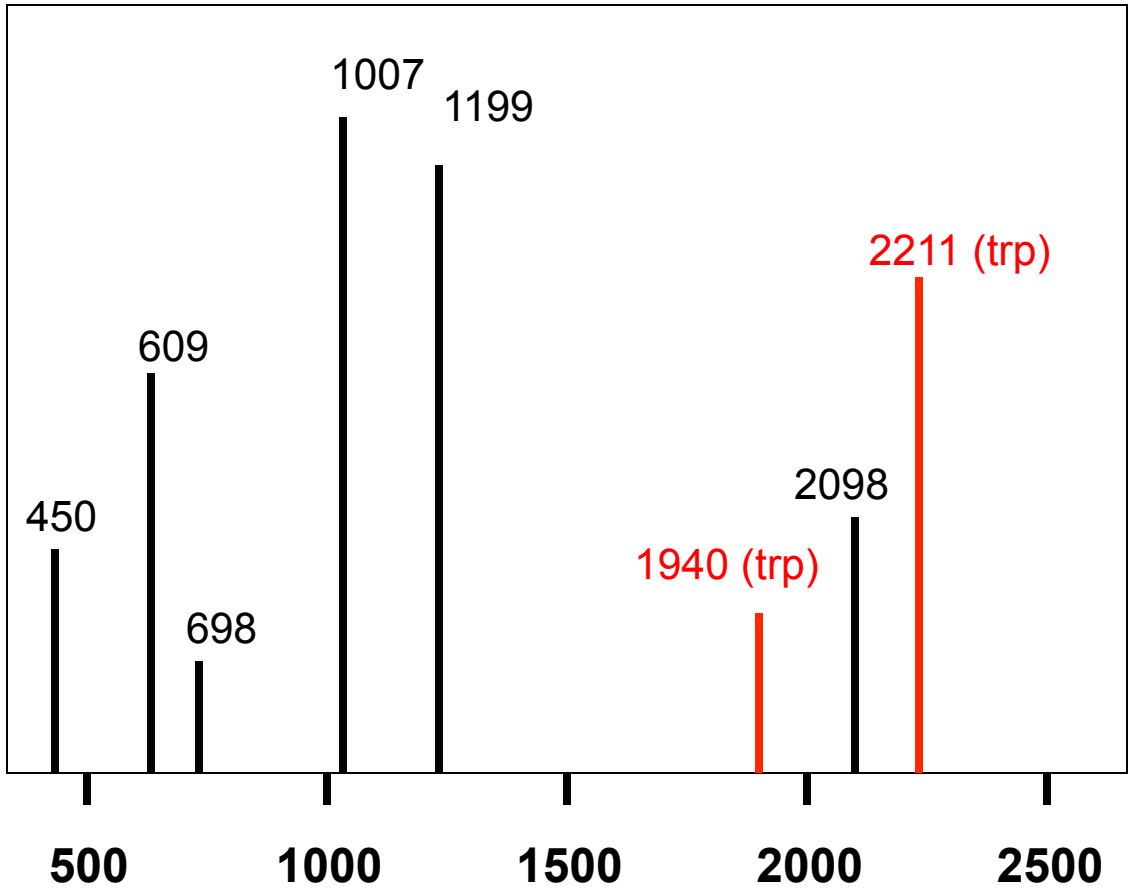
Mass List

```
450.2017 (P21234)
609.2667 (P12345)
664.3300 (P89212)
1007.4251 (P12345)
1114.4416 (P89212)
1183.5266 (P12345)
1300.5116 (P21234)
1407.6462 (P21234)
1526.6211 (P89212)
1593.7101 (P89212)
1740.7501 (P21234)
2098.8909 (P12345)
```

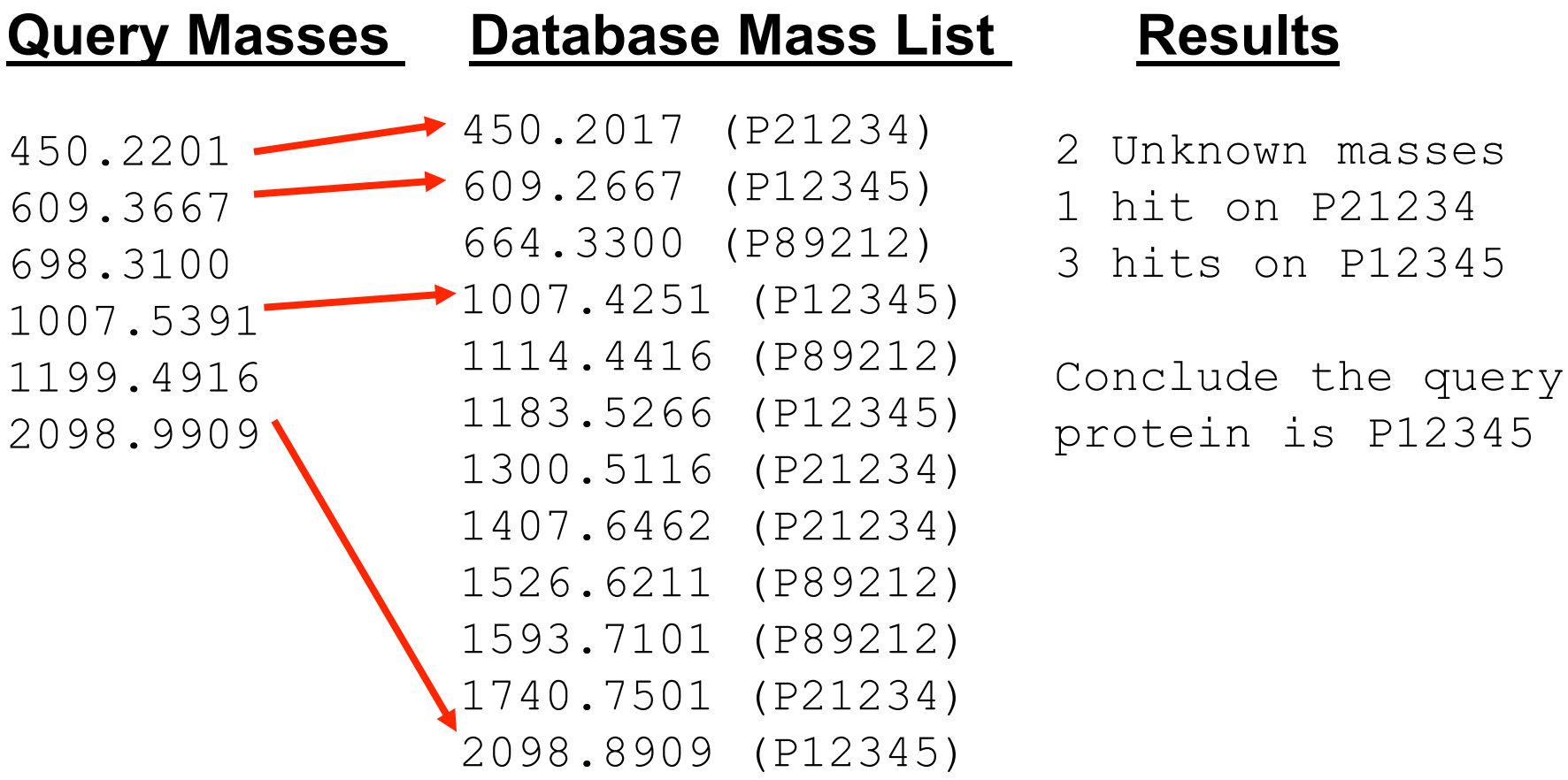
The Fingerprint (PMF) Algorithm*

- 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



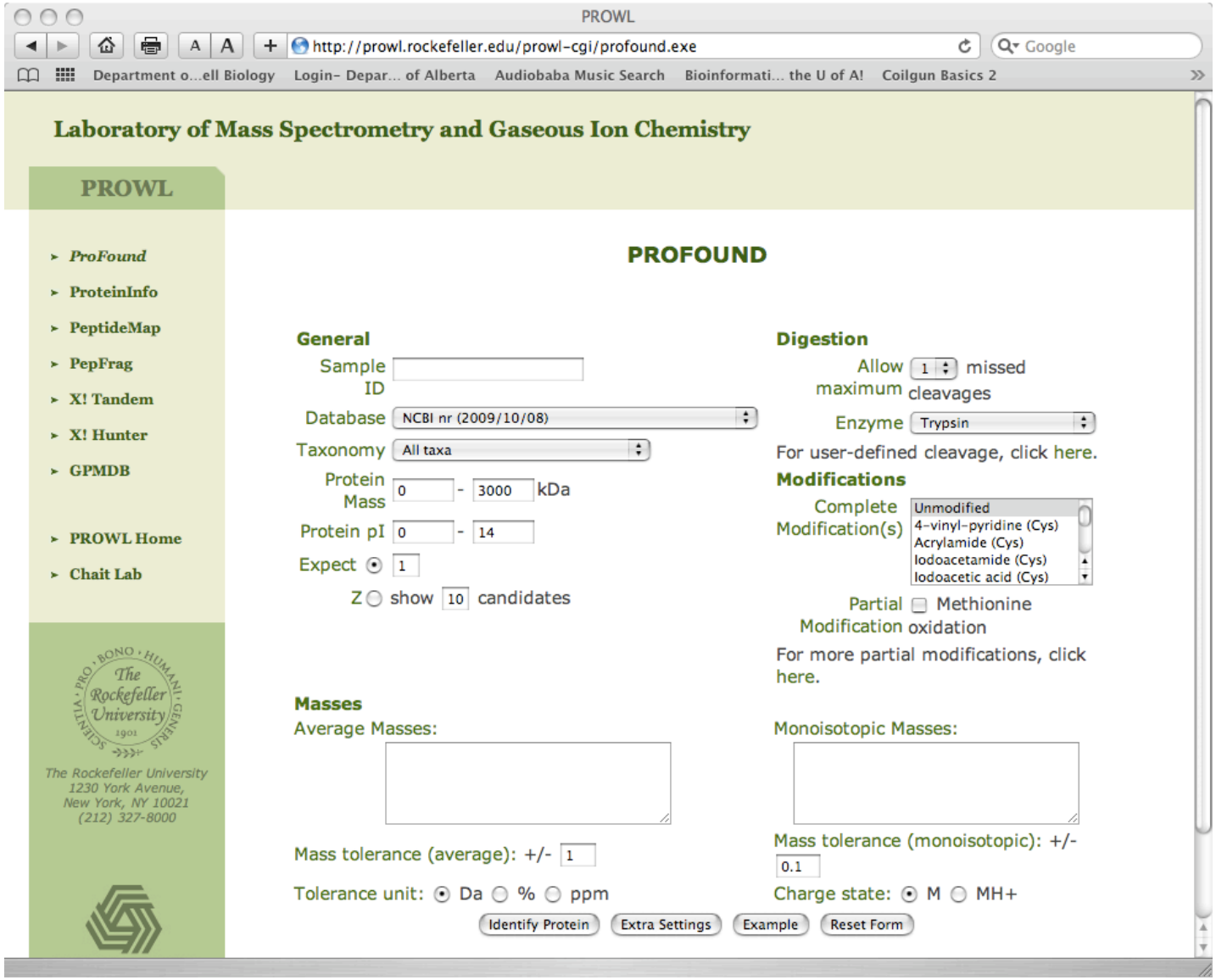
What You Need To Do PMF*

- 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 ± 0.1 Da)
- *A PMF website (Prowl, ProFound, Mascot, etc.)*

PMF on the Web

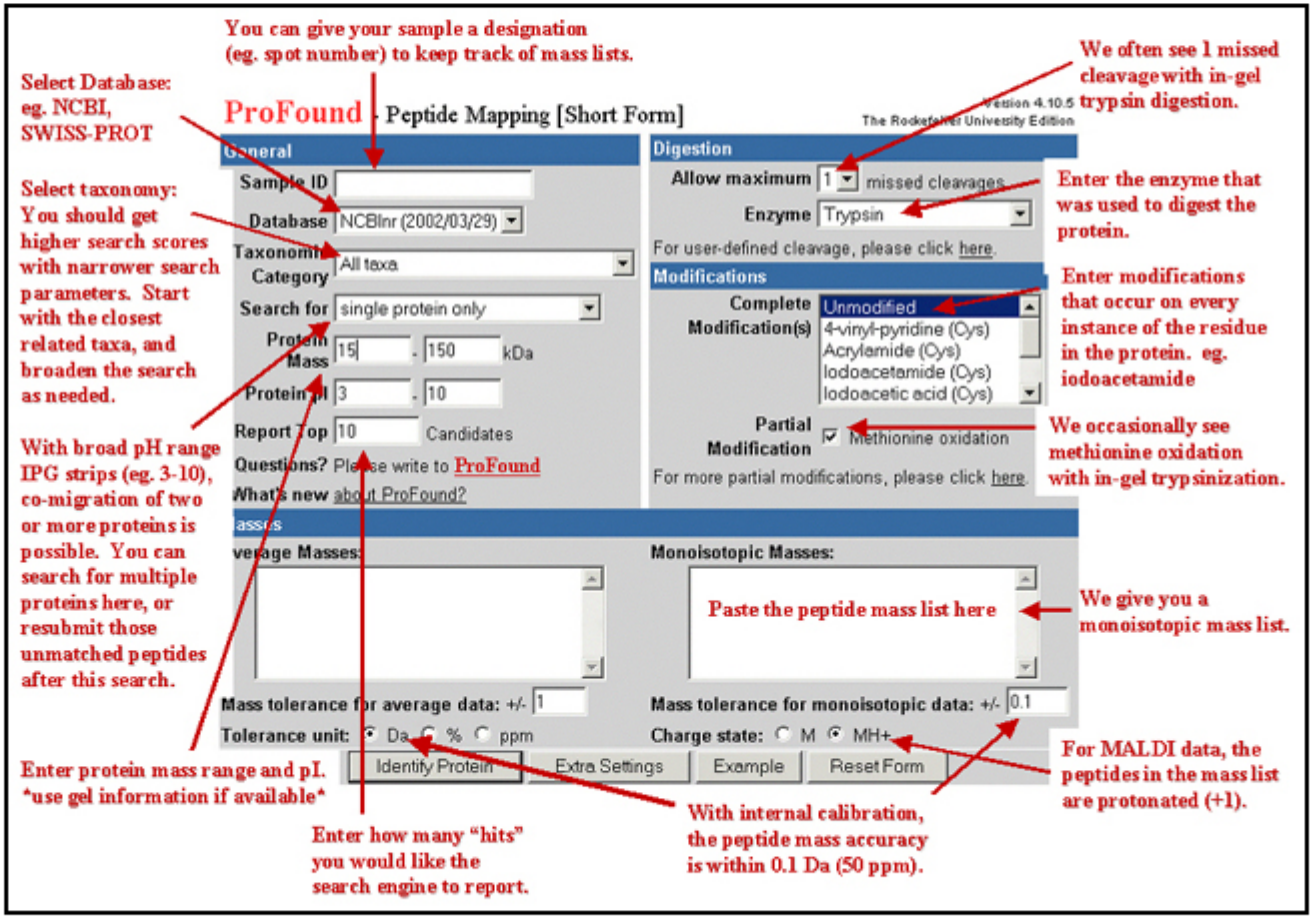
- **ProFound**
  - **http://prowl.rockefeller.edu/prowl-cgi/profound.exe**
- **Mascot**
  - **http://www.matrixscience.com**
- **ProteinProspector**
  - **http://prospector.ucsf.edu/prospector/mshome.htm**

ProFound


PROWL
http://prowl.rockefeller.edu/prowl-cgi/profound.exe
Google
Department o...ell Biology
Login- Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
Laboratory of Mass Spectrometry and Gaseous Ion Chemistry
PROWL
ProFound
ProteinInfo
PeptideMap
PepFrag
X! Tandem
X! Hunter
GPMDB
PROWL Home
Chait Lab
The Rockefeller University
1230 York Avenue,
New York, NY 10021
(212) 327-8000
PROFOUND
General
Sample ID
Database
NCBI nr (2009/10/08)
Taxonomy
All taxa
Protein Mass
0
3000
kDa
Protein pI
0
14
Expect
1
Z
show
10
candidates
Digestion
Allow maximum
1
missed cleavages
Enzyme
Trypsin
For user-defined cleavage, click here.
Modifications
Complete Modification(s)
Unmodified
4-vinyl-pyridine (Cys)
Acrylamide (Cys)
Iodoacetamide (Cys)
Iodoacetic acid (Cys)
Partial Modification
Methionine oxidation
For more partial modifications, click here.
Masses
Average Masses:
Monoisotopic Masses:
Mass tolerance (average): +/-
1
Mass tolerance (monoisotopic): +/-
0.1
Tolerance unit:
Da
%
ppm
Charge state:
M
MH+
Identify Protein
Extra Settings
Example
Reset Form

ProFound (PMF)

PROWL (ProFound)



What Are Missed Cleavages?

Sequence

```
>Protein 1
acedfhsakdfqea
sdfpkivtmeeewe
ndadnfekqwfe
```

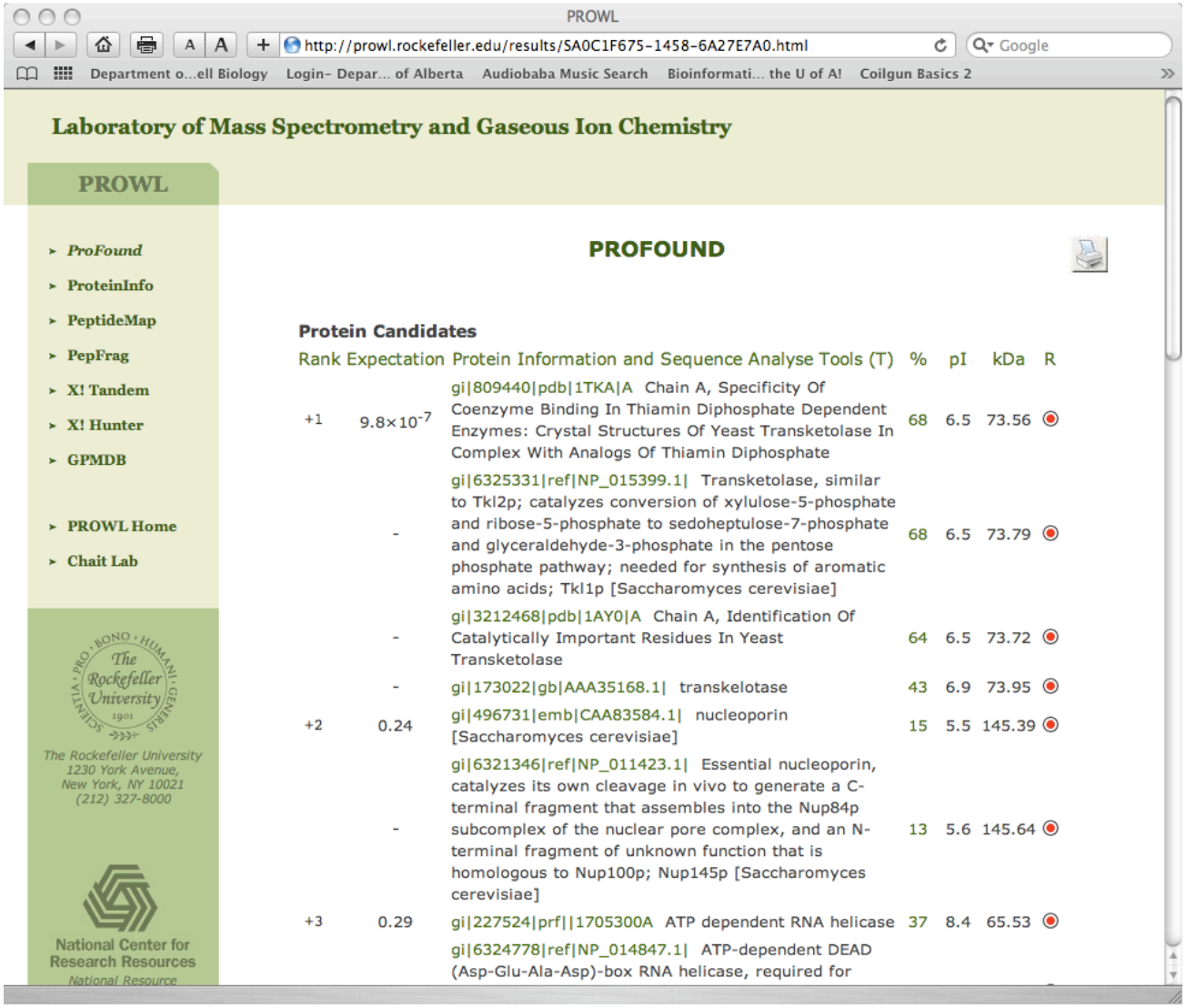
Tryptic Fragments (no missed cleavage)

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

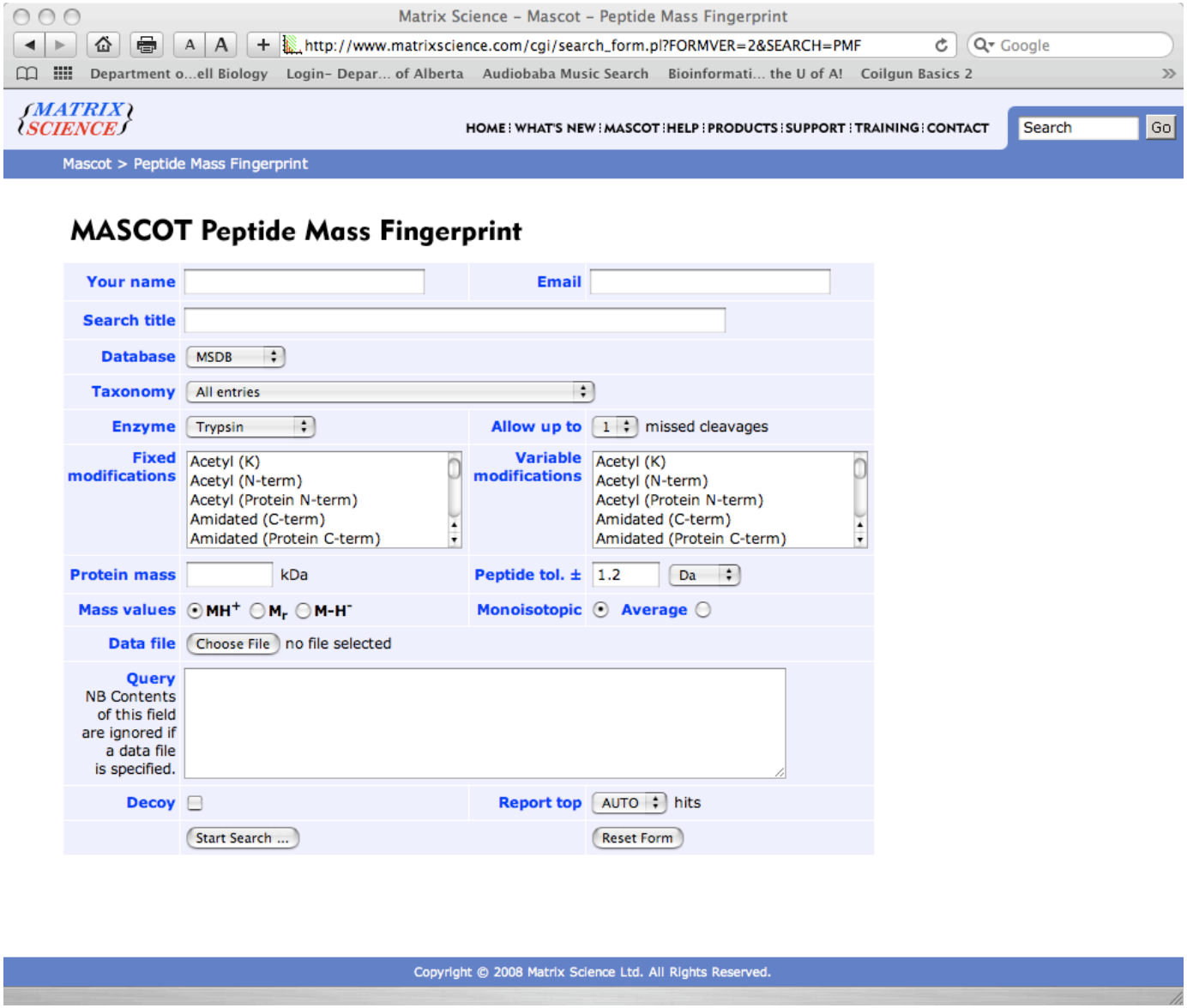
Tryptic Fragments (1 missed cleavage)

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

ProFound Results



MASCOT


Matrix Science – Mascot – Peptide Mass Fingerprint
http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF
Google
Department o...ell Biology
Login- Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
MATRIX SCIENCE
HOME | WHAT'S NEW | MASCOT | HELP | PRODUCTS | SUPPORT | TRAINING | CONTACT
Search
Go
Mascot > Peptide Mass Fingerprint
MASCOT Peptide Mass Fingerprint
Your name
Email
Search title
Database
MSDB
Taxonomy
All entries
Enzyme
Trypsin
Allow up to
1
missed cleavages
Fixed modifications
Acetyl (K)
Acetyl (N-term)
Acetyl (Protein N-term)
Amidated (C-term)
Amidated (Protein C-term)
Variable modifications
Acetyl (K)
Acetyl (N-term)
Acetyl (Protein N-term)
Amidated (C-term)
Amidated (Protein C-term)
Protein mass
kDa
Peptide tol. ±
1.2
Da
Mass values
MH+
Mr
M-H-
Monoisotopic
Average
Data file
Choose File
no file selected
Query
NB Contents of this field are ignored if a data file is specified.
Decoy
Report top
AUTO
hits
Start Search ...
Reset Form
Copyright © 2008 Matrix Science Ltd. All Rights Reserved.

MASCOT

Concise Summary Report (Peptide Mass Fingerprint Example)

http://www.matrixscience.com/cgi/master_results.pl?file=../data/F981122.dat

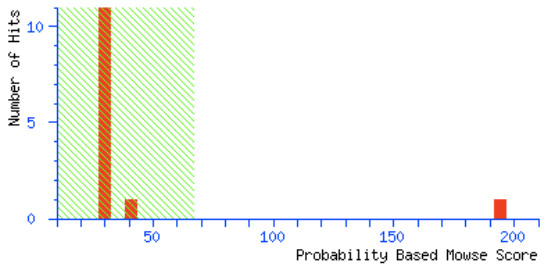
Department o...ell Biology | Login- Depar... of Alberta | Audiobaba Music Search | Bioinformati... the U of A! | Coilgun Basics 2

Mascot Search Results

```
User            :
Email           :
Search title    : Peptide Mass Fingerprint Example
Database        : SwissProt 51.6 (257964 sequences; 93947433 residues)
Timestamp       : 19 Feb 2007 at 14:08:10 GMT
Top Score       : 194 for PML_HUMAN, Probable transcription factor PML (Tripartite motif-containing protein 19) (RING fi
```

Probability Based Mowse Score

Protein score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 67 are significant (p<0.05).



Concise Protein Summary Report

Format As | Concise Protein Summary | Help

Significance threshold p< 0.05 | Max. number of hits AUTO

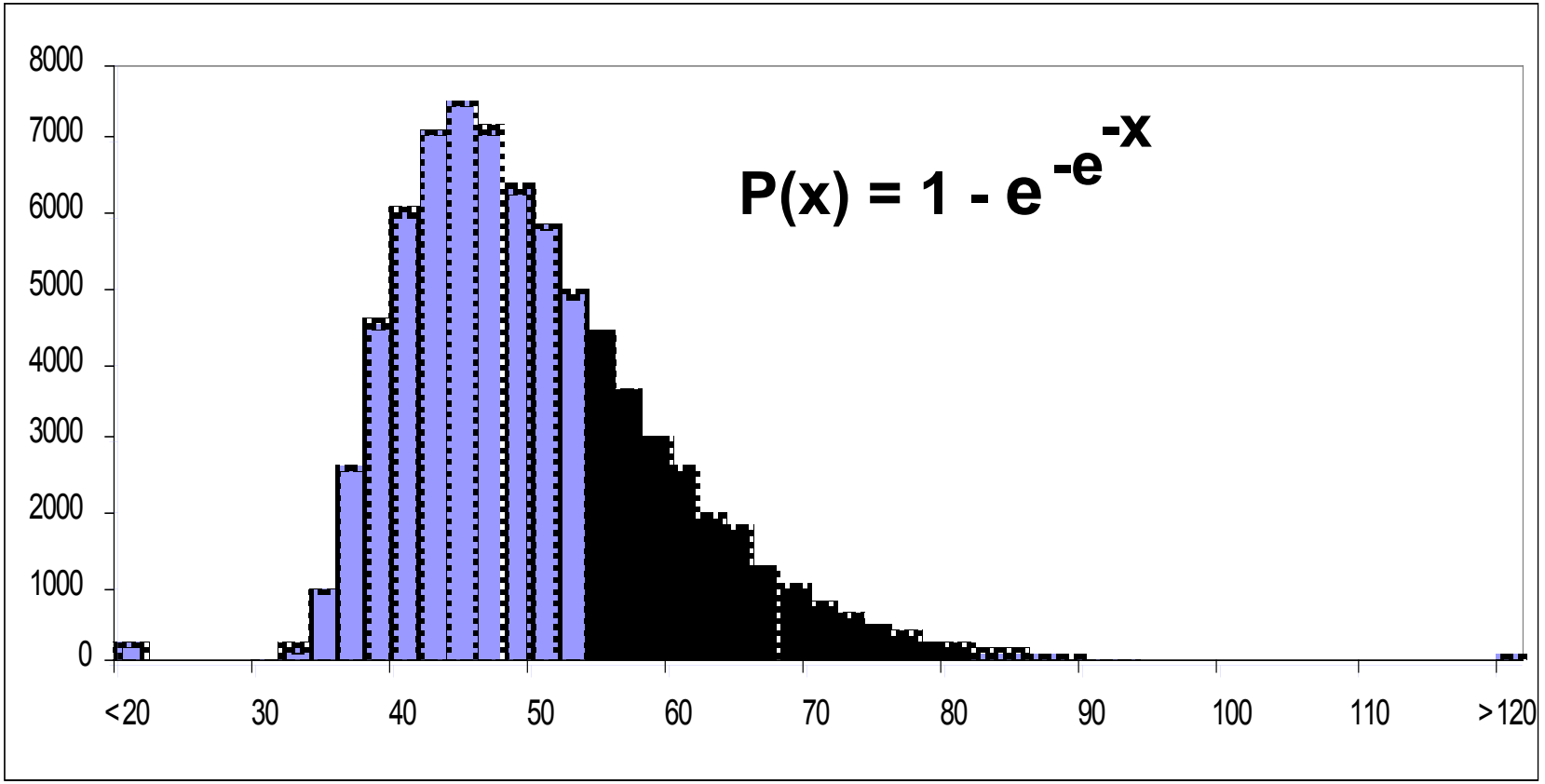
Re-Search All | Search Unmatched

```
1.    PML_HUMAN     Mass: 97455     Score: 194     Expect: 1e-14  Queries matched: 15
      Probable transcription factor PML (Tripartite motif-containing protein 19) (RING finger protein 71)
      MURC_IDILO    Mass: 52994     Score: 51      Expect: 2  Queries matched: 5
      UDP-N-acetylmuramate--L-alanine ligase (EC 6.3.2.8) (UDP-N-acetylmuramoyl-L-alanine synthetase) - I
      DPO1_RICHE    Mass: 104386    Score: 50      Expect: 2.8  Queries matched: 6
```

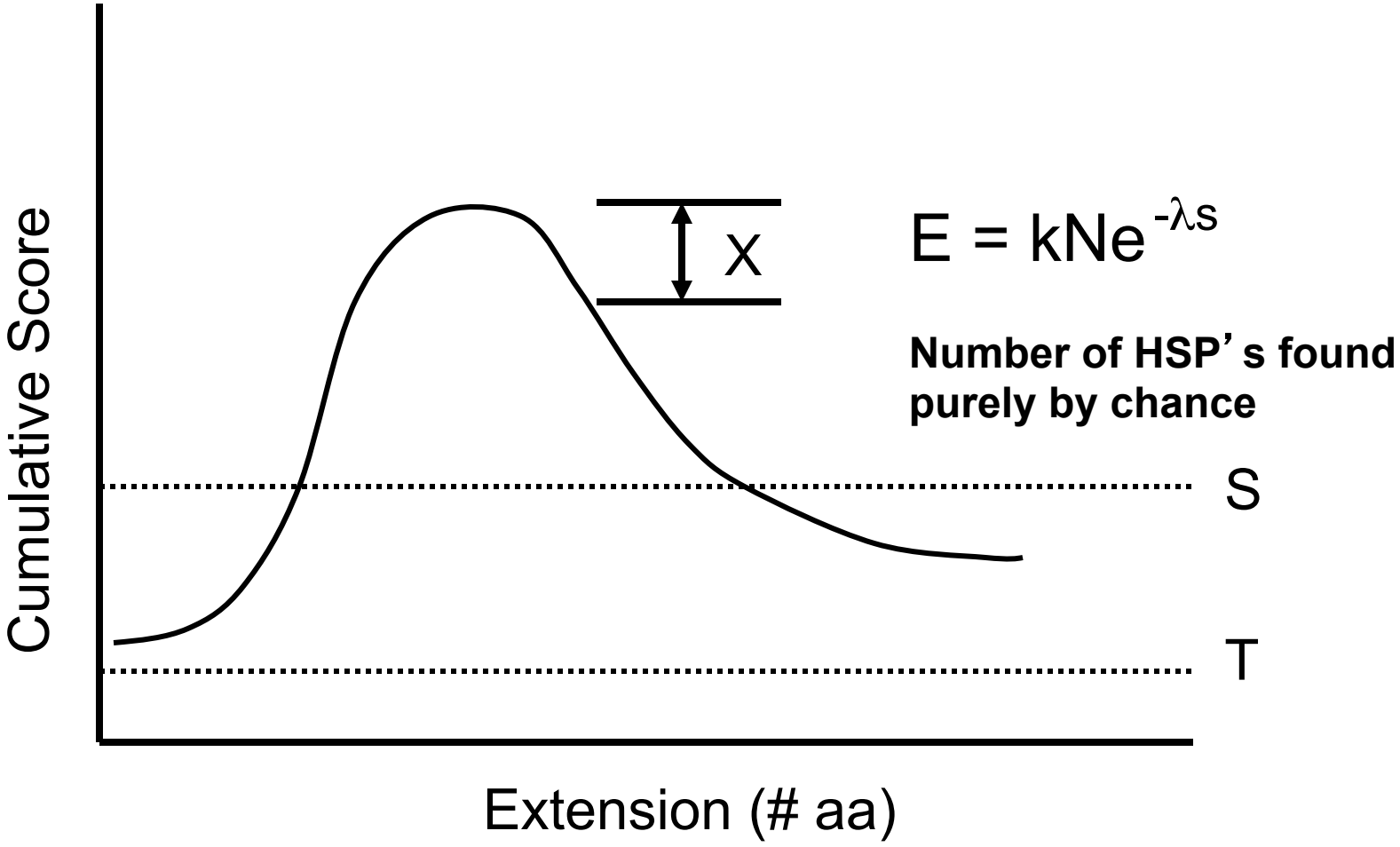
Mascot Scoring*

- 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*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 67 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)*

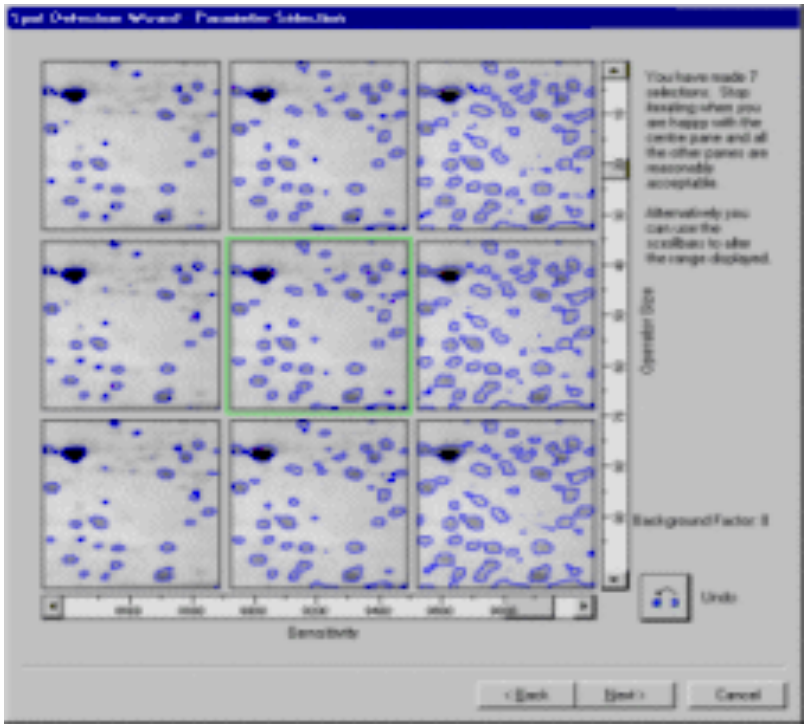
Limitations With PMF*

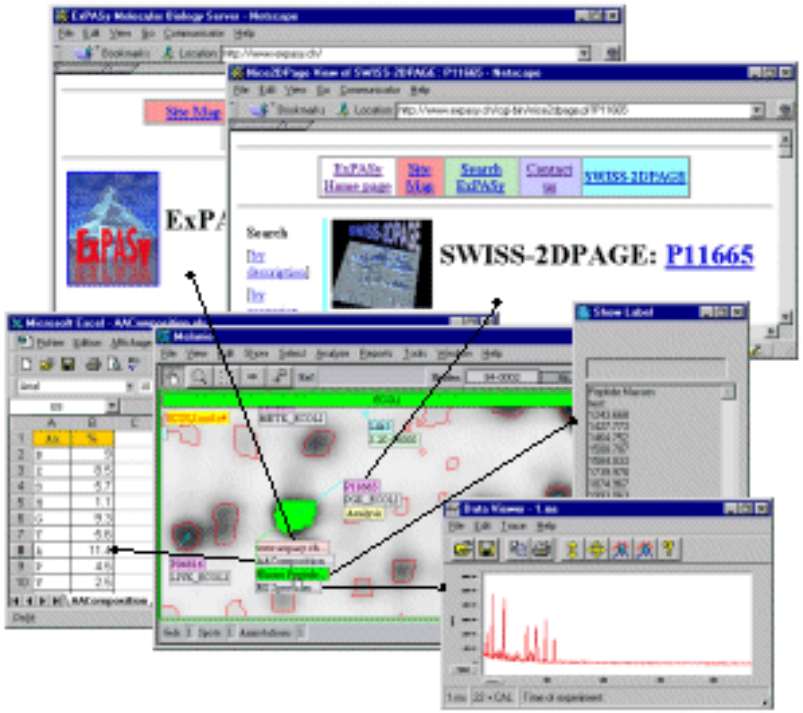
- Requires that the protein of interest already be in a sequence database
- Spurious or missing critical mass peaks always lead to problems
- Mass resolution/accuracy is critical, best to have <20 ppm mass resolution
- Generally found to only be about 40% effective in positively identifying gel spots

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 7 (GeneBio - Windows only)
  - http://world-2dpage.expasy.org/melanie/
- ImageMaster 2D Platinum (GeneBio)
  - http://www.genebio.com/products/melanie/
- Progenesis SameSpots
  - http://www.totallab.com/products/
- PDQuest 7.1 (BioRad - Windows only)
  - http://www.bio-rad.com

Common Software Features*

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

Expressional Proteomics Summary (1)

- 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