Today in HST.480/6.092

Gil Alterovitz

Harvard-MIT
Division of Health
Science & Technology

Announcements

- Homework 2/3 due this Fri 5p
- Projects: In progress

- Today
  - Intro to Proteomics II, Mass spec

Organization: Levels of Abstraction

- Part I: Sequence
- Part II: Expression
- Part III: Proteomics
- Part IV: Systems/Misc.

Computer Science/Algorithms Perspective

- 6.096 - Algorithms for Computational Biology (Spring)- Prof. Manolis Kellis
- This new course covers the algorithmic foundations of computational biology, combining theory with practice. We study the principles of algorithm design for biological datasets, analyze influential algorithms, and apply these to real datasets.
- Topics include:
  - Strings: biological sequence analysis, gene finding, motif discovery, RNA folding, global and local sequence alignment
  - Genomes: genome assembly, comparative genomics, genome duplication, genome rearrangements, evolutionary theory
  - Networks: gene expression, clustering algorithms, scale-free networks, machine learning applications to genomics

Harvard-MIT
Division of Health
Science & Technology

Evolution Perspective

- 6.891 Computational Evolutionary Biology (Fall)- Prof. Robert C. Berwick

- **Course Description**
- Evolution from a computational, modeling, and engineering perspective. Why has it been easier to develop a vaccine to eliminate polio than to control influenza or AIDS? Has there been natural selection for a 'language gene'? Why are there no animals with wheels? When does 'maximizing fitness' lead to evolutionary extinction? How are sex and parasites related? Why don't snakes eat grass? Why don't we have eyes in the back of our heads? How does modern genomics illustrate and challenge the field? Extensive hands-on laboratory exercises in model-building and analyzing evolutionary data.

Harvard-MIT
Division of Health
Science & Technology

HST-Perspective

- **HST.512/HST.513 Genomic Medicine**
- Subject studies the use of industrialized methods of data acquisition and analysis to improve medical care. Questions addressed are: What new benefits of genomics can be anticipated in the near future in terms of new drugs and treatments? How can diagnosis and the diagnostic process be changed today? How do our prognostic abilities change? How does one manage the deluge of clinically relevant genomic data? What constitutes a genomic clinical trial? What are the useful features of alternative genomic technologies today and for the near future? What are the different kinds of genomic informational resources and databases? Are they useful and how? What are the ethical individual and corporate challenges ahead? What are the key limitations we face? Enrollment limited. *I. Kohane, A. Butte, J. Drazen, T. Golub, S. Greenberg, J. Hirschorn, S. Lory, P. Park, M. Ramoni, A. Riva, Z. Szallasi, S. Weiss*

Harvard-MIT
Division of Health
Science & Technology

Mass Spec Lab Techniques

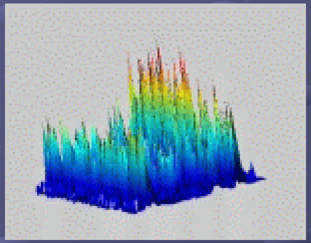
- **Harvard** Chemistry 165. Experimental Physical Chemistry
  Catalog Number: 0667
  Frank N. Keutsch
  Half course (spring term). Lectures: F., 1–2:30; laboratories M., or Tu., 1–5. EXAM GROUP: 6, 7
  Introduction to methods and techniques used in physical chemistry/chemical physics research laboratories. Nine of eleven laboratory assignments involve experiments conducted in current CCB Research Groups: molecular beams; mass spectrometry; Fourier transform infrared and NMR spectroscopies; laser ablation; laser spectroscopy; cavity ring-down spectroscopy; scanning tunneling and atomic force microscopy; kinetics. Computer-based methods of data acquisition and analysis are used throughout.
  Note: Recommended as an efficient preparation for research in experimental physical chemistry/chemical physics and related sciences.
  Prerequisite: Chemistry 160 or Physics 143a.

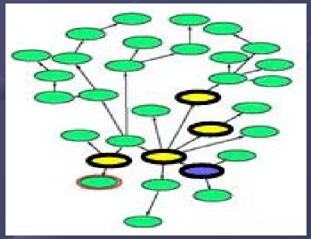
Proteomics/Mass Spec

- **Harvard: BCMP 301 (formerly *Genetics 327). High Throughput Functional Proteomics**
  Catalog Number: 1535
  Edward E. Harlow (Medical School) 2863
  BCMP = Biological Chemistry and Molecular Pharmacology

- **Harvard: Cell Biology 332. Mass Spectrometry and Proteomics**
  Catalog Number: 1568
  Steven P. Gygi (Medical School) 3939

Proteomics and Mass Spec





Gil Alterovitz

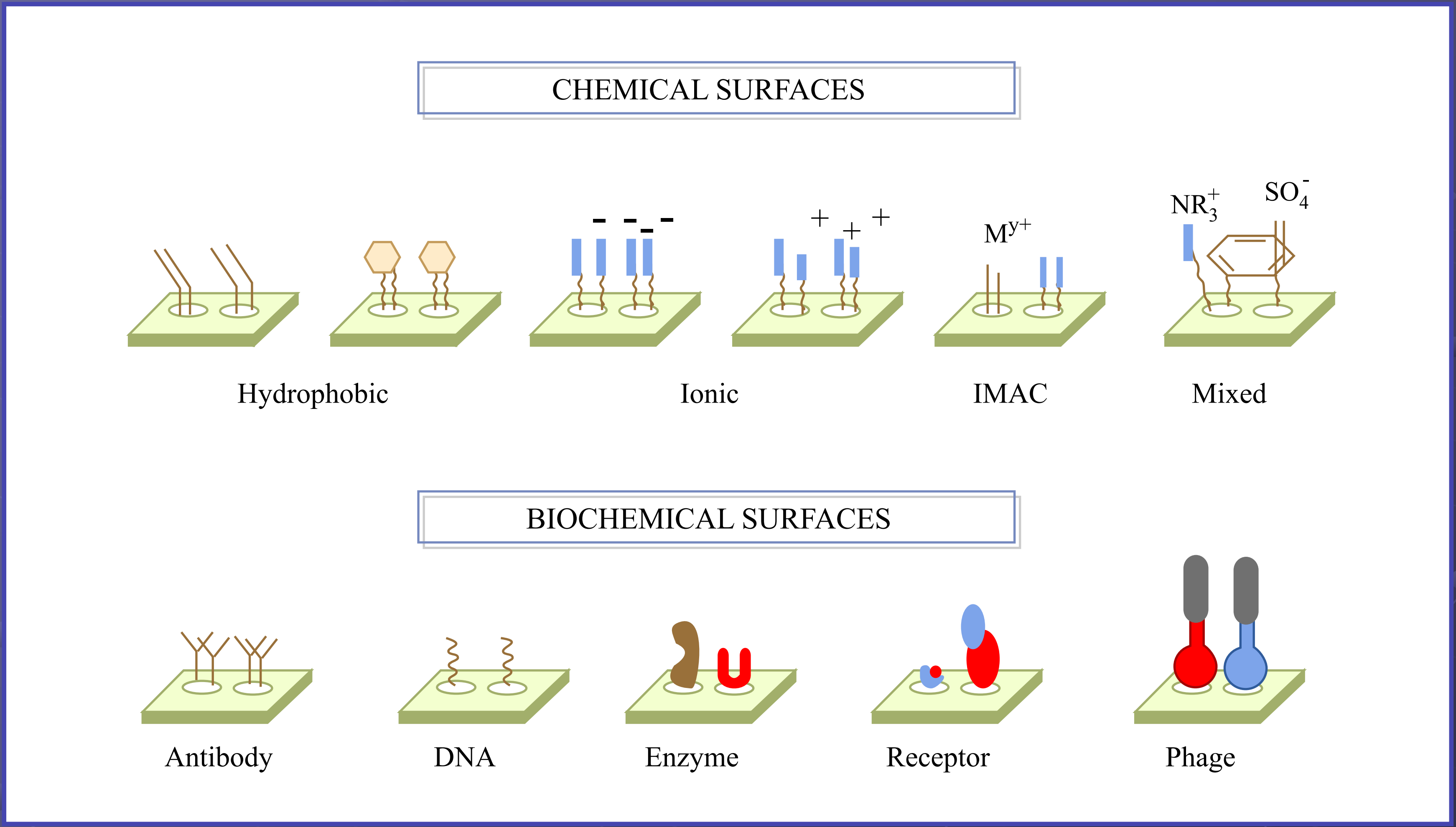
HST.480/6.092

Harvard-MIT
Division of Health
Science & Technology

Paradigm Shifts in Bioinformatics

- **Sequencing** (1980's to early 1990's)
  - DNA/RNA/Protein Sequence Analysis/sequence storage
- **3-D Protein Structure Prediction** (Mid-1980's-late 1990's)
  - Databases of Protein structures
- **DNA/RNA Microarray Expression Experiments** (Mid-1990's to 2000's)
  - Databases of expression data
- **Protein interaction experiments** (Early 2000's to Present)
  - Databases with pairwise interactions
- **Mass Spec proteomic pattern experiments** (Early 2000's to Present)
  - Databases with mass spec, protein identifications, proteomic patterns
- Integration of multiple modalities (Ongoing)

New Flexibility with SELDI-TOF



CHEMICAL SURFACES

Hydrophobic

Ionic

IMAC

Mixed

BIOCHEMICAL SURFACES

Antibody

DNA

Enzyme

Receptor

Phage

Figure by MIT OCW.

Fractionation

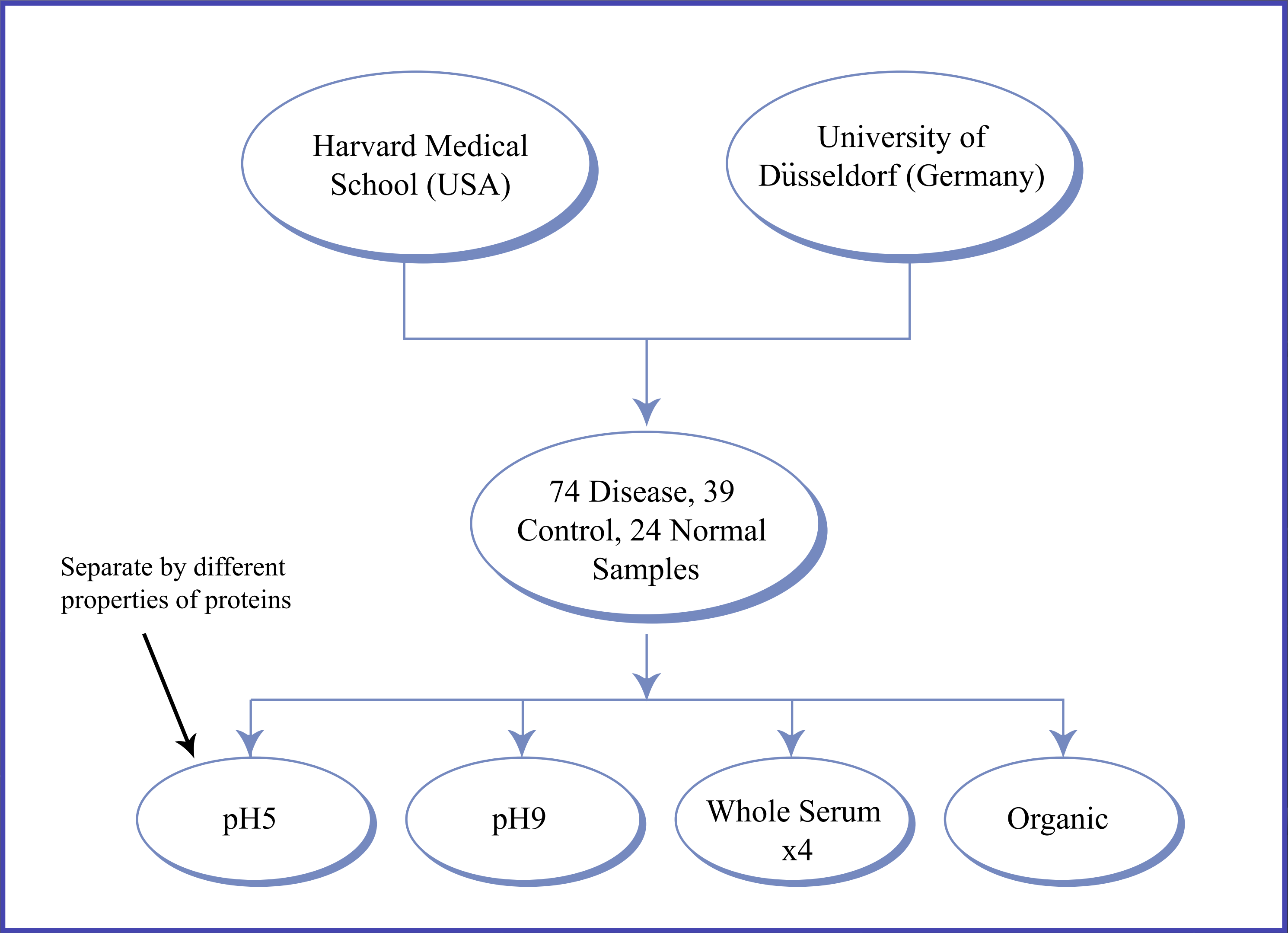
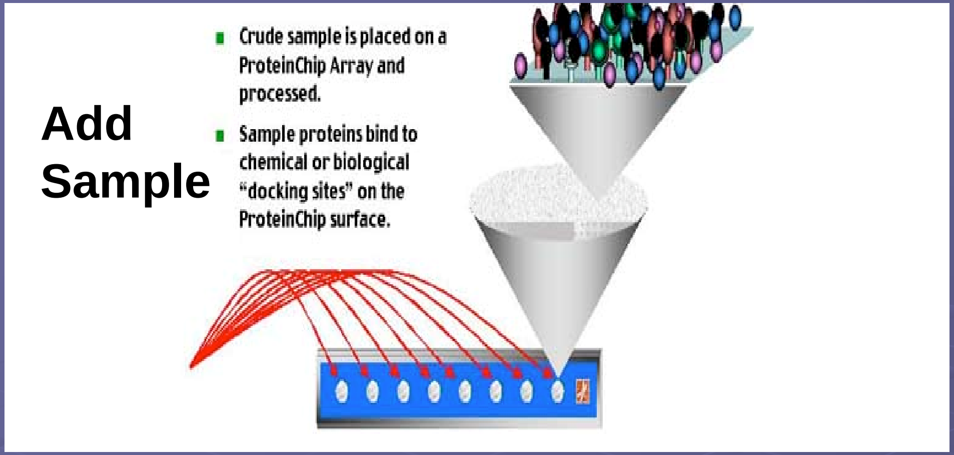

Harvard Medical School (USA)
University of Düsseldorf (Germany)
74 Disease, 39 Control, 24 Normal Samples
Separate by different properties of proteins
pH5
pH9
Whole Serum x4
Organic

Figure by MIT OCW.

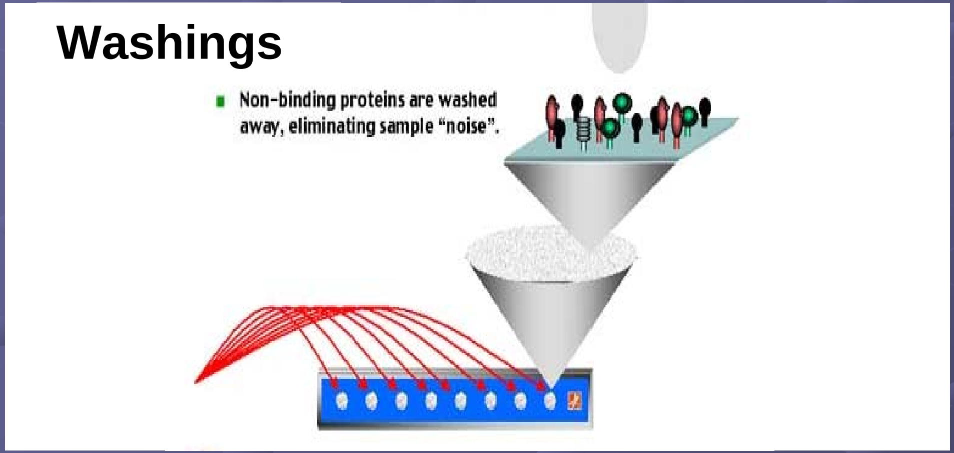
Alterovitz, G., et al., *Analysis and Robot Pipelined Automation for SELDI-TOF Mass Spectrometry.* Proceedings of the International Conference of IEEE Engineering in Medicine and Biology, San Francisco, CA, USA, 2004.

Steps in SELDI-TOF



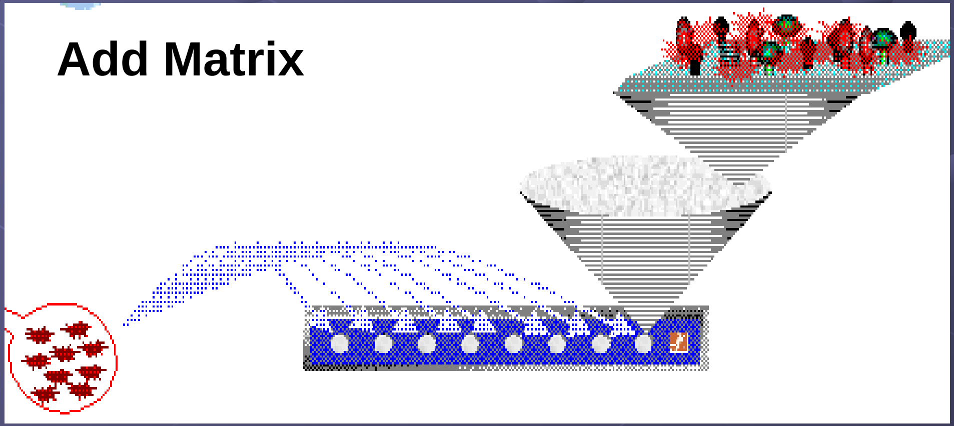
Add Sample

- Crude sample is placed on a ProteinChip Array and processed.
- Sample proteins bind to chemical or biological "docking sites" on the ProteinChip surface.



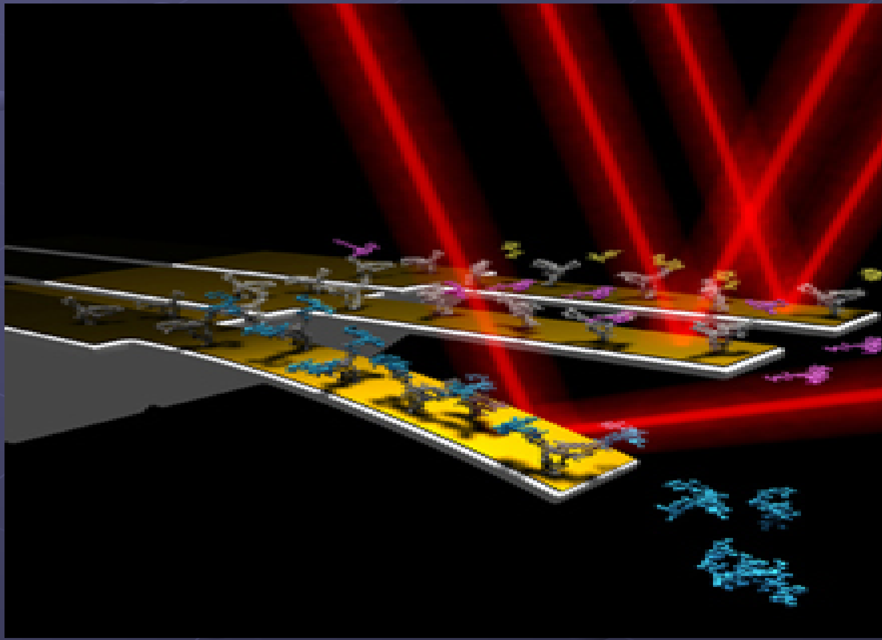
Washings

- Non-binding proteins are washed away, eliminating sample "noise".



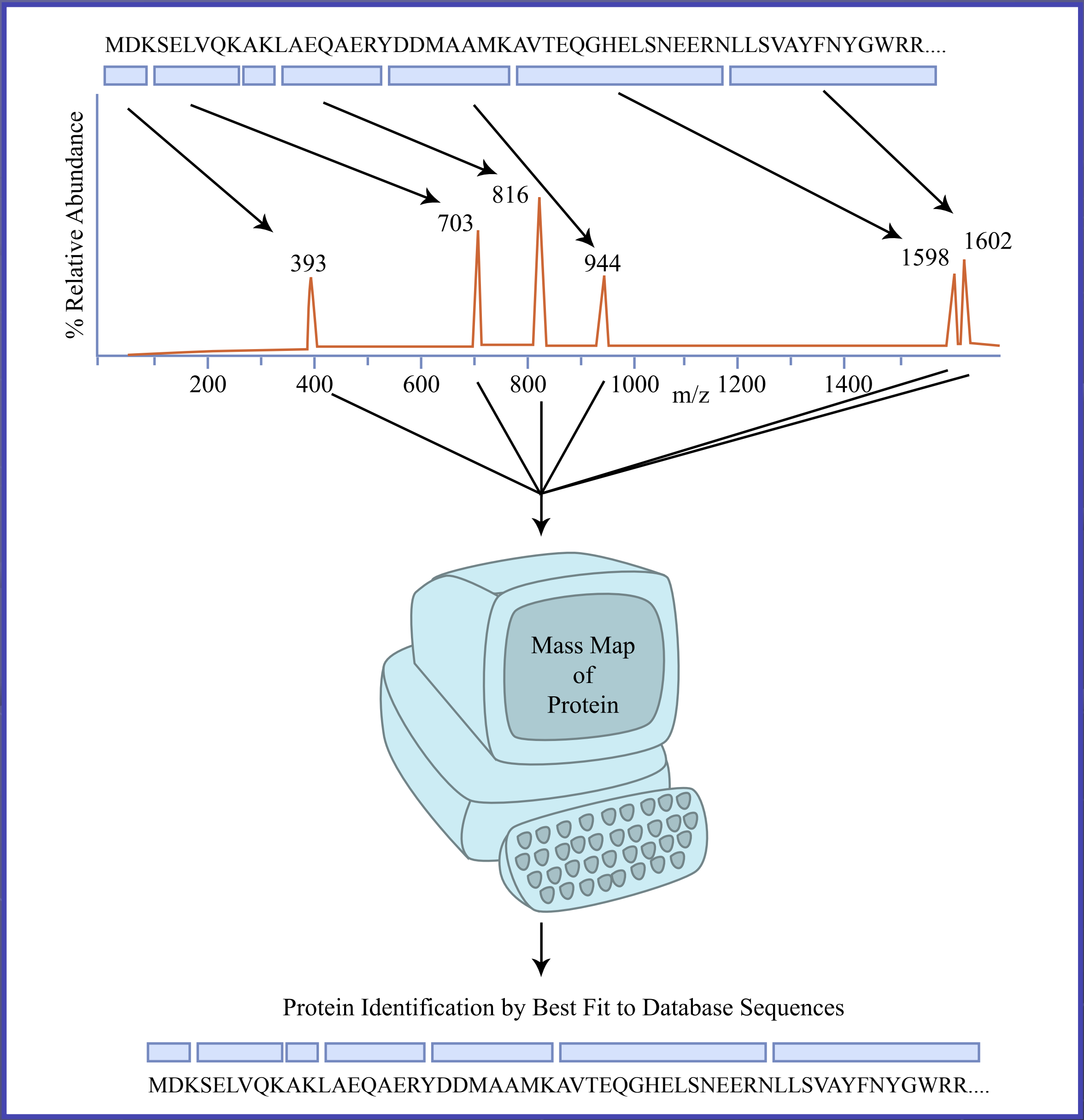
Add Matrix

SELDI-TOF Mass Spec



Harvard-MIT
Division of Health
Science & Technology

Computational Proteomics ≅ Bioinformatics for Genomics

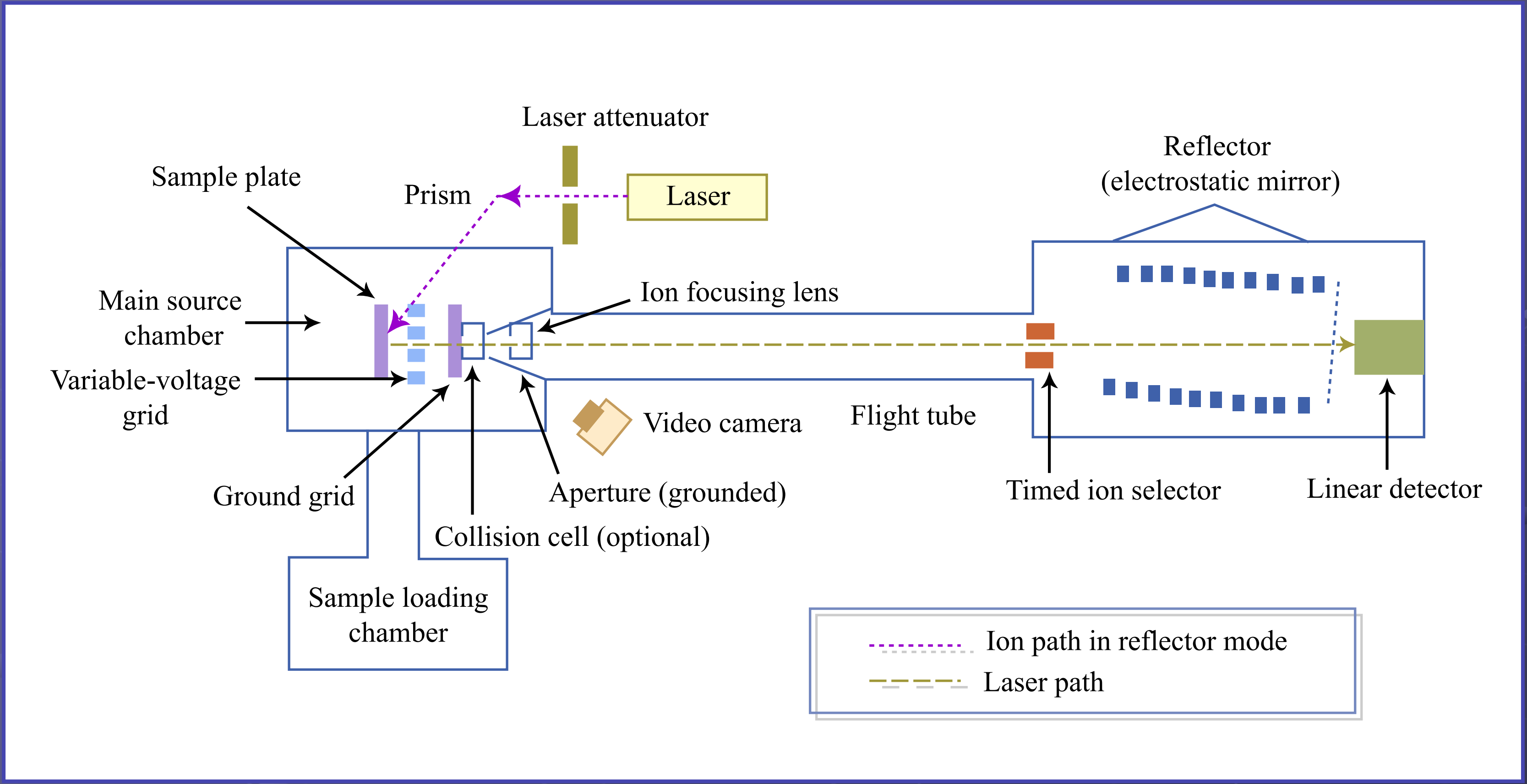


MDKSELVQKAKLAEQAERYDDMAAMKAVTEQGHELSNEERNLLSVAYFNYGWRR....
% Relative Abundance
393
703
816
944
1598
1602
200
400
600
800
1000
m/z
1200
1400
Mass Map of Protein
Protein Identification by Best Fit to Database Sequences
MDKSELVQKAKLAEQAERYDDMAAMKAVTEQGHELSNEERNLLSVAYFNYGWRR....

Figure by MIT OCW.

Harvard-MIT
Division of Health
Science & Technology

SELDI-TOF Mass Spec Schematic



$$\frac{m/z}{U} = a(t - t_0)^2 + b$$

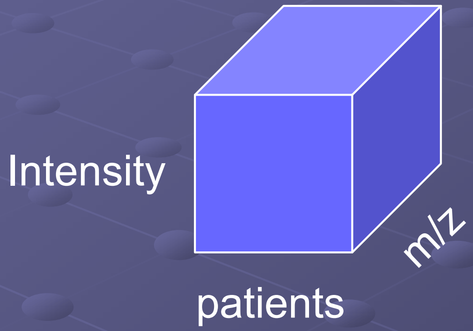
Figure by MIT OCW.

Where:
t = time of flight (μs)
m = mass (Da)
z = charge
U = 20,000 Volts
a = 0.272, b = 0, t_0 = 0.0038 are constants

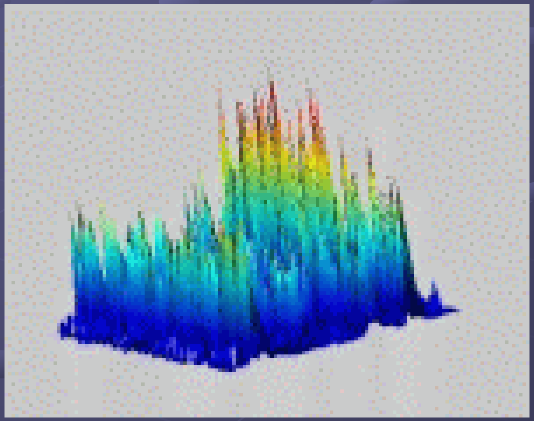
Harvard-MIT
Division of Health
Science & Technology

Data Axes

Axis



3-D Heat Map



Harvard-MIT
Division of Health
Science & Technology

Proteomic Pattern Clustering in N-Space

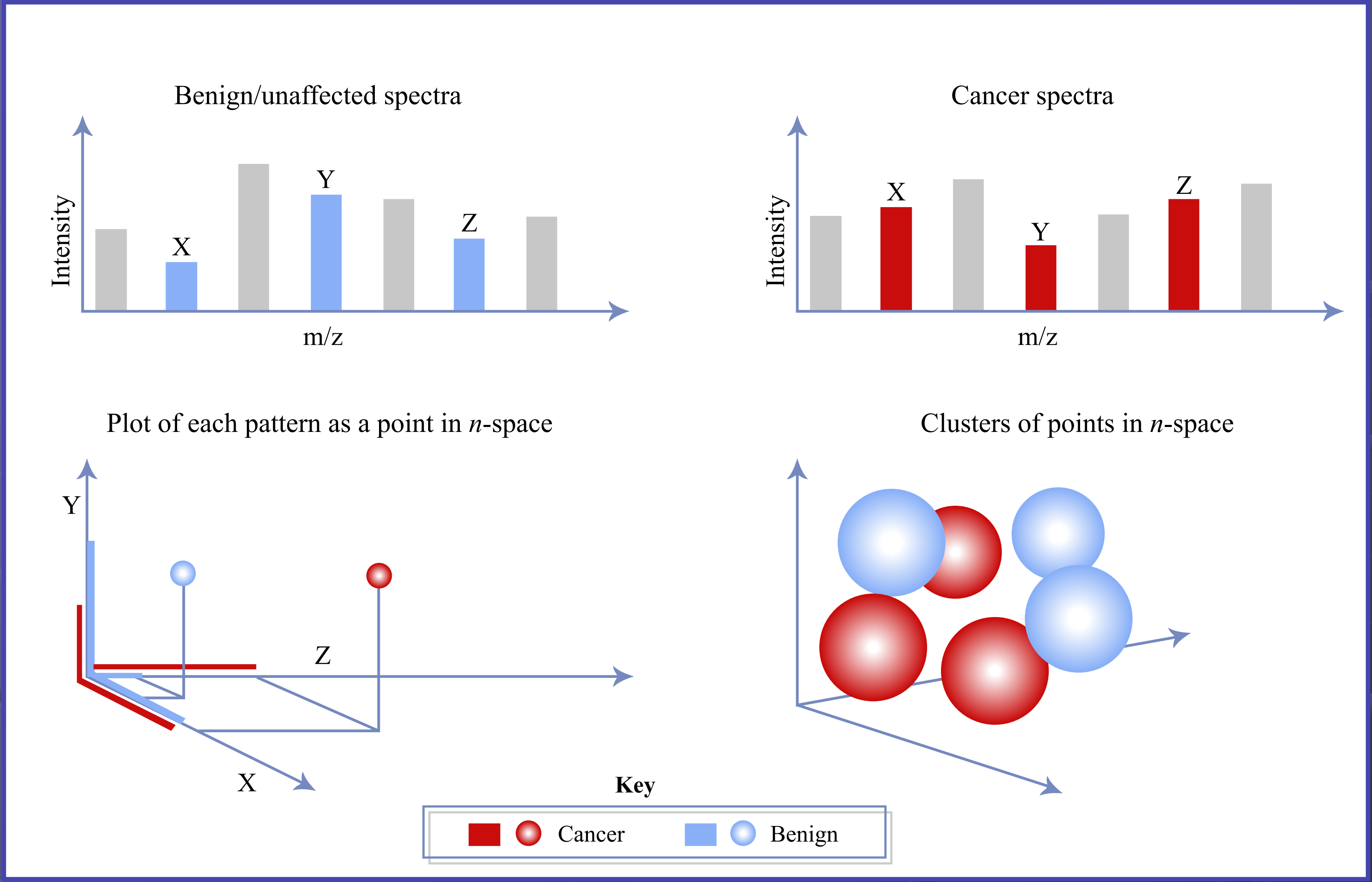

Benign/unaffected spectra
Intensity
X
Y
Z
m/z
Cancer spectra
Intensity
X
Y
Z
m/z
Plot of each pattern as a point in *n*-space
Y
Z
X
Clusters of points in *n*-space
Key
🟥 ● Cancer
🟦 ● Benign

Figure by MIT OCW.

The Challenges: SELDI Issues

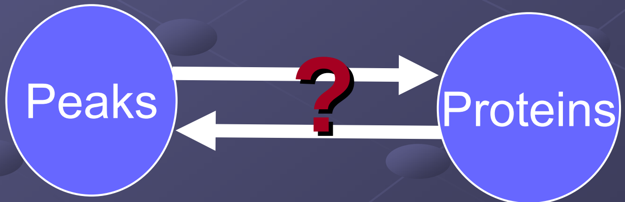
1. Different Operator (reproducibility), repeated measurements by same operator (repeatability) =>Hardware/software automation

2. Not one:one mapping.

   - Many Peaks ➔ 1 Protein (e.g. variability in machine measurement, different charges will appear on different parts of m/z axis). =>Hardware/software automation fractionation, Biological Validation

   - One peak ➔ Many Proteins (e.g. too many proteins with similar mass). => Biological Validation

3. Current models are typically ‘black boxes’: Proteomic profiles rather than protein identifications. Proteins are typically not identified.



Ambiguity in SELDI: A Linguistic Analogy

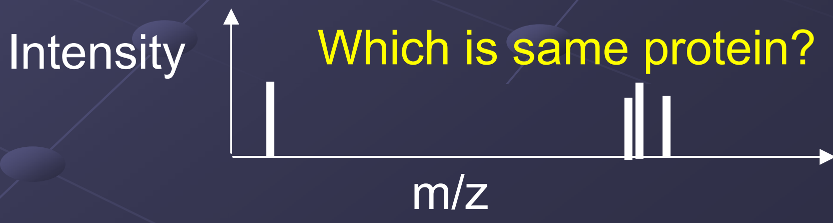
Language Representation: Spelling (& Pronunciation)

‘Meaning’

	= Homograph (Homophone) words spelled (& pronounced) alike (i.e. masses within +/- machine variability)	≠ Heterograph (Heterophone) words spelled (& pronounced) differently
= Synonym	Same word ✓	Synonym
≠	Homonym	Different words/meanings ✓

Key:

Language Representation = SELDI ‘Mass,’ m/z
Word meaning=‘protein’



Which is same protein?

Analyzing States and Control in Proteomics

Chameleon version:

1. Researcher wants to test hearing in chameleons.
2. But, how to get chameleon to respond?
3. Researcher remembers reading that 'A chameleon darker color can be a sign of distress or anger.'
4. Researcher's experiment: Test hearing of chameleon by provoking it (loud noises) -> color change to signal that animal can hear*.
5. Researcher add chemical that specifically kills chromatophore *cells**
6. Result: Chameleon does not respond to loud noises. Extrapolating to humans, researcher writes paper: 'Human Audition Potentially Mediated by Chromatophore Cells'
7. Our conclusion- we need to look at more than just analysis differential. We need to look at flow of control

Photograph courtesy of
Alastair de Wet and stock.xchng

Light Green ☺ <-> Brown ☹
Normal <-> 'Disease State'

Photograph courtesy of Christian Burger and stock.xchng

* Skin color changes are initiated by moving small, black granules (melanosomes) from
In *chromatophore cells*.

Key:
States (e.g. cancer vs. normal)= color vs no color change.
Control = hearing controls hormones->chromatophore cell receptors->release granules (melanosomes)

Quantifying Automation Reduction in Peak Intensity Variation

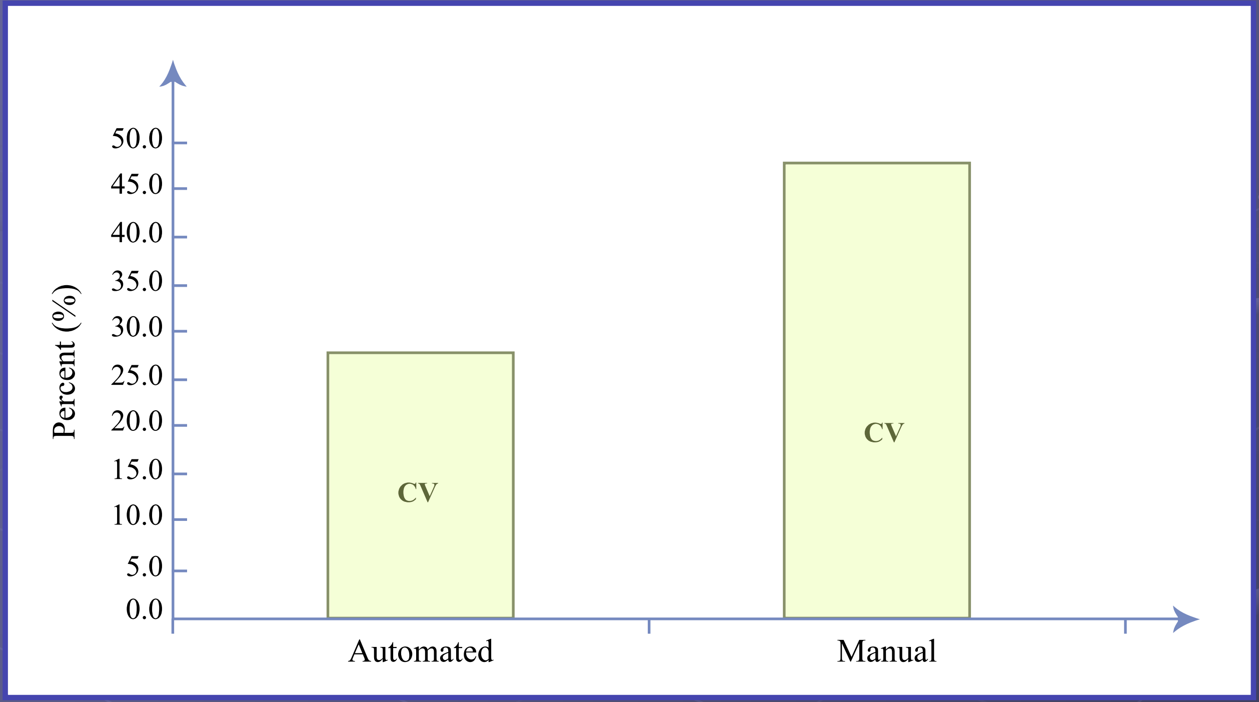


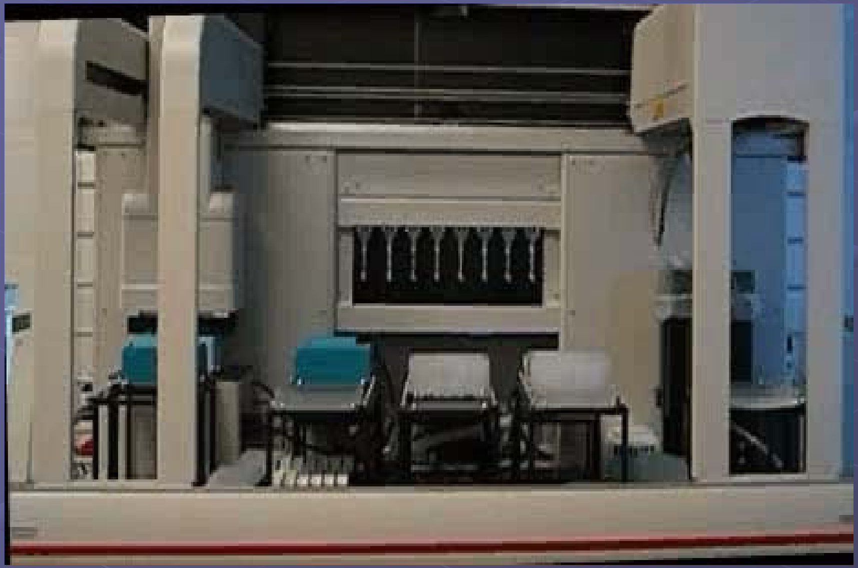
Figure by MIT OCW.

- One hundred (20 manual, 80 automated) biological samples done with replicates (2x). Coefficient of Variation (CV) is 27.8% for automation vs. 45.1% for manual.
- Statistically significant with *P* <0.001

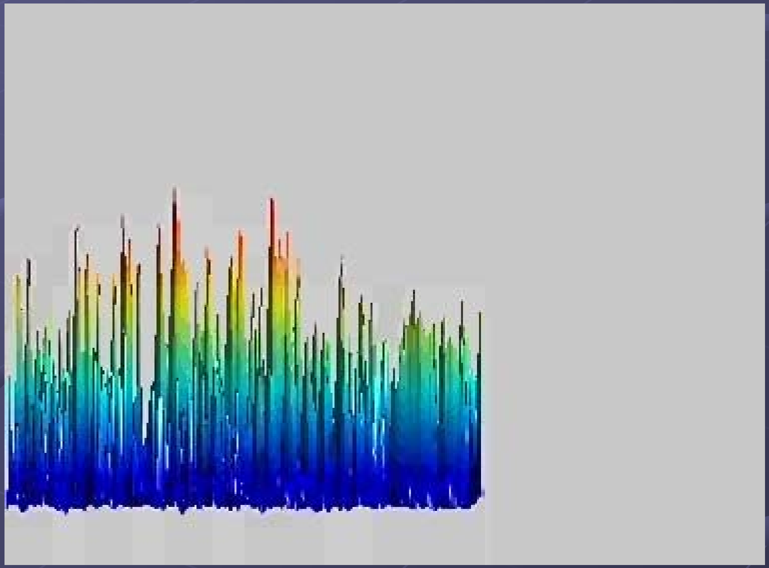
CV = σ **/** μ (standard deviation/mean)

Alterovitz, G., et al., *Analysis and Robot Pipelined Automation for SELDI-TOF Mass Spectrometry.* Proceedings of the International Conference of IEEE Engineering in Medicine and Biology, San Francisco, CA, USA, 2004.

Hardware and Analysis Automation Components



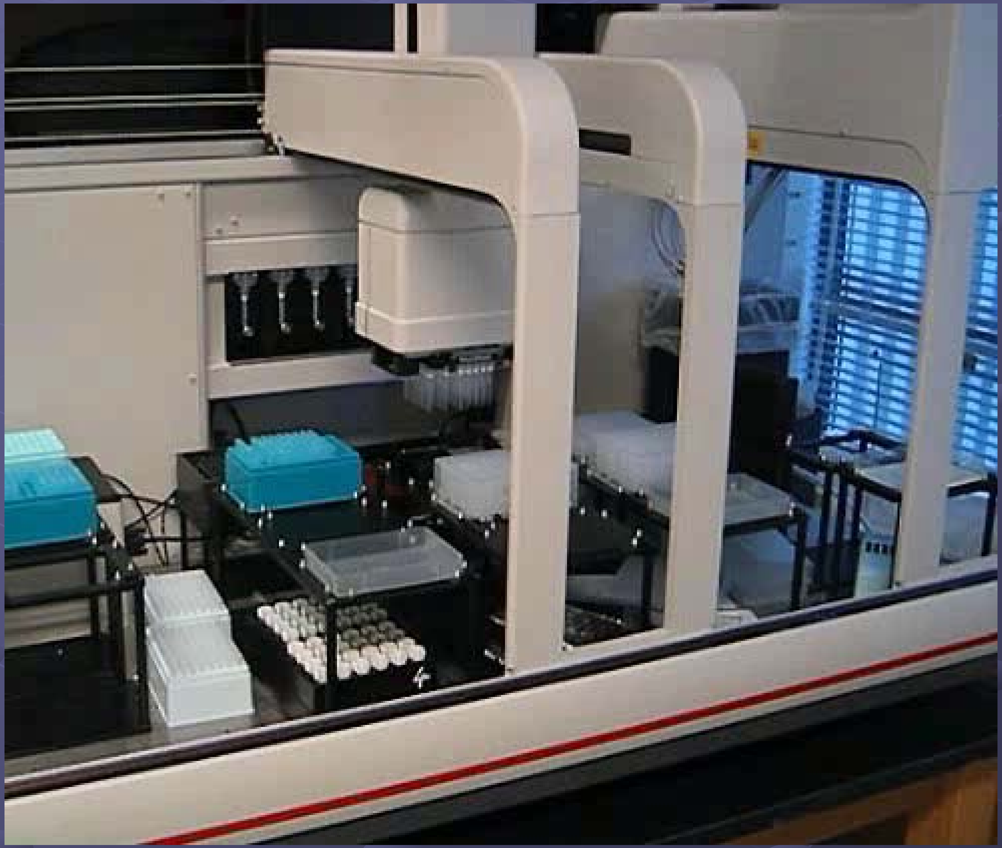
Robotic Automation



Analysis Pipeline

Harvard-MIT
Division of Health
Science & Technology

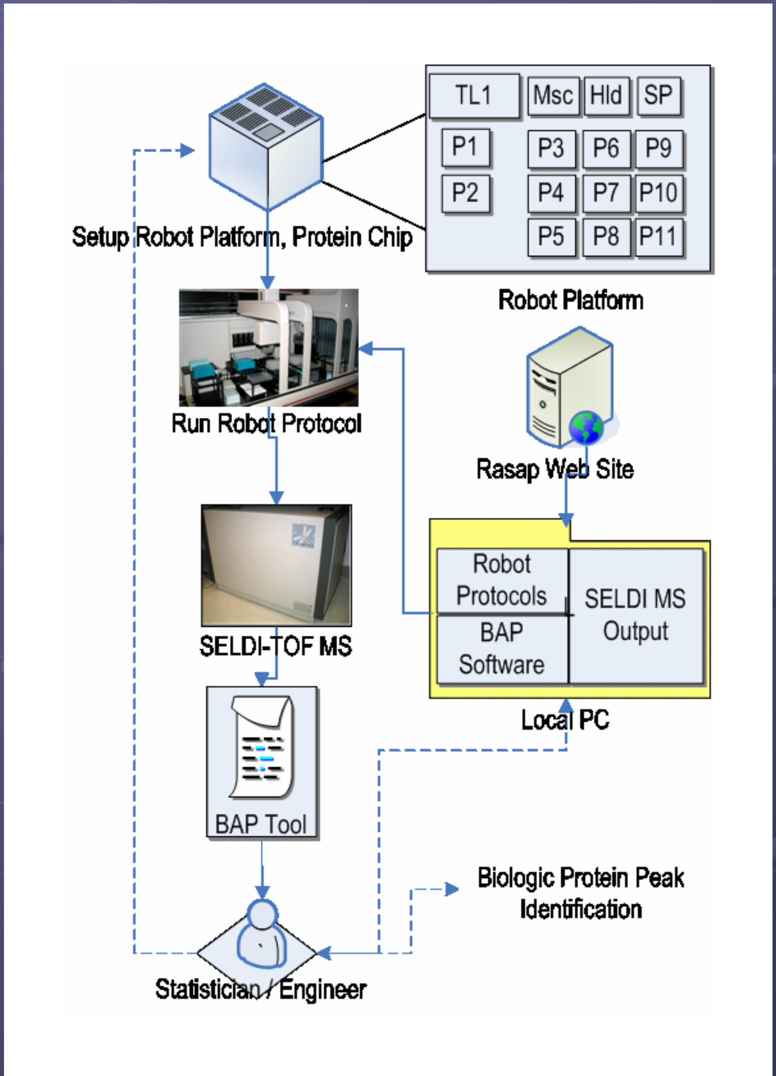
Robot Sample Preparation



Harvard-MIT
Division of Health
Science & Technology

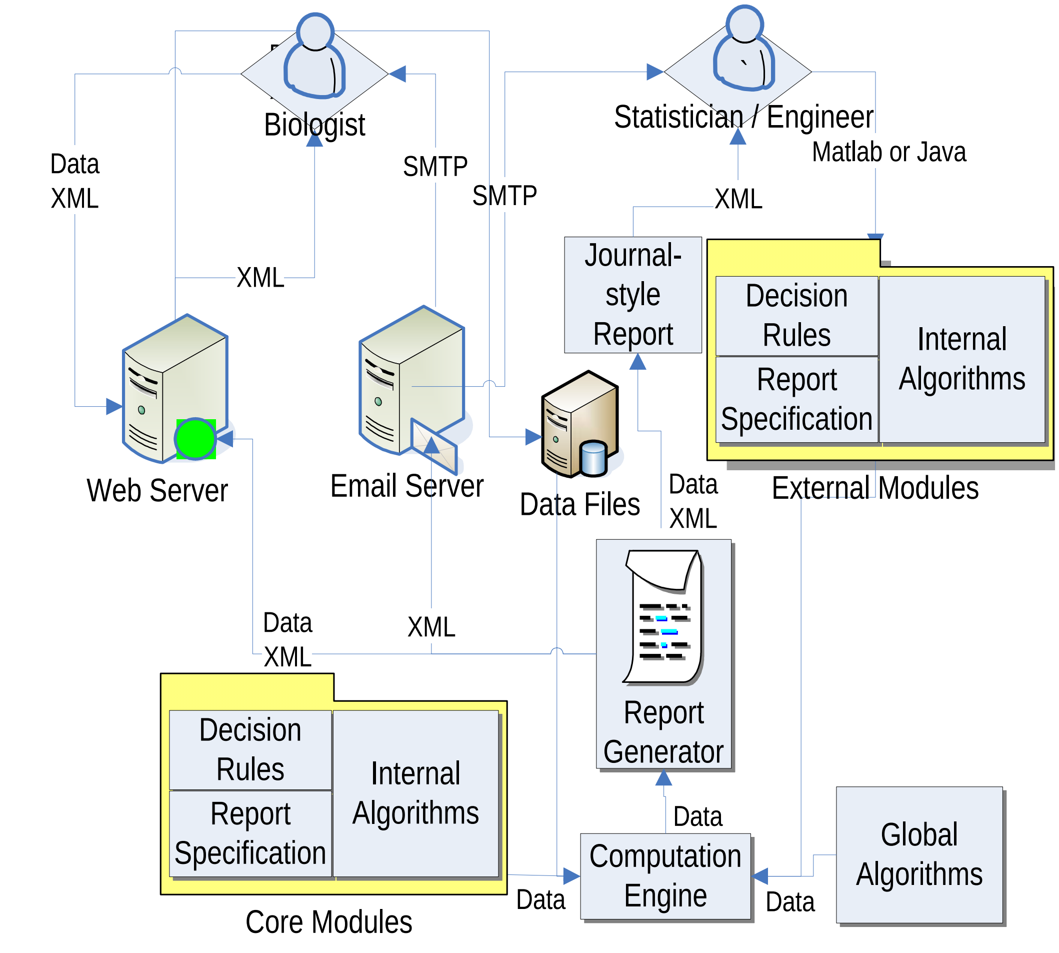
SELDI MS-TOF: Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry

Raspap: Robot Automated Sample Preparation and Analysis Pipeline for Proteomics



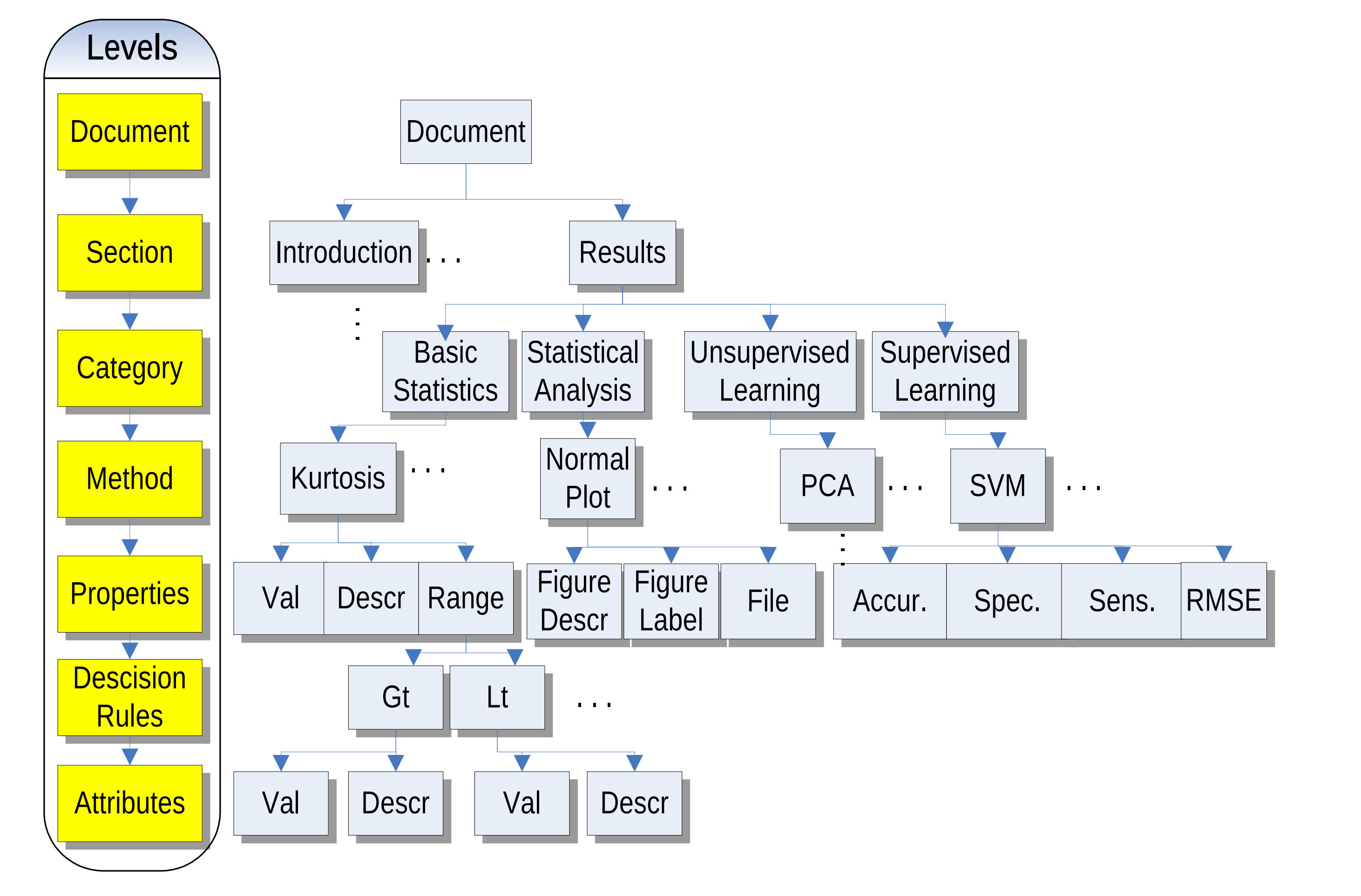
Harvard-MIT Division of Health Science & Technology

BAP: Bioinformatics Automated Pipeline



Harvard-MIT
Division of Health
Science & Technology

Object-Oriented Tree Structure of BAP



Harvard-MIT
Division of Health
Science & Technology

Machine Learning Results

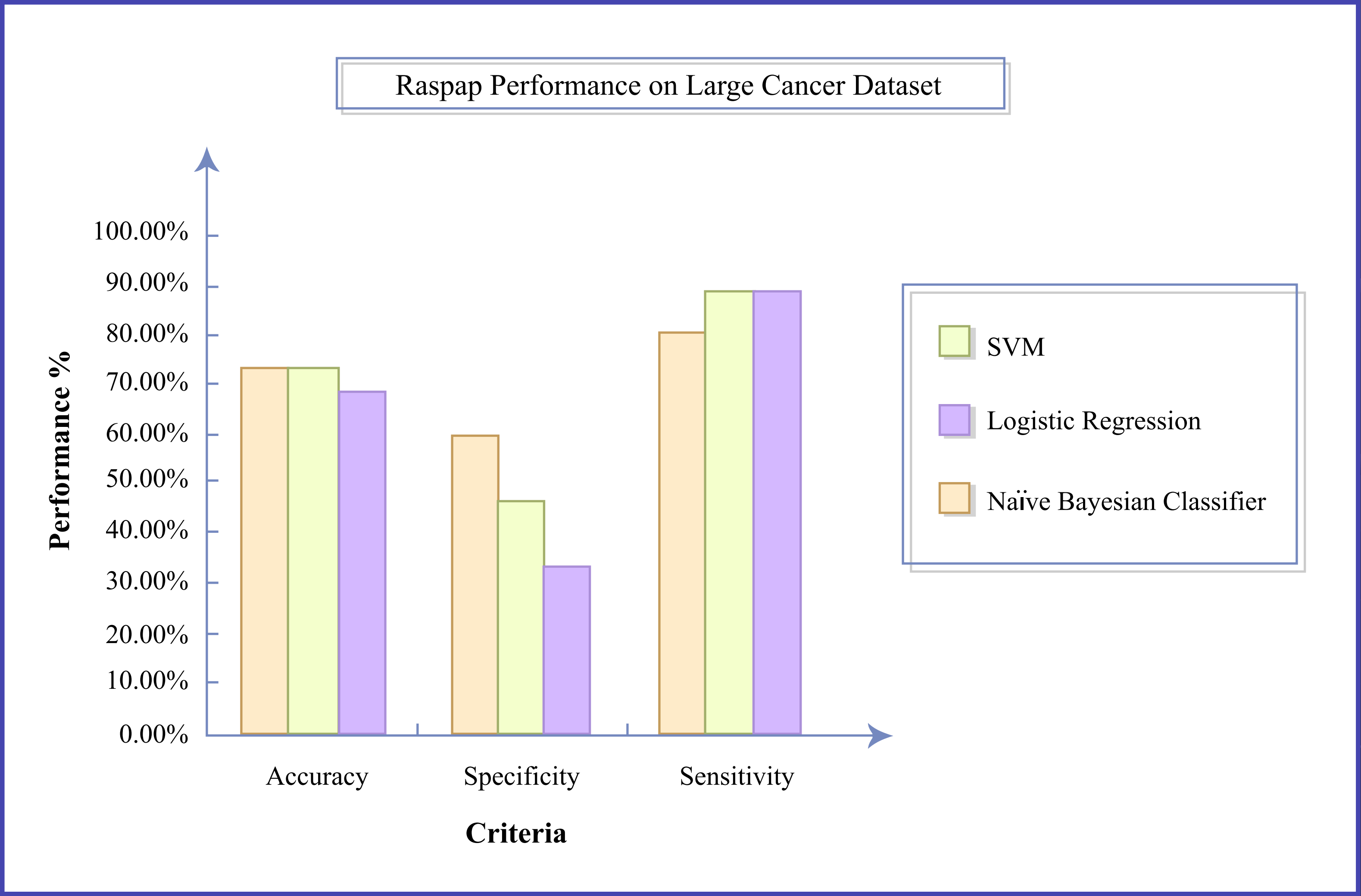
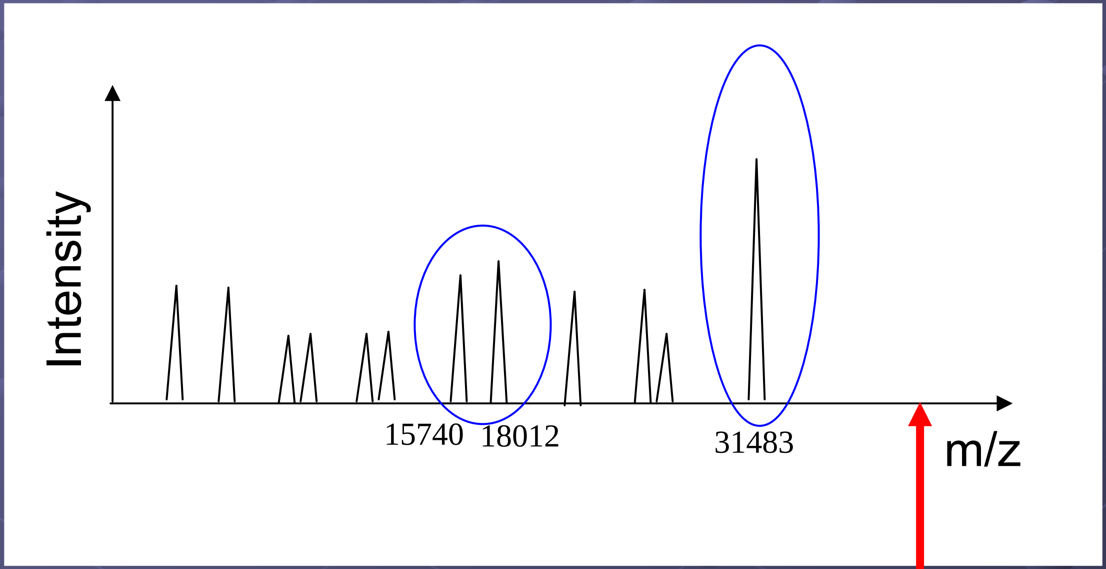


Figure by MIT OCW.

Alterovitz, G., et al., *Analysis and Robot Pipelined Automation for SELDI-TOF Mass Spectrometry.* Proceedings of the International Conference of IEEE Engineering in Medicine and Biology, San Francisco, CA, USA, 2004.

Example



Maximum range: 35000

Does 18,012 have a single charge?
What about 15,740?

Harvard-MIT
Division of Health
Science & Technology

Biological Protein Peak Identification

Image removed due to copyright considerations

Gel image

Tandem Mass Spec

- Take advantage of high sensitivity at low peptide resolution (without a matrix that can add irreproducible ‘noise’ in that region).
- Use this to sequence small cut bits of proteins (puzzle pieces)
- Compare cleaved proteins sequences with database to identify the protein in the sample (complete puzzle).
  - Via cross-correlation of spectra with hypothesized spectra of database entries
- Yields: protein identification and abundance (via peak area/intensity.

Challenges

- Protein may not be in database
- Cleaved protein may match several database entries

Amino Acid	Symbol	Average molecular weight (Da)
Alanine	A	71.0788
Arginine	R	156.1876
Asparagine	N	114.1039
Aspartic Acid	D	115.0886
Cysteine	C	103.1448
Glutamine	Q	128.1308
Glutamic Acid	E	129.1155
Glycine	G	57.0520
Histidine	H	137.1412
Isoleucine	I	113.1595
Leucine	L	113.1595
Lysine	K	128.1742
Methionine	M	131.1986
Phenylalanine	F	147.1766
Proline	P	97.1167
Serine	S	87.0782
Threonine	T	101.1051
Tryptophan	W	186.2133
Tyrosine	Y	163.1760
Valine	V	99.1326

Tandem MS/MS with HPLC

- In this approach, the proteins in a sample are first digested (cleaved into smaller peptides) using a protease such as trypsin.
- Trypsin cuts proteins on the carboxyl side of positively charged amino acid residues (e.g. lysine and arginine).

Image removed due to copyright considerations

Trysin 3-D Structure

High Performance Liquid Chromatography

- The chromatography involves a separation based on attributes such as:
  - Hydrophobicity: lacking attraction to water
  - Strong cation exchange: net positive charge
  - Strong anion exchange: net negative charge
  - Size separation: size/molecular weight
  - Special affinity: interaction with particular functional groups

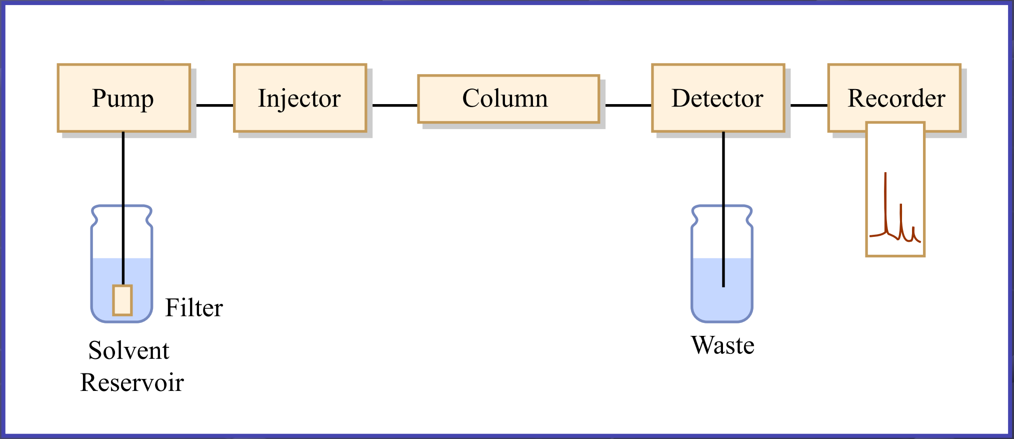


Figure by MIT OCW.

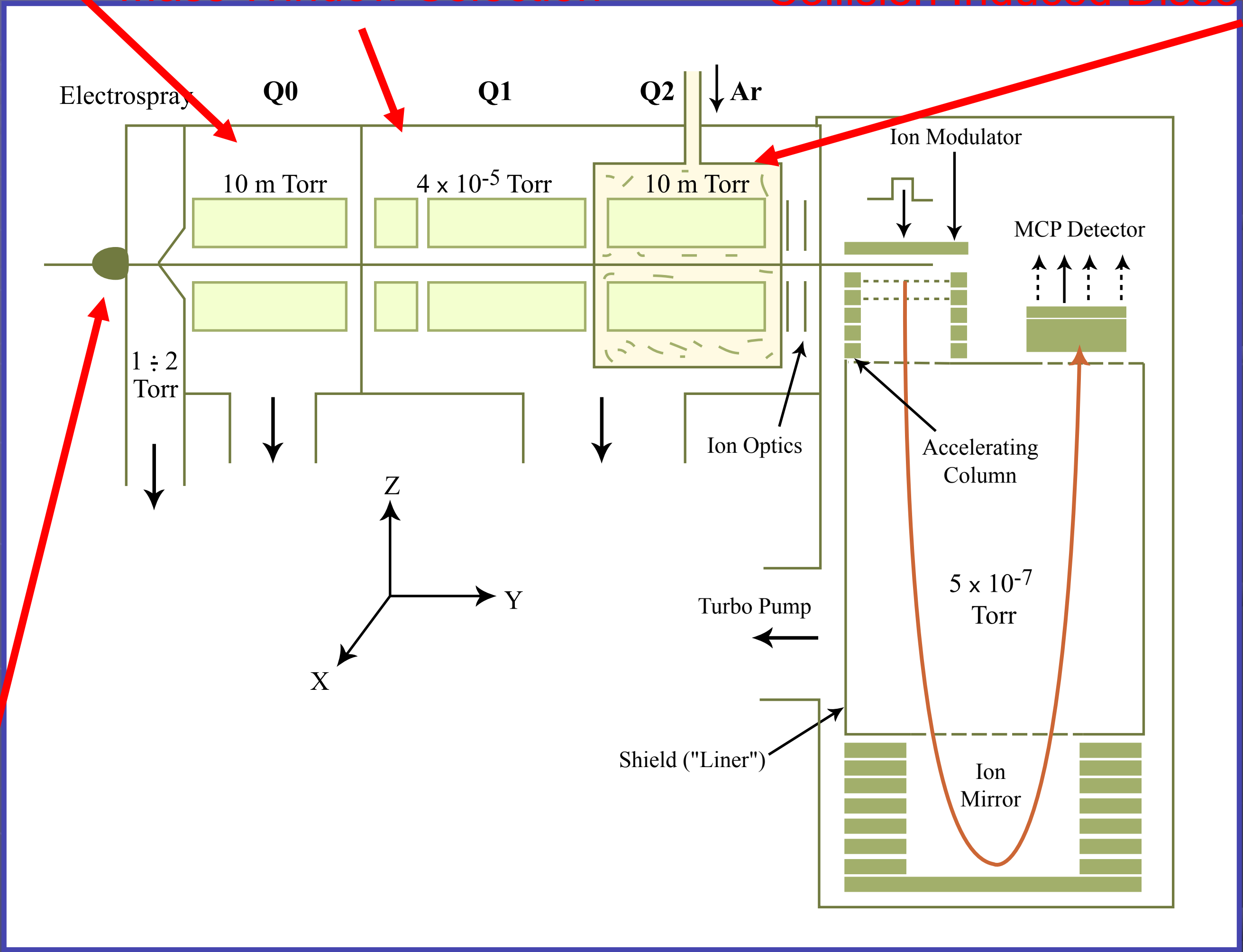
Harvard-MIT
Division of Health
Science & Technology

Schematic of Tandem QqTOF (quadrupole-time-of-flight) Mass Spectrometer

Pass-through (ion guide)

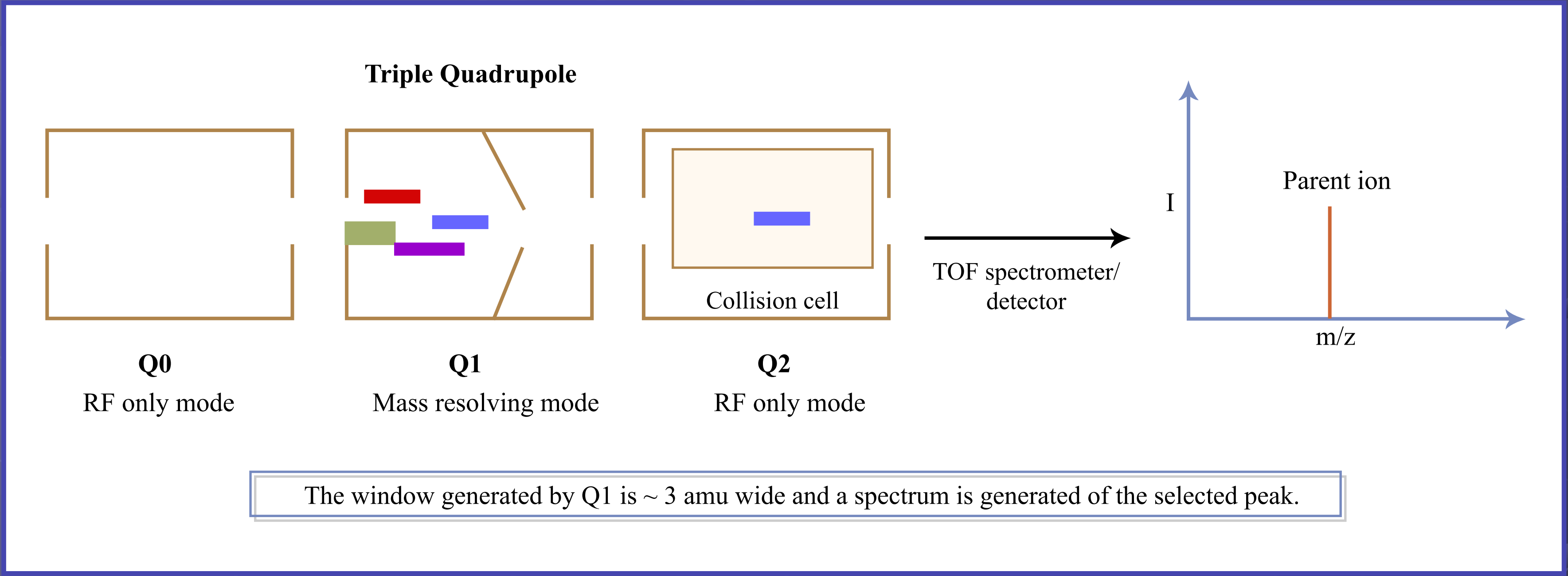
Mass Window Selection

Collision Induced Dissociation (CID)



Highly charged ions formed (>> 3 seen in SELDI)
(analyte solution pushed through needle into electric field->ionized droplets

Figure by MIT OCW.

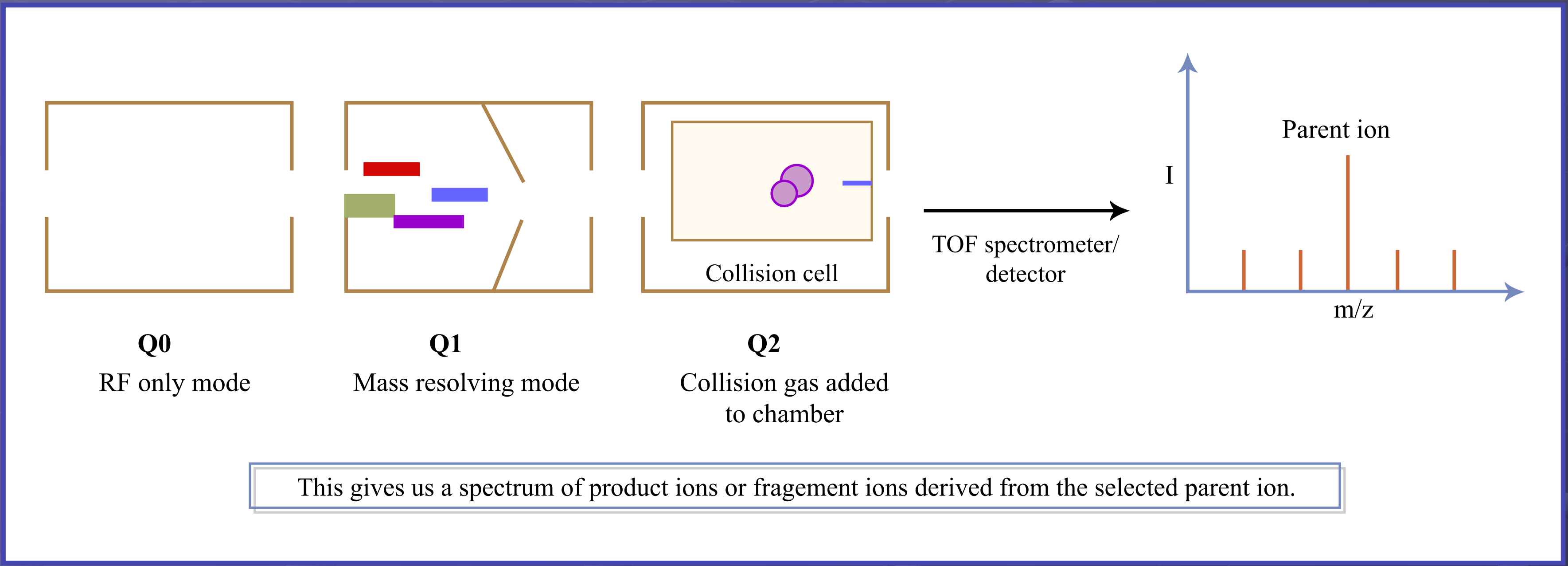


The window generated by Q1 is ~ 3 amu wide and a spectrum is generated of the selected peak.

Figure by MIT OCW.

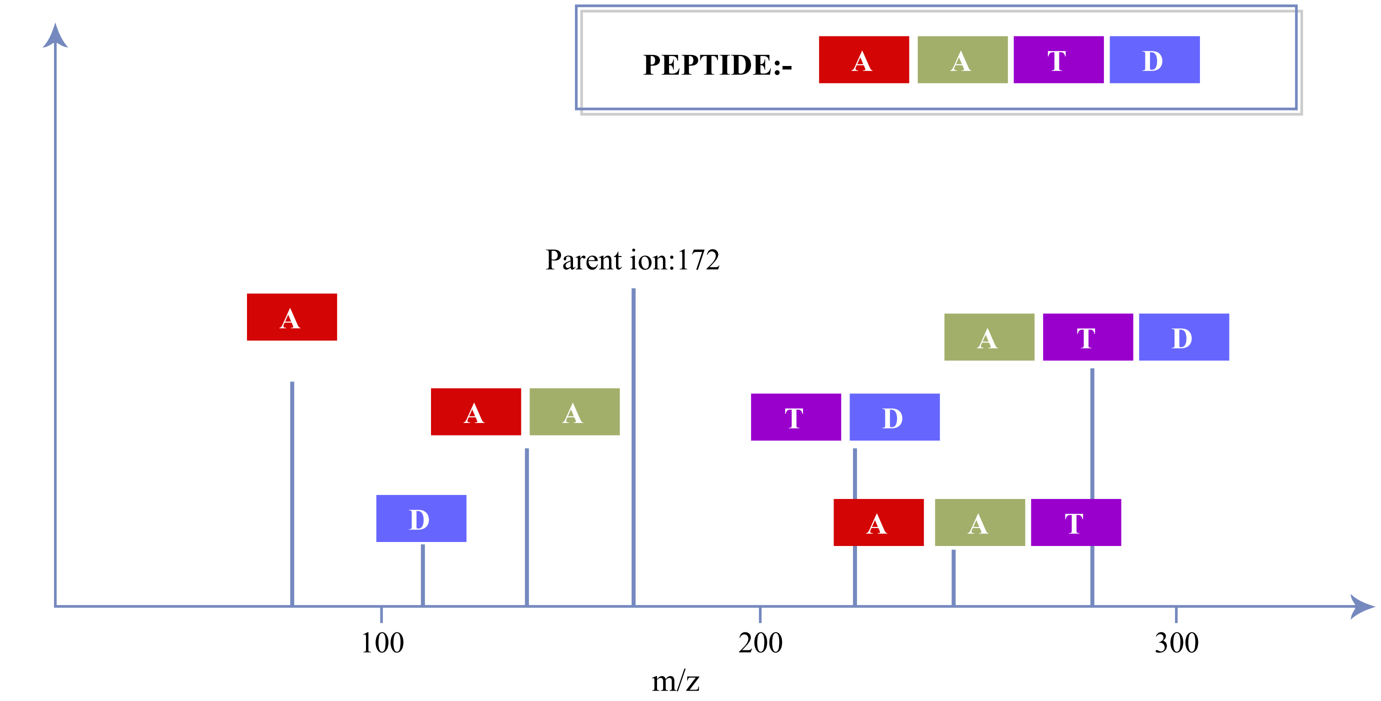
Source: Samuel Lunenfeld Research Center

Tandem Mass Spec



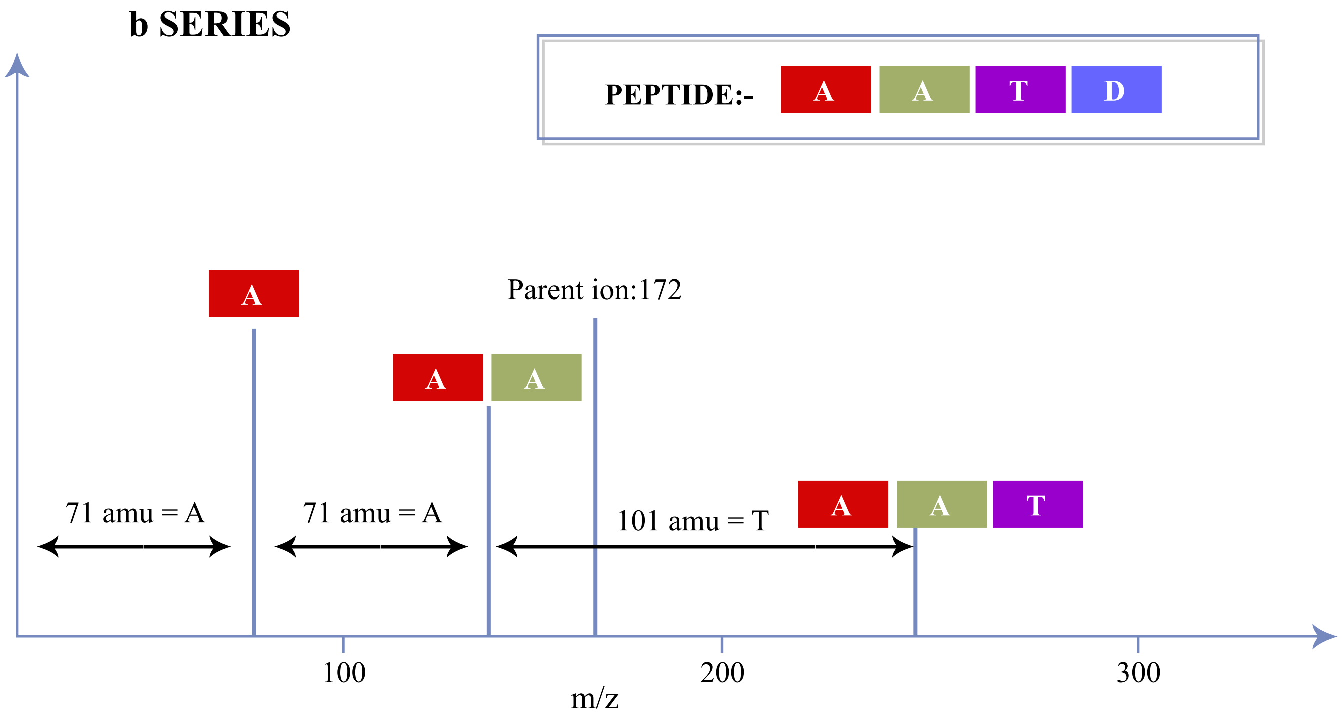
This gives us a spectrum of product ions or fragement ions derived from the selected parent ion.

Figure by MIT OCW.



Fragmentation between the C and the amino Ns are often used for sequencing. The N terminal portions of these fragmentations are referred to as the b-series, the C-terminal portions are the y-series.

Figure by MIT OCW.



If we focus on the b-series ion peaks, we can see how each pair of peaks is separated by the mass of 1 particular amino acid.

Figure by MIT OCW.

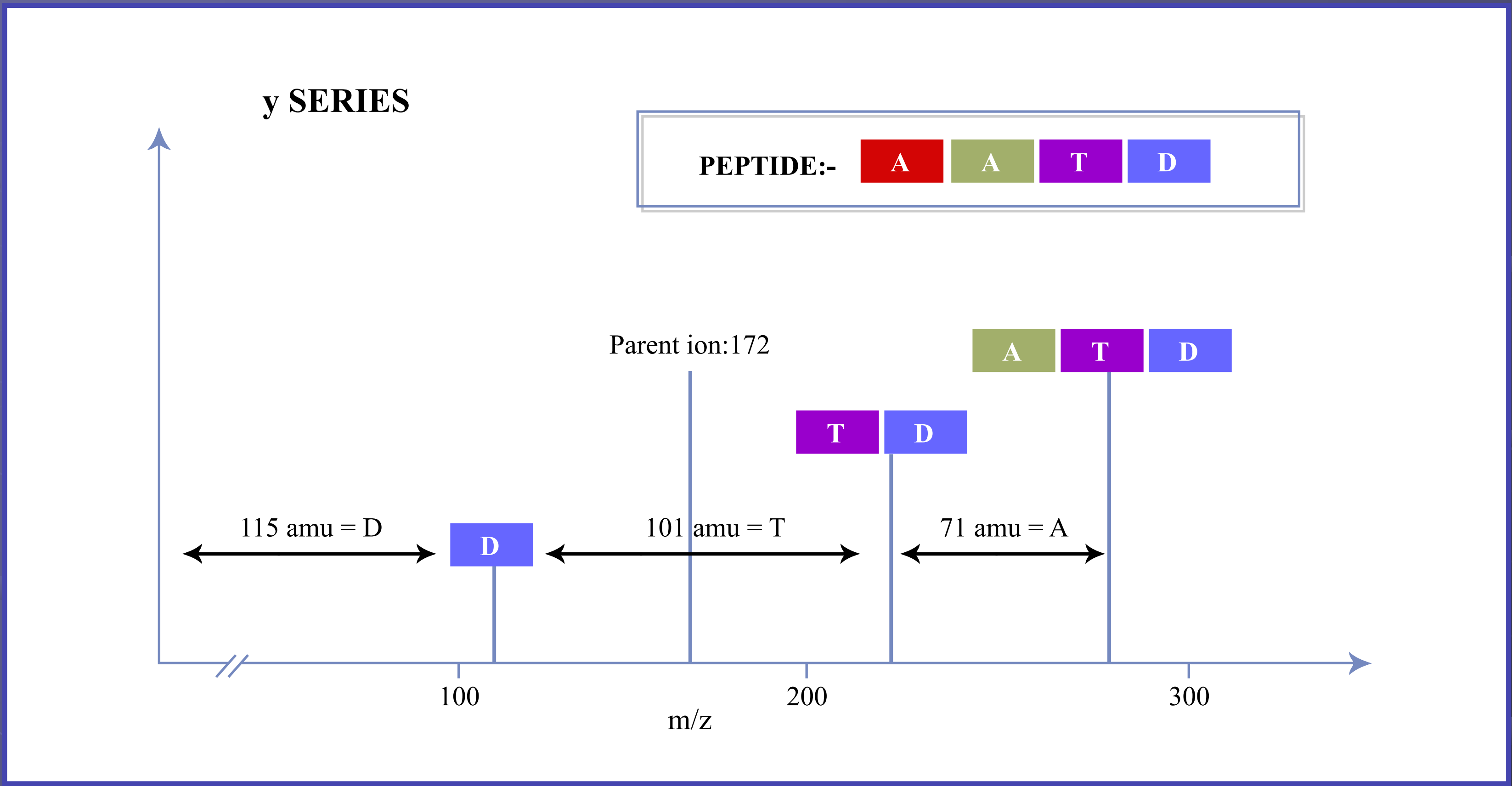


Figure by MIT OCW.

Harvard-MIT
Division of Health
Science & Technology