

Course Overview

BIO331/CS330

Safee Ullah Chaudhary

Lecture # 1

Department of Life Sciences, SBASSE, LUMS



1

Course Outline

1. Discuss course outline
2. Instruments of evaluation for Part A
 - 4 x Quizzes (4%)
 - 4 x Assignments (10%)
 - 1 Midterm (30%)
 - 1 Final (30%)
 - Class Participation (4%)
3. Introduce textbooks
 - a. Ingvar et al (Protein Sequences)
 - b. Mount et al (Protein Structure)



2

Getting to know each other!

- My Introduction
 - biolabs.lums.edu.pk/birl
- Code of Conduct
- Your Background
- Your Expectations from this course



3

Meeting Times

LECTURE

MW, 2:30-3:45 pm, SAHSOL CR 2-07

Introduce theoretical concepts

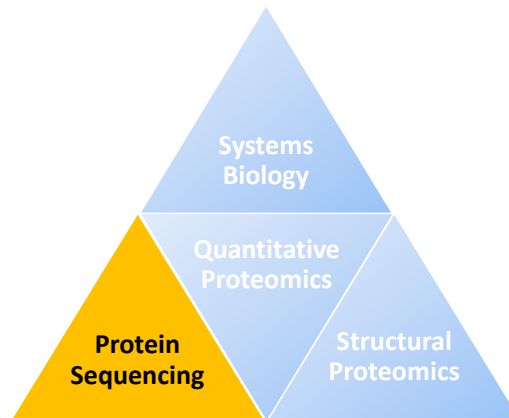
OFFICE HOURS

Fridays 2:30pm-3:00pm



4

The Great Pyramid of Proteomics!



5

Sequence Analysis

Introduction,
Experimental Techniques,
Mass & Charge Deconvolution,
Software Tools

Safee Ullah Chaudhary

Lecture # 2 - 3

Department of Biology, SBASSE, LUMS






















6

A GUIDE TO THE TWENTY COMMON AMINO ACIDS

AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC CODE ONLY DIRECTLY ENCODES 20. 'ESSENTIAL' AMINO ACIDS MUST BE OBTAINED FROM THE DIET, WHILST NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESISED IN THE BODY.

Chart Key: ● ALIPHATIC ● AROMATIC ● ACIDIC ● BASIC ● HYDROXYLIC ● SULFUR-CONTAINING ● AMIDIC ○ NON-ESSENTIAL ○ ESSENTIAL

 ALANINE (A) <small>Ala</small> <small>GCT, GCC, GCA, GCG</small>	 GLYCINE (G) <small>Gly</small> <small>GGT, GGC, GGA, GGG</small>	 ISOLEUCINE (I) <small>Ile</small> <small>ATT, ATC, ATA</small>	 LEUCINE (L) <small>Leu</small> <small>CTT, CTC, CTA, CTG, TTA, TTG</small>	 PROLINE (P) <small>Pro</small> <small>CCT, CCC, CCA, CCG</small>	 VALINE (V) <small>Val</small> <small>GTT, GTT, GTA, GTG</small>
 PHENYLALANINE (F) <small>Phe</small> <small>TTT, TTC</small>	 TRYPTOPHAN (W) <small>Trp</small> <small>TGT</small>	 TYROSINE (Y) <small>Tyr</small> <small>TAT, TAC</small>	 ASPARTIC ACID (D) <small>Asp</small> <small>GAT, GAC</small>	 GLUTAMIC ACID (E) <small>Glu</small> <small>GAA, GAG</small>	 ARGININE (R) <small>Arg</small> <small>CGT, CGC, CGA, CGG, AGA, AGG</small>
 LYSINE (K) <small>Lys</small> <small>AAT, AAG</small>	 SERINE (S) <small>Ser</small> <small>TCT, TCG, TCA, TCG, ACC, AGC</small>	 THREONINE (T) <small>Thr</small> <small>ACT, ACG, ACA, ACG</small>	 CYSTEINE (C) <small>Cys</small> <small>TGT, TGC</small>	 METHIONINE (M) <small>Met</small> <small>ATG, AAT</small>	 ASPARAGINE (N) <small>Asn</small> <small>AAT, AAA</small>
 GLUTAMINE (Q) <small>Gln</small> <small>CAT, CAA</small>					

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.

© COMPOUND INTEREST 2014 - WWW.COMPOUNDINTEREST.COM | Twitter: @compoundchem | Facebook: www.facebook.com/compoundchem
 Shared under a Creative Commons Attribution-NonCommercial-NoDerivatives license.



- Each amino acid can be presented with a single lettered amino acid tag

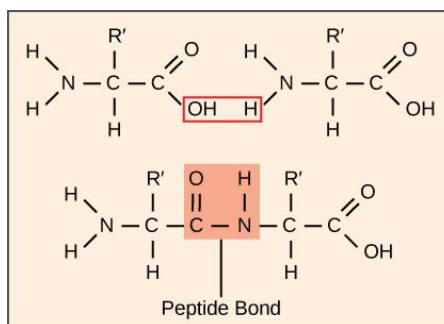
7



7

Amino Acids & Peptide bonds

- A polypeptide chain is generated by a series of condensation reactions, in vivo normally occurring within the ribosome during protein synthesis.



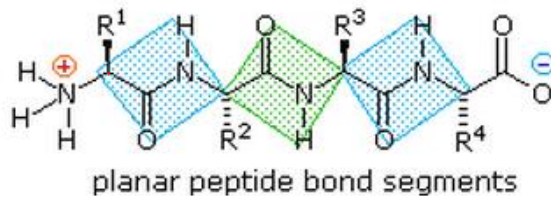
<https://www2.chemistry.msu.edu/faculty/reusch/VirtTxtJml/protein2.htm>

8



8

AND proteins are 3D, Right?



9



9



Peptide bond (Planar)



Pile of Peptide bonds



Peptide bonds supported by H-bonds



Unreal-Ray Crystallography of Paper Clip Protein



10

Angles to them all!

- Peptide bonds are planar (C=O and N-H)
- The Dihedral Angles
 1. Phi ϕ (*phi*, involving C'-N-C $^{\alpha}$ -C')
 2. Psi ψ (*psi*, involving N-C $^{\alpha}$ -C'-N)
 3. Omega ω (*Omega* C $^{\alpha}$ -C'-N-C $^{\alpha}$)
 - Controls the C $^{\alpha}$ - C $^{\alpha}$ distance
 - (Typically 180 degrees (planar) as peptide bond is planar)

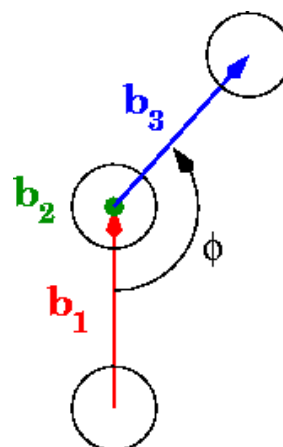


11

The Dihedral Angles

To visualize the dihedral angle of four atoms:

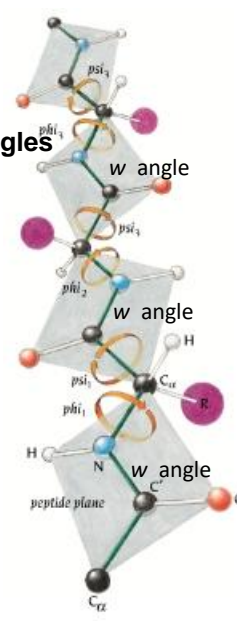
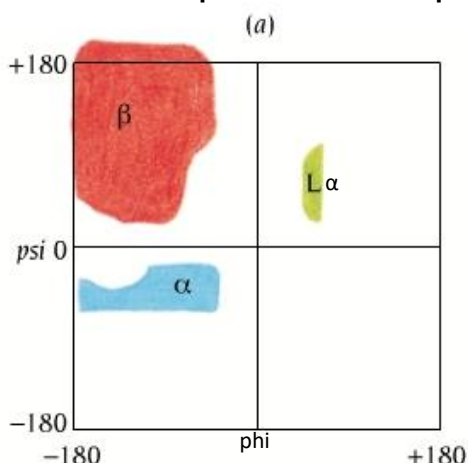
1. Look down the second bond vector
2. The first atom is at 6 o'clock, the fourth atom is at roughly 2 o'clock and the second and third atoms are located in the center.
3. The second bond vector is coming out of the page. The dihedral angle is the counterclockwise angle made by the red and blue vectors.



12

Certain side-chain configurations are energetically favored (rotamers)

Ramachandran plot: "Allowable" psi & phi dihedral angles

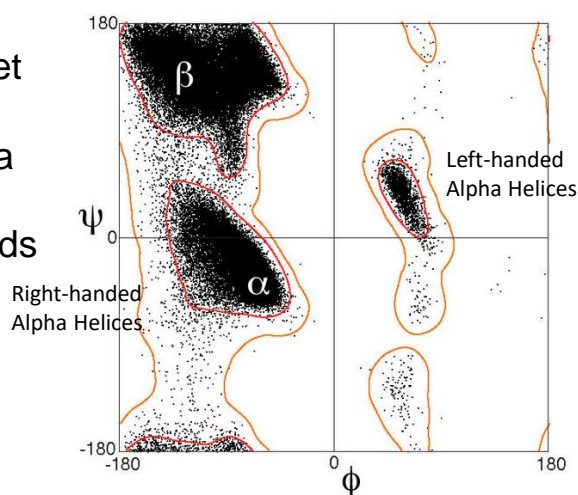


Right-handed Alpha Helix: If you hold it pointing away from you and it twists clockwise moving away, it is **right-handed**, otherwise it is **left-handed**. These models are mirror images and can not be converted into the other by rotation. The **helix** of normal DNA is **right-handed**.

13

Alpha-helix and beta-strand regions

Data as in (Lovell et al. 2003) showing about 100,000 data points for several proteins/amino-acids



<http://www.ocf.berkeley.edu/~asiegel/posts/?p=24>



14

First thing first!

Back to Protein Sequences

- So each amino acid can be presented with a single lettered amino acid tag
- And amino acids can join together, by formation of peptide bonds
- This process repeated for all codons in an mRNA molecule helps form a protein
- Hence, a protein sequence representation is essentially a concatenated list of several amino acid tags



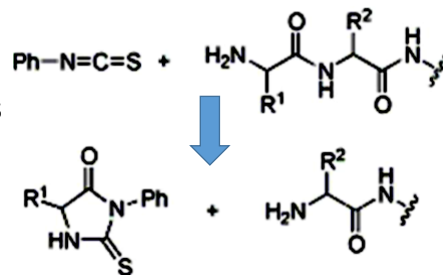
15

Protein Sequencing – Edman Degradation

- Edman degradation starting from the N-terminal and removing one amino acid at a time (details next).

- Drawback:

- Restricted to 60 residues
- Laborious: ~50 aa/day



- Modern technique: Tandem mass spectrometry

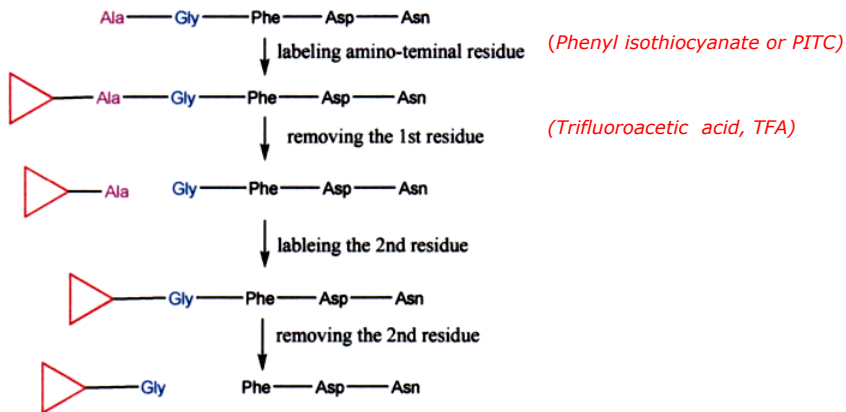
P. Edman, Acta Chem. Scand. 4, 283 (1950)



16

Proteins: Finding the Primary Structure

EDMAN DEGRADATION



(http://en.wikibooks.org/wiki/Structural_Biochemistry/Proteins/Protein_sequence_determination_techniques)

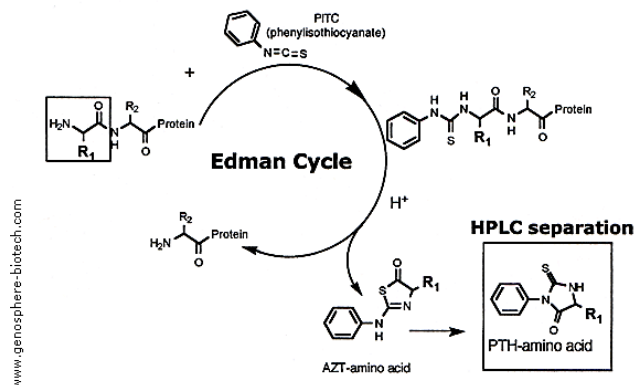
17



17

Edman Cycle

N-terminal sequencing cycle:



www.genosphere-biotech.com



18

Mass Spectrometry-based Proteomics

- **Objective:** Large-scale determination of gene and cellular function directly at the protein level
- **Challenge:** Complexity of cellular proteomes and the low abundance of many of the proteins necessitates highly sensitive analytical techniques
- Hence, mass spectrometry (MS) based proteomics has become the method of choice for analysis of complex protein samples
- High throughput MS-based proteomics is now an indispensable technology to interpret genome-wide information



19

Mass Spectrometry based Proteomics



20

What is a Mass Spectrometer?

Mass spectrometer is an analytical device that measures molecular masses within a sample.

How?

Mass spectrometer ionizes molecules and sorts them based on their **mass-to-charge** (m/z) ratio against their relative abundance.



21

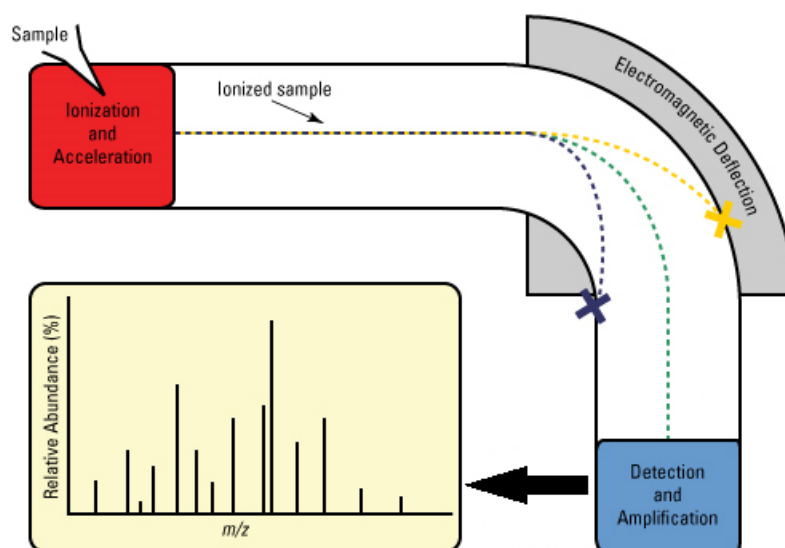
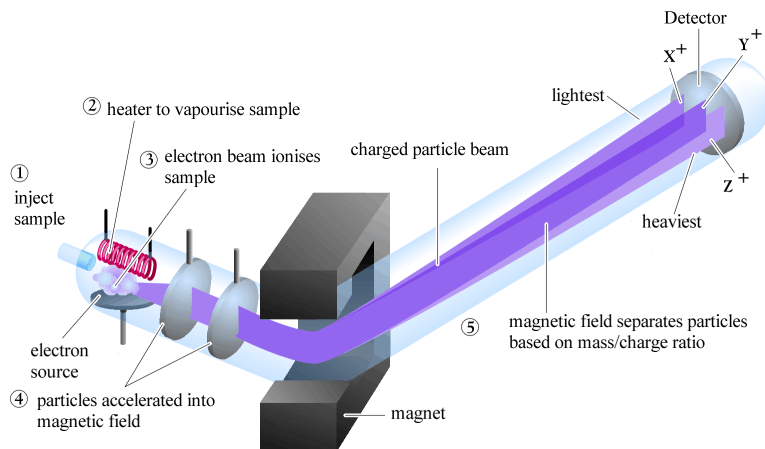


Figure - Mass Spectrometer Workflow



22

Physics behind Mass Spectrometry



www.mhhe.com



23

m/z Ratio

- Moving charged particles in a magnetic field experience forces given by

$$\mathbf{F} = m\mathbf{a} = m \frac{d\mathbf{v}}{dt} \quad (\text{Newton's second law of motion})$$

$$\mathbf{F} = Q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \quad (\text{Lorentz force law}) \quad \longrightarrow \quad \text{Force} \propto Q$$

F is the force applied to the ion, m is the mass of the particle, a is the acceleration, Q is the electric charge, E is the electric field, and $\mathbf{v} \times \mathbf{B}$ is the cross product of the ion's velocity and the magnetic flux density.



24

Components in a Mass Spec.

- **Ionization**
 - Proteomics typically involves addition of proton(s) to the protein or peptide
 - Protonation changes the mass by +1
 - Charged molecules are then transferred to Mass Analyzer
- **Mass Analyzer**
 - Separates the samples according to their m/z
- **Detector**
 - Selected molecules then hit the detector
- **Spectrum Assembly**
 - Proteomics software which is interfaced to the MS, assembles spectra

Lahore University of Management Sciences
(LUMS), Pakistan



25

Big Names in Proteomics



Top Down Proteomics
www.kelleher.northwestern.edu



26



GC-MS
Fred McLafferty in in his
lab at Cornell



Swiss Prot
Ralph Apweiler at European
Molecular Biology Lab



27



PST Approach (de novo),
Nano Electrospray
Mathias Mann at Max Planck



Quantitative Proteomics
Rudy Aebersold, ETH
Zurich



FT-ICR Development
Richard Smith at PNNL

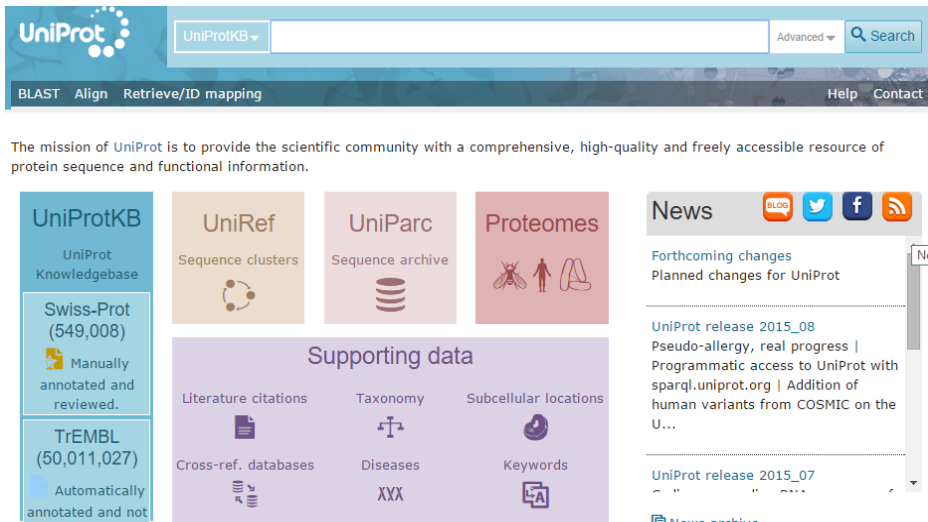


28

Protein Sequence Databases

- Searching for a protein by ID

<http://www.uniprot.org/>



29

Peptide Databases

- Peptide Atlas
- PepBank (Harvard)
- Cancer Peptide and Protein Database (CPPD)
- Antibacterial Peptides (Antibiofilms)
- Antiviral Peptides (Anti-HIV)
- Antifungal Peptides
- Antiparasitic Peptides (Antimalaria)
- Antiparasitic Peptides (Antimalaria)
- Anticancer Peptides
- Anti-protist Peptides
- Insecticidal Peptides
- Spermicidal Peptides
- Chemotactic peptides
- wound healing
- Antioxidant peptides
- Protease inhibitors



30

30

MS Spectral Data Processing – Charge State Deconvolution

- Charge needs to be estimated before calculation of the mass from m/z ratios $1 \text{ kg} = 6.022\text{e}+26 \text{ amu}$

Example Suppose we have a peptide with mass 2000.0 Da, and that the ionization yields peptide ions of charge +1, +2, and +3, by the attraction of one, two, or three protons, respectively. The peptide ions will then be detected at

ions with charge +1: $m/z = (2000 + 1)/1 = 2001$

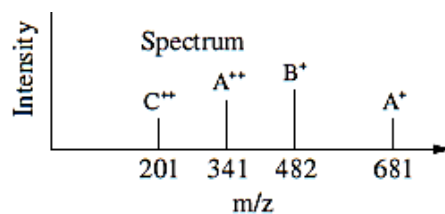
ions with charge +2: $m/z = (2000 + 2)/2 = 1001$

ions with charge +3: $m/z = (2000 + 3)/3 = 666.7$



31

Activity

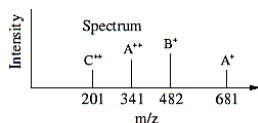


Calculate the MW of A, B and C



32

MS Spectral Data Processing – Isotopic Envelope Deconvolution



- Most abundant isotopic molecule will give the monoisotopic peak for small molecules
- However, for large molecules e.g. proteins, the most abundant peak will be the average peak

Element		Abundance (%)	Mass
Hydrogen	¹ H	99.99	1.007 83
	² H	0.01	2.014 10
Carbon	¹² C	98.91	12.000 0
	¹³ C	1.09	13.003 4
Nitrogen	¹⁴ N	99.6	14.003 1
	¹⁵ N	0.4	15.000 1
Oxygen	¹⁶ O	99.76	15.994 9
	¹⁷ O	0.04	16.999 1
	¹⁸ O	0.20	17.999 2
Phosphorus	³¹ P	100	30.973 8
Sulfur	³² S	95.02	31.972 1
	³³ S	0.76	32.971 5
	³⁴ S	4.22	33.967 6

33



33

Mass Isotopic Distributions

Calculating Isotopic Mass Distributions of



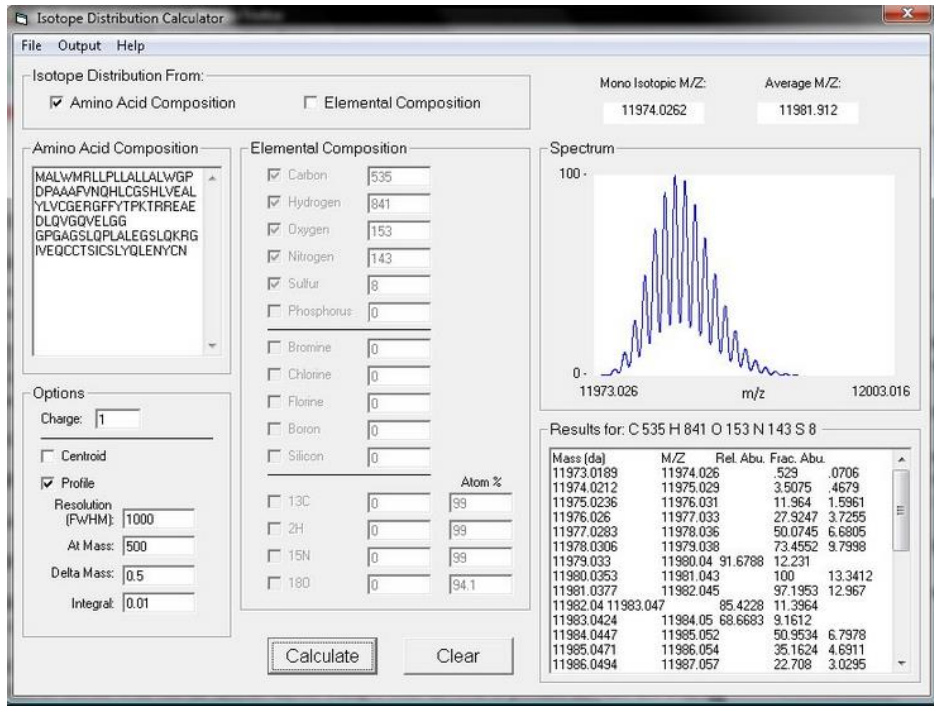
Calculate peaks and their intensity in each case and plot them!

Cookie Point (0.25)

34

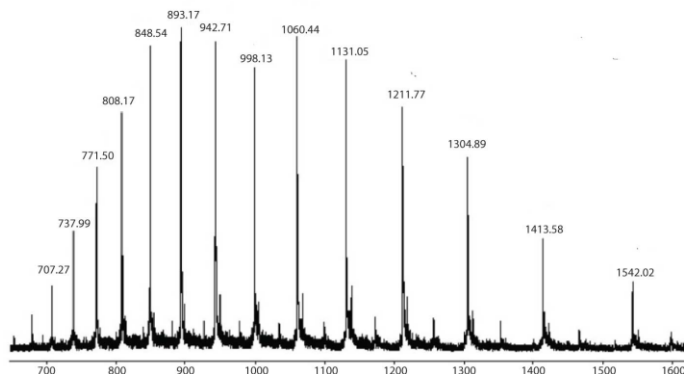


34



35

Excerpt from a Mass Spectrum



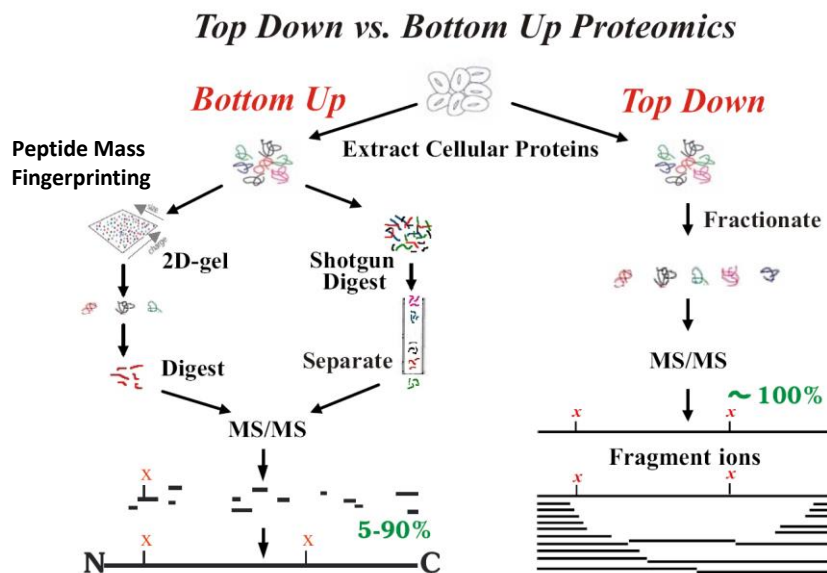
36

Approaches in High-throughput Proteomics

- Bottom up proteomics
 - Legacy protocol
 - Pros: MALDI-TOF/ESI-TOF desktop MS instruments
 - Cons: Low resolution and coverage
- Top down proteomics
 - Next-gen proteomics
 - Pros: High res, High Coverage, PTM discovery
 - Cons: Costly, Lack of search tools



37

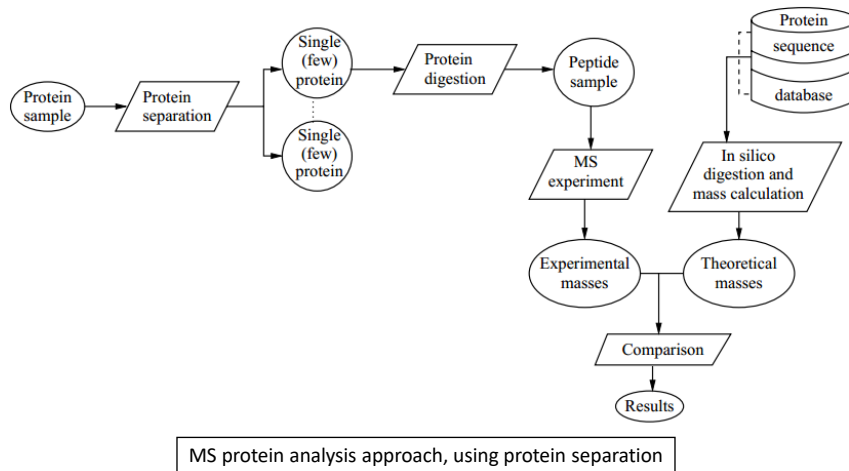


Le Duc et al



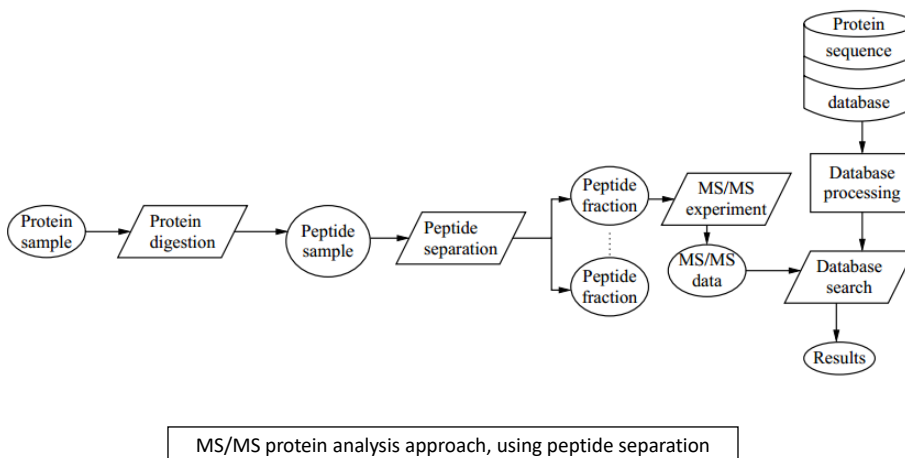
38

MS- Peptide Mass Fingerprinting



39

MS/MS- Tandem MS



40

2. Mass Analyzer – TOF

1. Ions are accelerated by an electric field, and then they enter a **field-free drift tube**
2. The ion **velocities reached are inversely proportional to the mass and the charge of the ion**
3. The time needed to pass through the drift tube is **dependent on the velocity**. When the ions hit the detector at the end of the drift tube, the **flight time is registered, and the m/z value can be calculated.**

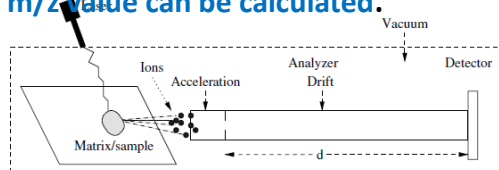


Figure 5.3 Illustration of a linear MALDI-TOF analyzer

41



41

Single mode and Tandem Mode

- **Single mode** MS can report m/z of an intact peptide
 - Advantage: Single charge on molecules
- **Tandem mode** (MS/MS) splits each peptide into two ions each
 - N-terminus and C-terminus fragments
 - BUT, to detect both the fragments, **each fragment needs to be charged**
 - So, **its advantageous to have multiple charges per fragment**, as it increases the probability that both ions will be detected

42



42

Hurdles in Application of MS

1. Hard ionization techniques
2. Low resolution of mass analyzers
3. Search Algorithms
 - Isotopic envelope deconvolution
 - Post-translational modifications detection



43

Soft Ionization to the Rescue!

The Nobel Prize in Chemistry 2002



John B. Fenn
Prize share: 1/4



Koichi Tanaka
Prize share: 1/4



Kurt Wüthrich
Prize share: 1/2

The Nobel Prize in Chemistry 2002 was awarded "for the development of methods for identification and structure analyses of biological macromolecules with one half jointly to [John B. Fenn](#) and [Koichi Tanaka](#) "for their development of *soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules*" and the other half to Kurt Wüthrich "for his development of nuclear magnetic resonance spectroscopy



44

Soft vs. Hard Ionization

• Soft Ionization

- Techniques which induce limited fragmentation during the ionization process
- Typically used on peptides and proteins (for generating molecular ion)
- MALDI, ESI

• Hard Ionization

- Frequently induces fragmentation during the ionization
- FAB (Fast Atom Bombardment), Electron Impact Ionization (EI) etc.

45



MALDI – Step by Step

- Matrix: small organic molecules which can absorb light
- The matrix and the sample are dissolved in an organic solvent
- Small drops of the solvent are spotted on plates and left for evaporation
- During this process, matrix forms crystals with the sample trapped inside
- Laser shots are delivered to the matrix. Crystals absorb the light and are ionized and their structure protects analytes
- Energy from the laser, ejects matrix/sample molecules from crystals
- Sample molecules receive the protons from the matrix (+1 for MALDI)
- An electric guide beam delivers the ions to the MS chamber.



46

ESI – Step by Step

- Peptides are brought in with a liquid flow (HPLC)
- The sample is sprayed by a heated needle into a strong electromagnetic field. A potential is also applied to the needle before spraying of the sample.
- As the solvent from the droplets evaporates, the drop size gets smaller, increasing the electromagnetic field
- Once the charge gets large enough, the charged sample molecules desorb from the surface
- Result: Molecules with 2+, 3+ etc. charge
- Primarily used for MS/MS analysis (Why?)



47

High Resolution Mass Spec!



Earth's field ranges between approximately 25,000 and 65,000 nT



48

SPECTRUM: A MATLAB Toolbox for Identifying Proteins from Top-Down Proteomics Data

Biomedical Informatics Research Laboratory,
LUMS



49

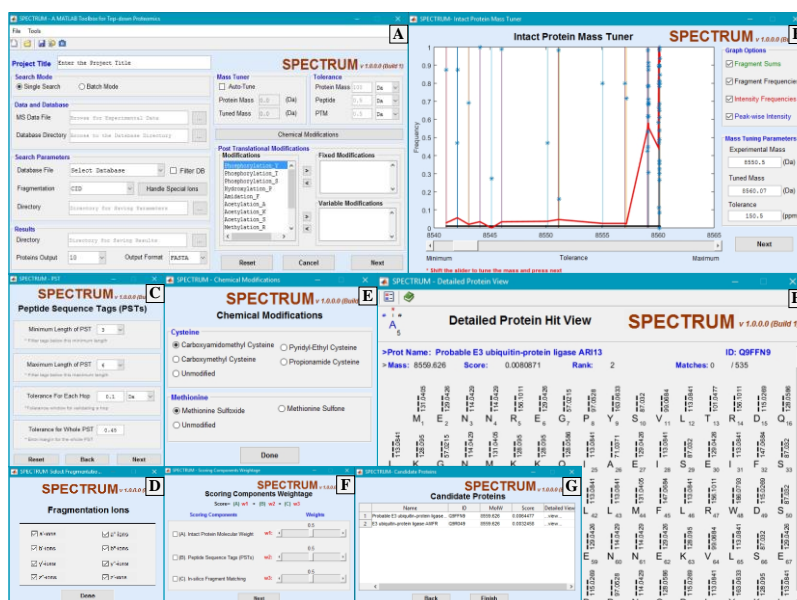


Fig. SPECTRUM GUIs. The set of graphical user interfaces in SPECTRUM created using MATLAB GUIDE to undertake the search process and visualize results. (A) Main SPECTRUM GUI to provide general search parameters, (B) GUI to tune intact protein mass, (C) GUI to provide PST search parameters, (D) GUI to include special fragmentation ions in the search process and (E) GUI to specify instrument based chemical modification. (F) GUI to tailor final scoring scheme, and (G-H) GUIs to provide the user with brief as well as detailed results.



50

Overview

- A toolbox for protein identification from top-down proteomics data built using MATLAB
- Open-source and open-architecture system for development, testing and benchmarking of top-down proteomics algorithms
- Search pipeline that seamlessly brings together key proteomics algorithms resulting in a lower FDR rates as compared to industry standard tools
- An intuitive yet comprehensive graphical user interface for convenient utilization as well as customization by the users



51

Key Features

1. Multiple data file format support including mzXML, MGF and flat text files
2. Intensity weighted sliding window protocol for intact protein mass tuning
3. *De novo* peptide sequence tag extraction and its scoring
4. PTMs identification (Site specific & Blind),
5. Abundance weighted *in silico* spectral comparison,
6. A multifactorial additive scoring scheme employing coefficient weighted constituent scores
7. A set of graphical user interfaces built using MATLAB GUIDE providing access to the aforementioned functionalities



52

Features Comparison - SPECTRUM vs other TDP tools

Top Down Proteomics Tools	Supported Features									
	Intact Mass Tuning/Filter	De novo Sequencing	Fragmentation Techniques	Variable PTM Search	Multiple PTM Search	Blind PTM Search	Truncated Protein Search	In silico Spectral Comparison	Graphical User Interface	Protein Quantitation
SPECTRUM	✓	✓	✓ (9 Types)	✓	✓	✓	✓	✓	✓	✗
pTop	✗	✓	✓	✓	✓	✗	✓	✓	✓	✗
MSPathFinder	✗	✓	✓	✓	✓	✗	✓	✓	✗	✓
MASH Suite Pro	✗	✗	✓ (2 Types)	✓	✓	✓	✓	✓	✓	✓
ProSightPC	✗	✓	✓ (7 types)	✗	✓	✓	✗	✓	✓	✗
TopPIC	✗	✗	✓ (4 types)	✗	✗	✓	✗	✓	✓	✗



53

Results – Case Study on HeLa H4 Histone

- Case Study I – Evaluation of SPECTRUM Search with Known Target Protein

Results Comparison for HeLa Dataset	Search Parameters					
	PST Search: Disabled Scoring Component: In silico Blind PTM Search: Disabled		PST Search: Disabled Scoring Component: In silico Blind PTM Search: Enabled		PST Search: Enabled Scoring Component: In silico Blind PTM Search: Disabled	
	SPECTRUM	ProSight PC*	SPECTRUM	TopPIC*	SPECTRUM	pTop*
Protein Spectral Matches	10	10	10	8	8	0
Proteins Identified	3	3	3	2	1	0
True Positives (out of 10)	8	8	8	7	8	0
Not Reported	0	0	0	2	2	10
Search Time (in seconds)	15	24	16	2350	15	13

* ProSight PC v4.0, TopPIC v1.1, pTop v1.2

PST : Peptide Sequence Tag



54

Results – Case Study *E. coli* Dataset

• Case Study II – Evaluation of SPECTRUM Search with Unknown Target Protein

Results Comparison for <i>E. coli</i> Dataset	With PSTs/TagSearch			Without PSTs/TagSearch		
	SPECTRUM	MSPahtFinder	pTop	SPECTRUM	MSPahtFinder	TopPIC
No. of PrSMs Identified	1739	1458	1181	1911	1319	1262
No. of Proteins Identified	245	128	128	305	110	128
Total Search Time (in seconds)	2228 [†] (4456*)	344 [†]	619 [†]	1678 [†] (3356*)	304 [†]	751 [†]
Average No. of PrSMs per Protein	7	11	9	6	12	10
Average No. of matched fragments for each PrSM	39.4	55	41	41.7	56	36

[†] Average compute time for 1 target and 1 decoy search

* Average compute time for 1 target and 3 decoy searches

Note: Benchmarking performed using a desktop machine with Intel® C7 7700 @ 4.2GHz and 32 GB RAM



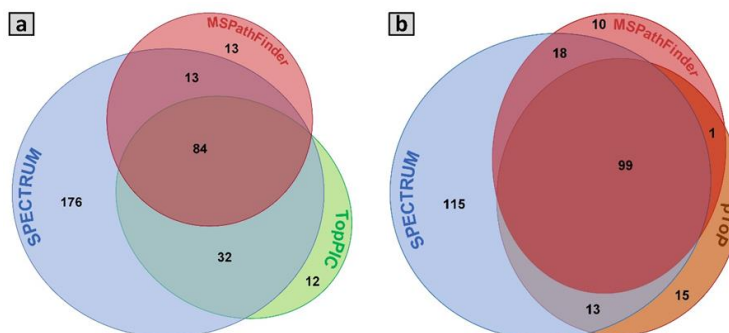
55

Results – Case Study *E. coli* Dataset

❑ Validation/benchmarking of the platform was performed against:

1. **Published datasets**
2. **TDP tools** (ProSight PC, pTop, TopPIC, MSPahtFinder)

❑ SPECTRUM enhanced protein identification rates from 91% to over 177%




56

MENU ▾

SCIENTIFIC REPORTS

Article | [OPEN](#) | Published: 02 August 2019

SPECTRUM – A MATLAB Toolbox for Proteoform Identification from Top-Down Proteomics Data

Abdul Rehman Basharat, Kanzal Iman, Muhammad Farhan Khalid, Zohra Anwar, Rashid Hussain, Humnah Gohar Kabir, Maria Tahreem, Anam Shahid, Maheen Humayun, Hira Azmat Hayat, Muhammad Mustafa, Muhammad Ali Shoaib, Zakir Ullah, Shamshad Zarina, Sameer Ahmed, Emad Uddin, Sadia Hamera, Fayyaz Ahmad & Safee Ullah Chaudhary 

Scientific Reports **9**, Article number: 11267 (2019) | [Download Citation](#) 