BIO312
Protein Structure and Function

Class Today

- Self Introduction
- Handbook Highlights
- Lecture 1

About Dr. Han

- **Research Keywords**: Virulence regulators; Fructose-induced fatty liver; Protein crystallography; Drug design and screen

- Email address: Lanlan.Han@xjtlu.edu.cn
- Office telephone number: (0512)89165331
- Room number and office hours: SA344, by appointment

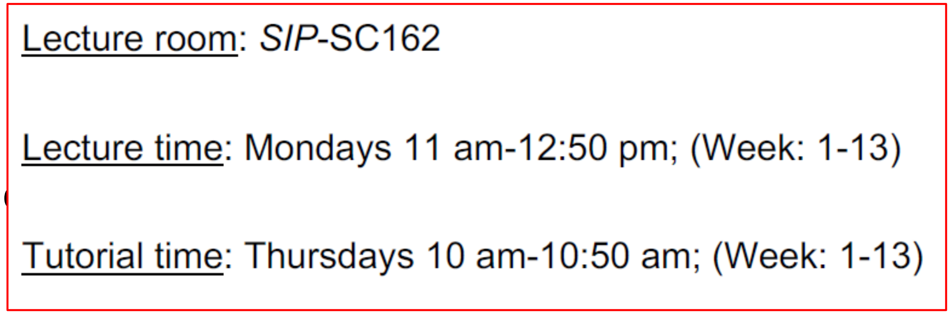
Handbook Highlights -- Contents

Methods and Applications of Proteome Analysis. (Lecture 1-6);

Lecture 1: Mischaracterization of proteins
Lecture 2: Identification of proteins in complex mixtures
Lecture 3: Bottom-up and top-down proteomics
Lecture 4: Quantitative mass spectrometry-based proteomics
Lecture 5: Protein post-translational modifications
Lecture 6: Targeted proteomics

Acquisition of proteins and protein structure determination (lecture 7-14);

Lecture 7: Protein folding and unfolding
Lecture 8: Overview of techniques and applications
Lecture 9: Circular dichroism and infrared spectroscopy
Lecture 10: X-ray crystallography
Lecture 11: Nuclear magnetic resonance: Structural Determination
Lecture 12: Overview of Techniques for protein-ligand interactions
Lecture 13: Modelling: Computational Prediction of Structure & Dynami
Lecture 14: Relating structure to function



Lecture room: *SIP*-SC162

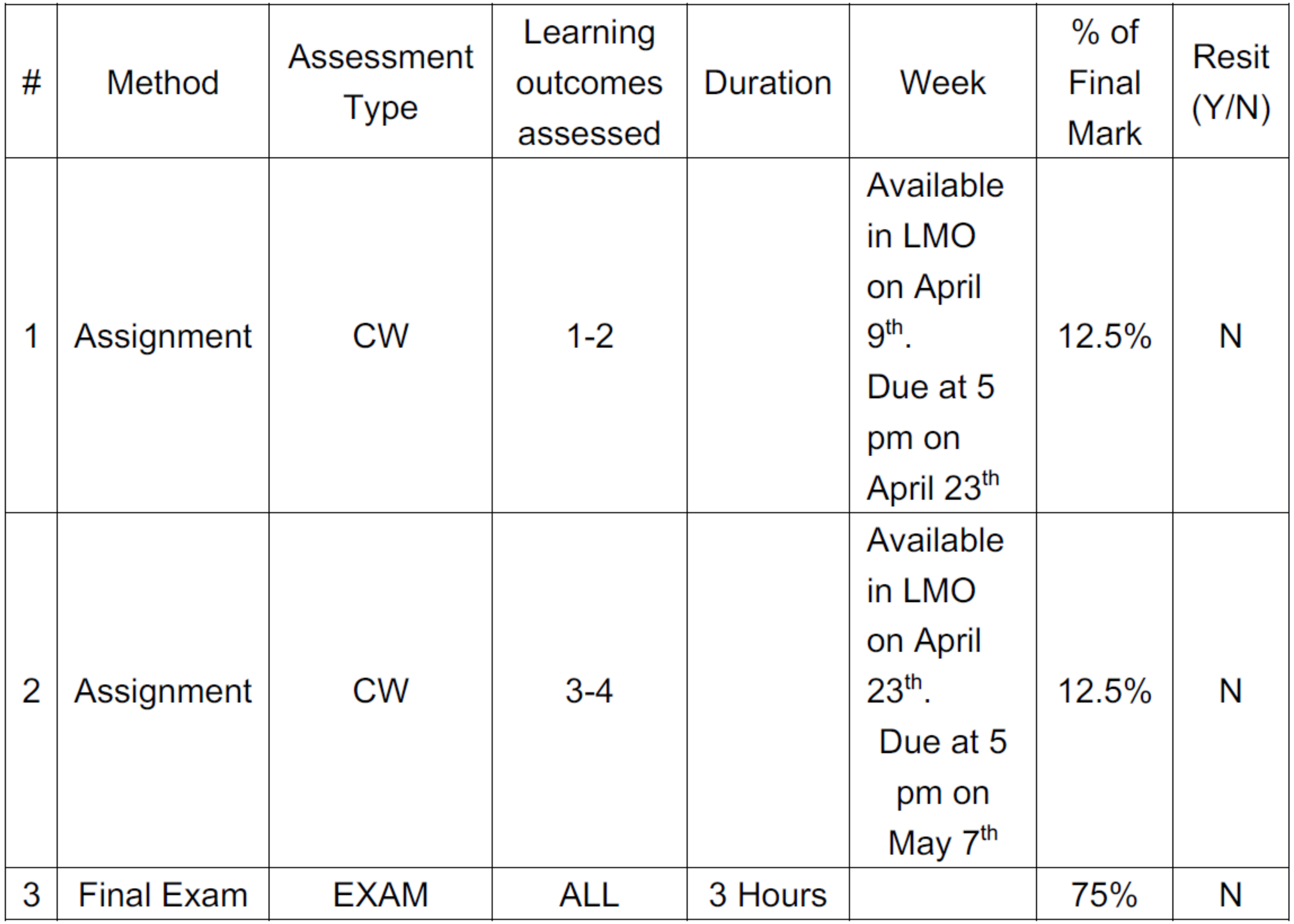
Lecture time: Mondays 11 am-12:50 pm; (Week: 1-13)

Tutorial time: Thursdays 10 am-10:50 am; (Week: 1-13)

Handbook Highlights -- Learning Outcomes

1. Know the methods used for the extraction, enrichment and analysis of proteins
2. Discuss how proteomics-based approaches can be used to study fundamental and applied biological problems
3. Describe methods of analysis of post-translational modifications of proteins and implications for cell function
4. Understand how technologies such as nuclear magnetic resonance (NMR), Xray crystallography and other physical methods can be used to determine the detailed fine structure of proteins
5. Explain how the knowledge of protein structure can be used to explain function, in particular the structural basis of receptor signaling and the function of molecular motors.

Handbook Highlights -- Assessment



#	Method	Assessment Type	Learning outcomes assessed	Duration	Week	% of Final Mark	Resit (Y/N)
1	Assignment	CW	1-2		Available in LMO on April 9th. Due at 5 pm on April 23th	12.5%	N
2	Assignment	CW	3-4		Available in LMO on April 23th. Due at 5 pm on May 7th	12.5%	N
3	Final Exam	EXAM	ALL	3 Hours		75%	N

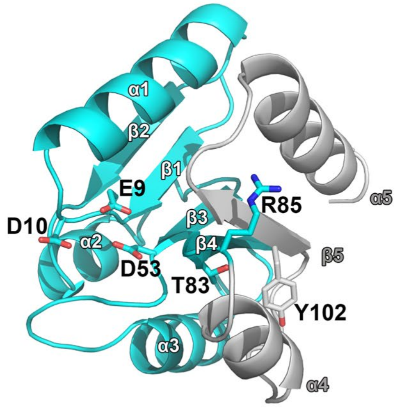
LECTURE 1: MICRO-CHARACTERIZATION OF PROTEINS

What are proteins?

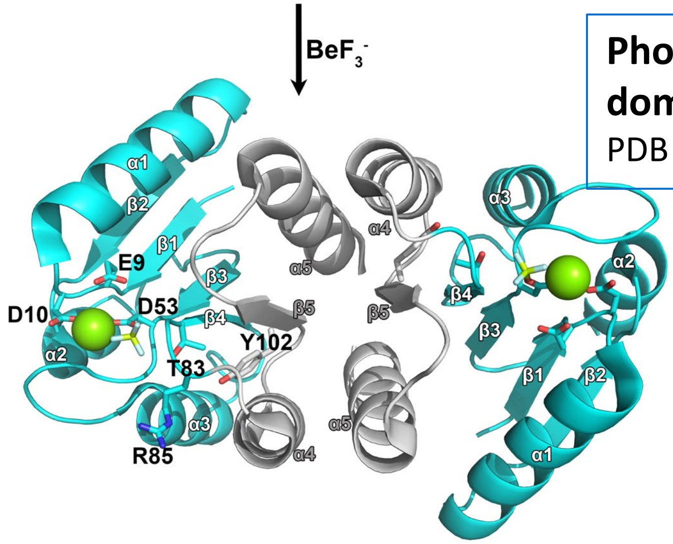
- **Proteins are polymers of amino acids**
- Proteins play a crucial role in biological processes and have many important biological functions
    - **Enzymes** – biological catalysts
    - **Defense proteins**
        - **Antibodies** produced in response to **antigens**
    - **Transport proteins**
    - **Regulatory proteins**
    - **Structural proteins**
    - **Movement proteins**
    - **Nutrient Proteins**

How do Proteins Act?

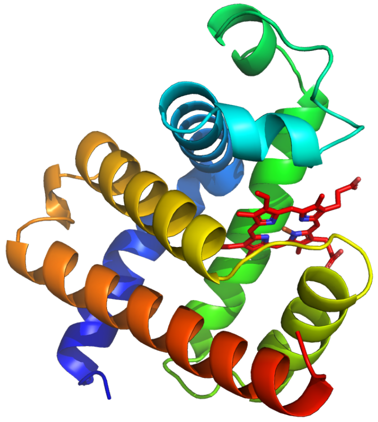
- Proteins have many important biological functions
  - Binding
  - Catalysis
  - Molecular switches
  - Structural component



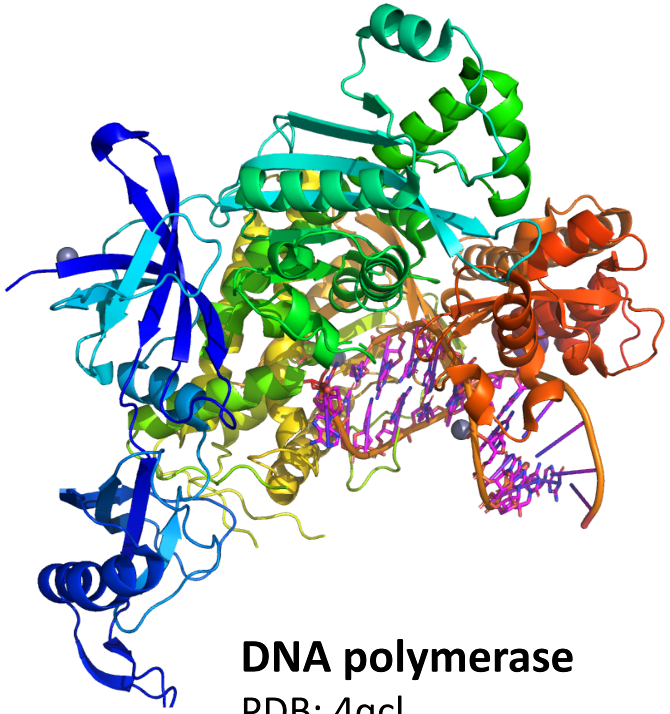
PhoB REC domain monomer
PDB ID: 1B00



PhoB REC domain Dimer
PDB ID: 1ZES

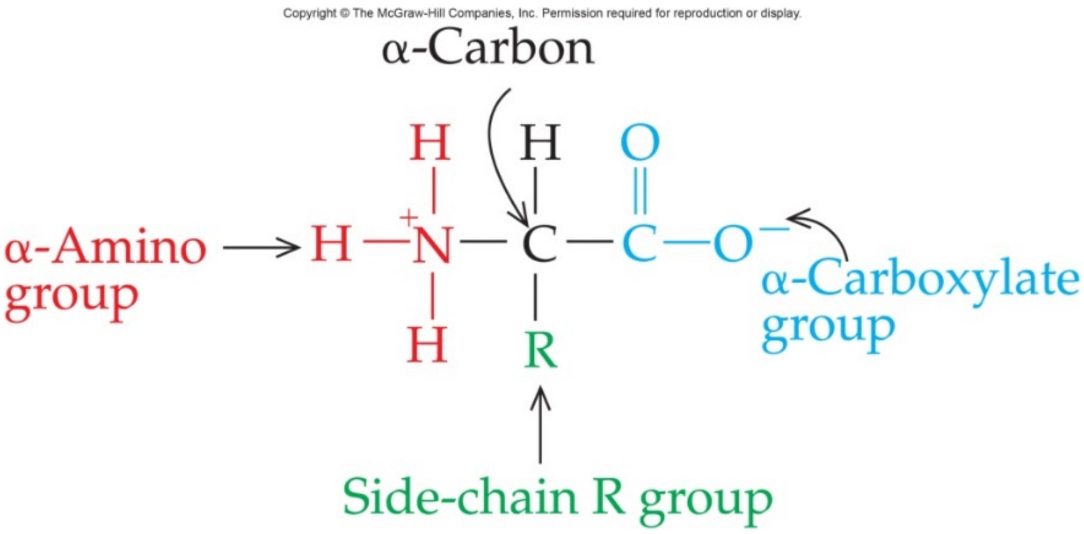


Myoglobin
PDB: 1a6k



DNA polymerase
PDB: 4qcl

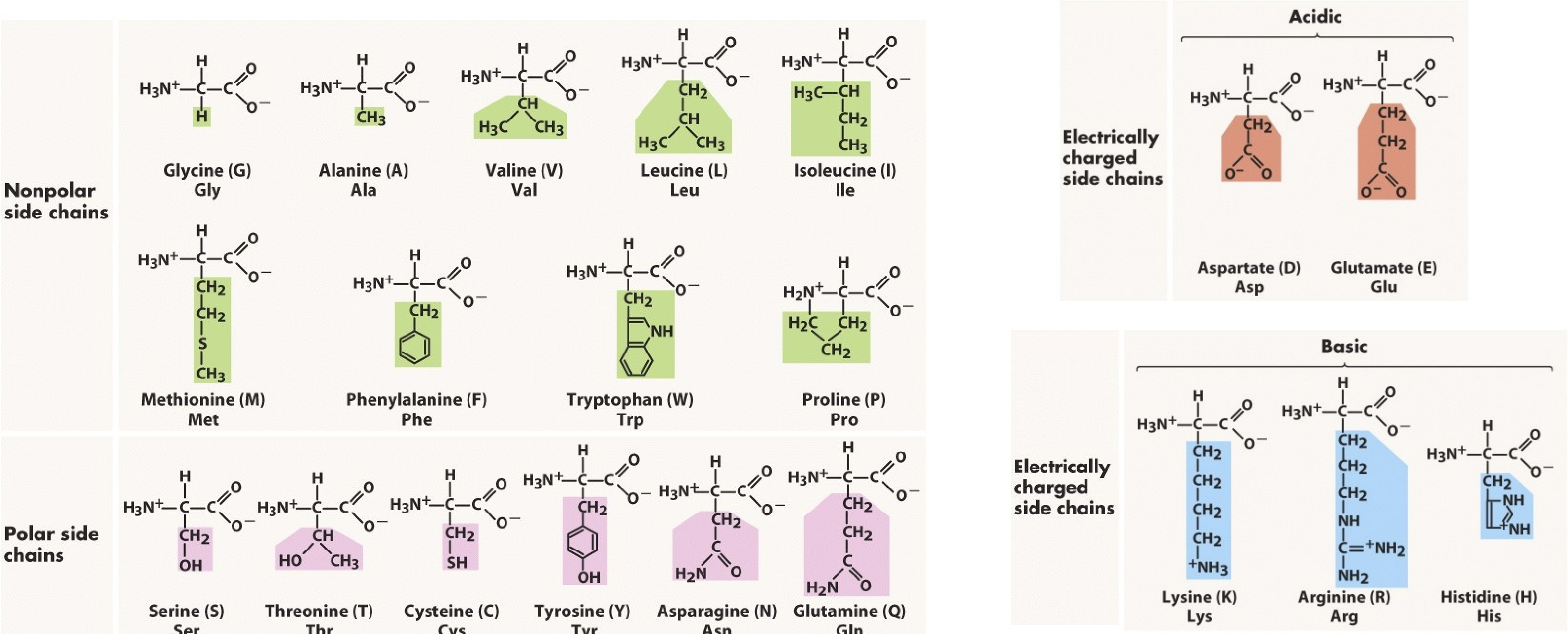
Protein Building Blocks - Amino Acids



- **α-Amino acids** contain both an amine and an acid
- 20 common amino acids in nature
  - Differ in R group
- At physiologic pH the amino acid has:
  - Carboxyl group in $–COO^-$
  - Amino group in $–NH_3^+$
- Neutral molecule with equal number of

  + and − charges is a **zwitterion**

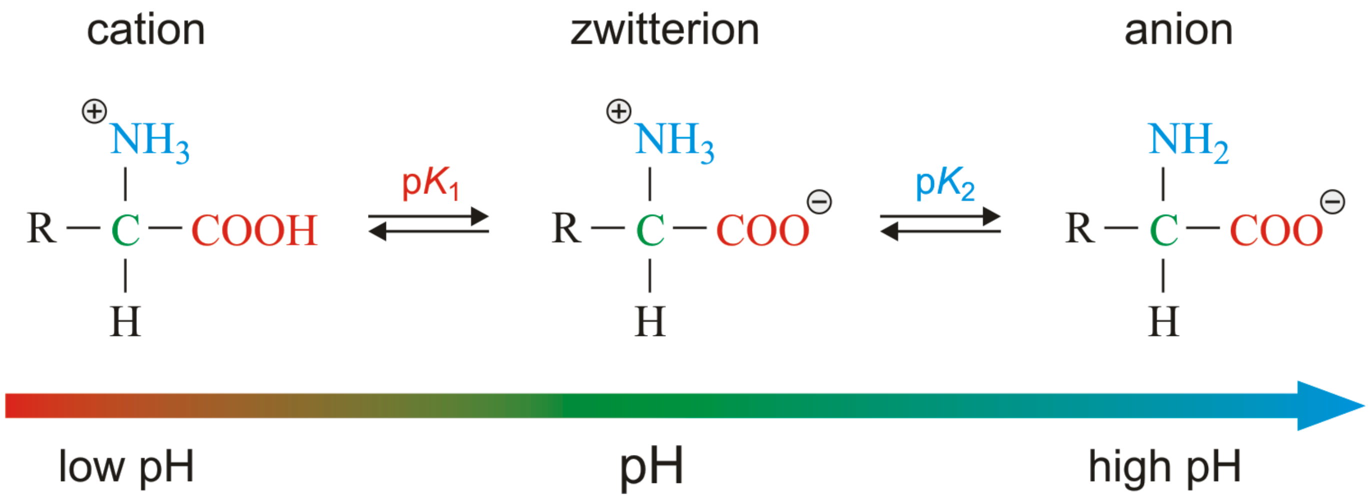
Structures and Abbreviations of the 20 Common Amino Acids



© 2005 Pearson Prentice Hall, Inc.

Zwitterions

- By changing the pH, you can affect the net charge on the zwitterions
- The pH point at which there is no net charge on the zwitterions is called the ***isoelectric point (pI)***

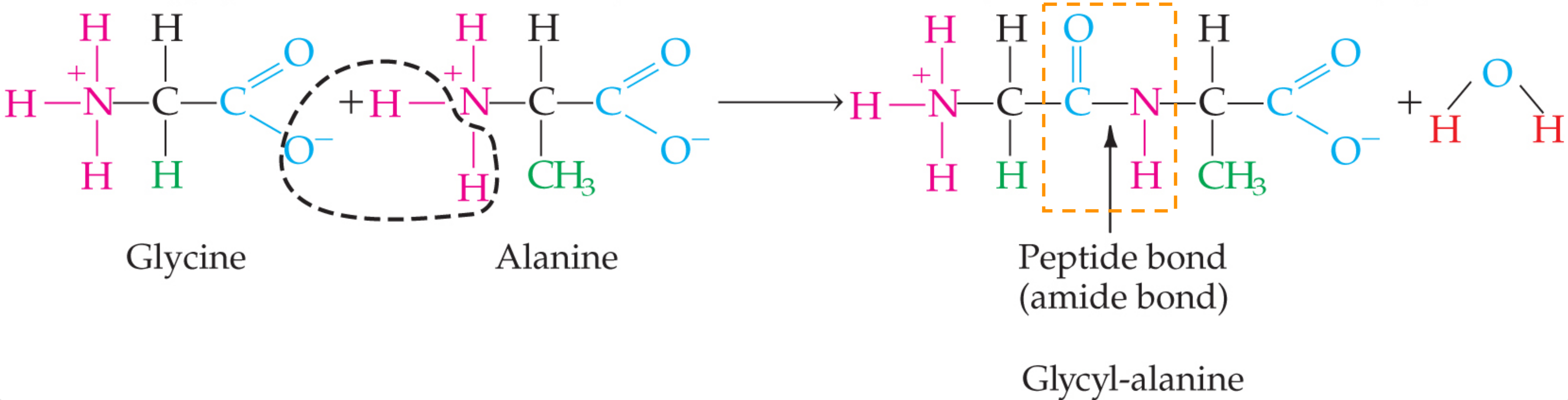


- pH < pI, positively charged
- pH > pI, negatively charged

https://qph.fs.quoracdn.net/main-qimg-9c4ec7925ae4465111a7cb10476d58b8

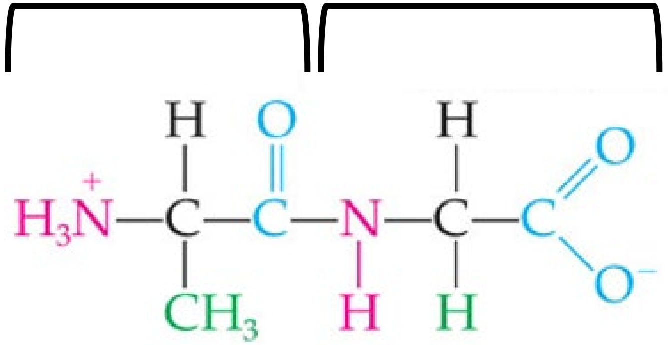
The Peptide Bond

- Proteins are polymers of L-α-amino acids
  - Carboxyl group of one amino acid is linked to the amino group of another amino acid
  - Linkage is an amide bond or **peptide bond**
  - This reaction is a dehydration reaction as water is released



Dipeptides

- Condensing or dehydrating two amino acids produces a dipeptide
  - Amino acid structures are written with the N-terminal on the left



residue 1

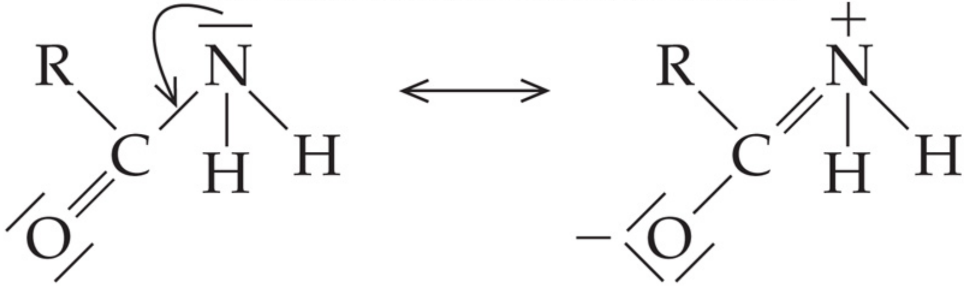
residue 2

N terminus

C terminus

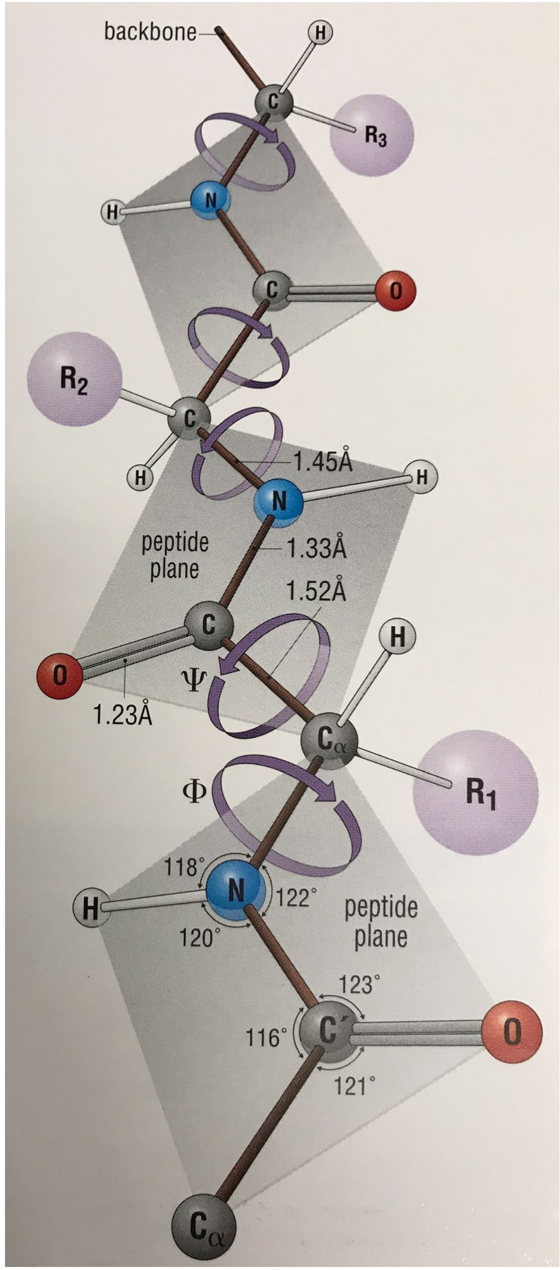
Alanyl-glycine
(ala-gly)

Structure of Peptide Bond



- Resonance
  - Increase the polarity of the peptide bond
  - Partial double bond property: the carbonyl O, Carbonyl C, and the amide N) are **coplanar**

- N-C_α and C_α-C bonds are free rotatable



Protein structure and Function. Gregory A Petsko, Dagmar Ringe.

From Sequence to Structure: Primary Structure

- Central Dogma:
  - Genetic information flows: DNA → RNA → Protein

- **Primary structure** is the amino acid sequence of the polypeptide chain connected by the peptide bonds
  - Is determined by the sequence of a gene (DNA)
  - Between proteins, the more similar the sequence the more similar:
    - Their function
      - Between species; within species
    - Their evolutionary history

From Sequence to Structure: Primary Structure

- Sequence alignment is a way of arranging primary sequences (of DNA, RNA, or proteins) in such a way as to align areas sharing common properties.

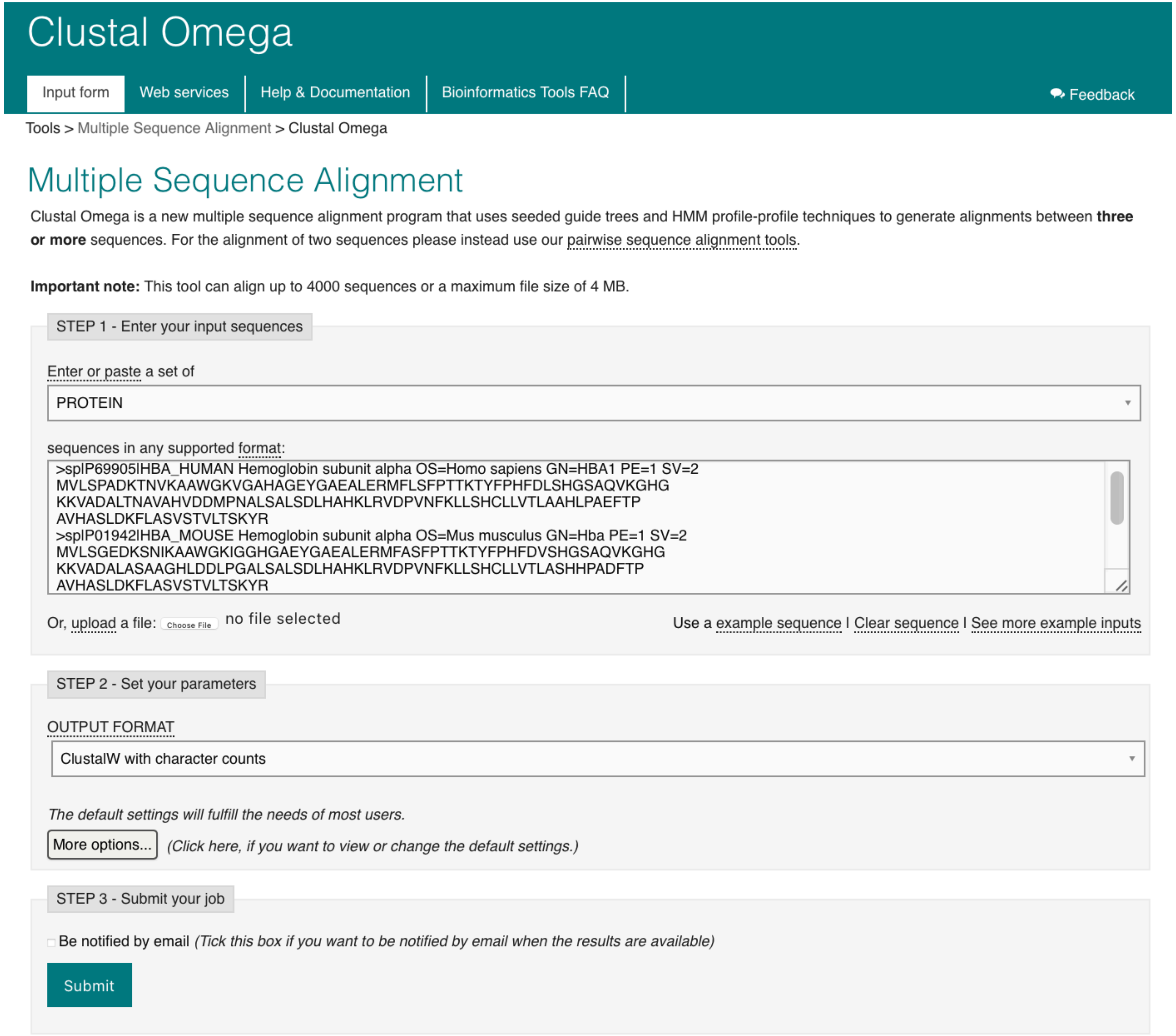
```
CLUSTAL O(1.2.1) multiple sequence alignment (Partial)


RitR      MGKRILLLEKERNLAHFLSLELQKEQYRVDLVEEGQKALSMALQTDYDLILLNVNLGDMM
PhoB      MARRILVVEDEAPIREMVCFVLEQNGFQPVEAEDYDSAVNQLNEPWPDLILLDWMLPGGS
          *.:***::*.*  : .::.: *::: ::   .*: :.*:.   :   *****:  *

RitR      AQDFAEKLSRT---KPASVIMILDHWEDLQEELEVVQRFAVSYIYKPVLIENLVARISAI
PhoB      GIQFIKHLKRESMTRDIPVVMLT-ARGEEEDRVRGLETGADDYITKPFSPKELVARIKAV
          . :* ::*.*    :   *:*:     : ::.:. ::  * .** **.  ::*****.*:
```

The degree of relatedness, similarity between the sequences is predicted computationally or statistically

Many software tool used for general sequences alignment tasks, e.g., **Clustal Omega** or **BLAST**

Clustal Omega

https://www.ebi.ac.uk/Tools/msa/clustalo/



Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01942|HBA_MOUSE Hemoglobin subunit alpha OS=Mus musculus GN=Hba PE=1 SV=2
MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
AVHASLDKFLASVSTVLTSKYR
```

Or, upload a file: Choose File no file selected

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

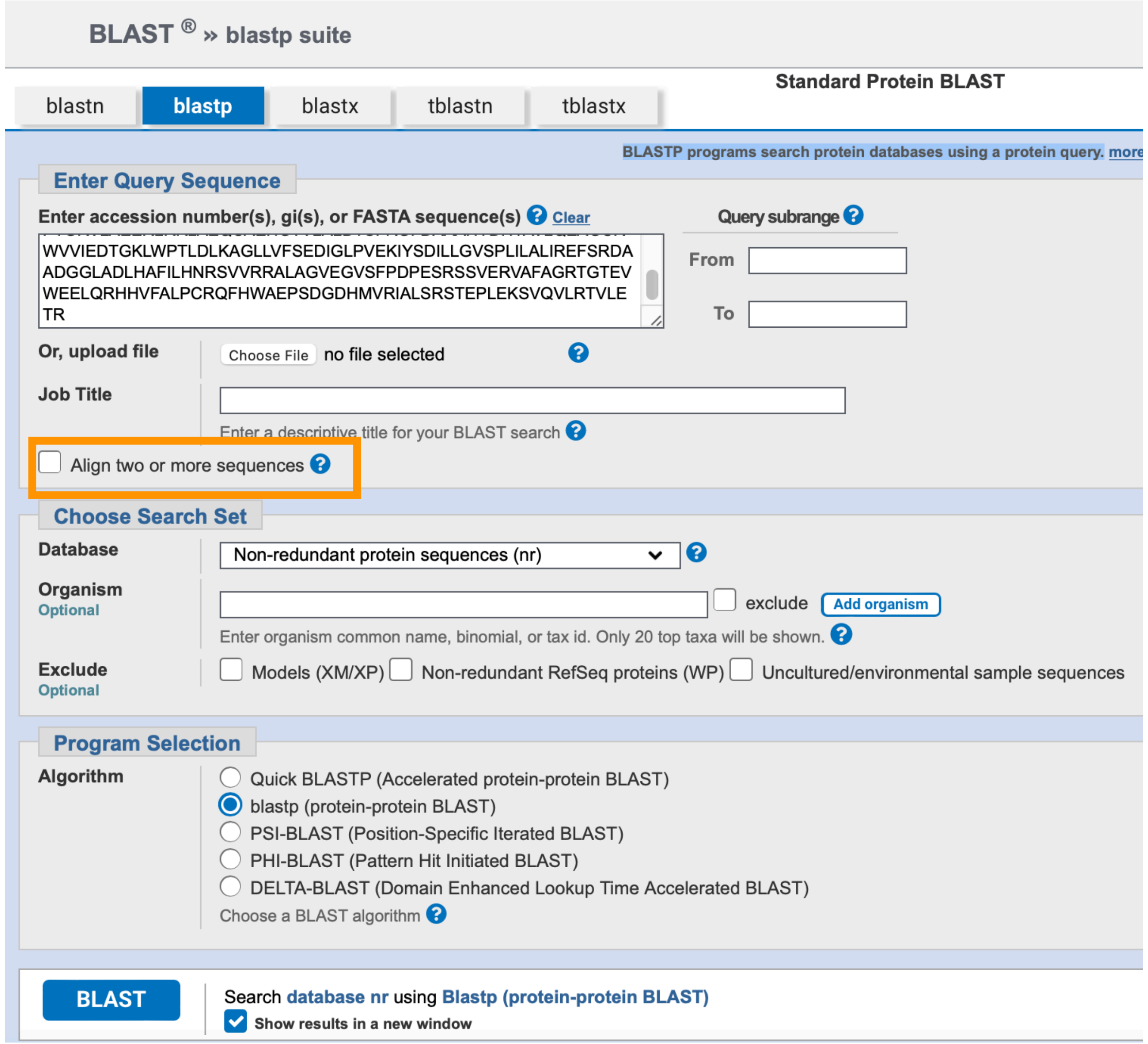
Submit

Basic Local Alignment Search Tool (BLAST)

- The BLAST finds regions of local similarity between sequences.
- The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.
- BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families in the databases.
- Website: https://blast.ncbi.nlm.nih.gov

- Introduction to BLAST homepage: https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf

Basic Local Alignment Search Tool (BLAST)



BLAST ® » blastp suite

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
WVVIEDTGKLWPTLDLKAGLLVFSEDIGLPVEKIYSDILLGVSPLILALIREFSRDA
ADGGLADLHAFILHNRSVVRRALAGVEGVSFPDPESRSSVERVAFAGRTGTEV
WEELQRHHVFALPCRQFHWAEPSDGDHMVRIALSRSTEPLEKSVQVLRTVLE
TR
```

Query subrange
From
To

Or, upload file: Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: ☐ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

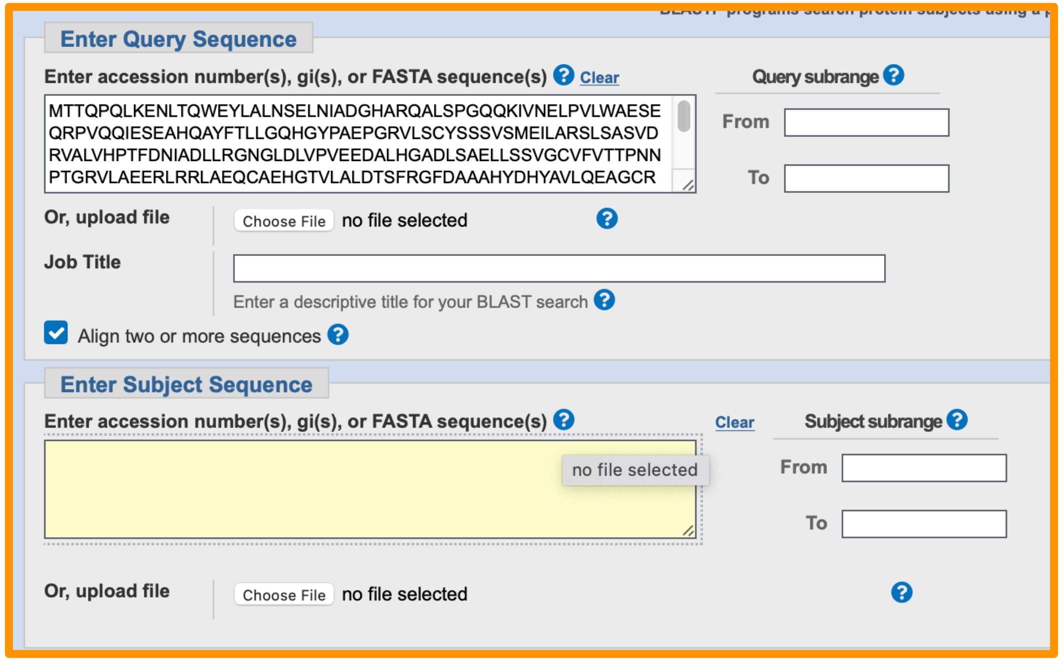
Algorithm

○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database nr using Blastp (protein-protein BLAST)

☑ Show results in a new window



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
MTTQPQLKENLTQWEYLALNSELNIADGHARQALSPGQQKIVNELPVLWAESE
QRPVQQIESEAHQAYFTLLGQHGYPAEPGRVLSCYSSSVSMEILARSLSASVD
RVALVHPTFDNIADLLRGNGLDLVPVEEDALHGADLSAELLSSVGCVFVTTPNN
PTGRVLAEERLRRLAEQCAEHGTVLALDTSFRGFDAAAHYDHYAVLQEAGCR
```

Query subrange
From
To

Or, upload file: Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

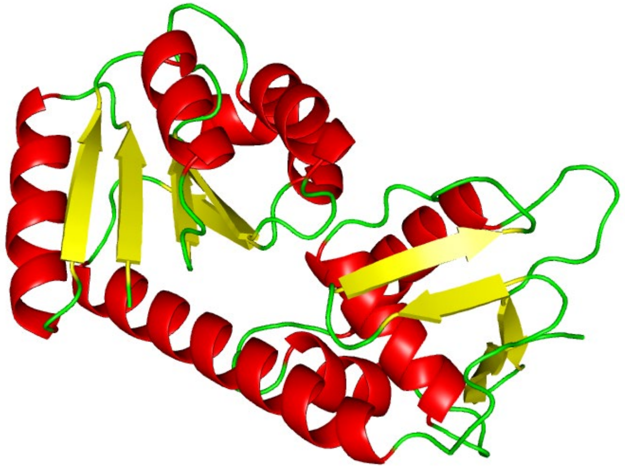
no file selected

Subject subrange
From
To

Or, upload file: Choose File no file selected

From Sequence to Structure: Secondary Structure

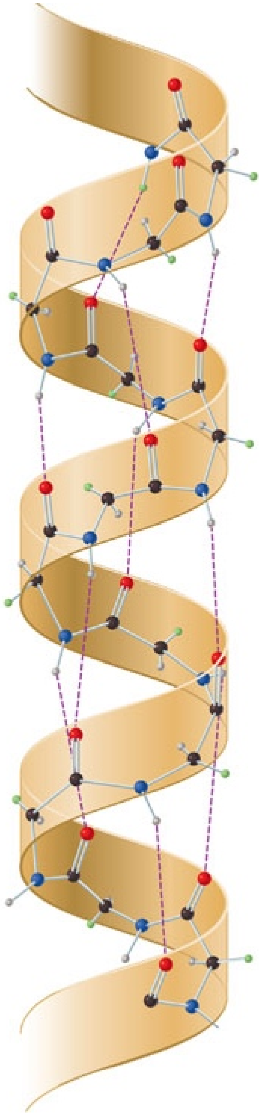
- When the primary sequence of the polypeptide folds into regularly repeating structures, **secondary structure** is formed
- Secondary structure results from *hydrogen bonding between the amide hydrogens and carbonyl oxygens* of the peptide bonds

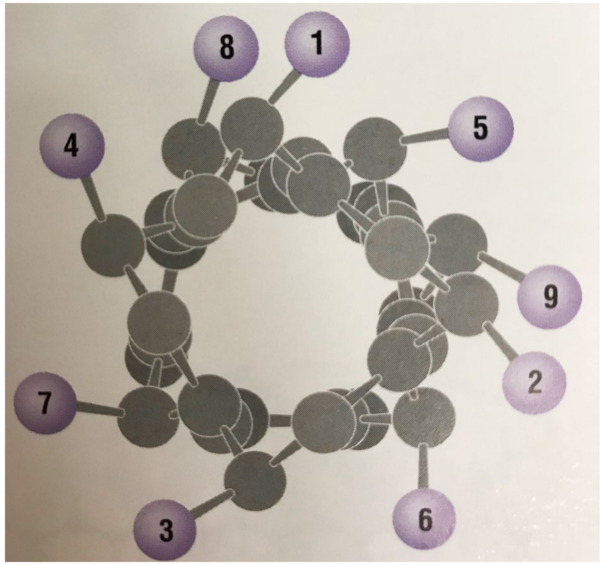


- Common secondary structures:
  - α-helix
  - **β strand**
- Not all regions have a clearly defined secondary structure, some are random or nonregular, e.g., turns, loops.

α-Helix

- Amino acids coil “upward”
  - Forms a right-handed helix
  - 3.6 amino acids per turn
  - Repeat distance or pitch is 5.4 angstroms(Å)
- H bonds between $C{=}O_n$ ----- $H\text{-}N_{n+4}$
  - Within the structure, each amide and carbonyl forms a hydrogen bond.
  - All the hydrogen bonds are parallel along the helical axis
- Amino acid side chains stick “out” of the helix.





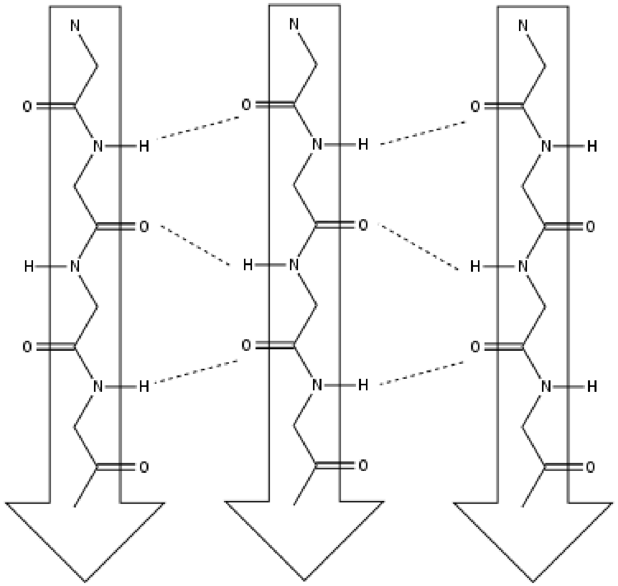
Top view

β Sheet

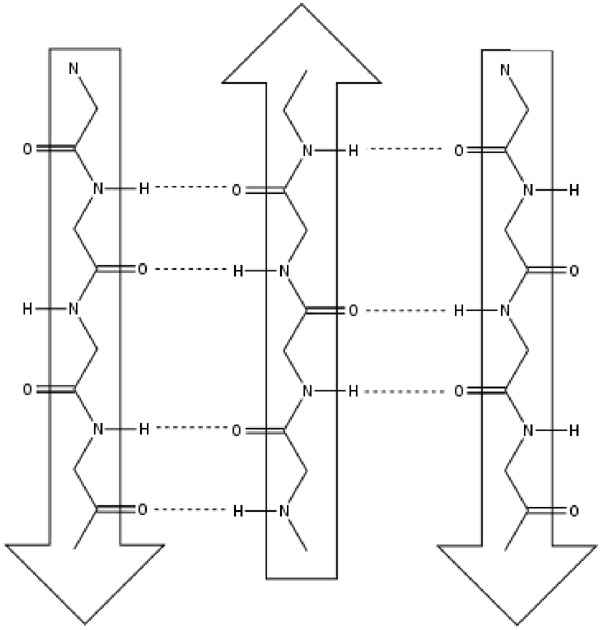
- All of the *carbonyl O and amide H* are involved in the H bonds with the chain nearly completely extended

- Two possible orientations
  - **Parallel** if the N-termini of β strands are head-to-head
  - **Antiparallel** if the N-terminus of one chain is aligned with the C-terminus of the other
    - Common in proteins because it’s more stable.

- β sheets are never flat. They always right-handed twist.



Parallel b-sheet



Antiparallel b-sheet

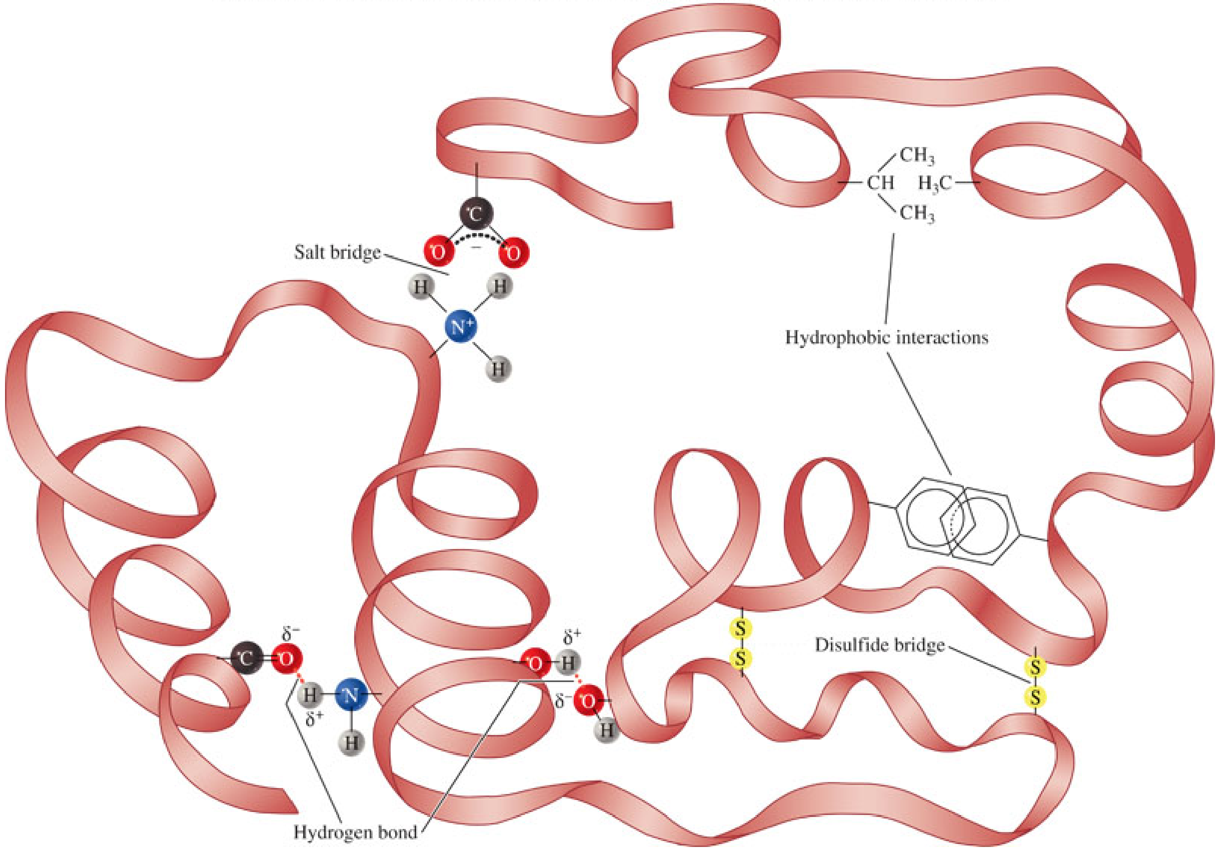
https://upload.wikimedia.org/wikipedia/commons/thumb/9/9e/Beta_sheets.svg/880px-Beta_sheets.svg.png

From Sequence to Structure: Tertiary Structure

- In a folded protein, the secondary structure elements fold into a compact form is called the **tertiary structure**

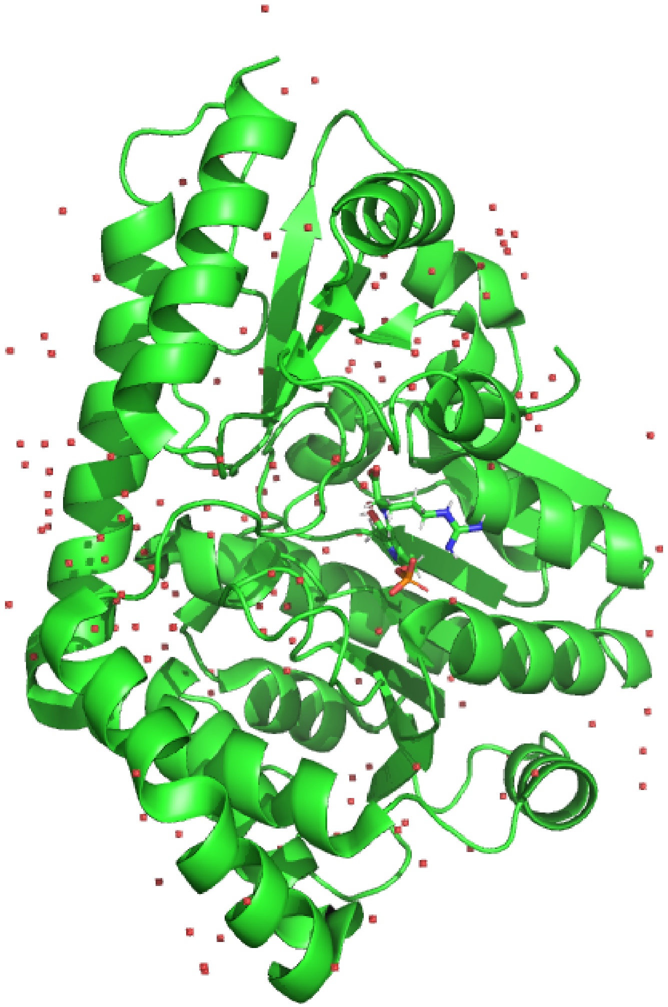
- Occurs due to interactions ***between amino acid <u>side chains</u>***.
  - **Disulfide bridges** between two cysteine residues
  - Non-covalent interactions
    - **Salt bridges** between ionic side chains $-COO^-$ and $-NH_3^+$
    - **Hydrogen bonds** between H donor and acceptor
    - **Hydrophobic interactions (van der Waals interaction)**: two nonpolar groups are attracted by a mutual repulsion of water

Interactions Involved in Tertiary Structure



- Weaker interactions in proteins are dynamic. They form and break constantly, giving protein a globular appearance of an “oil droplet” in a solution.
  - Hydrophobic residues are buried inside
- Often these movements are vital for ligand binding and catalytic function.

Bound water molecules of a folded protein are an important part of the structure



- May be found on the surface
- Bond to end caps of secondary structure

- Dissociate and exit only to satisfy the H bonding or salt bridges requirements of residues that will ultimately contact ligands
- Form water network to assist in acid/base chemistry

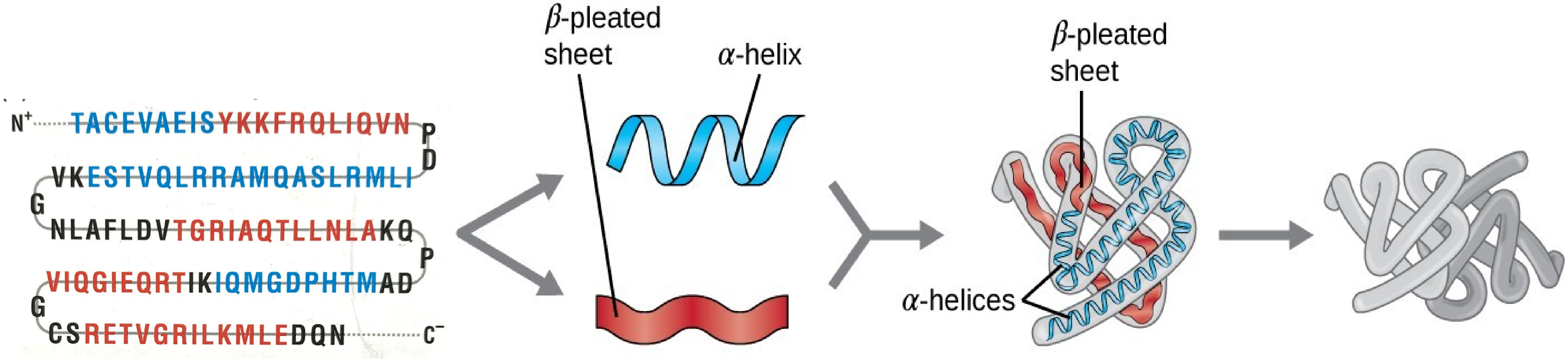
The Quaternary Structure of Proteins

- When more than one polypeptide come together to make a functional protein.
- **Quaternary structure**: the arrangement of subunits or peptides that form a larger protein
  - **Subunit** is a polypeptide chain having primary, secondary, and tertiary structural features
  - Subunits can be identical polypeptides, nearly identical polypeptides or very different polypeptides.
    - One chain: monomer
    - Two chains: dimer (same chains: homodimer; different chains: heterodimer)
    - Trimer, etc.

Types of Interactions in Quaternary Structure

- Quaternary structure is maintained by the same forces which are active in maintaining tertiary structure
  - **Disulfide bridges** between two cysteine residues
  - Non-covalent interactions
    - **Salt bridges** between ionic side chains $-COO^-$ and $-NH_3^+$
    - **Hydrogen bonds** between H donor and acceptor
    - **Hydrophobic interactions (van der Waals interaction)**: two nonpolar groups are attracted by a mutual repulsion of water

Proteins Structures - Overview



Primary Protein Structure
Sequence of a chain of amino acids

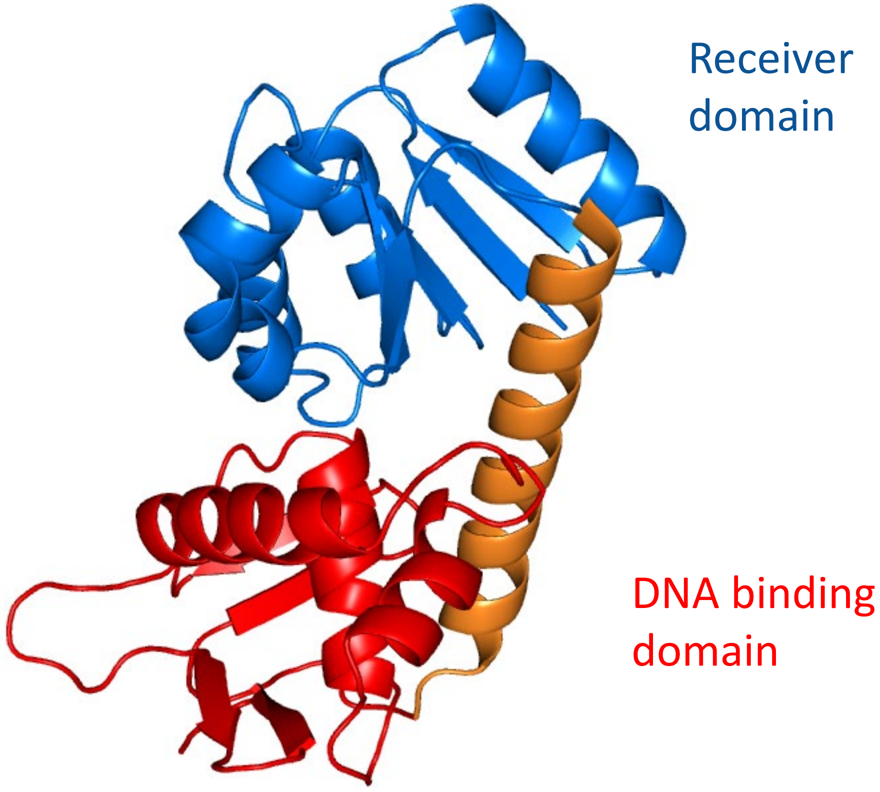
Secondary Protein Structure
Local folding of the polypeptide chain into helices or sheets

Tertiary Protein Structure
three-dimensional folding pattern of a protein due to side chain interactions

Quaternary Protein Structure
protein consisting of more than one amino acid chain

Domains and Motifs

- **Domains**: part of the sequence that appear as conserved modules in proteins that are not related, in global terms.
    - Usually with a distinct three-dimensional fold, carrying a unique function and appearing in different proteins



RitR mutant C128D
PDB: 5VFA

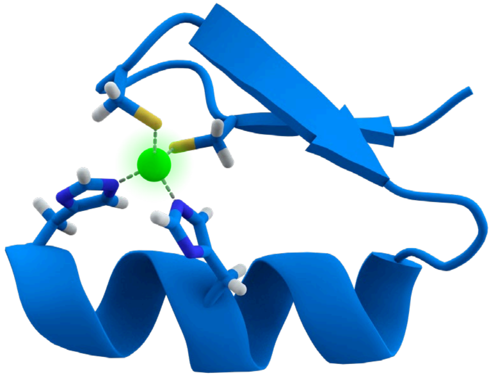
Domains and Motifs

- **Motifs** are commonly used in two ways.

1. Short conserved sequences, which appear in a variety of other molecules.

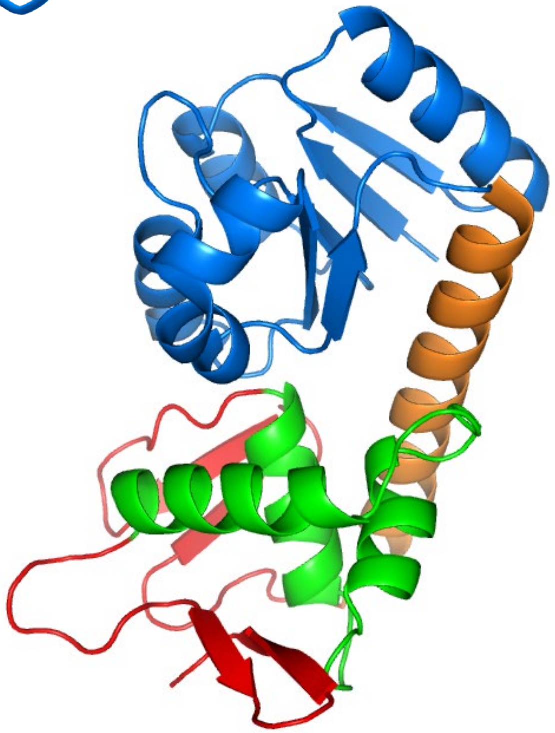
   e.g., Zinc finger motif in many DNA binding proteins

   CXX(XX)**C**XXXXXXXXXXXX**H**XXX**H**.



2. A set of contiguous secondary structure elements that either have a particular functional significance or define a portion of an independently folded domain.

   e.g., Helix-turn-helix motif in DNA binding domain

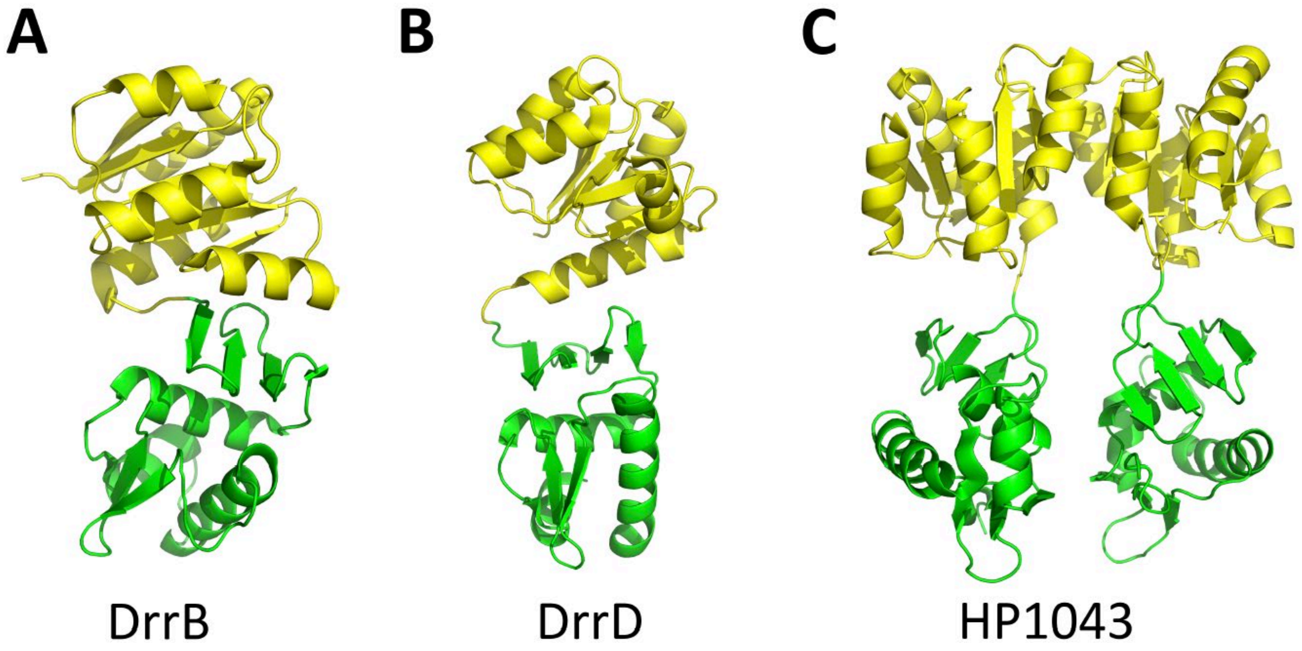


RitR mutant C128D
PDB: 5VFA

Top figure: By Thomas Splettstoesser (www.scistyle.com) - self-made, based on PDB structure 1A1L, the open source molecular visualization tool PyMol and Cinema 4D, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=3106866

Proteins can be grouped into families based on the domains they contain

- Structural alignment: a method for discovering significant structural motifs.
- But often it is not the case that structural families share a common function.



Structural comparison of various classes of full-length OmpR/PhoB subfamily structures. Receive domain (REC domain) colored in yellow, DNA binding domain colored in green.

Protein Modification

- Most proteins start folding during the translation process!
- Many proteins require post-translational modifications:
  - Some proteins need to be cleaved to be active: zymogens, insulin
  - Some proteins require cofactors and/or form complexes
  - Some require oxidation of cysteines to form disulfide bonds.
  - Some require carbohydrate or lipid modifications
  - etc.

Stability of Proteins

- Proteins must fold to a globular conformation to carry out the most important tasks in living organisms

- The unfolding of proteins is called ***denaturation***.
  - Temperature
  - pH
  - Organic Solvents/Detergents
  - Heavy Metals
  - Mechanical Stress

- Peptide bonds between amino acids can be broken through ***hydrolysis***.
  - This changes the primary structure!

  - Strong acid or strong base
  - Certain enzymes

Methods for Stabilization of Proteins

- Correct pH (buffer)
- Maintain temperatures (usually low)
- Minimize processing times
- Minimize agitation
- Minimize denaturing chemicals
- Add reducing agents (Oxidation can cause inactivation typically intracellular proteins)
- Add protease inhibitors
- Add stabilizer

Examples of Stabilizers

A. These reduce free water levels by hydrogen bonding with H_2O

- Glycerol
  - Sugar
    - Polyethene Glycol

B. BSA (Bovine serum albumin)

- added to proteins which are at LOW concentration

Storage of Proteins

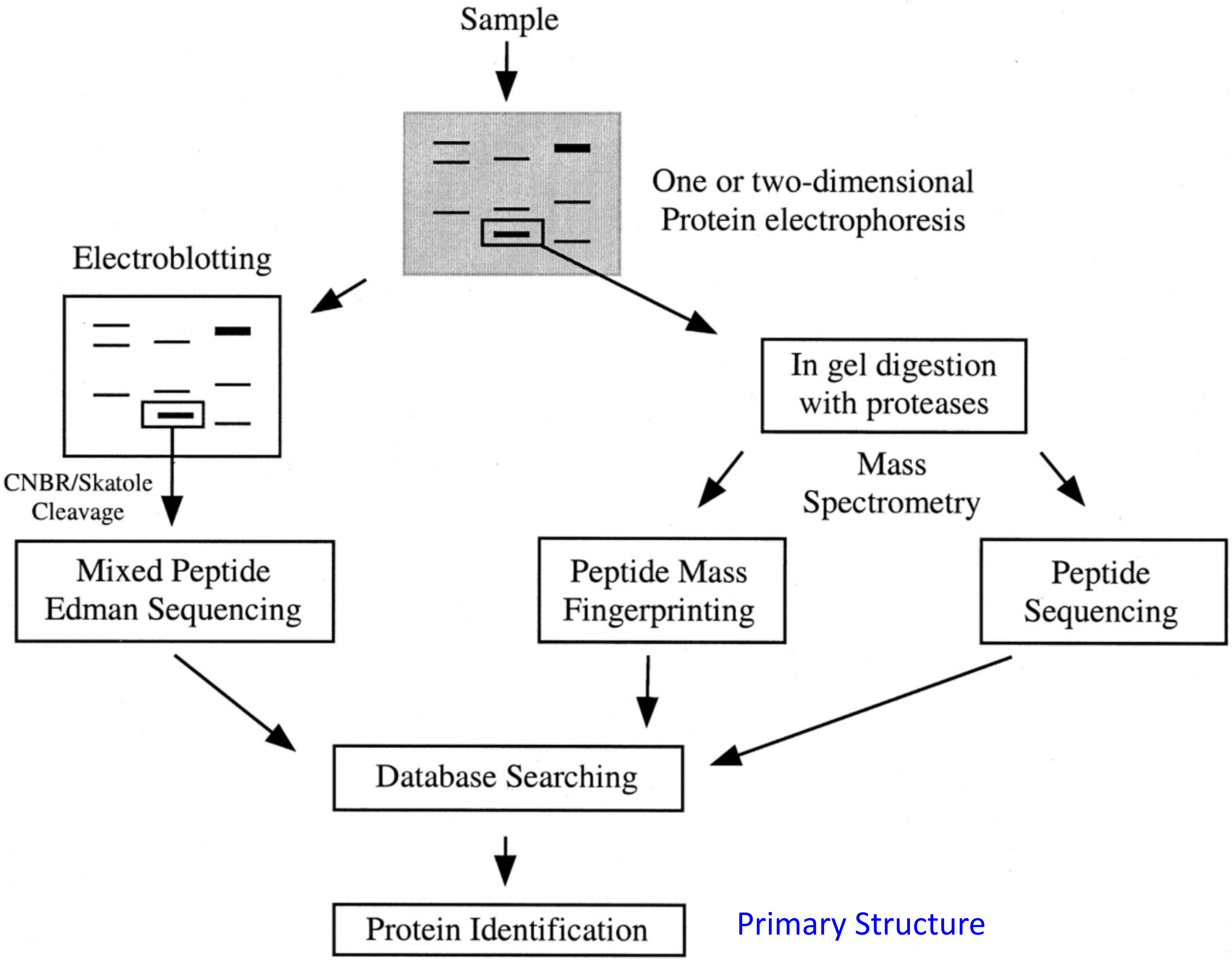
- Similar conditions apply as with Stability
- Freezing with Liquid N_2 (and thaw) is typically OK
- Lyophilization
- Note: Concentration of contaminants

Lyophilization Freeze Dryer System

Drying of protein

- Freeze protein
- Increase temperature
- Apply vacuum
- Remove water vapor

Strategies for Protein Identification:



Sample

One or two-dimensional Protein electrophoresis

Electroblotting

CNBR/Skatole Cleavage

Mixed Peptide Edman Sequencing

In gel digestion with proteases

Mass Spectrometry

Peptide Mass Fingerprinting

Peptide Sequencing

Database Searching

Protein Identification

Primary Structure

Graves PR, Haystead TA. Molecular biologist's guide to ,
Rev. 2002;66(1):39-63. doi:10.1128/MMBR.66.1.39-63.2002

Protein Identification: Primary Structure

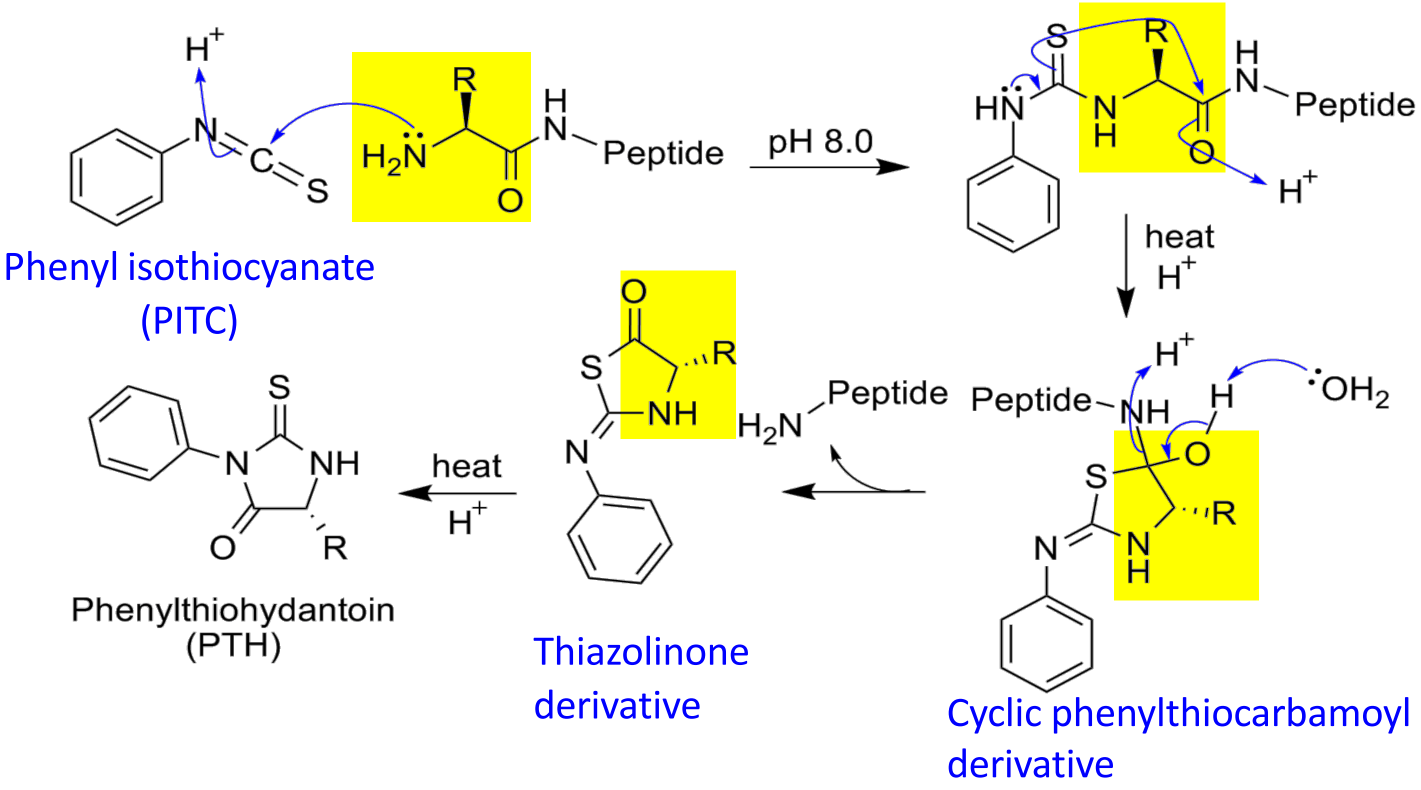
Edman Sequencing

- Obtain N-terminal amino acid sequences
- A variable alternative to MS
- Determine the true start amino acid of a protein

Mass Spectrometry

- Obtain peptide masses or amino acid sequences, and thus identify the protein by searching databases
- Determine the type and location of protein modifications

Edman Sequencing



Works with 30 aa fragments (in practical)

Cut protein with cyanogen bromide (CNBr)

- cuts on carboxyl side of Met

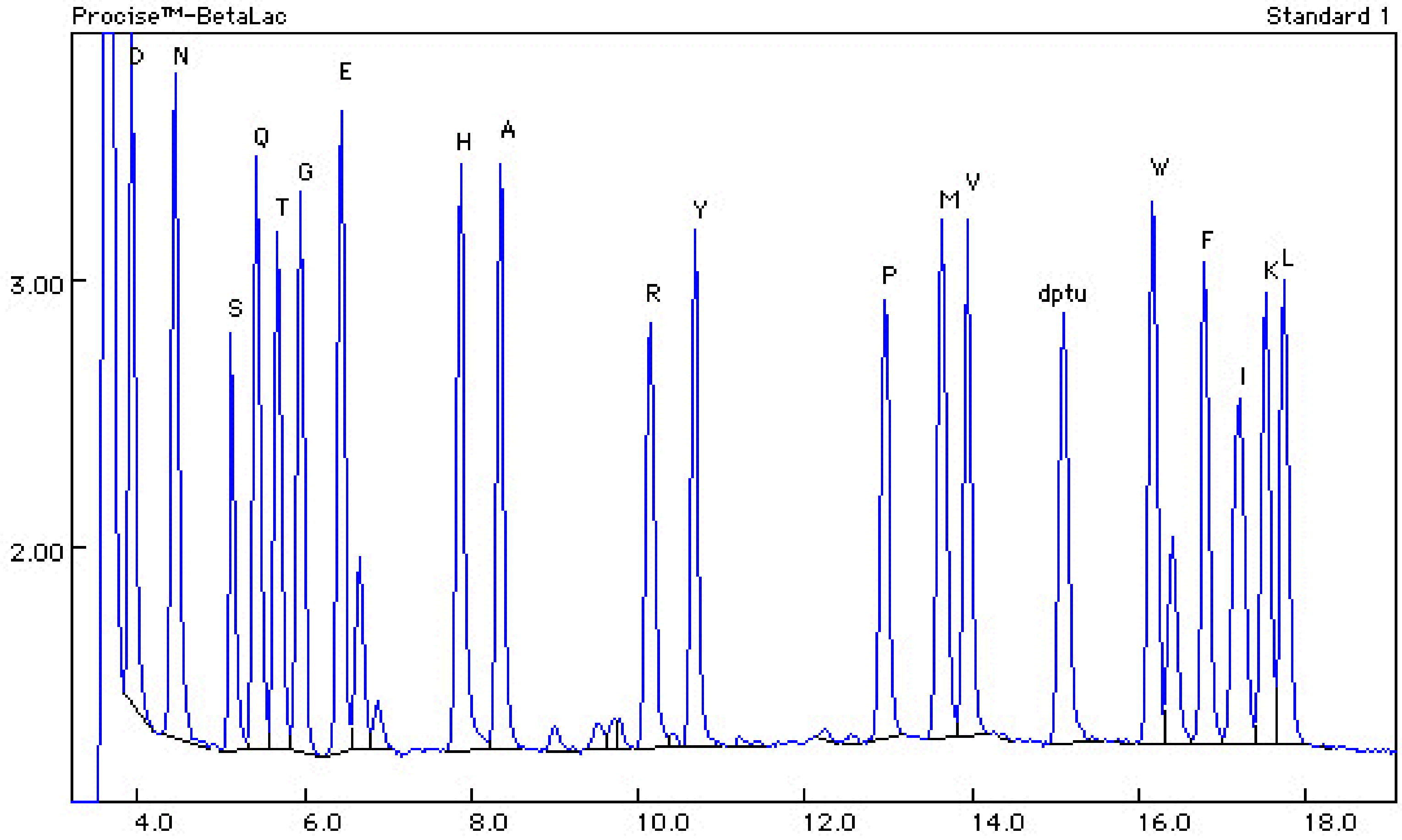
Phenylisothiocyanate (PITC)

- binds to and releases N-terminal residue

Chromatography against known standards

Reagent	**Cleavage Properties**
Chemical agents	
70% formic acid	Asp-↓-Pro
Cyanogen bromide in 70% formic acid	Met-↓
2-nitro-5-thiocyanobenzoate, pH 9	↓-Cys
Hydroxylamine, pH 9	Asn-↓-Gly
Iodobenzoic acid in 50% acetic acid	Trp-↓
Endoproteases	
Trypsin	Arg/Lys-↓
Lys-C	Lys-↓
Arg-C	Arg-↓
Glu-C (bicarbonate)	Glu-↓
Glu-C (phosphate)	Asp/Glu-↓
Asp-N	↓-Asp
Chymotrypsin	Phe/Tyr/Trp/leu/Met-↓ (also Ile/Val-↓)

HPLC Chromatogram of Known Standard Amino Acids



Retention time

Mass Spectrometry

Sample Preparation

- 1-D or 2-D gel electrophoresis
- In gel digestion
- Reverse-phase chromatography or HPLC

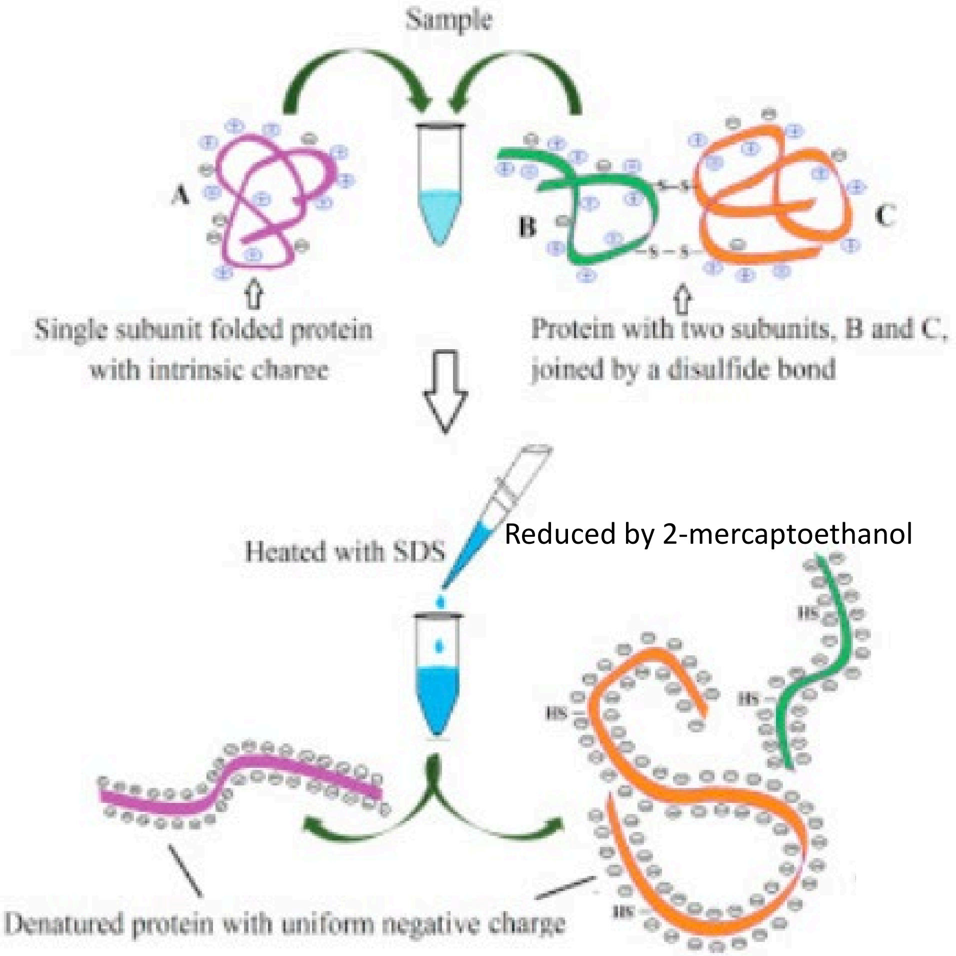
Sample Ionization

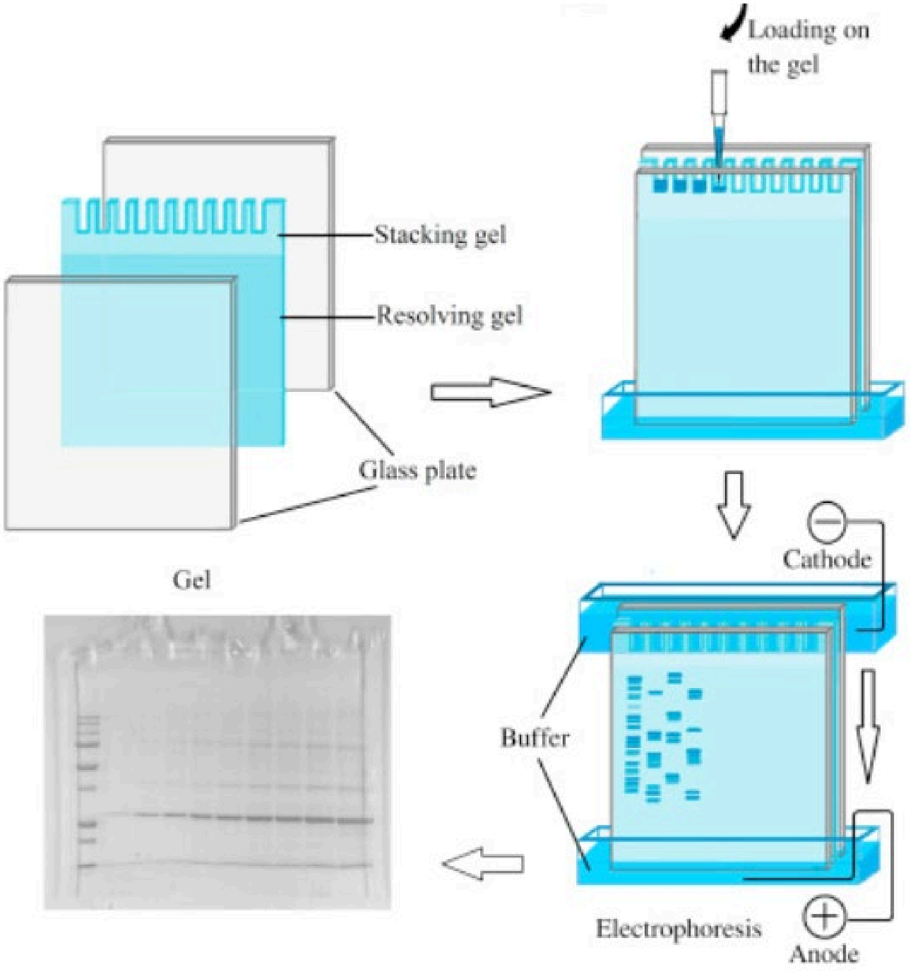
- The sample molecules must be charged and dry
- By addition or loss of one or more protons (H^+)
- Electrospray ionization (ESI) or Matrix-assisted laser desorption/ionization (MALDI)

Separation and Mass Analysis

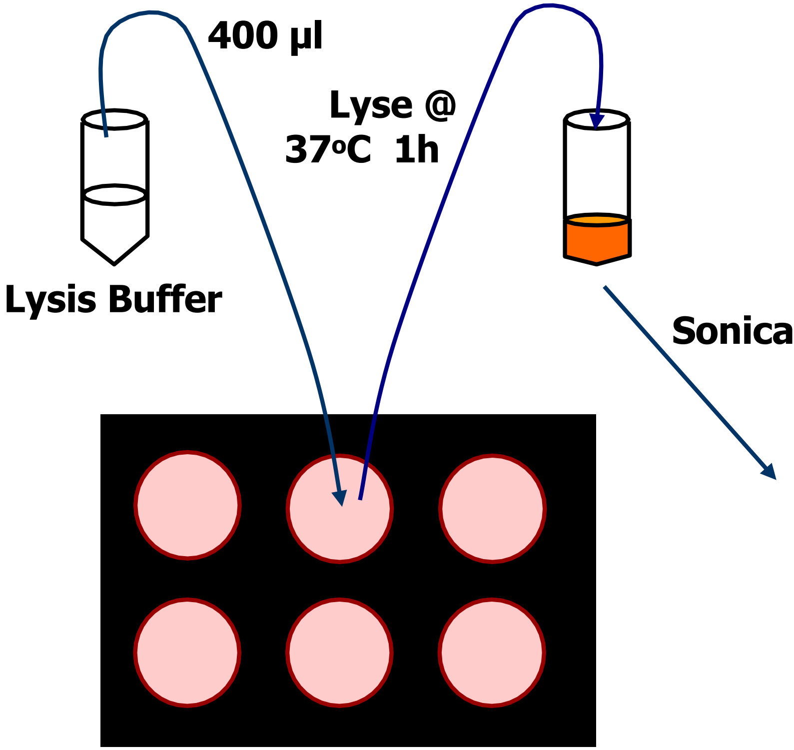
- Resolve the molecular ions by their mass and charge in a vacuum
- Quadrupole mass analyzers or Time of flight (TOF) or Ion trap

SDS-PAGE (sodium dodecyl sulphate–polyacrylamide gel electrophoresis)

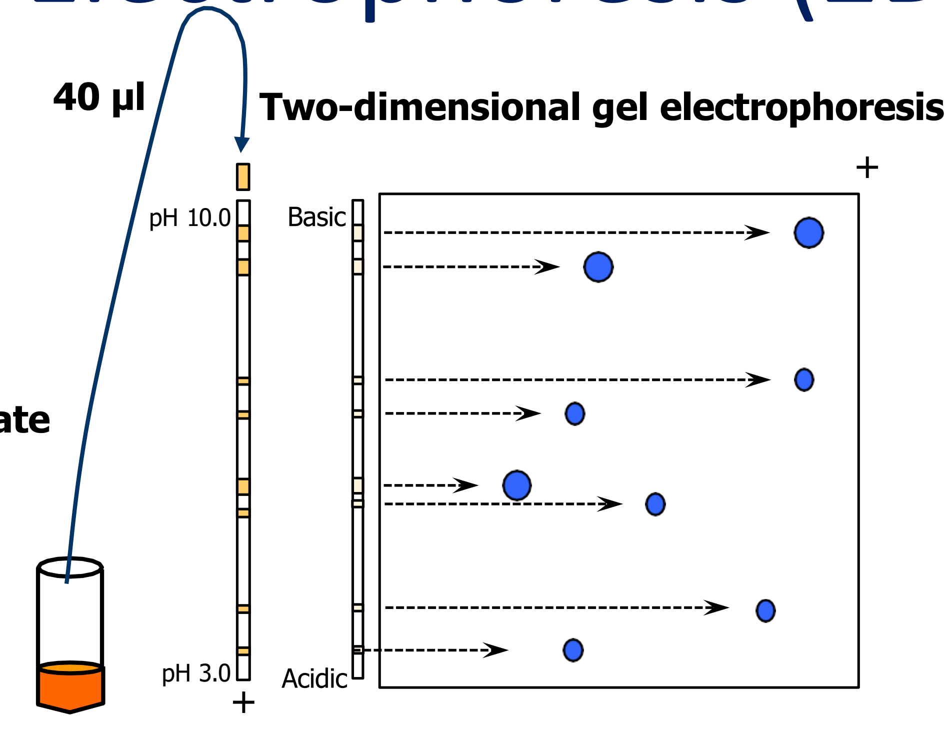


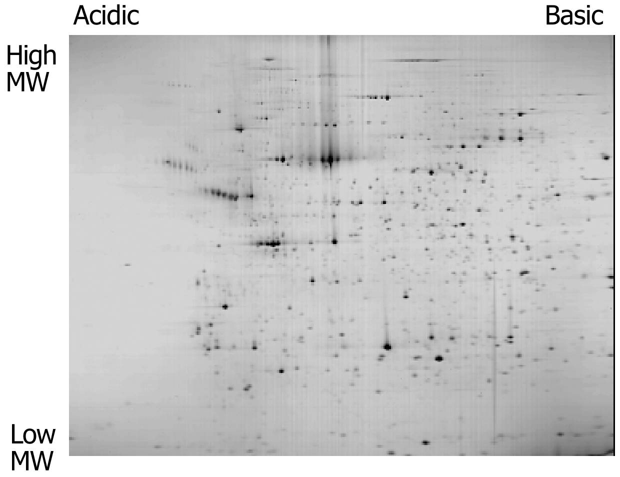


Two-dimensional Gel Electrophoresis (2DE)



Primary hepatocyte cell culture
(~1 x 10^6 cells/well)





One gel per well

Slides from Dr. Mu Wang

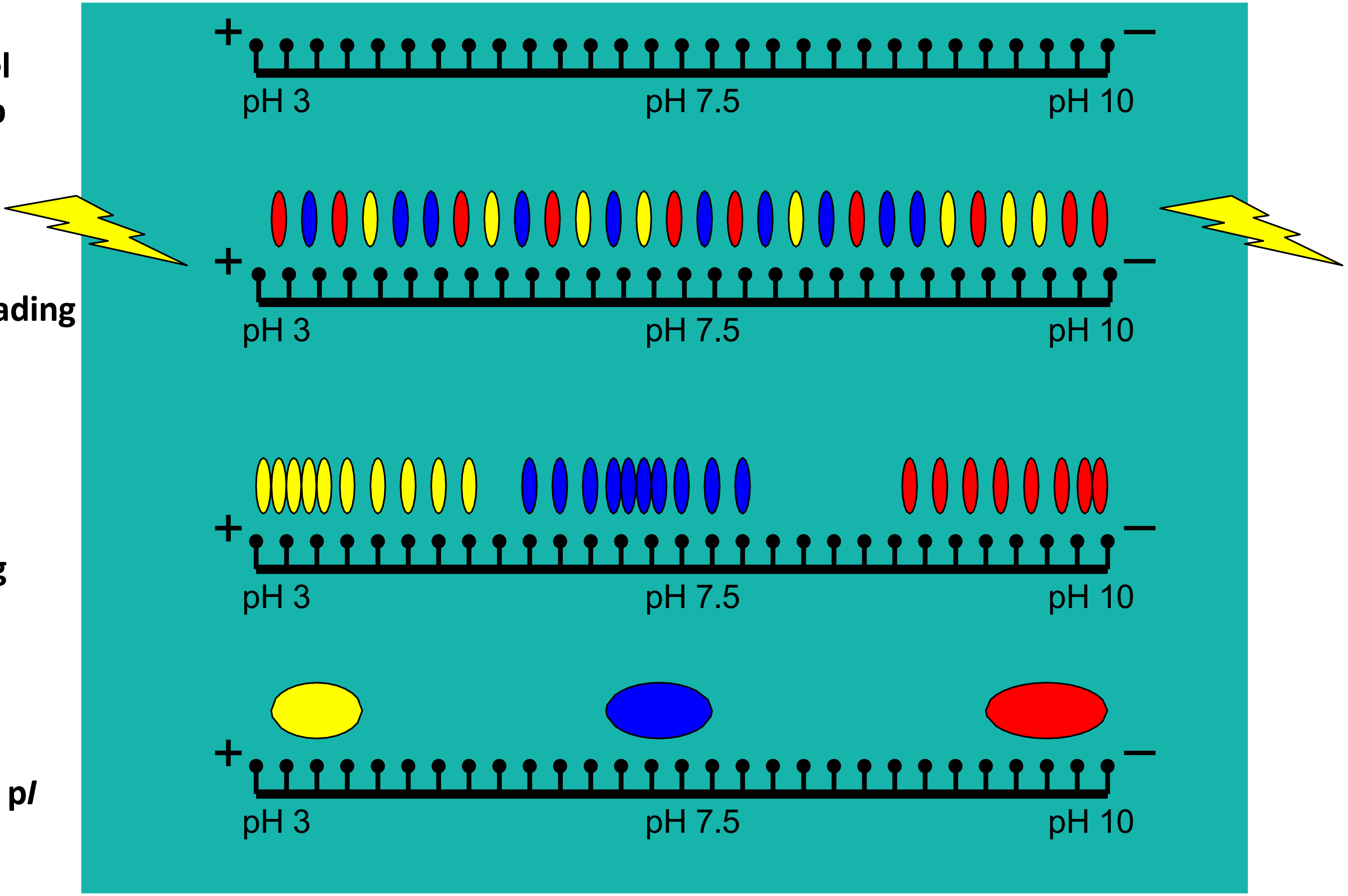
First Dimension: Isoelectric Focusing

IEF tube gel or IPG strip

Sample Loading

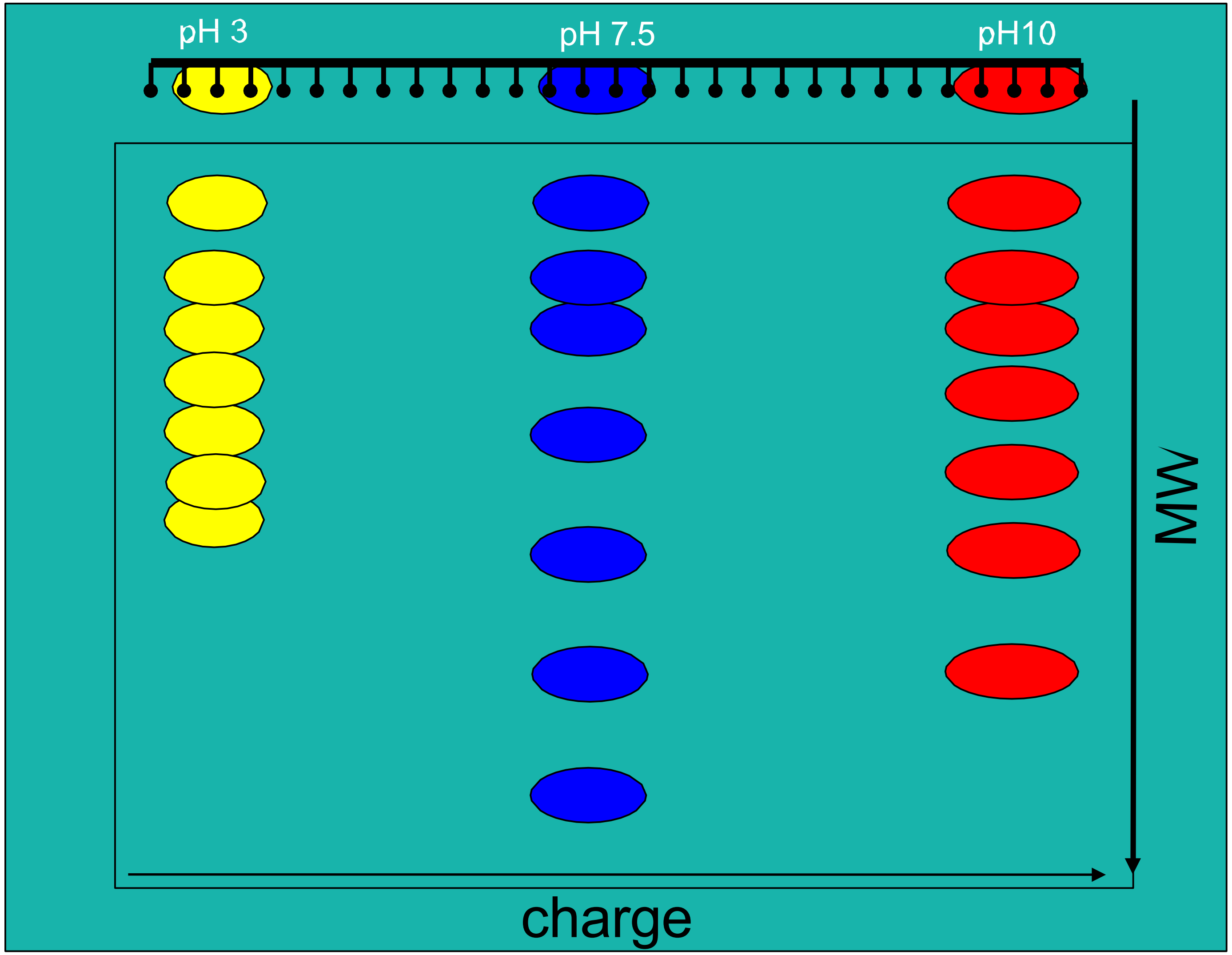
Focusing

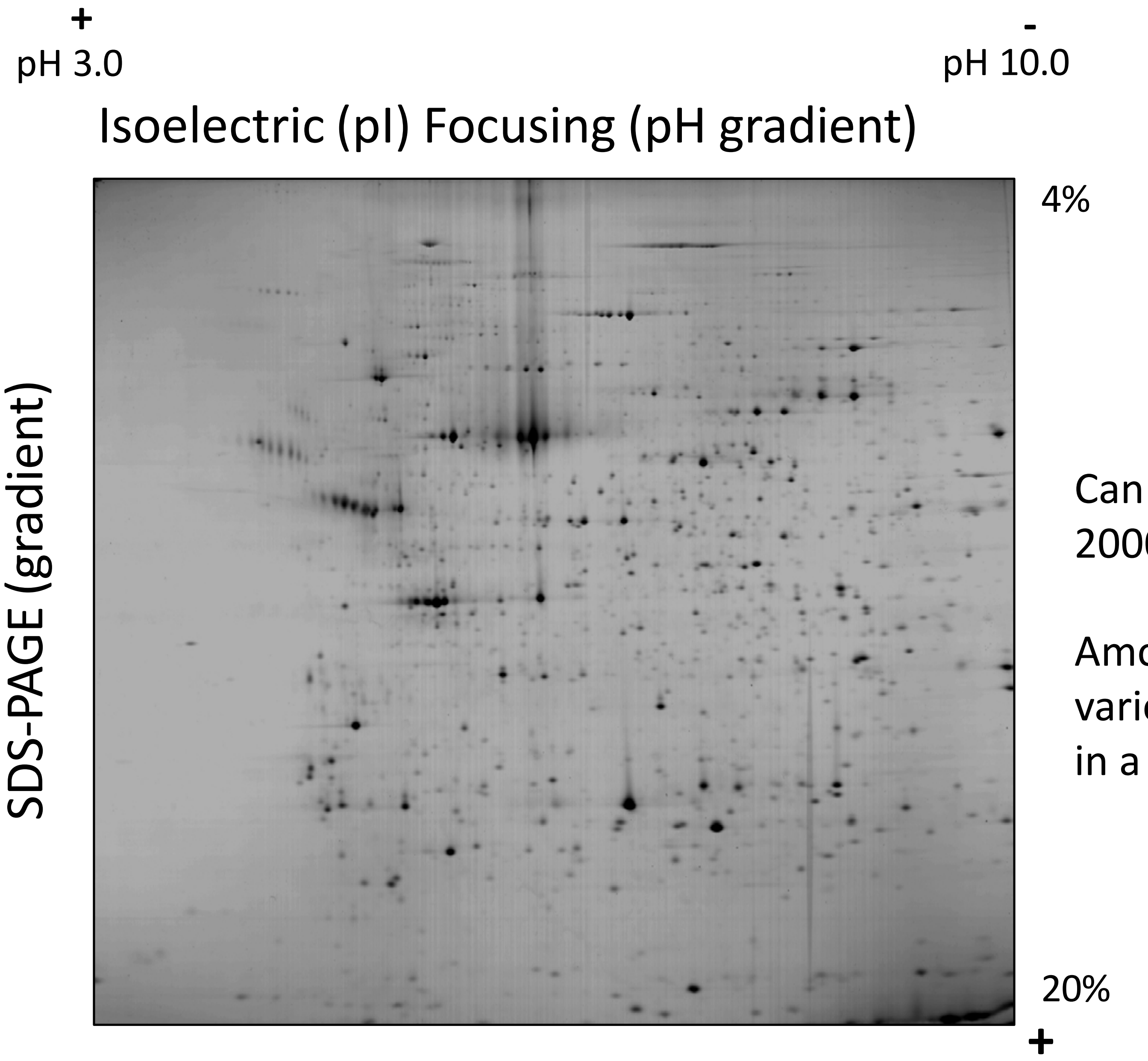
Focused at p*I*



Slides from Dr. Mu Wang

Second Dimension: SDS-PAGE





Can resolve up to 2000+ proteins /gel

Amount of proteins varies up to 10^9-fold in a cell lysate.

Slides from Dr. Mu Wang

Application and Weakness of 2DE

- **Broad-based Screening of Protein Expression**
  - up- and down-regulation
  - post-translational modification
  - protein identification

- **WEAKNESSES**
  - Labor Intensive and time-consuming (~2 days, only one sample per gel)
  - Many large or hydrophobic proteins can not enter the 1st dimension gel or proteins with extreme acidity or basicity are not well represented.
  - Dynamic range of protein expression is problematic (e.g. low copy proteins)

Two-dimensional Gel Electrophoresis (2DE)

Groups

2D Gels

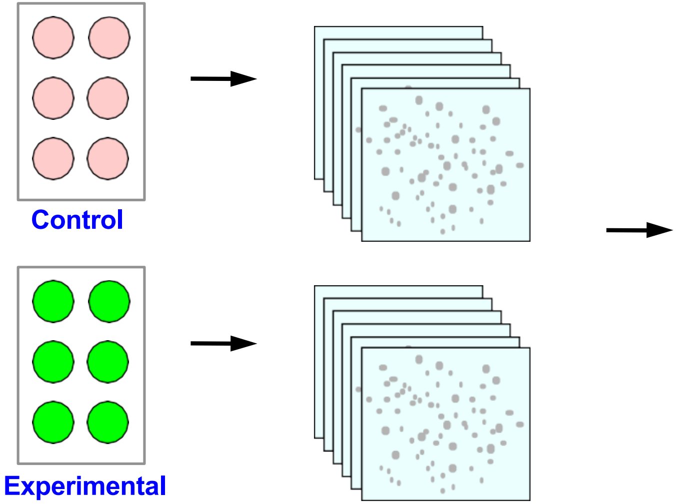
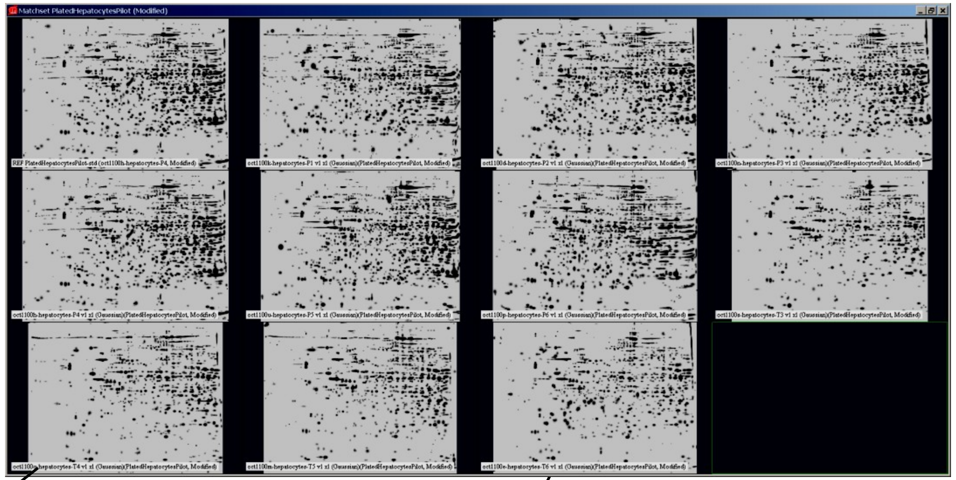
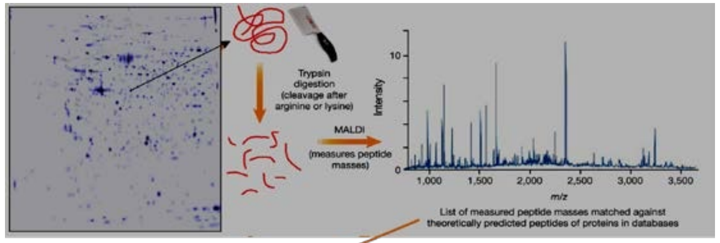
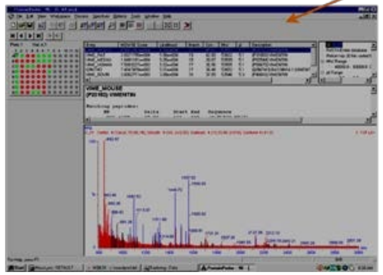
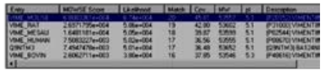


Image Analyses – MatchSet Construction

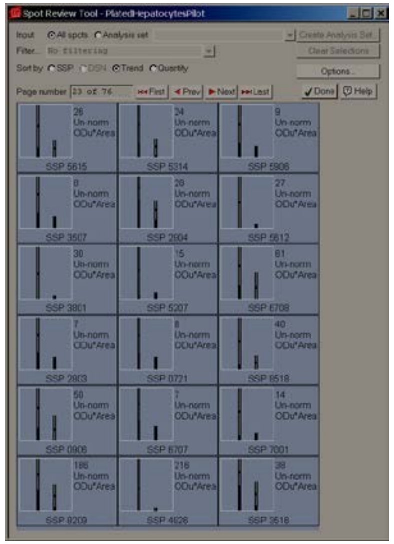


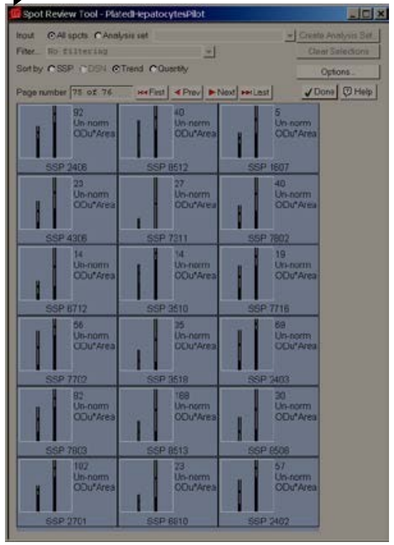






Protein ID





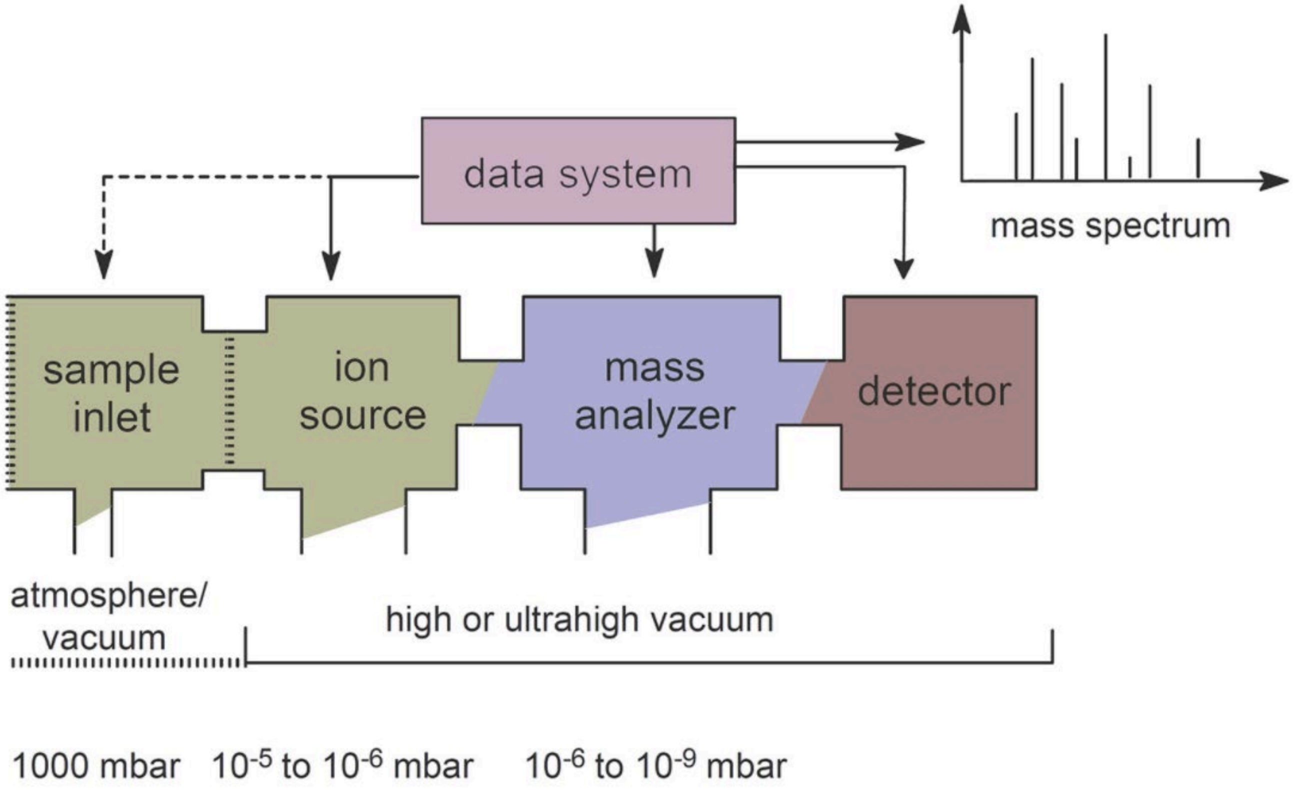
Data Export – e.g. ANOVA

Statistical Analysis of Protein Expression

Mass Spectrometry

- “It is an analytical technique that identifies the chemical composition of a compound or sample based on the **mass-to-charge ratio (m/z)** of ions”
  - Pharmaceutical industry
  - **Metabolite Characterization Studies in Drug Discovery**
  - **Biomarkers - metabonomic approaches**
  - Natural Products
  - Agriculture
  - Food analysis
  - Environmental
  - **Proteomic** and Genomic studies
- *Qualitative analysis*: Structural Characterization, Exact Mass, Molecular weight
- *Quantitative analysis*: Concentration level down to ppt (ng/L) level

General Scheme of a Mass Spectrometer

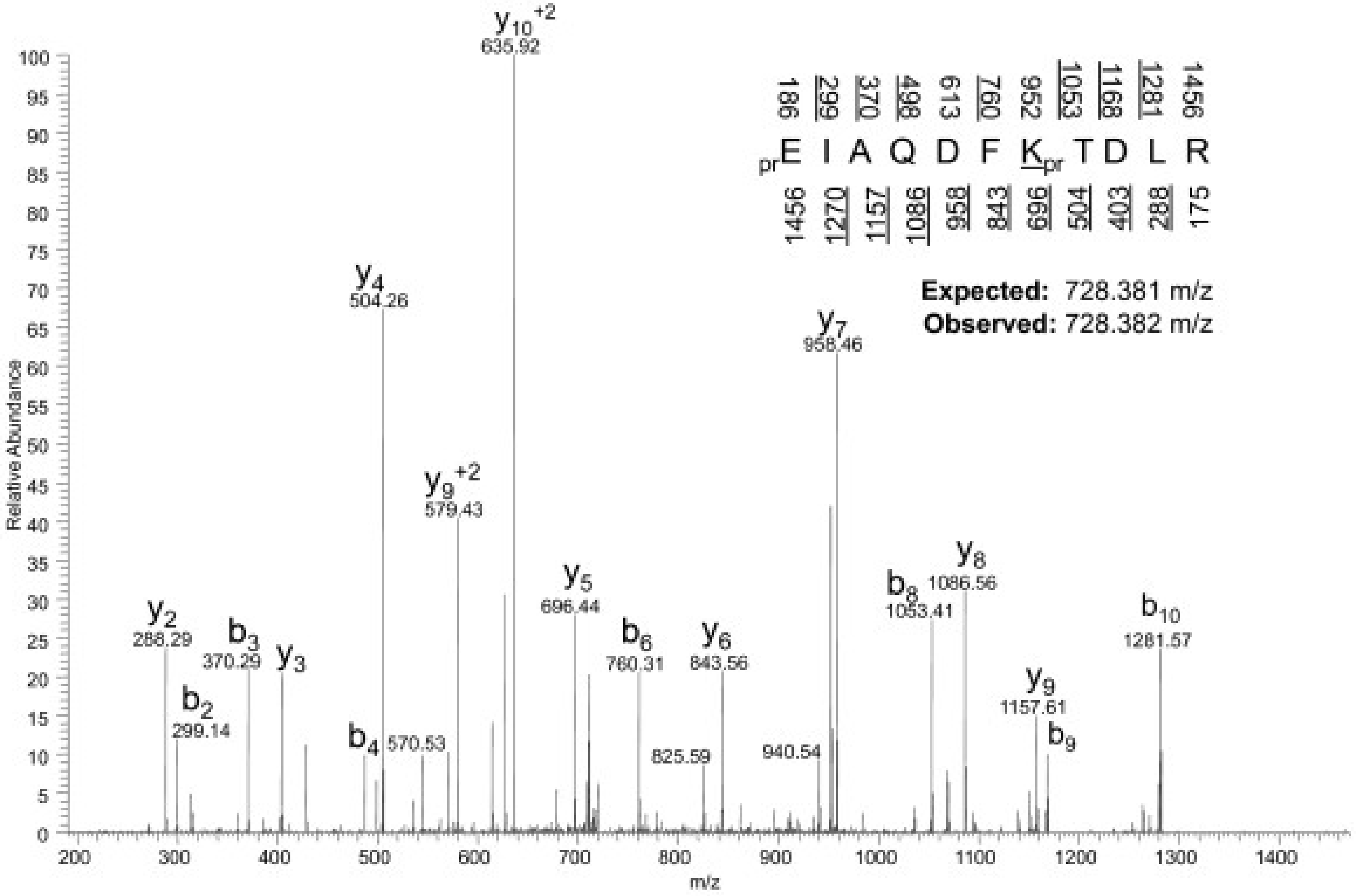


- Mass spectrometry is an extremely **sensitive** method requiring less than 1 µg of sample
- Samples are needed to be transformed into **ions**, for MS to work.
  - Ions can be subjected to acceleration, deceleration, deflection or oscillation
- Ions are delivered in the dilute gas phase

https://dgms.eu/en/about-dgms/this-is-mass-spectrometry/mass-spectrometers/

Mass Spectrum

- A mass spectrum is the two-dimensional representation of ion abundance versus *m/z*.
- Ion abundance is reflected by the signal area, more simply by signal height.
- The numerical value of *m in m/z* equals the numerical value of the ionic mass on the atomic mass scale.
    - For example, the molecular ion of ethanol (CH_3CH_2OH), 46 u, is detected at *m/z* 46 (when z=1).



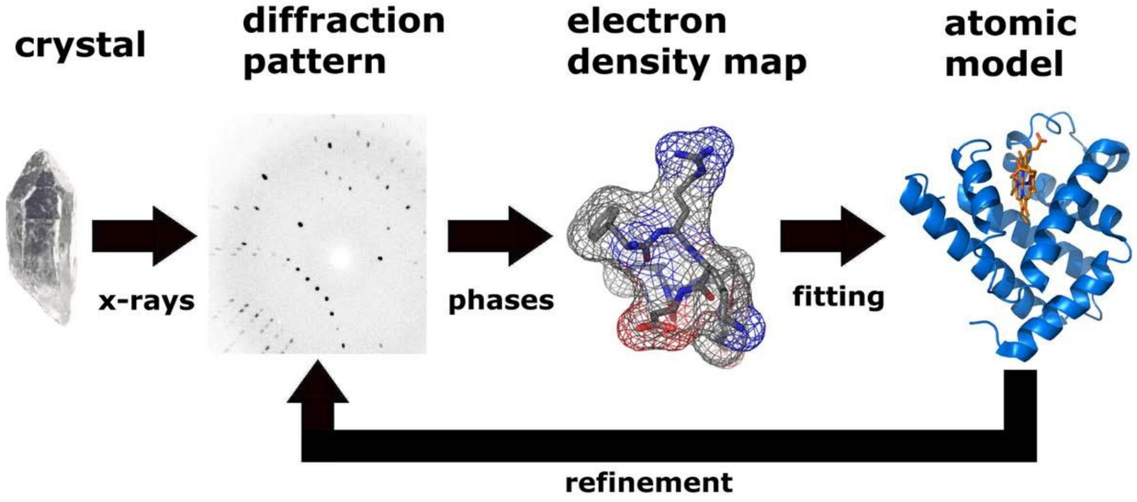
https://www.researchgate.net/profile/Rebecca-Levin-2/publication/49660823/figure/fig2/AS:214315966701573@1428108316420/Tandem-mass-spectrometry-MS-MS-spectrum-of-labeled-lysine-peptide-MS-MS-of-the-H3.png

3D Structure Determination

- X-Ray Diffraction (aka X-ray Crystallography)
  - Crystalize proteins (can be difficult)
  - Pummel them with X-rays
  - Reconstruct their shape from diffracted light
- NMR (Nuclear Magnetic Resonance Spectroscopy)
  - small proteins (<40 kD)
- Cryo-EM (Cryo-Electron Microscopy)
  - Larger molecules (>50 KD)
  - Does not require protein crystals
  - transforming the images into sharper molecular structures.

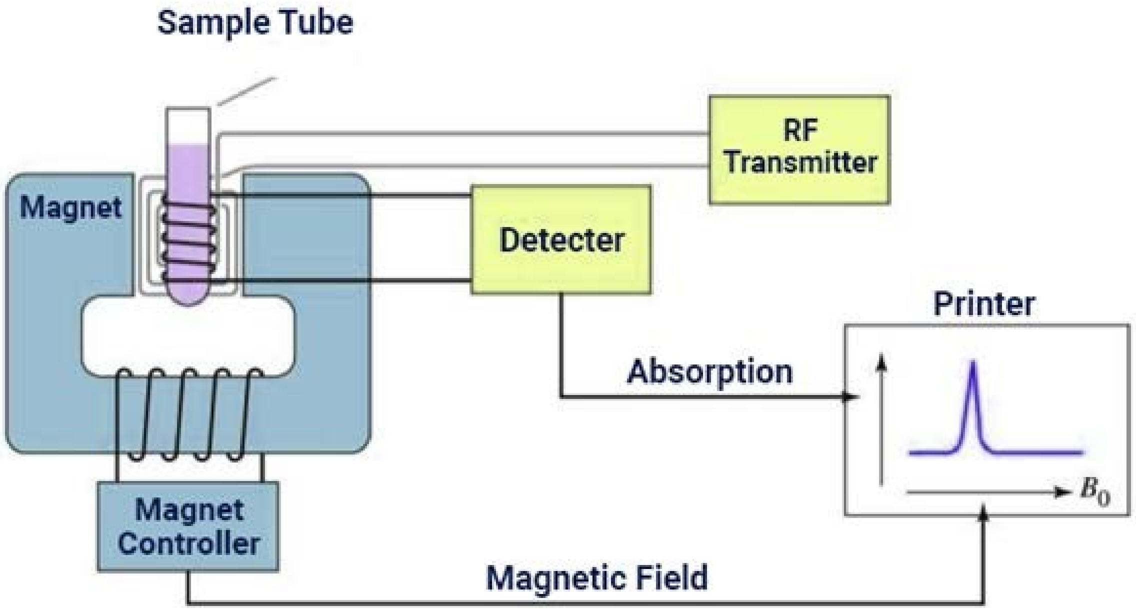
X-ray Crystallography

- Beam of X rays directed at protein
- X-rays are diffracted into different directions by the e- cloud of the atoms
- These beams hit a film detector
- By analysis of the angles and intensities of these diffracted beams
- computer analysis to create ***electron density map***



NMR

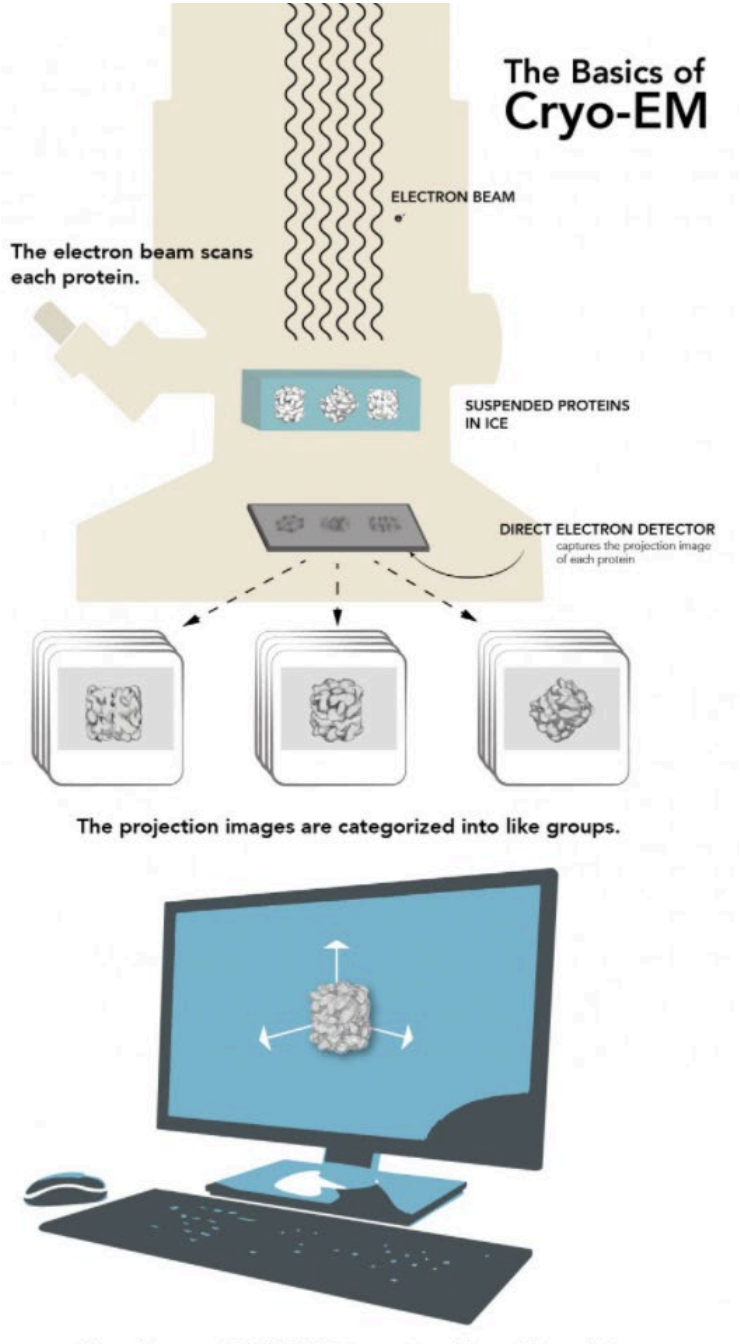
- Apply magnetic field to protein
- atomic nuclei spin - create their own magnetic field
- Pulses perturb the nuclei
- Relax and emit radiation
- Measures the interaction of atomic nuclei (through bonds and distance)



Cryo-EM

- Use beams of electrons rather than light to form images of tiny samples (proteins)
  - >50 kD
  - Highest resolution is ~3 Å
  - Small amount of sample
- Freeze samples to preserve the natural structure of biological specimens and reduce damage from the electron beam
- Generates 3-D images at nearly atomic resolution

- In 2017, three scientists, Jacques Dubochet, Joachim Frank and Richard Henderson, were awarded the Nobel Prize in Chemistry for developing a technique that would image biomolecules



The projection images are categorized into like groups.

Then they are COMBINED to create a hi-res 3D model.

https://cryoem.slac.stanford.edu/what-is-cryo-em