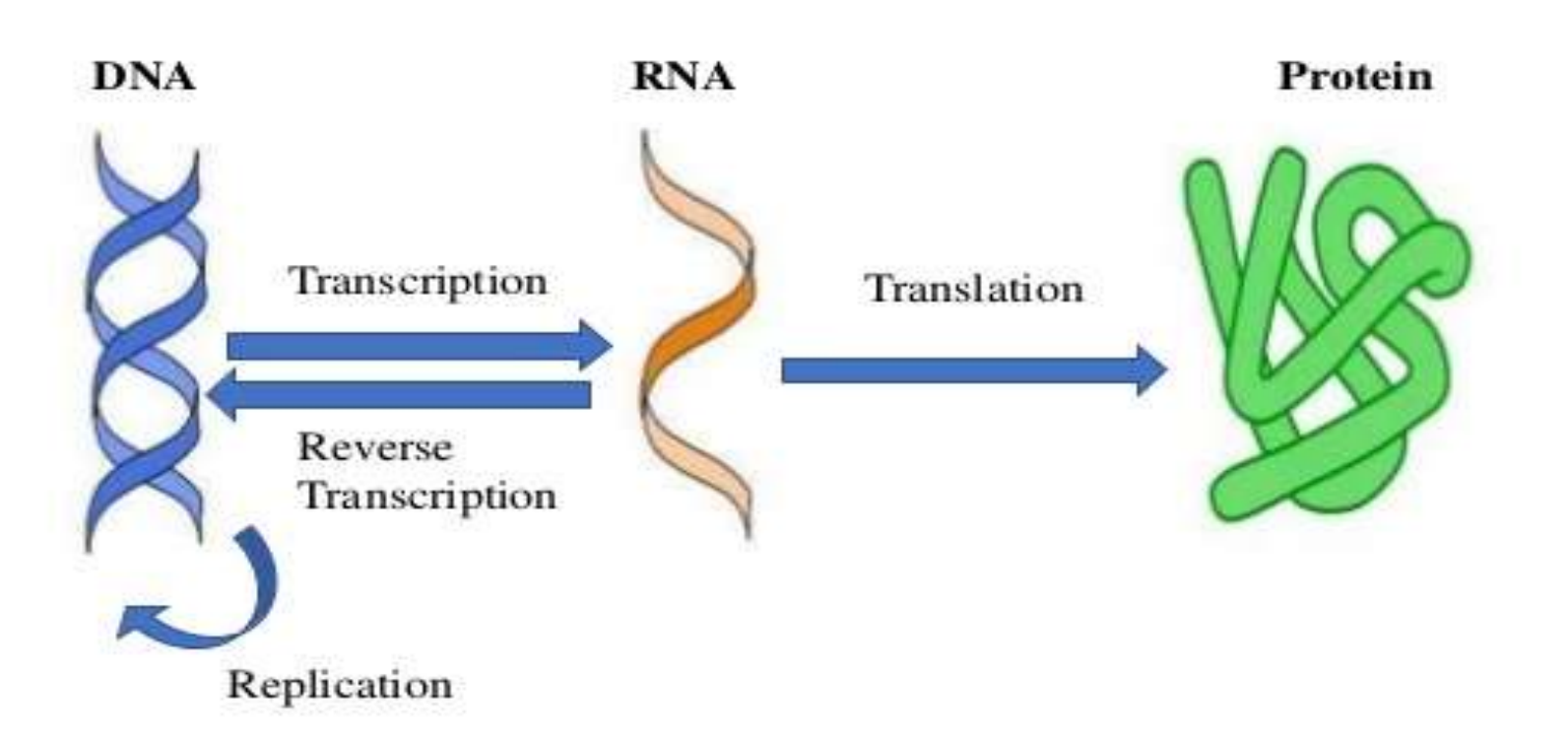
PROTEOMICS

By-

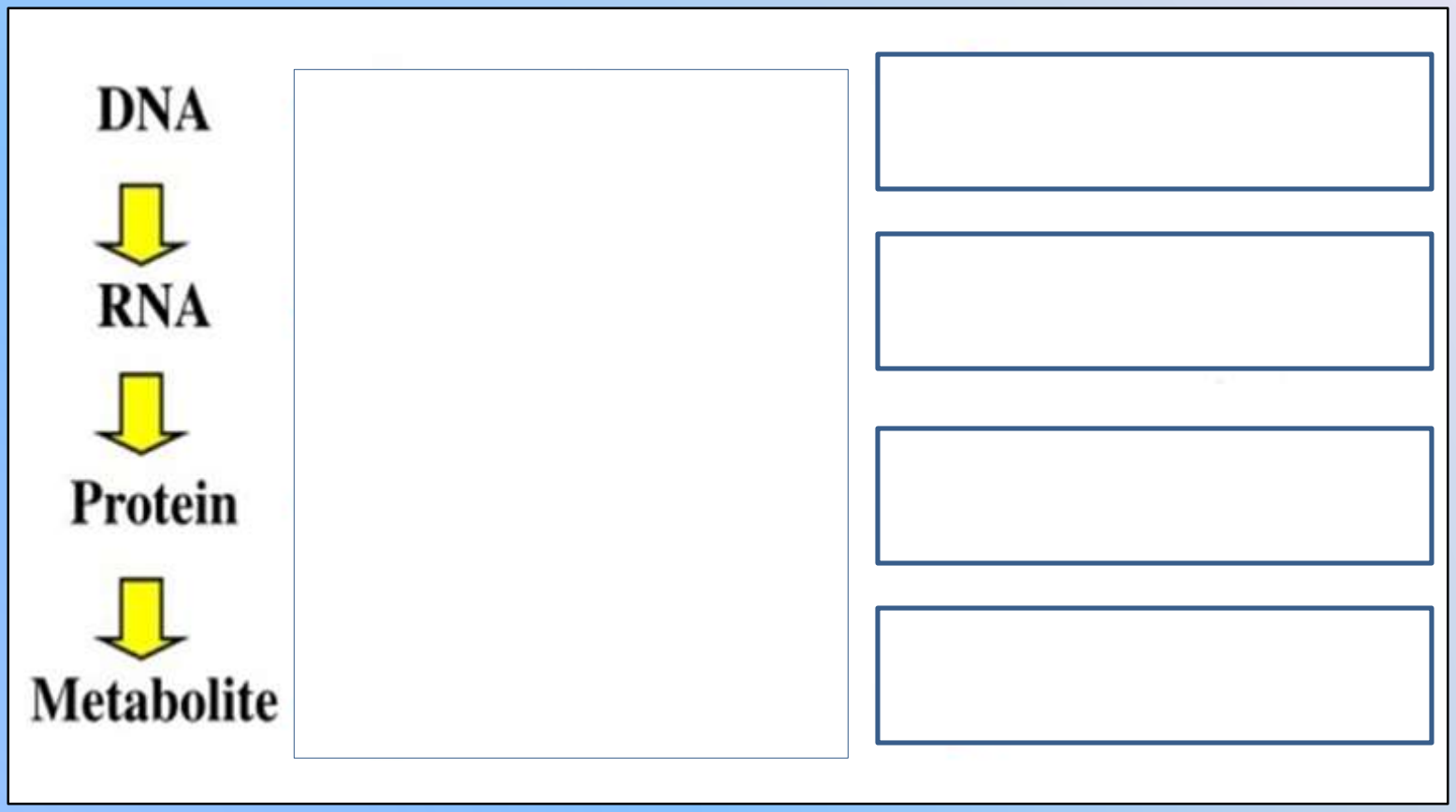
Ms. Rupal Mishra

What is ….. ??

Central Dogma of Life



Omics Terminology

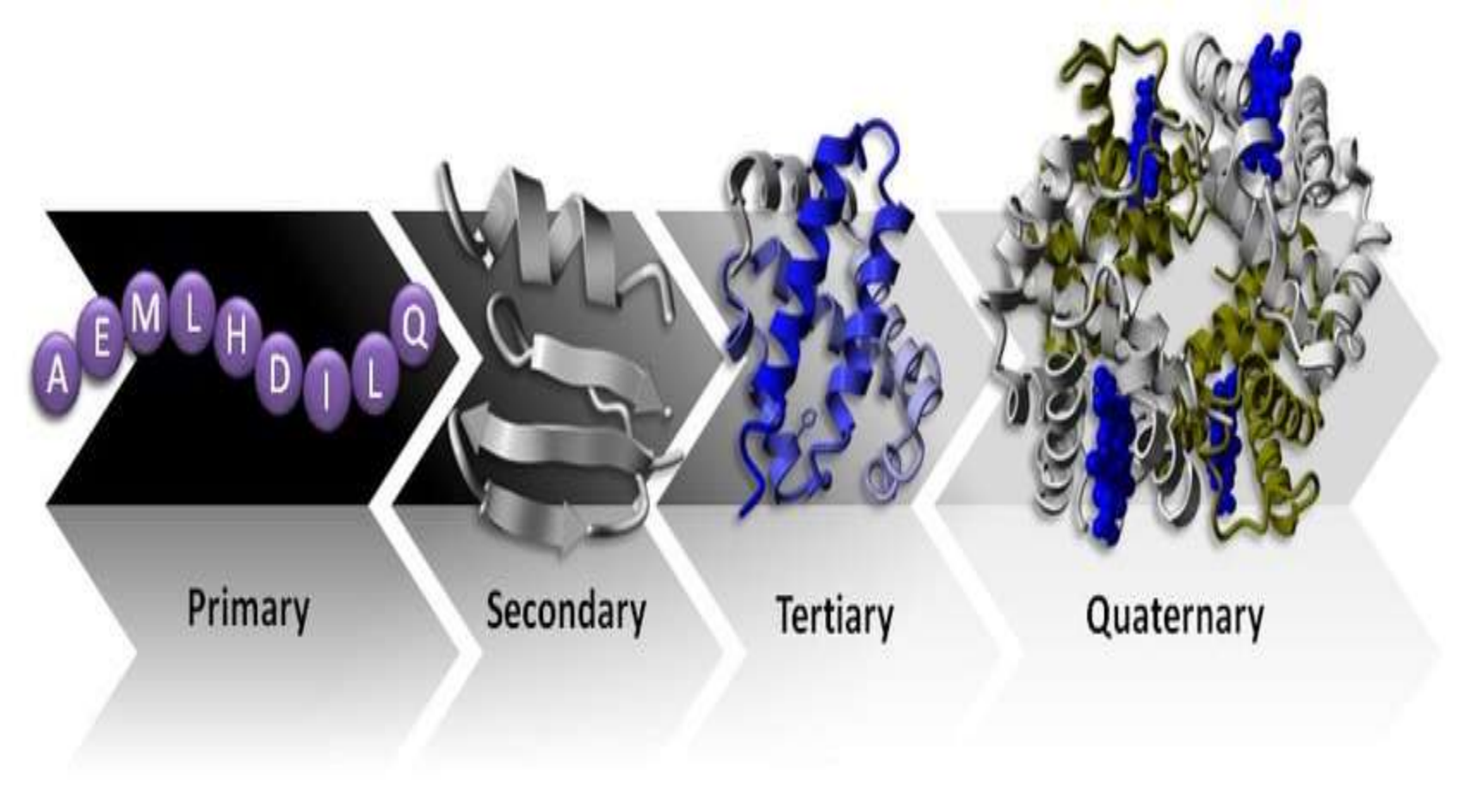


Protein

- Proteins are Building blocks, large & complex molecules, are made up of one or more long chains amino acids, required for the structure rigidity, proper functioning and regulation of pathways

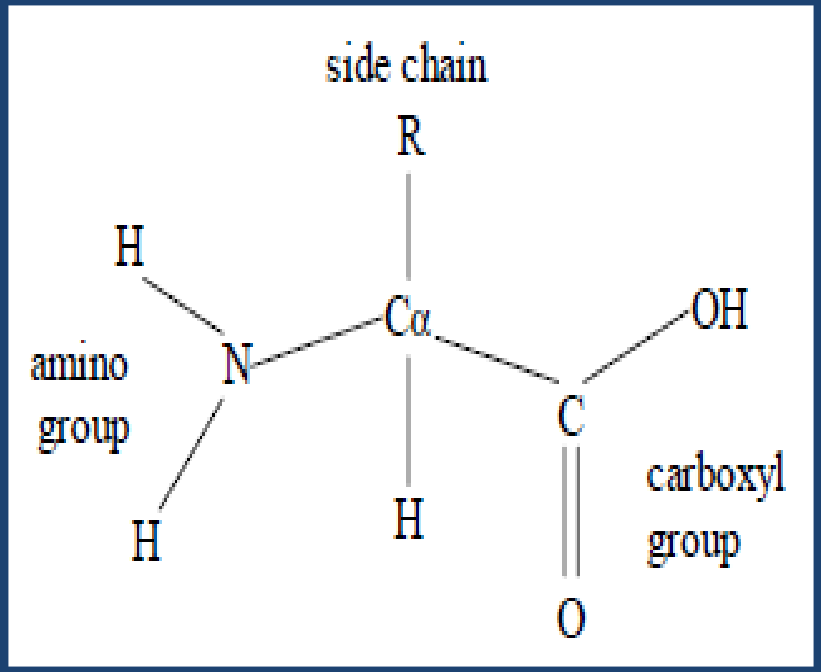
- Functions:
  - ✓ Growth and Maintenance (Tissues)
  - ✓ Causes Biochemical Reactions (Enzymes)
  - ✓ Acts as a Messenger (Insulin)
  - ✓ Provides Structure (keratin)
  - ✓ Bolsters Immune Health (antibodies)
  - ✓ Transports Molecules (hemoglobin)
  - ✓ Stores Nutrients (Ferritin)

Protein Hierarchy

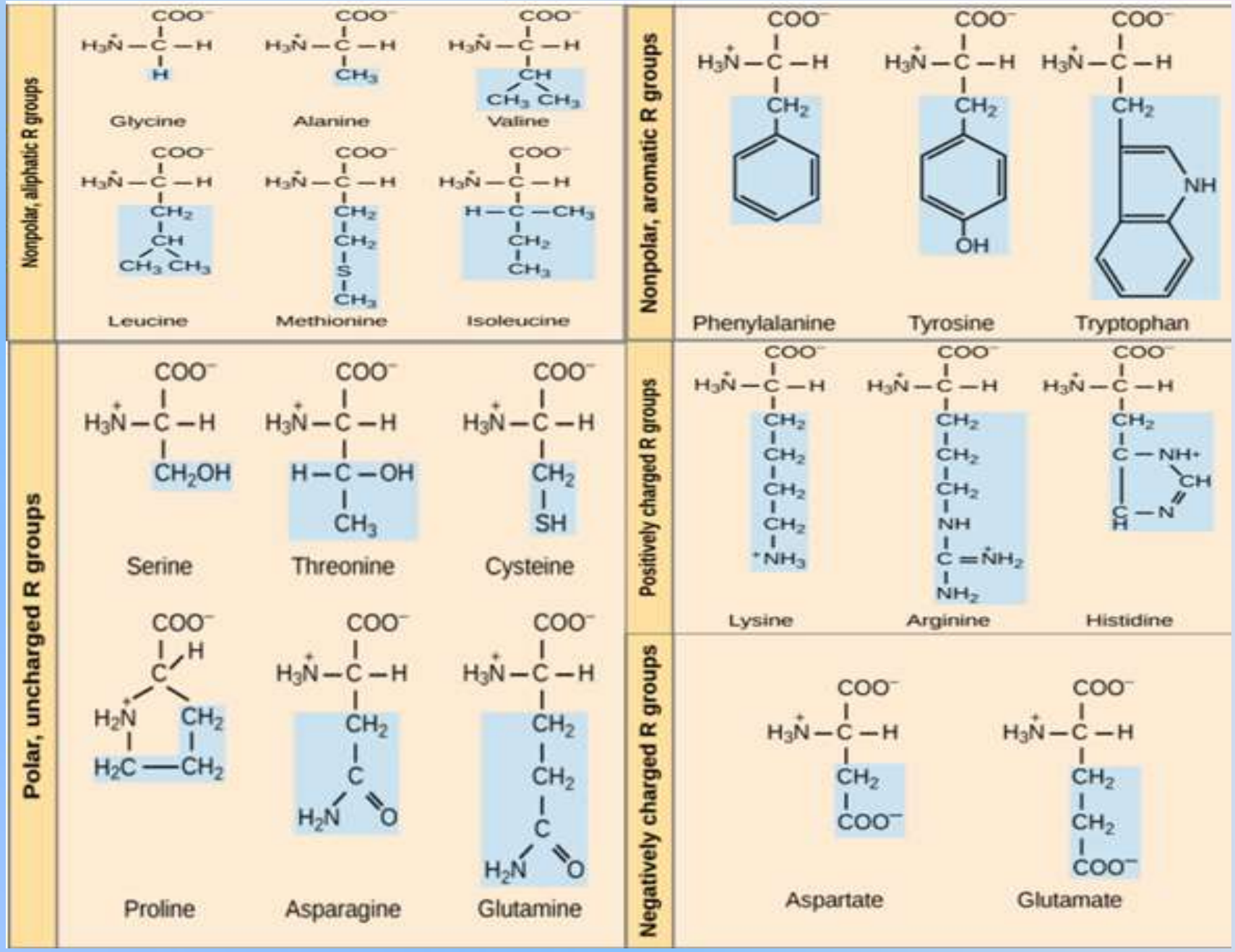


Amino-Acid

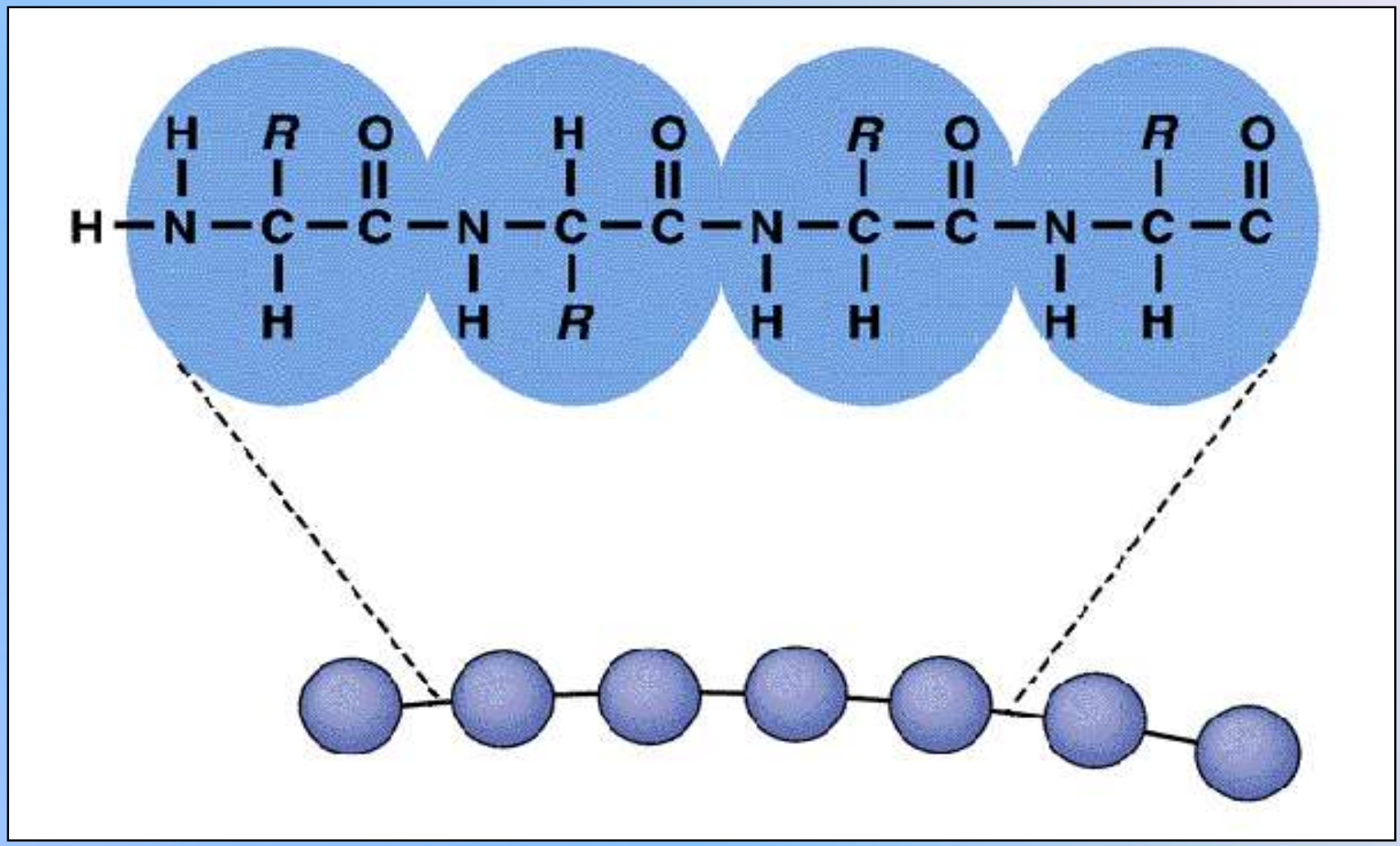
- Amino acids in the interior of the protein molecule come from the hydrophobic class while amino acids from the hydrophilic class are at the surface of the molecule.



AMINO ACID	THREE-LETTER ABBREVIATION	ONE-LETTER SYMBOL
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

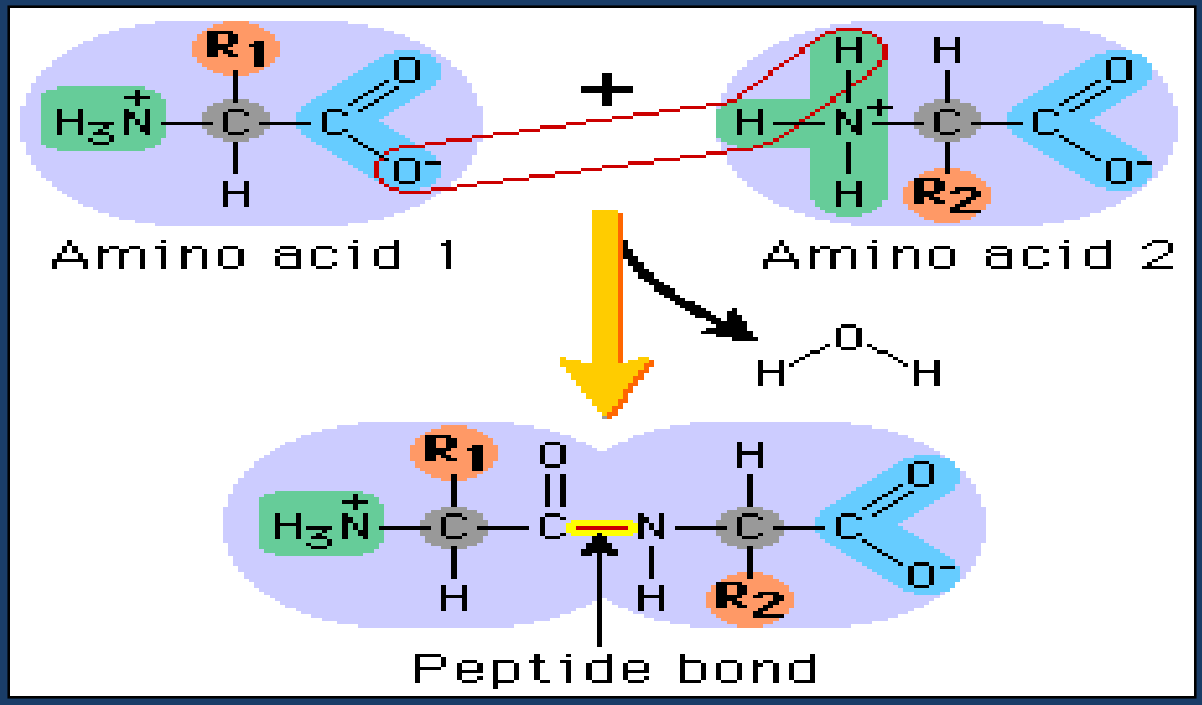

Nonpolar, aliphatic R groups
Glycine
Alanine
Valine
Leucine
Methionine
Isoleucine
Nonpolar, aromatic R groups
Phenylalanine
Tyrosine
Tryptophan
Polar, uncharged R groups
Serine
Threonine
Cysteine
Proline
Asparagine
Glutamine
Positively charged R groups
Lysine
Arginine
Histidine
Negatively charged R groups
Aspartate
Glutamate

Primary Struc of Protein



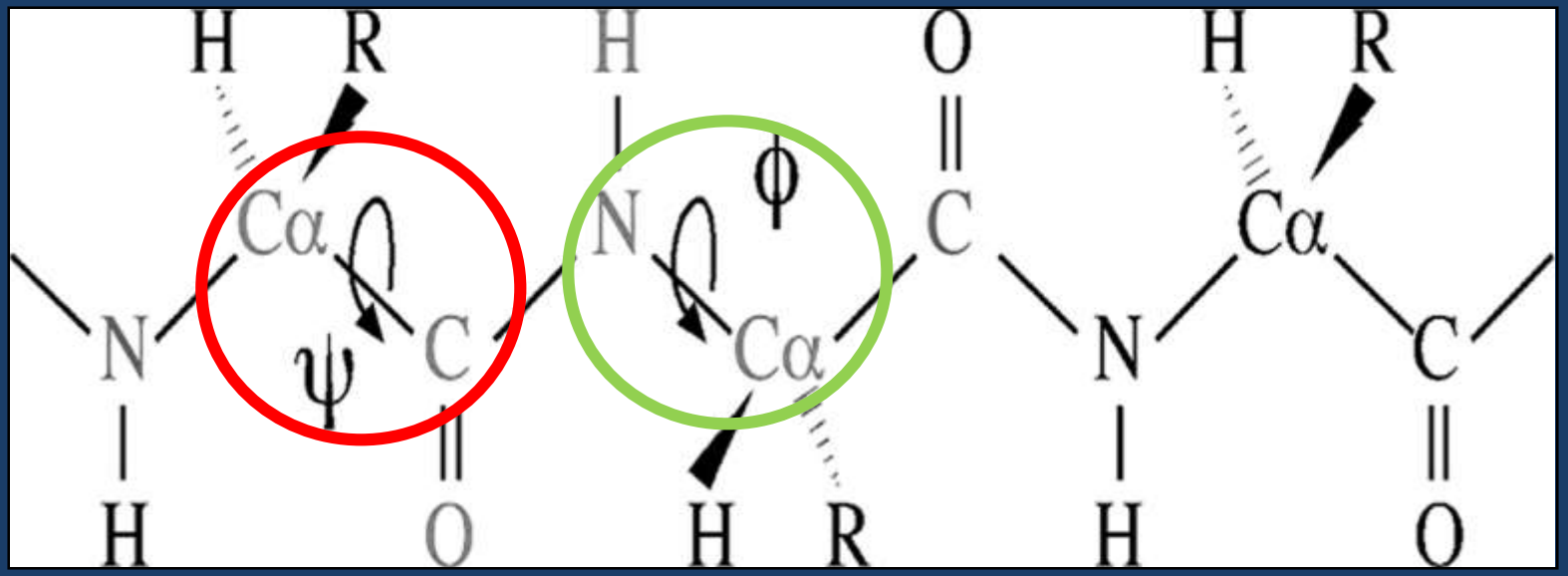
Peptide bond

- A peptide bond is a covalent chemical bond linking two consecutive amino acids from Carboxyl group of one amino acid and Amino group of another along a peptide or protein chain.



Dihedral Angle

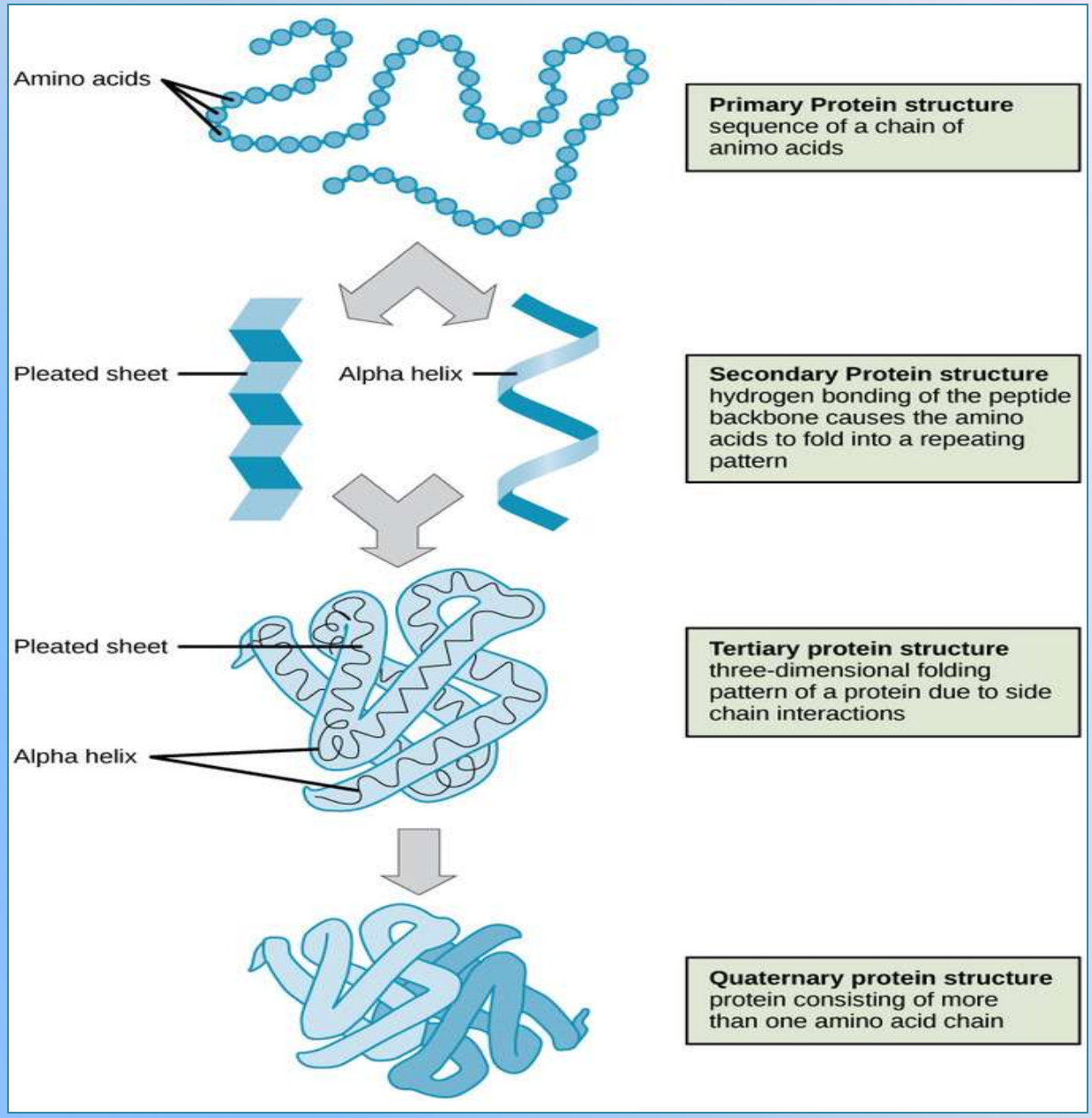
- The angle of rotation about the bond is referred to as the dihedral angle (also called the tortional angle).
- N–Cα bond, which is called as phi (φ).
- Cα–C bond, which is called as psi (ψ).



Stabilizing Forces

- Protein structures can be organized into four levels: primary structure, secondary structure, tertiary structure, and quaternary structure.
- Protein structures from secondary to quaternary are maintained by non-covalent forces.
- These include electrostatic interactions, van der Waals forces, and hydrogen bonding.
- Disulfide bridges, which are covalent bonds between the sulfur atoms of the cysteine residue, are also important in maintaining some protein structures.
- For certain types of proteins that contain metal ions as prosthetic groups, non-covalent interactions between amino acid residues and the metal ions may play an important structural role.

Protein Structure & Folding



Post-Translational Modifications

- **Proteins differ from one another primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes, and which usually results in folding of the protein into a specific three-dimensional structure that determines its activity.**
- Shortly after or even during synthesis, the residues in a protein are often chemically modified by posttranslational modification, which alters the physical and chemical properties, folding, stability, activity, and ultimately, the function of the proteins.

Proteome

- **What Is A Proteome?**

A proteome is the sum of all the proteins in an organism, a tissue, or the sample under study

Proteome refers to the entire set of expressed proteins in a cell

- **What is Proteomics?**

Proteomics is the study of the proteome

It is the study of composition, structure, function and interaction of the proteins directing the activities of each living cell

Proteomics

- **Marc Wilkins** coined the word **'Proteome'** in **1994**.
- The **goal** of proteomics is to analyze the varying proteomes of an organism at different times, in order to highlight differences between them.
- Proteomics on the whole can be divided into **three types**: Functional, structural and differential or Expression proteomics.

Proteomics objectives

1) Protein/peptide separation
2) Identification and characterization of resolved proteins by MS
3) Data analysis and applications.

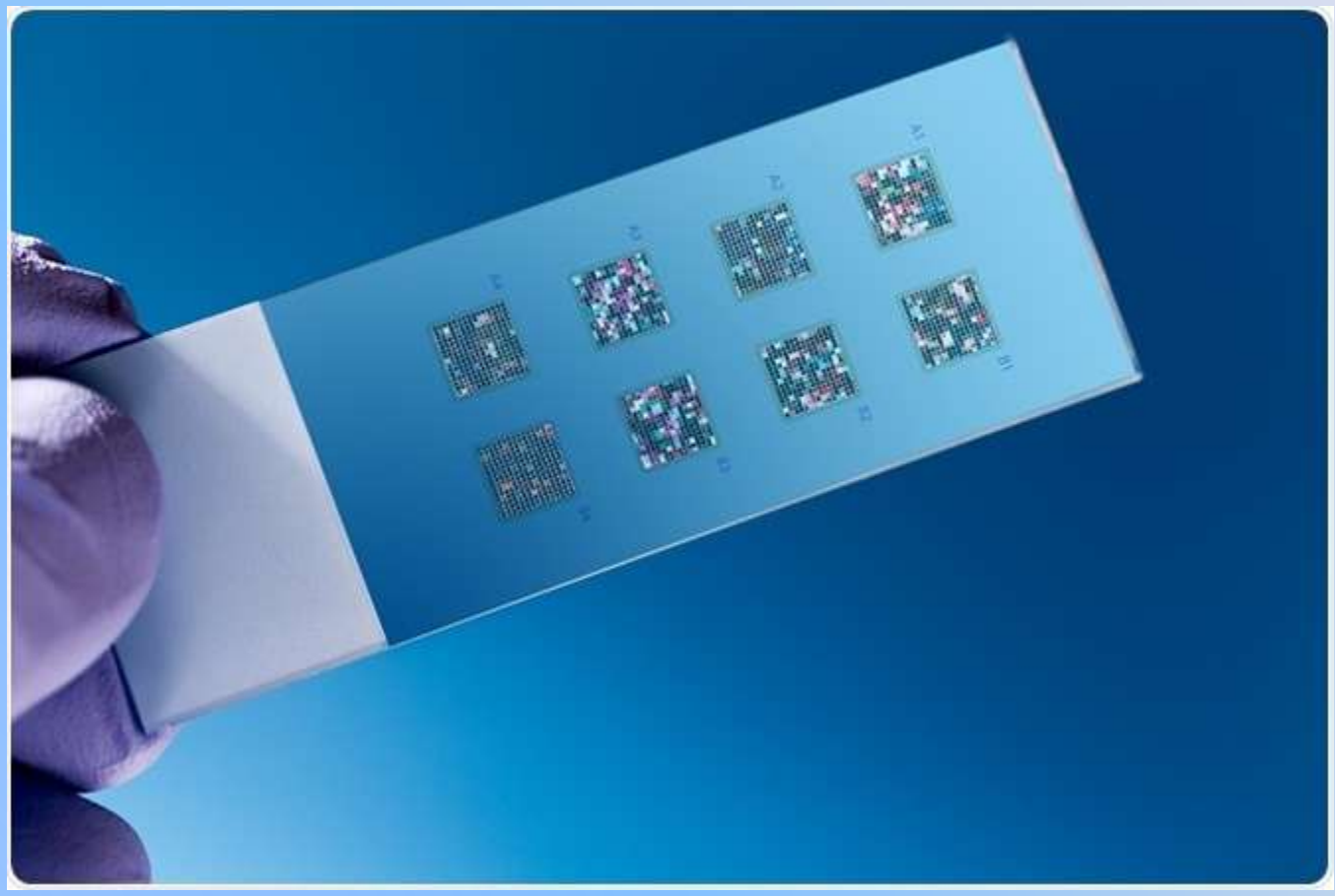
Types of Proteomics

- Functional:- Identification of Protein-Protein, Protein-DNA & Protein-RNA interactions affecting function
- Structural:- Identification of all interactions by metal ions, toxin , drugs etc. affecting protein structure
- Differential:- Determination of differences in protein expression

Protein Microarray

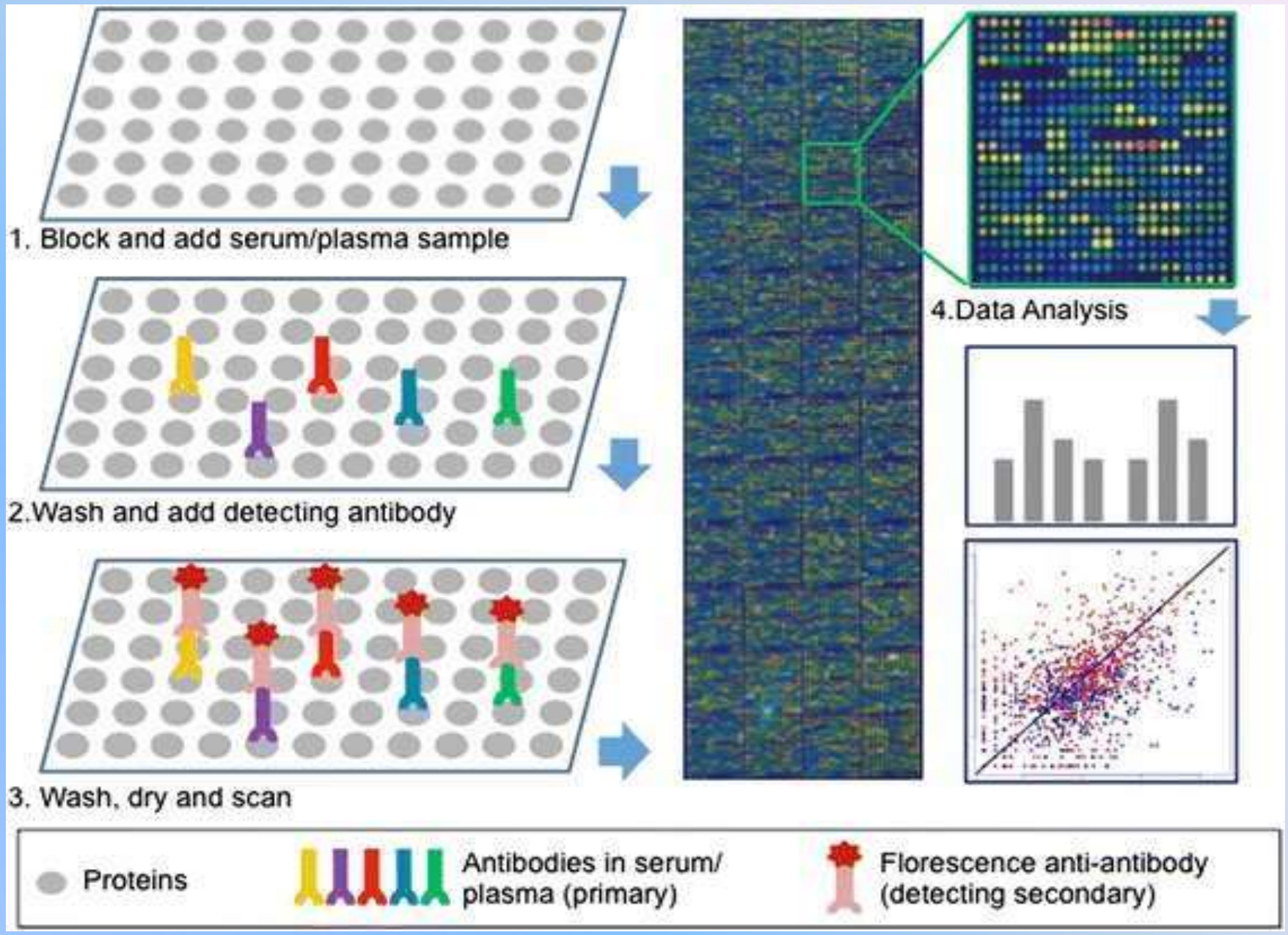
- Protein micro array(or Protein chip) is a high-throughput method used to track the interactions & activities of proteins, to determinate their functions.
- Protein micro array are rapid, automated, economical and highly sensitive consuming only small quantities of samples and reagents.
- Microarray technology is a term that refers to the miniaturization of thousands of assays on one small plate that contain small amounts of purified proteins in a high density format.
- They allow simultaneous determination of a great variety of analyte's from small amounts of samples within a single experiment.

Protein Microarray Chip

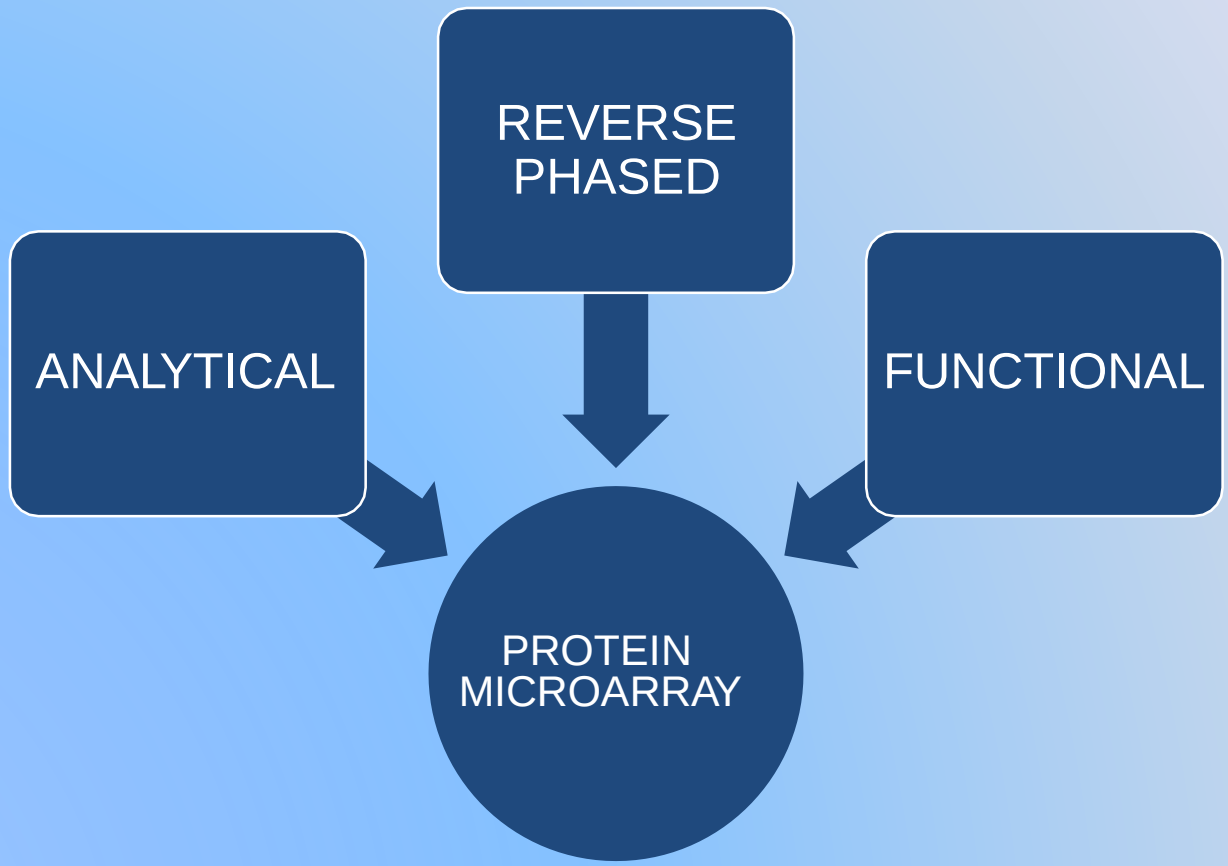


Protein Microarray

- Protein microarrays are typically prepared by immobilizing proteins onto a microscope slide using a standard contact spotter or non contact microarray
- Different methods of arraying the proteins:
  - Robotic method
  - Ink jetting method
  - Piezoelectric spotting
  - Photoliyhography.
- In these methods, robotic is contact microarray method while the other three are non contact microarray methods


1. Block and add serum/plasma sample
2.Wash and add detecting antibody
3. Wash, dry and scan
4.Data Analysis
Proteins
Antibodies in serum/ plasma (primary)
Florescence anti-antibody (detecting secondary)

Types Of Protein Microaarays

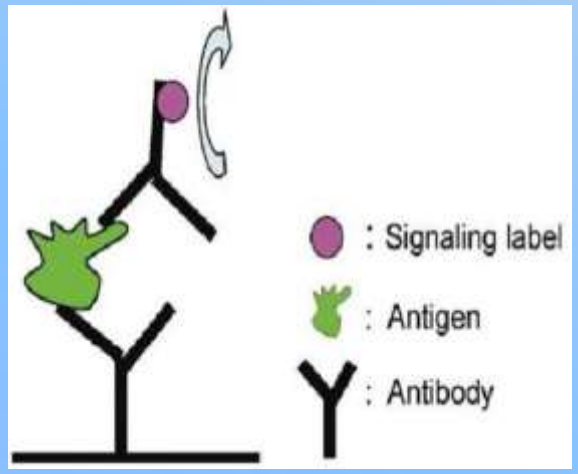


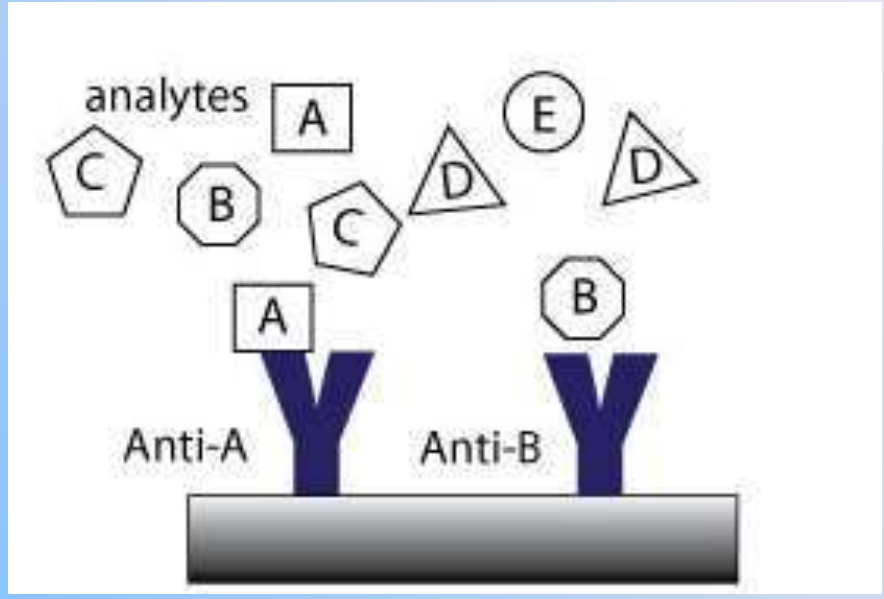
Analytical Protein Microarray

- The first model to demonstrate the application of antibody arrays was the "**analyte-labeled**" assay format. In this format, proteins are detected after antibody capture using direct protein labeling .

- Some limitations have to be considered because this method lacks specificity in protein target labeling and has poor sensitivity for low abundance proteins. Moreover, targeted protein labeling may lead to the epitope destruction due to some chemical reaction.

Analytical Protein Microarray





Used to understand:

- expression levels,
- binding affinities and specificities,
- response of the cells to a particular factor,
- identification and profiling of diseased tissues.

Analytical Protein Microarray

- Another model of antibody array provides higher sensitivity using the "**sandwich**" assay format. This format employs two different antibodies to detect the targeted protein .
- One antibody, called the **capture** antibody, immobilizes the targeted protein on the solid phase, while the other antibody, called the **reporter** or **detection** antibody, generates a signal for the detection system. Using two antibodies significantly increases the specificity and sensitivity of the "sandwich" assay format, even at femtomolar levels .
- These assays offer a multiplexed format of the original Enzyme-linked Immunosorbent Assay (ELISA).

Analytical Protein Microarray

- Analytical microarrays(or antibody microarrays) have antibodies arrays on solid surface,and are used to detect proteins in biological samples.
- Often a second is used to detect a protein that is captured by the antibody attached to the solid phase,in a principle similar to that of **sandwich immunoassay,** in which the first antibody is spotted on the array and then a captured antigen on the chip is detected with a second antibody that recognises a different part of antigen.
- Analytical protein arrays can be used to monitor protein expression levels or for bio- marker identification, clinical diagnosis, or environmental/ food safety analysis.

Limitations

- Antibodies are the most popular protein capture reagents, although their affinity and/or specificity can vary dramatically .
- Many antibodies may cross-react with proteins other than their expected target proteins when tested on functional protein microarrays, especially when multiple analyte detection is employed.
- The need for highly specific antibodies has become a major challenge in analytical protein microarrays because nonspecific binding will lead to large numbers of false positive result.
- Another challenge comes from producing a large number of antibodies in a high-throughput fashion. Recombinant antibodies have become a promising means to overcome this.

Functional Protein Microarray

- Also known as **target protein array**. With functional protein microarrays purified recombinant protein are immobilized onto the solid phase.
- Functional protein microarrays have recently been applied to many aspects of discovery based biology, including protein- protein, protein- lipid, protein-DNA, protein-drug, &protein – peptide interactions.
- These can be used to identify enzyme substrates.

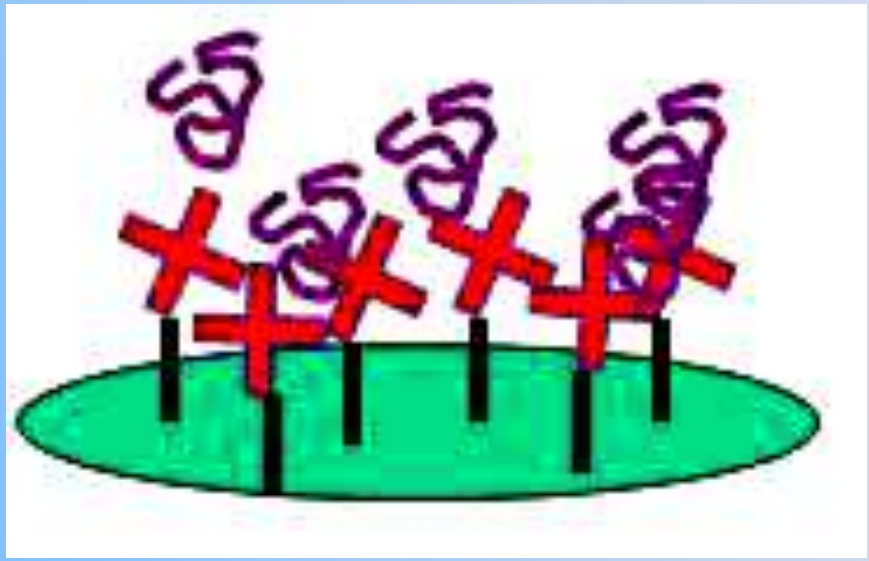
Functional Protein Microarrays

- Immobilized purified proteins are used to:

  - identify protein-protein/DNA/RNA/PL/SM,

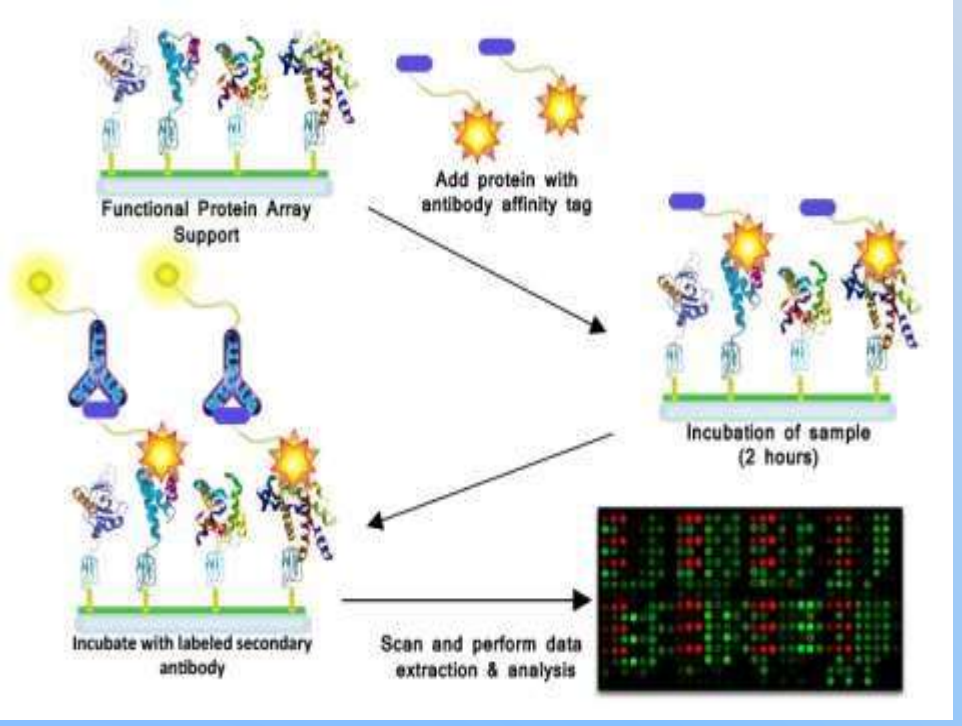
  - assay enzymatic activity.

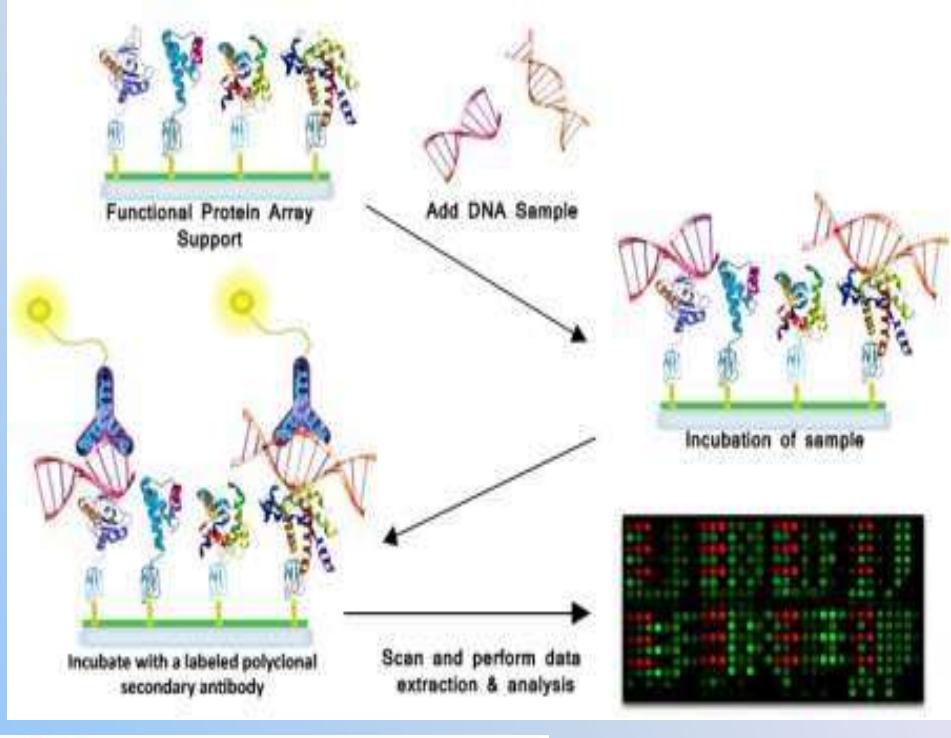
- They differ from analytical arrays in that they contain full length functional proteins.

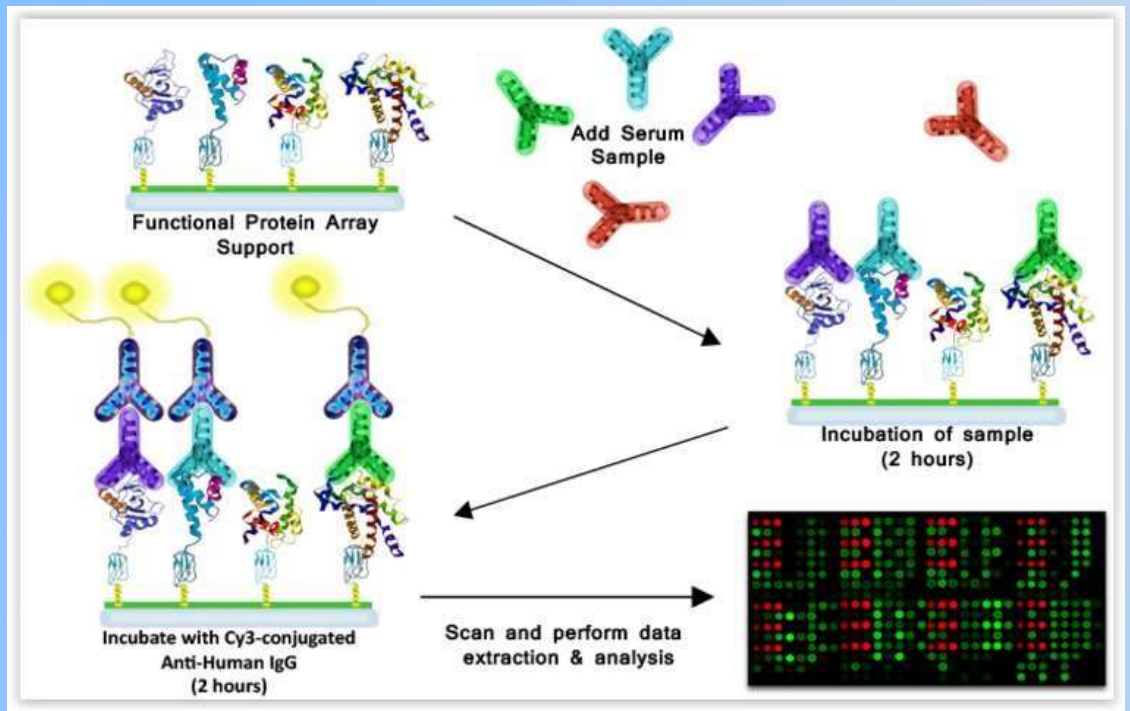


Functional Protein Microarrays

- These can also be used to detect antibodies in a biological specimen to profile an immune response.
- The first use of functional protein microarrays was demonstrated by **zhu et al.** (2001) to determine the substrate specificity of protein kinases in yeast.
- Protein microarrays enable us to study many post-translational modifications (i.E., Phosphorylation, acetylation, ubiquitylation, S- nitrosylation) in a large-scale fashion, which is critical for understanding cellular protein synthesis and function.


Functional Protein Array Support
Add protein with antibody affinity tag
Incubation of sample (2 hours)
Incubate with labeled secondary antibody
Scan and perform data extraction & analysis


Functional Protein Array Support
Add DNA Sample
Incubation of sample
Incubate with a labeled polyclonal secondary antibody
Scan and perform data extraction & analysis


Functional Protein Array Support
Add Serum Sample
Incubation of sample (2 hours)
Incubate with Cy3-conjugated Anti-Human IgG (2 hours)
Scan and perform data extraction & analysis

Reverse-Phase Protein Microarray

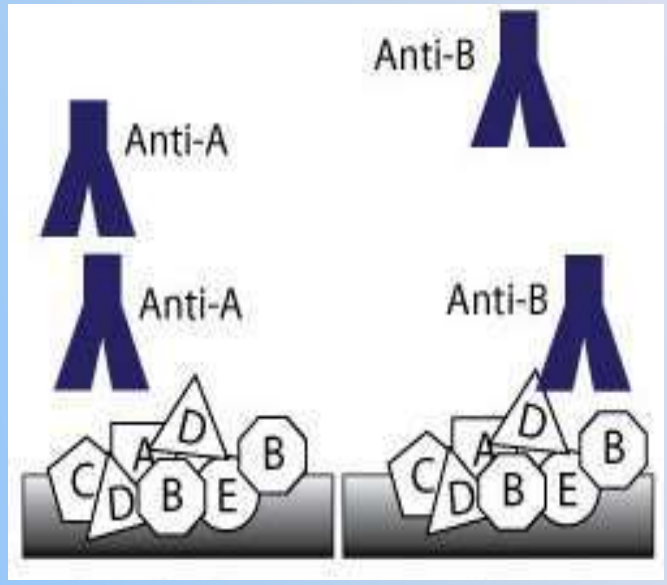
- Involves complex samples, such as **tissue lysates**.cells are isolated from various tissues of interest and lysed.
- The lysate is arranged onto the microarray & probed with antibodies against the target protein of interest.
- These antibodies are typically detected with **chemiluminescent, fluoresent or colorimetric assays**.
- This type of microarrays was first established by **Paweletz** and colleagues to monitor histological changes in **prostate cancer** patients.

Reverse-Phase Protein Microarray

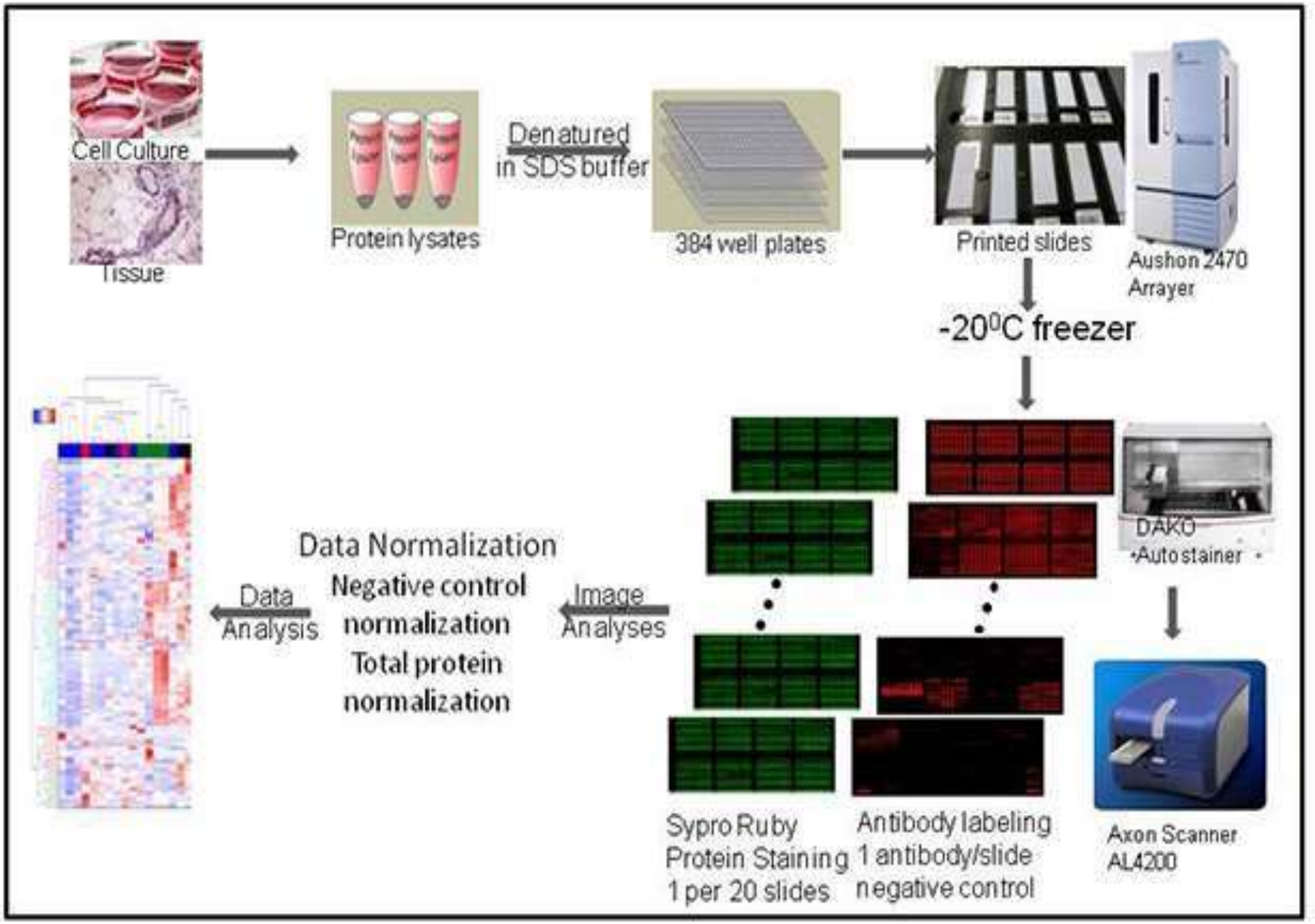
- Using this method, they successfully detected microscopic transition stages of **pro-survival checkpoint protein** in three different stages of prostate cancer: normal prostate epithelium, prostate intraepithelial neoplasia, and invasive prostate cancer.
- The high degree of sensitivity, precision and linearity achieved by reverse-phase protein microarrays enabled this method to quantify the **phosphorylatio**n status of some proteins (such as Akt and ERK) in these samples; phosphorylation was statistically correlated.

Reverse Phase Protein Microarray

- Involve complex samples, such as tissue lysates probed with antibodies against the target protein of interest.
- These antibodies are typically detected with chemiluminescent, fluorescent or colorimetric assays.
- Used to: determination of the presence of altered proteins or other agents as a result of disease.
- Specifically, post-translational modifications, which are typically altered as a result of disease.



Reverse Phase Protein Array (RPPA) Overview



Reverse Phase Protein Microarray

- RPAs allow for the determination of the presence of altered proteins or other agents that may be the result of disease.
- Specifically, post translationl modifications,which are typically altered as a result of disease can be detected using RPAs.
- Harnessing this sophisticated technology, Ciaccio et al. (2010) profiled EGF receptor signaling dynamics using micro-western arrays (MWA), which combine **western blotting** and **reverse- phase protein microarrays** to produce better sensitivity by **separation** of whole lysate sample components.

Reverse Phase Protein Microarray

- This method allowed them to precisely measure **91** phosphosites of **67** proteins at **6** different time points with five EGF concentrations in **A431** human carcinoma cells to analyze the
- A significant drawback of this approach however is that it is highly dependent on the availability and specificity of commercially produced antibodies. Because of this bias, it is has limited applications.

Commercially available protein microarray

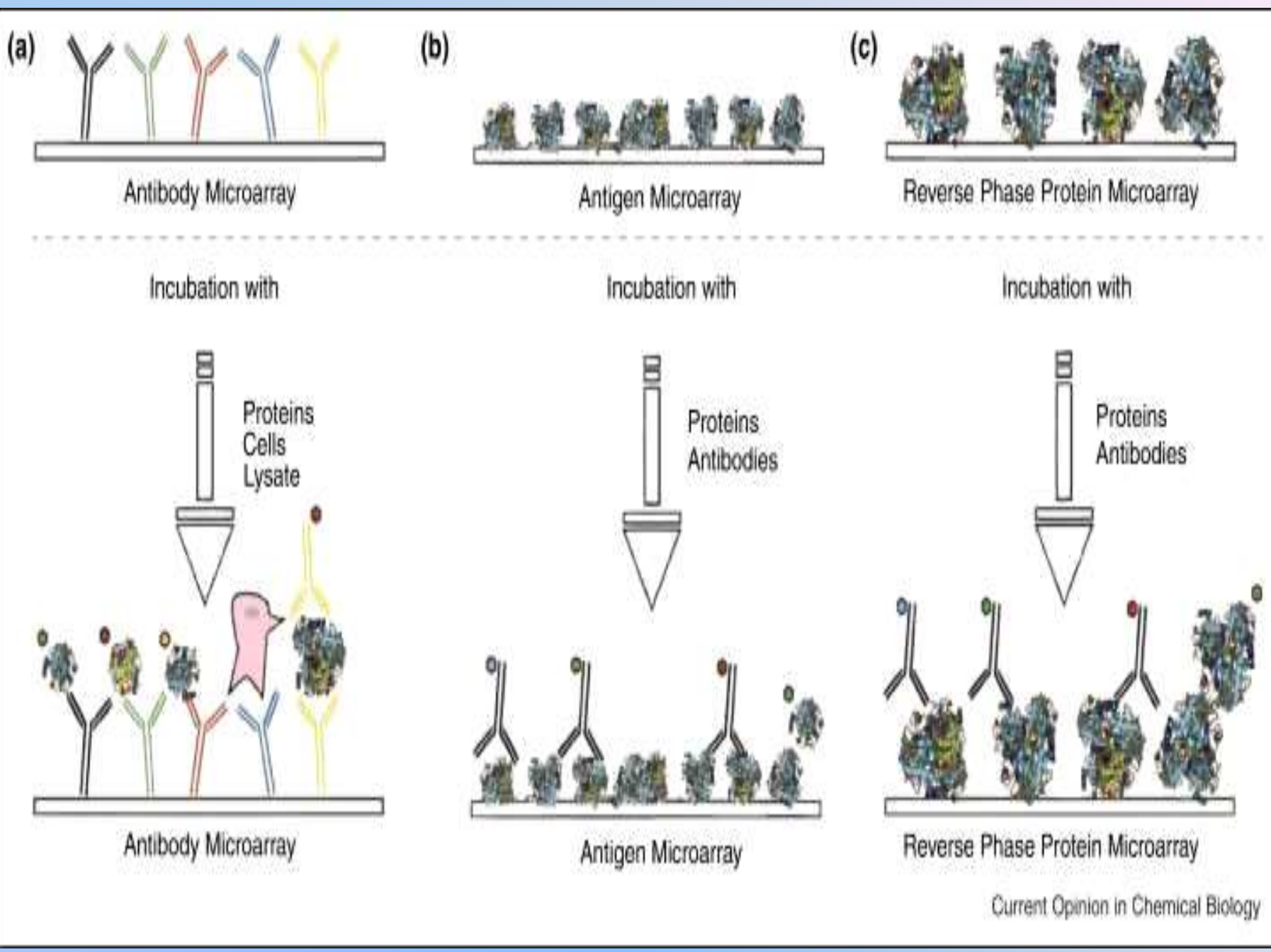
Product Type	Product Name	Company	Type of Array	Protein Content
Human protein	ProtoArray®	Invitrogen	Functional	9000 human proteins
Kinase	Kinex™	Kinexus Bioinformatics	Functional	200 human kinase proteins
Pathogen	Arrayit Pathogen Antigen Microarray	Arrayit Corporation	Functional	Essential proteins of different pathogens
Antibody for specific group of proteins	RayBio® Human RTK Phosphorylation Antibody Array	RayBiotech, Inc	Analytical	Antibodies against 71 human kinases
	RayBio® Human Cytokine Antibody Array	RayBiotech, Inc	Analytical	Antibodies against various human cytokines
	PlasmaScan™ 380 Antibody Microarray	Arrayit Corporation	Analytical	Antibodies for human plasma detection
	Cytokine Antibody Microarray	Full Moon BioSystems, Inc	Analytical	Antibodies against 77 human cytokines
	Kinase Antibody Microarray	Full Moon BioSystems, Inc	Analytical	Antibodies against 276 human kinases
Antibody for pathway detection	MAPK Pathway Phospho Antibody Array	Creative Bioarray	Analytical	185 antibodies against phospho-proteins in the MAPK pathways
	Signaling Explorer Antibody Microarray	Full Moon BioSystems, Inc	Analytical	1358 antibodies for multiple pathways
	Wnt Signaling Phospho Antibody Microarray	Full Moon BioSystems, Inc	Analytical	227 phospho-antibodies for cell growth, movement and development pathways
Cell lysate	SomaPlex™	Protein Biotechnologies	Reverse-phase	A variety of human cancer cell lysates

Applications

- There are five major areas where protein arrays are being applied: diagnostics, proteomics, protein functional analysis, antibody characterisation & treatment.
- Diagnostics involves the detection of antigens & antibodies in blood samples; to discover new disease biomarkers; the monitoring of disease states & responses to therapy in personalised medicine; the monitoring of environment & food.
- Proteomics pertains to protein expression profilling i.e; which Proteins are expressed in the lysate of a part of cell

Applications

- Protein functional analysis is the identification of protein-protein interactions, protein- phospholipid interactions, small molecule targets, enzymatic substrates & receptor ligands.
- Antibody characterization is characterizing cross reactivity, specificity & mapping epitopes.
- Treatment development involves the development of antigen- specific therapies for autoimunity, cancer & allergies; the identification of small molecule targets that could potentially be used as new drugs.


(a)
(b)
(c)
Antibody Microarray
Antigen Microarray
Reverse Phase Protein Microarray
Incubation with
Incubation with
Incubation with
Proteins
Cells
Lysate
Proteins
Antibodies
Proteins
Antibodies
Antibody Microarray
Antigen Microarray
Reverse Phase Protein Microarray
Current Opinion in Chemical Biology

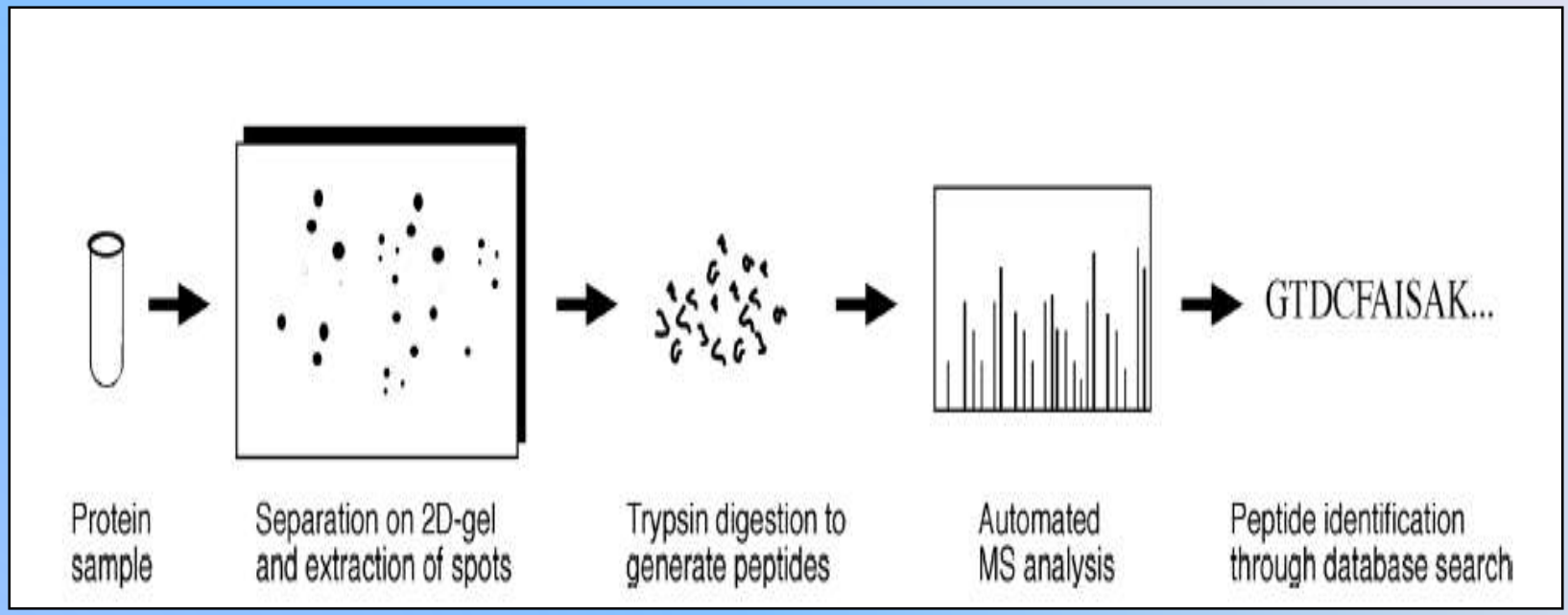
Applications of Proteomics

I. Protein sample identification/ confirmation.
II. Protein sample purity determination.
III. Detection of amino acids substitution.
IV. Nutrition Research
V. To identify unknown protein of interest.
VI. Quantify protein and peptide.
VII. Protein Biomarker.
VIII. Difference in expression of protein.
IX. Protein to gene prediction.
X. Drug development.

Steps - Proteomics Study

- Purification of Proteins
- Separation of Protein
- Extraction of Protein
- Digestion into peptides
- Determination of Mass by Mass spectrometry
- Determination of Amino-acid sequence of peptides by database search
- Protein Identified

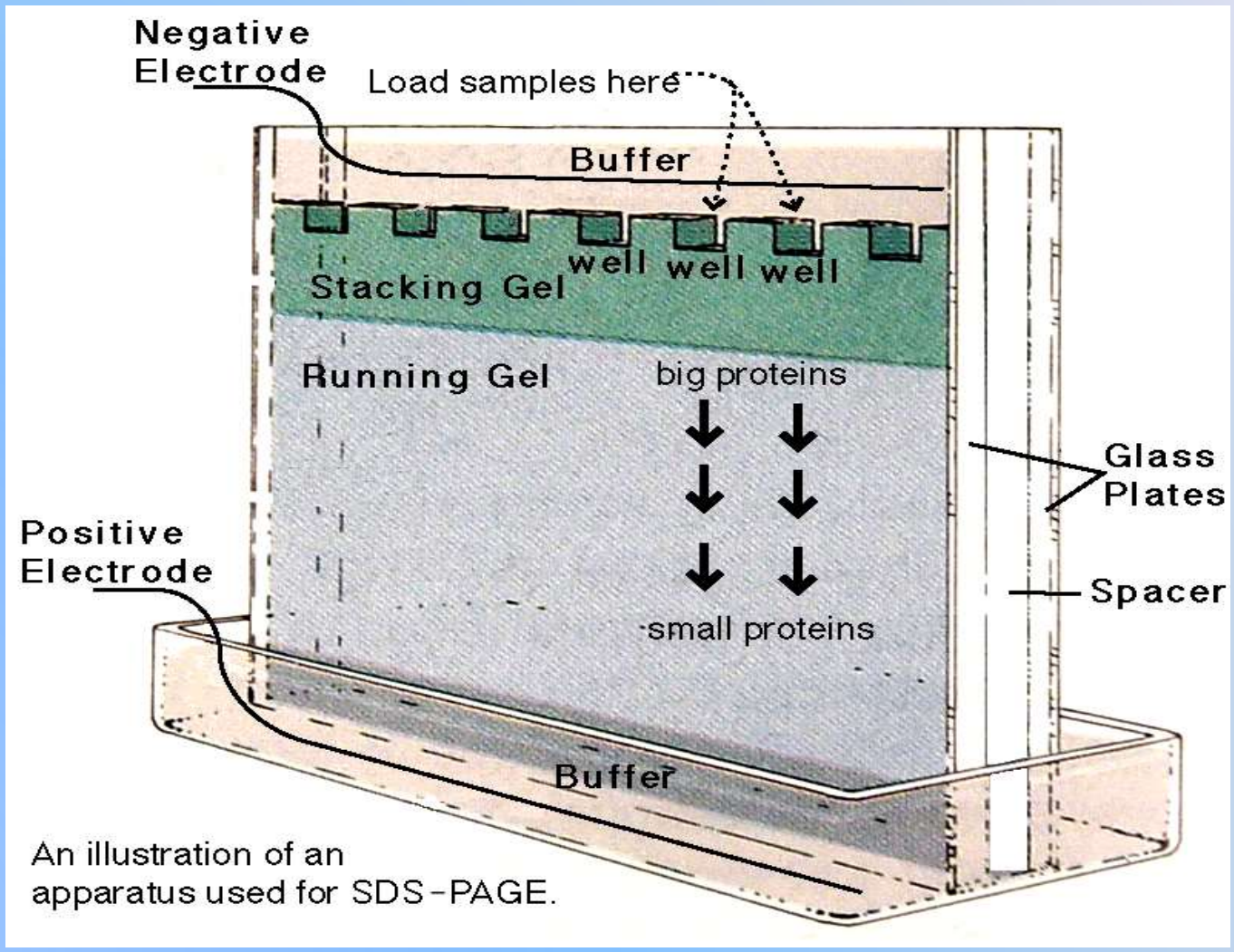
Techniques Involved in Proteomics Study



Separation Technique

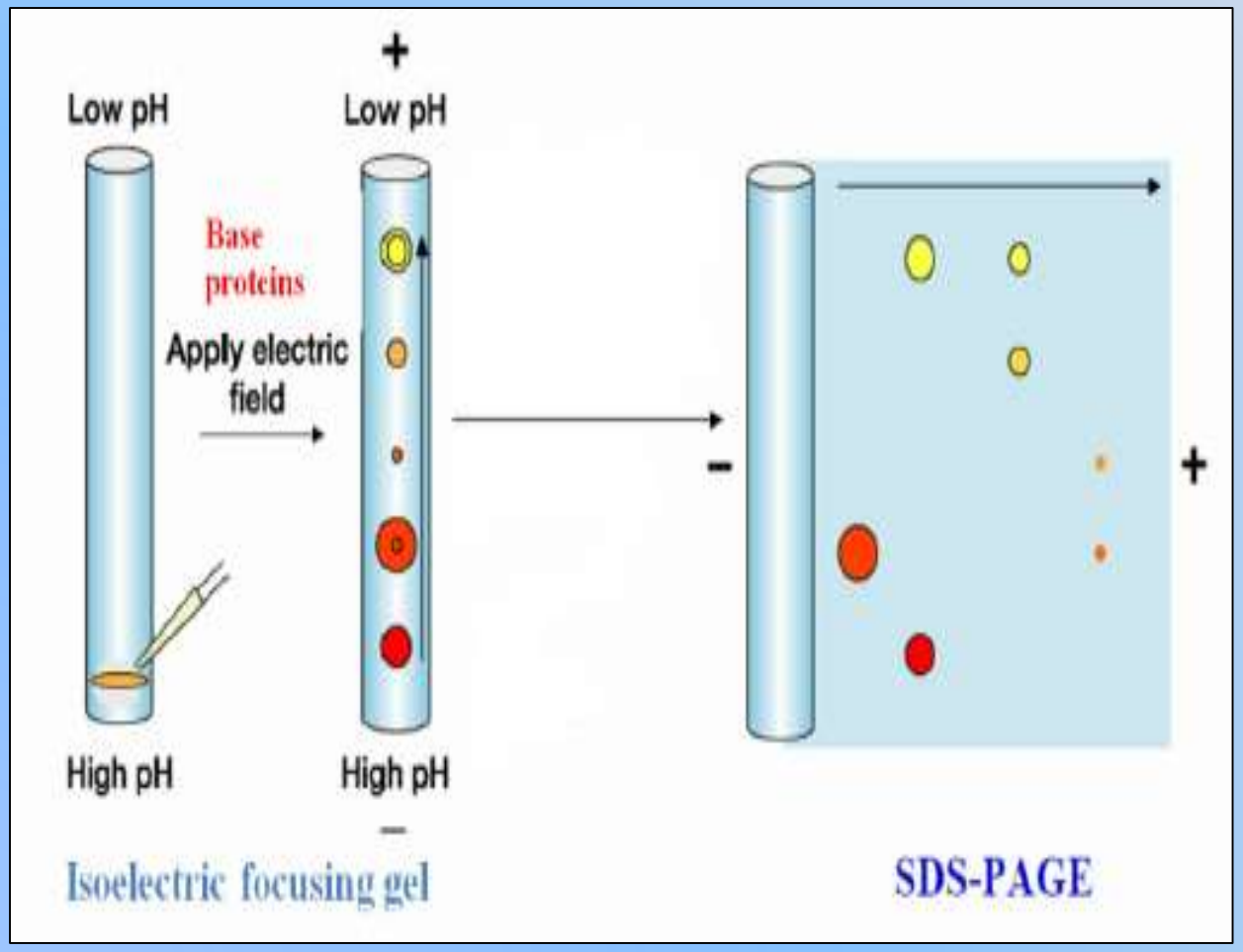
- One - Dimensional SDS-PAGE
- Two - Dimensional SDS-PAGE
- What is SDS-PAGE? – SDS-PAGE a type of gel electrophoresis.
- What is the purpose of doing gel electrophoresis? – It has been seen that by running a gel we are able to identify more proteins from the sample.
- An electric current is applied across the gel, causing proteins will differentially migrate based on their molecular mass.

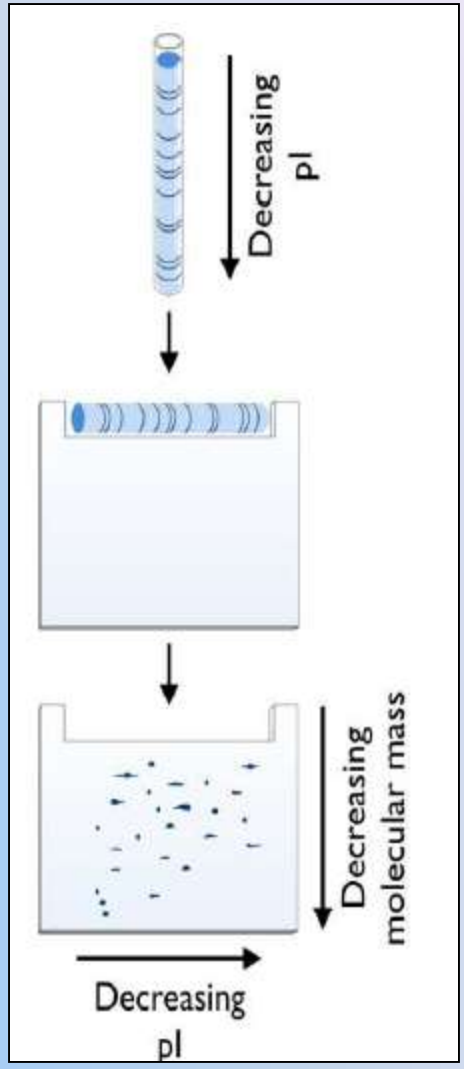
1-D SDS Page



An illustration of an apparatus used for SDS-PAGE.

2-D SDS Page


+
Low pH
Low pH
Base proteins
Apply electric field
High pH
High pH
−
Isoelectric focusing gel
−
+
SDS-PAGE


Decreasing pI
Decreasing molecular mass
Decreasing pI

	1D Gel Electrophoresis	2D Gel Electrophoresis
DEFINITION	1D gel electrophoresis separates proteins based on the molecular weight of the protein using polyacrylamide gel electrophoresis	2D gel electrophoresis separates proteins based on both the iso-electric point and the molecular weight of the protein
SEPARATION BASED ON	Molecular weight only	Iso-electric point and molecular weight
RESOLUTION	Low	High
COST	Low	High

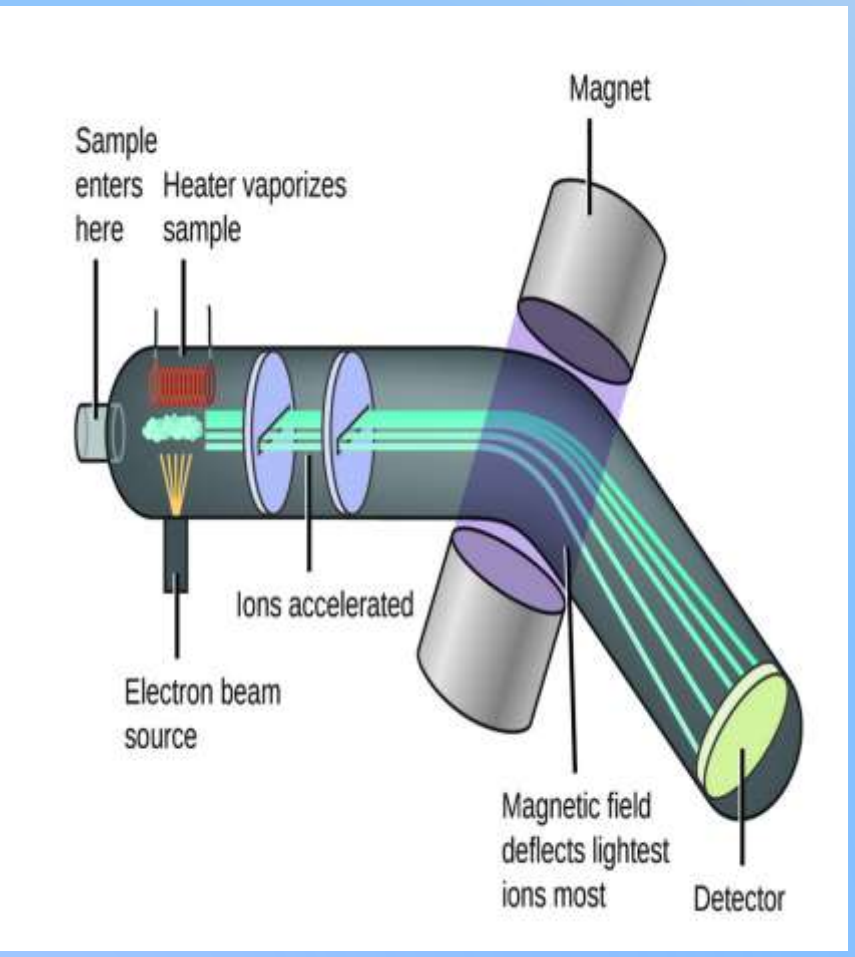
Mass Spectrometry

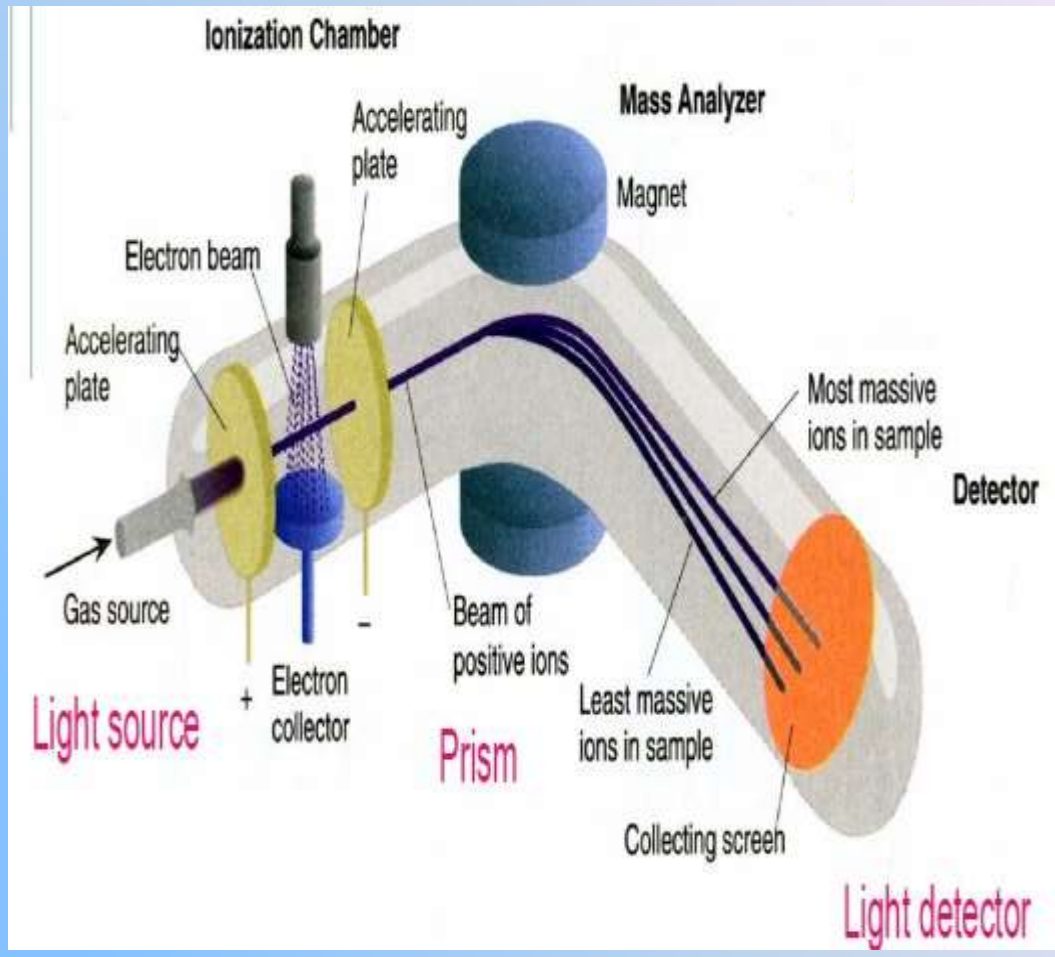
- Mass spectrometry is an analytical technique that is used to measure the mass-to-charge ratio of ions. The results are typically presented as a mass spectrum, a plot of intensity as a function of the mass-to-charge ratio.
- Mass spectroscopy is the most accurate method for determining the molecular mass of the compound and its elemental composition.
- In this technique, molecules are bombarded with a beam of energetic electrons. The molecules are ionised and broken up into many fragments.
- Each ion has a particular ratio of mass to charge, i.e. m/e ratio(value). For most ions, the charge is one and thus, m/e ratio is simply the molecular mass of the ion.

Mass Spectrometry

- Mass spectrometry in proteomics used for protein identification.
- It is useful for obtaining structural information like peptide mass & identifying type and location of protein modification.
- A mass spectrometer separates proteins according to their mass-to-charge(m/z) ratio.
- The molecule is first ionized. The process of ionization of proteins forces them to move towards the analyzer because of the charges on ions.
- Two types of MS instruments:
  1) MALDI-TOF [matrix assisted laser desorption ionization-time of flight.]
  2) ESI-MS-MS [ESI Tandem mass analyzer]

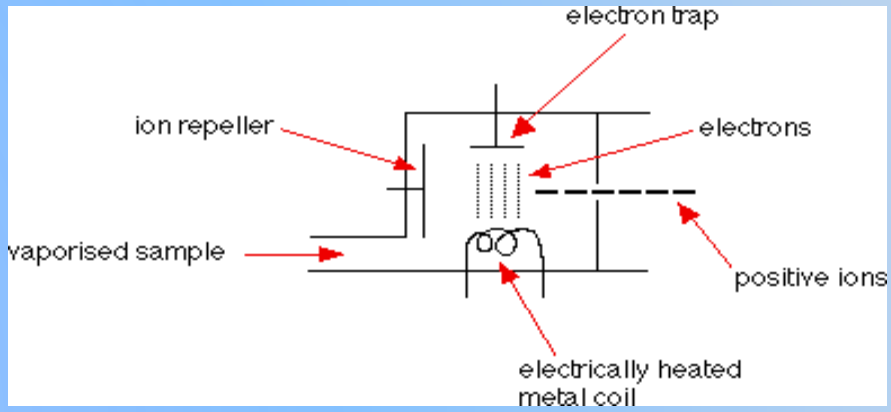
Mass Spectrometry





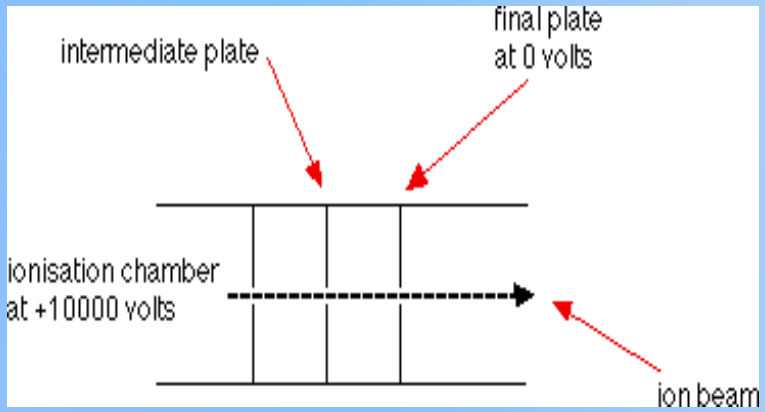
Ionization

- The atom is ionized by knocking one or more electrons off to give a positive ion. (Mass spectrometers always work with positive ions).
- The particles in the sample (atoms or molecules) are bombarded with a stream of electrons to knock one or more electrons out of the sample particles to make positive ions.
- Most of the positive ions formed will carry a charge of +1.
- These positive ions are persuaded out into the rest of the machine by the ion repeller which is another metal plate carrying a slight positive charge.



Acceleration

- The ions are accelerated so that they all have the same kinetic energy.
- The positive ions are repelled away from the positive ionisation chamber and pass through three slits with voltage in the decreasing order.
- The middle slit carries some intermediate voltage and the final at '0' volts.
- All the ions are accelerated into a finely focused beam.

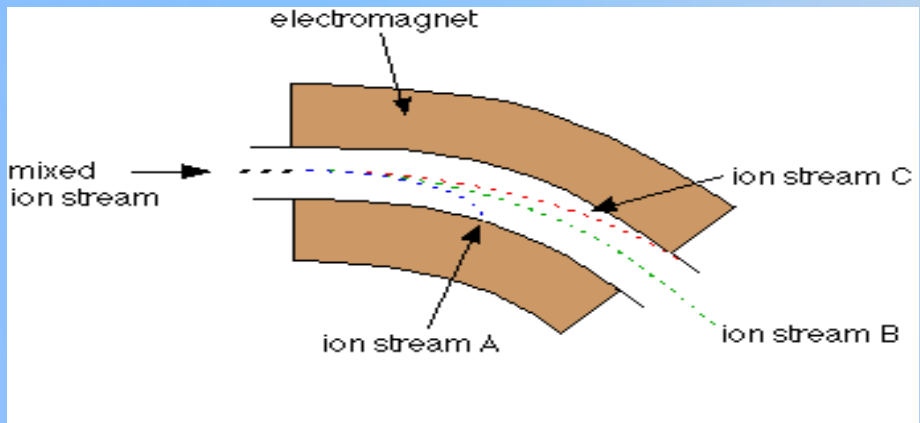


Deflection

- The ions are then deflected by a magnetic field according to their masses. The lighter they are, the more they are deflected.
- The amount of deflection also depends on the number of positive charges on the ion -The more the ion is charged, the more it gets deflected.
- Different ions are deflected by the magnetic field by different amounts. The amount of deflection depends on:

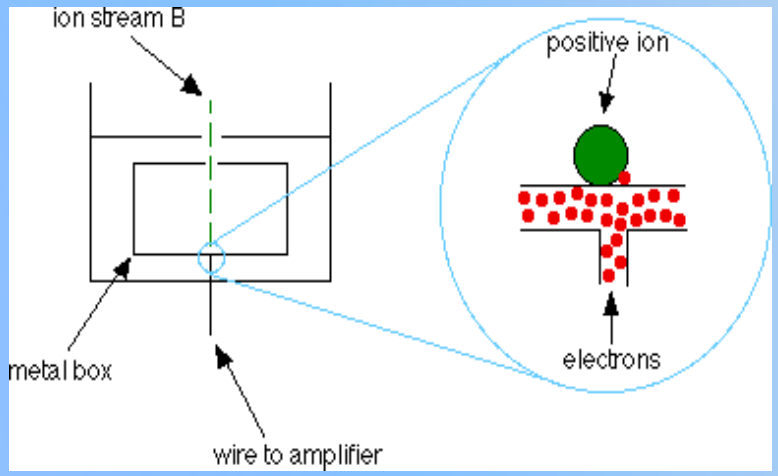
  1) The mass of the ion: Lighter ions are deflected more than heavier ones.

  2) The charge on the ion: Ions with 2 (or more) positive charges are deflected more than ones with only 1 positive charge.



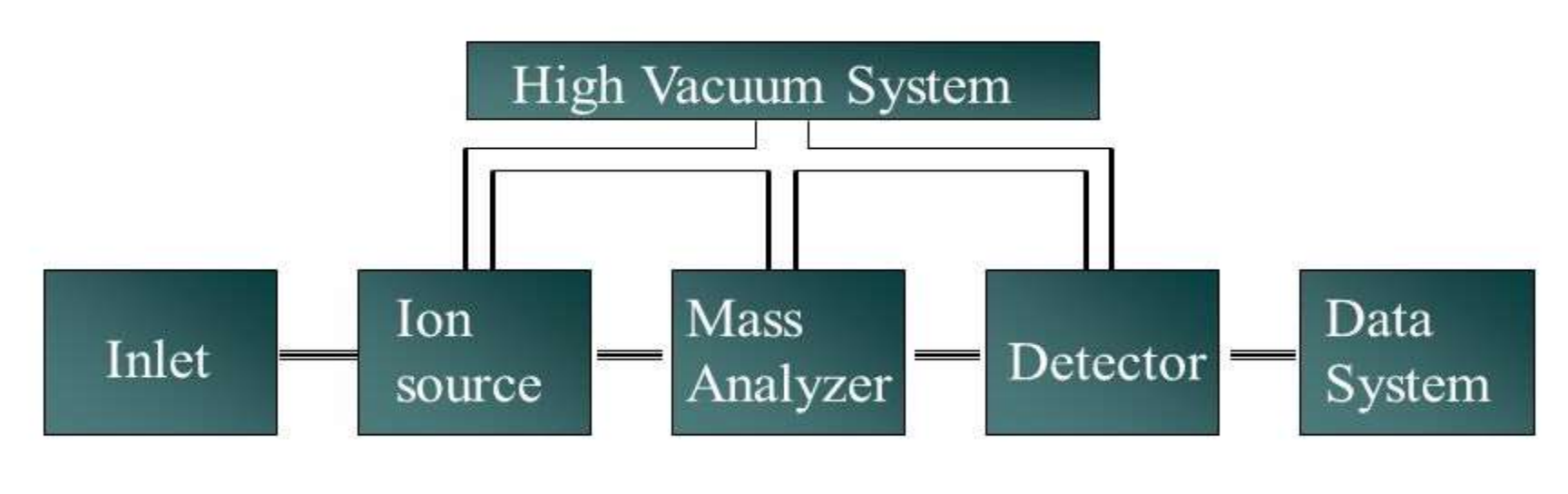
Detection

- The beam of ions passing through the machine is detected electrically.
- When an ion hits the metal box, its charge is neutralised by an electron jumping from the metal on to the ion.
- That leaves a space amongst the electrons in the metal, and the electrons in the wire shuffle along to fill it.
- A flow of electrons in the wire is detected as an electric current which can be amplified and recorded. The more ions arriving, the greater the current.



Components of MS

1) Inlet – Introduction of sample
2) Source - Produces gas-phase ions from the sample.
3) Mass analyzer - Resolves ions based on their m/z ration.
4) Detector - Detects ions resolved by the mass analyzer.
5) Data System - Result



Inlet System

- SOLIDS SAMPLES with lower vapour pressure

  directly inserted into the ionization chamber and volatilization is controlled by heating the probe.

- LIQUIDS are handled by hypodermic needles

  injection through a silicon rubber dam.

- GASES SAMPLES are leaked into the ionisation chamber

  directly by the help of mercury manometer.

The Sample Inlet System

Batch Inlets

- The batch inlet system is considered the most common and simplest inlet system. Normally, the inside of the system is lined with glass to elude losses of polar analyte by adsorption.
- This system externally volatizes the sample which leaks into an empty ionization region. Boiling points up to 500 degrees C of gaseous and liquid samples can be used on typical systems.
- The system's vacuum contains a sample pressure. Liquids are introduced using a microliter syringe into a reservoir; gases are enclosed in a metering area that is confined between two valves before being expanded into a reservoir container.

The Sample Inlet System

- Liquids that have boiling points lower than 500 degrees C can not be used in the system because the reservoir and tubing need to be kept at high temperatures by ovens and heating tapes.
- This is to ensure that the liquid samples are transformed to the gaseous phase and then leaked through a metal or glass diaphragm containing pinholes to the ionization area.

The Sample Inlet System

The Direct Probe Inlet:

- A direct probe inlet is for small quantities of sample, solids, and nonvolatile liquids. Solids and nonvolatile liquids are injected through a probe, or sample holder.
- The probe is inserted through a vacuum lock. Unlike the batch inlet, the sample will need to be cooled and/or heated on the probe.
- The probe is placed extremely close (a few millimeters) to the ionization source, where the slit leads to the spectrometer.

The Sample Inlet System

Electrophoretic Inlets

- Chromatographic systems and Capillary Electrophoretic units are often coupled with mass spectrometers in order to allow separation and identification of the components in the sample.
- If these systems and units are linked with a mass spectrometer, then other specialized inlets, Electrokinetic and Pressure injection, are required.
- Electrokinetic and pressure injection controls the amount of volume injected by the duration of the injection, which typically range between 5 to 50 nL.

Ion Source

- Since the mass analyzer utilizes only **gaseous ions i.e.,** starting point of mass spectrometric analysis is formation of gaseous analyte ions.
- Non –Volatile solids are first converted in to gases and from the gaseous sample the ions are produced in a **Box like enclosure called Ion Source.**
- **Function -** Produces ion without mass discrimination of the sample and Accelerates ions into the mass analyzer.

Ion Source

Desorption- A phenomenon whereby a substance is released from or through a surface.

Sorption- A process whereby one substance attached to another. It can be of two types

1) **Adsorption-** Adhesion of atoms ions or molecules from a gas liquid or dissolved solid to a surface. This process create a adsorbate on the surface of adsorbent.
2) **Absorption-** A process in which atoms ions or molecules are taken up by a bulk phase i.e solid liquid or gas. Molecules are taken up by the volume not by the surface.

Categories Of Ion Sources

Gas Phase Sources

- Electron Impact Ionization (EI)
- Chemical Ionization (CI)
- Field Ionizations (FI)

Desorption Sources

- Field Desorption (FD)
- Electrospray Ionization (ESI)
- Matrix assisted desorption / Ionisation (MALDI)
- Plasma desorption (PD)
- Fast Atom Bombardment (FAB)
- Thermospray Ionization (TS)
- Secondary Ion Mass Spectrometry (SIMS)

Mass Analyser

An ion, after leaving ion source, the ions are separated according to their m/e ratio.

In this area, the ions are accelerated by both electrostatic and magnetically

Types:-

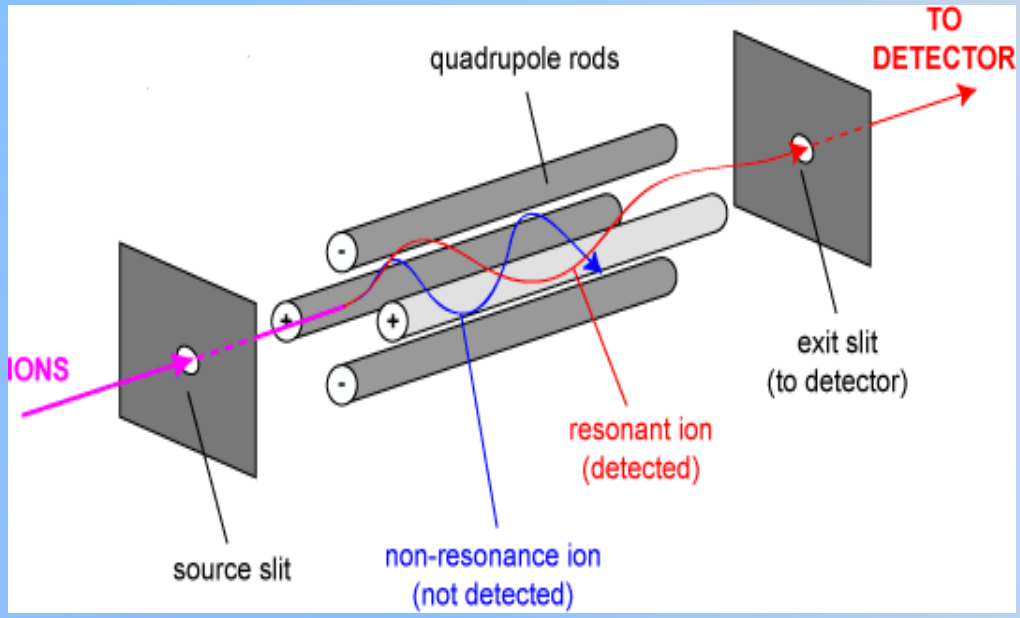
- **Magnetic sector mass analysers**
- **Double focussing analysers**
- **Quadrupole mass analysers**
- **Time of Flight analysers (TOF)**
- **Ion trap analyser**
- **Ion cyclotron analyser**

Quadrupole mass analyzer

- A quadrupole mass spectrometer contains four parallel cylindrical rods which can scan or filter sample ions based on their mass-to-charge ratio.
- Opposing rods are connected electrically and a radio frequency voltage is applied between the pairs of rods.
- Ions travel between the rods and only ions with a specific mass-to-charge ratio will exit the quadrupole; other ions will collide with the rods.
- The desired mass-to-charge ratio can be altered by changing the applied voltage.

Quadrupole mass analyzer

- Triple quadrupole mass spectrometry makes use of the same technology, but uses a linear series of three quadrupoles to improve sensitivity and selectivity.
- This type of spectrometry is useful when studying particular ions of interest since it is able to stay tuned to a single ion for extended periods of time.



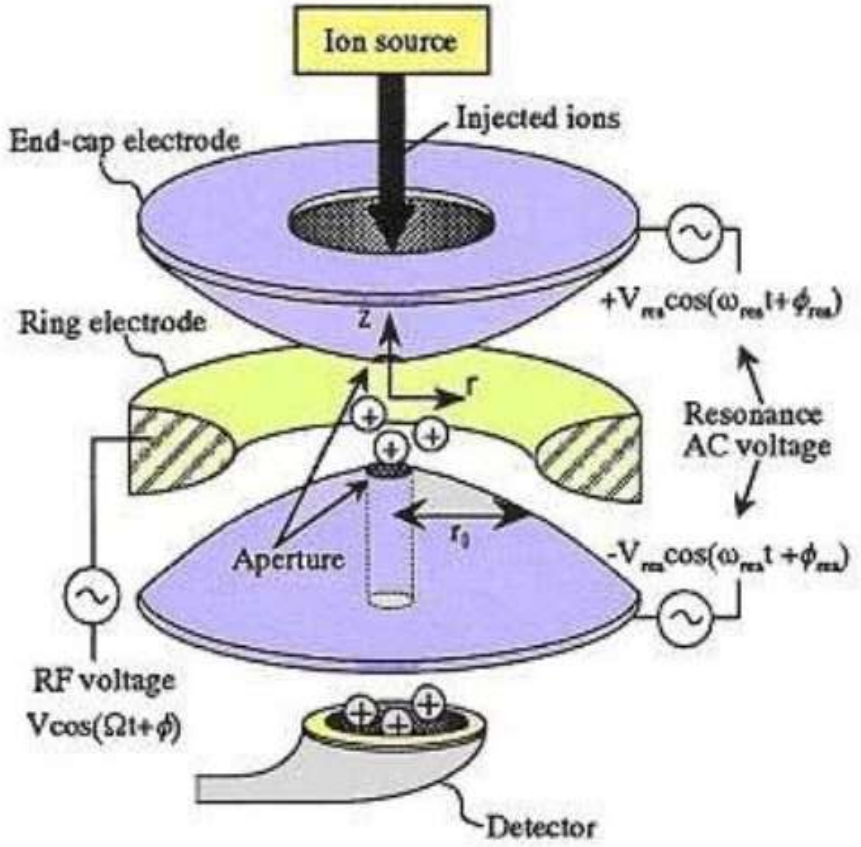
Ion Trap Mass Analyzer

- This analyzer employs similar principles as the quadrupole analyzer mentioned above, it uses an electric field for the separation of the ions by mass to charge ratios.
- The analyzer is made with a ring electrode of a specific voltage and grounded end cap electrodes.
- The ions enter the area between the electrodes through one of the end caps. After entry, the electric field in the cavity due to the electrodes causes the ions of certain m/z values to orbit in the space.
- As the radio frequency voltage increases, heavier mass ion orbits become more stabilized and the light mass ions become less stabilized, causing them to collide with the wall, and eliminating the possibility of traveling to and being detected by the detector.

Ion Trap Mass Analyzer

Ion traps are ion trapping devices that make use of a three-dimensional quadrupole field to trap and mass-analyze ions

Offer good mass resolving power



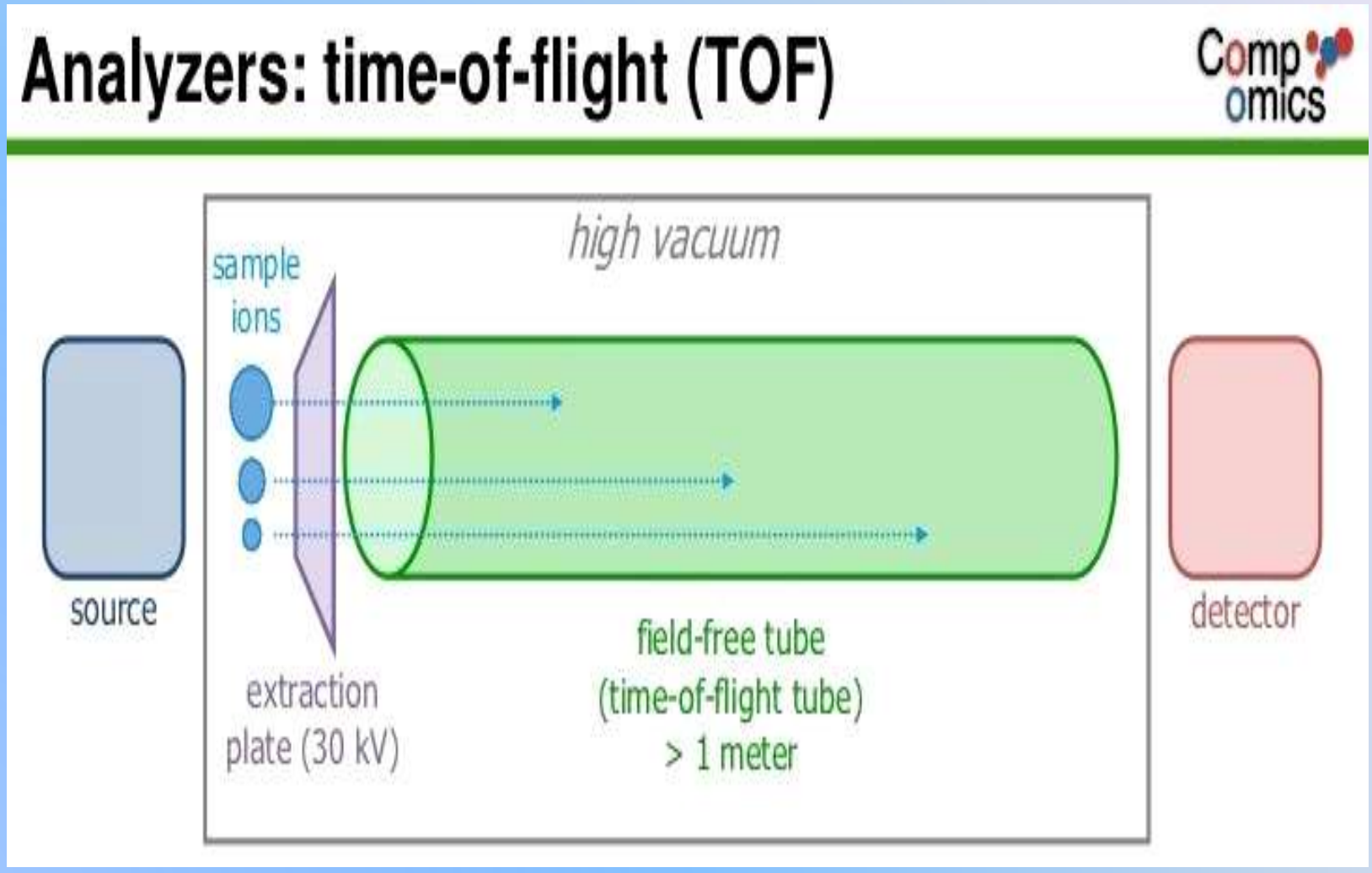
TOF Analyzers

- TOF Analyzers separate ions by time without the use of an electric or magnetic field.
- In a crude sense, TOF is similar to chromatography, except there is no stationary/ mobile phase, instead the separation is based on the kinetic energy and velocity of the ions.
- Ions are accelerated by an electric field of known strength.
- This acceleration results in an ion having the same kinetic energy as any other ion that has the same charge.

TOF Analyzers

- The velocity of the ion depends on the mass-to-charge ratio (heavier ions of the same charge reach lower speeds)
- The time that it subsequently takes for the ion to reach a detector at a known distance is measured.
- This time will depend on the velocity of the ion, and therefore is a measure of its mass-to-charge ratio.
- From this ratio and known experimental parameters, one can identify the ion.

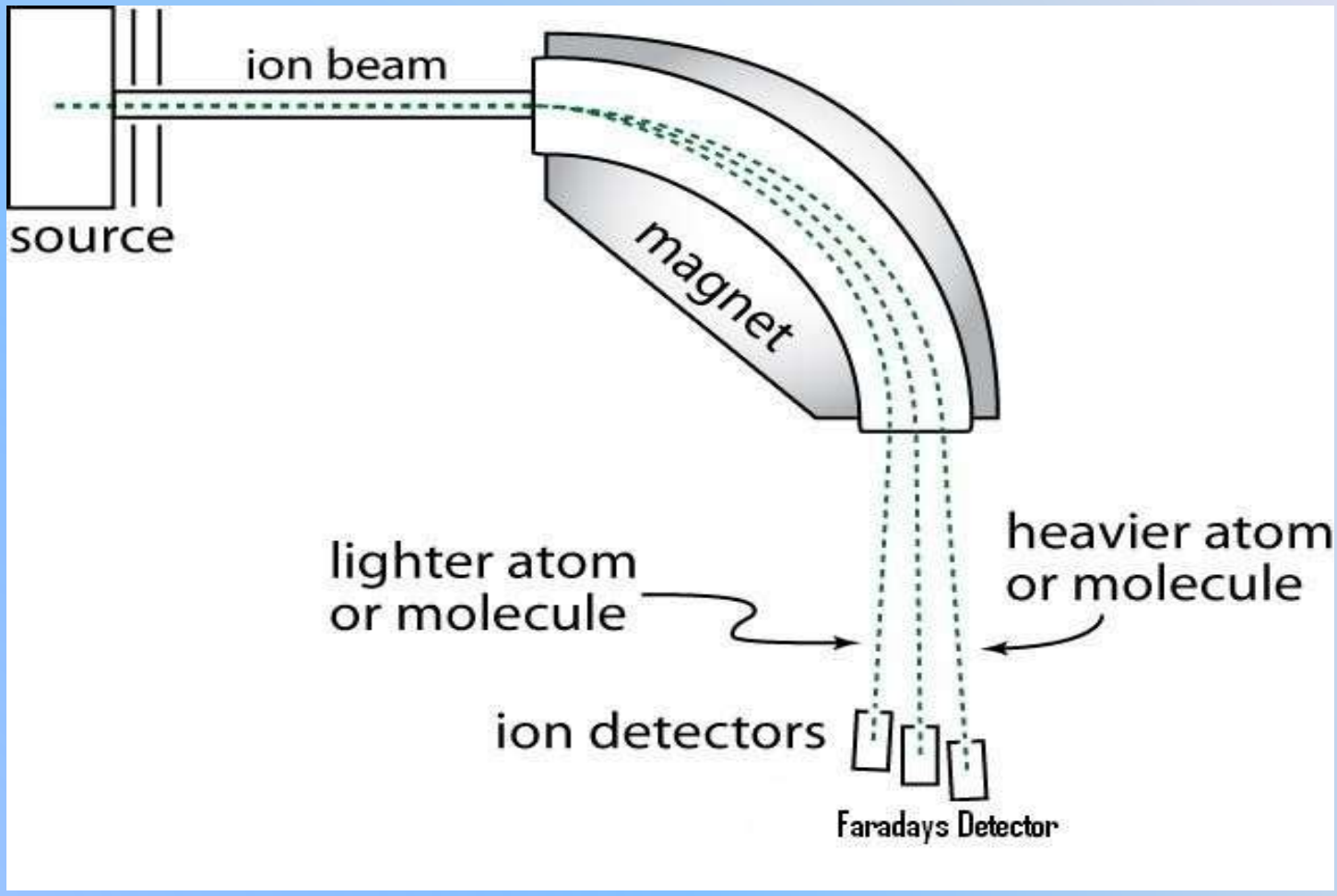
TOF Analyzers



Magnetic sector analyzers

- Similar to time of flight (TOF) analyzer mentioned earlier,
- In magnetic sector analyzers ions are accelerated through a flight tube.
- Where the ions are separated by charge to mass ratios. The difference between magnetic sector and TOF is that a magnetic field is used to separate the ions.
- As moving charges enter a magnetic field, the charge is deflected to a circular motion of a unique radius in a direction perpendicular to the applied magnetic field

Magnetic sector analyzers



Detectors

- Faraday cup
- Electron Multiplier
- photomultiplier
- Micro Channel Plate

Computational Methods

Broadly Classified Steps:

1. Number of polypeptide chains (subunits).
2. If subunits are too large, fragment them into shorter polypeptide chains.
3. Determine the amino acid composition of each polypeptide chain.
4. Complete the sequence by comparing overlaps of different sets of fragments.
5. For reference sequence, check it out in BLAST, UNIPORT, NCBI protein sequencing method.

Computational Methods

- Using a mass-based approach, each protein in a database is theoretically subjected to the same experimental conditions as the protein to be identified.
- Typically, this will involve an enzymatic digestion and possible secondary fragmentation.
- This produces a theoretical mass spectrum (or spectra) for each protein in the database.
- These theoretical mass spectra are compared with the experimental spectrum.
- Method of protein identification using mass spectrometry –
  - ✓ **Peptide mass fingerprinting**
  - ✓ **Peptide fragment fingerprinting**
  - ✓ **Tag-Based Approaches**

Peptide mass fingerprinting (PMF)

- **Peptide mass fingerprinting (PMF)** was the first available method of protein identification using mass spectrometry, and is still widely used.
- This method uses theoretical spectra each comprising the list of masses expected by an enzymatic digestion of each protein sequence in the reference database.
- PMF is popular and works well in practice because it is relatively fast to compute PMF scores against a database.
- For good quality samples belonging to well-characterized model organisms, PMF can in many cases produce protein identifications with high confidence, especially in organisms with smaller genomes.
- Sometimes a sample spectrum does not resemble any theoretical spectra in the protein database closely enough to make a confident identification

Peptide mass fingerprinting (PMF)

- This can happen for many reasons, such as unexpected post-translational or chemical modifications, splice variants, individual sequence variants (single nucleotide polymorphisms [SNPs], etc), or omissions and errors in the database.
- Steps –
  1) Scoring PMF
  2) Deciding on a threshold
- One of the limitations of PMF is its sensitivity to database size.
- A larger database has an elevated chance of the experimental masses randomly matching theoretical peptide masses in these databases, thereby decreasing the confidence of protein identifications using PMF.
- PMF Packages-
  - ✓ Aldente
  - ✓ Mascot
  - ✓ MS-Fit
  - ✓ Profound

Peptide Fragment fingerprinting (PFF)

- PFF approaches using PFF data are the current mainstream of high-throughput protein identification. Proteins are first digested with an enzyme, and then individual peptides are selected to undergo further fragmentation to yield PFF spectra.
- The set of these spectra, along with information such as the parent mass of these fragmented peptides, are then used in the database search.
- There are many dozens of scoring systems described in the literature, but in most cases these consist of two steps:
  1) Attributing a score for each protein in the database and
  2) Calculating a measure of confidence that the top-ranking identified protein is not a false positive.

Peptide Fragment fingerprinting (PFF)

- PFF is the method of choice for high-throughput applications due to the additional information gained from secondary fragmentation.
- This information makes the protein identification process less sensitive to effects such as protein modifications and can generate higher statistical confidence in the correct identification than traditional PMF.
- Some of the more popular PFF packages are
  - ✓ **Sequest**
  - ✓ **Popitam**
  - ✓ **Sonar**
  - ✓ **Protein Prospector**
  - ✓ **TANDEM**
  - ✓ **Phenyx**
  - ✓ **Spectrum Mill**

Tag-Based Approach

- Tag-based approaches begin with an attempt to extract peptide sequence information directly from the peptide fragmentation spectra.
- These methods are based on casting the problem into one of finding a maximum path length through a graph.
- There are 2 Tag-Based approaches-
  1) **De novo sequencing**
  2) **Tag-based search algorithms**
- Packages for De Novo Sequencing of MS Data and Tag-Based Protein Identification Engines are-

✓ **GutenTag**
✓ **Lutefisk**
✓ **InsPecT**
✓ **PEAKS**
✓ **MSBLAST**
✓ **FASTA**

Tag-Based Approach

De novo sequencing

- The process of inferring protein sequence from MS/MS data is known as de novo sequencing.
- Due to the high complexity of most MS/MS spectra, de novo sequencing tools often return short, ambiguous sequences known as “tags.”
- These tags are then searched against a database.
- Although many of these tags may randomly align with sections of protein sequence right across the genome, the correct protein identification is expected to have multiple alignments with sequence tags derived from the unknown protein.
- De novo sequencing quality is highly dependent on the precision of the mass spectrometer and the quality of the spectra.

Tag-Based Approach

Tag-based search algorithms

- Most current tag-based methods use a basic adaptation of the BLAST or FASTA algorithms.
- These are already in common use in the life sciences for gene and protein sequence alignments.
- For use in tag-based searching, the algorithms are modified for the much shorter peptide sequences usually generated by MS/MS, typically in the order of eight to 15 amino acids.
- Tag-based approaches are much faster than PFF searches

THANK YOU

ANALYSIS OF PROTEOMICS DATA

On Expasy Server

-Ms. Rupal Mishra

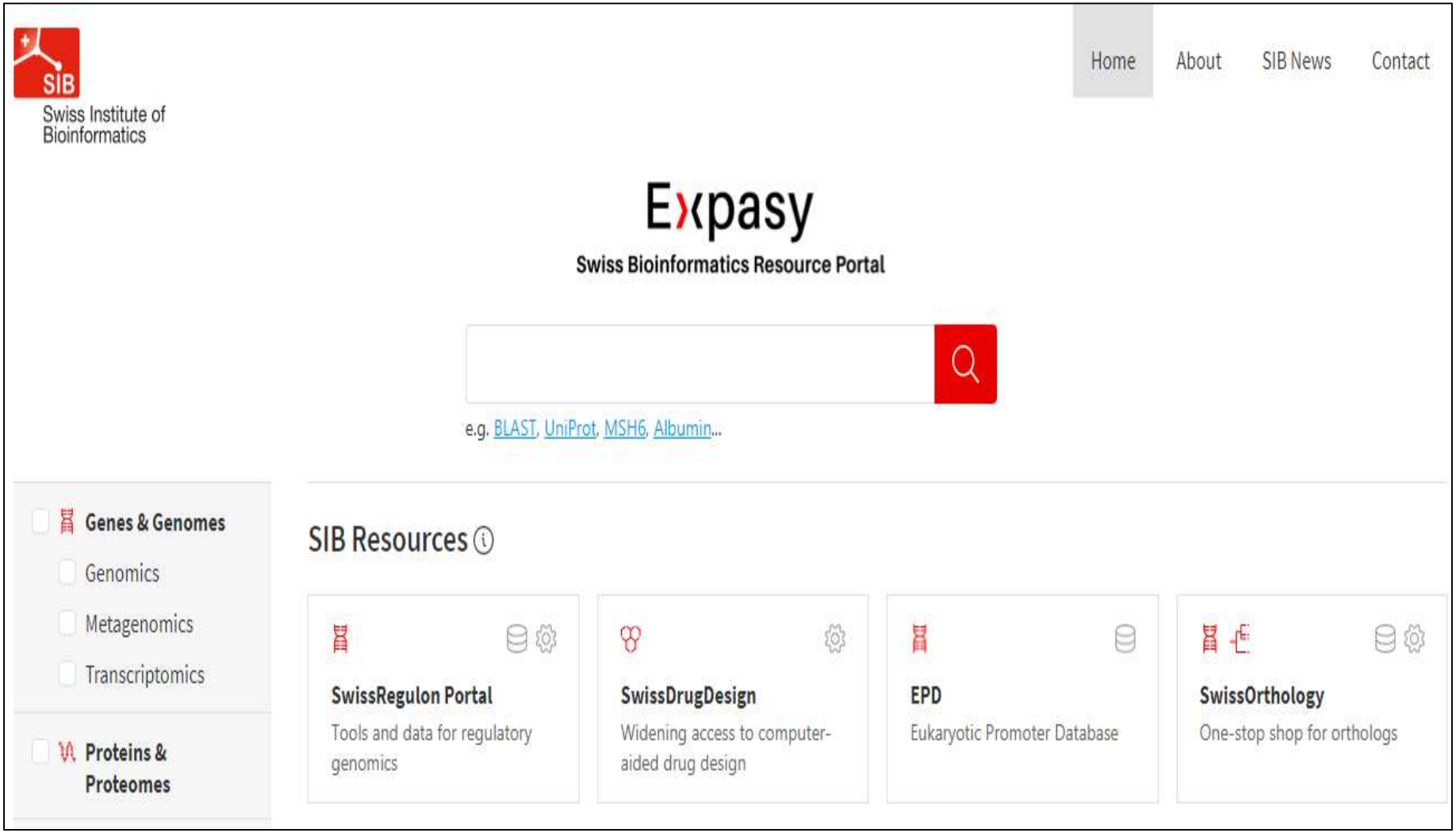
Expasy

- Expasy was created in August 1993 - the dawn of the internet era.
- At that time, it was referred to as '**ExPASy**, the Expert Protein Analysis System' as proteins were its primary focus.
- First life science website & among150 very first websites in the world.
- In June 2011, it became SIB Expasy Bioformatics Resources Portal (a diverse catalogue of bioinformatics resources developed by SIB Groups).
- The current version of Expasy (Expasy 3.0) was released in July 2020 following a massive user study and taking into account design, user experience and architecture aspects.

Expasy

- It is an extensible and integrative portal which provides access to over 160 databases and software tools, developed by SIB Groups and supporting a range of life science and clinical research domains, from genomics, proteomics and structural biology, to evolution and phylogeny, systems biology and medical chemistry.
- User-friendly search engine of Expasy allows you to seamlessly
  1) Query in parallel a subset of SIB databases through a single search
  2) Surface related information and knowledge from the complete set of >160 resources on the portal.
- Expasy provides information that is automatically aligned with the most recent release of each resources, thereby ensuring up-to-date information.

Expasy


SIB
Swiss Institute of Bioinformatics
Home
About
SIB News
Contact
Expasy
Swiss Bioinformatics Resource Portal
e.g. BLAST, UniProt, MSH6, Albumin...
Genes & Genomes
Genomics
Metagenomics
Transcriptomics
Proteins & Proteomes
SIB Resources
SwissRegulon Portal
Tools and data for regulatory genomics
SwissDrugDesign
Widening access to computer-aided drug design
EPD
Eukaryotic Promoter Database
SwissOrthology
One-stop shop for orthologs

Introduction

- Protein identification and analysis software performs a central role in the investigation of proteins from two- dimensional (2-D) gels and mass spectrometry.
- For protein identification, the user matches certain empirically acquired information against a protein database to define a protein as already known or as novel.
- For protein analysis, information in protein databases can be used to predict certain properties about a protein, which can be useful for its empirical investigation.
- The two processes are thus complementary.

Introduction

Protein Analysis tools on Expasy include:

1) **Compute pI/Mw** - a tool for predicting protein isoelectric point (pI) and molecular weight (Mw).
2) **ProtParam** - to calculate various physicochemical parameters.
3) **PeptideMass** - a tool for theoretically cleaving proteins and calculating the masses of their peptides and any known cellular or artifactual posttranslational modifications.
4) **PeptideCutter** - to predict cleavage sites of proteases or chemicals in protein sequences.
5) **ProtScale** - for amino acid scale representation, such as hydrophobicity plots.

Compute pI/MW

- This tool calculates the estimated pI and Mw of a specified Swiss-Prot/TrEMBL entry or a user-entered AA sequence.
- https://web.expasy.org/compute_pi/
- These parameters are useful if you want to know the approximate region of a 2-D gel where a protein may be found.
- To use the program, enter one or more Swiss-Prot/TrEMBL identification names (e.g., LACB_BOVIN) or accession numbers (e.g., P02754) into the text field, and select the “click here to compute pI/Mw” button.
- If one entry is specified, you will be asked to specify the protein’s domain of interest for which the pI and mass should be computed.

Compute pI/MW

- The domain can be selected from the hypertext list of features shown, if any, or by numerically specifying the domain start and end points.
- If more than one Swiss-Prot/TrEMBL identification name is entered, all proteins will automatically be processed to their mature forms, and pI and Mw values calculated for the resulting chains or peptides.
- If only fragments of the protein of interest are available in the database, no result will be given and an error message will be shown to highlight that the pI and mass cannot be returned accurately.

Compute pI/MW

- Some database entries have signal sequences or transit peptides of unknown length (e.g., Q00825; ATPI_ODOSI).
- In those cases, an average-length signal sequence or transit peptide is removed before the pI and mass computation is done.
- In Swiss-Prot release 35, the average signal sequence length is 22 amino acids for eukaryotes and viruses, 26 amino acids for prokaryotes and bacteriophages, and 30 for archaebacteria.
- Transit peptides have an average length of 55 amino acids in chloroplasts, 34 for mitochondria, 29 for microbodies, and 51 for cyanelles.

Compute pI/MW

If your protein of interest is not in the Swiss-Prot database, you can enter an AA sequence in standard single letter AA code into the text field, and select the “click here to compute pI/Mw” button. The predicted pI and Mw of your sequence will then be displayed.

```
LACB_BOVIN (P02754)


DE   BETA-LACTOGLOBULIN PRECURSOR (BETA-LG).

OS   BOS TAURUS (BOVINE).


The parameters have been computed for the following feature:

FT   CHAIN        17    178       BETA-LACTOGLOBULIN.


Considered sequence fragment:

LIVTQTMKGL DIQKVAGTWY SLAMAASDIS LLDAQSAPLR VYVEELKPTP EGDLEILLQK

WENGECAQKK IIAEKTKIPA VFKIDALNEN KVLVLDTDYK KYLLFCMENS AEPEQSLACQ

CLVRTPEVDD EALEKFDKAL KALPMHIRLS FNPTQLEEQC HI


Molecular weight: 18281.00


Theoretical pI: 4.83
```

Compute pI/MW

Alternatively to the verbose html output, the result for a list of Swiss-Prot/TrEMBL entries can also be retrieved in a numerical format, with minimal documentation. A file containing four columns—ID, AC, pI, and Mw—is generated and can be loaded into an external application, such as a spreadsheet program.

ASNA_MOUSE_1	O54984	4.81	38691.61
ARSA_MOUSE_1	P50428	5.50	52145.20
ARSB_MOUSE_1	P50429	6.39	55458.91
ARX_MOUSE_1	O35085	5.14	58490.34
ARY1_MOUSE_1	P50294	5.10	33713.36
ARY2_MOUSE_1	P50295	5.63	33701.41
ARY3_MOUSE_1	P50296	6.07	33685.69
ASAH1_MOUSE_1	Q9WV54	6.11	13797.05
ASAH1_MOUSE_2	Q9WV54	8.87	29017.27
ASAH1_MOUSE_1	Q9WV54	6.11	13797.05
ASAH1_MOUSE_2	Q9WV54	8.87	29017.27
ASCL1_MOUSE_1	Q02067	8.56	24740.54

Compute pI/MW

Expasy

Compute pI/Mw

Home | Contact

Compute pI/Mw tool

Compute pI/Mw is a tool which allows the computation of the theoretical pI (isoelectric point) and Mw (molecular weight) for a list of UniProt Knowledgebase (Swiss-Prot or TrEMBL) entries or for user entered sequences [reference].

Documentation is available.

Compute pI/Mw for Swiss-Prot/TrEMBL entries or a user-entered sequence

Please enter one or more UniProtKB/Swiss-Prot protein identifiers (ID) (e.g. *ALBU_HUMAN*) or UniProt Knowledgebase accession numbers (AC) (e.g. *P04406*), separated by spaces, tabs or newlines. Alternatively, enter a protein sequence in single letter code. The theoretical *pI* and *Mw* (molecular weight) will then be computed.

Or upload a file from your computer, containing one Swiss-Prot/TrEMBL ID/AC or one sequence per line: Choose File No file chosen

Resolution: Average or Monoisotopic

Click here to compute pI/Mw Reset

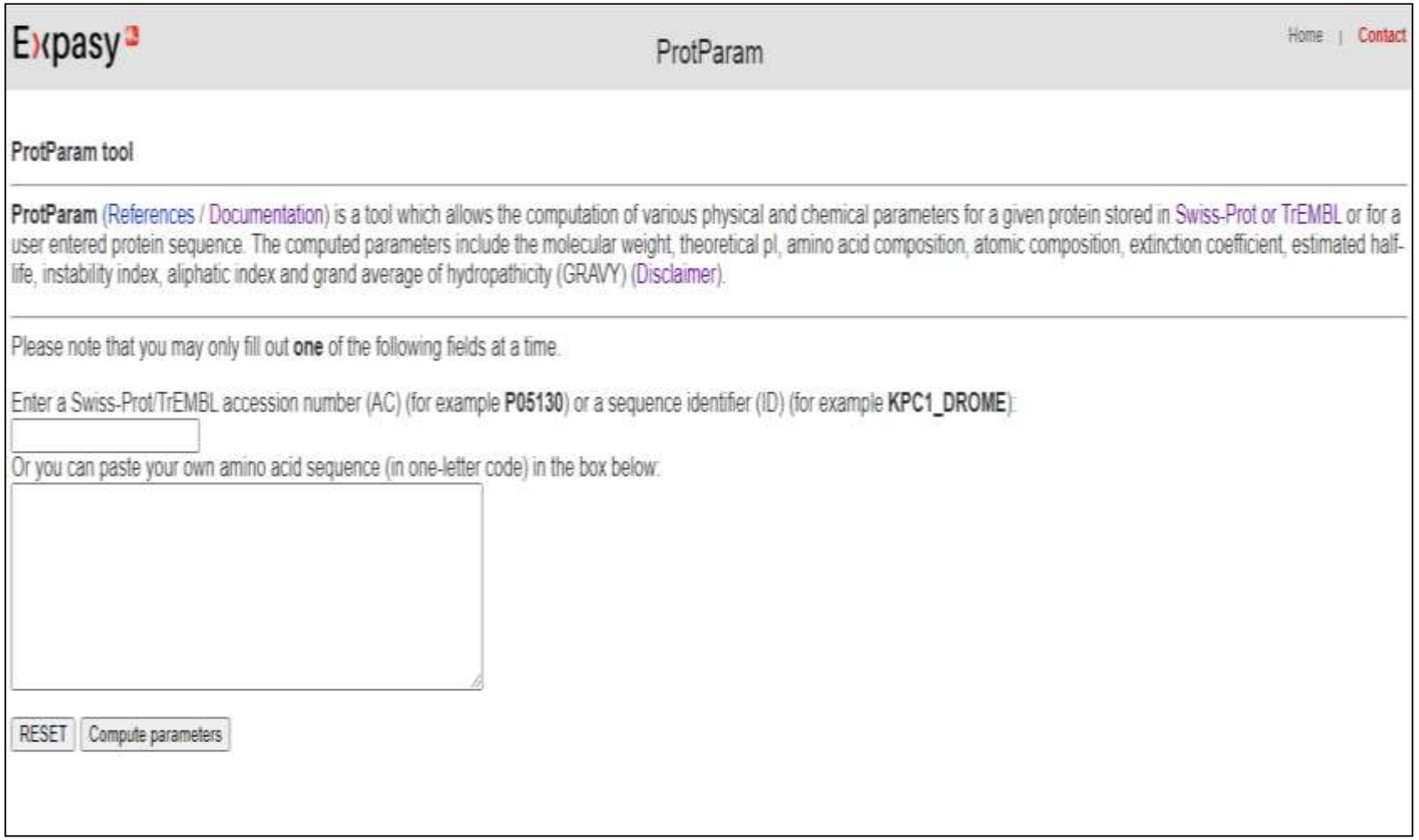
ProtParam

- ProtParam computes various physico-chemical properties that can be deduced from a protein sequence.
- https://web.expasy.org/protparam/
- No additional information is required about the protein under consideration.
- The protein can either be specified as a Swiss-Prot/TrEMBL accession number or ID, or in form of a raw sequence.
- White space and numbers are ignored.
- If you provide the accession number of a Swiss-Prot/TrEMBL entry, you will be prompted with an intermediary page that allows you to select the portion of the sequence on which you would like to perform the analysis.

ProtParam

- The parameters computed by ProtParam include –
  - molecular weight,
  - theoretical pI,
  - amino acid composition,
  - atomic composition,
  - extinction coefficient,
  - estimated half-life,
  - instability index,
  - aliphatic index
  - grand average of hydropathicity (GRAVY).

ProtParam


Expasy
ProtParam
Home | Contact
ProtParam tool
ProtParam (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Disclaimer).
Please note that you may only fill out one of the following fields at a time.
Enter a Swiss-Prot/TrEMBL accession number (AC) (for example P05130) or a sequence identifier (ID) (for example KPC1_DROME):
Or you can paste your own amino acid sequence (in one-letter code) in the box below:
RESET
Compute parameters

PeptideMass

- This program is designed to calculate the theoretical masses of peptides generated by the chemical or enzymatic cleavage of proteins, to assist in the interpretation of peptide mass fingerprinting and peptide mapping experiments.
- https://web.expasy.org/peptide_mass/
- Protein sequences can be provided by the user or can be a code name for a protein in the UniProt Knowledgebase.
- When proteins of interest are specified from UniProtKB/Swiss-Prot, the program considers all annotations for that protein in the database, and uses these in order to generate the correct peptide masses and warn users about peptides that are not likely to be found when undertaking peptide mass fingerprinting.

PeptideMass

- Many protein from Swiss-Prot has annotations that describe discrete posttranslational modifications the masses of these modifications will be considered in peptide mass calculations.
- The mass effects of artifactual protein modifications such as the oxidation of methionine or acrylamide adducts on cysteine residues can also be considered.
- The program can supply warnings where peptide masses may be subject to change from protein isoforms, database conflicts, or mRNA splicing variation.

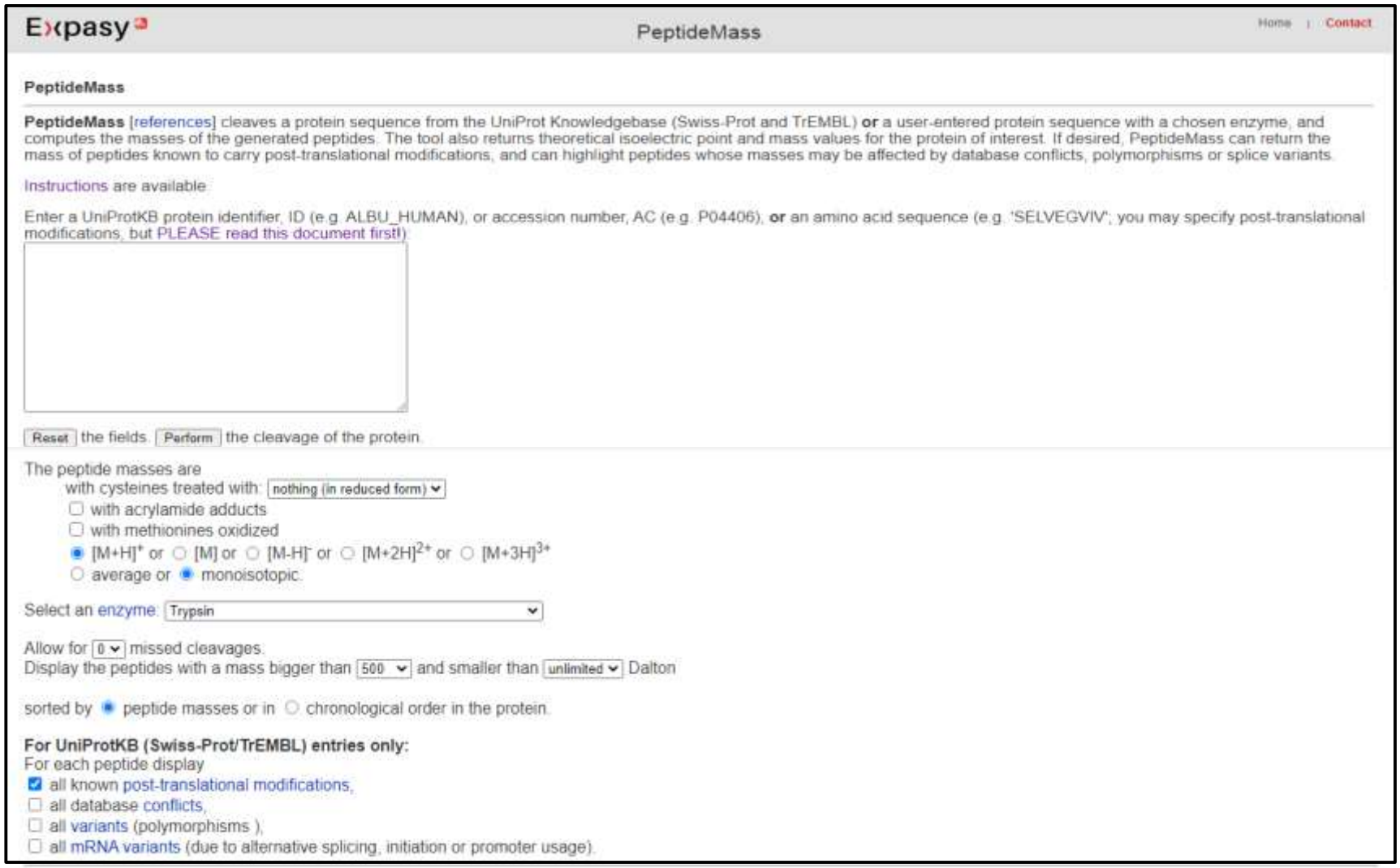
PeptideMass

- To use the program, enter one or more Swiss-Prot identification names (e.g.,TKN1_HUMAN) or any Swiss-Prot/TrEMBL accession number (e.g., P20366) into the text field, or enter a protein sequence of interest using the standard one-letter AA code.
- User-specified sequences should not contain the character X, but can contain the character J, to represent either Ile or Leu, which are of the same mass.

PeptideMass

- You can select to exclude masses below a certain threshold (e.g., 500 Daltons) which might be too small to be visible in a mass spectrum.
- The PeptideMass output will include the portions of the sequence covered by only the fragments that are above that threshold.
- Finally, click on the “Perform” button to send data to the program.

PeptideMass



Expasy — PeptideMass — Home | Contact

PeptideMass

PeptideMass [references] cleaves a protein sequence from the UniProt Knowledgebase (Swiss-Prot and TrEMBL) **or** a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides. The tool also returns theoretical isoelectric point and mass values for the protein of interest. If desired, PeptideMass can return the mass of peptides known to carry post-translational modifications, and can highlight peptides whose masses may be affected by database conflicts, polymorphisms or splice variants.

Instructions are available.

Enter a UniProtKB protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SELVEGVIV'; you may specify post-translational modifications, but PLEASE read this document first!)

Reset the fields. Perform the cleavage of the protein.

The peptide masses are
with cysteines treated with: nothing (in reduced form)
- with acrylamide adducts
- with methionines oxidized
- [M+H]$^+$ or [M] or [M-H]$^-$ or [M+2H]$^{2+}$ or [M+3H]$^{3+}$
- average or monoisotopic.

Select an enzyme: Trypsin

Allow for 0 missed cleavages.
Display the peptides with a mass bigger than 500 and smaller than unlimited Dalton

sorted by peptide masses or in chronological order in the protein.

For UniProtKB (Swiss-Prot/TrEMBL) entries only:
For each peptide display
- all known post-translational modifications,
- all database conflicts,
- all variants (polymorphisms),
- all mRNA variants (due to alternative splicing, initiation or promoter usage).

PeptideCutter

- PeptideCutter predicts potential substrate cleavage sites, cleaved by proteases or chemicals in a given protein sequence. The tool returns the query sequence with the possible cleavage sites mapped on it and/or a table of cleavage site positions.
- https://web.expasy.org/peptide_cutter/
- Protease digestion can be useful if one wants to carry out experiments on a portion of a protein, separate the domains in a protein, remove a tag protein when expressing a fusion protein, or make sure that the protein under investigation is not sensitive to endogenous proteases.

PeptideCutter

- Different forms of output of the results are available:

1) The first list is arranged alphabetically according to enzyme names.
2) The second list displays sequentially all cleavage sites in the sequence and the respective cleaving enzymes from the N- to the C-terminus.
3) A third option for output is a map of cleavage sites.

PeptideCutter

- The protein sequence can be entered in the form of a Swiss-Prot/TrEMBL accession number, a raw sequence, or a sequence in FASTA format, in one-letter amino acid code.
- Letters that do not correspond to an amino acid code (*B, J, O, U, X, or Z) will* cause an error message, and the user is required to correct the input. Please note that only one sequence can be entered at a time.
- You have the possibility to select one or a group of enzymes and chemicals.
- You can also ask the program to consider only enzymes that cut the sequence a chosen number of times, which may be of particular interest if you have selected a large number of cleavage agents.

PeptideCutter

PeptideCutter [references / documentation] predicts potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. PeptideCutter returns the query sequence with the possible cleavage sites mapped on it and /or a table of cleavage site positions.

Enter a UniProtKB (Swiss-Prot or TrEMBL) protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SERVELAT'):

Perform the cleavage of the protein. Reset the fields.

Please, select

- (•) all available enzymes and chemicals
- ( ) only the following selection of **enzymes and chemicals**

☐ Arg-C proteinase	☐ Asp-N endopeptidase	☐ Asp-N endopeptidase + N-terminal Glu
☐ BNPS-Skatole	☐ Caspase1	☐ Caspase2
☐ Caspase3	☐ Caspase4	☐ Caspase5
☐ Caspase6	☐ Caspase7	☐ Caspase8
☐ Caspase9	☐ Caspase10	
☐ Chymotrypsin-high specificity (C-term to [FYW], not before P)	☐ Chymotrypsin-low specificity (C-term to [FYWML], not before P)	
☐ Clostripain (Clostridiopeptidase B)	☐ CNBr	☐ Enterokinase
☐ Factor Xa	☐ Formic acid	☐ Glutamyl endopeptidase
☐ GranzymeB	☐ Hydroxylamine	☐ Iodosobenzoic acid
☐ LysC	☐ LysN	☐ NTCB (2-nitro-5-thiocyanobenzoic acid)
☐ Neutrophil elastase		
☐ Pepsin (pH1.3)	☐ Pepsin (pH>2)	☐ Proline-endopeptidase
☐ Proteinase K	☐ Staphylococcal peptidase I	☐ Tobacco etch virus protease
☐ Thermolysin	☐ Thrombin	☐ Trypsin

( ) for the following enzymes an additional, more **sophisticated model** can be applied that attributes a probability of cleavage to each site :

Chymotrypsin
Trypsin

Please enter the lowest cleavage probability that you would like to be displayed: [ ]%

Please indicate the way you would like the cleavage sites to be displayed

- ☑ Map of cleavage sites. Please select the number of amino acid within one block: 60
- ☑ Table of sites, sorted alphabetically by enzyme and chemical name
- ☐ Table of sites, sorted sequentially by amino acid number

Please indicate which enzymes to include in the display

- (•) All enzymes and chemicals
- ( ) Enzymes and chemicals cleaving exactly [ ] times
- ( ) Enzymes and chemicals cleaving at least [ ] times, and at most [ ] times

ProtScale

- ProtScale allows to compute and represent (in the form of a two-dimensional plot) the profile produced by any amino acid scale on a selected protein.
- https://web.expasy.org/protscale/
- An amino acid scale is defined by a numerical value assigned to each type of amino acid.
- The most frequently used scales are hydrophobicity scales, most of which were derived from experimental studies on partitioning of peptides in apolar and polar solvents, with the goal of predicting membrane-spanning segments that are highly hydrophobic, and secondary structure conformational parameter scales.
- In addition, many other scales exist which are based on different chemical and physical properties of the amino acids.

ProtScale

- ProtScale can be used with 50 predefined scales entered from the literature.
- The scale values for the 20 amino acids, as well as a literature reference, are provided on ExPASy for each of these scales.
- To generate data for a plot, the protein sequence is scanned with a sliding window of a given size.
- At each position, the mean scale value of the amino acids within the window is calculated, and that value is plotted for the midpoint of the window.
- You can set several parameters that control the computation of a scale profile, such as the window size, the weight variation model, the window edge relative weight value, and scale normalization.

ProtScale

Interpreting Results

- The method of sliding windows, and hence ProtScale, only provides a raw signal and does not include interpretation of the results in terms of a score.
- When interpreting the results, one should consider only strong signals.
- In order to confirm a possible interpretation, one could slightly change the window size, or replace the scale by another similar one (e.g., two different hydrophobicity scales), and ensure that the strong signal is still present.

ProtScale

ProtScale [Reference / Documentation] allows you to compute and represent the profile produced by any amino acid scale on a selected protein.

An **amino acid scale** is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but many other scales exist which are based on different chemical and physical properties of the amino acids. This program provides 57 predefined scales entered from the literature.

Enter a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL accession number (AC) (e.g. **P05130**) or a sequence identifier (ID) (e.g. **KPC1_DROME**):

Or you can paste your own sequence in the box below:

Please choose an amino acid scale from the following list. To display information about a scale (author, reference, amino acid scale values) you can click on its name.

- ○ Molecular weight
- ○ Bulkiness
- ○ Polarity / Grantham
- ○ Recognition factors
- ○ Hphob. OMH / Sweet et al.
- ◉ Hphob. / Kyte & Doolittle
- ○ Hphob. / Bull & Breese
- ○ Hphob. / Guy
- ○ Hphob. / Miyazawa et al.
- ○ Hphob. / Roseman
- ○ Hphob. / Wolfenden et al.
- ○ Hphob. HPLC / Wilson & al
- ○ Hphob. HPLC pH3.4 / Cowan
- ○ Hphob. / Rf mobility
- ○ HPLC / TFA retention
- ○ HPLC / retention pH 2.1
- ○ % buried residues
- ○ Hphob. / Chothia
- ○ Ratio hetero end/side
- ○ Average flexibility
- ○ beta-sheet / Chou & Fasman
- ○ alpha-helix / Deleage & Roux
- ○ beta-turn / Deleage & Roux
- ○ alpha-helix / Levitt
- ○ beta-turn / Levitt
- ○ Antiparallel beta-strand
- ○ A.A. composition
- ○ Relative mutability
- ○ Number of codon(s)
- ○ Polarity / Zimmerman
- ○ Refractivity
- ○ Hphob. / Eisenberg et al.
- ○ Hphob. / Hopp & Woods
- ○ Hphob. / Manavalan et al.
- ○ Hphob. / Fauchere et al.
- ○ Hphob. / Janin
- ○ Hphob. / Rao & Argos
- ○ Hphob. / Tanford
- ○ Hphob. / Welling & al
- ○ Hphob. HPLC / Parker & al
- ○ Hphob. HPLC pH7.5 / Cowan
- ○ HPLC / HFBA retention
- ○ Transmembrane tendency
- ○ HPLC / retention pH 7.4
- ○ % accessible residues
- ○ Hphob. / Rose & al
- ○ Average area buried
- ○ alpha-helix / Chou & Fasman
- ○ beta-turn / Chou & Fasman
- ○ beta-sheet / Deleage & Roux
- ○ Coil / Deleage & Roux
- ○ beta-sheet / Levitt
- ○ Total beta-strand
- ○ Parallel beta-strand
- ○ A.A. comp. in Swiss-Prot

Window size: 9

Relative weight of the window edges compared to the window center (in %): 100

Weight variation model (if the relative weight at the edges is < 100%): ◉ linear ○ exponential

Do you want to normalize the scale from 0 to 1? ○ yes ◉ no

If you need more information about how to set these parameters, please click here.

Submit Reset

THANK YOU

INTERPRO

-Ms. Rupal Mishra

Protein Signatures

- Protein Signature - an amino acid sequence associated with a protein characteristic.

Pea	legumin(leg A)	TCTCTATAAATTA	CGCATCA	CTCTTCATGGCT
Pea	legumin(leg B)	TCTCTATAAATTA	CGCATCA	CTCTTCATGGCT
Pea	legumin(leg C)	TCTCTATAAATTA	CGCATCA	CTCTTCATGGCT
Vicia faba	legumin (LeB4)	TTCCTATAAATCA	TTCACCA	GTCACAATGTCC
Soybean	lectin (Le1)	CTAGTATAAATAG	TGCATAC	AAAGCAATGGCT
Kidney bean	lectin (pPVL134)	GTTGTATAAATAG	AATGCAT	GAATGCATGATC
Castor bean	lectin (ricin)	TCTGTATTAATTT	GACAGCC	TCAAGGATGAAA
Kidney bean	lectin (Lec1)	GTTGTATAAATAG	TGCATGA	GCATACATGGCT

What's value of signatures?

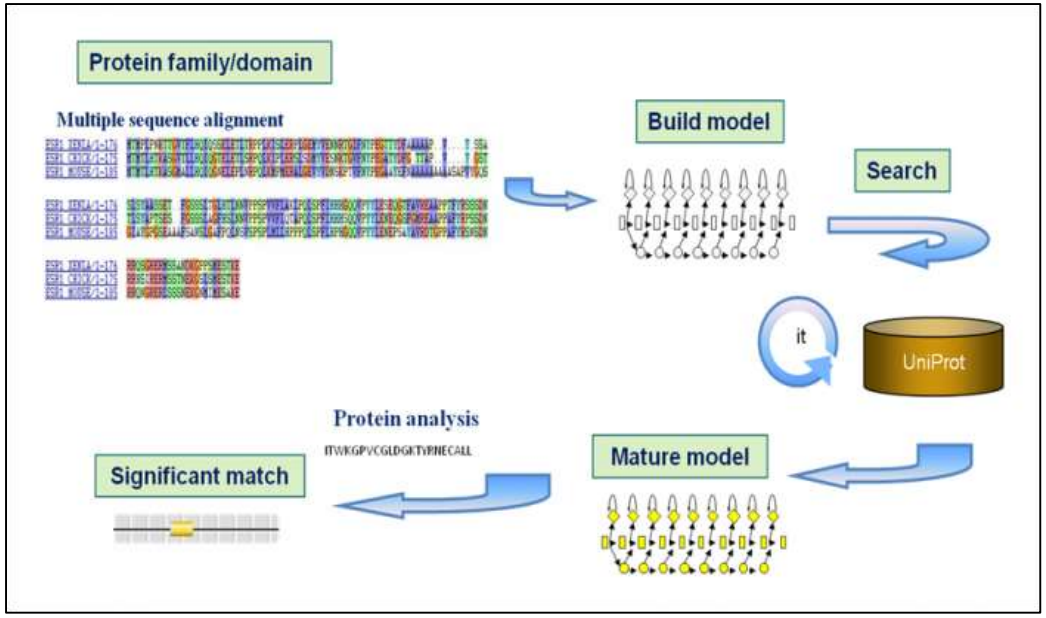
- Better at finding proteins with common function
  - Find more distant homologues than BLAST
- Classification of proteins
  - Associate proteins that share: { Function, Domains, Sequence, Structure }
- Annotation of protein sequences
  - Define conserved regions of a protein
    - e.g. { location and type of domains, key structural or functional sites }

How are protein signatures made?

- In order to classify proteins into families and to predict the presence of important domains or sequence features, we require computational tools. One set of such tools are the predictive models known as protein signatures.
- There are different types of signatures, built using different computational approaches.
- However, their common starting point is a multiple sequence alignment of proteins sharing a set of characteristics (e.g. belonging to the same family or sharing a domain).
- When building the initial model, the level of amino acid conservation at different positions in the alignment is taken into account.

How are protein signatures made?

- The model is then used to search a protein database in an iterative manner, refining the model as more distantly related sequences in the database are identified.
- Once the model is mature, the signature is ready and can be used for protein sequence analysis.



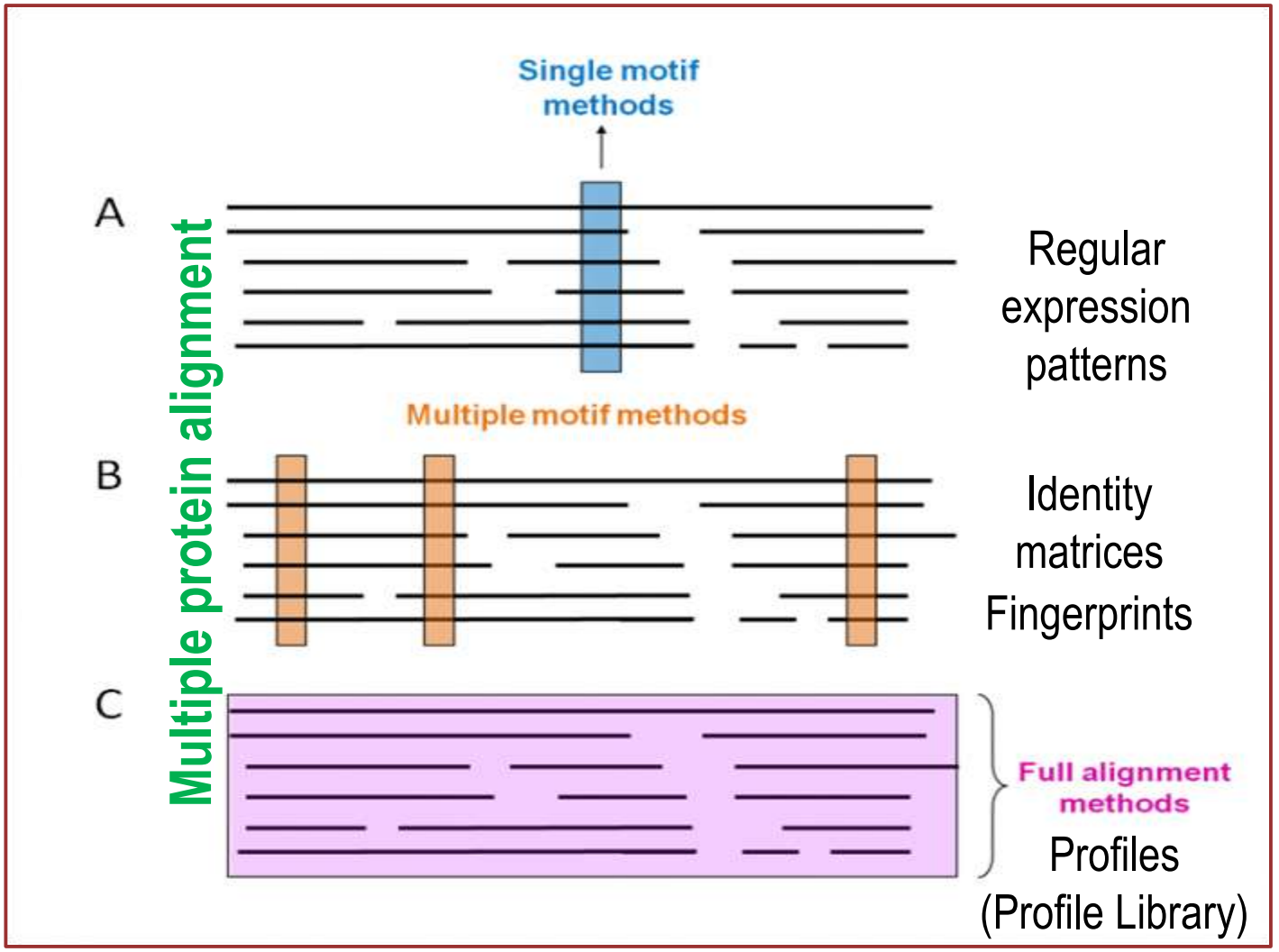
Terms to recollect

- **Patterns-** Many important sequence features, such as binding sites or the active sites of enzymes, consist of only a few amino acids that are essential for protein function. Patterns are very good at recognizing such features. They are built by identifying these regions in multiple sequence alignments.
- **Fingerprints-** A fingerprint is a group of conserved motifs taken from a multiple sequence alignment - together, the motifs form a characteristic signature for the aligned protein family.
- **Profiles -** are used to model protein families and domains. They are built by converting multiple sequence alignments into position-specific scoring systems (PSSMs). Amino acids at each position in the alignment are scored according to the frequency with which they occur.

Terms to recollect

- **Motif-** A **motif** is a short conserved sequence pattern associated with distinct functions of a protein or DNA. It is often associated with a distinct structural site performing a particular function.
- **Domains-** Distinct functional and/or structural units in a protein. Usually they are responsible for a particular function or interaction, contributing to the overall role of a protein. Domains may exist in a variety of biological contexts, where similar domains can be found in proteins with different functions.
- **Family-** A protein **family** is a group of evolutionarily-related proteins. Proteins in a **family** descend from a common ancestor and typically have similar three-dimensional structures, functions, and significant sequence similarity.

Types of Protein signatures



Introduction

- InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites.
- https://www.ebi.ac.uk/interpro/
- To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several collaborating databases (referred to as member databases) that collectively make up the InterPro consortium.
- A key value of InterPro is that it combines protein signatures from these member databases into a single searchable resource, capitalizing on their individual strengths to produce a powerful integrated database and diagnostic tool.

Introduction

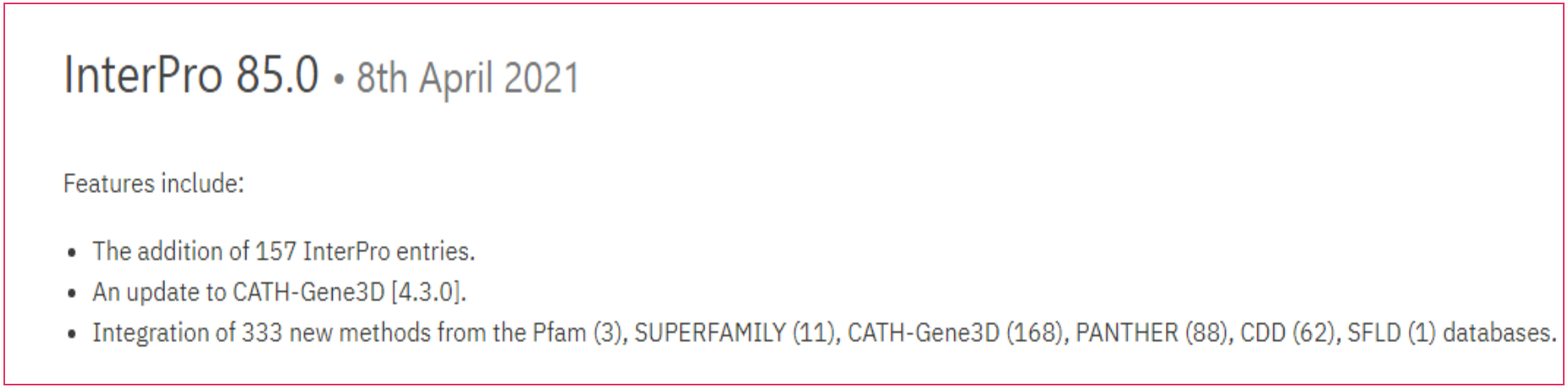
- They add more value to InterPro entries by providing detailed functional annotation as well as adding relevant GO terms that enable automatic annotation of millions of GO terms across the protein sequence databases.
- A **G**ene **O**ntology annotation represents a link between a gene product type and a molecular function, biological process, or cellular component type.
- A link, in other words, between the gene product and what that product is capable of doing, what biological processes it contributes to, and where in the cell it is found.

Introduction

- InterPro integrates signatures from the following 13 member databases: CATH, CDD, HAMAP, MobiDB Lite, Panther, Pfam, PIRSF, PRINTS, Prosite, SFLD, SMART, SUPERFAMILY AND TIGRfams.
- The member databases use a variety of different methods to classify proteins.
- Each of the databases has a particular focus (e.g. protein domains defined from structure, or full length protein families with shared function).
- InterPro integrates the signatures from the member databases into InterPro entries and to identify where different member database entries are the same entity.

Introduction

- You can use the InterPro website to obtain information about individual protein families, domains, important sites, perform a sequence search or browse through InterPro annotations.
- InterPro is updated approximately every 8 weeks. The release notes page contains information about what has changed in each release.
- All information in InterPro is freely available.

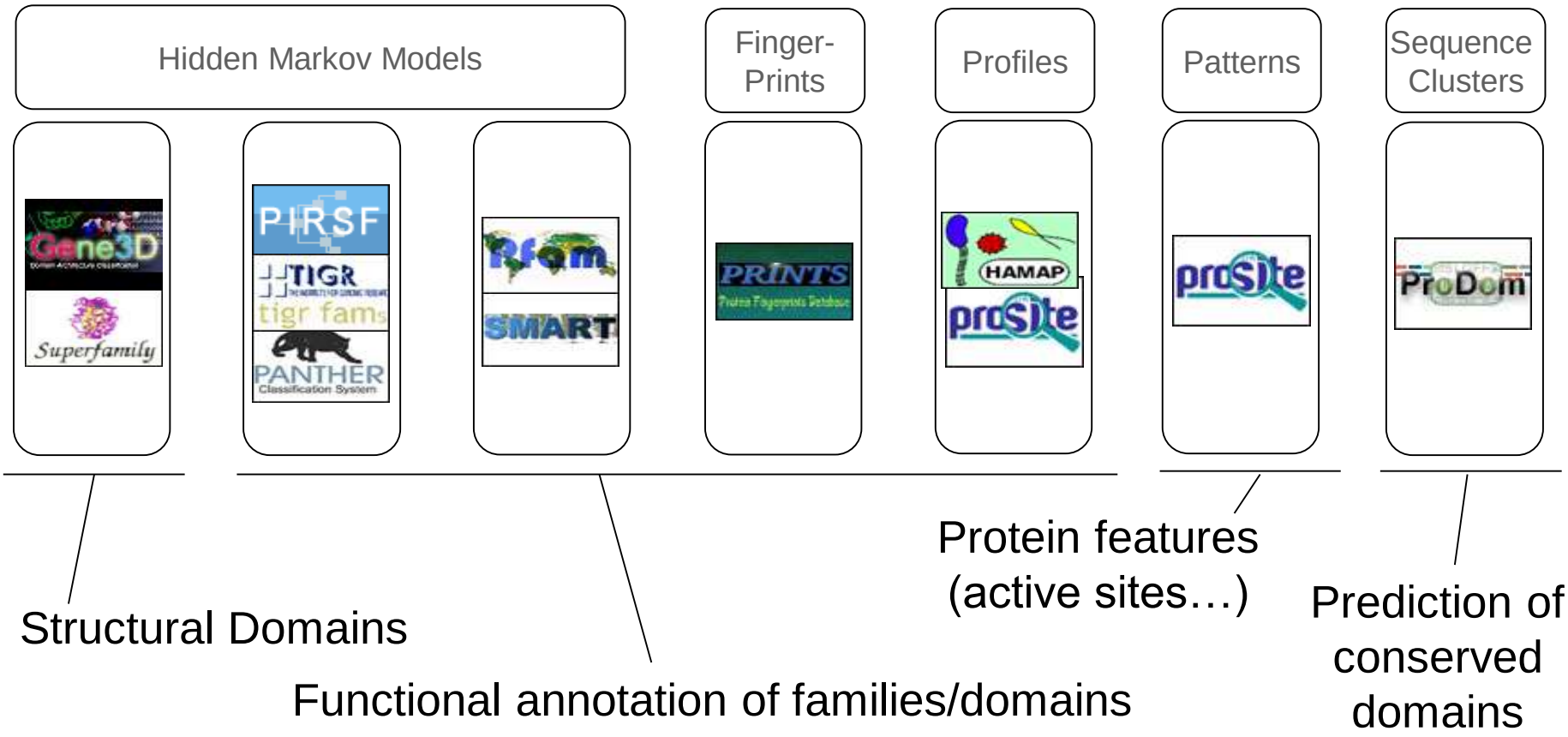


InterPro 85.0 • 8th April 2021

Features include:

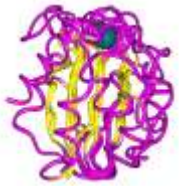
- The addition of 157 InterPro entries.
- An update to CATH-Gene3D [4.3.0].
- Integration of 333 new methods from the Pfam (3), SUPERFAMILY (11), CATH-Gene3D (168), PANTHER (88), CDD (62), SFLD (1) databases.

Member database information



Member database information

SIGNATURE DATABASE	VERSION	SIGNATURES	INTEGRATED SIGNATURES
PROSITE patterns	2019_11	1,311	1,285 (98.0%)
PROSITE profiles	2019_11	1,265	1,225 (96.8%)
PRINTS	42.0	2,106	1,948 (92.5%)
SMART	7.1	1,312	1,264 (96.3%)
TIGRFAMs	15.0	4,488	4,434 (98.8%)
Pfam	33.1	18,259	17,722 (97.1%)
SUPERFAMILY	1.75	2,019	1,620 (80.2%)
CATH-Gene3D	4.3.0	6,631	2,596 (39.1%)
PIRSF	3.10	3,285	3,236 (98.5%)
PANTHER	15.0	139,691	10,070 (7.2%)
HAMAP	2020_05	2,346	2,342 (99.8%)
CDD	3.18	16,212	3,698 (22.8%)
SFLD	4	303	157 (51.8%)



InterPro Curation Principles

- To represent MDBs signatures as closely as possible to what they intended.
- To reflect biological reality as accurately as possible in the entry they create by using types, relationships, GO mapping.
- To provide as much information to the end user as possible about the signature by annotating signatures and providing links to other databases.

InterPro entry types

InterPro entries are created for protein families, domains, sites, repeats and homologous superfamilies, defined as follows:

F **Family** - a group of proteins that share a common evolutionary origin reflected by their related functions, sequence homology or similarities in their structure.

D **Domain** - a distinct functional, structural or sequence unit often found associated with other types of domains.

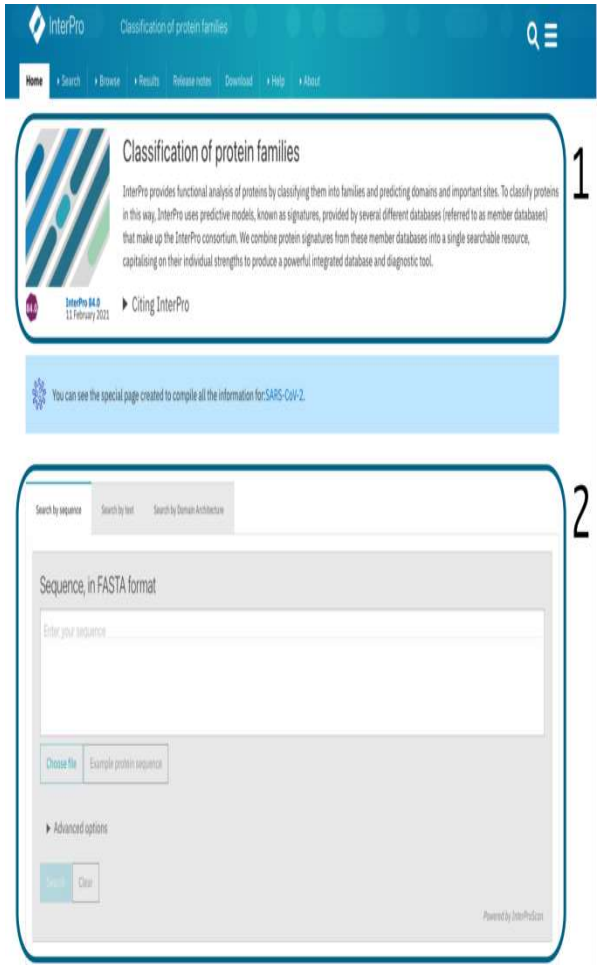
S **Site** - a short sequence containing one or more conserved residues, including: active sites, binding sites, conserved sites and sites of post-translational modification.

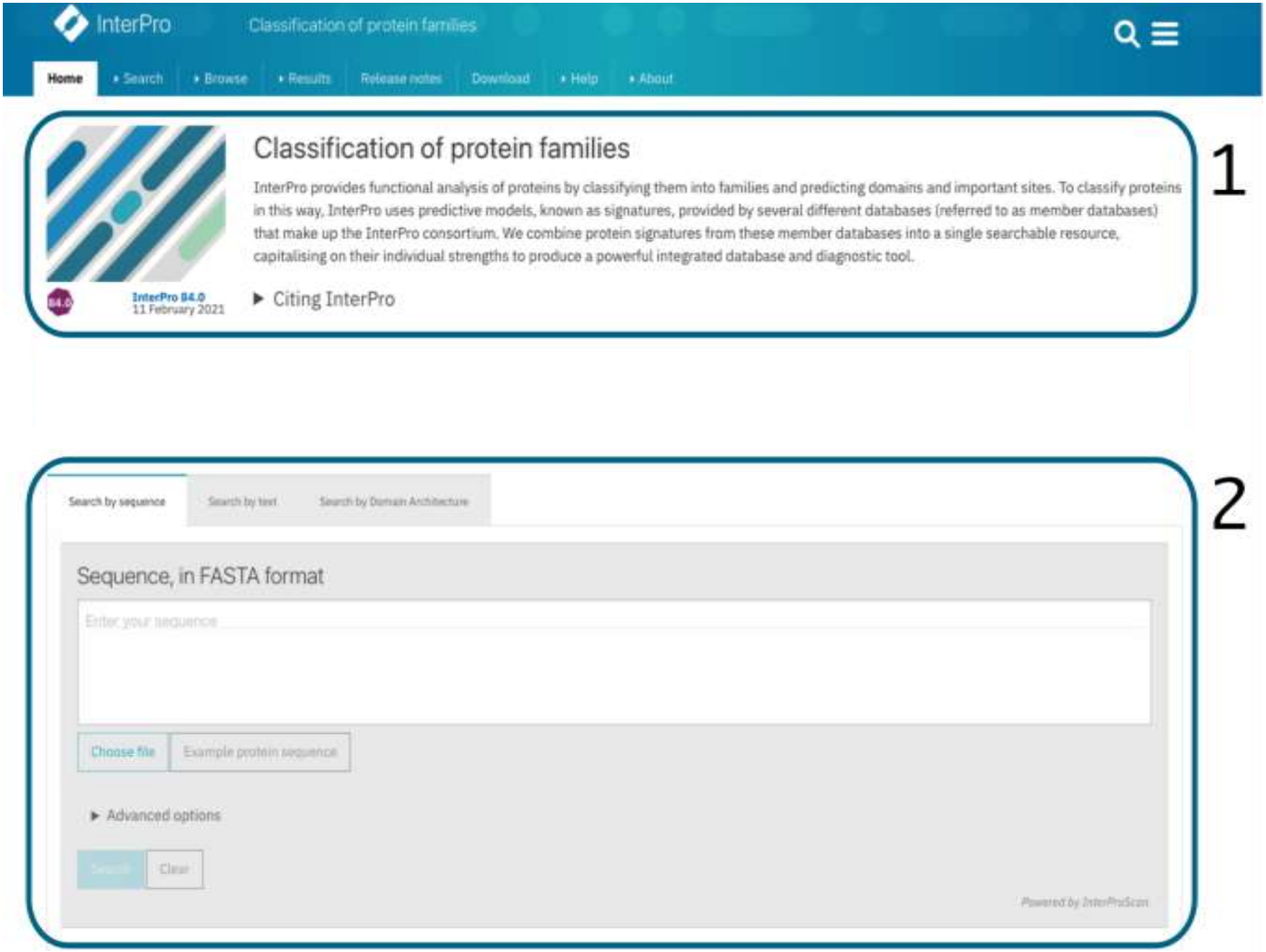
R **Repeat** - A short sequence (usually <50 amino acids) typically repeated many times within a protein.

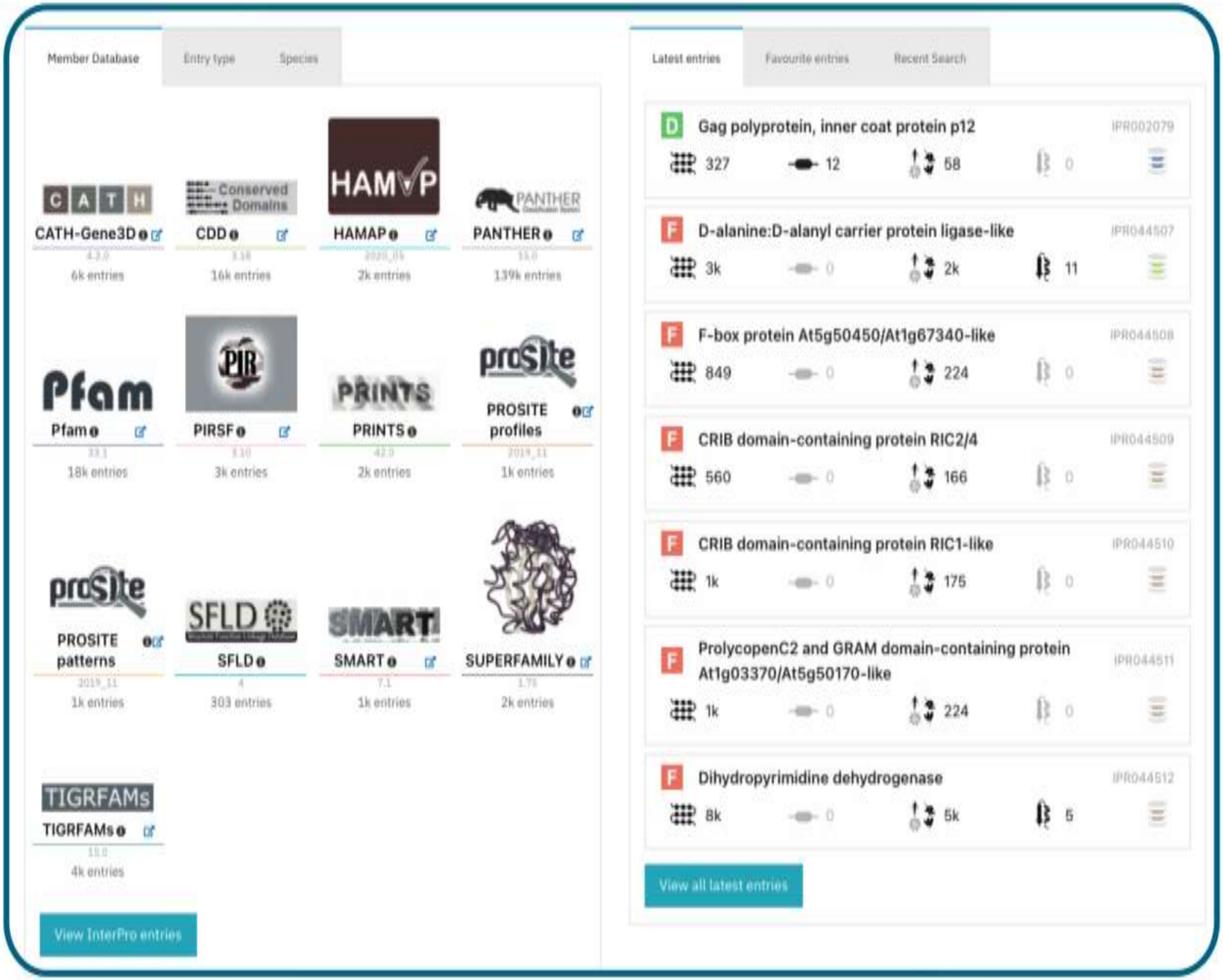
H **Homologous Superfamily** - a group of proteins that share a common evolutionary origin, reflected by similarity in their structure, even if sequence similarity is low. This entry type contains signatures from the CATH-Gene3D and SUPERFAMILY member databases exclusively.

U **Unintegrated** - member database signatures that might not yet be curated in InterPro, or might not reach InterPro's criteria for integration, but may still provide useful information.

InterPro HomePage




InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
About
Classification of protein families
InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.
InterPro 84.0
11 February 2021
Citing InterPro
1
Search by sequence
Search by text
Search by Domain Architecture
Sequence, in FASTA format
Enter your sequence
Choose file
Example protein sequence
Advanced options
Search
Clear
Powered by InterProScan
2


Member Database
Entry type
Species
CATH-Gene3D
6k entries
CDD
16k entries
HAMAP
2k entries
PANTHER
139k entries
Pfam
18k entries
PIRSF
3k entries
PRINTS
2k entries
PROSITE profiles
1k entries
PROSITE patterns
1k entries
SFLD
303 entries
SMART
1k entries
SUPERFAMILY
2k entries
TIGRFAMs
4k entries
View InterPro entries
Latest entries
Favourite entries
Recent Search
D Gag polyprotein, inner coat protein p12 IPR002079
327 12 58 0
F D-alanine:D-alanyl carrier protein ligase-like IPR044507
3k 0 2k 11
F F-box protein At5g50450/At1g67340-like IPR044508
849 0 224 0
F CRIB domain-containing protein RIC2/4 IPR044509
560 0 166 0
F CRIB domain-containing protein RIC1-like IPR044510
1k 0 175 0
F ProlycopenC2 and GRAM domain-containing protein At1g03370/At5g50170-like IPR044511
1k 0 224 0
F Dihydropyrimidine dehydrogenase IPR044512
8k 0 5k 5
View all latest entries

In the spotlight

BLOG

Interpro 84.0 New And Updated Features

Typhaine Paysan-Lafosse

The latest release of InterPro comes with a large number of improvements to the website. We hope that these changes

Read more →

PROTEIN FOCUS

The Beauty Of Domain Architecture And Evolution

Hsin-Yu Chang

Gene duplication serves as a major force in evolution. Duplicated genes can become a playground for evolution, and

Read more →

BLOG

Interpro 83.0 Key Features

Typhaine Paysan-Lafosse

InterPro 83.0 has been released with plenty of new InterPro entries, but also with many new features and improvements. This

Read more →

BLOG

Technical View Of The Interpro Client

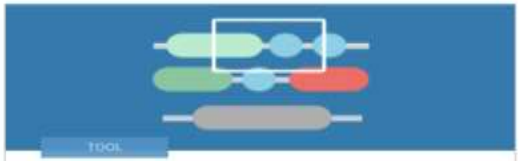
Gustavo A. Salazar

In the last post we discussed the technical details of our API. We talked about how the website influenced some of the

Read more →

Read all articles

Tools & libraries



TOOL

InterProScan

InterProScan is the software package that allows sequences (protein and nucleic) to be scanned against InterPro's signatures. Signatures are predictive models, provided by

Read more →

TOOL

A new API for InterPro

You can now skip URL and use this JSON interface to work with your data directly. Currently there are 6 main endpoints: entry, protein, structure, taxonomy, proteome and set.

Read more →

LIBRARY

Nightingale

Nigthtingale is a monorepo containing visualisation web components, including the formerly known Protvista, a powerful and blazing-fast tool for handling protein sequence

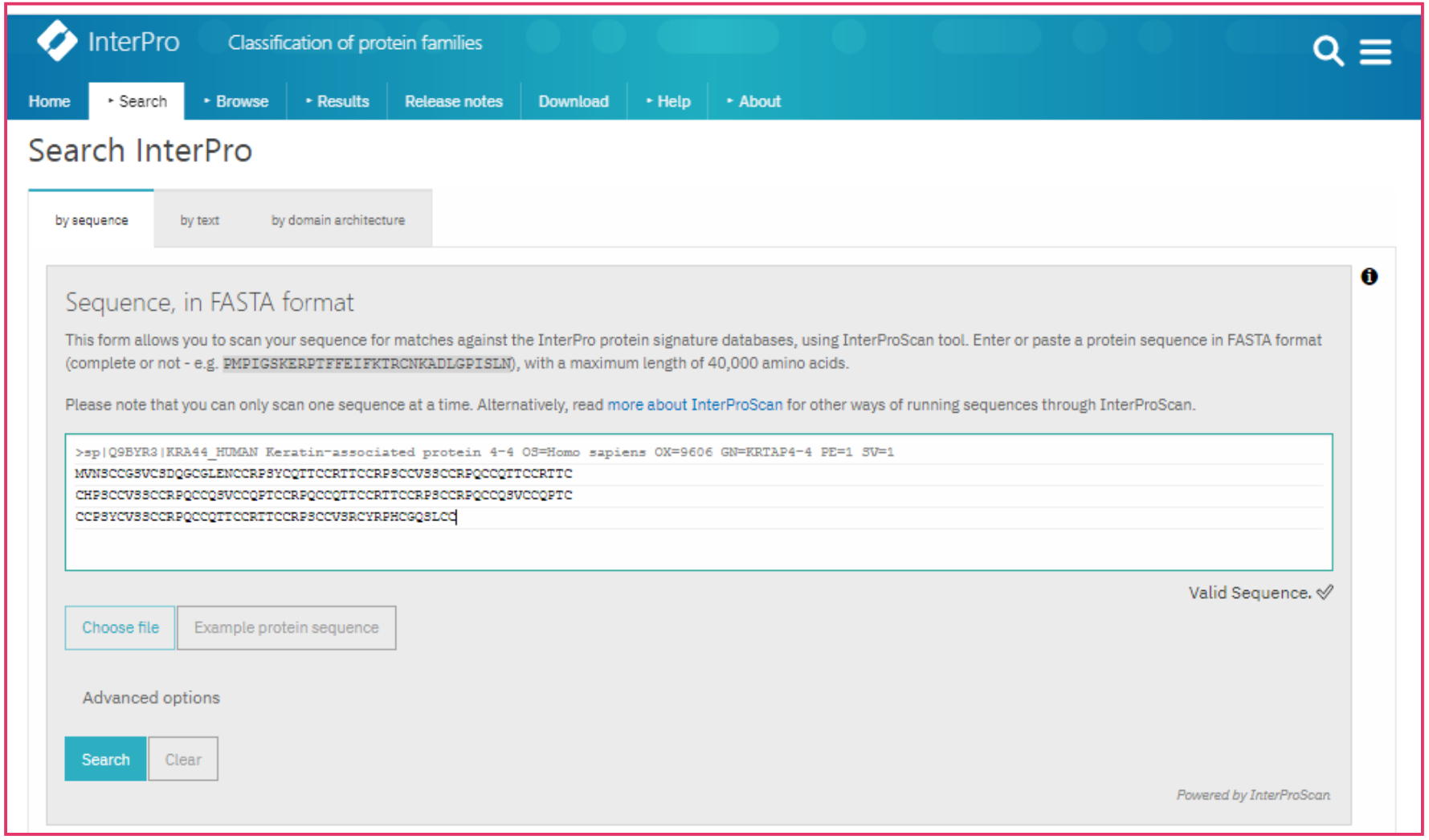
Read more →

InterPro
@InterProDB

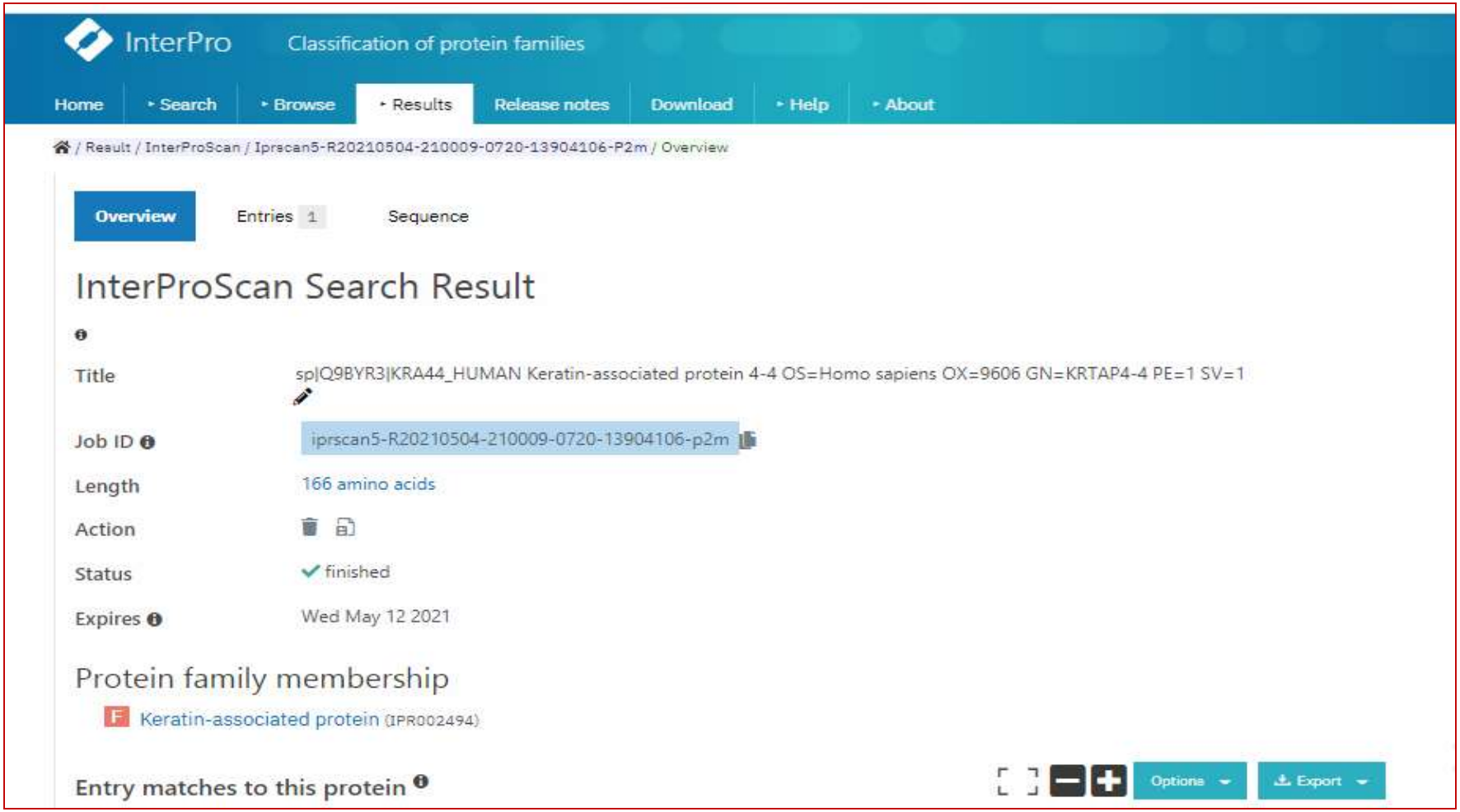
InterProScan 5 (version 5.50-84.0) is now available. For more details please visit: interproscan-docs.readthedocs.io

3h

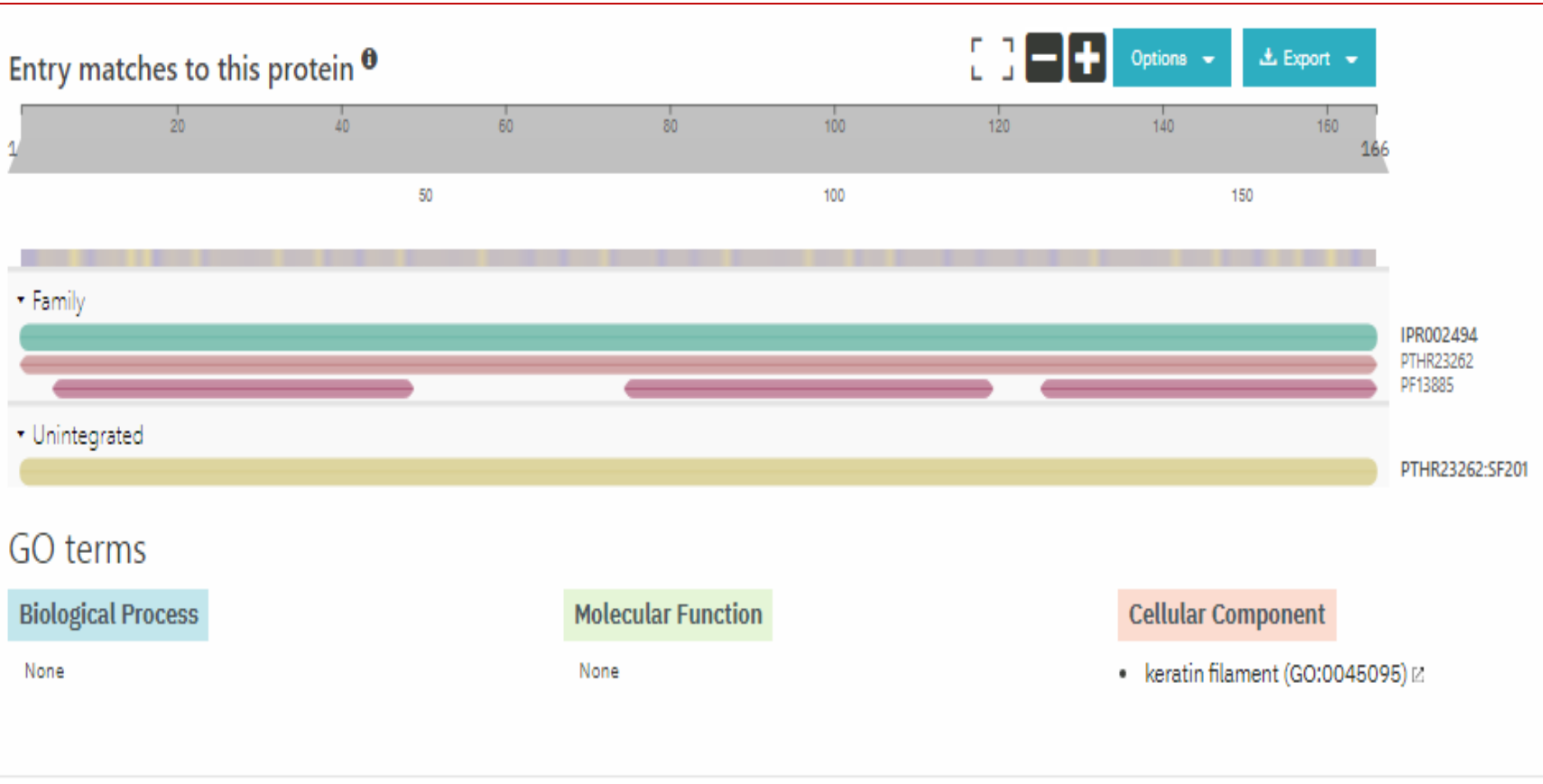
Example


InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
About
Search InterPro
by sequence
by text
by domain architecture
Sequence, in FASTA format
This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool. Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFFEIFKTRCNKADLGPISLN), with a maximum length of 40,000 amino acids.
Please note that you can only scan one sequence at a time. Alternatively, read more about InterProScan for other ways of running sequences through InterProScan.
>sp|Q9BYR3|KRA44_HUMAN Keratin-associated protein 4-4 OS=Homo sapiens OX=9606 GN=KRTAP4-4 PE=1 SV=1
MVNSCCGSVCSDQGCGLENCCRPSYCQTTCCRTTCCRPSCCVSSCCRPQCCQTTCCRTTC
CHPSCCVSSCCRPQCCQSVCCQPTCCRPQCCQTTCCRTTCCRPSCCRPQCCQSVCCQPTC
CCPSYCVSSCCRPQCCQTTCCRTTCCRPSCCVSRCYRPHCGQSLCC
Valid Sequence.
Choose file
Example protein sequence
Advanced options
Search
Clear
Powered by InterProScan

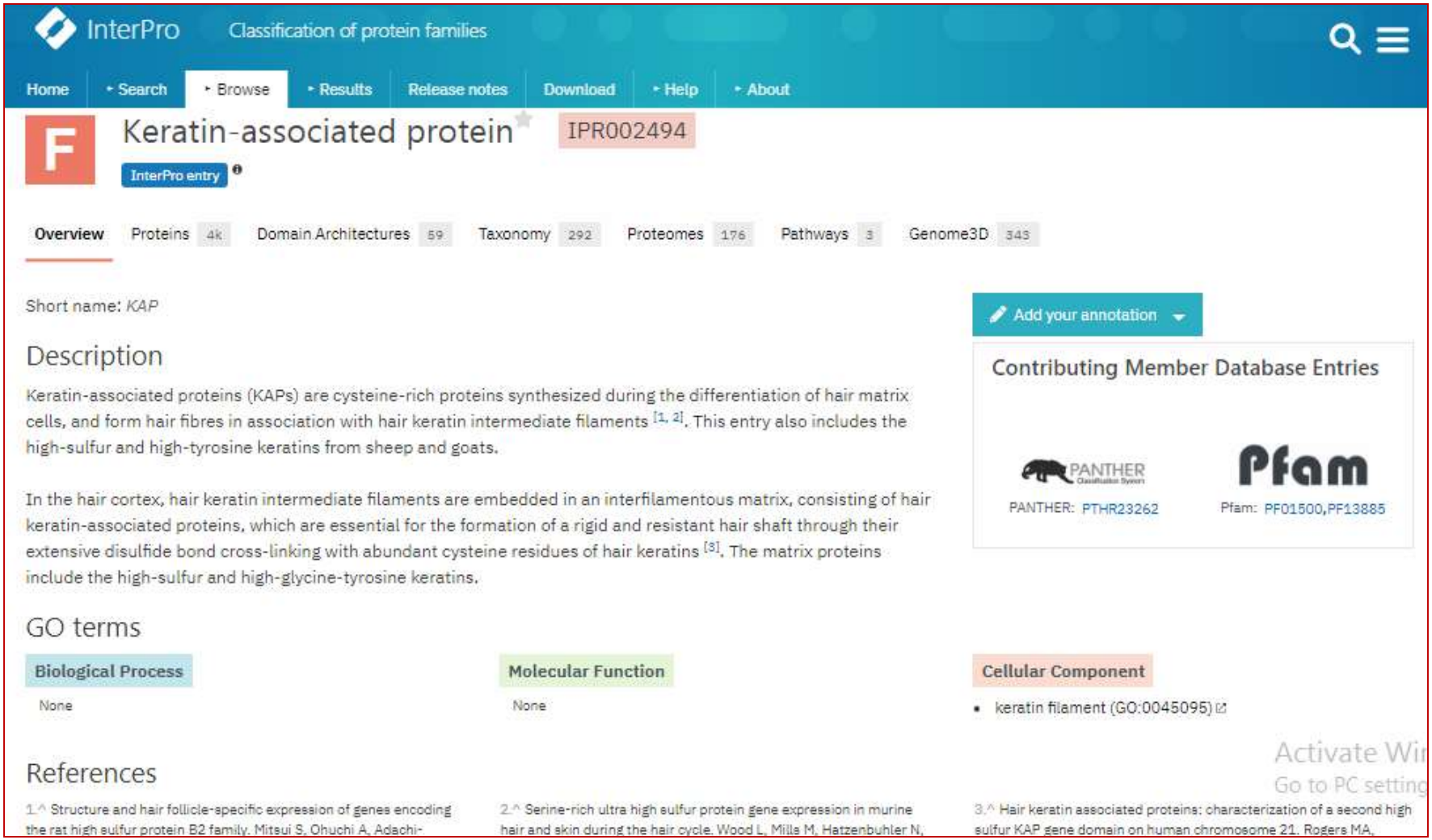
Example


InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
About
/ Result / InterProScan / Iprscan5-R20210504-210009-0720-13904106-P2m / Overview
Overview
Entries 1
Sequence
InterProScan Search Result
Title
sp|Q9BYR3|KRA44_HUMAN Keratin-associated protein 4-4 OS=Homo sapiens OX=9606 GN=KRTAP4-4 PE=1 SV=1
Job ID
iprscan5-R20210504-210009-0720-13904106-p2m
Length
166 amino acids
Action
Status
finished
Expires
Wed May 12 2021
Protein family membership
F
Keratin-associated protein (IPR002494)
Entry matches to this protein
Options
Export

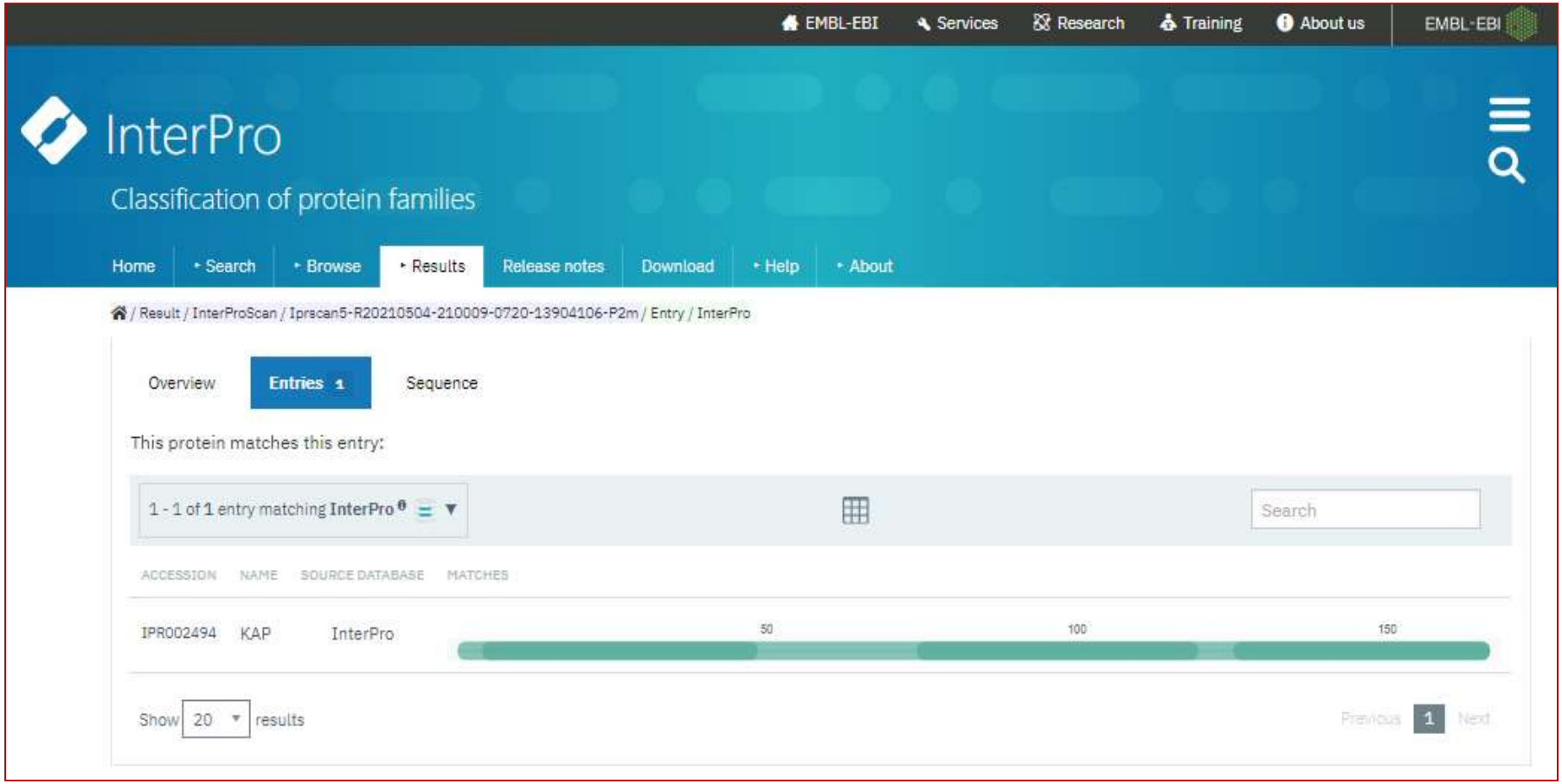
Example


Entry matches to this protein
Options
Export
20
40
60
80
100
120
140
160
1
166
50
100
150
Family
IPR002494
PTHR23262
PF13885
Unintegrated
PTHR23262:SF201
GO terms
Biological Process
None
Molecular Function
None
Cellular Component
keratin filament (GO:0045095)

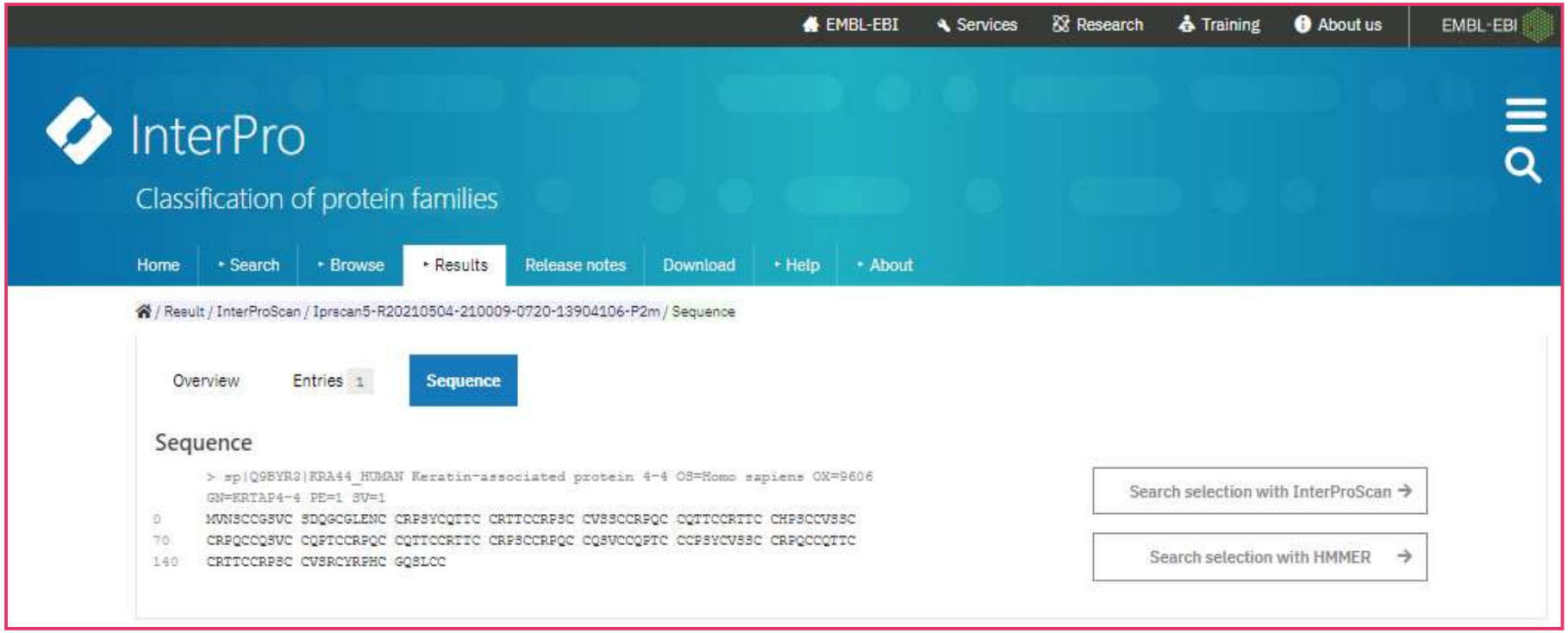
Example


InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
About
F
Keratin-associated protein
IPR002494
InterPro entry
Overview
Proteins 4k
Domain Architectures 59
Taxonomy 292
Proteomes 176
Pathways 3
Genome3D 343
Short name: KAP
Add your annotation
Description
Keratin-associated proteins (KAPs) are cysteine-rich proteins synthesized during the differentiation of hair matrix cells, and form hair fibres in association with hair keratin intermediate filaments [1, 2]. This entry also includes the high-sulfur and high-tyrosine keratins from sheep and goats.
In the hair cortex, hair keratin intermediate filaments are embedded in an interfilamentous matrix, consisting of hair keratin-associated proteins, which are essential for the formation of a rigid and resistant hair shaft through their extensive disulfide bond cross-linking with abundant cysteine residues of hair keratins [3]. The matrix proteins include the high-sulfur and high-glycine-tyrosine keratins.
Contributing Member Database Entries
PANTHER
PANTHER: PTHR23262
Pfam
Pfam: PF01500,PF13885
GO terms
Biological Process
None
Molecular Function
None
Cellular Component
keratin filament (GO:0045095)
References
1.^ Structure and hair follicle-specific expression of genes encoding the rat high sulfur protein B2 family. Mitsui S, Ohuchi A, Adachi-
2.^ Serine-rich ultra high sulfur protein gene expression in murine hair and skin during the hair cycle. Wood L, Mills M, Hatzenbuhler N,
3.^ Hair keratin associated proteins: characterization of a second high sulfur KAP gene domain on human chromosome 21. Rogers MA,
Activate Wi
Go to PC setting

Example


EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
About
/ Result / InterProScan / Iprscan5-R20210504-210009-0720-13904106-P2m / Entry / InterPro
Overview
Entries 1
Sequence
This protein matches this entry:
1 - 1 of 1 entry matching InterPro
Search
ACCESSION
NAME
SOURCE DATABASE
MATCHES
IPR002494
KAP
InterPro
50
100
150
Show 20 results
Previous
1
Next

Example


EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
About
/ Result / InterProScan / Iprscan5-R20210504-210009-0720-13904106-P2m / Sequence
Overview
Entries 1
Sequence
Sequence
> sp|Q9BYR3|KRA44_HUMAN Keratin-associated protein 4-4 OS=Homo sapiens OX=9606 GN=KRTAP4-4 PE=1 SV=1
0 MVNSCCGSVC SDQGCGLENC CRPSYCQTTC CRTTCCRPSC CVSSCCRPQC CQTTCCRTTC CHPSCCVSSC
70 CRPQCCQSVC CQPTCCRPQC CQTTCCRTTC CRPSCCRPQC CQSVCCQPTC CCPSYCVSSC CRPQCCQTTC
140 CRTTCCRPSC CVSRCYRPHC GQSLCC
Search selection with InterProScan
Search selection with HMMER

THANK YOU

PROTEIN - PROTEIN INTERACTIONS

-Ms. Rupal Mishra

Protein-Protein Interactions (PPI)

- Proteins control and mediate many of the biological activities of cells by these interactions.
- Protein–protein interactions occur when two or more proteins bind together. They improve our understanding of diseases and can provide the basis for new therapeutic approaches.

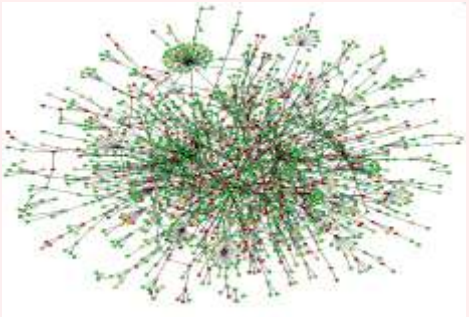
Gene is the basic unit of heredity.

Genome

Proteins, the working molecules of a cell, carry out many biological activities

Proteome

Proteins function by interacting with other proteins.



Interactome

Protein-Protein Interactions (PPI)

- **Protein–protein interactions** (**PPIs**) are physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect.
- Proteins rarely act alone as their functions tend to be regulated.
- Many molecular processes within a cell are carried out by molecular machines that are built from numerous protein components organized by their PPIs.
- These physiological interactions make up the interactomics of the organism.

Protein-Protein Interactions (PPI)

PPIs are involved in many biological processes:

- Signal transduction
- Electron transfer proteins
- Membrane transport
- Cell metabolism
- Muscle contraction
- Protein complexes or molecular machinery
- Protein carrier
- Protein modifications

Types of protein-protein interactions

There are 3 categories-

1) **On the basis of their Composition**
2) **On the basis of their duration of interaction**
3) **On the basis of their bonding**

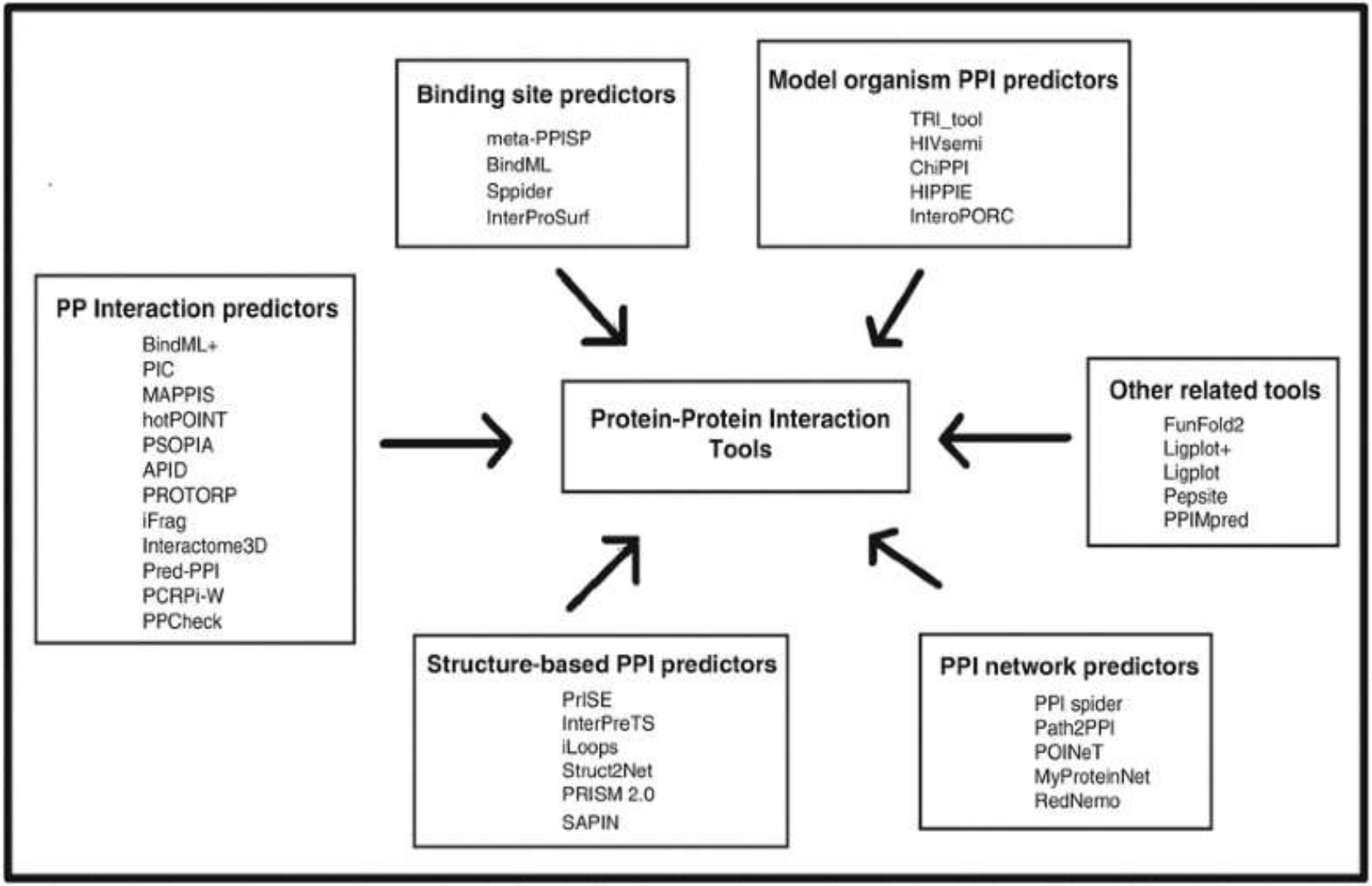
- **On the basis of their Composition**
  - **Homo-Oligomers**: These are macromolecule complexes having one type of protein subunit. e.g. : PPIs in Muscle Contraction
  - **Hetero-Oligomers**: These are macromolecule complexes having multiple types protein subunits. e.g.: G protein-coupled receptors.

Types of protein-protein interactions

- **On the basis of their duration of interaction**
  - **Stable Interactions**: These comprise of interactions that last for a long duration. These Interactions carry out Functional or Structural roles. e.g.: Hemoglobin structure
  - **Transient Interactions**: Interactions that last a short period of time. e.g.: Muscle Contraction

- **On the basis of their Bonding**
  1) **Covalent :** Strongest association - disulphide bonds or electron sharing (Post translational modifications).
  2) **Non-covalent :** Established during transient interactions by the combination of weaker bonds (Hydrogen bonds, Ionic interactions, Van der waals forces, or Hydrophobic bonds).

Tools for PPI Analysis



PPIs Identification Methods

Experimental *(In vivo)*	• **Yeast two-hybrid system** • split ubiquitin system • split lactamase / split galactosidase system • split yellow fluorescent protein (YFP) system
Experimental *(In vitro)*	• **Co-immunoprecipitation** • Tagged Fusion Proteins • X-ray Diffraction • Biacore • Phage display
Computational *(In silico)*	• BIND • DIP • MINT • IntAct

DIP Database

- The DIP (Database of Interacting Proteins) database lists protein pairs that are known to interact with each other.
- Interact here means that two amino acid chains were experimentally identified to bind to each other.
- The database lists such pairs to aid those studying a particular protein-protein interaction but also those investigating entire regulatory and signaling pathways as well as those studying the organization and complexity of the protein interaction network at the cellular level.
- The DIP database is composed of nodes and edges.

DIP Database

1) **DIP Nodes** (*proteins*) –

- Each protein participating in a DIP interaction is idetified by a unique identifier of the form <***DIP:****nnn****N***> and cross-references to, at least, one of the major protein databases.
  (PIR, SWISSPROT and / or GENBANK)
- In addition, some basic information about each protein, such as name, function, subcellular localization and cross-references to other biological database is stored locally (if available) in case the cross-referenced databases are not accessible.
- Node search allows one to find DIP entries using cross-references to other databases or by searching protein annotation local to the DIP database.

DIP Node Annotation

- **DIP node ID**
  This is a unique identifier for each protein described within the DIP database. It has a format <**DIP:**nnn**N**>.
- **Description/Name**
  The common name of the proten and/or its short description.
- **Primary Database reference(s)**
  At least one reference to PIR, SWISSPROT and/or GENBANK entry is provided.
- **Cross-references**
  References to other database entries related to the protein of interest as retrieved from SwissProt and NR BLAST entries. The list of cross-references is subdivided into (P)rotein, (D)omain and (F)eature (eg motifs) categories.
- **Superfamily**
  The superfamily specified in the PIR entry.
- **Organism**
  The organism producing the protein. A cross-reference to the TaxonID database is provided for additional information.
- **Function**
  A short description of the protein's function within the cell.
- **Localization**
  Localization of the protein within the cell (if known)
- **Keywords**
  Keywords associated with the protein. They might describe its structure and/or sequence features, biological activity, function, cellular localization, etc.

DIP Database

2) **DIP Edges** (*interactions*)

- The information about each DIP interaction is identified by a unique identifier of the form <***DIP:****nnn****E***> that provides access to information such as the region involved in the interaction, the dissociation constant and the experimental methods used to identify and characterize the interaction.

DIP Edge Annotation

- **DIP interaction ID**
  This is a unique identifier for each interaction described within the DIP database. It has a format <***DIP:****nnn****E***>. For example, DIP:1234E referes to the interaction between actin and foofoo in *S. cerevisiae*. Please, use the DIP interaction ID when referring to an interaction described within the DIP database.
- **Residue Ranges**
  The range is meant to specify more accurately which parts of the proteins are directly involved in the interaction.
- **Protein Domain**
  The interacting domain name. An incomplete list of domain names is available.
- **Dissociation Constant**
  The dissociation constant of the protein-protein complex in mols/liter (M).
- **Experimental Method(s)**
  The experimental method(s) used to identify and/or characterize the interaction.
- **Reference(s)**
  A literature reference to the experimental work studying the interactions.
- **Comments**
  Additional comments supplied by the person who deposited the interaction. Comments could include an explanation (evidence) why the proteins are thought to interact, and what the biological significance of the interaction is.

Searching DIP

- The DIP database can be searched in a number of ways in order to identify the starting point for exploration of the protein interaction network.
- It is possible to search for entire groups of proteins fulfilling certain criteria, such as sequence similarity to a given protein, specific function or cellular localization, the presence of a specified domain or well-known sequence motif.
- Search Types-
  1) Node,
  2) BLAST,
  3) Motif,
  4) Article,
  5) IMEx
  6) pathBLAST.

Search Types

- **Node** - Search the database for matches within the various fields describing protein (node) entries. The results are returned as a list of proteins that fulfill the criteria specified.
- **BLAST** - Search the database for sequences matching a particular sequence or its fragment. The results are returned as a list of proteins sorted by the BLAST significance score (*p-value*).
- **Motif** - Search the database for proteins containing a domain or motif defined by one of the domain/motif databases: Prosite, InterPro, Pfam, PRINTS or SMART, or as a user-specified regular expression.

Search Types

- **Article** - Search the database for interactions described in selected article(s). The results are returned as a list of interactions that were described by at least one experiment from the selected article(s).
- **IMEx** - Search the database for interactions supported by experiments annotated according to the curation rules adopted by IMEx (a consortium that was established by DIP) together with other consortium partners, in order to standardize annotation of experiments demonstrating protein interactions. The results of the search are returned as a list of interactions that are supported by at least one experiment annotated according to the IMEx Consortium standard.
- **pathBLAST** - Search the database for protein-protein interaction network of the target organism to extract all protein interaction pathways that align with the query pathway.

DIP Database

Database of Interacting Proteins

Jobs
Help
News
Register
Statistics
Satellites
SEARCH
SUBMIT
Software
Services
Articles
Links
Files
MIF

Search by: [protein] [sequence] [motif] [article] [IMEx] [pathBLAST] [Help][LOGIN]

THE DIP DATABASE

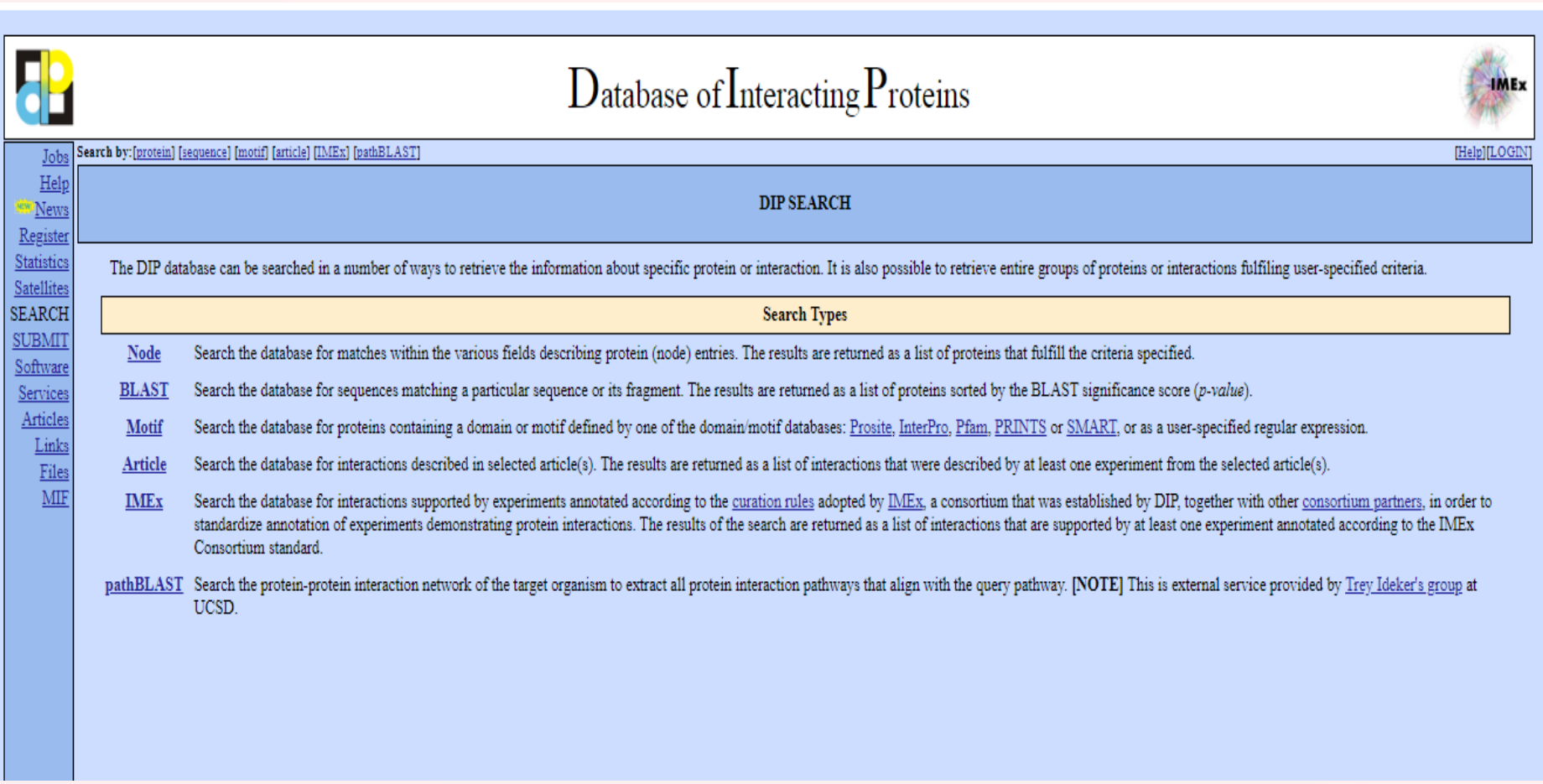
The DIP™ database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data. Please, check the reference page to find articles describing the DIP database in greater detail.

This page serves also as an access point to other projects related to DIP, such as The Database of Ligand-Receptor Partners (DLRP) and JDIP.

DIP PAGES

NEWS	Announcements about the most recent additions and changes to the database.
REGISTRATION/ ACCOUNT	Registration and account maintanance. Registration is required to gain access to most of the DIP features. Registration is free to the members of the academic community. Trial accounts for the commercial users are also available. Please, consult Terms of Use for further details.
STATISTICS	Detailed information about the current state of the database as well as some statistics on server usage.
SATELLITES	DIP-related projects, such as DLRP and JDIP.
SERVICES	DIP-derived services.
ARTICLES	DIP in press. Both, papers published on DIP as well as a list of publications referring to DIP.
SEARCH	Database search. This is the starting point of the database exploration. Once the initial protein is found through keyword or sequence searches the interaction network can be explored by interactively following the interaction links.
LINKS	Links to other protein interaction databases and related sites.
FILES	Download the complete DIP dataset as well as specialized DIP subsets and additional data (*registration required*).
HELP	A short description of the DIP database.

DIP Database



Database of Interacting Proteins

IMEx

Jobs
Help
News
Register
Statistics
Satellites
SEARCH
SUBMIT
Software
Services
Articles
Links
Files
MIF

Search by:[protein] [sequence] [motif] [article] [IMEx] [pathBLAST] [Help][LOGIN]

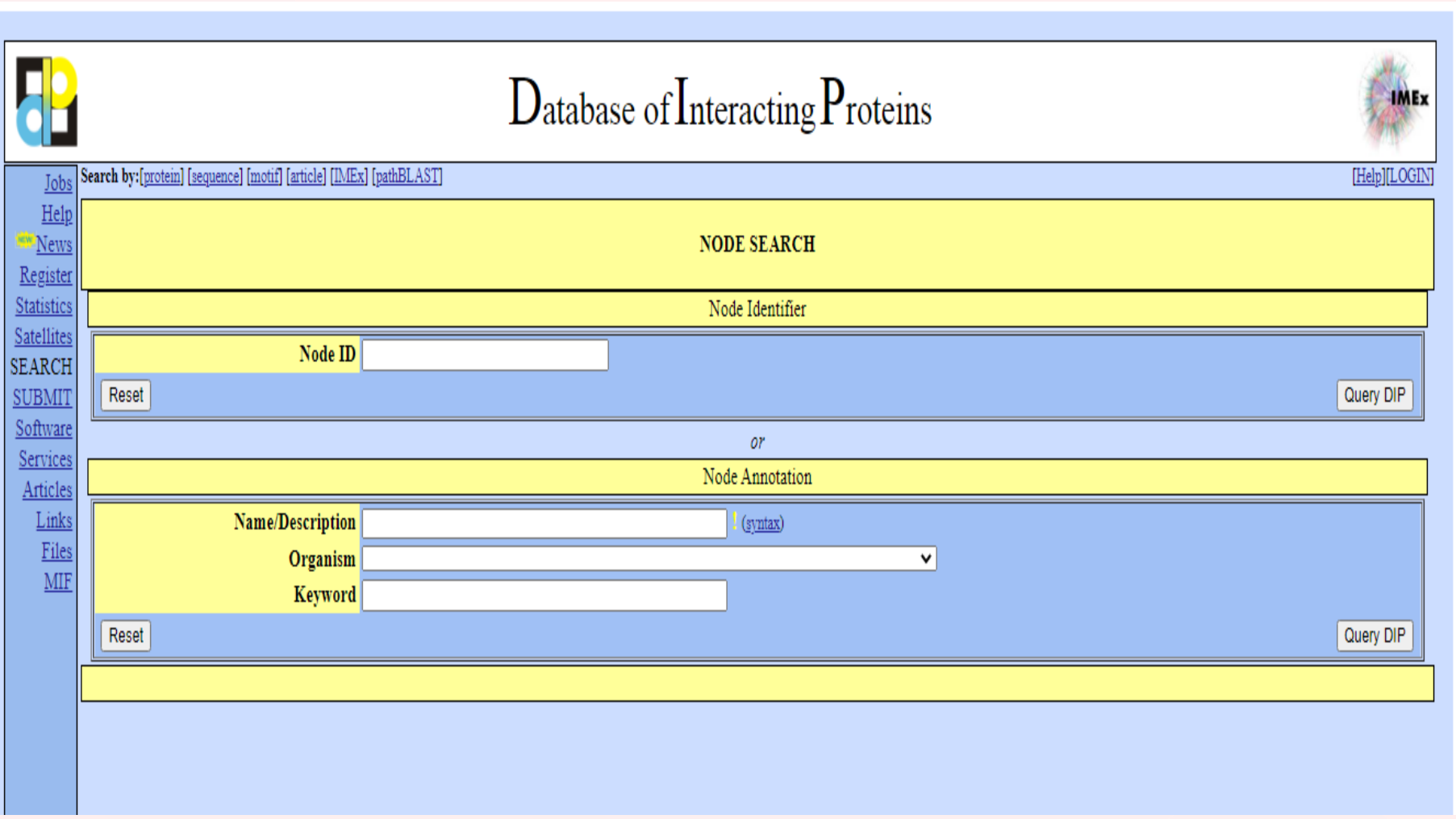
DIP SEARCH

The DIP database can be searched in a number of ways to retrieve the information about specific protein or interaction. It is also possible to retrieve entire groups of proteins or interactions fulfiling user-specified criteria.

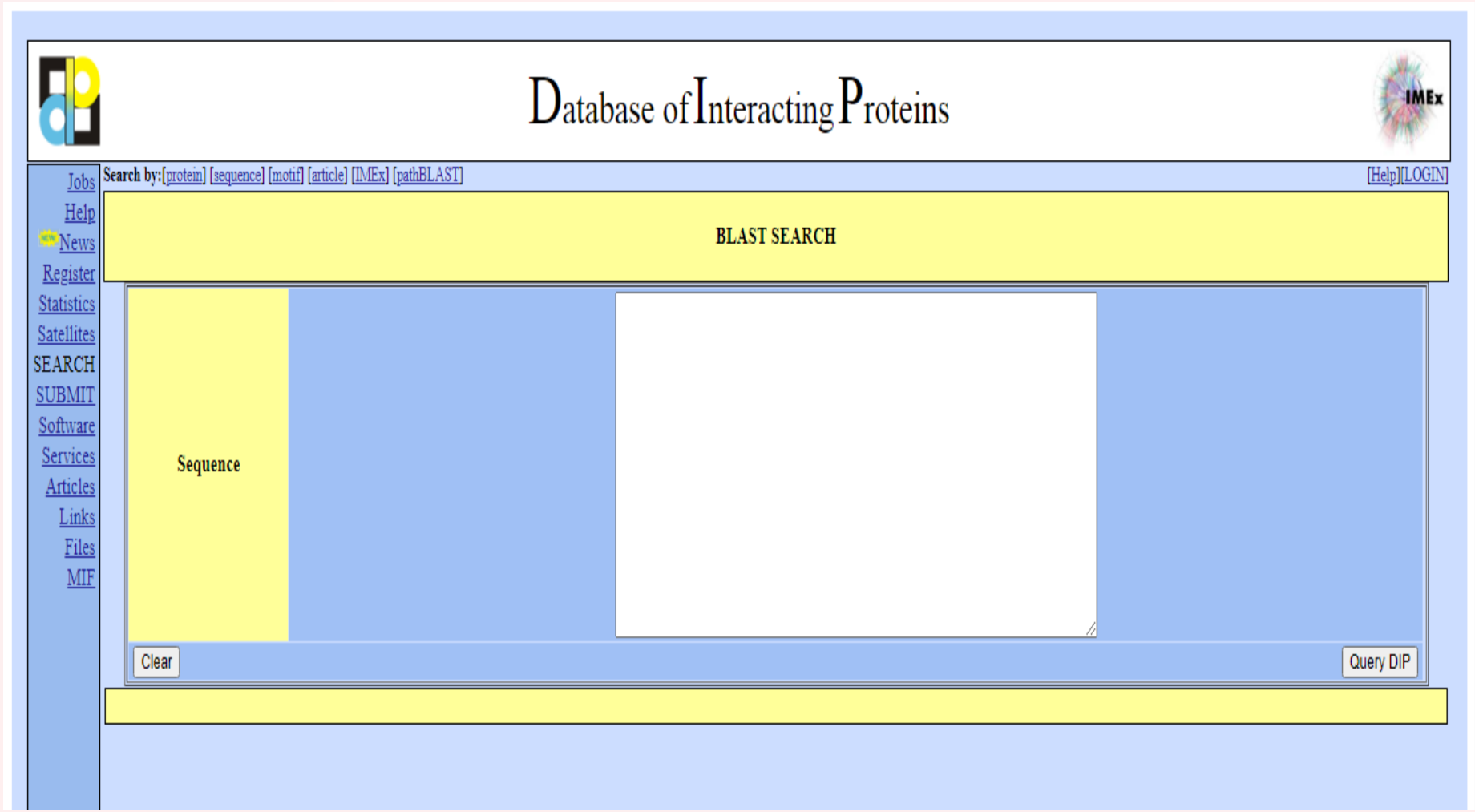
Search Types

Node	Search the database for matches within the various fields describing protein (node) entries. The results are returned as a list of proteins that fulfill the criteria specified.
BLAST	Search the database for sequences matching a particular sequence or its fragment. The results are returned as a list of proteins sorted by the BLAST significance score (*p-value*).
Motif	Search the database for proteins containing a domain or motif defined by one of the domain/motif databases: Prosite, InterPro, Pfam, PRINTS or SMART, or as a user-specified regular expression.
Article	Search the database for interactions described in selected article(s). The results are returned as a list of interactions that were described by at least one experiment from the selected article(s).
IMEx	Search the database for interactions supported by experiments annotated according to the curation rules adopted by IMEx, a consortium that was established by DIP, together with other consortium partners, in order to standardize annotation of experiments demonstrating protein interactions. The results of the search are returned as a list of interactions that are supported by at least one experiment annotated according to the IMEx Consortium standard.
pathBLAST	Search the protein-protein interaction network of the target organism to extract all protein interaction pathways that align with the query pathway. **[NOTE]** This is external service provided by Trey Ideker's group at UCSD.

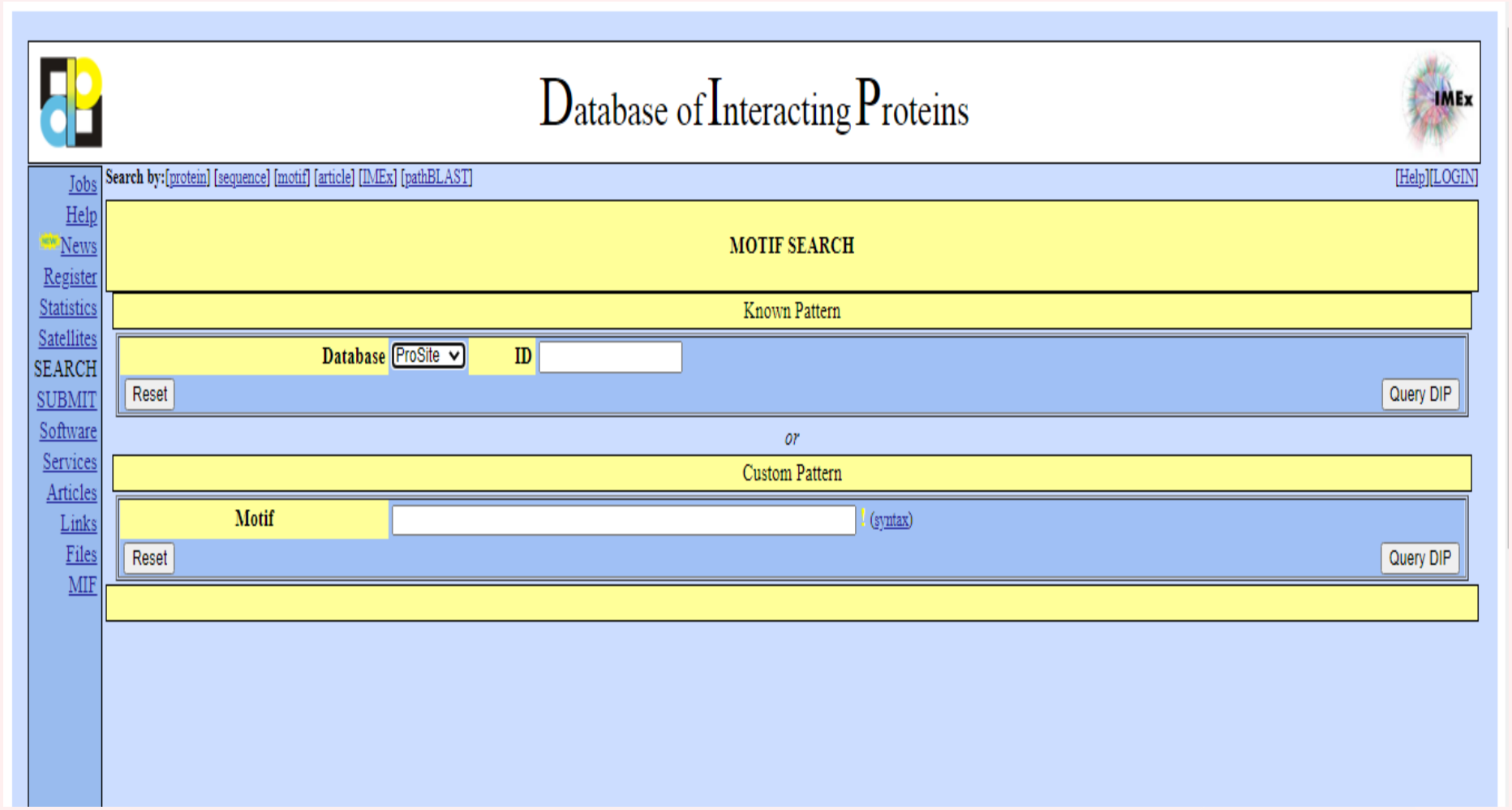
NODE Search


Database of Interacting Proteins
IMEx
Jobs
Help
News
Register
Statistics
Satellites
SEARCH
SUBMIT
Software
Services
Articles
Links
Files
MIF
Search by:[protein] [sequence] [motif] [article] [IMEx] [pathBLAST]
[Help][LOGIN]
NODE SEARCH
Node Identifier
Node ID
Reset
Query DIP
or
Node Annotation
Name/Description
(syntax)
Organism
Keyword
Reset
Query DIP

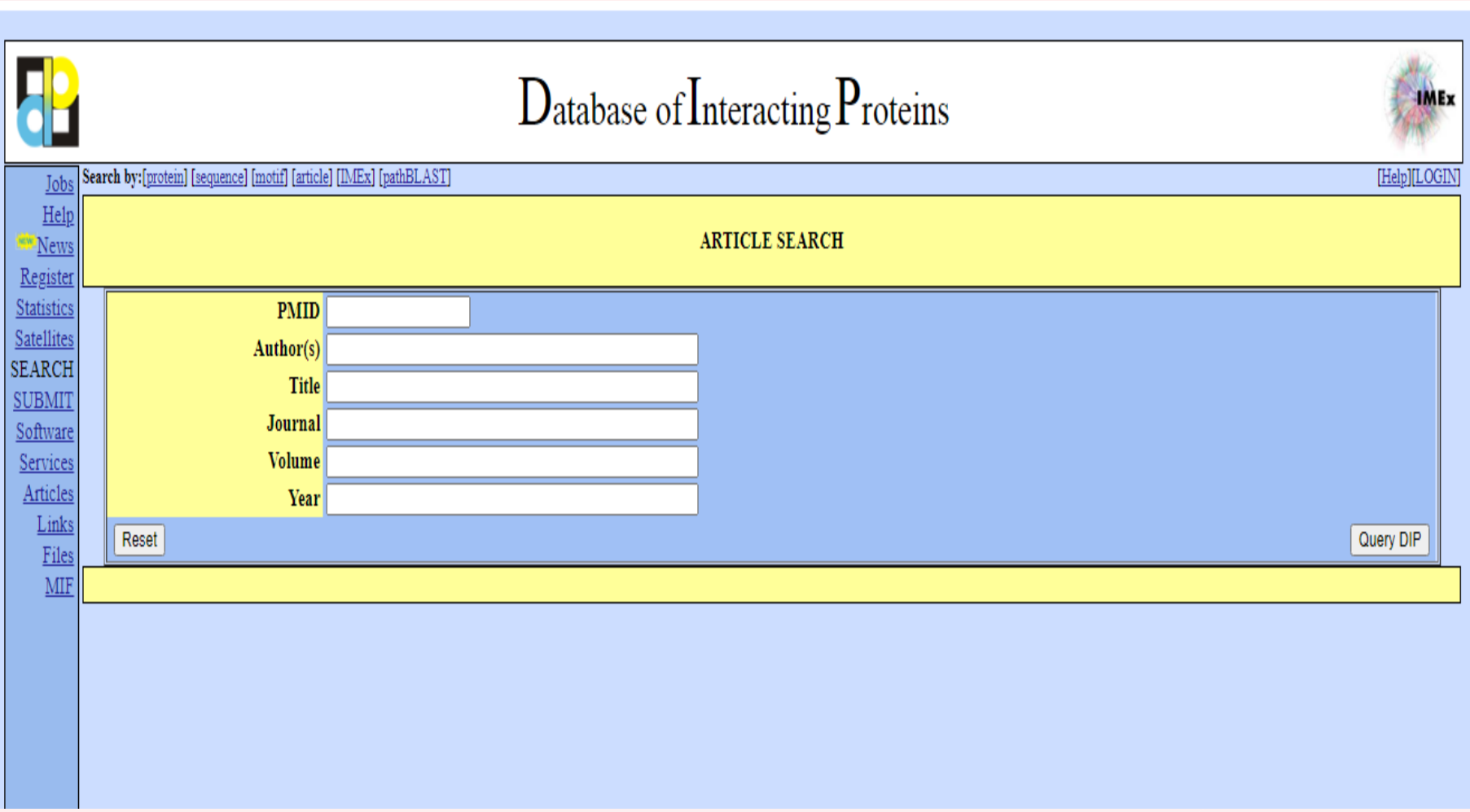
Blast Search



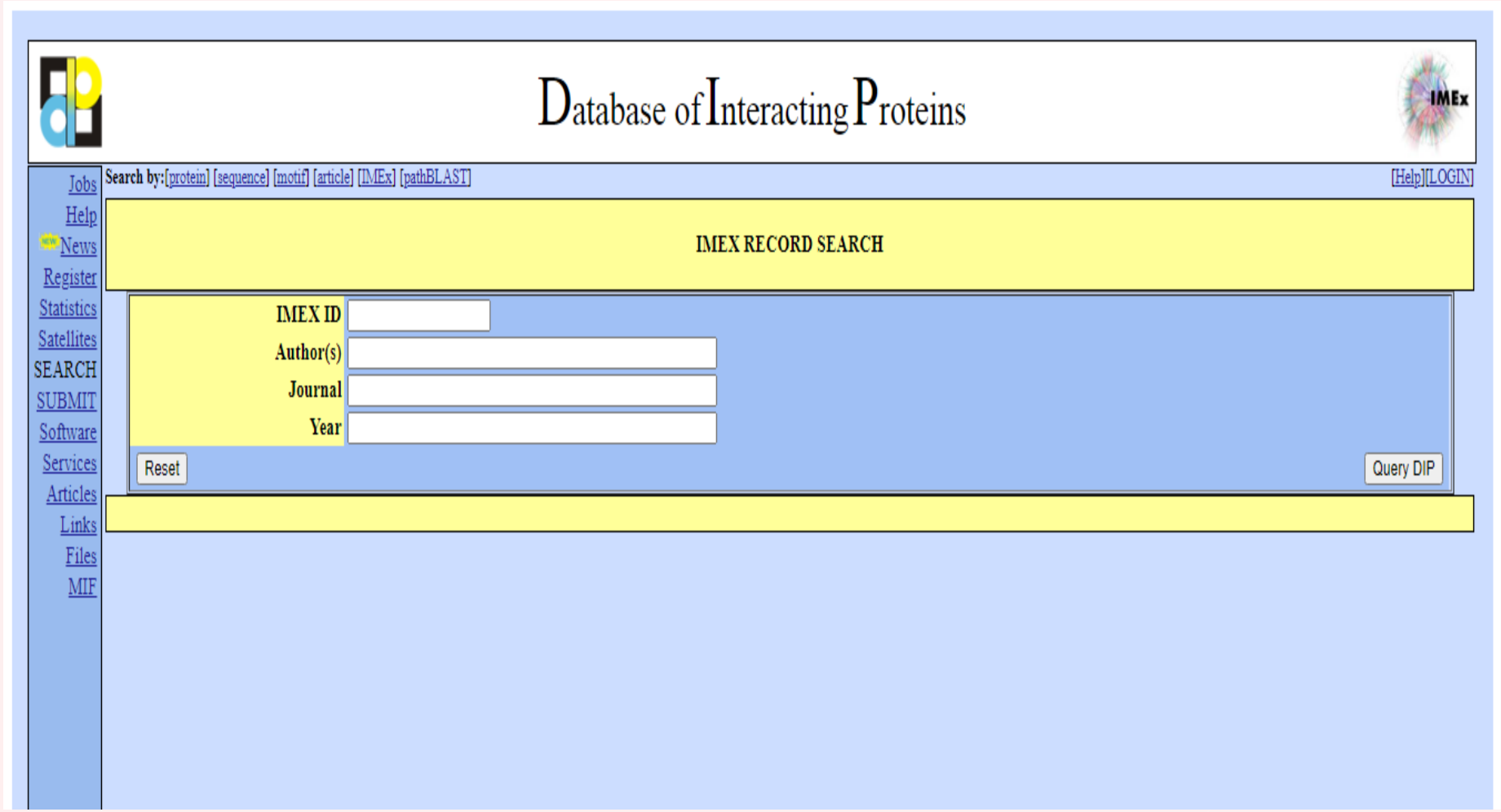
Motif Search



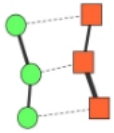
Article Search



IMEX RECORD Search



PathBLAST Search



PathBLAST

Collaborating Labs: UC San Diego UC Berkeley Tel Aviv University Whitehead Institute

About PathBLAST

PathBLAST searches the protein-protein interaction network of the target organism to extract all protein interaction pathways that align with a pathway query.

To learn more about PathBLAST read: FAQ Selected Publications

PathBLAST Search

Example Input Pathways:

1 2 3 4 5

Please enter the proteins in your pathway query:

	Protein ID		Protein Sequence
A		and/or	
B		and/or	
C		and/or	

Add a Protein Remove a Protein

Please select the Target Organism Network: Saccharomyces cerevisiae

Show Advanced Options

BLAST! RESET

THANK YOU