Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 5a
Protein sequence databases

In this lecture, we will mainly discuss on Protein Sequence Databases. So, let us refresh ourselves about the last class. So, you remember what did we discuss in the last class, this mainly about the protein structures. So, what is the building block for protein structures?

Student: Amino acid.

Amino acids, right. So, amino is a classified in 2 major groups what are 2 major groups.

Student: (Refer Time: 00:41).

Hydrophobic residues and hydrophilic residues and the hydrophobic residues we subclassified into different groups; what are the sub-classifications?

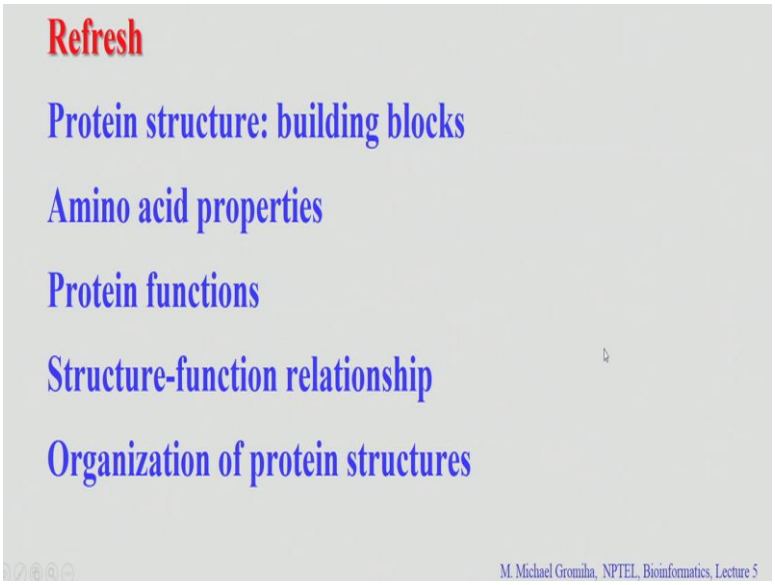
Student: (Refer Time: 00:48).

So, you can (Refer Time: 00:50) take.

Student: (Refer Time: 00:50).

Amino acids, aromatic amino acids, Sulphur-containing amino acids, and glycine; for the case of hydrophilic residues.

(Refer Slide Time: 00:59)

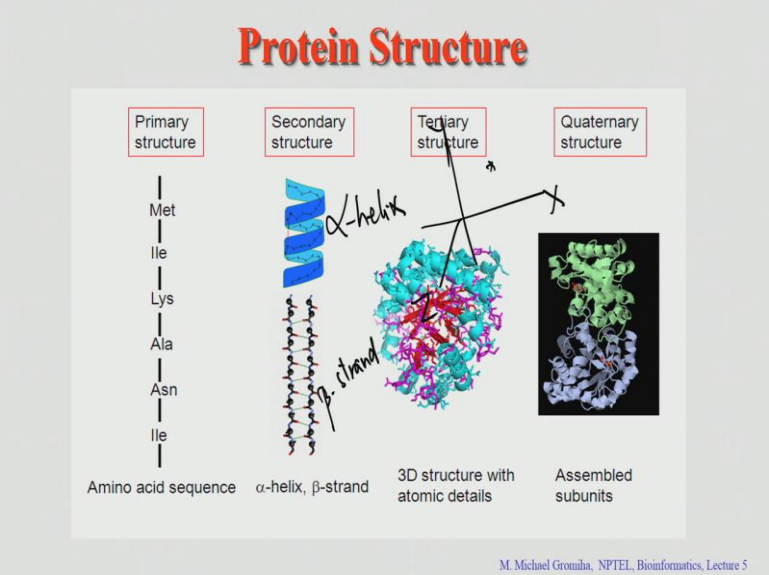


We classified into positive charge, negative charge and polar, right. So, you have classified based on the amino acids properties, then we discussed about protein functions do you remember what are the different functions we discussed in the last class; enzymes, antigens, antibodies, structural proteins.

Student: Structural proteins, regulatory proteins.

Regulatory proteins and the sweetener, blood clotting proteins, thrombin, and transporters hemoglobin. So, we discussed about various proteins with they perform different functions; then we discussed about the structure-function relationship to understand the functions of proteins, right the structures will help and mainly the structure of a protein dictates its function. So, it is very important to understand the structural level. So, if you look into the various organization of protein structures.

(Refer Slide Time: 01:44)



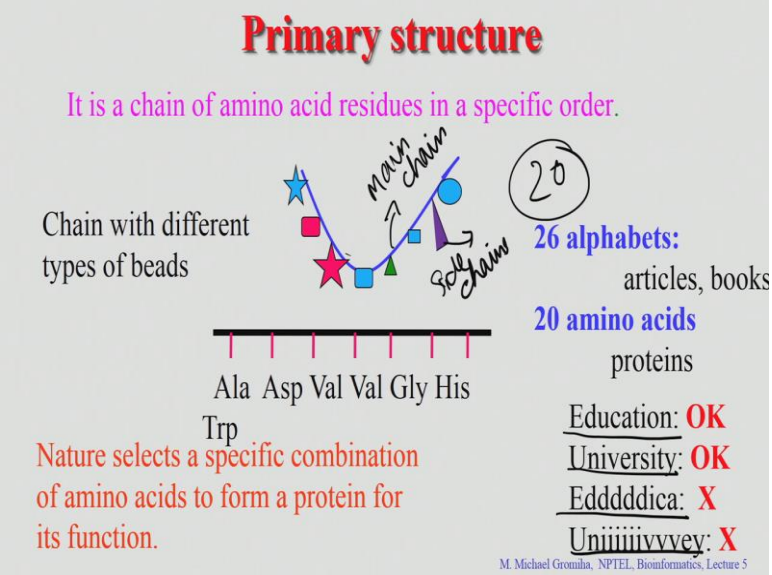
They are mainly classified into 4 different groups like primary structure, secondary structure, tertiary structure and quaternary structure.

In addition, there are 2 structures in between them; they are super-secondary structures, the combination of different secondary structures plus domains that is in a protein some part of this residues; they can fold and they can perform functions they are called domains. So, what is a primary structure? It is the arrangement of different amino acid residues, right. So, primary structure gives the sequential order of this amino acid residues and secondary structure provides the arrangement of these residues; in some specific shape; for example, here this is a kind of spiral shape. So, this is called alpha helix. So, here you can see; this is a kind of ladder. So, this is called beta strand, right. So, here the next one is tertiary structure.

The tertiary structure will provide the atomic coordinates of all the atoms in each residue. So, you can get the; get the location of each atom, right if you take the x y z coordinates we will see where is the location of each atom in a residue, right, in the x y z coordinate, Then the assembled subunits of these tertiary structures provide the coordinate structures. So, let today we will discuss mainly on protein primary structure and where shall we get the information from; for the protein primary structures, are their databases available? And how the extract the data? And how to utilize the data? Right. So, I

discussed earlier primary structure; is a specific combination of amino acid residues in a protein.

(Refer Slide Time: 03:28)

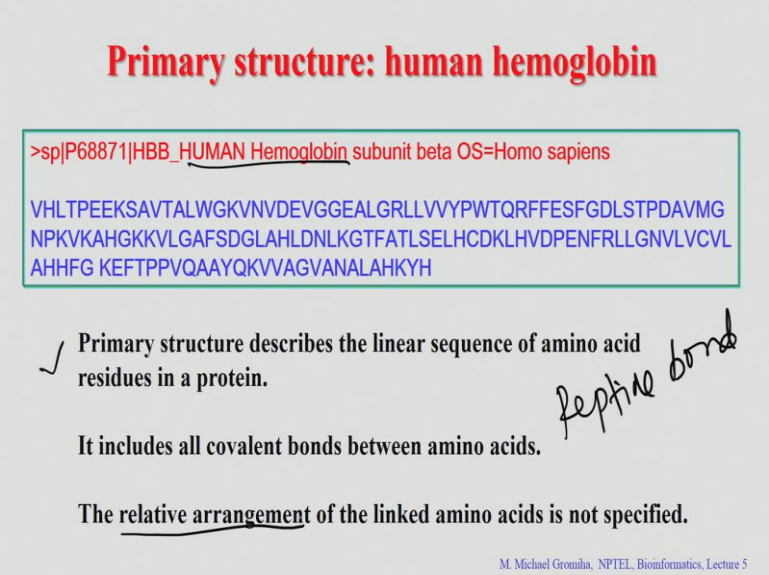


Totally how many naturally occurring amino acid residues 20 amino acid residues right 20 amino acid residues, but this 20 amino acid residues can make various combinations right you can see the combination of various possibilities right. So, you said find to have any possible combinations or only specific combinations are allowed or possible for protein function. So, this is kind of English dictionaries now if there are 26 alphabets right. So, where this 26 alphabets you can make various words sentences right and you can put this articles and books and so on.

Likewise, there are 20 different amino acids right they can also be used to make several functional proteins in the case of this 26 alphabets in English right. So, only some specific words are present in a dictionary. For example, if you take right like this education right this is present in the dictionary and this university this also present in the dictionary, but if you write like this it is as no meaning. Likewise, it will have write like this right. So, I put U N and all this 4 5 6 Is and some Vs right. So, it has no meaning, it is not available in the dictionary likewise the proteins formed by different amino acid residues, but there should be some specific combinations right and then again if you see in this chain, this is the main chain. So, like the varying chain.

So, these are the beads. So, this is common for all the amino acids; this is called the main chain right, this is called the main chain and here the beads are different. Right, this is called as side chains. So, if you see here; the same bead can repeat again, they can come close to each other or they are far away from each other, but nature selects a specific combination of amino acid residues to form a functional protein like this case you can make any polypeptides, but all polypeptides are not proteins; only the specific combination of this polypeptides, they have formed a functional protein.

(Refer Slide Time: 05:43)



So, I show some an example. So, this is the amino acid sequence for human hemoglobin. So, they have the specific combination amino acid residues and even if we describe any of these amino acids; for example, we discussed earlier right the; for the case of; right the; glutamic acid 6 to valine, it causes the disease sickle-cell anemia. So, likewise if you even a small change in a protein, that can alter the function of a particular protein, sometime they may lead to disease too. The combination of this amino acid residues is very important for a functional protein.

So, what is the information we know from the primary structure? So, the primary structure gives the linear sequence of amino acid residues. So, it will give you the specific combination of these residues; that is we know and the second one it includes all covalent bonds because when we make 2 different amino acids. So, what will happen; how to combine the two different amino acids?

Student: Covalent.

Covalent; by the elimination of the water molecule, we form a bond; what is the name of a bond?

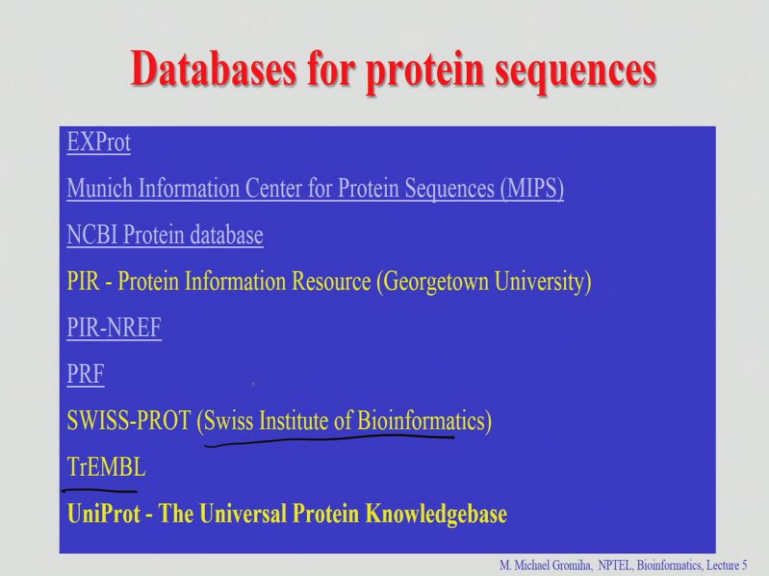
Student: Peptide bond.

Peptide bond right; so you have the peptide bonds. So, when we know the all the covalent bonds between all the amino acids right, then if you the relative arrangement of this amino acids are not specified, we do not know where they are located. So, we know only the combinations and we know the covalent bonds. Now, there are various sequences available in the literature when they synthesize the protein, right somewhere or they get this protein sequences right they publish the data in the literature.

Now, to understand the features are there any important characteristic features among these sequences, it is essential to have a set of sequences; if you have the collection of sequences, then we can easily know what is the preference of is specific residues in a protein, right; are there any bias of any specific amino acids, because if we have 20 different amino acid residues right are there any amino acid with prefer to occur more in any proteins or prefer to occur less any protein and so on.

So, what happened; then the Margaret Dayhoff right from the Georgetown University. So, she collected the sequences of various proteins and published a book of Atlas of protein sequences, she publishes very large volumes of books which contain this sequences of proteins, later on, the change this information the converted the information available in the book in the form of a database.

(Refer Slide Time: 08:11)



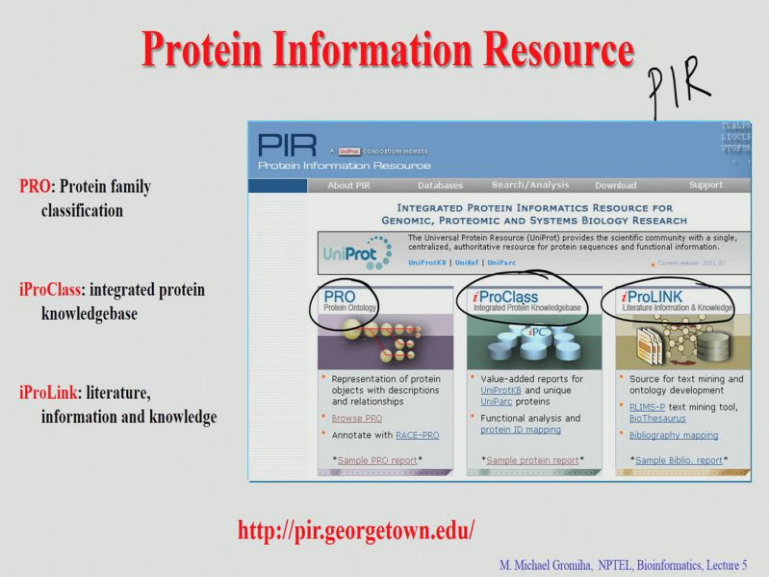
This is how to regard the database called the Protein Information Resource right; there are various databases which contain the information regarding protein sequences like the Munich information center for protein sequences, right NCBI protein database and PIR protein information resources developed by Margaret Dayhoff from Georgetown University.

The mean time; the PRF; Protein Research Foundation, they also developed to collect the data is from the Osaka University, then SWISS-PROT; it is the unique resource for the protein sequences and they created and maintain from Swiss Institute of bioinformatics. So, they have 2 types of information; one is the manually curetted sequences and also they translated sequences from the DNA sequences, this is called the TrEMBL, later on they combine every both this major databases like PIR; PIR and SWISS-PROT, right and they formed a consortium called the UniProt, right this is the universal protein knowledgebase which contains the information regarding protein sequences plus other information; what we can use as tools. Although, these 2 major databases PIR and SWISS-PROT; they merge together to share the information regarding protein sequences.

They develop tools by these groups; they maintain by themselves. So, you can share and you can get these sequences from UniProt, but if you want to use any of these tools, there you can use the SWISS-PROT or PIR. So, they have specific tools to analyze these

protein sequences. So, then I will explain what are the information earlier they started to collect and how they emerge together to form the UniProt. Protein Information Resource this is PIR.

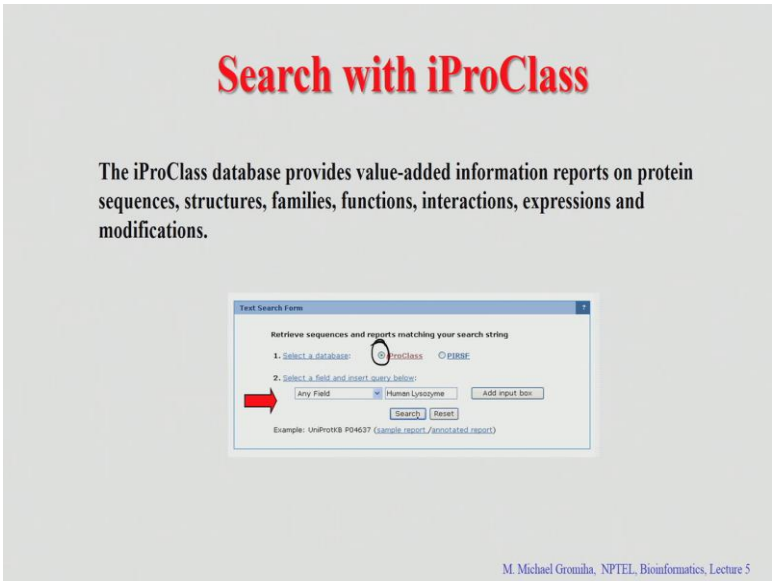
(Refer Slide Time: 09:56)



This is the one which initially started from the Georgetown University right. So, they collect the data. So, they translate the data available in the book in the form of the web resource, right and they classified into 3 different groups; one is the protein ontology, right this PRO and iPro class is contains integrated protein knowledge base and the iPro link is contains literature information and knowledge. Essentially, this is the integrated protein information resource.

For understanding the ergonomics data; proteomic data as well as systems biology research; so try to pull together. So, they collect the data and they classified into 3 different categories and each one present in different ways. For example, in the case of protein ontology, mainly what they give; they represent the protein objects with descriptions as well as the relationship with they give the different function information. Here the iPro class, right they give the functional analysis and the mapping as well as mainly the sequence information; the link they give this source for the text mining and as well as the ontology development; we can have the different links with the different aspects like bibliography mapping and so on.

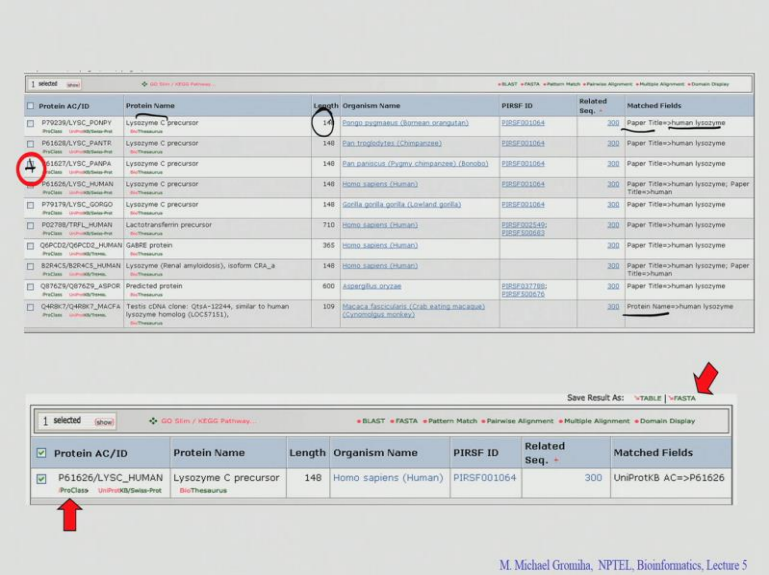
(Refer Slide Time: 11:11)



Then how to use this database; so the iPro class, they provide a various order of information on various aspects like protein sequence, protein structures, families, functions, interactions, expressions, as well the modifications, right. So, they give the protein sequence and for each protein sequence, they added information regarding other resources. So, they have various options to search a data only say simple search in this case you can give any keywords and again get the data using specific keywords or they developed advanced search options, here they have various conditions. So, you can give any of the conditions based on user choice and the database uses all the information as a query and then provide you the decide output.

So, here for this is the simple text; so you can select a database at the PIRSF or iPro class because iPro class contains the mainly sequences. So, you click on this iPro class, right and here there are various options for the query; whether you want to search with any specific field or you want to search on any field. So, here we use any field. So, then I put human lysozyme. Let us say you want to get the information regarding human lysozyme. So, if click human lysozyme and search, right you can click on search.

(Refer Slide Time: 12:36)

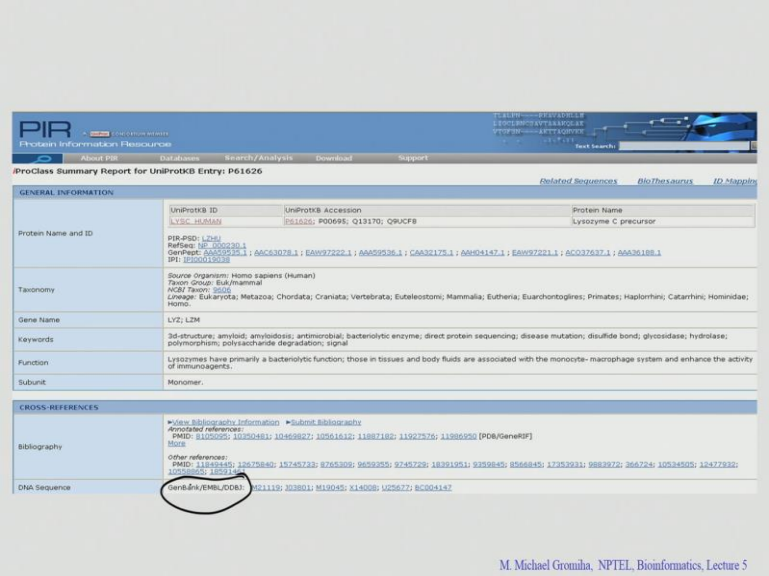


So, this will give you the whole picture. So, we will take your query, right. So, last time we discuss about databases. So, what are they generally list biological databases relational database?

So, they use the query and what is the language they use? Structured Query Language the SQL that is structured query language to pick up the data from the database. So, they use this lysozyme as a query in any of the fields right and wherever it finds the name; the lysozyme in the data that will display the entries. So, if you have there are several entries. These are the protein id and this is the protein name. So, they will give the length of the protein; what is the meaning of 148; 148 amino acid residues in the particular protein. So, then the organism; what are the organism/species for the particular protein and this PIRSF id and related sequences and what is the match field where the match field because human lysozyme is the field we gave. It matches to the paper title.

Sometimes they have the paper title right somewhere we can have the protein name work for here it match the protein name right. So, it matches wherever it matched, it showed where they found the match. Now, if you take this one. So, click on this then we can go get to more details on that particular specific protein. So, you go the iPro class and they link with the different other resources.

(Refer Slide Time: 14:02)

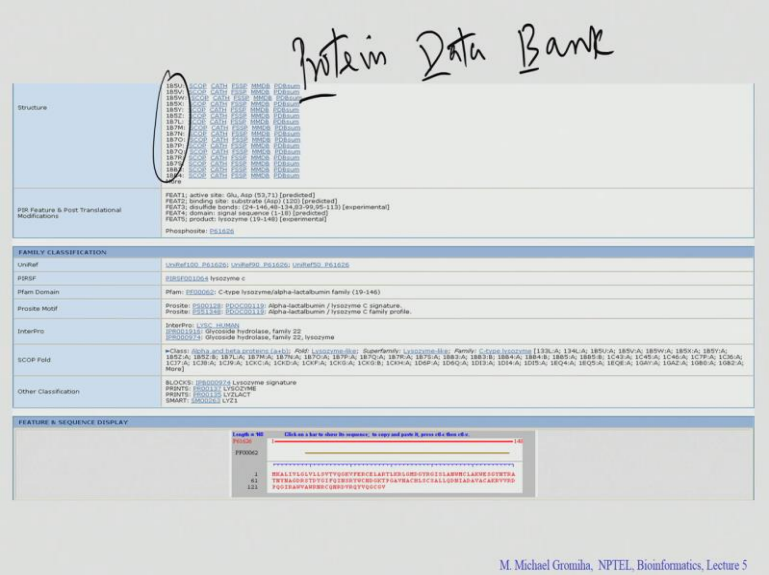


So, go with the iPro class. So, then we will get the information. So, here we first we give the protein name. So, then give the all the other synonyms all these things right then we get taxonomy, for example, this is from Homo sapiens from the human right.

This is the Mammalia and you can see the lineage is Eukaryota, from Metazoa right and go up to the last one that is homo sapiens, they give all the lineage of the particular protein, then again the gene name and the different keywords they give right because it is important because of several keywords. So, that one can obtain the decide information with any different searches right you can get from 3D structures or amyloids or the disease mutation, we can contain several disease mutations, disulfide bond, and so on. So, you can obtain a specific protein sequence based on various search options this is where they give various keywords in this a particular protein then give the function right.

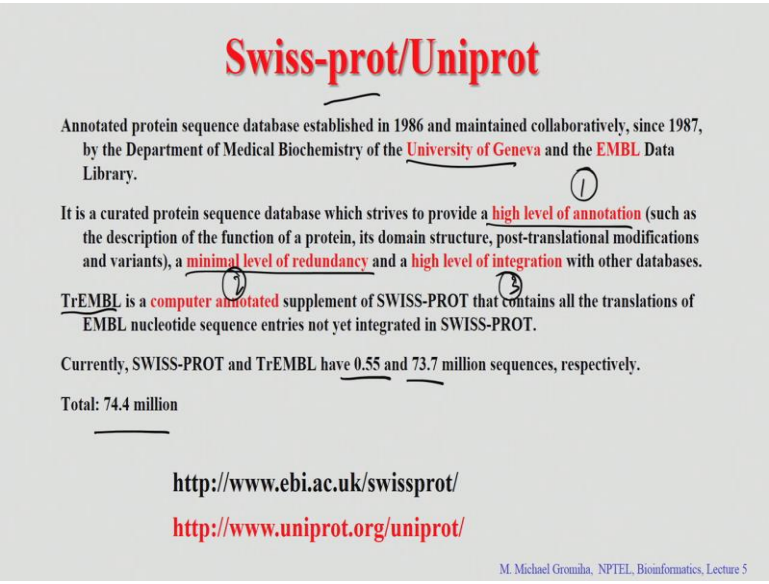
So, this is the mainly a bacterial function right. So, here the tissues in body fluids are associated with the monocyte-macrophage system. So, they give the specific function of a particular protein then you can class references right. So, they give the bibliography DNA sequence what are the different DNA sequence databases DDBJ EMBL and GenBank right. So, they give the link to the all the databases right they give GenBank EMBL or the DDBJ and so on. So, now, I give the structures I will discuss in later classes when we have the 3 d structures.

(Refer Slide Time: 15:23)



So, they give the codes for the protein databank codes right. So, they give the codes for the protein databank this is called the PDB right they have they give the ids then give the family classification right based on the folds and there are the classifications PFAM domains and so on right, then give a sequence here this is the sequence of their particular a protein. So, for the human lysozyme that is why they can see the number of residues right in this particular protein.

(Refer Slide Time: 16:02)



So, now this is the iPro class we can get from the PIR database. In the same time, this is instead of bioinformatics they also try to collect the data and they develop database called the SWISS-PROT right later in UniProt comes mainly from this “prots” because they have the various functional aspects and they collected a lot of information regarding protein sequences plus the analysis.

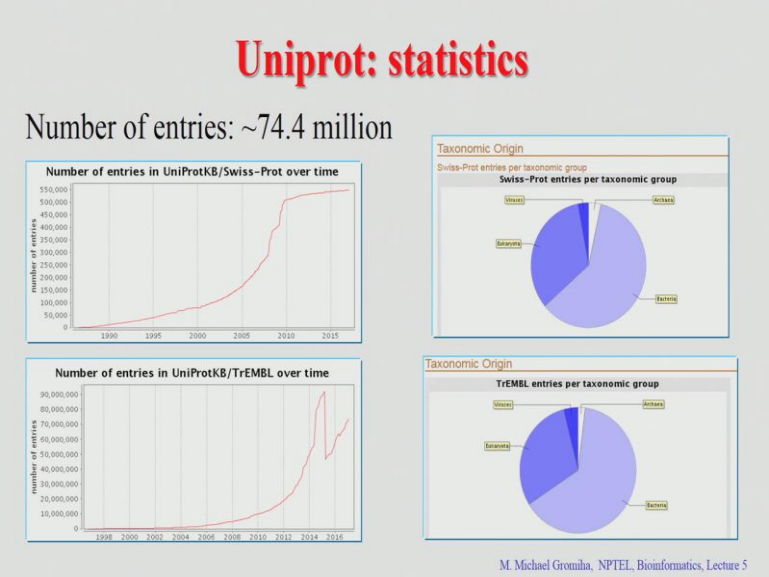
So, this SWISS-PROT is established in 1986 right from the University of Geneva and EMBLEMBL data library if this is the curated database right and the major aspect of UniProt this has high level of annotation they give much information regarding particular protein. In the later slides, I will explain some of the major aspects of this UniProt in the provides a high level of annotation, for example, they give the description of the protein various functions right and the structures are they post-translational modifications and different variants, diseases and the interactions and so on.

So, give wealth of information for a particular protein then second aspect if this is the first aspect let us say high level of annotation and the second aspect is it has minimum level of redundancy the redundancy if you see this sequences they provide minimal sometimes you can get the same information from different resources they tried to put in the resources, but try to minimize the redundant information right and the third aspect it also has high level of integration with the other databases. So, they give the links to various other databases in the literature. So, there are 3 major aspects of this UniProt what are 3 major aspects of the UniProt high level of annotation minimum level of redundancy as well as high level of integration

Now, these are the major aspect this is the reason why several a users for this particular database called the UniProt database then they have another supplementary data this is trEMBL this is translated sequences right this is the computer-annotated supplement which contains all translations of the EMBL nucleotide sequence entries which are not integrated in SWISS-PROT mainly SWISS-PROT they have manually curated the sequence. And they provide very accurate information, in addition, they also provide the computer annotated sequence from the DNA sequences this is the reason if you see it has the this is the manual accurate data is 0.55 million and this is 73.7 million, this is the computer annotated sequences and totally we have 75 million sequences, right.

This is a good amount of data right of protein sequences. So, you can use this database UniProt database. So, the website the UniProt dot org you can get all the information regarding a particular protein.

(Refer Slide Time: 18:37)



So, we look into the UniProt. So, if you want to use a database as I discussed in the previous classes if you want to develop a database. So, you have to maintain the database it is easy to develop a database but is very difficult to maintain a database right. So, there should be a uniform increase of data and there should be a very reliable data right. So, and then divided very clean data right this is very important. So, if you see these UniProt you can this is the data for the UniProt a SWISS-PROT and this is for the TrEMBL right this TrEMBL, SWISS-PROT.

This is the manual curated one. This is the computer-annotated one. So, once we gradually increase of these sequences they started working on these sequences and they spent several efforts of input, to curate the data. So, you can see gradually increase in data likewise the TrEMBL also you can see the gradual increasing data right maybe they material you deleted some of this entries this is the reason there is this lot and then again it's growing up. So, they have main gradually maintained the data continuously maintaining the data and the gradually increasing the data they show the growth of this database and if you see the citations will get the citations also improving every year because for any protein sequence information one has to use this UniProt database.

So, now if you look into the classification which type of data is dominant in UniProt database right if you see this is the classification. So, mainly you can see the data dominated with which type of organism which kingdom of life this is mainly the bacteria you can see this is the bacteria data, this highly dominant and followed by (Refer Time: 20:08) and then virus and then a little bit and then about the archaea. So, you can see the similar preference for the viruses and the archaea.

In both cases, even for the SWISS-PROT entries as well as the TrEMBL entries you can see a similar level of data mainly dominated by the bacteria followed by (Refer Time: 20:25) Eukaryota and viruses. Now the question is there are many proteins in the database, what is the average length of the protein? So, there are many proteins right. So, there what is the average length of the protein?

So, what is the average length of the protein?

Student: 3 1 5.

315 amino acids based on the all the proteins deposited in the UniProt database, usually, if you look into the sequence length in different proteins in UniProt.

(Refer Slide Time: 20:53)

So, you can see this is mainly this region this region means the that is about a 100 to 400 residues right most of the proteins are small protein like proteins are medium range proteins or 200 and 300 proteins, right you can see the higher peak are on 250 300; 300

proteins, right and some proteins which are quite long which is more than 1000 residues right when they finally, if you take the average length it is about the 315 amino acid residues in each protein, but there are most of the proteins they are in this region. So, that is about a 100 to 400 residues.

So, now if you look into these different proteins right are; they are any bias of different amino acid residues totally. How many amino acid residues?

Student: 20.

20 residues right and different classifications polar, non-polar, charge and so on, is there any bias any specific residues are predominantly occurring in protein sequences right. So, here I show the data. So, this data shows the gradual decrease and here this is classified based on the different groups, right. If you see, there are some amino acid residues which are highly occurring in protein sequences right what is the residues occur dominantly in protein sequences leucine, alanine, glycine, right. So, here these are the residues which occur predominantly in protein sequences when looking into the rarely occurring amino acids, what are rarely occurring amino acids.

So, tryptophan. This data obtained from few releases ago. This is the current data, release one. So, if you see the tryptophan there is just one point 2 percent and for the sixteen the same 1.2 percent. So, highest occurring if you see leucine is nine point eight percent right and the alanine is the eight point nine percent if all the 20 residues are randomly distributed like; what is the percentage of each amino acid residues?

Student: 5 percent.

5 percent right, but here if you see there is not like that right some residues are ten there is the double amount and some residues are only 1.2. So, this is mainly due to the structure and function aspect is very important right for to fold a particular protein in their globular shape as well as to maintain the stability and for the functions. These hydrophobic residues they tend to form a hydrophobic core that is initiating the protein folding and maintain the stability likewise you can see the sufficient number of charge residues like for example, lysine and arginine, right; so here the lysine and arginine. So, what is a percentage of lysine and arginine a 5 percent lysine and the arginine?

Student: 5.6

5.6 percent as well as the polar residues that are serine, threonine. You can see sufficient percentage because they tend to form electrostatic interactions as well as form the hydrogen bonds. Now, these residues have specific bias in the unique protein sequences.