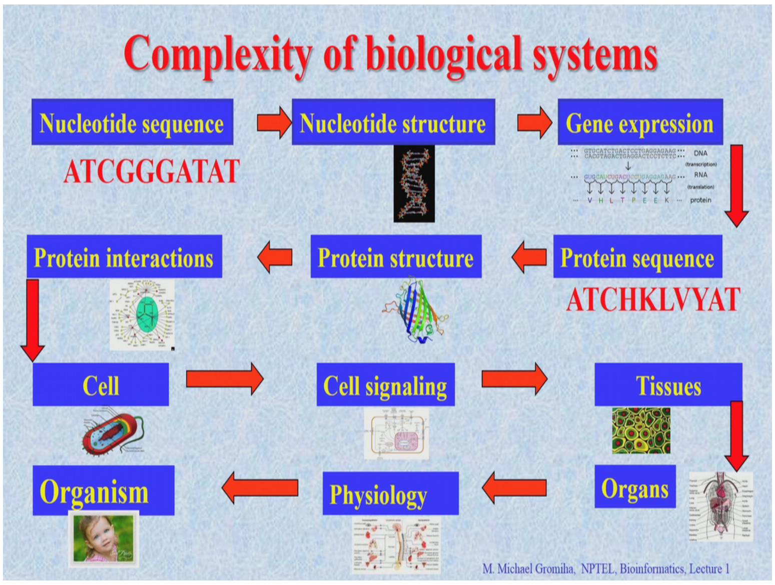
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 1b
Complexities in biological systems

So, I will explain how these bioinformatics is important in different complexity of living systems.

(Refer Slide Time: 00:21)



Complexity of biological systems

Nucleotide sequence → Nucleotide structure → Gene expression

ATCGGGATAT

Protein interactions ← Protein structure ← Protein sequence

ATCHKLVYAT

Cell → Cell signaling → Tissues

Organism ← Physiology ← Organs

... GTGCATCTGACTCCTGAGGAGAAG ... DNA
... CACGTAGACTGAGGACTCCTCTTC ...
(transcription)
... GUGCAUCUGACUCCUGAGGAGAAG ... RNA
(translation)
... V H L T P E E K ... protein

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 1

So, if you talk about the complexity of biological systems. So, we have the different ways first we start with DNA nucleotide sequence right I put a nucleotide sequence here. So, then we develop with the nucleotide structure. So, if we started with the order different nucleotide bases ATCG. So, we go the nucleotide sequence, then we go with the nucleotide structure and from nucleotide structure, then we go with the gene expression. So, we go with the DNA, DNA to RNA and RNA to protein.

So, you can go to the gene expression and finally, we get the protein sequence. So, once we protein sequence then we go into the next level, this is called protein structure and then from protein structure to protein interactions and protein interactions then this is say opening the cell right and then we go for the cell signaling and go with the tissues, and develop to the organs and the physiological system and finally, the organism.

So, this is the complexity of the biological systems. So, how bioinformatics contribute to understand the complexity of the biological systems? So, if we look into the applications of the bioinformatics, you can see there are various aspects to contribute in different stages of these biological systems. For example, if we go into the nucleotide sequence, how bioinformatics can contribute to nucleotide sequence? Storage of large number of sequences and analysis nucleotide sequences right. So, for example, if you look into this nucleotide sequence, there are millions of sequences are available now. So, they are stored in the kind of databases for example, you can say the few types of few databases can you list few names of the databases.

Student: NCBI.

NCBI; so GenBank you can get right EMBL we have the data and mainly the DNA data bank of Japan DDBJ right. So, there are various databases. So, they collect the data right put everything together and they organize a specific way right. All these databases are freely available and you can keep have several search options to extract the data and you can use this information for analysis. The first one is we have the clear database for nucleotide sequence and when this database is ready, then we can do the analysis right which type of analysis you can do from the nucleotide sequence? You can say interaction analysis, you can say a thickened content, you can say a GB content or the any specific properties for example, whether the DNA regions are highly bending or is highly flexible because flexibility is important for the binding right. So, we can derive various features for example, if you have a DNA sequence what is a melting temperature, this can withstand for higher temperature or low temperature. So, all the information you can get from the bioinformatics analysis.

Now, we go with the nucleotide sequence to nucleotide structure right, what are the information you can bioinformatics contribute to nucleotide structure? So, if you have the structures right first you can make a database, right for can you one database for nucleotide structure?

Student: (Refer Time: 03:11).

PDB contains a data or the PDB protein databank anyway. So, nucleus database also contains the information regarding the nucleotide structure. So, then you can also as you mentioned you can also predict the structure from the sequence right there are many bioinformatics tools available, to predict the nucleotide structure from the sequence. Then from the structures you can analyze various parameters for example, various based up parameters for the nucleotides right how all the how about the base tracking energy or the or different bases, how is the base pairing energy, how about the hydrogen bonds right various types of interactions.

So, all these information you can get from the nucleotide structures and we have different algorithms available in the literature to analyze the nucleotide structures as well as to predict the nucleotide structures from the nucleotide sequence. Then go with the next step right from the nucleotide structure to the gene expression. So, how the bioinformatics contribute to gene expression?

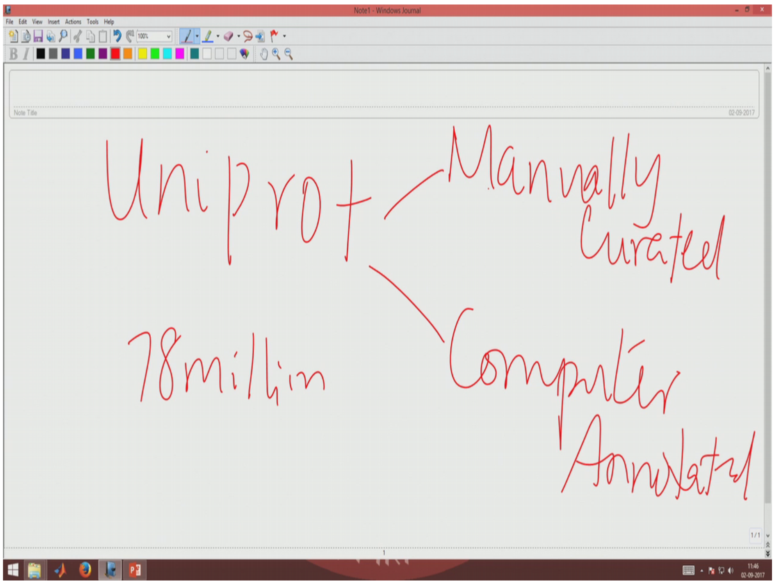
Student: You can predict the gene expression using bioinformatics means, how many proteins, how many RNA sequence are transcribed from the gene.

Correct. So, you can the coding region, you can get non coding regions and you can translate the sequences, the DNA sequences to the protein sequence then so on right. So, you can have the algorithms to convert this translate these DNA sequences to protein sequences and so on. Then go with the protein sequence. So, how the bioinformatics help in protein sequence?

Student: From the gene expression it can again predict the protein sequence, what the protein is probably making. And from protein sequence it can identify what is the contribution of each residue in different kind of protein.

Correct. So, if you go with the first two protein sequence right. So, many sequences which are deposited right are accumulated and they are accurate and they deposited in a form of a database right; can you tell a the database which contains the protein sequences?

(Refer Slide Time: 05:05)



Uniprot right; uniprot is a database right, Uniprot database which contains protein sequence information right. So, earlier it was started with the pir protein information resource right in the meantime the switchport also developed sequence databases, they merge together right we are tend to 15 minutes ago. So, they com were the combined effort of the consortium called the Uniprot consortium.

So, they collect all the information regarding protein sequences, and they put together in the form of the database called Uniprot database. Currently commonly sequences are known in Uniprot database? 17 million sequences right. So, currently if you see the Uniprot have two types of databases, one is the computational annotated and the second is the manually curated. So, we have the manually curated database and the computer annotated right.

So, kindly 55000 sequences from the manually curated and the 78 million sequences using computer annotated sequences right. So, totally you get about 78 million sequences right. So, you have the bioinformatics database right which contains all the information. This database is widely used in the literature, because this database contains lot of information right what are the information you can obtain from the Uniprot database? You can see the function notation, (Refer Time: 06:27) structures, interaction with the other proteins, linking mutation with diseases post translational modifications.

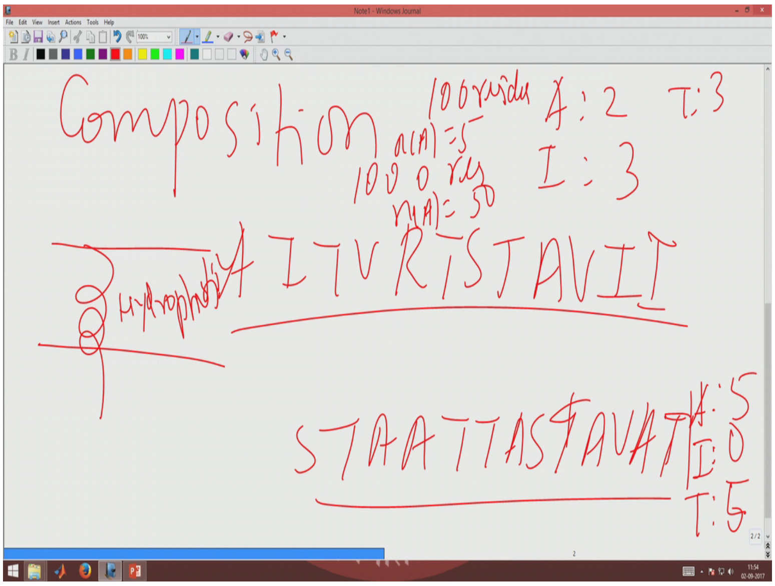
Student: Isoforms.

Isoforms. Amino acid sequences in different formats in faster format and the different format styles. So, it started with the protein names and then this synonym is and what are the various sources and organisms and finally, go with the functions right and they give almost all the information regarding the particle regarding particular protein as well as they give the literature and linked with the all the databases. So, it is high integrated ones, less redundancy and highly reliable that highly annotated.

So, there is a reason why the Uniprot is the unique resource for any protein sequences, because continuous even the protein sequence database it has all the information regarding a particular protein. So, if you are interested in any particular protein, you can go to Uniprot right and you can use it to get the information regarding your particular protein. Also it has high level of annotations and linked with the other databases right.

So, then we can use the information available in the Uniprot. So, if you go here. So, you have the protein sequence now you have the database right (Refer Time: 07:32) sequences. Now what can we do with the sequences? You can also use the sequences to extract information. So, what are the various informations you can extract from protein sequences.

(Refer Slide Time: 07:52)



So, now if you get the protein sequence right what are the various information you obtain from the protein sequence, using bioinformatics you can calculate the composition.

Student: Observation.

Right what is a composition?

Student; Number of a minute thing.

So I will explain the details in later classes. So, just I will have a brief explanation I will give you now.

So, if you have the sequence for example, if you have this sequence right. So, there is another sequence you can see like this. So, if we look into these sequences, there are some specific preference of amino acid residues right. If you see the first sequence, which residue is highly preferred can see the common times A occurs 2 times, I 3 times T.

Student: T 3 times.

Three times right.

So, we see this sequence. So, how many times A occurs?

Student: 5 times.

5 times.

Student: I 0 times

0 t.

Student: T 5 4.

4 times, right 1, 2, 3, 4.

Student: 4

This is various T then this is 5. Right, 5 times. So, then we can see a different sequences you can see the difference some residues are preferred in some in some sequences, some residues are not preferred in some sequences right. So, if you have the sequences what are the all several types of proteins? So, you can see the number of times each amino acid residue occur right and you can see this has some bias in this in the protein structures. If we got different types of proteins for example, DNA binding proteins; you know DNA binding proteins the proteins interact with the DNA, see DNA is the negative charge because of phosphate group. So, here we have that the binding region you can see the preference of positive charged residues right what are the positive charged residues?

Student: Arginine.

Arginine lysine, right

You can see the preference of positive charged residues.

So, if you have any small stretch of regions, with the occurrence of more number of positive charged residues then you can see that these sequence could be a DNA random bending protein right. Likewise if any protein contains a stretch of hydrophobic residues for example, more than 15 to 16 hydrophobic residues, then we can see that this protein let us say its sequence could probably be transformed from helical proteins; because in a transform of helical proteins you can a stretch of hydrophobic residues inside the membrane; so if for you see the helical proteins. So, if this is a membrane right then this is the helical regions here, this is highly dominated by the hydrophobic residues right. So, you can see that.

So, likewise if you have a sequence in the Uniprot, you can extract various information. So, these are currents. Now here in this occurrence the numbers depend on the length right for example, if you have 100 residues and another protein 1000 residues then if you see the currents right. So, a will be around 10 or d will be around 10 or to so on. If you have the 1000 residue protein, you probably divide it by multiply it by 20 right randomly if you distributed right how many residues you need in the in the protein right for example, if you have 100 residues the protein have 100 residues if all the 20 residues are randomly distributed right what is the occurrence of a particular residue?

Student: 5 times.

Student: 3.

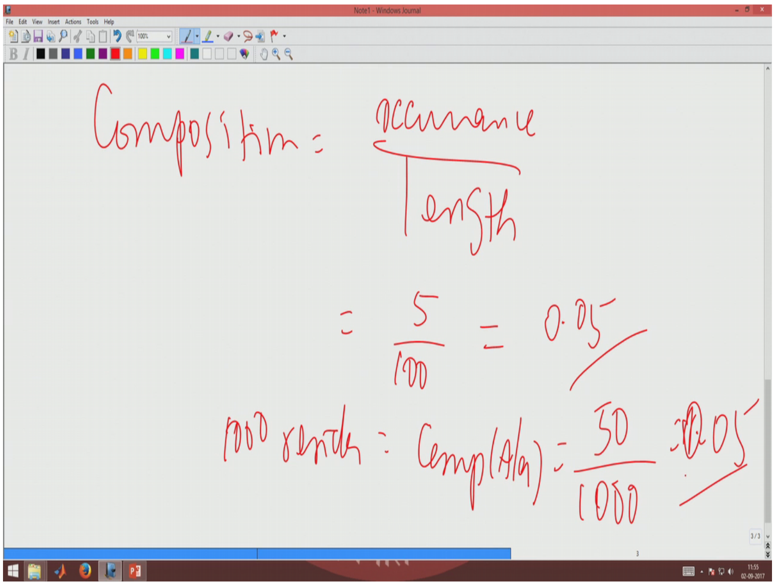
5 right; on the other hand if there is 1000 residues. So, n of a equal to.

Student: 50.

50 right.

So, the length matters if you have different proteins at different lengths the number will be different. So, by comparing these two you cannot say that this a is highly dominant in the case of second protein. So, in this case we need to normalize; then how do we normalize it with length.

(Refer Slide Time: 11:47)



$$\text{Composition} = \frac{\text{occurrence}}{\text{length}}$$

$$= \frac{5}{100} = 0.05$$

$$1000\ \text{length} : \text{Comp}(\text{Ala}) = \frac{50}{1000} = 0.05$$

So, in this case composition is the occurrence divided by length. In this case if you have 100 residues right randomly distributed right what is the composition? This is 5 by 100 right this equal to.

Student: 5, 1 by 0.

Point.

Student: 0, 5.

0, 5.

For the second case right you get thousand residues. So, what is the composition for a (Refer Time: 12:23).

Student: 50.

50 by.

Student: 1000.

Thousand right, this equal to.

Student: 0.05.

0.05.

So, in this case if you see; if you look into the just a sequence length, you can see the difference, but if you normalize with a length. So, they are same in this case you may have 100 residue protein or 1000 residue protein. So, the occurrence the composition is the same right in this case there is no no bias right. This is the reason why we need to normalize with the chain length to get a composition.

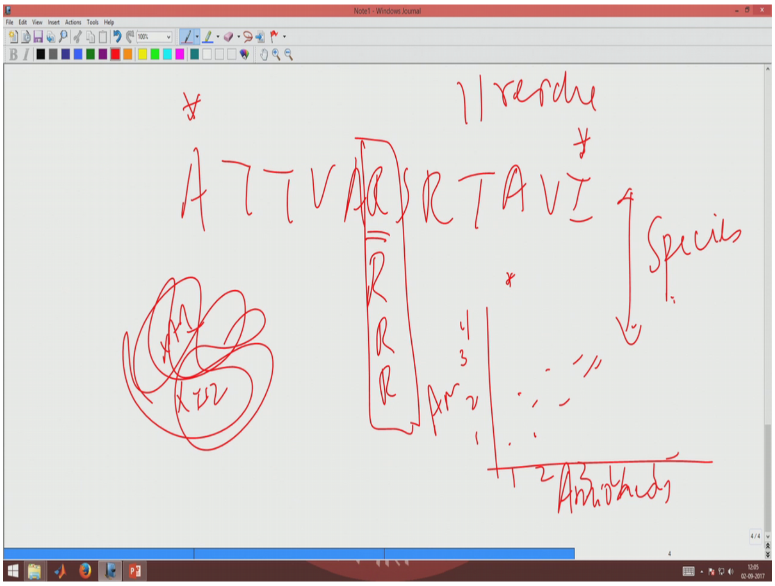
So, looking to this Uniprot database there are 78 million sequences, and you try to obtain this composition for the 20 different problem as residues right you can a see some sort of bias right for all 20 amino acid residues. If all the 20 residues are equally distributed right then what is the percentage of each residue is 5 percent right. If they are randomly distributed, but we look into this unipro database and the sequences that is not the same right some residues are higher occurrence and some are less right can you tell me some which residues are highly dominant and residues are less preference.

Student: Up to 5(Refer Time: 13:33).

Till to 5 and 16 this residues are less preferred and if you see the hydrophobic residues right well in a (Refer Time: 13:39) right you can say highly preferred, they have more than about 10 percent right in the case of 16 or the (Refer Time: 13:44) to 2 to 3 percent right. Because nature of the selection is in such a way that the proteins from the unfolded state or folded state they because of hydrophobic collapse right these residues come close to each other. So, there is dominance of the hydrophobic residues, and then once you are done then there is maintained the stability by other interactions like the hydrogen bonding, electrostatic interactions right. So, you can see that the union and the setting and the charges residues, they are next level preference in the whole amino acid sequence. So, if you have the sequence the bioinformatics contribute to various levels to understand the different types of proteins, based on the structure and based on the function.

Right we discussed about two types of the features a occurrence of the composition, there are various features you can you can derive. For example, you can construct some profiles for properties, you can get the average property of a particular sequence and you can see whether any specific residue is concerned.

(Refer Slide Time: 14:38)



That means. So, if we have a sequence right. So, we see several organisms several species you check the data, and see there are any particular residue which is present in all the places in all organisms. If it is present in all the organisms, then you can say that this particular residue is highly concerned and this plays an important role in the structural function and if you disturb this residue, then that will create several problems in the particular protein and cause some diseases right probably.

So, in this case you can see this information from the sequences because the lot of sequences are available. So, you take any sequence make a same similar sequences together, and see any residue is present in the same position in all the organisms and how they are important for the function. So, protein sequence you can see the application of bioinformatics to derive the various features of the protein sequence. Then there are various other factors, that I will explain later right into the future classes. Then next level we go to the protein structure right how the bioinformatics contribute to protein structure?

Student: Protein structure protection can be done apart from that a structure, can be used in various functions for example, in enzymes the interaction pattern uses and enhances catalytic sides.

So, catalytic sides you are right.

So, you have lot of ways you can use the protein structures using bioinformatics. First we try to have a database right which is database, which contains protein structures protein databank right. Protein databank is a unique resource for obtaining protein structures right. So, currently how many structures are deposited in protein databank? Approximately 1, 30,000 structures are deposited in the protein databank. So, we have the protein structures right and how many sequences are deposited in the Uniprot database?

Student: Million.

78 million sequences, right.

So, if you see the difference of known information between the sequences and the structures, there is about 600 to 700 folds.

So, mainly the structure determines the function. So, in this case it is very important to get the structure for any sequence if the structure is not known right that is called a protein folding problem. So, we predict the structure of a protein from the sequence, because chain sequence are known right for 78 million and the structure (Refer Time: 16:59) are very less. So, you can have the relationship; when you determine the structures it takes time right at times it involves lot of manpower and the time and the money right its cost effect cost effect. So, we try to predict. So, within bioinformatics tools, we can predict the structure of any protein from its sequence. So, if we have a sequence we can try to predict let the accuracy levels are about 8 somewhat 80 percent now. So, in later classes I will explain how to predict the structure from a sequence.

There are various methods available to predict the structure; right the foremost one is homology modeling. Because if two sequences are similar, then we assume that the structures are similar right using this concept that you can derive with same methods based on homology modeling, to predict the structure from the amino acid sequence and you do the fold recognition, because if you have any specific fold, you can use the fold recognition technique and you can use threading as well as you can use the (Refer Time: 17:55) tech method (Refer Time: 17:57) modeling.

So, we start from the scratch right starting from the bond length, bond angle torsion angle. And finally, we use calculate the energy and we minimize the techniques to predict the 3 D structures. The only the disadvantage of the (Refer Time: 18:12) modeling is it can handle some limited number of residues, because if takes enormous computational power and the time. In this case they cut into pieces and then do the modeling and they join together. And finally, they minimize the energy to get the final folded structure.

So, now if we have the structure, we can predict the structure from the sequence and if you have the structures what are the various information you can obtain. You can use the bioinformatics tools right or you can use the bioinformatics algorithms, to derive various features from the structures right. For example, what are the various features you can derive from the protein structures? Several types of similarity you can do for example, if you have a protein right. So, if you have a protein here a pseudo protein. So, some residues which are here alright, these residues are not easily accessible. So, if you there in the outer surface if you see here. So, these residues are highly accessible. So, if we have the structure, we can easily see which residues are buried that is inside the core and which residues are outside the surface. Where did you get that information we can see if we from this we can see which residues get interact with other systems, other proteins or ligands or nucleotide acids and so on.

If that is at the interior core, they are responsible mainly for keeping the protein stable right they are mainly contributing to the folding and the stability. See if you look into these residues which are at the outer surface right these residues, they tend to interact with other molecules such as proteins or nucleic acids or small molecules right for the interactions right and for the function, and this is also important for the structure of a drug design just I discussed earlier. So, if you have a target and to identify the ligands to interact. So, that should have some pockets right if you know this structure, you can identify the bending pockets the bending sides, where is small molecule or ligand can interact and with the amino acid residues in a protein. And if you have the 3 D structures right there are various many residues in a sequence, and these residues interact with each other and form a structure right and this can easily tell you which residues are close to each other.

So, for example, if you have the 3 D structures right this is a sequence. So, let us form a 3 D structure like this, there are various occasions where these two residues which are faraway in the sequence, how many residue is far apart? 11 residue is far apart right, but this can be possible these residues are close to each other. Now for example, you say a is here a one is here and the I 12 is here right. So, if we have the 3 D structures, you can get this information. Where how these residues interact with each other and how they are located right. So, in this case you can see the conducts oh if. So, this is the amino acid sequence, here also amino acid sequence and this residue 1, 2, 3, 4, 5 like this right. So, how whether they are close or they are far. In this case you can see for example, if this and this are very far, they can be close in space some case the close residues they are always close.

So, you can see the information that how far the residues in the amino acid sequence are distributed very far away and close or they come together in the case of protein 3 d structures right. How the residues in a space are distributed in the sequence right there you get the information regarding. Then we can see the residues, which contribute from far away and which residues which make a short range contacts. And all the information you get from the 3 D structures.

Now, if you have the structures then also you can derive various features like for example, you can see the conduct order, long range order and the hydrophobic behavior of the residues and various aspects you can obtain from the protein 3 D structures. Then we translate this information to predict the structures to understand various functions of these proteins right this is the reason why we try to get several parameters from the protein structures.

So, in our next level we go to the protein interactions, what are the various protein interactions?

Student: Protein nucleic acid.

Protein nucleic acid interactions.

Student: Small molecule.

Protein small molecule interactions.

Student: (Refer Time: 22:14).

Protein interactions right.

So, now we can the protein interactions, also the bioinformatics contribute significantly to indentify the residues right. So, because crystographers, they tried several structures and they deposit the structure the acid.

So, to understand the residues and you call the uniform information, why are the proteins can protein interact and which residues are preferred to interact and these residues are preferred to interact right are any specific patterns available for it to proteins to interact. So, in this case we use bioinformatics to understand this specific binding sides right now various criteria to define the binding sides right what are the various criterias to define the binding sides?

Student: Distance.

You can use the distance right whether they are close to each other, then the proteins try to interact and you can use the accessible surface area right how far the surface area reduced upon binding. So, based on that you can see whether any these residues interact or not. We can calculate the interaction energy right and from that you can define whether these protein two proteins interact or not.

Likewise, you can see the protein ligand interactions, protein RNA interactions, protein DNA interactions all these 3 you can do. Then there is the various databases available for the protein interactions right on various aspects. First one is you can see the interacting patterns interacting pairs, which two proteins interact right and the interacting sides right and specific to some of the genomes like human genomes. If we take the human genome currently there are how many proteins in the human genome, how many proteins? 20 on to 20,000 plus right; so in earlier it was assumed to 1 to 100000 10 30,000 50,000 right currently the Uniprot if you see around 20,000 plus sequences in the Uniprot database right.

So, if all these sequence then which proteins interact with each other. So, what are the two pairs interact right. So, this information we have collected the information, there are various databases. Database for interacting proteins right and for the binding affinity of these proteins, if the protein a interacts with protein b what is the binding affinity of these two proteins to interact. So, there are various databases available for the case of the protein interaction right. So, from this interaction studies, we can try to derive the hypothesis how the two proteins interact with each other.

Likewise we can see the other the other aspect like cell right we can have the various databases as well as analysis right to understand the different activities for a cell. And cell signaling and different pathways right and have the tissue data to see the organs and physiology as well as the organisms; for example, neuroscience and so on right. So, you can use the bioinformatics to understand the complexity of the various levels of complexity in biological systems.

In the next classes I will start with a different aspect starting with nucleotide sequence to up to these different types of interactions.