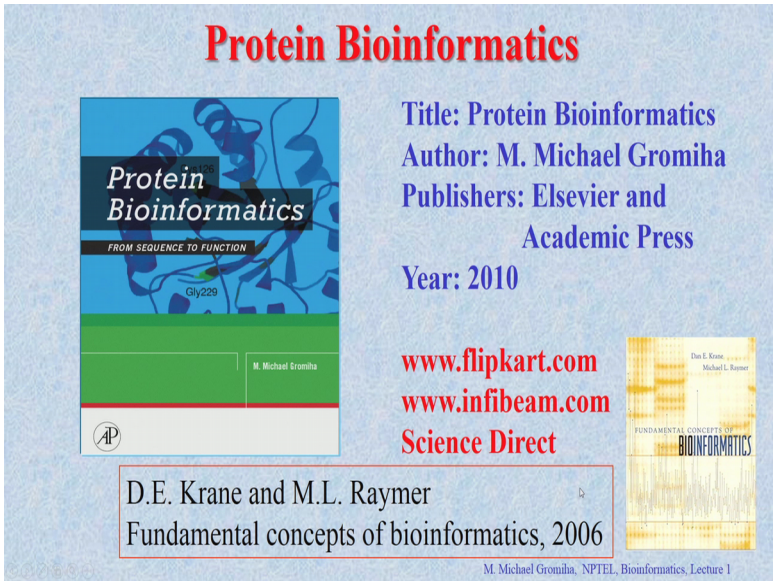
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 1a
Bioinformatics

Welcome to the NPTEL course on Bioinformatics I am Doctor Michael Gromiha from the Department of Biotechnology, IIT Madras. In this lecture I will provide an overview on bioinformatics different aspects of bioinformatics and the applications of bioinformatics and different complexities of biological systems. In this subsequent lecture I will describe the details on various aspects.

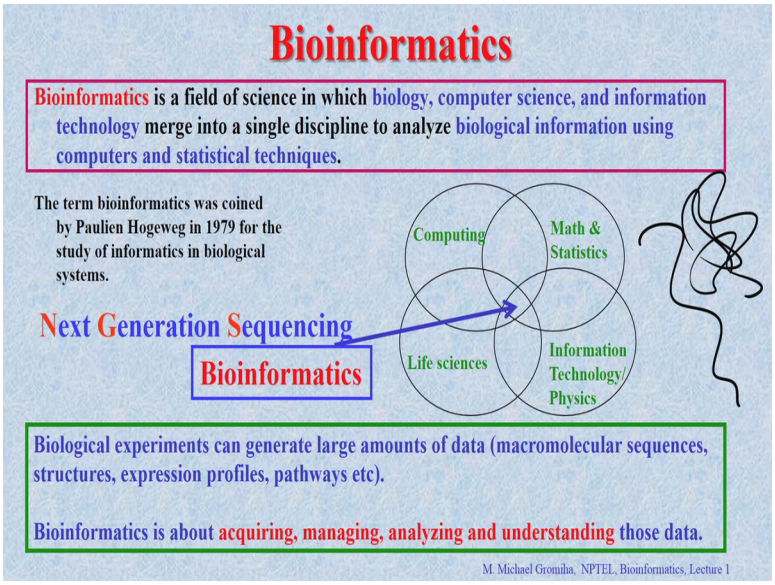
(Refer Slide Time: 00:41)



In this course, I mainly follow the books protein bioinformatics written by me and published in 2010 by Elsevier and academic press and the book by Krane and Reymer fundamental concepts of bioinformatics and published in 2006. For the general concepts of bioinformatics I will use the Krane and Reymer book and for the applications on various aspects of proteins; such as protein sequence analysis protein structure analysis protein structure prediction and protein folding, I use my book on protein bioinformatics

So, what is bioinformatics? If you split it into two parts: bio plus informatics, so applications of the informatics on biological systems.

(Refer Slide Time: 01:26)



So, I put the bioinformatics in the central part. So, it is a field of science in which biology plays a major role right and combined with other applications from different fields of science, such as computer science, information technology and others to merge into a simple discipline to analyze the biological data using statistical techniques as well as the computer algorithms.

So, if you see this diagram. So, I put the computer scien the bioinformatics at the middle with all other fields, which are linked with bioinformatics. See if you into this width of bioinformatics in fact, the analysis of biological systems have been carried out for the past few several decades using a small scale analysis; and in 1979 Pauline Hogeweg coined the word bioinformatics to deal with the applications of the different science in biological systems.

So, we look into the different fields of science, how the different fields contribute to the birth and the development of bioinformatics. For example, computer science can you tell one example how the computer science contribute to bioinformatics.

Student: Machine learning can be used probably and in computer science, there are a lot of computer algorithms have been used for solving the biological problems.

Correct. So, you can develop download several programmings and you can extract the hidden data, available in biological informations. Also you can use different machine

learning techniques and the algorithms, to understand and to capture the information as well as for the prediction purposes. So, can you tell one example how the mathematics or statistics is used to the development of bioinformatics?

Student: Example would be use of statistics for an example in the use of (Refer Time: 03:11) plot they plot the phi and psi angles of protein by whole (Refer Time: 03:15).

Correct. So, you can use mathematics right to derive some principles and relate the data for example, the protein sequences or how the distribution of the rest use in the (Refer Time: 03:30) plot and so on using this mathematics and you verify the data whether you obtain any model. So, whether they are statistical significant or not you can relate with the correlation analysis you can relate the regression techniques, as well as you can see whether the data are statistically significant or not. So now, if you use the information technology; how the information technology contribute to the development of mathematics.

Student: There has been a increased computational resources over the decades, which is used to do a large scale data analysis (Refer Time: 03:57).

Correct. So, you can use for a logic analysis, you can develop the online resources right you can see the computer storage and so on in this case that will enhance the applications of bioinformatics to various fields; likewise physics if you talk about physics. So, you can see the concept of various types of interactions like electrostatic interactions (Refer Time: 04:15) interaction auto (Refer Time: 04:16) interactions, how these interactions are important to understand to the folding mechanism proteins. For example, if you take protein folding in the unfolded state of a protein. So, it is like wobbling and you can see it is very like a random coil confirmation.

When this protein folds in the specific three dimensional structures, it can form a specific three d structures. And how a protein can attain a specific 3D structures from its sequence right this can be explained by various types of interactions like disulfide bonds electrostatic interactions, frontoverse interactions. So, understand the principles governing the folding state of a protein, it requires the physical concepts. So, you can use physics to understand the mechanism of the folding of in a (Refer Time: 05:03) of a proteins. So, we consider all the fields; if we say life science, computing maths, stat

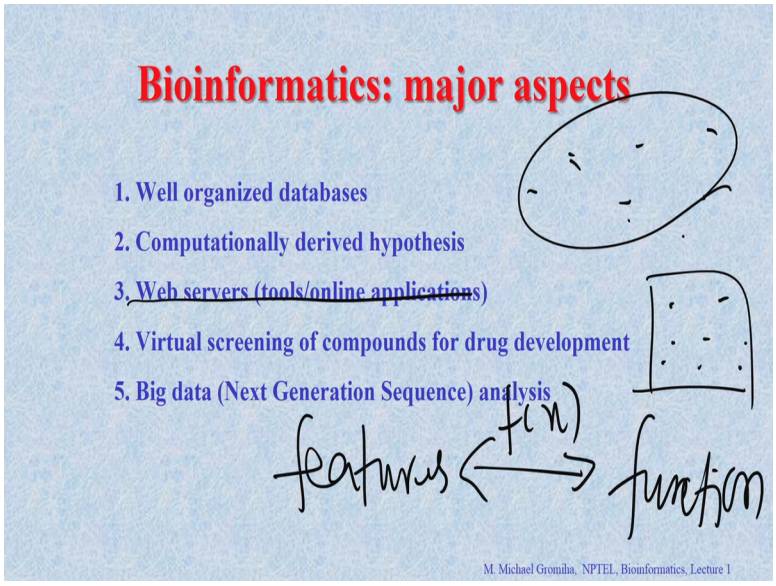
information technology or physics or chemistry, which one is the some major field for the birth of bioinformatics.

Student: Life sciences.

Life sciences right because we need the data. So, without data even if you have several fields, we cannot apply it to different data right. So, we need a specific data. So, data where shall we get the data, we can get the date from biological experiments. So, if we look into this life sciences there are produced a lot of data right on a various aspects such as the macromolecule sequences for example, DNA sequence or protein sequence right and the structures, protein structures, DNA structures, complex structures and so on different expression profiles different pathways and so on.

So, how the bioinformatics is used to understand the data? So, bioinformatics is help to acquire the data, to manage the data and to analyze the data and to understand the data right. So, bioinformatics is the major field right to understand the concepts and understand to the hidden information available by the experiments produced in life sciences.

(Refer Slide Time: 06:07)



So, now what are the various aspects, how the bioinformatics is grown what are the various applications of bioinformatics, how the bioinformatics contribute a society that there are various aspects. So, I briefly I put into 5 bullet points, the first one is well

organized databases. The biologists they do the experiments and they produce data and they publish the data and the literature it is very important to collect all the information and put in the form of database. For example, if data are available scatteredly here and there right. So, it is very important to collect all the information and put in a proper form right and then you give some options to extract the data from this database. This is what the bioinformatics do to develop a plenty of databases which are well organized or in computable form

Once we have to data right submission number of data, which is very essential and it is required for any analysis right otherwise you do not get any statistical significant data. So, second option is once you have the database, you can derive hypothesis what will happen what is the relationship between some specific features as well as any function. If you know the function or if you know any specific characteristics of any biological systems, what makes these characteristics to a particular systems right? So, here we try to develop some features right and using these features you can link the function of any biological system right for example, if you take a protein sequences or protein structures there are different types proteins say for example, if you get a protein structures globular proteins, membrane proteins and different types of globular proteins say alpha proteins beta proteins and so on and whether we can identify these types of proteins from a pool of sequences right.

So, you can left side you can see the sequences; because if you have sequence you know the information regarding the amino acid residues right. So, you can get the number of residues of which type. So, there are 20 different types of residues right present in these proteins, and you can relate these fields there is dominant of some specific residues for example, hydrophobic residues, then you can say that this could be a (Refer Time: 08:24) protein; likewise for any proteins having different functions or different type of diseases right. So, you can derive the hypothesis what is the basic principle for having this biological systems once you derive the hypothesis to understand these are the major factors for any specific systems right the next step is whether we are able to describe any algorithm, whether we are be able to derive any algorithm right.

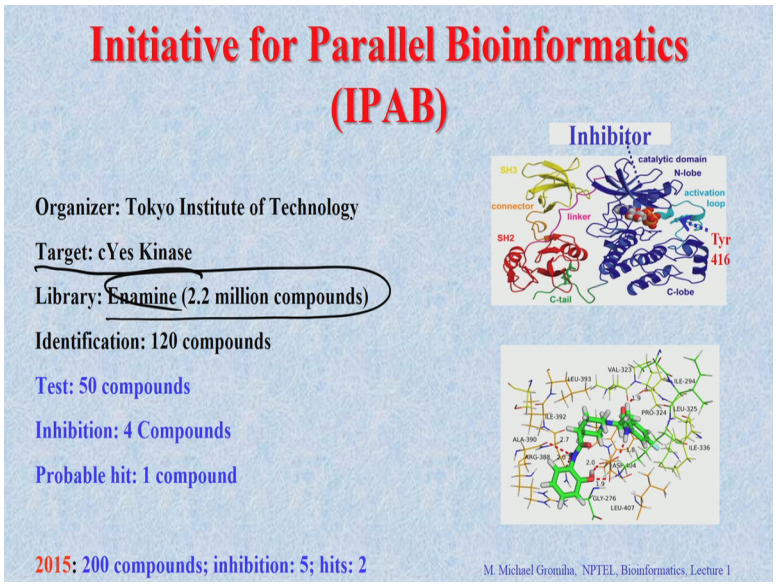
So, if you see the features and if you see the functions, and if you carry the relationship; here you have the features this side here you have the function, whether we can relate these features and function then what is the mathematical equation to characterize the

function in terms of features right. So, we can do a function right to understand the function from the features. These when you do these we can make the algorithm once algorithmic study, then we can use it for public in this case we use web servers or online applications right with the earlier days when the internet was not fast enough.

So, at that time see everyone they create the servers, they create their own algorithm they keep themselves it is difficult to transfer. But currently due to the advancements of these computers and biology and computational techniques, and fast internet facilities several observers have been developed to give the applications to the others right in our laboratory also we have developed various tools, which are widely used in the literature. So, many people use these databases and as well as the tools, try to understand any biological systems.

So, the fourth one, I will little bit discuss about the virtual screening. So, for example, in the case of drug design currently it is very popular, because there are a lot of small molecules are available in the literature and the people are affected with the several types of diseases like the cardiovascular diseases, cancer and so on and currently were developed with the Chikungunya, Dengue, and so on. So, in all these cases to identify drugs, they try to find a target and then we see what are the functions of the particular target, what are the actions, what are the important residues right and then we try to inhibit that activity so that we can reduce these disease.

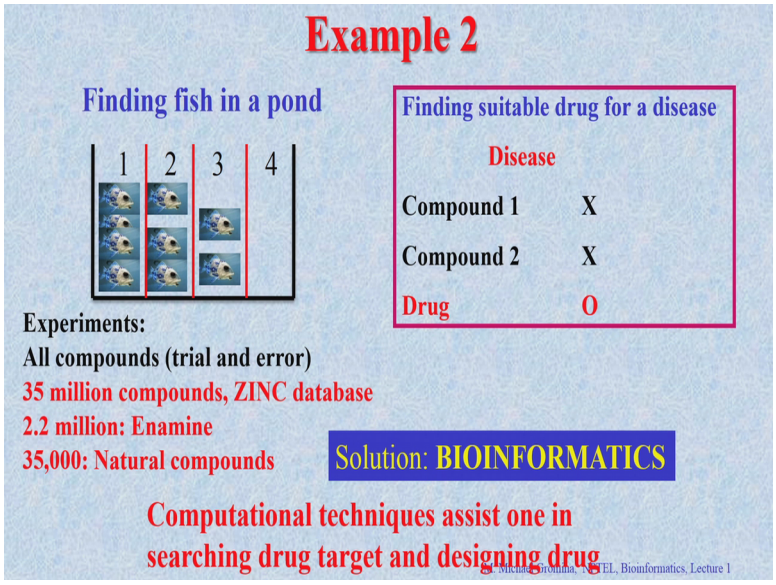
(Refer Slide Time: 10:51)



So, here is one example for the structure based drug design. So, if you have a protein. So, this is a target. So, here I show a target of c Yes Kinase, because these c Yes Kinases are very important for several cellular activities right.

So, there are several kinases one is the c Yes Kinase here this is very important for the colorectal cancer. So, this is an attractive target to define the inhibitor for the colorectal cancers. To do this there are various options how to derive a particular (Refer Time: 11:21) to be an inhibitor right. So, it is a very large pool, how to derive it. So, in this case I will tell one example.

(Refer Slide Time: 11:26)



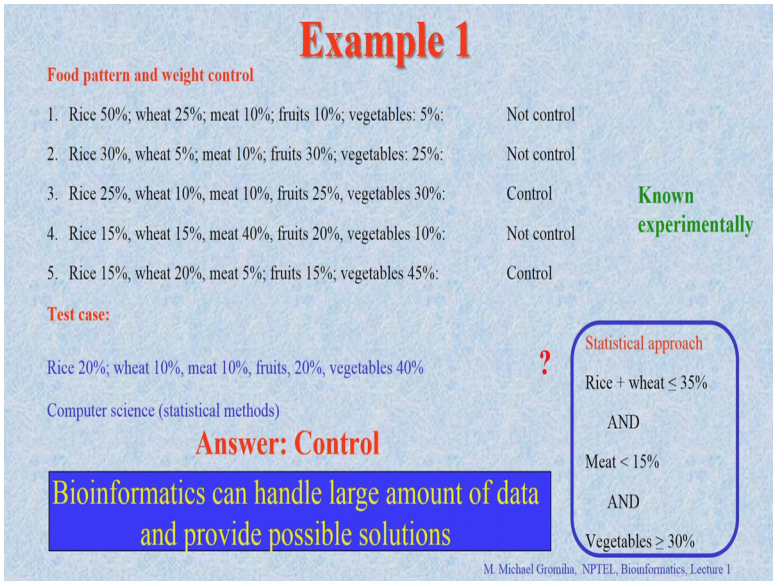
So, finding a fish in a pond; so if you see it is a pond here. So, can you see different fishes in a different ponds? So, if you want to catch a fish where will you put your net in the number one right number 1or number 2 if you put you will get a fish if you put your net in number 4. So, you will not get anything you only will spend much time you will not get anything right. So, if some of you will tell you that you are catching trying to catch a fish for long time. So, you try to use this one, to put that in a particular side. Then if you do it and if you get a fish then you will be very happy right because you do not have to waste your time.

So, this is the case, if for any disease. So, there are different compounds for example, compound 1, compound 2, compound 3 and so on if we take compound 1 and you try this is failed. So, it is not a drug and you have compound 2 this is also not a good drug

and then compound 3 is probably a drug right there are millions of compounds, if you look into the literature right. So, there are a lot of compounds for example, if you go to a zinc database, there are 35 million compounds and in the enamine database 2.2 million compounds and in the natural compounds in the Chinese medicine 35000 compounds. So, if you want to try one by one right when you try everything, then the patient will die at that time. Second case if you try to use each compounds experimentally, if will take long time it needs long manpower and also it needs it is lots of money.

So, in this case how to do it; so among these 35 million compounds if someone can reduce to 35000 compounds, then the number of experiments will be reduced by one 1000 times. So, if you instead of these 2.2 million compounds we have 100 compounds or 1000 compounds, then you can reduce enormously the search option how to do it. So, here is a solutions bioinformatics can do it because currently we have very fast computers and we have very good techniques. So, it can assist one to searching drug target and designing drug for many millions of compounds, and with that second one is how to use hypothesis? Here I show you an example.

(Refer Slide Time: 13:28)



So, here I have 5 of known values; so experimentally known. So, we take the example of number one the 0.1. So, rice 50 percent, wheat 25 percent, meat 10 percent, fruits 10 percent and vegetables 5 percent like. If you do this, then we can see that this is not controlled for example, take the food pattern and weight control and go over the second

one. So, rice 30 percent, wheat 5 percent, meat 10 percent fruits 30 percent and vegetable 25 percent in this case also it is not controlled.

And if you go for the third one rice 35 percent, wheat 10 percent, meat 10 percent, fruits 75 percent and vegetables thirty percent here it is controlled likewise the 5 examples. So, now, I have a test case this is a test case right to another one consumes 20 percent, rice 10 percent wheat fruits 20 percent meat 10 percent vegetable 40 percent right. So, now, the question is here it is controlled or not what is the answer is controlled right it is correct. So, why it is controlled, how do you know this is controlled can you tell one example.

Student: The fifth point vegetables are (Refer Time: 14:28).

Vegetables; vegetable here also are 25 percent, but it is.

Student: (Refer Time: 14:31).

This is not controlled, but here vegetables is less. So, we can derive some principles you can derive some equation series statistics, right. So, show one example right we can say that if answer is controlled you are right.

So we can see the right hand wheat is less than 35 percent and meat is less than 15 percent and vegetables more than 30 percent. So, likewise you can derive several equations right, you can properly study the initial data sets experimental data sets, from this experimental data sets you can derive some equations right some conditions where this will fit and apply these conditions to any set of data and then you can see whether that is controlled or not. So, likewise with bioinformatics can handle large amount of data and provide possible solutions. So, I will explain a little bit more about the virtual screening of the compounds, here is show one example how we use the virtual screening to understand the drug design. So, it is shown an example here is the protein right this is the c Yes Kinase in a protein.

So, there are different domains. So, we can see one domain left side that is a domain SH3 domain and SH2 domain and here this is the (Refer Time: 15:39) and the here is the catalytic domain and here is a phospho relation side the tyrosine four and six and this c lobe the question is this is very is very important for the colorectal cancer this is a target.

So, it is important to identify a probable hit target for this particular c Yes Kinase enzyme. So, this is one aspect here one side we have the protein.

So, you have the target to c Yes Kinase and the other side. So, how to design a inhibitor. So, they have a library of enamine library you have 2.2 million compounds; among the 2.2 million compounds how to choose? The probable compounds which can be a lead compound for a drug. So, and if I do it for a 2.2 million compounds, it can take long time it takes lot of money because it is compound cost of thirty to 40000 rupees right. So, it will do this spend time to do for all the 2.2 million compounds. So, how to do that? So, in this case you can derive some methodology, first you see whether the structure is known if the structure is known then you can use the particular structure. If the structure is not known then you may need to model this structure and then we have to stick for the activation sides, where are the activation sides they will again combine and see this pockets which are the binding sides right here is the side.

So, now we see 2.2 million compounds you check the features of all the compounds and make some conditions to fit with this particular a pocket right then you can use some molecular weight, you can use the hydrogen (Refer Time: 17:08) or the acceptors right various options you take and then you eliminate the compounds. Finally, you can use a virtual screening like docking you can do with these compounds and finally you derive some compounds.

So, in 2004 the Turkey Institute of Technology organized a competition, to identify a inhibitors for this particular target. We also contributed in that right we it is identified about 120 compounds, and they tested 50 compounds summing to 120 that and we showed that 4 compounds showed inhibition and one is the probable hit compound. They continued the same in the next year 2015 right there are about 2000 compounds right they found 5 showed inhibitions and to are hits.

So, in the down side of floor we can see this figure, I show they how they (Refer Time: 17:52) and interact with the protein you can see the green ones right. So, the green shows the (Refer Time: 17:55) and the surroundings ones are the protein side. So, these (Refer Time: 18:01) they have some specific interactions. You can see the hydrogen bonds or the hydrophobic interactions and the (Refer Time: 18:08) interactions and they because of these interactions they tightly binds with the protein and they act as an inhibitor for

these particular Kinase. In the later classes I will explain about the more details on the structure based drug design.

So, till now we discussed few aspects of bioinformatics. So, what are the asp different aspects we discussed? Databases the one is well organized databases bioinformatics contribute to organized databases right. So, then the second one computationally derived hypothesis when you have the data, then you can develop several function right.

So, to relate the features and the functions right and once we derived the hypothesis then we can develop several algorithms right for the prediction, and then once we predict then we can make it as online applications in the form of web servers, there are several web servers for example, protein structure prediction, protein function prediction right. So, how the DNA can bend, and how the DNA can interact with the proteins and so on. Then the fourth one we discussed now regarding the virtual screening of compounds how they do the screening for the drug development. And currently if you see the bioinformatics is widely applied in next generation sequence analysis. Now all are interested in personalized medicine few years ago it was very expensive to sequence it you know; now it is very cheap to get a sequence right.

So, everyone wants to see there (Refer Time: 19:24) you know and what are the proteins they have and what are the functions they will do, and to understand what are the probability of having any specific mutation to your protein and so on. So, in this case now currently we have a lot of data right from obtained from next generation sequences right for example, the illumine sequencing right. So, we can due to the advancements in the sequencing techniques, now there are a lot of data available in the literature. But the question is how to analyze the data. So, how to extract information from this specific sequences right.

So, there are several ways to get the sequences, they have a usage short reads and to get the final sequence. For example, if you some patients which are affected with a cancer or affected with the Parkinson disease and Alzheimer disease and so on; how they (Refer Time: 20:04) different from healthy individuals. So, they get the data for the patients and (Refer Time: 20:11) they get the full sequence and they get the data from the individual healthy individuals and they compare. So, how are the features, what are the variations or the mutations right where the mutations are in the protein coding regions or non coding

regions. And then they relate how these mutations are or these residues are important, why are they are involved in different pathways and how they are influencing the different diseases. They try to see the information and then they go with the treatment.

For example, if you are affected with a cancer or any specific diseases they are treated. So, they go to the hospital, they get the patients data as well as they get the information regarding drug and the drug response and they make a database you can do it right. So, from that information you can see that if in the specific variations, this specific drug will work. So, we have this information, then this will be helpful for the personalized medicine for different for different diseases.

Likewise which is bioinformatics plays a major role on different aspects in human health as well as for medicine.