Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. Bing Zhang
Department of Biosciences and Bioengineering
Baylor College of Medicine
Indian Institute of Technology, Bombay

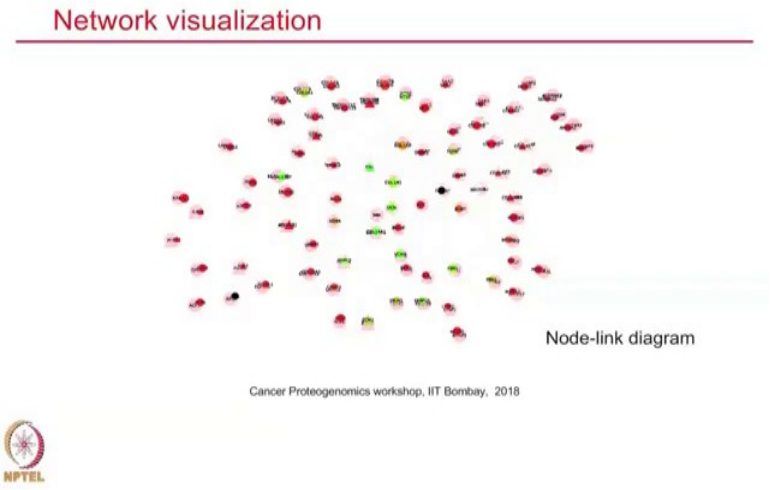
Lecture - 52
Network Analysis– II

Welcome to MOOC course on Introduction to Proteogenomics. In the last lecture by Dr. Bing Zhang you were introduced to the concept of Network Analysis with emphasis on protein-protein interactions. In today's lecture you will be explained the need for network visualization and the various resources available to make sense of the complex data. I am sure you appreciate that not only analysis, but also the data presentation and visualization is very crucial.

And, in this light Dr. Bing Zhang today's lecture is going to provide you good insight and opportunity to look at various tools available for data visualization. So, let us welcome Prof. Bing Zhang for today's lecture.

Next I am going to talk about I mean for our data analysis and what can we use this network to help us and one way is we use this to visualize for visualization and the other way is to do some data analysis.

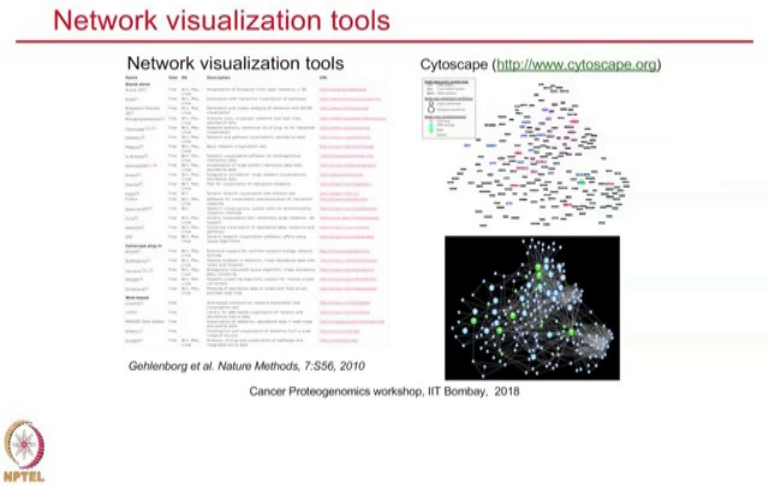
(Refer Slide Time: 01:25)



So, I will talk about the network visualization first. So, network visualization is usually through the node link diagram as we said I mean. So, each gene or proteins are node and they are connected to back edge this is called the link. So, it is called node link diagram and the beauty of this is not only you can see this and then you can overlay data to the network in order to understand your data right.

For example, I mean interesting a certain type of proteins in the network, let us say I am interested in transcription factors and then I can change the shape of the nodes to a triangle shape and then immediately I say ok. So, that is transcription factors and then I can overlay my differential expression result to the network and then we can say ok, these genes are up regulated and these are down regulated.

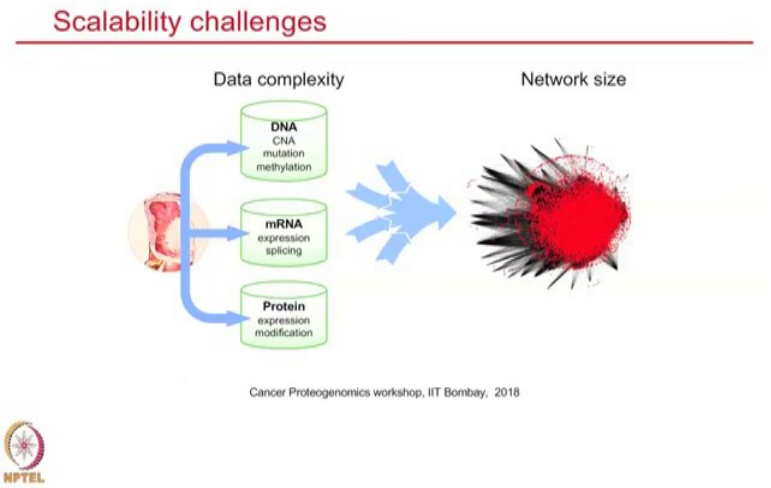
(Refer Slide Time: 02:15)



So, and they are tools that have been developed to facilitate the network based both the network visualization and the network based the data visualization. And, this review article in nature measures summarizes a lot of tools that you can use to visualize biological networks. And, this is the most popular one is the cytoscape it is very easy to use and very powerful.

This is one network I did in earlier publication and if you can also put some artistic, I mean change to the network. I mean actually this was also done in the cytoscape and after I did this and this was became a cover of that issue of the journal. So, within the cytoscape you can actually do a lot of questions to visualize your network and your data.

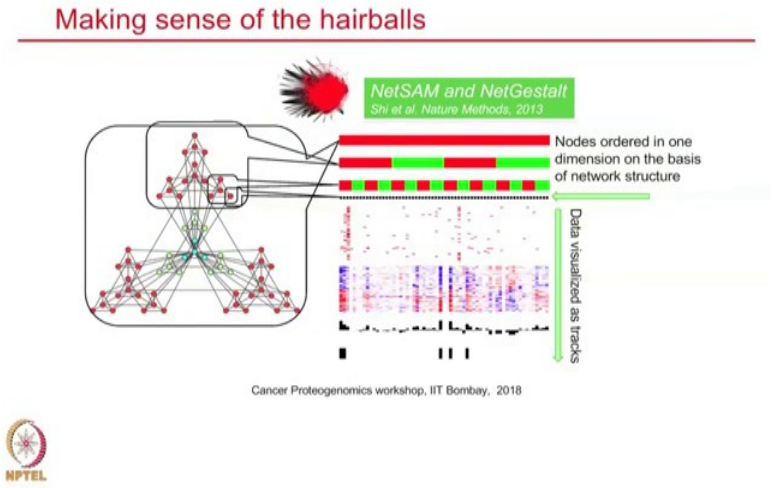
(Refer Slide Time: 03:17)



But, there is also challenge I mean although the cytoscape is very good for small networks, but when you have the whole protein-protein interaction network you want to know where is I mean denote and maybe a hundred thousand edges. It is very difficult to say anything; it becomes a hair ball again right.

And, also if you want to overlay a lot of different types of data like in multi-omics study, we want to overlay proteomics, genes expression, copy number all this data to the network it becomes challenging. I mean you can only change the shape, color, but I mean the lot size, but they are lot a lot you can do.

(Refer Slide Time: 03:58)

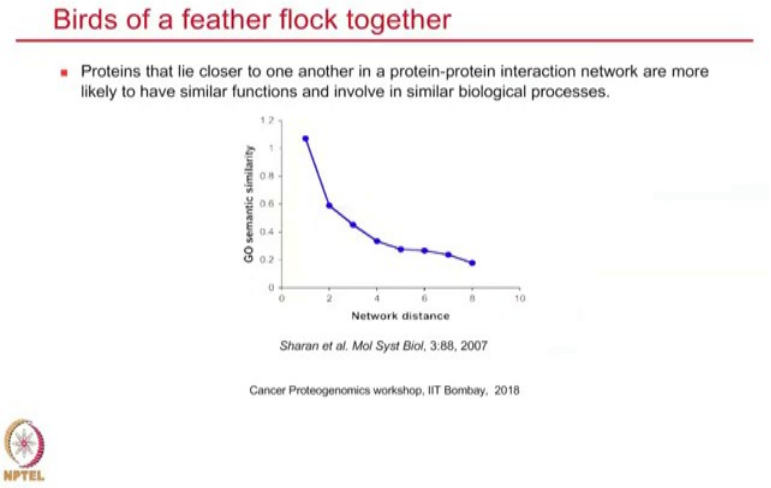


So, one thing we did was to use the hierarchical modular organization property, we learned from the biological network and then try to use that property. So, if we still can use this bar who represents the whole network and then we can use a smaller bars to represent each of these sub networks. And, then we can each of these sub networks can be further organized at smaller modules and eventually they have each of the nodes located in under these modules.

So, as you can see this is because the way we can do this is because, the network is organized in a hierarchical modular way right; that is what we learn from the network property. The beauty of that is now instead of using two dimensions to visualize the network, we only use one dimension and this way we can visualize the data and in the second dimension is on the paper or in the browser.

So, I will not go into the detail, but we have the method to convert a network into this; I mean modular structure hierarchical modular structure. And then to also a tool, but NetGestalt that you can use to explore the data next the TCGA data under the this framework.

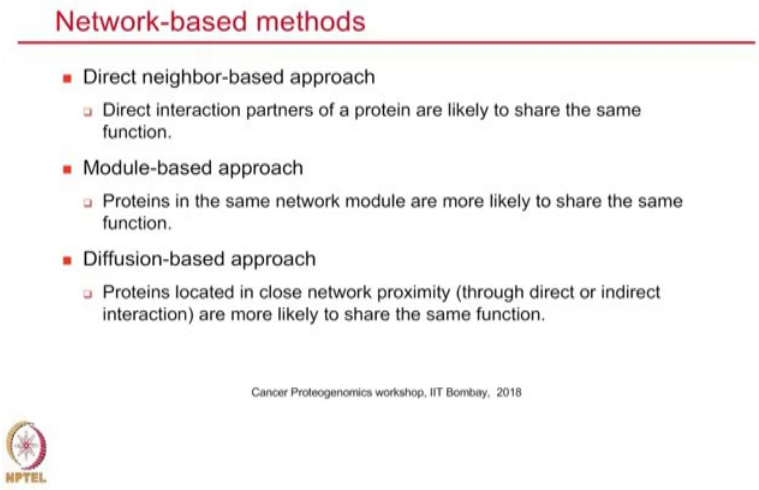
(Refer Slide Time: 05:31)



And finally, I want to talk about I mean we talk about the network visualization network based data visualization right. So, at the end I want to talk about how we can use this network to help analyze our own data. So, this is personally based on the observation that the nodes in the network are not random again not randomly connected, usually genes or proteins that are functionally similar to each other are more likely to be connected to each other in the network.

As you can see in this plot like if we have a way to quantify the functional similarity between two proteins and then we can see the protein pairs, that are directly connected to each other. Meaning they are one step from each other they have much higher average functional similarity than the protein pairs that have shortest path lengths of 8 or above, as you can see this relationship is very obvious.

(Refer Slide Time: 06:42)

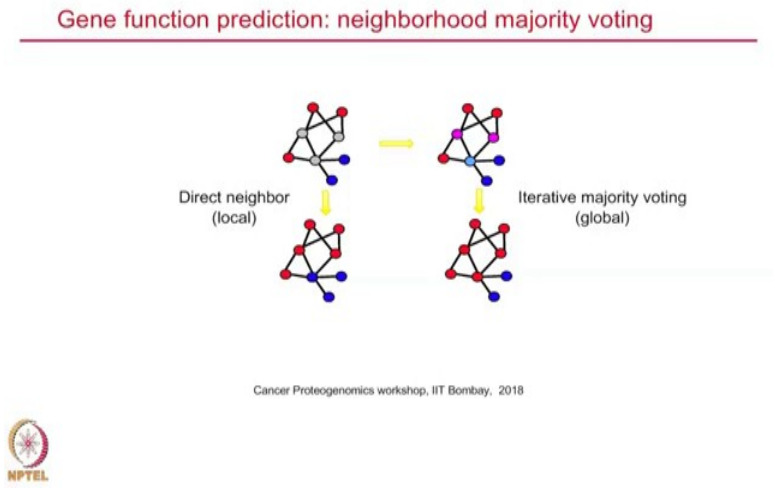


So, by leveraging this observation so, we can come up with a few different ways help us to either predicting gene functioning or prioritizing genes in our studies. So, first we can use a because we know that direct labels are likely to share the same function. They can explore the directly interaction partners of the proteins in the network and then the second approach is we can divide use some graph algorithms to try to separate the networks into modules.

We know they are modules in the network right. And, then we expect that the proteins in the same module were likely to share a similar structure or share similar function and the last method is called the diffusion based approach. So, I think this is a very local method that you only focus on the direct relationship.

This is a relatively more global method you explore the module, but the diffusion based approach basically try to explore the whole network as a whole system; I think this is probably a more powerful approach. So, we can I can show you a few examples to help you to understand this approach.

(Refer Slide Time: 08:03)



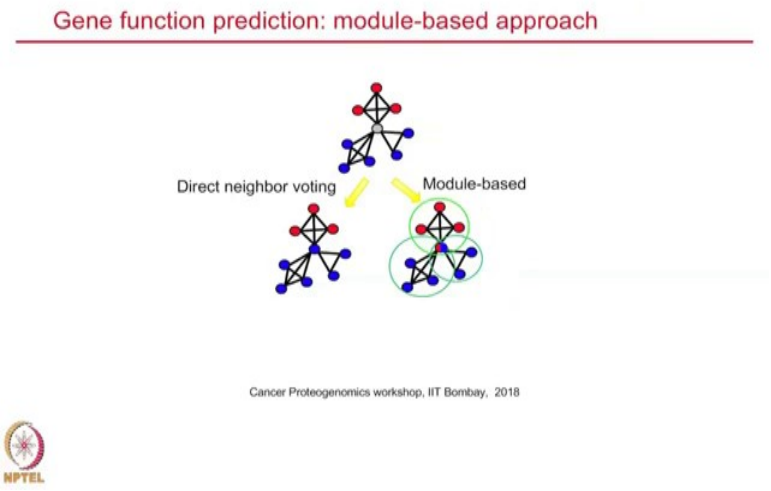
Let us say if way this is a small protein-protein interaction network let us say and then we know the red proteins meaning protein have function a and then the pro proteins did that the protein associated with function b. And, then we have other proteins in the network that we do not know what they do in the network. And, then we try to use this network to help us to predict what are the functions of this network do they likely to have the red function or have the blue function and the if we do direct neighborhood analysis.

So, basically for each node we can count how many red labels it has and how many blue labels it has and through this we can assign the function through a majority voting algorithm. So, basically if there are more red labels then it is red and the if there are more blue label labels then it is blue. So, this is very easy to implement and you can quickly guess a function of the proteins.

But, one limitation of this approach is that I mean because the this three proteins we basically have no idea about their functions at the beginning. But, maybe for some proteins like this guy we are pretty sure it is more likely to be a red protein, but for others it is less clear right. So, you can also think about doing this in an iterative way. So, let us say you do you have a intermediate step here, you make a temporary assignment. For example, this is two pink meaning it is more like to need to be red and this is a light blue meaning is more likely to be blue than red.

But, after that you read after this you read for the green proteins you recount to the recounting and in this case we can see this protein care to learn to red function rather than a blue function. So, it so this iterative process can better leverage more information than just using the original neighborhood analysis.

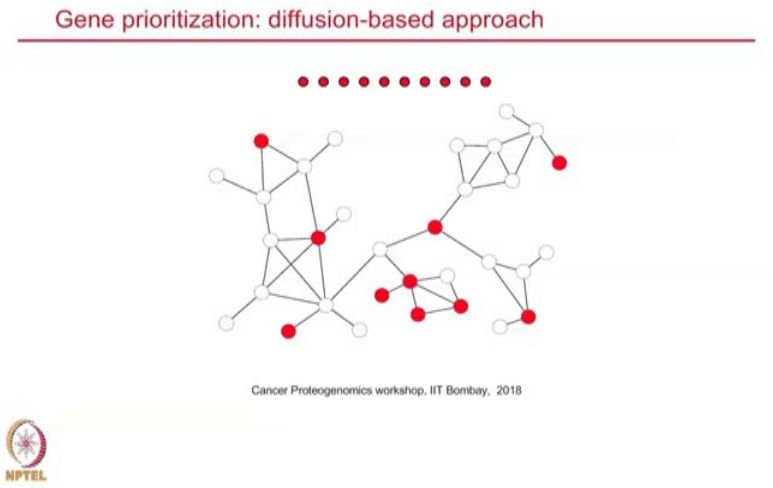
(Refer Slide Time: 10:13)



And, the we can also use a module-based approach to do this for example, in this case we only have one protein with unknown function and then we want to assign a function here. And, if we were to the neighborhood counting and then you would say 1 2 3 4 here and 1 2 3 here you were saying think maybe this protein is a blue protein. But, if you do a module based approach and especially if you have a method that can allow the module to have overlapping members.

And, then you can have a module like this 1 2 3 modules and then this pathway is a module dominated by red proteins. And, this is a module dominated by blue proteins then you can probably think this protein might have both functions, if we use this. And, that is more likely to be true right a proteins a lot of times may have multiple functions depending on what it interact with in a specific condition.

(Refer Slide Time: 11:15)



So, and diffusion based approach, this is particularly used for in gene prioritization; let us say if you do a high throughput study. And, then you do lets say a GWAS study you may come up with multiple SNPs or SNPs associated genes that are associated with the phenotype. Or, if you do differential proteomic analysis you come up with 20 proteins or 10 proteins that are very likely to be associated with the phenotype.

And, then you want to do experiments next step right, but which protein to choose to do the knockout experiment; if you have a 100 candidates it is very difficult to know. So, the network-based approach can help us to prioritize. So, let us say these are the candidates we have and then we can map them to the network and after you map to the network you can use a process called random walk process. So, basically we can imagine an each node is a person, let us say I am a person.

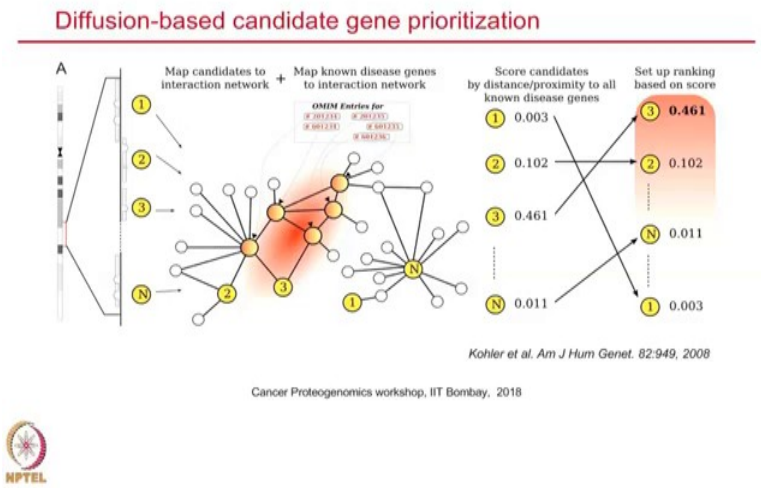
So, I start from here from the red node and then if you just take random walk at each step to the next node. For example, at the first step I can go either here or here or here or here right, but after you go here I have the next step, I can go here or here or here or here or here. So, but you can use some iterative updating process and then you at the steady state they should end up, they have the opportunity to end up somewhere.

So, let us say if you start from here then at the steady state I may have higher probability to end here or here or here right. And, then we can calculate if we start from count from start from all the each of these 10 nodes, but the probability of ending on a particular

protein. And, then you sum that up and then you get the steady state probability for ending at that protein and then I use color shade to indicate the probability.

And, as we can say here if we start with this 10 possible positions, well we are likely to end in this area, but it is also possible to end in these areas. So, anyway I mean for the 10 proteins you can probably say ok, this 4 proteins are more important than the other proteins. And, also this might sometimes help you to identify new genes in not included especially in a say in proteomics; we have a lot of missing identifications right. For example, this might be a low abundant protein and now you say maybe this is I mean somewhat possibly very important protein.

(Refer Slide Time: 14:08)

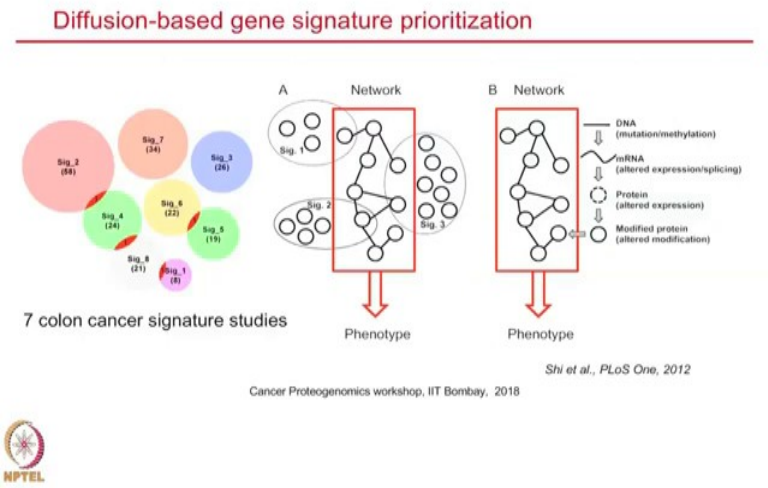


So, let us see some real world application of this method. So, in this study I through a GWAS study lets say you identified a lot of genes or the SNPs associated with these genes, that have protein you need to be important or candidate genes for disease. But, let us say you also have prior knowledge on which genes you already know to be associated with disease.

Now, you have a protein-protein interaction network and then you can map all the known genes and the new genes in the network. And, then if we go through the diffusion process we can estimate, if we just start from this node the known this related genes and then what is the probability of ending and this proteins. And, then we get a new score for these proteins and the based on this process, we can rank I mean you would expect the

genes 3. Well, I were likely to be disease related protein than this gene 1 right; so, this is very easy to understand.

(Refer Slide Time: 15:23)



And, the also we used this is for in a study for gene signature prior prioritization. So, I have a worked based another group. So, basically to try to develop gene expression signatures for colon cancer, but we are just one group to do this study. And, this are many group because, this is an important question colon cancer gene expression signature and the 7 published the studies on this topic.

But, if we look at the gene signatures reported by these 7 studies; they do not overlap actually you see where neither overlap about of the gene signatures reported by their studies. So, then we was thinking I mean what is causing this discrepancy is it because just the all these are incorrect identifications or there are some possible other explanations.

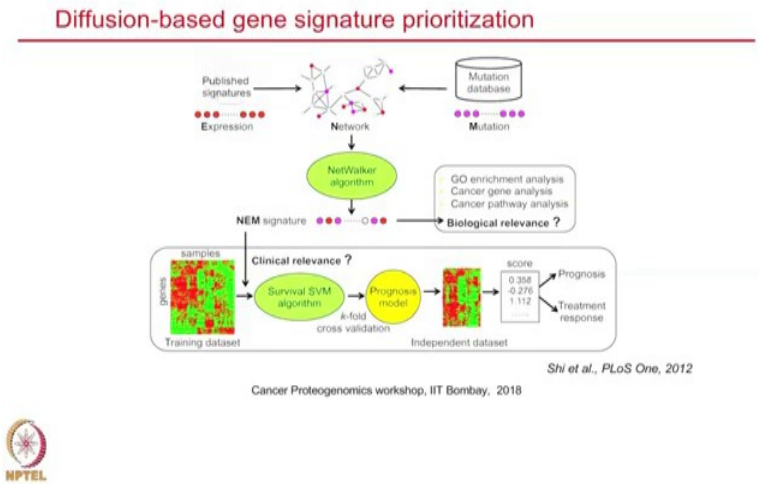
So, one way to think about this is maybe if this is the network and each study may and this is the network that is actually driving colon cancer poor prognosis and each study may identify some of the important component in the network. But, maybe they did not get all the important nodes, but also they observe some other proteins that are not critical to this network, but they just co-vary with those nodes.

So, for example, gene signature 1 may only identify this and gene signature 2 only identify this. So, if you only to the overlapping they have no overlap, but if you map all those signatures to the network, we will be able to identify maybe it is this region that is important. And, the similar idea is I mean not only this is only for mRNA based gene signature study right.

But, a protein activity can be ordered at multiple levels, at the DNA level the for example, this protein it can be the activity of the protein can be altered by mutations or copy number alterations or mRNA this the expression change or protein expression change or PTM modifications.

All these can put in alter the protein activity and the if protein at protein activity in this important network is altered you know as sample, then you are going to see potentially a phenotype right. So, that means, if we have multi-OMICS studies we can also map all the observations to the network, that can also help us to prioritize the findings from those studies.

(Refer Slide Time: 18:08)



One example is a way again collected all the gene precious signatures from the 7 studies and also we collected all the mutations in colon cancer. And, then we mapped those to a protein-protein interaction network and then we have through the network algorithm. So, basically it is a network diffusing algorithm, we talk about and then we get a new list of proteins that I mean are important not only because of the differential expression

mutation, but also because of the kind of centralized the locations certain part of the important part of the network.

And, then we were able to come up with a prediction models based on those signatures and then we were able to show that the signature you get this where he has better reproducibility, when you apply to a new cohort then the I mean gene expression signatures you study I mean from individual studies.

Student: Dr. Zhang most of the network analysis whatever to it we are getting the individual for the networks is being taken from for the experimental and computational and all these tools to it. So, most of the many a times these interactions are transient interactions. The interactions between proteins they are not permanent interactions

That is right. That is right

Student: So, how robust are the computational tools to pick them out , because there are chances you may miss them out in experiments how well represented are these databases as far as these interactions are concerned.

That is a very good question. So, if you go to about biogrid for example, you download all the protein-protein interaction in the database and that is basically the collection of all the possible interactions that has ever been reported in any of the publications. It could be in a diseased state, it could be in after EGFR treatment or if it is not condition dependent. So, basically you get is a map, it is just like Google map with everything on the map, but you do not have context dependent information.

So, there are few different ways to address this for example, you can try to build your own condition or specific networks through experiments, you can do pull down or like you have to do pull down in the specific condition. Or, you can also try to manage some gene co-expression information for example, if you are interested in colon cancer. You can take a look at the colon cancer co-expression and try to overlay or integrate the co expression information in that condition with a big protein-protein interaction network.

And, try to get some conditional specific network that is actually a very active research area and the people are trying to develop algorithms to come up with context specific interaction networks rather than just the this global. But, I think that provides still

provides you a reference map that you can to start to derive condition specific networks, but that is a very important question.

Student: But, can we predict the interaction from the basis of a primary sequence.

Well, I think of course, you can do the prediction based on the sequence I mean that is for example, one approach I talked about that domain based approach. And, basically if you say to I mean and I think people are start to use deep learning actually some people you might have doing this type of analysis. They try to leverage the deep learning and then when you have enough training models and then you can use that as a training examples.

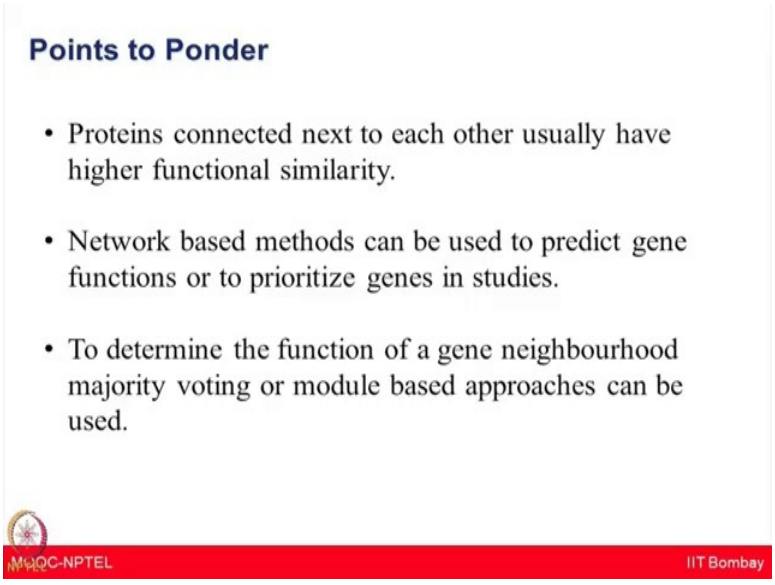
And, then to say how can we use deep learning or typical machine learning approach to capture those different features, that can help you to predict. Yes, I mean the sequences are also important in predicting.

Student: So, both primary and 3D structures are useful

Exactly yes.

Of course, I think people are and there are some recent deep learning based methods, they try to incorporate the, I mean the linear sequence. And, also the protein structure information and some domain information as features to predict.

(Refer Slide Time: 22:34)



In today's lecture we learnt about the importance of network visualization and interpretation of complex data. We also learned that proteins connected next to each other usually have a higher functional similarity. This forms the basis for network analysis tools. If the aim is to predict gene function, the network based methods like module based approach and neighborhood majority voting can be used to get important leads.

The diffusion based approach is used to get information on gene priority in a network and network visualization tools like cytoscape and NetGestalt are widely used in clinical studies. I hope these two lectures and various tools which Dr. Bing Zhang showed have been helpful to you to now try use your own data set or publicly available data set and try to create the various networks and visualize them using these tools.

Thank you.