Introduction to Proteogenomics

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

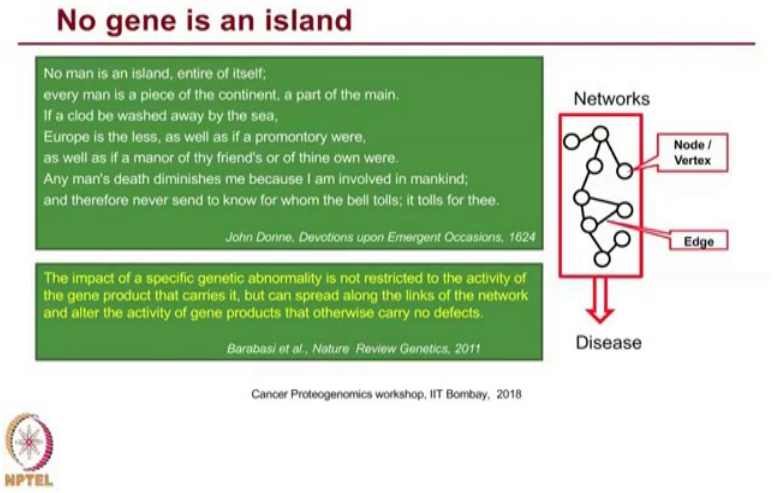
Lecture - 51
Network Analysis – 1

Welcome to MOOC course on Introduction to Proteogenomics. It is well known fact by now, that most diseases are caused due to dysregulation of pathways or networks and they are not caused just because of effect of a single gene. It could be just in very few rare cases; but otherwise it is a group of genes or proteins or a pathway is actually going to affect a given disease.

Hence it is important to understand, how proteins interact with each other. In today's lecture Dr. Bing Zhang will introduce you to the concept of Network Analysis and various tools available for it is use. I hope this information will be very helpful for your own projects when you are looking at how proteins interact with other protein and form a given network. So, let us welcome Dr Bing Zhang for today's lecture.

So, in this lecture we are going to talk about Biological Network Analysis. So, I think we had plenty of science and the formulas and things like that . So, I will start today's lecture with a poem, actually it was a very beautiful poem from Poet John Donne and this was written, I mean 400 years ago.

(Refer Slide Time: 01:43)



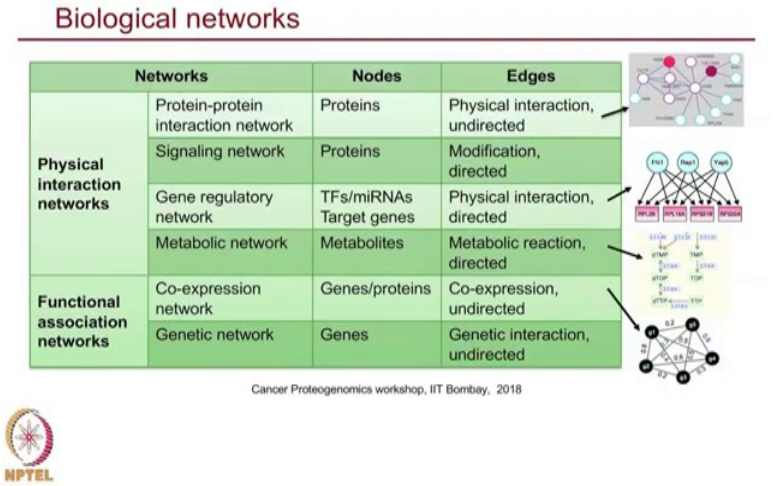
It is titled; No man is an island and I do not want to repeat this, but it is a very beautiful piece of work. I mean basically express the idea that in the holistic view of the societies there are no mankind survive or strive without support from the community. So, I think it is still true all those 400 years has passed, it is still true in today's society. And you can think about, I mean people around you; for example, through connections where is WhatsApp, I think that is the one you guys use in India; you have a connections through that social network, right.

And so, a few years ago when I was reading a review article in this nature review genetics from the Albert Barabasi and the so, basically this sentence reminded me about this poem and the idea is that; if there is a change in a gene in the network, it will have the impact not only to the gene itself, but the impact will pass through the whole network. So, I was thinking, ok. So, it is not only, no man is an island; actually no gene is an island. So, that is a same we are going to talk about today. So, I mean we need to understand, biology understand the genes in the context of the networks.

So, indeed during the past decade also and a lot of studies have shown that complex phenotypes including most of the disease phenotypes are the result of the dysregulation of the networks, rather than individual genes like. Network dysregulation cause disease, not individual gene cause disease. And in order to understand the network, I would first introduce two terms one is called a node and the vertex or vertex.

So, those are basically the individual elements are in the network and they are connected by edges.

(Refer Slide Time: 03:49)



So, with these two terms then, we can look at some typical biological networks where you have; they can be divided into two categories the physical interaction networks. So, these are the types of networks, the nodes are connected to genes or proteins that physically interact with each other. For example, the protein-protein interaction network and in this network each node is a protein and the edge represents the interaction between two proteins.

And the signaling networks, its a kind of a specific type of protein-protein interaction network; I mean it is not only the interaction of the proteins, but the protein A can kind of regulate protein B, right. For example, a kinase as we just talked about kinase could regulate the downstream target. So, in this way you can imagine or networking exists, but the edges are directly ; for example, this is the kinase, this is a target this is called undirected protein-protein interaction network.

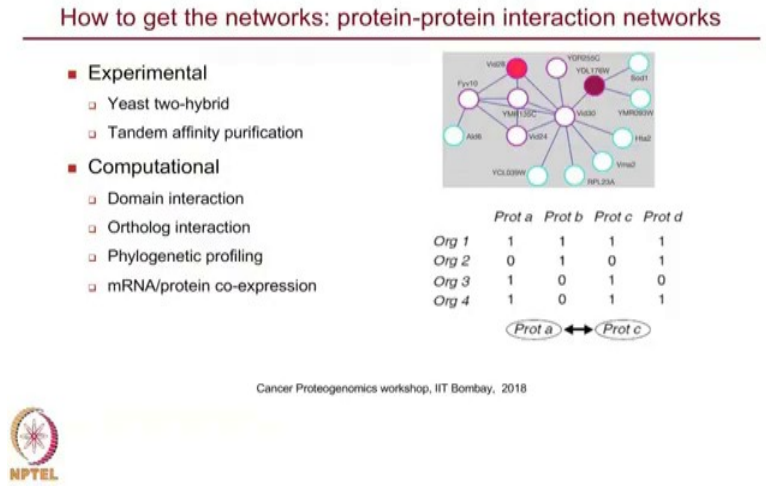
But, if in signaling networks, it is this protein has a can modified other proteins. So, there is a direction associated with the edge. And also the gene regulatory networks, in this networks the nodesare either the transcription factors or it can also be the micronized and their target genes, is also a directive network. So, it represents the in the physical interaction between the TFs or miRNAs and their targets. And then metabolic networks, in this one the nodes are the metabolites and the edges are the reaction going from the substrate to the product, also it is a directed network.

And another type of network is called functional association networks. In this type of network, I mean we do not really know whether the two genes in the network interact with each other or not, they may or may not interact with each other. For example, the co-expression network, if we have a; I mean, a lot of experiments; am I obviously, through genes keep going up and down together, we can infer the expression relationship between them.

And that can actually be quantified for example, based on the Pearson correlation. And in this way, I mean this we can also call this a weighted network; meaning the edges can be weighted by the co expression level alright and the, it is also undirected. And the genetic network and other than that those are the genes and the relationship indicates the genetic interaction, meaning you do some perturbation experiment. If you knock out the gene A or gene B you gets a same phenotype, you can indict guess maybe there is some relationship between the genes.

So, the first question I want to talk about is, how we get all those networks; because we talk about the network, we need to get the network first in order to do something random, right.

(Refer Slide Time: 06:53)



So, first we talk about the protein-protein interaction network. And in order to get protein-protein interaction networks, basically we want to establish the relationships or interaction relationship between two proteins, as you can see here right. So, the experimental approach that can help us to get this type of relationships, including the yeast two hybrid experiment or the pull down experiment, pull down followed by the mass spec analysis.

And they are also computational approaches that can help us in further protein-protein interaction relationships. For example, we can start from the known protein-protein interactions and then we can try to infer, I mean which domains actually interact with each other. And then we can generalize to do protein pairs, if they have those interacting domains, we can case maybe they interact way each other. And then we can, there are lot of studies in the model organisms; we can also through the ortholog relationship, we can also map those to human and then for example, guess the interaction relationship in human.

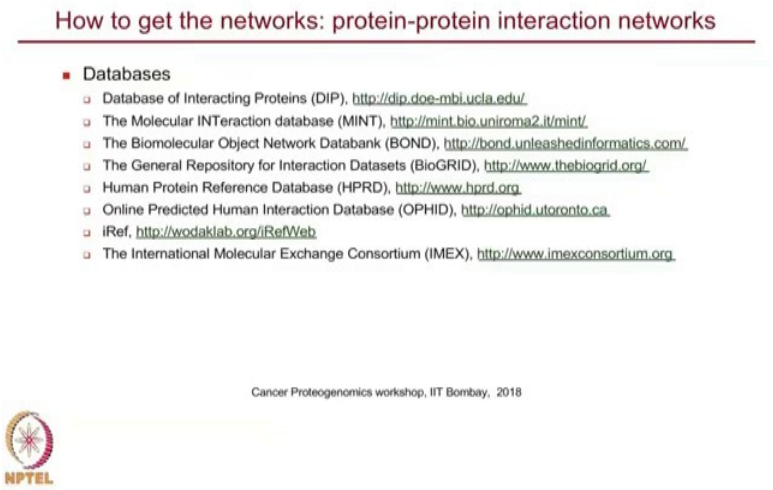
And we can also do phylogenetic profiling. So, meaning and you have a lot of proteins in each organism right, and then you look at the existence whether this protein exist in organism a and b or it is that alright. Then after you do this for a lot of hundreds of organisms you will be able to see some proteins tend to co-occur together; for example,

in this map, in this table I mean can you tell me which two proteins are more likely to be interact with each other than the other proteins pairs. Is that B and C.

Student: b and a organism; protein a and protein c.

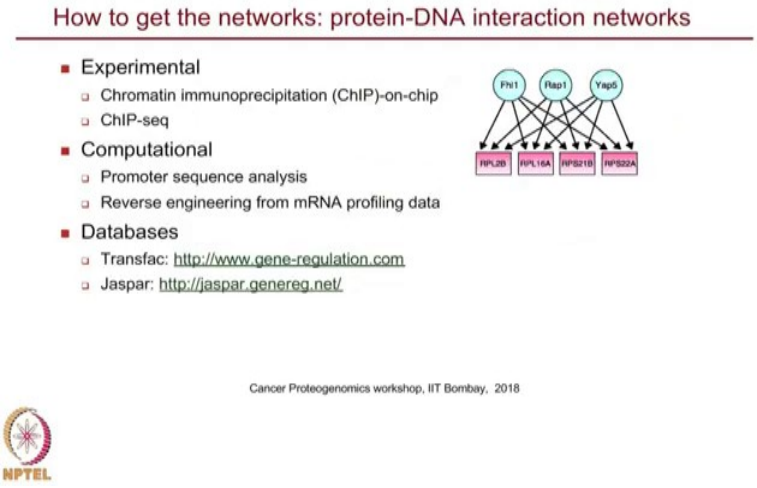
Yeah exactly, a and c; because always appear together, right. So, if you need two protein to interact, they have to coexist in that organism in order to interact. And similarly and through the gene expression or protein expression experiments, but if they say two proteins always come up together and then they come infer the interaction relationship; but of course, those are computational approaches just help us to make inference needs to be validated.

(Refer Slide Time: 09:22)



Yeah and the another way is, I mean you do not want to do experiments and you do not know how to do the computational inference; but there are plenty of protein-protein interaction databases that you can use to download those information and here I use the I mean, quite a few, I mean databases that you can, I do not want to go through them individually; but it is in the handout and you can get those, get to know those resources by yourself after the class.

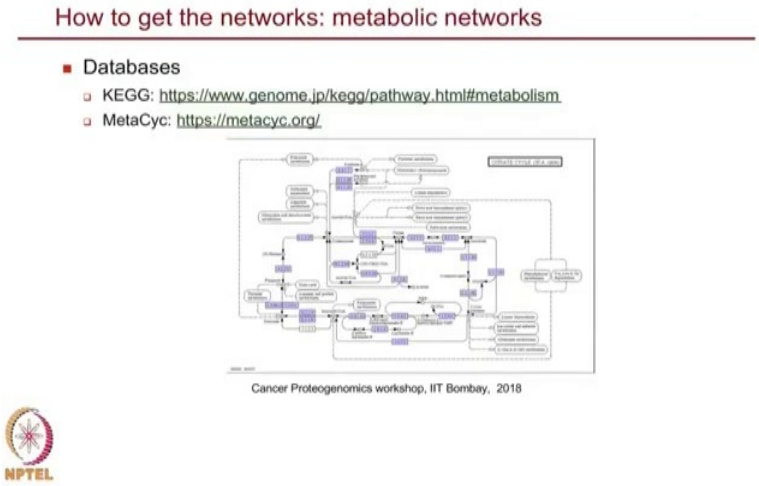
(Refer Slide Time: 09:52)



And for the protein DNA interaction; so, basically we want to establish the relationship between the transcription factors and the target genes. And the experimental approach includes chip which is early version of the studied; now it is people are you only doing ChIPs-seq. And the computational approach and we can do some promoter sequence analysis through the motif analysis or we can do reverse engineering from mRNA profiling data and also there are databases that we can get this type of information from the Transfac.

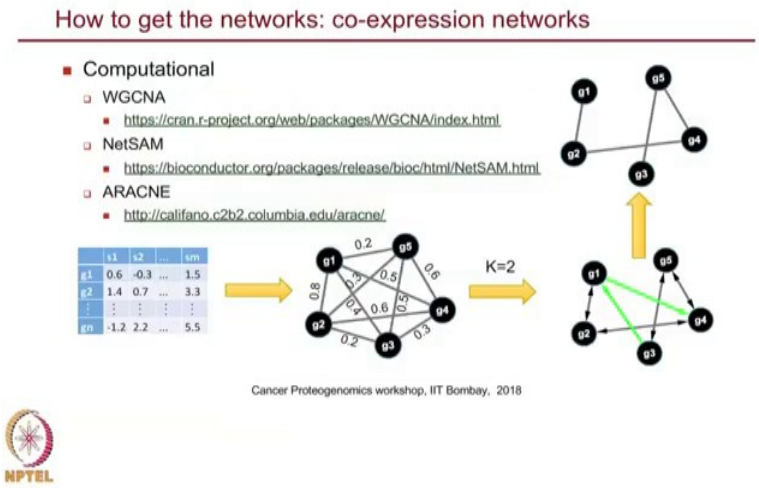
I think this is now commercialized, but the Jaspar is the open source resource that you can use.

(Refer Slide Time: 10:38)



And the metabolic networks those are the networks that have been very well studied for long time and very well established database for this type of networks. And the two commonly used ones includes the KEGG and Meta Cyc, did two well use the metabolic network or pathway databases.

(Refer Slide Time: 10:58)



And for the co expression network which is also typically used. I mean this is mostly from the computational analysis, you start with the gene expression or protein expression matrix; each row is a gene, each column is the sample, then you can use a one of this

method to do co expression network inference. For example, the WGCNA; what you do is to, you calculate for each pair of genes, you calculate the co expression relationship and then for examples through the Pearson correlation and then you get a score and you get way to the network. And then what they did was to raise correlation through a certain power, to further discriminate the inter highly correlated ones from the lowly correlated ones.

And the netSAM package we developed; I mean so, basically try to convert this weighted network into some unweighted networks; because there are certain graph algorithms that we can use, can apply to the unweighted network, but cannot easily apply to the weighted network. In this case I mean, we can think about a K nearest neighbor approach.

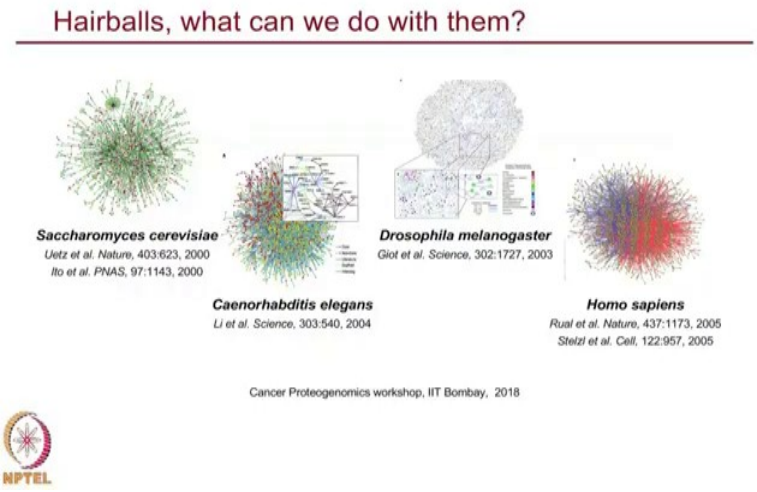
So, basically for each of the node in the network, we can ask what are the I mean let us say, we are talking about a two nearest neighbor network. And we are asking in this network; what are my two best friends and then we get for each gene, we will vote for the two best friends for each gene.

And then when you think someone is your best friend, the other guy may not think the same right and then we will also remove those relationships, we only keep the ones that are mutual. I mean you think I am your best two friends and you also think so; so that means, this can give you from this network to a very robust relationship based, but it is unweighted network now.

And so, ARACNE is also another very popular tool to use two beautiful expression network rather than using Pearson correlation or Spearman correlation that deriving the relationship between two genes based on mutual information. So, that is also a good idea, because I mean it is the mutual information can capture different types of relationships, not only the linear relationship or monotonic relationship are there more types of relationships that can be captured by the mutual information.

So, let us see, you one goes through all these and then you were able to build a network, right.

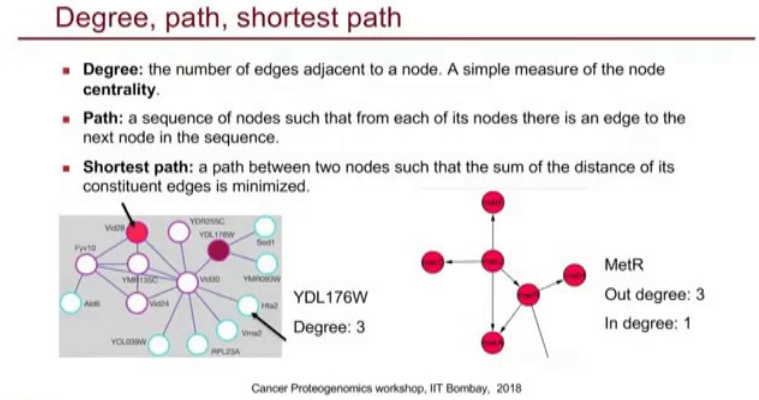
(Refer Slide Time: 13:30)



This was what happened like maybe, early 2000 there are a lot of experiments that have been done and the people start to build the protein-protein interaction networks in from the early time for the yeast, and then also the human protein-protein interaction networks three experiments have been published.

At the very beginning people get very excited all this looks great and we get a lot of information; but then if you look at this, I mean people start to realize this is just the hairballs right, I mean it is beautiful to look in a way, but what can we do come from this. So, then the next question is, if you have the network what can you learn from the network.

(Refer Slide Time: 14:19)



And in order to do that maybe I will introduce a few more terms, in order to better understand this networks; the first is the degree, the degree means the number of links or edges each node has. For example, we can look at this node, this purple node in this network and basically it has three links and it has degree of three, this is for undirected network. But if it is within a directed network, each node have incoming degrees and also outcoming degrees; and for example, this Met R this gene it has out degree of 1, 2, 3; but in degree is 1.

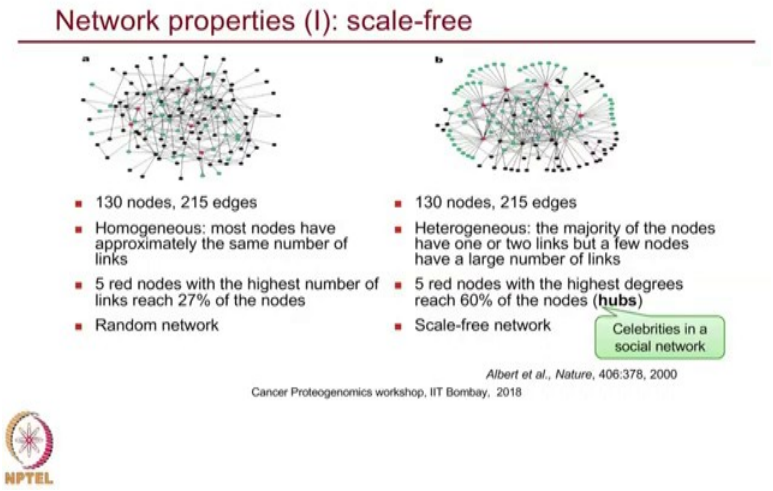
So, degrees are very simple, but important measurement for the nodes or genes in the network, because if you look at this, you will think this is kind of in most interpretation of the network, if it has higher degree, right. So, this it is a very simple indication of the centrality of each node in the network.

And the second thing we want to talk about is a path. So, that is how it is not to explore the relationship between two nodes in the network, right. For any pair of nodes you pick and you can find paths that link these two nodes; for example, here and we can find this path which include two edges and then we can also go 1, 2, 3 here or we go 1 2 3 4 5 here, right. So, that a lot of different paths that you can find in your network and this is the one with the shortest the total length of the path.

For example, from this node to this node, I mean I think this a path with two links is a shortest path this is called the short path between them. So, with the understanding of the

degree and the path then we can start to explore some of the property of these networks; I think the after we created network, the first thing we want to do is to understand; I mean, the how the network is organized, then what are the characteristics we can learn from those networks.

(Refer Slide Time: 16:27)



The first one I want to yeah, is the first property we want to talk about; so, maybe I will give to you this example, I mean these are two networks, each of them have 130 nodes and 215 edges. So, the number of nodes and the edges are the same; but if you look at these two networks, there are not verse that actually quite different, right. So, one thing is that I mean this network is more homogeneous; meaning every node is very similar to each other. So, basically they have the same number of connections.

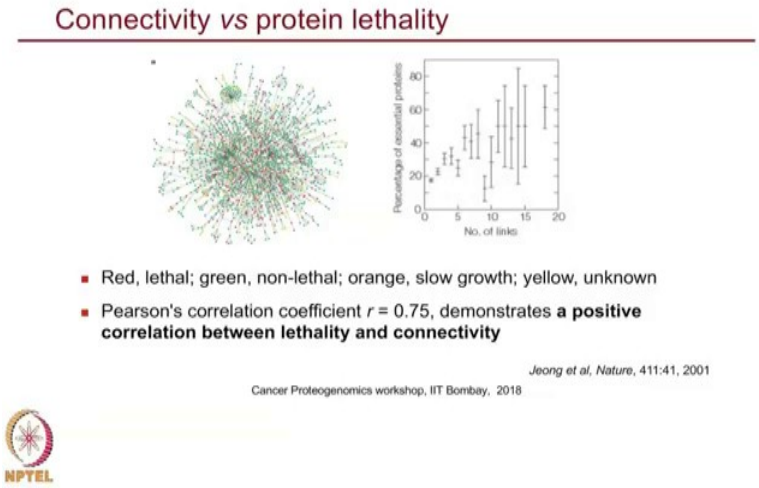
The 5 red nodes basically with the highest number of links only reach 27 percent of theother nodes. But if you look here, and the nodes are very heterogeneous; meaning some of them have a lot of links, but most of them only have one or two links. So, in this network the 5 red nodes with the highest earnings that they can reach 60 percent of the other nodes in the network. So, this is I; that means, the nodes are very similar in this network and the nodes are not similar in this network.

And in reality most of the real life networks like social networks or the even biological networks, they usually have this organization rather than this one. And the Barabasi who is the kind of very important person in network analysis, he named this network scale

free network and this is basically a random network; if you randomly connect the nodes this is what you get. But, the real life network are not like this, they are born exist; this is called the scale free network. So, this can be probably best understood in social networks and we can immediately understand the, what are the hubs in the network, right.

Of course those are the people that are celebrities like the stars, and for example, if you look at the social network, they of course, have a lot of connections and the for us guys I mean maybe we do not know too much too many people and they only have a few connections. So, the hubs are the celebrities in the social network; but biological networks are also have this scale free organization. So, of course, and people interesting to know, I mean what are the hubs in the biological networks do they play a different role in the biological processes than the other nodes in the network.

(Refer Slide Time: 18:57)

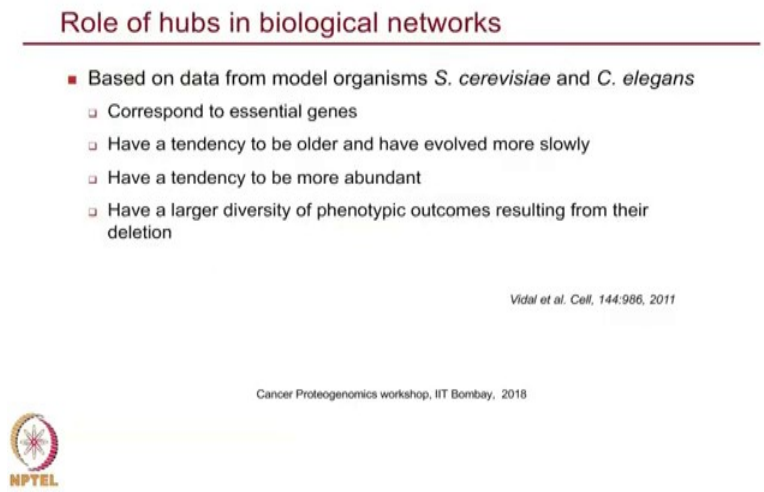


So, in early 2000, like 2001 there was an interesting study. So, at that time and they were I mean, genetic study in yeast; basically try to delete each individual protein in the yeast proteome and try to see the impact of those proteins diminishing of those proteins. And they found some of these red proteins had an lesser impact meaning, if you delete that protein, the cell will die and some of them, I mean does not cause much impact or some of them only cause slow growth, but at that year I think the protein interaction network of the yeast have also been published.

So, this group tried to combine this two types of data and see and now I group the nodes in the network in based on the number of links or edge or degree they have in the network, and here is nodes with only one link and here is a node with twenty links. So, basically you group them based on the number of links and then they look at that the percentage of the nodes in that category that has a lesser impact after when the proteins deleted. And you can see a very nice positive correlation between the number of links of a node and the percentage of the list of proteins in that category.

That means the hub proteins meaning, the nodes that have more connections in the network, when it is deleted it will have a stronger impact to the cell itself.

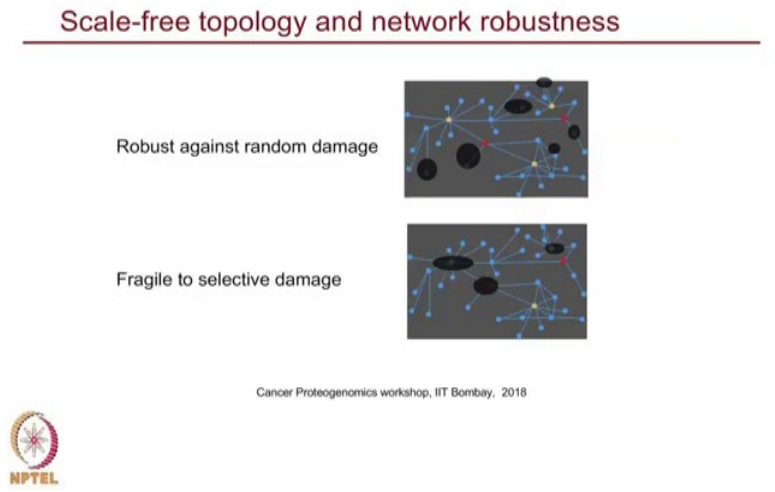
(Refer Slide Time: 20:42)



And during after that publication and people keep exploring; what are the other possible properties of the hubs in the biological network. And in this review article by Mark Vidal in 2011 and he summarized the major findings. So, basically in the first study showed the hubs correspond to the essential genes.

And they tend to be older proteins and usually they evolve more slowly than other proteins in the network. They have tendency to be more abundant and they have a larger diversity of phenotypic outcome, when it is deleted it is they could have different types of functional; that means, they may be involved in more different types of functions.

(Refer Slide Time: 21:26)

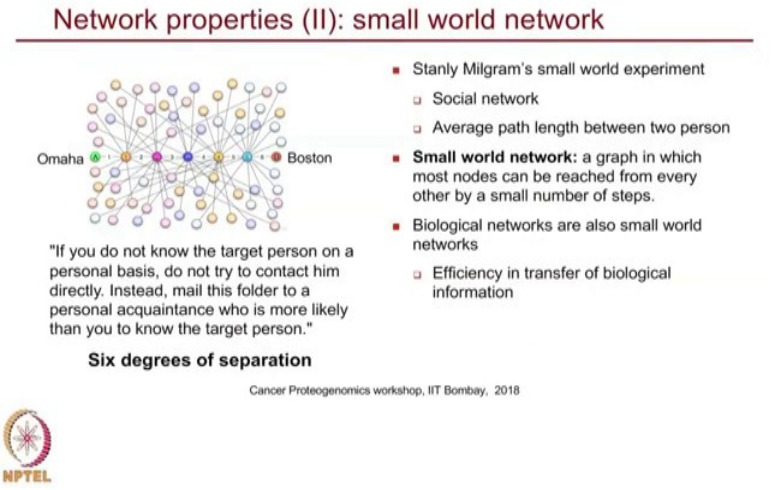


So, and then you may wonder why the power, the cells choose to establish or the evolution has I mean evolved into the scale free network rather than a random network right; does this give the cell any benefit.

And indeed if we think about a network like this and if it is a scale free network; let us say the mutations, we know that the mutations or random attacks and genes, right. But if the mutations occurred randomly across the genome, it could hit any of the proteins; but most of the proteins are the proteins only with one or two links. So, like mutation in these proteins will not affect the cell either system, right.

So, it provides us cell, the robustness to survive. So, meaning mutations typically do not have a important consequence, because it does not make significant impact to the network as a system. But this also gives us some; we can also think it this way; now if we want to try out try to kill bacteria in the cell and what should we do and in that way we can think about target, the attack of the central node, meaning the hubs. For example, in the bacterial network, because if we attack this important nodes; then it will destroy the bacteria, and in that way we can think about it how to prioritize genes when we want to treat the disease we were trying to kill fungi or kill bacteria in human.

(Refer Slide Time: 23:09)



That is the first property of the network is the existence of hubs or the scale free property of the network. The next thing I want to talk about the small world network. So, this was originally studied by the in the social network context the by a scientist called them Stanly Milgram. I think he was a social scientist in Harvard and he did this experiment in 1967. So, the idea is, he want to understand how people in the are connected, right to each other.

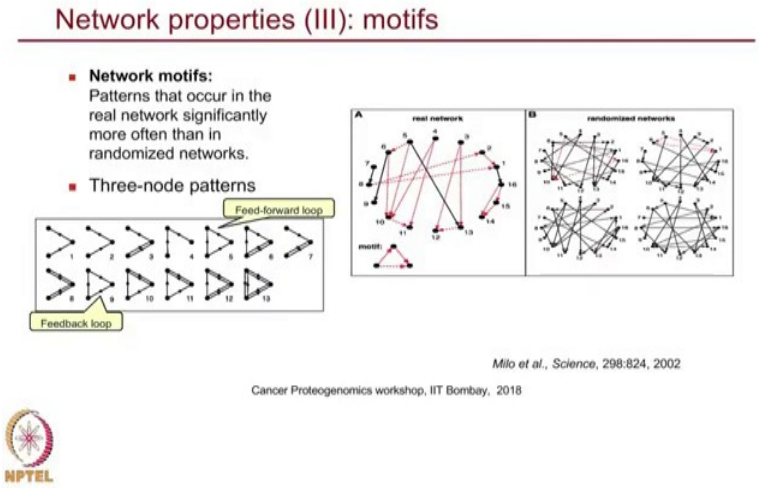
Of course and now it is easy because we have the internet and since we know how to do this; but at that time I mean it is actually very difficult to do this, how can you even think about I mean to a way to estimate or to understand how people connected to each other. I think he come up with a very interesting experimental design. So, he prepared the 160 packages and he gave those packages to random people in a very small city Omaha and Nebraska in the US. And then he gave these packages to random person in this small city and they asked them to try to send the package to a stock broker in Boston which is far away from Omaha, right.

And then you cannot directly send this to him, because you do not know him; you have to every time you have to pass this to somebody you now, this is his instruction. So, and then at the end he got collected all the letter from this stock broker and then he counted how many times is required for the letter to reach the stockbroker from the original place. And surprisingly I, the average number of passes to reach him was 6.

So, that is a famous 6 degree of separation, I think probably some of you have heard about of this phrase came from his study. And the then through this study we understand, well although it looks like everyone so far away from each other; especially when you think about there was in 1967 right and maybe nobody knows each other. But now you see for any person you do not really have a good connection and still you can reach him in 6 steps.

And so, this is what we called the small world network and if we look at the biological networks it is also, I mean the average path lengths between any two nodes; and it is between next three to four or actually it is even smaller than the social network he estimated at that time. I think one reason with the, I mean it is a biological networks use a small network structure is probably, because it can pass information more efficiently.

(Refer Slide Time: 26:15)



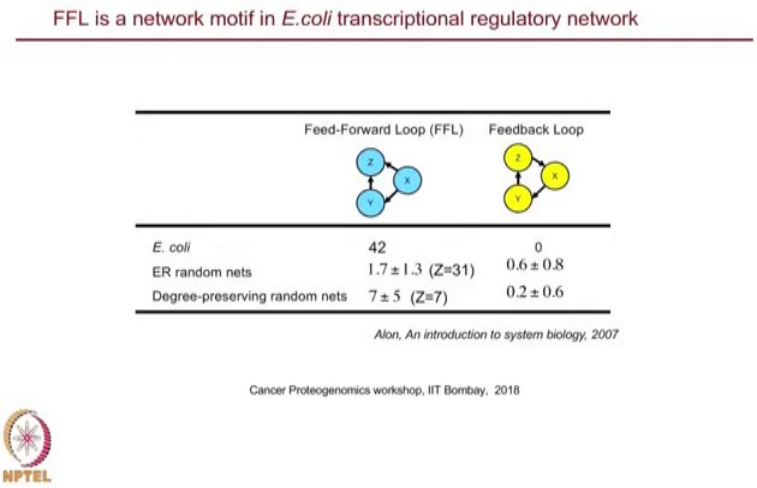
And the next property I want to talk about is the motifs. So, when we talk about motifs, it is a patterns that keep occurring in a system more often than random chance, right. So, if we think about three nodes in, I mean, what are the possible relationship between them. There actually is 13 different types of relationship between three nodes, if they are somehow connected. And then for example, this pattern is very well studied it is for Feed-forward loop, right.

Node A can have a positive relationship with this node B and it also through node C it also A C B, right. So, there is it is called the feed forward loop and this is called a

feedback loop. So, let me go around and then for each of this maybe let us say this is the real network; and then you can count all the, how many times you see a feed-forward loop, the three node motif in this more network.

And then you want to know whether this is about if, I mean because the definition is statistically more frequent than random chance, right. What you can do is to randomly shuffle the nodes in this network and build some random networks; and then you can, after you do the shuffling, you can count the number of the feed forward motifs again. And then for example, here we see a lot more feed-forward motifs in this network than the random networks.

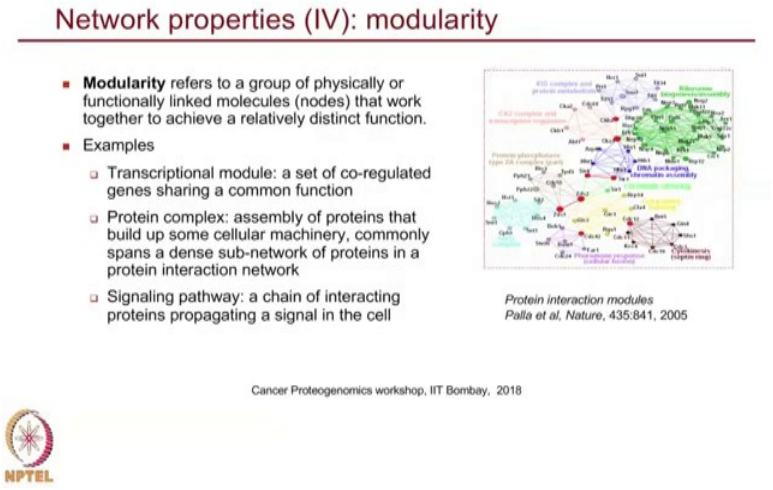
(Refer Slide Time: 27:56)



And there is a study in E.coli transcriptional regulatory network; because in engineering we know that the feed-forward loop and the feedback loop both are very commonly used, I mean to regulate the system, right. But, in the biological network connect the transcriptional network, they actually found that the feed-forward loop, they observed 42 in the E.coli network.

But they only did not observe any feedback loop in that network and this is significantly more than what you would expect by chance. So, that means, this is a kind of a motif or the way of organizing the nodes that the biological system actually use; but not I mean the feedback loop.

(Refer Slide Time: 28:49)



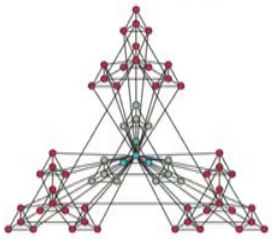
And the next one is the modularity. So, this basically says, I mean the genes and the, or proteins in the network tend to form groups rather than randomly connected to each other. And so, in the transcriptional networks, I mean these are of course, the I mean transcriptional modules meaning the target genes of course; I mean they are, if they are targeted by the same set of transcription factors, they form a group, right.

And in the protein-protein interaction network like this one, we can see I mean the there are groups of proteins in the network and these are usually the protein complexes in thenetwork. And the signaling networks, we can have the signaling pathways that the represents the modules in those networks.

(Refer Slide Time: 29:45)

Network properties (V): hierarchical organization

- Modules combine in an iterative manner to generate hierarchical networks.

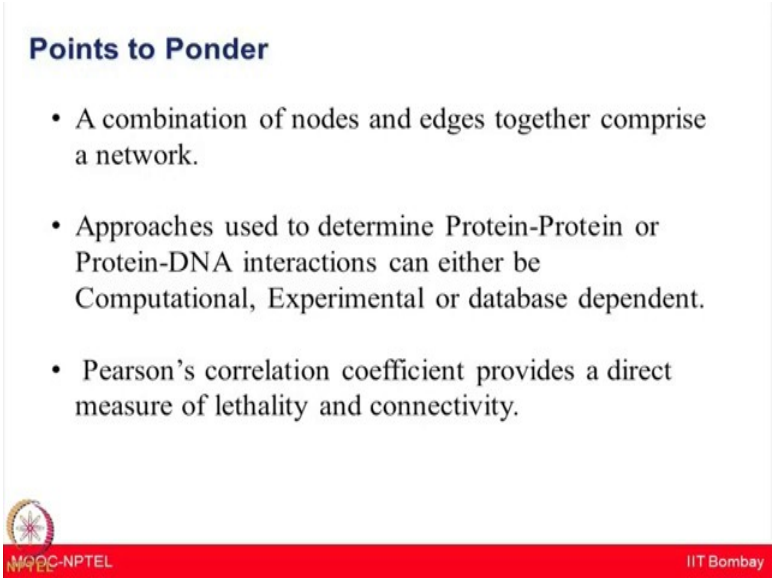


Barabasi and Oltvai, Nat Rev Genet, 2004

And also the modules are not only I mean occurred a like separately, they also organized in a hierarchical way. Meaning there are small modules that get connected to each other to form a relatively larger modules and then eventually and reach the whole network. And we can think about the protein complexes and the two complexes may interact which with each other to form a, I mean relatively larger network and the eventually we know the 6 degree separation right, everything is actually connected.

So, those are the five different properties associated with network that have been revealed through a lot of studies.

(Refer Slide Time: 30:34)



In today's lecture, you were introduced to various types of protein-protein, protein DNA and other bio molecular network analysis. These interactions can be either experimentally determined or through the use of computational software predictions. Databases like DIP, MINT, BioGrid, they contain information of various protein-protein interactions. Metabolic networks can be studied using databases like KEGG or Meta Cyc; while co expression networks can be computationally determined.

Pearson correlation coefficient is the direct indicator of lethality and connectivity. It is important to note that biological networks generally follow the characteristics of small world networks. I hope this information is giving you more clues and ideas, how you can utilize these available resources and do more protein-protein and protein bio molecular interaction and network analysis for your own data set.

In the next lecture, we will learn more about the visualization of network and Dr. Bing Zhang will continue his lecture and show you how to use various tools to do data visualization for network analysis.

Thank you.