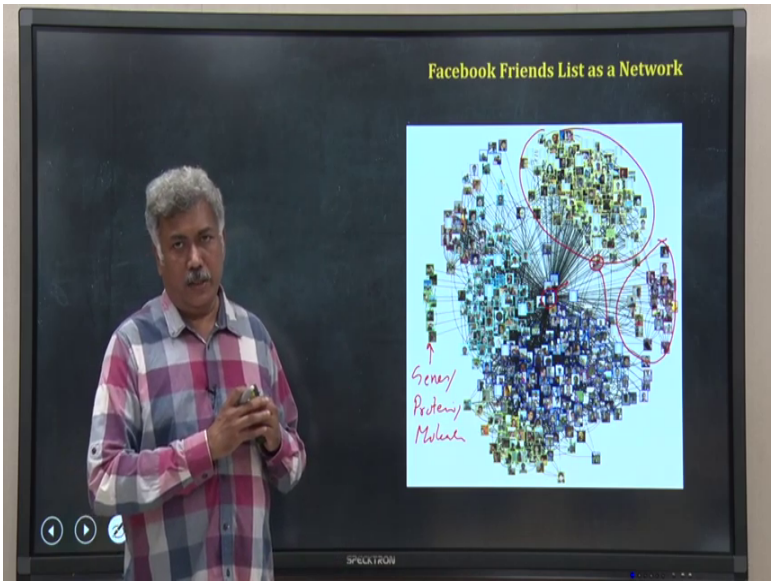
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences and Bioengineering
Mehta Family School of Data Sciences and Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture 23
Network Visualization

Hello, welcome back. You must be having a Facebook account and there may be hundreds of friends in your friends list. And many of these people who are your friends, maybe a friend to each other. So, now you know hundreds of people and they also know hundreds among them. How can you visualize this information neatly in one single diagram? The simplest way to visualize it is to use a network.

(Refer Slide Time: 01:06)



Let me give you an example, here is my Facebook network, I must be somewhere here in the middle and then all these square boxes the photos that you see are people are in my, those are my friends in my friends list in Facebook, and I have connections, edges between each of these photo each of these people. So, I have edge from me, let me take a color and show you from me, I am somewhere here from here, I have edges going to people like someone here and then that person also to another person, so, there is edge going from that person to this one.

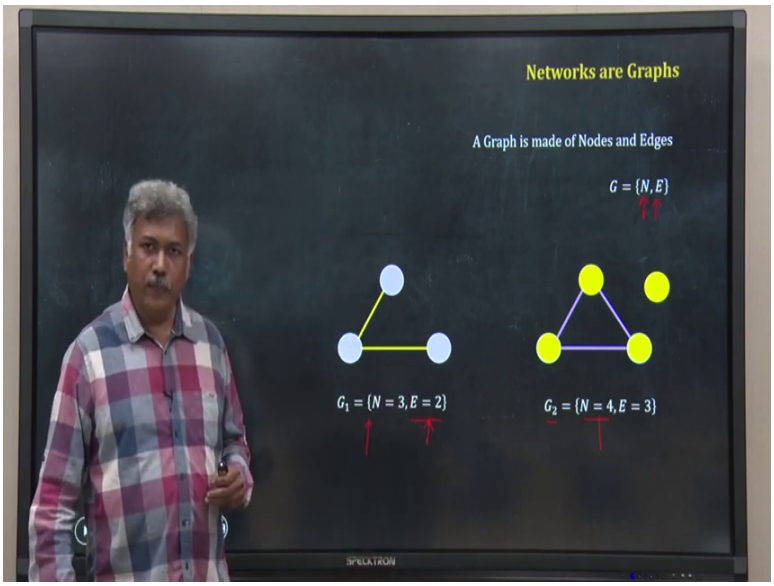
So, by using these object these squares within photographs of each of this person, and by connecting them by edges I am showing who is connected to whom, by connected I mean, whether they are friends on the Facebook friend list or not. Now, once you have represented this data, I have used a tool after downloading the data from my Facebook account and that software actually organizes these diagrams in a very nice way.

You can see clearly there are something called cluster. Like for example, this is one cluster you can see people who are connected with each other, similarly there may be another cluster. So, these are the cluster of people because these are coming from different places. For example, I may have studied in a particular school. So, all my friends in there from that school will belong to one particular cluster. I have worked in some other organization.

People from that organization, they know each other and I know them so they belongs to another cluster. So, in this way, you can create a Facebook network and visualize the whole data. This one will be called a Facebook friends network. And I can use this type of network diagram to visualize large numbers of large scale data in biology also. What you have to do? You have to do a simple change.

You replace these people, each of these people by suppose genes or proteins or suppose molecules. So, in place of human being in these networks, I will put molecules, genes or proteins, something like that, which are relevant for a particular biological analysis. And the edges between them, the connection between them will represent some sort of relation and what you will get will be called a network. Now networks are actually mathematical objects.

(Refer Slide Time: 03:51)



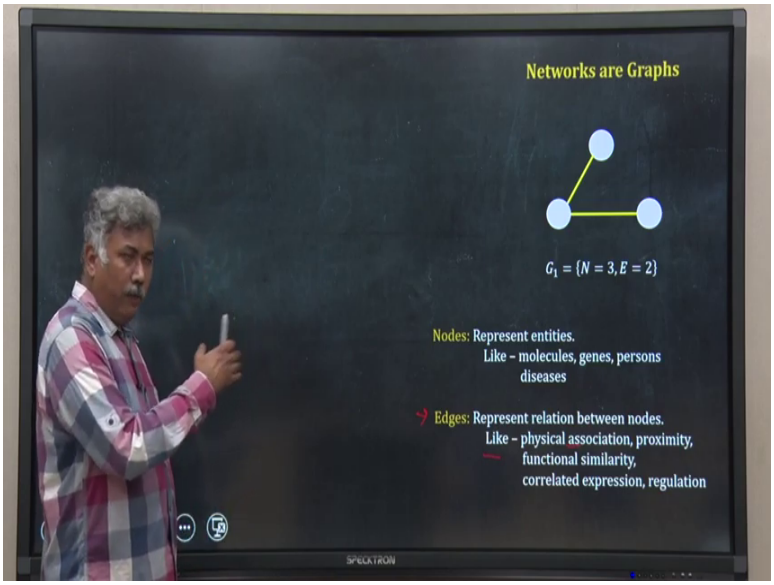
$$G = \{N, E\}$$

$$G_1 = \{N = 3, E = 2\}$$

$$G_2 = \{N = 4, E = 3\}$$

They are called graphs. What is a graph? Graphs are set of nodes and edges. In my previous example, in the Facebook network, each of these individual was a node. So a graph will have certain number of nodes and those nodes will be connected with certain number of edges. For example, if you see here, in this first graph, graph one, I have 3 nodes, and they are connected by 2 edges whereas in that second one, graph two, I have 4 nodes, and 3 edges. So what does these nodes and edges represent?

(Refer Slide Time: 04:34)



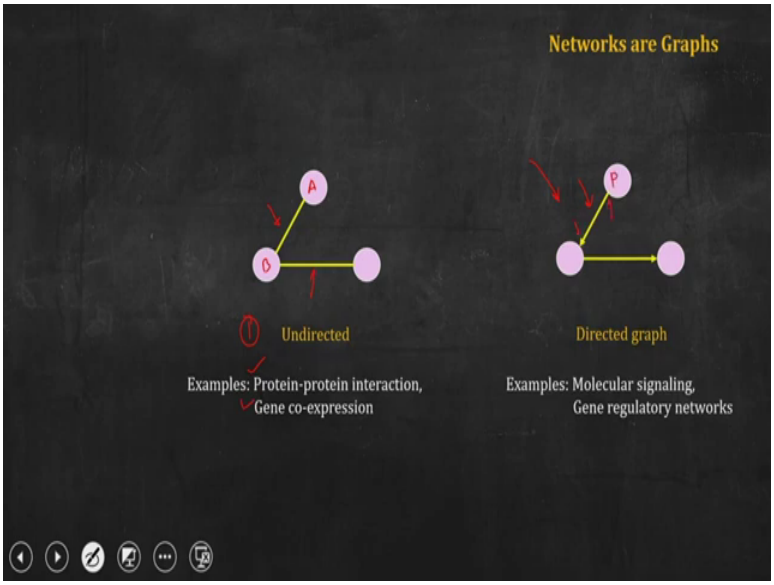
$G_1 = \{N = 3, E = 2\}$

In a network when you are creating, a graph you are creating. Nodes will represent entities elements or objects. For a biological network when you create a graph, your nodes will represent molecules, genes, they may represent species, they may represent cells, whereas the edges they will represent some sort of relation between these nodes between these objects.

For example, suppose two genes have similar expression. So, if you create a graph or network where each of these nodes are genes, then you will put a link between them, edge between them, these two genes. Or suppose you are creating a protein-protein interaction network, you know 2 protein interact physically they bind to each other physically.

So, when you create a network or graph, each of these nodes will be protein and you will put a link between them, edge between two proteins a pair of proteins if they interact physically. So, that means, the edges represent physical association they may represent proximity, they represent functional similarity, they sometimes may represent correlation in gene expression or may represent even regulations.

(Refer Slide Time: 05:49)



Now, I have shown the graphs to a few examples and then there can be different types of graphs also. For example, the one common type of graph which you will encounter very frequently are called undirected graphs. What is undirected graph? A undirected graph will have nodes and the edges have no direction as you can see here, this is one edge and this is one edge. Suppose this is A and this is B, I can say the edge is from A to B, I can also say the edge is from B to A.

So, there is no directionality in this connection between A and B. So, it is undirected graph, few examples of undirected graphs or networks are protein-protein interaction network, I will show examples, then gene co expression network that is also a undirected graph and I will show you a few example of that. On the other hand, these diagram shows a directed graph.

Notice the edges, these edges are not a straight line, they are arrows . So, the arrow has a start and arrow has a end. So, suppose this is P and this is Q and maybe in this case, I am representing P and Q are two molecules, two proteins and P controls production of Q. So, I am giving an arrow from P to Q. So, that means, there is a directionality of this connection, directionality of this edge.

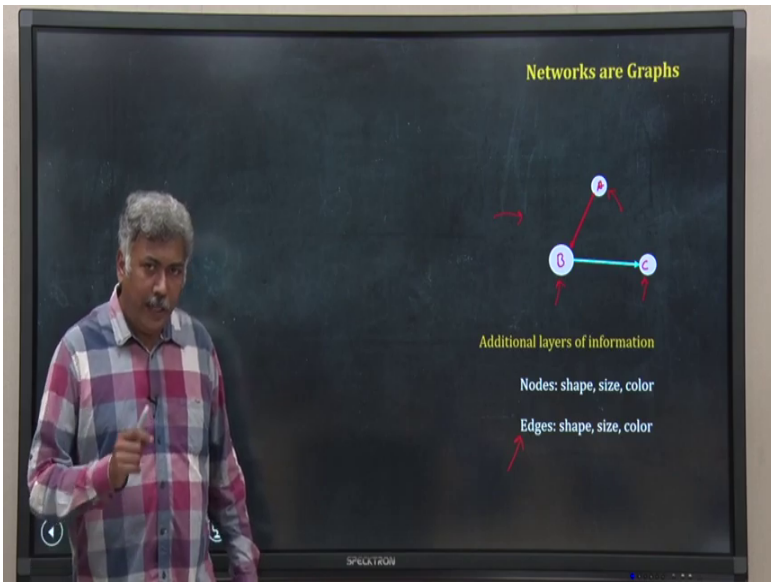
So, that is why it is called directed graph, where you will find directed graph? When you will create network for molecular signaling pathways or maybe you are creating a network for gene regulation. So, you are creating gene regulatory networks, those networks or graphs will be

directed graphs. Now, the interesting part is that graph are mathematical objects, and a network that you build for visualization are actually graphs.

And once you have the graph, then you can actually use the mathematics of graph theory to also analyze the network mathematically quantitatively. So, visualizing network is not only for visualization. At the same time, you can also do quantitative analysis of those networks using graph theory. Although in this lecture, we will not go in quantitative analysis of networks or graph, we will focus primarily on the visualization.

Now, what I have discussed till now is that the networks are actually graphs and they are made up of nodes and edges and node represent objects or entities like molecules, human being species, anything like that, whereas the edges represent the connection between them.

(Refer Slide Time: 08:37)

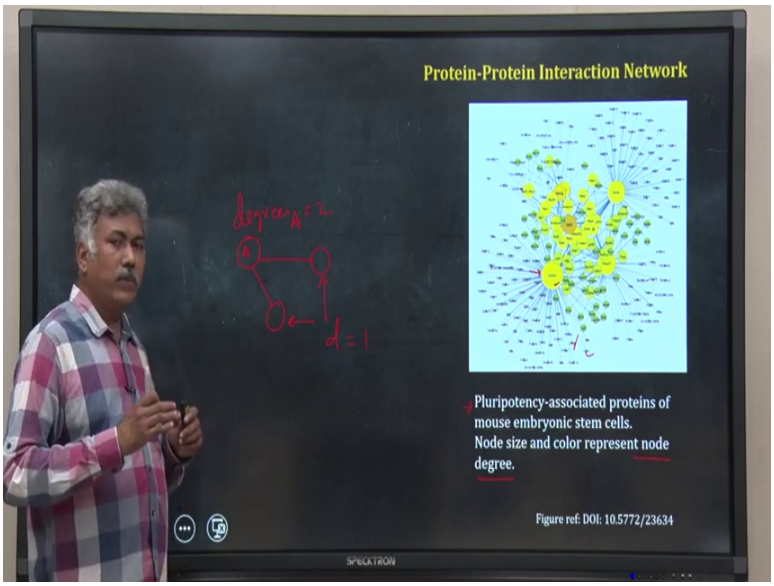


But that is not the end of the story; I can actually add additional layers of information. For example, you can change the shape of this shape and size and color of these nodes. By changing the shape and size, you can convey certain amount of information to the person who is seeing these network or graph. At the same time, the edges the shape, size and color of the edges can have, those edges or connections can also be changed to convey certain additional information.

Take the example here, suppose take this is a directed network, directed graph of 3 nodes. So, I can say suppose this is A this is B, and this is C. And suppose from a Pathway Database, I know that A controls B and B control C. Now, I have done some gene expression analysis and a particular system I have seen expression of A has some amount whereas the expression of B is double of that amount, whereas the C has the same amount of expression as A.

So, how can I convey that additional information? What I have done, I have changed the size of the circle. You see, A and C has same size whereas B is double in size. So by this size, change in size and conveying that they have different level of expression, level of production. So, in this way you can add different layers of information in your graph or network.

(Refer Slide Time: 10:12)



Now, let me start with one of the commonest network that we see. This is called protein-protein interaction network. There are high throughput experiments, like two hybrid system by which you can actually investigate which protein in genome scale study, which proteins interact physically or physically associate with each other inside the cell. Lots of experiments has been done, there are databases where this information is curated.

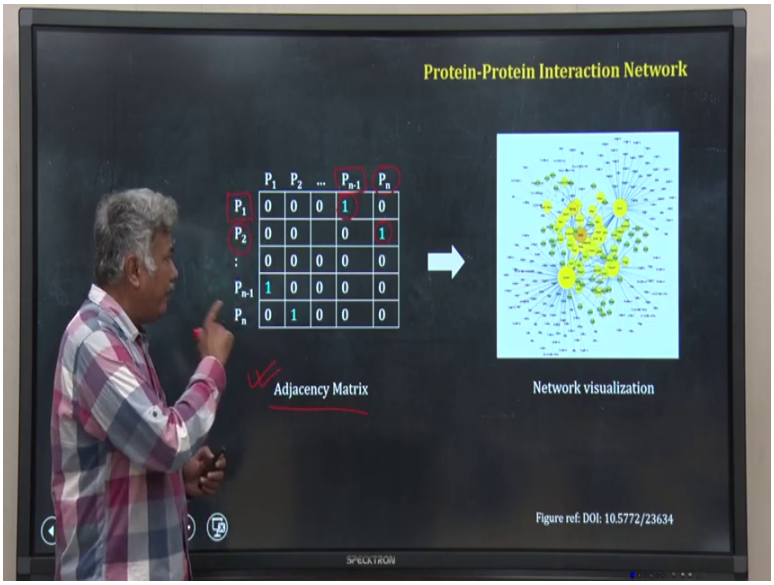
At the same time, there are computational methods also to identify partners or physical association in a set of proteins. So, suppose you have done some experiments or you have collected the data from a database. Now, you want to represent these data or protein-protein interaction as a network or a graph. So, each of these nodes of your graph or network will be protein. And if two proteins A and B, they interact physically with each other, you will put an edge between them.

That is what has been done in this graph, what they have done, they have taken proteins, which are involved in pluripotency of a mouse embryonic stem cells, and each of these node are those proteins, and then what they have done, they have put edges between like this between two protein here this one and this one, to represent that, they know these two protein physically associate with each other.

Now, what the size? At the same notice that at the same time, they have changed the size and color of each of these node, they are of the circles of different size and color. So, what they are trying to convey by this color and size, they are trying to convey the degree of each of these node. Now, what is degree? Degree is very simple quantitative measure of how many edges are there connected to a node.

So, if I take a graph like this, where this one is suppose A, A has two edges connected to this, so, the degree of this one will be degree of A will be 2, whereas the degree of these 2 is equal to 1. So, in this diagram, they have modified this color, and the size of each of these nodes, depending upon the degree. As you can see, oct 3 and 4 is actually a hub molecule. It connects to lots of proteins and controls the function of lots of protein. And that is why it has a very high degree. And that is why they have used a large circle with a deep color to represent oct 4.

(Refer Slide Time: 13:20)



	P1	P2	...	Pn-1	Pn
P1	0	0	0	1	0
P2	0	0		0	1
:	0	0	0	0	0

Pn-1	1	0	0	0	0
Pn	0	1	0	0	0

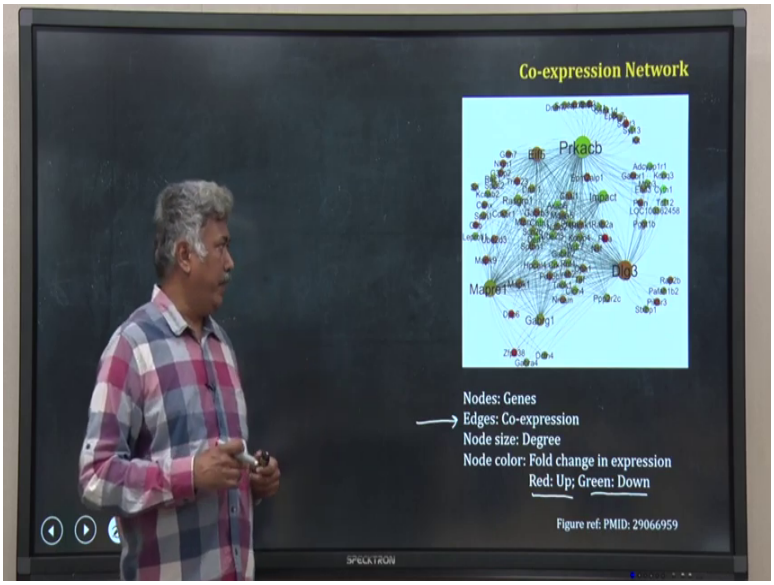
Now, when you want to create this type of diagram, how will you store the data, because you have to feed these data of physical interaction between the multiple proteins to the computers to the software, so, that it can create the plot. The data is usually stored in something called adjacency matrix. Let me explain that. So, this is a adjacency matrix for n number of proteins and they are interacting with each other.

So, this is a square matrix as you can see here, I have n number of columns and n number of rows. Now, I have filled this matrix with zero and then in specific cells, I have put 1 for example, this one is 1 what does that represent? That represent that P1, I know that P1 physically associates with P n minus 1. So, these two protein either from my experiment or from database information I know they physically bind to each other.

So, I have put 1 there. Similarly, I know P2 physically associate with P n, that is why I put a 1 there in that cell. So, in this way you create a matrix and that matrix is called adjacency matrix. Think for a moment try to create this adjacency matrix yourself. Yes, and then see, you can easily now create a graph on network using this information, you do not need any other additional information.

So, mathematically a graph or network that we visualize is actually nothing but a adjacency matrix. Now, I have discussed till now about protein-protein interaction network and from that I am explaining what is the adjacency matrix.

(Refer Slide Time: 14:57)

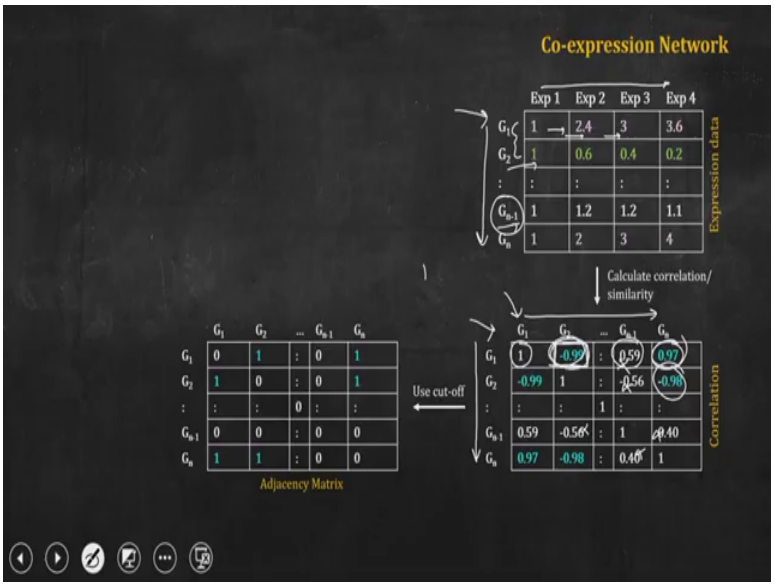


So, now I will move to another very popular network visualization that is called gene co-expression network. That is also a graph and so, it is made up of nodes and edges, but now the nodes are actually nothing but genes. So, here I have multiple genes and each of them are represented by circles of different color. Why? Because this is coming from a microarray experiment and they have measured the fold change in expression of these genes in those experimental conditions.

So, those genes which are upregulated are marked by red color, whereas, those genes which are downregulated in the experiment are actually green color. So, they have used a color range from red to green to represent the level of expression either upregulation or downregulation of these genes. And that is why each of this circle is colored by that.

So, that is one layer of information and then their size are also different, because they are representing the degree of each of these nodes, how many edges are there connected to these each of these node, there is a degree of each node by the size of a circle. And now, you notice each of these nodes are connected by edges to some other nodes. So, what does a particular edge represent in this case, in this example, in this plot, the edges represent co-expression of two genes. What do I mean by co-expression of two genes?

(Refer Slide Time: 16:30)



Let me explain that, suppose you have done an experiment you have 4 experimental condition, you are doing a microarray experiment or suppose, you are doing large scale RNA seq experiment to observe the, measure the change in expression of n number of genes in different experimental condition. In this particular case, I have shown a data table on the top I have 4 experimental condition, experimentation condition 1 to 4 and I have n number of gene.

The first gene has some gene upregulation in its expression. So, you can see in experimental condition 1 it is 1 then it is fold change has become 2.4, then experimental condition three with a 3 fold increase in expression, then in the fourth experiment, it has 3.6 fold change in expression with respect to experimental condition 1, whereas, for gene 2, the expression has got downregulated.

As we move from experiment 1 to 2, to 3, to 4, the expression level fold change has become fractional. On the other hand, for this one, n minus one gene, as you can see, gene expression level has not changed much it remained ballpark close to 1 and 1.1. So, now, if you have this type of tabular data, then what you can do, you can now try to find some similarity measure, which will measure the similarity in change in gene expression

for these genes across the experiment. There can be many type of similarity measure, one of the easiest one you can imagine is Pearson correlation. So, you can imagine G1 G2 G3 up to GN, and these genes and their expression level in all four experimental condition are random

variables. So, you can calculate Pearson correlation between them. And you can show that in a tabulated fashion.

So, that is what I have done, I am using Pearson correlation coefficient to measure the correlation between these genes in their expression level across the experiment. So, as you can see, obviously, this will be a square matrix where G1 to GN on the column G1 to GN on the row, and the diagonal element will be always one because this one is for correlation between gene one and gene one. So, that will be obviously one.

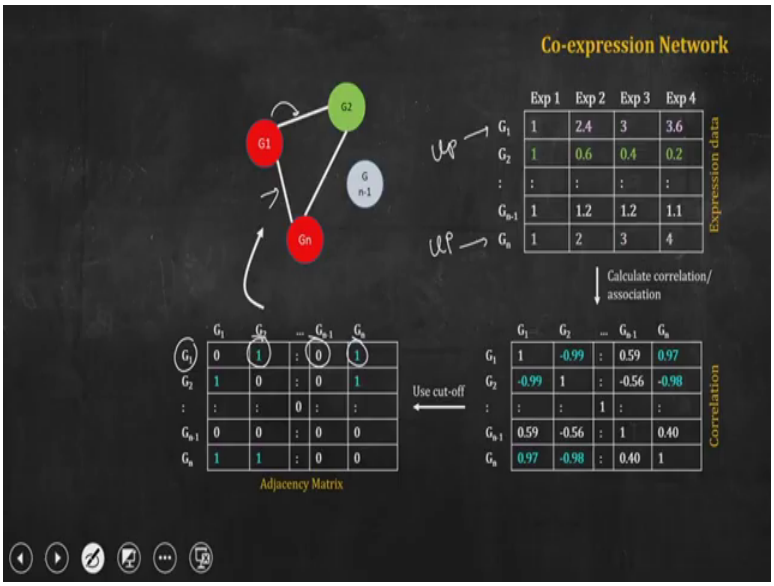
But look at these value, this is the correlation between gene one and gene two. As you can see here, in gene one, as I am doing different experiments, 1 to 2 to 3 and 4, there is a increase in gene expression. Whereas for gene two, in those experiments, the gene expression has dropped, it is downregulated. So, that means there is a negative correlation, and that is what I have got here, minus 0.99.

Similarly, G1 and Gn, if you see, both of them has actually increased in my experiment, that is why they have a positive and very high correlation of 0.97, whereas, Gn minus 1 has remained almost unchanged across the experiment. That is why its correlation with gene one is low 0.59. So in this way, I create a correlation matrix. So that correlation matrix in a way is showing me the similarity in expression of different genes.

Now, what I will do, I will use some sort of threshold or cutoff. So, I am I have decided here that if the correlation coefficient is above 0.9, whether minus 0.9 or plus 0.9 whatever it is above 0.9, then I will consider them as a correlation, as a effective correlation, rest of the values will be converted in 0. So, then what I get, I get this one, this one, this one those blue colored one, those with low value get chucked off.

And that is what I do here. So, in this case, this is point minus 0.99 so I make it 1. This one is 0.97 so, I make it 1, in this way I fill this matrix with 1 and 0, what do I end up with, I get the adjacency matrix. As you can see, these matrix has one and zero, those cases where two genes are very high correlation in expression either positive or negative does not matter I have 1 otherwise 0. So, then I can use this adjacency matrix to create a network.

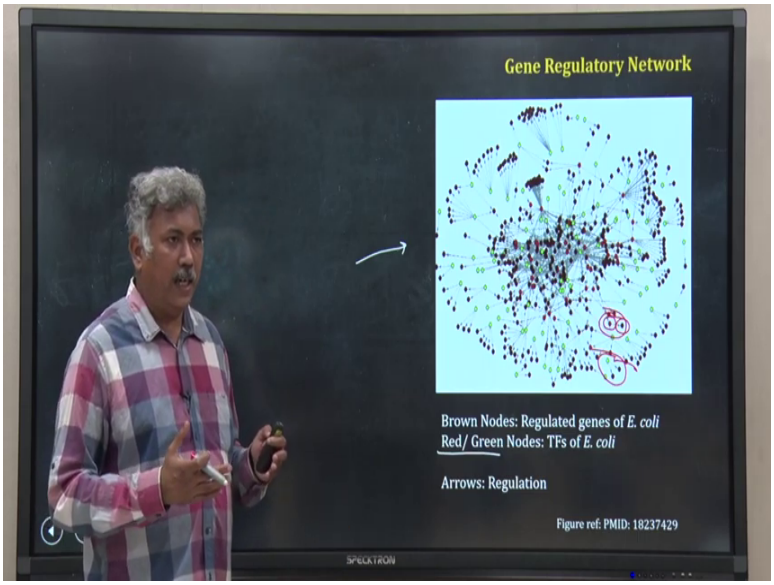
(Refer Slide Time: 21:18)



The network will look like this as gene1 has high correlation, it has 1 in the adjacency matrix with respect to 2 and also here for GN. So, that means I should have an edge between G1 and G N and G1 and G2. G1 and Gn1 if you look into that cell, this one is zero. So, there is no edge between G one and Gn1 and in fact Gn1 has no edge connecting to any other any of the other nodes either. Now, I have added another level of information in this graph, I have color coded the nodes.

So each node is color coded based upon their change in gene expression. As I know gene one and gene two has upregulation I have used red color for them whereas for Gene two there is downregulation. So, I have used green color for that. This is how you create a gene co-expression network. Now remember, a gene co-expression network is a undirected graph.

(Refer Slide Time: 22:33)



There is another type of genetic network which is a directed graph and that is called gene regulatory network, you should not confuse between these gene co-expression network and gene regulatory network. Gene regulatory network is a directed graph which represent the information of regulation, by that I mean that who regulates whom, who activates the production of which gene, which gene controls the expression of which gene that information is stored in gene regulatory network.

So as you know, transcription factors control the expression of other genes. So, in this network, there will be at least two types of node, one type of nodes will be transcription factor from those nodes, arrows will come out, and the arrows will end up on those nodes, which are genes controlled by this transcription factor.

And that is what has been shown here, in this diagram is a gene regulatory network of E. coli. And these brown nodes, which are majority here, are actually nodes, which are non-transcription factor, they are not transcription factor, these are gene which are controlled by transcription factor. And those genes which are transcription factor in these diagrams are either red or green.

What is the difference between red and green here, the green nodes are transcription factor, who controls others, but are not controlled by some other gene, whereas the red transcription factor transcription factors, which control other genes, but at the same time, those transcription factor

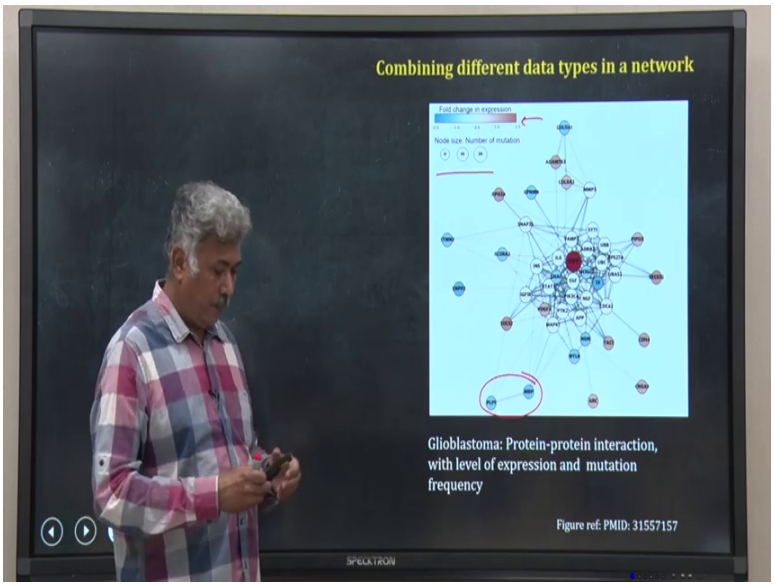
also are controlled by another transcription factor. Take a example. In this case, I have a green one, controlling two brown nodes.

And you can see I have a arrow here starting from the green node ending at the brown node. So that means this green node is a transcription factor. It controls the expression of those two brown nodes. You can notice another thing, there is no arrows pointing towards this green node. That means there is no transcription factor that controls the expression of that particular transcription factor.

That is why it is marked by green whereas if you see this red one, in this red one, red one is a transcription factor. It is controlling the expression of these brown one, but at the same time, this red one is controlled by a green transcription factor. So that is how in this directed graph, we are representing an E.coli gene regulation, which gene is controlled by which transcription factor.

Now, we are not limited to representation of biological data in these two fashion only, that I only show gene co-expression data and what we say protein-protein interaction data, we can actually mix and match different layer of information in one single graph.

(Refer Slide Time: 25: 38)

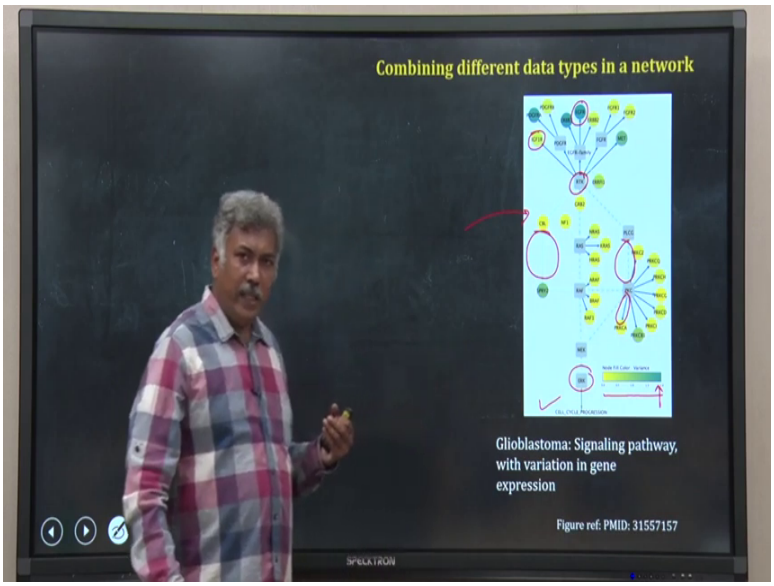


So, let us take this example, here we have multiple layer in this particular plot. But one important thing is that each of these nodes are genes, which are involved in somehow involved in Glioblastoma, Glioblastoma, a type of cancer. Now, what they have done, they have fetched the protein-protein interaction information from a database, and what they are doing each of these edges are representing that protein-protein interaction.

So that means, their database says these two protein has interaction among themselves. Now, apart from that, the color of the each of these nodes for each of these protein represent the fold change in expression that they have observed in those Glioblastoma patients. So, the color is representing now, the change in expression of this protein. At the same time, they have varied the size of each of the circle, represented the node, what do they represent? They represent the frequency of mutation.

So, those proteins which are very frequently mutated in Glioblastoma, in this diagram, it has a very large diameter, is a large circle, whereas those proteins which does not get mutated in Glioblastoma, are with the smallest circle. So in this diagram, I have three types of data hidden, one is protein-protein interaction data, second is level of expression data, and the third one is the mutational frequency data.

(Refer Slide Time: 27:16)



Let us take another way of representing multiple layers of data I have taken from the same paper. We know there are signaling pathways and in cancer, certain signaling, canonical signaling pathway are overactive and they are predominant, for example, here what they have shown, they have shown the RTK ERK pathway, which control cell cycle progression.

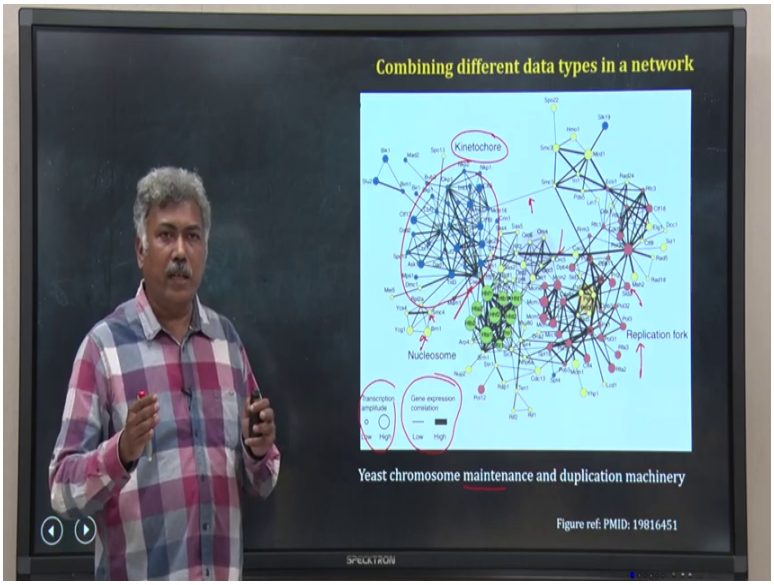
So you can get this pathway information from different databases for example, K, you can go and get the curated pathway diagram for a particular pathway. Now, once you have taken that pathway, what they have done, they have overlaid certain gene expression information on that path. So, this diagram is a diagram of RTK ERK pathway, which is involved in cell cycle progression.

Each of these molecules are involved in that pathway, and the arrows are of different shape and color. For example, these arrow represent inhibition, whereas, these arrows represent activation, whereas these edges represent their part of that system complex. Now, this information they have got from a database and now on that they have layered the gene expression data.

They have observed the variation in expression of this gene across different Glioblastoma patient sample, and that information is color coded like this from yellow to green, and each of these nodes, each of these proteins are now color coded by that. So, you can see EGFR is on this side. So that means it has a high variance in expression different sample, whereas IGF1R has lower

variation in expression across different samples. So, in this way, they have layer two set of different data in one network diagram.

(Refer Slide Time: 29:09)



Here is another one. In this case, what they have done, they have tried to show the proteins which are involved in chromosome maintenance and duplication in yeast. So, what do we have each of these nodes in this diagram are actually proteins and they have color coded them based on their chromosomal location, where they associate with the chromosome and at the same time, they have layered another information, the edges if you notice are thick and thin, this is very thick, whereas this one is thin.

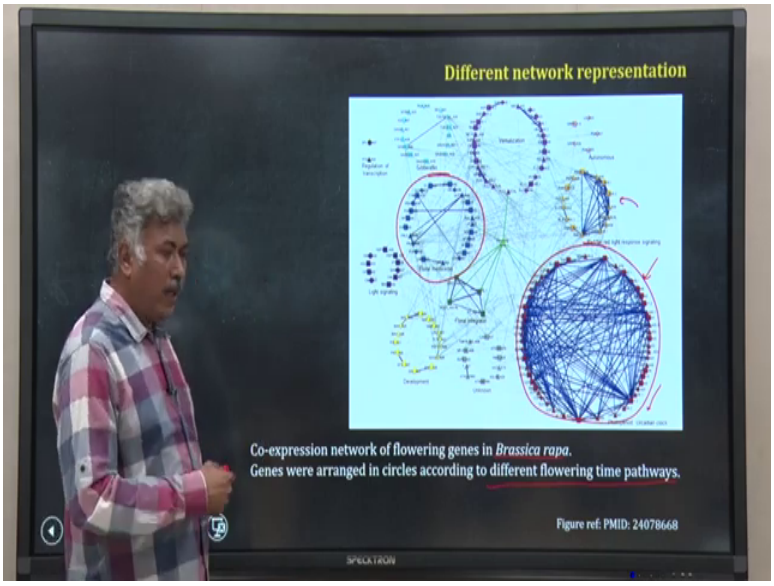
So the thickness of these edges represent gene expression correlation. So, if two genes has very high co-expression correlation just like suppose Pearson correlation, then that will have a very thick line between them whereas if two genes, two protein has very low correlation, they will get connected by a thin line. Apart from that, they have also changed the size of each of these nodes as you can see here. So, these proteins are involved in a duplication machinery.

So, they have measured their expression amount across a cell cycle and that was encoded in the size of each of this node. So, in this way they have layer multiple information in one network diagram. One interesting thing here you can see that those proteins which belongs to the same, work in the same chromosomal location, for example, replication fork they are coded by red

color and they are all clubbed together in one portion, whereas, these things, these blue colored which are involved in this one, they are clubbed in one place.

Actually, that is what the software has done, the software forces the diagram to segregate these module in different sections. And that helps us in visualization, this type of change in layout is very crucial, when you want to show a large network otherwise, everything will collapse on each other. So, you will get a hodgepodge of everything. So, there is a multiple, there are multiple types of layout.

(Refer Slide Time: 31:35)

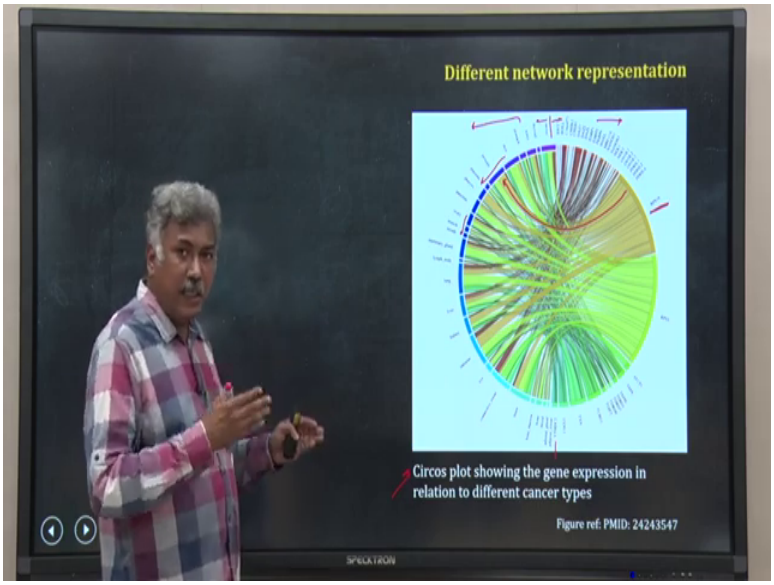


I will show another type of layout we segregate different modules or different part of the diagram in different section. So that it become much easier for us to understand. Take this example, this is called circle layout, as you can see, genes are arranged in circle. This I have one circle here, I have another circle here.

So, what is this diagram, this is a gene co-expression network of a flowering, of flowering genes of a flowering plant and what they have done, they have created these different circles according to different genes which are involved in different flowering time pathway. So, all those genes which are involved in these circadian clock, they are together in this particular circle, the edge between each of these gene represent the correlation in their expression.

Whereas, some other genes which are involved in suppose here, I have another module where the genes are arranged in circles and they also have correlation among them between two module between two circle there are also edges representing correlations. So this is one way of actually forcefully arranging a layout in a particular fashion to give clarity in the figure.

(Refer Slide Time: 32:44)



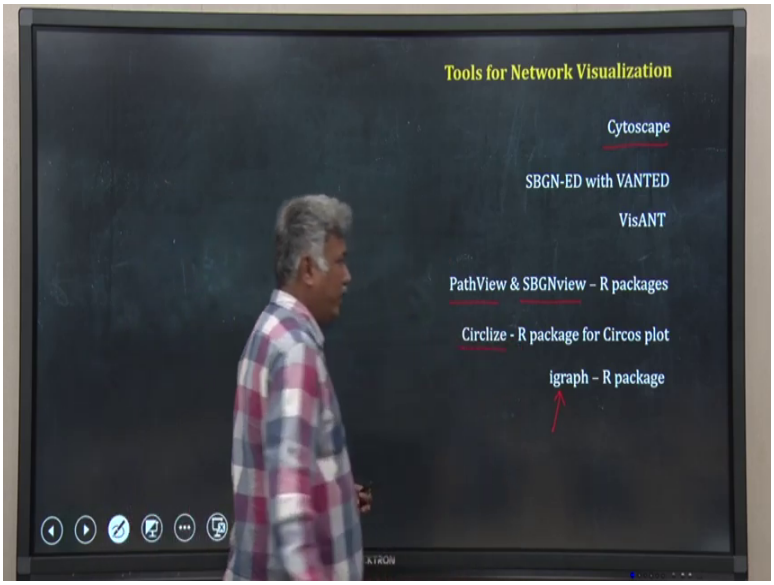
This one although somebody can argue is not a network but in a way it is actually network but represented in a different fashion and it has now become very popular in genomics proteomics community because it can beautifully visualize a large and diverse data set. This called a Circos plot and I have shown one of the simplest one, what we have in this plot.

If you see if I divide in two part, on this side I have different types of cancers and what the authors have done they have collected curated data in level of expression of certain genes in those cancers and then so you can imagine each of this gene for example this one gene and you can consider that as a node actually and whereas these cancers are also node in my network and now I have two types of nodes one set of node represent cancer one set of node represent genes and if one gene is over expressed in one particular type of cancer.

I will put a ribbon between these two and that is what they have done, for example if you take this gene, this gene has over expression in this cancer, prostate cancer that is why there is a ribbon between these two and the thickness of this ribbon how much over expression is there. So

in this way in a circular fashion with ribbons we are visualizing the over expression of certain genes in diverse type of cancer and this type of plot is called Circos plot.

(Refer Slide Time: 34:33)



So what I have discussed till now, I have discussed starting from a general network, a facebook network. I discussed about that graphs and basic concept of graph and we under understood that every network that we draw are actually graphs and in those graphs we can overlay different types of information and visualize them. So you must be wondering are there any software to do this type of, create this type of diagram or not. There are lots of such softwares.

I have listed here few only few of them but if you are involved in creating this type of graph or network diagram please, explore further. The most common one, the most popular one is obviously cytoscape. It is a very stable software with a large module libraries creating different types of network so its library is very large so you have to explore that library and you have to find the right module that can do the job for you and then using cytoscape is quite easy.

Apart from that if you are familiar with R and you want to work on R then there are lots of R packages to create a graph, visualize them and quantitatively analyze them. For quantitative analysis of graphs and networks and visualization to some extent igraph package is a standard, gold standard package in R.

So if you want to do lots of calculations quantitative analysis of the network then you have to look into igraph package, otherwise if you are more focused on visualization of networks you want to incorporate multiple pathway information, gene expression information, mutation information in one network diagram then you may look for this type of R packages like PathView, SBGNview or if you want to create a Circos plot you may use this Circlize R package which can create nice Circos plot for you. That is all for this lecture.

Thank you for learning with me today.