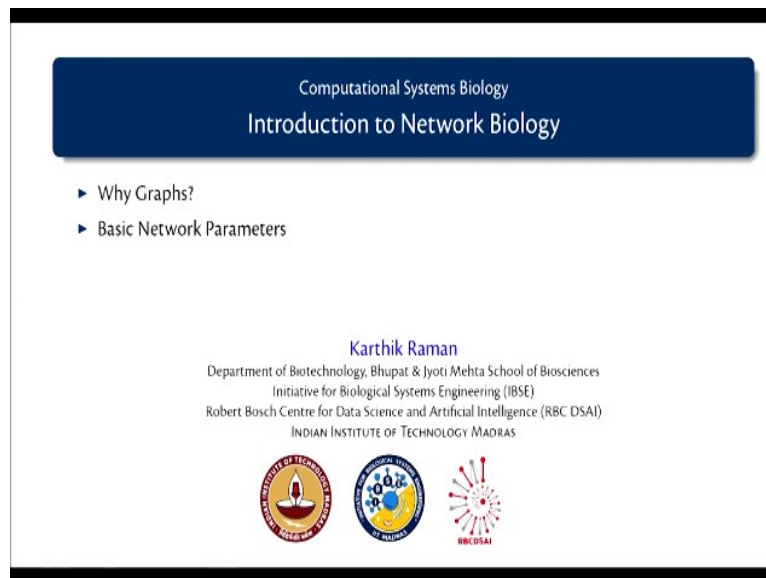


Computational Systems Biology
Karthik Raman
Department of Biotechnology
Indian Institute of Technology – Madras

Lecture - 15
Introduction to Network Biology

So, we will continue to review the basic concepts of network biology and I will motivate you today as to why we need to use graphs to study biological systems and we will also look at some basic network parameters that are used to characterize and understand biological networks, networks in general and also biological networks.

(Refer Slide Time: 00:11)




The slide features a dark blue header with the text "Computational Systems Biology" and "Introduction to Network Biology". Below the header, there is a list of topics: "Why Graphs?" and "Basic Network Parameters". The slide also includes the name "Karthik Raman" and his affiliations: "Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences", "Initiative for Biological Systems Engineering (IBSE)", and "Robert Bosch Centre for Data Science and Artificial Intelligence (RBC DSAI)", all at "INDIAN INSTITUTE OF TECHNOLOGY MADRAS". At the bottom, there are three logos: the IIT Madras logo, the IBSE logo, and the RBC DSAI logo.

So, welcome back. Let us look further at graph theory. We will look at particularly we will start moving to network biology today, right. How are graphs and networks applied in biology? The first thing is, why graphs?

(Refer Slide Time: 00:47)


History Introduction Why Graphs? Network Biology

Many interesting questions can be asked of graphs



Social Networks

- ▶ Do I know someone who knows someone ... who knows X?
 - ▶ *existence of a path*
- ▶ How long is that chain to X?
 - ▶ *shortest path problem*
- ▶ Is everyone in the world connected to one another?
 - ▶ *identification of connected components*
- ▶ Who has the most friends?
 - ▶ *most connected nodes/centrality analyses*




You know in a social network scenario, one asks a lot of questions in general. Do I know someone who knows someone who knows X? Now how often how easily do people know other people and so on. So, this is basically something like the existence of a path, right. How long is that chain to X, right? How many hops is that person away from you? You normally see this on linked in right.

It says that somebody is outside of your network, somebody is you know 3 hops away from you and so on and so forth. And shortest path, right, so this is like a shortest path. How long is the path to that someone? Is everyone in the world connected to everyone else, right. So social scientists often ask questions of this sort which boils down to identification of connected components in a network.

Or who has the most friends or who is the most influential person in a network, right. So, we look at what are the most connected nodes, different types of centrality analysis and so on. If you look at biological networks, one literally asks the identical questions, right, is there a way to produce a metabolite X from A which is an existence of a path kind of question.

(Refer Slide Time: 01:53)




History
Introduction
Why Graphs?
Network Biology

Many interesting questions can be asked of graphs

Biological Networks


- ▶ Is there a way to produce metabolite X from A?
 - ▶ *existence of a path*
- ▶ How long is that chain to X from A?
 - ▶ *shortest path problem*
- ▶ Are all proteins connected to others by a path?
 - ▶ *identification of connected components*
- ▶ Which is the most influential protein in a network?
 - ▶ *most connected nodes/centrality analyses*



How long is that chain? Is that the best chain? Is that the most biologically plausible pathway? You can ask all sorts of questions of this type. And are all proteins connected to one another by a path? Right, which boils down to identifying connected components in a network or which is the most influential protein in a network?

You may want to target that if you want to make a change in the network. Maybe you want to kill a pathogenic organism, so you then try to target the most centrally connected protein and so on. So, if you see from social network to biological networks, the kind of questions we ask are pretty much similar. In fact, the same algorithms are used to predict partners in protein networks like the same algorithm that keeps bothering you whether is this person your friend on Facebook, right. So, the algorithms that are used are pretty much similar.

(Refer Slide Time: 02:53)




History
Introduction
Why Graphs?
Network Biology

Graph Algorithms

Many many problems in science and engineering can be cast back on to a graph!

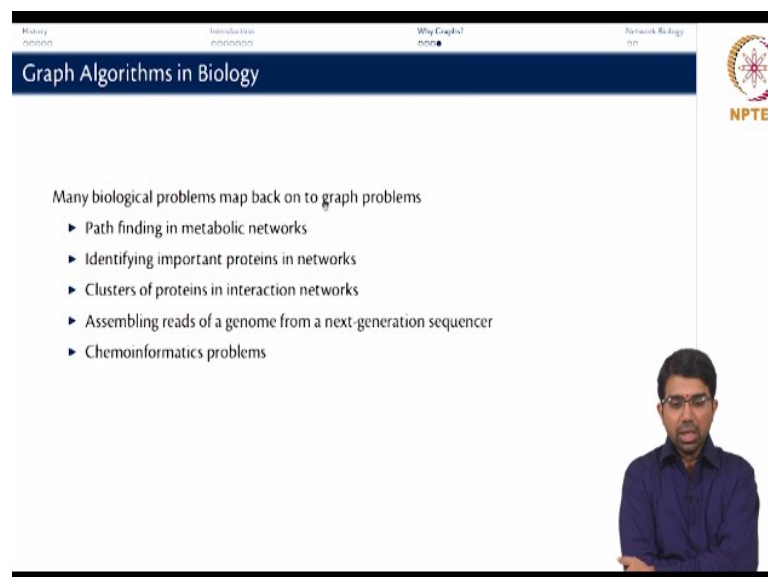
- ▶ Shortest path problem
- ▶ Travelling salesperson problem
- ▶ Finding [strongly] connected components
- ▶ Graph isomorphism
- ▶ Vertex cover problem
- ▶ Minimum spanning tree problem
- ▶ Hamiltonian path problem
- ▶ Eulerian path problem
- ▶ k-shortest path problem
- ▶ Centrality measures



And why graphs? Again, there are several graph algorithms or several problems in science and engineering which map back very nicely on to graph problems. So, these are some of the most popular graph algorithms on Wikipedia actually. What is the shortest path problem, traveling salesperson problem, how do you find strongly connected components, graph isomorphism, vertex cover, minimum spanning tree, Hamiltonian path, Eulerian path, k-shortest path problem and of course centrality measures.

So, we will not look at practically any of these in this course. This is what you would learn in a classic graph theory course. In this course, we will try to look particularly at the last item, right. We will try to study centrality measures and the importance of these centrality measures, different types of centrality measures to essentially study or rank proteins or nodes in any given biological network.

(Refer Slide Time: 03:49)



The slide is titled "Graph Algorithms in Biology" and features the NPTEL logo in the top right corner. The main content is a bulleted list of biological problems that map back to graph problems:

- ▶ Path finding in metabolic networks
- ▶ Identifying important proteins in networks
- ▶ Clusters of proteins in interaction networks
- ▶ Assembling reads of a genome from a next-generation sequencer
- ▶ Chemoinformatics problems

A small inset image of a man in a blue shirt is visible in the bottom right corner of the slide.

So, there are many biological problems also. We just talked about general science and engineering problems. There are many biological problems that map back on to graphs like path finding in metabolic networks, identifying the most central important proteins in a protein interaction network, clustering of proteins, identifying communities, groups of proteins and so on.

And of course, you know in computational biology this is slightly outside the ambit of this course of computational systems biology but how do you assemble reads from a next generation sequencer or chemoinformatics problem and so on. You can consider as we were briefly talking about in a previous class, you can even treat a single molecule as a graph and

then look at you know how different molecules can react with one another and so on just using graph theory.

So, we now switch to network biology and basically we have to understand a lot of concepts with relation to network biology. So, what are these concepts?

(Refer Slide Time: 04:50)

Network Jargon

- ▶ Node/Edge/Edge Weight
- ▶ Density
- ▶ Degree
- ▶ Shortest path/geodesic
- ▶ Diameter
- ▶ Characteristic path length
- ▶ Degree distribution
- ▶ Clustering coefficient
- ▶ Closeness centrality
- ▶ Betweenness centrality
- ▶ Edge betweenness
- ▶ Connected component
- ▶ Strongly connected component in directed graphs
- ▶ Acyclic graphs
- ▶ Motifs

So, first is the concept of, I think you all understand the concept of a node and an edge and an edge weight. The immediate next question to ask is what is density?

(Refer Slide Time: 05:03)

NETWORK BIOLOGY

	Undirected	Directed
Density	$\frac{ E }{ V \cdot V }$	$\frac{ E }{ V \cdot (V - 1)}$
Degree	k	$k = k_{in} + k_{out}$
Shortest path		
Diameter	asp	max?

So, what is density? It is essentially the fraction of edges that exist in a network. If you have, this is the number of nodes in a network, right, the size of the vertex set. So, what is the total

number of possible edges? $|V|^2$, especially in the case of an undirected graph, in a directed graph it will actually be, right and the numerator is. So, this is in the case of directed graph.

The next concept is that of degree. What is degree? It is the number of neighbours of any particular node, right. So here you would just say it is some value k and here you will say it is, $k = k_{in} + k_{out}$, right. You can separate the number of incoming edges and the number of outgoing edges, right and totally you called it degree. So, if you were to think of the adjacency matrix, how would you compute degree in a adjacency matrix?

Just row sums, just do row sums or column sums, they would be the same in case of an undirected graph, right assuming you have a full representation and not just the upper triangular representation and in a directed graph, you can get k_{in} and k_{out} by summing the rows or columns appropriately. So, that is the concept of degree. Degree is actually one of the most important, surprisingly the most important property of the network.

Maybe not surprisingly, it is one of the simplest properties and also turns out to be one of the most useful properties, especially in relation to biology as we will see very shortly. Then there is this notion of shortest path. What is the shortest path in a network? So, it is that path of either lowest weight or lowest number of edges between any 2 points such that there is no other part that is shorter than it, right.

So, you can't, so, this is not a short path, this is not a shortest path, right. I hope you can see that yellow, but that is not a short path, whereas you know the shortest path will basically be this, between A and B. It need not be unique. You can have multiple shortest paths between the same pair of nodes. So, if you had let's say, you can have multiple shortest paths and it is also the immediate other thing you need to understand is k-shortest paths right.

So, this is the shortest path or what you could potentially call the 1 shortest path, 2 shortest path is the second shortest path, third shortest path, k shortest path. So, first k shortest path, this is usually a very useful metric to understand. How many other paths are there between your points of interest? And in the directed network you have a path from A to B, but there is no path from B to A in this network.

Because there is no path from this intermediate node to A, right. So, in a directed network you need to worry about existence of paths in either direction. Just because there is a path from A to B doesn't mean that there is a path from B to A. Whereas, in an undirected network you never worry about something of that sort. The next is the notion of what's known as diameter. What is the diameter of a graph?

The longest shortest path, right, so it is just like what is the diameter of a circle? It is the maximum separation between 2 points that lie on the circle, right. The farthest, right and you won't go across the circumference, right. You will only take the shortest distance. Same logic here, you will take the shortest distance between any 2 points on the graph, right. So, it is the longest of all the shortest distances in the graph.

So, if you actually made a matrix, this is in fact you will end up computing it in MATLAB. This is the all shortest path matrix, right. In fact, we will look at it during the lab session of how do you compute this. So, in the all shortest path matrix you will have obviously zeros in the diagonal, no distance between the node and itself. What do the infinities represent? Unconnected, disconnected nodes that are not reachable from any given node, right.

Other than this, what is the max? That is your diameter, right. So, then there is this notion of characteristic path length.

(Refer Slide Time: 11:36)

What is characteristic path length? It is usually written like an L in this fashion, right, a calligraphic L. It is the average separation between 2 nodes in a graph. So, what is it in terms

of the shortest path matrix that we saw on the previous page? Average of non-zero non-infinity elements. You take all the non-zero, non-infinity elements, average them up, you get characteristic path length.

Then comes the second most, you know, important concept which is that of degree distribution. So, what is the degree distribution? It is essentially a histogram of degrees, right. Your x-axis is of course degree itself and y-axis can be either $N(k)$ or $P(k)$, $N(k)$ is the number of nodes with degree k , $P(k)$ is the probability of finding a node with degree k .

It is essentially a normalised $N(k)$, right. How about this look? So, let's consider a very simple graph to begin with. What is the number of vertices? 5. Number of edges? Completely connected, so 10, $5*4/2$, 10, right. What is the degree of each of these nodes? 4. So, what does the degree distribution look like? Single point essentially, that's it. This is the degree distribution, right.

But you can have different types of degree distribution for different networks. So, it is not uncommon to find degree distributions like this or like this or like this or any other kind of degree distributions. So, you can find different kinds of degree distributions in different networks. So, this becomes an important quantity or you know an important way to classify different networks.

How does the degree distribution look like? So, I can group networks based on their degree distribution. Does it have this kind of a degree distribution, then I will call it some kind of a network. Does it have this kind of a degree distribution, I will call it another kind of network. So, what would you call a network that has a degree distribution that look like this? This one, so just it's essentially empty with one point where with all the another synonym for uniform, regular, right.

A regular graph is a graph where your all nodes are somewhat identical. Strongly connected need not be, well, so first of all this is not a directed graph so you need not bring in the concept of strongly connected, so is this a regular graph? Yes, right, so a regular graph is a graph in which you can basically all nodes have the same degree and so on. So, that's another useful concept.

So, degree distribution is really an important concept. We will discuss it in much greater depth when we look at different types of topologies. The other very useful measure is something known as clustering coefficient.

(Refer Slide Time: 16:26)

The slide content includes:

- Clustering coefficient** (handwritten title)
- A diagram of a triangle with nodes A, B, and C circled together.
- A diagram of a complete graph K_4 with nodes A, B, C, and D.
- Handwritten text: "Cliquishness", "Clique? || sub complete graph", and the calculation $k=4, C_2=1.0, HC_2=4C_2$.
- Another diagram showing two triangles sharing a node, with handwritten text "2 Cliques" and "n(C2)".
- Handwritten text: "C3, C4, C5" and the calculation $\frac{1}{3C_2} = n(C_2)$.
- A small diagram of a triangle with handwritten text $\frac{3}{3C_2} = 1.0$.
- The NPTEL logo is visible in the top right corner.

What is a cluster? So, there is relatively more connection between a bunch of nodes, right. And if you were to like really reduce the definition, it would be, what is the simplest cluster that you can have? This is a cluster, right. So, A knows B, B knows C, C knows A, right, or A interacts with B, B interacts with C, C interacts with A. So, this is a very simple cluster. Clustering coefficient is essentially a count of the number of triangles of this sort.

So, if you look at let us take this network, again. So, clustering coefficient is also described as, what is a clique? Yeah, it is the same as a complete graph or rather a complete subgraph right. For example, you will say that this is obviously not a complete graph, but it has 2 cliques right. So, what is cliquishness? It is how cliqued a particular neighbourhood is, right. So, let us take this node, it's clustering coefficient basically will tell you how much of a clique do you find in it is neighbourhood.

Or in other words, why clustering coefficient actually came about is social scientist started asking this question if I know 2 people, how often do they know each other? I know A, I know B, how often are A and B friends? That is essentially this triangle we are talking about, right. So, now if you look at this network, this node A has, how many neighbours? 4 neighbours, it is all fully connected right.

So, B, C, D, E, so you have 4 neighbours. What is the total number of connections possible between the neighbours? $4C2$ and how many connections actually exist? Not 4, you will find many more right. So 1, 2, 3 yeah, you will essentially find $4C2$, 1.0. So, it is heavily clustered right, but for any, so for this network, you can now see what is the clustering coefficient of this node? What is the clustering coefficient of this node that I have circled? How many neighbours does it have? 3.

So the denominator is $3C2$, numerator is 1, right, because there is, these 2 nodes are actually connected to each other. So, the other way of looking at it is how many triangles intersect at this point, right. So, if you have a network like this you will find that there are 3 triangles right and this actually a complete node. So, we will say $3/3C2$ equals 1.0 again. So, we will look at another example for clustering coefficient when it is a little more interesting.

When we study regular lattices and so on, right, it is very easy to enumerate the clustering coefficient in some of these examples, but it happens to be a very important way to classify networks right. How clustered are particular nodes? So, you will see that important proteins will have a higher clustering coefficient. So, you will find that there are pathways right.

Pathways you can imagine may form some sort of cliques in a protein interaction network. **“Professor - student conversation”** Very good question, we will actually come back to that in a moment right. Once we finish that entire slide, I will come back to discussing about every single metric that we have talked about whether it corresponds to nodes or networks.

Here, it clearly corresponds to a node, but you can obviously average it for a network and there are other ways to average it as well which we will see shortly. Cliquishness is basically, this is the definition right, so this denominator tells you what is going to be the ideal, if there is a clique what is going to be its size and if the numerator is the same as the denominator you will find it is a clique.

So, you can actually say that there are more cliques also right, potentially you can say this is another clique, this is another clique, this is another clique as well. Yeah, but essentially these 3 form a clique, these 3 form a clique, any triangle can be considered a clique, no? So, it's basically what you would call a complete graph of size 3. You can also look for complete graphs of size 4, size 5 and so on, right.

You will probably get a little more understanding of this when we look at motifs in networks, which is a couple of lectures down the line. Of clustering, see it is still basically the same concept, no? In a weighted graph, the notion of clustering doesn't really change right because you only look at how many nodes are connected to each other, you essentially ignore the weights.

I am not sure if there is a weighted version of clustering coefficient, but there are weighted versions of some other centrality measures, but for clustering coefficient I am not really sure, I will have to check. In directed graphs, you can extend the same logic essentially. So how many connections, you can look at connections in one direction, but usually you know clustering coefficients make a lot of sense for undirected graphs.

But I am sure there is you can extend to directed graphs as well. Good, is clustering coefficient clear? Because these are very important ways to study any single network. These are the first things you would do when you get a new biological network. You first compute the degrees of every single node, you then compute the degree distribution.

You then compute the clustering coefficient of every single node, then you may be look at a distribution of clustering coefficients and so on and so forth. What is the characteristic path length? what is the average separation between any 2 nodes on my graph so on and so forth.

(Refer Slide Time: 24:24)

The slide is titled "Recap" and is divided into two sections. The first section, "Topics covered", lists "Why Graphs?" and "Basic Network Parameters". The second section, "In the next video ...", lists "Centrality Measures", "Closeness Centrality", and "Betweenness Centrality".

Recap

Topics covered

- ▶ Why Graphs?
- ▶ Basic Network Parameters

In the next video ...

- ▶ Centrality Measures
- ▶ Closeness Centrality
- ▶ Betweenness Centrality

In today's video, I hope I motivated why we need to use graphs to study biological systems and we also looked at some basic network parameters. In the next video, we will move on to more interesting and more complex network parameters particularly looking at centrality measures such as closeness and betweenness centrality.