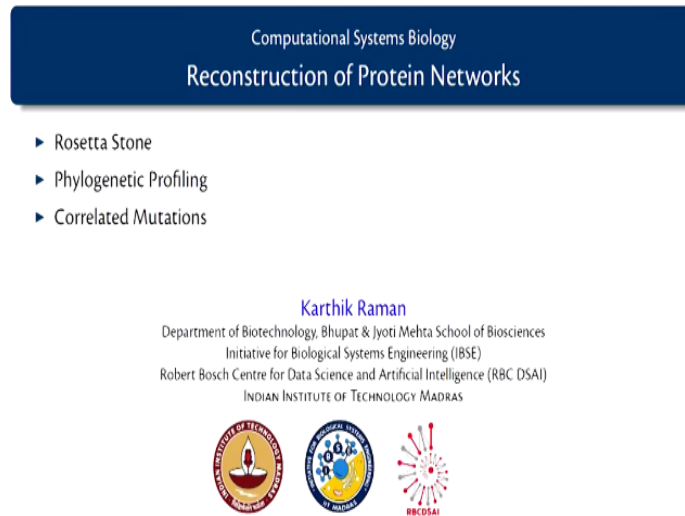


Computational Systems Biology
Karthik Raman
Department of Biotechnology
Indian Institute of Technology – Madras

Lecture - 32
Reconstruction of Protein Networks


(Refer Slide Time: 00:11)



Computational Systems Biology
Reconstruction of Protein Networks

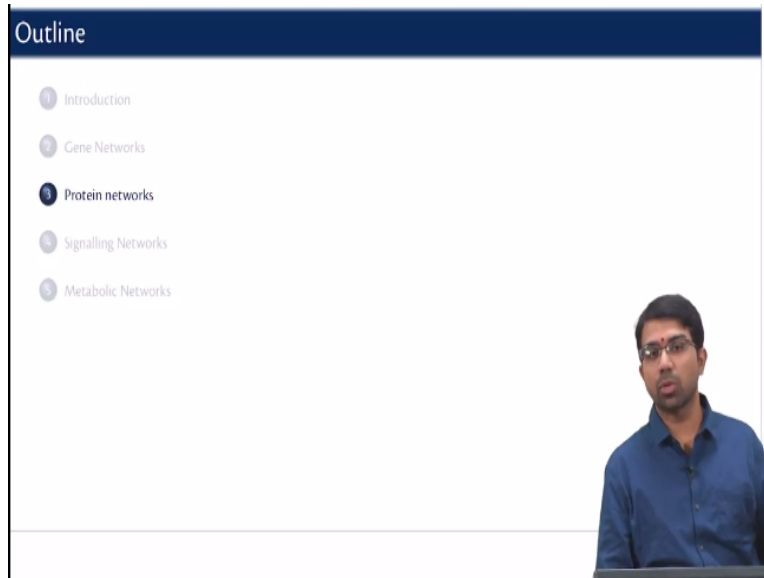
- ▶ Rosetta Stone
- ▶ Phylogenetic Profiling
- ▶ Correlated Mutations

Karthik Raman
Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences
Initiative for Biological Systems Engineering (IBSE)
Robert Bosch Centre for Data Science and Artificial Intelligence (RBC DSAI)
INDIAN INSTITUTE OF TECHNOLOGY MADRAS



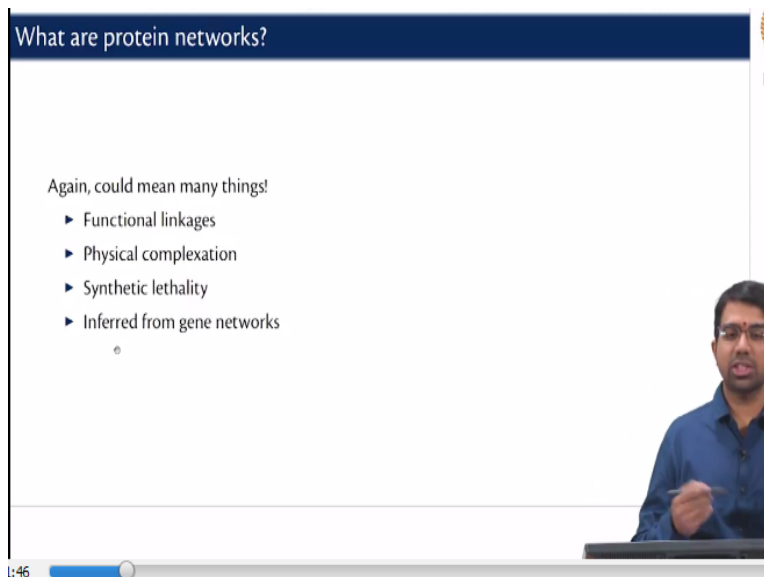
In today's video, let us look at reconstruction of protein networks or protein interactions and there are few interesting methods that are mostly sequence based that have gone into building these kind of networks which is Rosetta Stone, Phylogenetic Profiling and Correlated Mutations.

(Refer Slide Time: 00:28)



So let us look at Protein networks which are very commonly analyzed and often used for predicting different kinds of things so one who uses protein networks predict even essentiality, function and so on.

(Refer Slide Time: 00:42)



So what are protein networks? Again it could mean different things just like it could be a co-expression network or a transcriptional related network or a synthetic lethality network in case of genes again for proteins it could mean just functional association or functional linkages. What does this mean? This means that the function of two proteins is somehow linked, they basically part of the same pathway, they may compensate from one another they might physically interact which is a different thing.

So physical compensation is strong kind of interaction but it could also be just that they are functionally associated. A and B always express together, A and B always finding themselves in the same pathway A and B are always in the same Operon, so maybe you know they are doing some related function, right. You have a Beta Glycoside and lactose important protein in the same operon obviously, to be in Synthetic lethality or it could just be a projection from your gene networks. So you just start with your gene network and projected to the protein space.

(Refer Slide Time: 01:48)

Protein-protein Functional Linkages/Interactions

- ▶ Basis for several signalling pathways and regulatory networks
- ▶ Experimental methods of identification:
 - ▶ Yeast two-hybrid
 - ▶ Affinity purification/Mass spectrometry
 - ▶ Protein microarrays
 - ▶ Further reading: Shoemaker BA & Panchenko AR (2007a) *PLoS Comput Biol* 3:e42+
- ▶ Computational methods of identification:
 - ▶ Domain fusion
 - ▶ Conserved neighbourhood
 - ▶ Phylogenetic profiles
 - ▶ Co-evolution
 - ▶ Further reading: Shoemaker BA & Panchenko AR (2007b) *PLoS Comput Biol* 3:e43+
- ▶ Databases:
 - ▶ DIP, STRING, HPRD, Predictome



So protein interactions, proteins you must have all read it a multiple time that proteins are the work hours in the cell and all functions everything is mediated through proteins and particularly protein-protein interactions. So every Gene Regulatory Network actually at heart is a protein interaction network or a protein is technically speaking a protein DNA network but you can always think of it is a gene, gene network or as a protein-protein network.

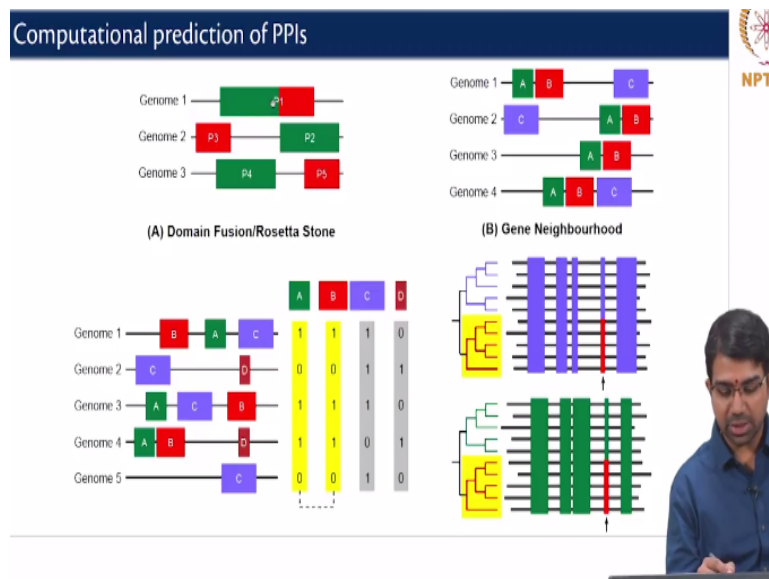
The very interesting signalling pathways that relay on different kinds of protein interactions. There are few enzymes, there will be phosphatases, kinases and a complex network of these along with small molecules which gives you very interesting signalling cascades which have which can be very tunable as well. So the response of a signalling cascade can vary across several orders of magnitude depending upon the architecture of the signalling pathway itself.

And many experimental methods to identify these protein interactions and a lot of new techniques have been developed which are very high (()) (02:50) and so on. Like Yeast two-hybrid, Affinity purification, Mass spec of course you can also have Protein Microarrays and so on. This paper is a slightly old paper now but it is still a very good paper it gives you all the experimental methods for studying protein interactions, okay.

But in this course we are more interested in computational methods and there happen to be many computational methods relying on different aspects typically based on just the genes sequences because they are abounded. So can you infer protein interactions based on operating functional associations, based on the gene sequences for an organism. We will see how that is done. And there is a lot of databases available.

Database of interacting proteins, human protein database and the string, we will look at all these in a later class where we will try to study what are all the different types of databases which are very helpful for building, studying and analyzing these kinds of networks.

(Refer Slide Time: 03:53)



So how do you computationally predict protein-protein interactions? There are many interesting methods to predict protein-protein interactions and these are all essentially based on Genome context or genome sequence. The first of them is what is called Domain Fusion or Rosetta Stone.

So you may all have heard a Rosetta Stone which was popular archeological discovery, right. So there was this particular stone which several encryptions which were discovered.

This is a spin on that the name. So the scientist basically predicted that when you have a protein that exists as polypeptides two separate polypeptides and two from genomes and then few as single polypeptides in another genome it maybe because they are likely to interact. Okay. So this maybe a lower organism where you find two separate proteins P2 and P3 or P5 or P4 right. But then some other organism perhaps the higher organism you find them fuse together in a single polypeptide.

This maybe two domains in some other cases but they are fuse together in this case. So this is Rosetta Stone wherein you have two polypeptides that are present separately in any organism that can be fuse in another organism. Why? It could be that in a simple organism one argument is it in a simple organism these proteins are likely to be closed to one another, right so the volume is very small.

But in a larger organism the cell volume is so large that if you want P2 and P3 to interact the probability becomes very low. So if you want them to interact you might have them stick them together, right. So the two domains are fuse so then now they are already easy to build on those interactions. The next concept involves gene neighborhood. So what we find is that some proteins share the same neighborhood.

So if we see A and B always occur in the same neighborhood. C may or may not occur with A and B but when A occurs B always occurs somewhere close to it. So they could be in the same operon or whatever. So you find that there is proximity in across several genomes. So remember whenever I say here genome 1, genome 2, genome 3 and so on, there are different genome sequences. So you find a particular stretch that is common in different organisms.

So A and B occur together in organism 3; A and B occurs together in organism 2 and they also occur together in organisms 1 and 4. So based on this strength you may say that A and B actually interact. Another very interesting concept is that of Phylogenetic Profiles and this was this study

actually was performed in 1999 with only 17 sequence genomes. And they were already able to find map functions of several equally proteins and so on.

So what—how does Phylogenetic Profiles work? So—consider this as genomes these are the genome sequences and these are particular polypeptides chains in those genome sequences. So a profile is basically nothing but the presents or absents of a particular protein in augment sequence, right. So you have in this case A is present in genome 1, absent in genome 2 so you put a 0, present in genome 3 you put a 1, present in genome 4 you put a 4, absent put a 0

D on the other hand is absent-present-absent-present-absent, right. And B has the same Phylogenetic Profile as A; whenever A is present B is present so you have a 1. Whenever A is absent B is also absent. So maybe this A and B is form a pathway that is only required together, I mean they are part of pathway so they, they will only function together. If A is not there, there is no rule for B as well.

You could have an alternative scenario wherein you can have complimentary profiles. We do not have an example of that sort here. But if you see in at least a few cases if you look at these three genomes, genome 1, 2 and 3 whenever A is present D is absent whenever D is present A is absent. Okay. So maybe they are alternate proteins that can do the same function and so on. So you can essentially infer protein interaction or functional associations based on several genomes sequences.

But they would not be a synthetic lethal part definition because they have to exists together right. They are functionally equivalent so in that sense there is a similarity synthetic lethality because in at least some cases we expect that if A and B are doing the same function; if you remove A, B will compensate, if you remove B, A will compensate but if we remove A and B the organism might die, so that argument might be extendable here but the definition we are saying that only A exists but not B exists when you are saying talking about complementary profiles.

“Professor – student conversation starts” But genome 5 (()) (09:24) so A and B both are there the organism is existing. Yeah. So I am not saying that you can infer that, this is actually weak

evidence, right. **“Professor – student conversation ends”** So here D and A are not correlated so there A, B and D are present. So I am not given an example of complimentary Phylogenetic profiles here. But you can essentially think about different kinds of interactions that can be there. And you may also want to see the strength of the signal, right. So in how many of the genomes is this present?

Do you want an absolutely 100% of the genomes, you have-- today you have 1000+ sequence genomes, so do you want this the profile to be the same across 1000 bits, so if you look each of it as a bit—this is a bit vector in 101010 so how many of these bits are matching. So you may just compute a threshold and say as long as a 90% of genomes or 80% of genomes I find a similar profile I might consider these proteins to interact.

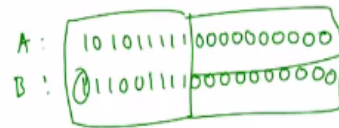
So Phylogenetic Profiling is a very powerful method to infer protein-protein functional associations. **“Professor – student conversation starts”** So but these genomes for different organisms? Yes, yes different organisms. So there might be organism in which sequence does not exist also? Yes. So then get a 0. No, no not anything, if B, A, C nothing is there. Fine. But you are only looking at pair-wise right.

You look at A and B and pair-wise you see that this is 10110 this is also 10110 it could have several 0s this could be paired with another five 0s in the end and you will still say that they seem to always co-occur. My question is, if that entire sequence is not there so all the four will be 0, if you are going to pair five more 0s with all A, B, C, D 0s you cannot tell that is correlated, right?

Yeah, you will have to yeah so if I added another 100 genomes with none of these proteins existing so you can so you will say that all B is absent yeah that would not be a strong signal so you have to do a statistical correction and so on, right certainly. So it may not be easy but like. Where is lethality come from here? So those are (()) (11:51) right so.

Yeah, yeah fair enough but what he is saying is just having 0s is not good enough to infer Phylogenetic profiling you know, like the functional associations. **“Professor – student conversation ends”**

(Refer Slide Time: 12:10)



So in other words for A the phylogenetic profile infer B it is, we will say this vary only two bits so these are likely to interrupt, right. But, now would you still say that they will interact? Yes, it means you have a stronger evidence but it is all negative evidence, right. So it is not very interesting that way whereas up till this point you have relatively strong evidence, here one will say this evidence is somewhat diluted. So you will have to actually carefully compute this.

And this also need not be 1s and 0s you can have extent of similarities, right so 0.9 what is the closest match you typically look for something like a bi-directional, best hits so on. All these are essentially pointing towards the presence of orthologs. So if you look up your basic Evolutionary biology so these are all trying to indicates orthologs. So last of these is quite interesting. This— what is this look like dendrogram, yes from what we did in previous classes.

But it is also essentially a tree of life, right. So it tells you which two sequences are related to each other and so on. So what is this look like? If each of this is a sequence this is a cartoon representation of a multiple sequence alignment. You must have seen many of these right

wherein you find that this particular protein is residue is conserved this bunch of residues is conserved so whatever is color bend in this picture they all represent of conserved residues, right.

And what you find is there are two such proteins, right and two such families of proteins so 1,2,3,4,5,6,7,8,9,10 sequences in family 1, 10 sequences in family 2 and they are multiple aligned, right. But what you find is that whenever you have a mutation in this family there is a mutation in this family as well, mean to say there is a correlated mutation between these two families.

So the assumption the forecast is whenever these proteins makes a small change this protein makes a corresponding change so that it can continue to interact. So it is like sort of the lock and key. So whenever there is a small change in the lock the key adapts to the lock, right so to speak. So you have correlated mutations that happen in two conserved families, two families of proteins, so now your hypothesis that, well these proteins are likely to be interact.

Maybe this is some enzymes this some enzymes which was you know and therefore they both interact. So these are together different ways to infer edges between proteins in a protein functional association network. So you may some evidence from domain fusion, you may have some evidence from gene neighborhood, you may have some other evidence from Phylogenetic profiles and you may also have evidence from correlated mutations or from other aspects.

You could do text mining; you can have experiment data. How do we integrate all of these that something? we look at in the next class.

(Refer Slide Time: 15:54)

Recap

Topics covered

- ▶ Rosetta Stone
- ▶ Phylogenetic Profiling
- ▶ Correlated Mutations

In the next video ...

- ▶ Basic Concepts
- ▶ Reconstruction

So I hope you had an overview of some of the interesting concepts underlying the reconstruction of protein interactions basically method such as Domain Fusion or Rosetta Stone, Phylogenetic Profiling, Correlated Mutations and so on. In the next video, we will move over to Signalling networks where I will give you some basic concepts and talk to you about how we go about reconstructing signalling networks.