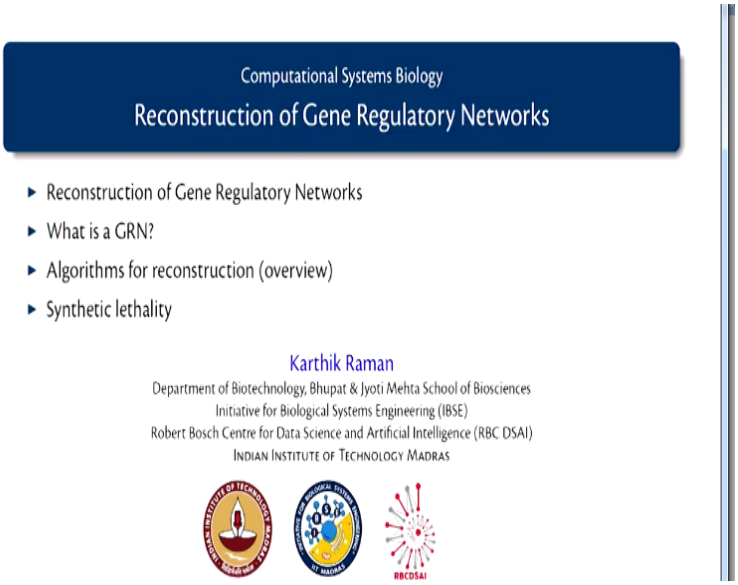
Computational Systems Biology
Karthik Raman
Department of Biotechnology
Indian Institute of Technology – Madras

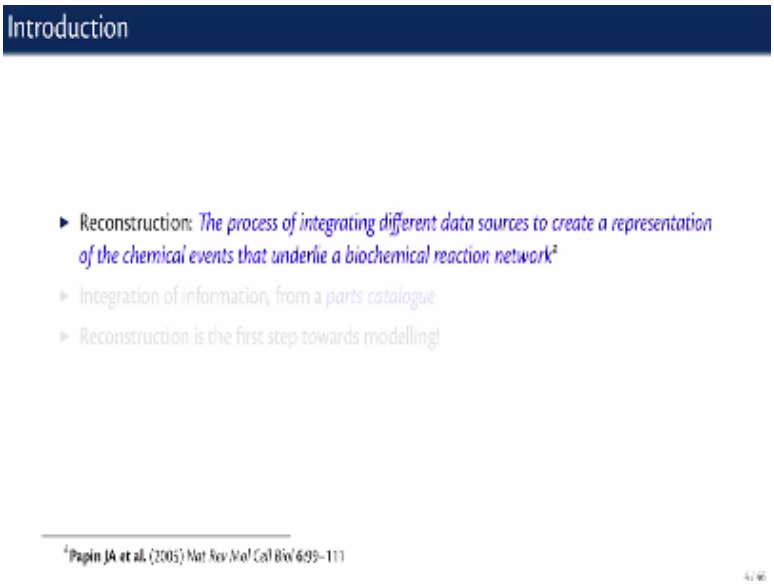
Lecture - 31
Reconstruction of Protein Networks

(Refer Slide Time: 00:11)



In today's video, we move onto the concept of reconstruction or building these biological networks and the first network we will look at GRNs or Gene Regulatory Networks. I will give you an overview of some different algorithms that exists for reconstruction but we will not go into the details. I will also give you some details on the concept known as Synthetic lethality.

(Refer Slide Time: 00:36)



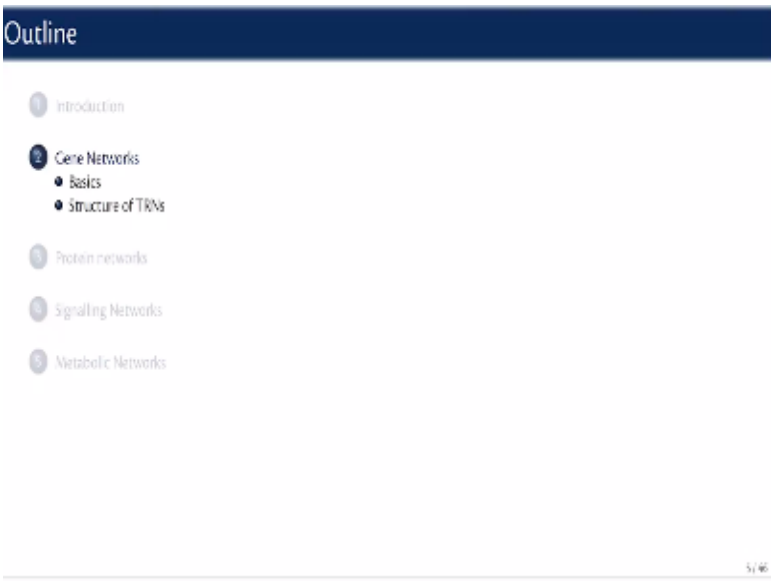
So what is reconstruction, it is the process of integrating different data sources there are lot of different types of data sources there are actual biochemical experiments, there are several published manuscripts papers on these kinds of topics, there are other existing databases and so on. So how do you integrate data from all of these different sources to build a single network? This network basically catalogs all kinds of events biochemical events but underlay any cell.

So you have, so a cell is nothing but a sort of a consorted mixture of all these kinds of networks, right. So there are several chemical events you may signaling, you may have gene regulation all of these involved finally at the basic stage some level of molecular binding, unbinding recognition and so on. So this essentially involves integration of information from a parts catalogue. So the parts catalogue is nothing but your genome itself, right.

Your genome has all the parts; you can identify all the genes all the-- a lot of proteins from the genome. Then how are these wired together? How are these parts tide to one another? that the process of you know, that if you will edge inference is reconstruction. And it is typically the first step towards any modelling exercise, particularly the kind of modelling we looked that in the recent lectures.

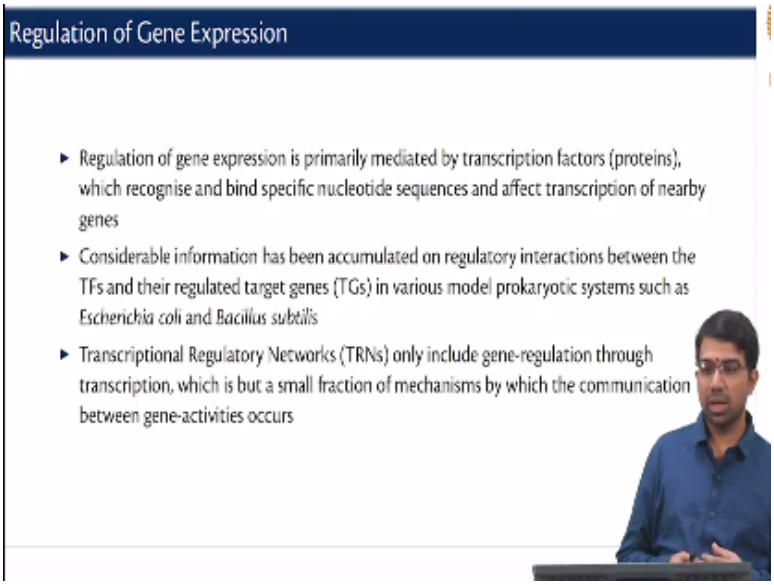
Of you construct a network then you study several properties of the network, what is the degree distribution, what is the characteristic path length; how does may network behave so on and so forth, okay.

(Refer Slide Time: 02:09)



So let us now first start off with Gene networks.

(Refer Slide Time: 02:12)



So the expression of gene if you recall your biology is controlled by other genes, right there are transcription factors that will bind to DNA that will help or prevent the binding of other proteins, polymerase so on and so forth. So you have a specific flow of information in biology, you have

DNA to RNA to protein and this process is facilitated by several enzymes and several other helping molecules such as your transcription factors.

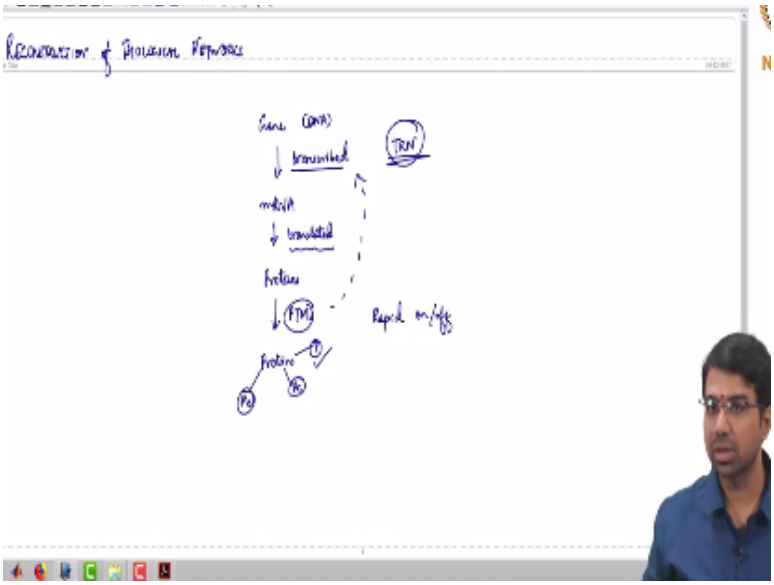
And the regulatory relationship between these kinds of genes form very complex networks and these are usually elucidated through various kinds of experiments like Chromatin immunoprecipitation, ChIP-seq and so on. Most important thing is these transcription factors will bind to particular places on the DNA, right. So these are sequence motives, right. So they will bind to an AT, AT, A something of that sort, a particular sequence stretch on the DNA.

And reconstruction involves identifying all these regulatory relationships. So who interacts with whom and what is the nature and strength of this interaction. So it is essentially trying to identify the edges, the direction of the edges and the weights of the edges. So these are the three things that you want to build in any reconstruction.

So how does regulation of gene expression work, it is primarily mediated by transcription factors which are essentially other proteins or other genes as we will generalize which bind to specific nucleotide sequences and affect the transcription of one or more genes, right, you may have one or more genes in an operon which will be affected, right. And there is a lot of information accumulated about these regulatory interactions between the TFs and their target genes.

In different kinds of organisms these are available across different databases where is transcriptionfactor.org which is a particularly popular database. And we will look at different types of these gene regulatory networks. The most commonly talked about would be the transcriptional regulatory networks, which involves just the transcription regulation, right. So if you look at the flow of information in biology--

(Refer Slide Time: 04:33)



You have gene, let us call this DNA, this is transcribed to mRNA, translated to proteins and then you have post translation modification where your proteins could be phosphor related, assert related and so on and so forth. Typically, phosphor relation is the most important operation, right. So then you must have heard the central dogma and so on about how information flows predominantly in one direction but occasionally can also flow in the opposite direction, right.

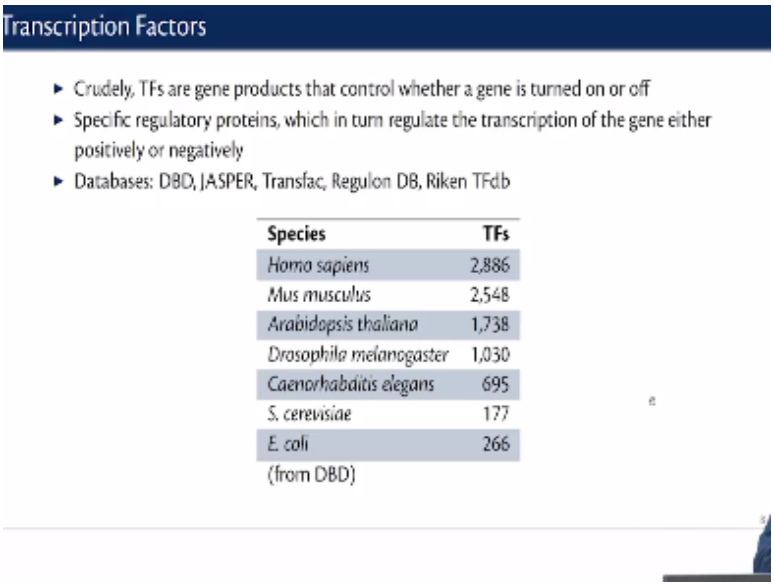
You can control exercise control at any of these steps, right. When you exercise control here that is transcriptional regulation or TRN, you can also exercise control in any of the other steps. This is not that common translational. This is very, very useful. This is used for Rapid on/off. Phophor relation is excellent switch and this is the most common device if you will in signaling as we will see shortly, right. So transcriptional regulation essentially looks at this part.

And the--we will look at other concepts such as you know how there is amplifications that happens in the response and so on at a later stage. But in general there are various levels at which a cellular network can be regulated and particularly important is the first part which is transcription. So how does how does—how is gene regulation regulated, gene expression regulated, right.

Regulation of gene expression is basically you know mediated by this transcription factors and we mostly look at Transcriptional Regulatory Networks which include only regulation through

transcription. But remember this is small fraction of mechanism by which you know variations in gene expression can occur.

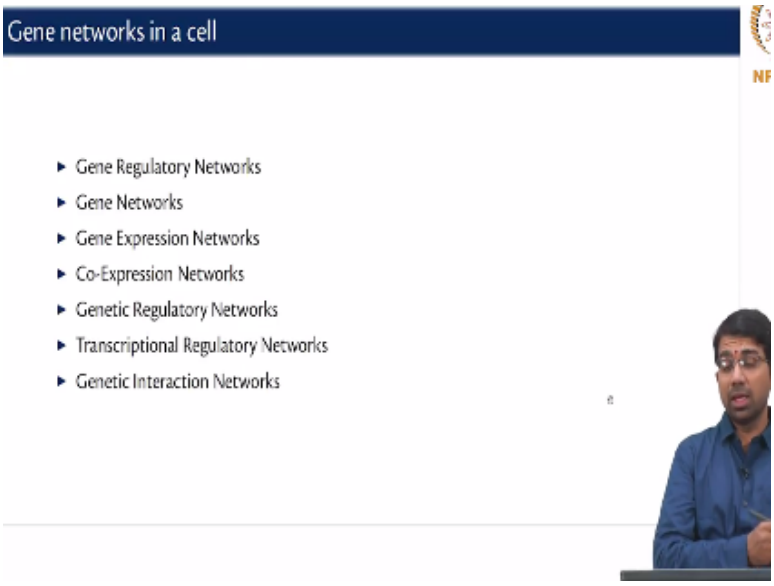
(Refer Slide Time: 06:58)



What are transcriptional factors? Crudely, they are genes or proteins they are actually proteins but you can generalize them to genes as we will see in the moment that control which gene is turned on and off, right. They can either influence a particular gene positively or negatively. It can activate the transcription of another gene or inhibit, repress the transcription of another gene. And there are many databases such as DBD, JASPER, Transfac, Regulon and so on.

And these are some statistics which are you know, for Homo sapiens for humans you have nearly 3000 transcription factors whereas in E.coli you have < 300 transcription factors. But the thing is if you look at a gene interaction gene regulatory network you will find that most of these transcription factors are Hubs. They will influence the function of several other proteins, right. There will be a distribution but in general you will find that they influence several other proteins.

(Refer Slide Time: 07:58)

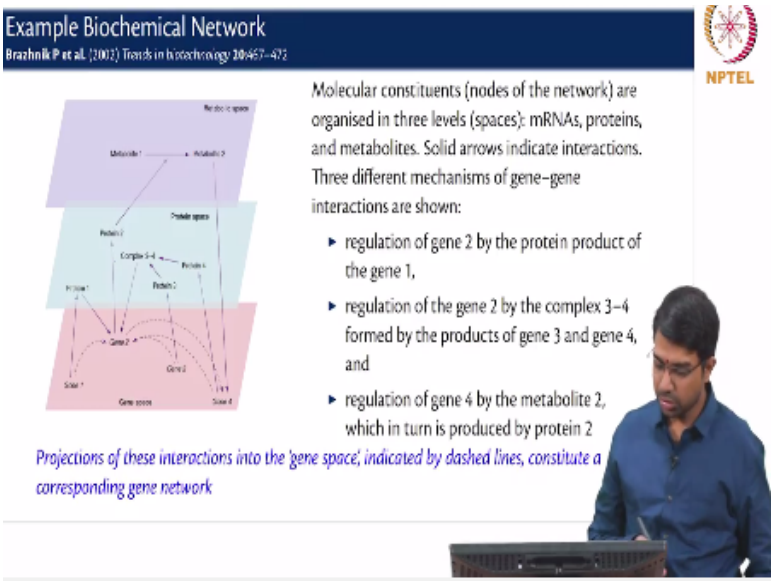


What are the different gene networks within a cell? So typically, you find all these terms being used in various papers and you will have to rap your head around these, so you may find all of these are essentially different nomenclatures for gene regulatory networks, so may you call these gene regulatory networks, gene networks or gene expression networks and so on, but more specifically is that if you have a right paper you should try to make sure that you carefully (()) (08:26) it, what is it that you talking about.

Are you talking about some just genetic interaction or are you talking about transcriptional regulation or are you talking about co-expression, right? So you can imagine that gene co-expression network will connect genes that are expression together under one condition or under many conditions. Genetic Regulatory is a little more generic turn, transcription regulatory is very specific, right.

So genes that transcriptionally regulate another gene will find the place in this network, and network genetic interaction networks.

(Refer Slide Time: 09:00)

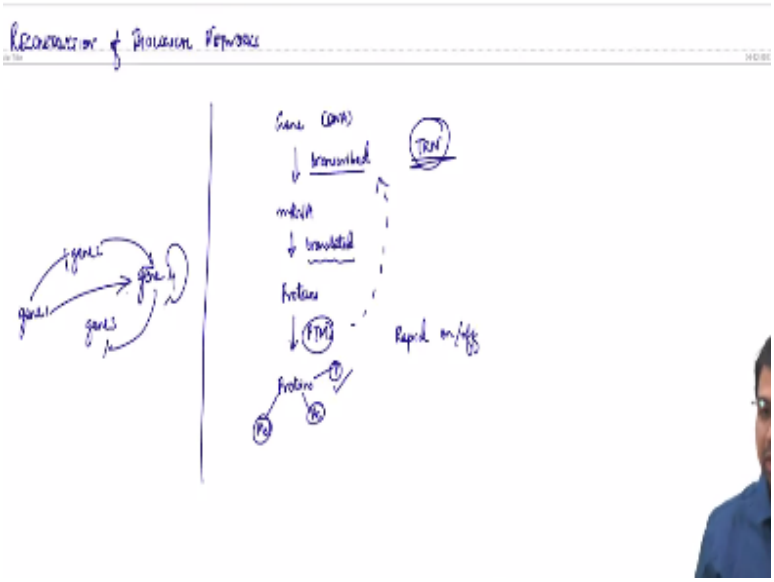


So let us look at a simple biochemical network here. So you have a bunch of genes, so you have gene 1, gene 2, gene 3, gene 4 this is what you can call this the genes space and then you have the proteins that is genes encode for followed by a metabolite reaction and enzymatic reaction that protein to catalysis. So you have you can think of these as either genes or there mRNAs. What are the different interactions shown?

The regulation of gene 2 by the protein product of gene 1. The regulation of gene 2 by the complex 3, 4 which is in turn found by gene 3 and gene-- protein 3 and protein 4 which in turn comes from gene 3 and gene 4. And then the regulation of gene 4 by metabolite to which is in turn produced by protein 2. Protein 2 is the enzymes that involved in the production of metabolite 2 which has some sort of an effect on gene 4.

It could be allosteric inhibitor or whatever and so on, right. So these are the different kinds of the behaviours that you see here, okay. So what is the gene network now? You will have to project all these interactions that happen in the metabolic space, the protein space into the gene space. You finally want to draw something like.

(Refer Slide Time: 10:28)



Gene 1, gene2, gene 3, gene 4 and you want to basically say something like this, something like this, something like this and so on. What is it really, it looks like this. So you project all these interactions to the gene space which constitute the corresponding gene network which will look somewhat like this in this case, right.

But how does it different from differ from many of the other networks we saw, there are self-interactions, right. That is because most genes will have a negative effect on their own concentration because of the mRNAs degradation rate, so higher the concentration faster the degradation. So how many nodes Gene is a node, how many edges you normally would have had n choose to instead you have n squared here.

Because the diagonal also has, so the diagonal will carry all -1 in this case in some sense, right. Or in fact your numbers in the matrix can be any number between 0 and 1 or -1 and 1. -1 would be perfect repression if you just consider a normalized case whereas +1 would be perfect activation, 0 is no interaction at all, right that is how your matrix might look at for a gene regulatory network.

So basically the level at which a gene is present is going to you know higher the level of a gene it is going to get degraded faster. **"Professor – student conversation starts"** Transcribing very

fast. No, even though it is trying to transcribe very fast it is going to—the mRNAs is going to degrade very fast so because of that the rate of the protein is going to come down.

So a Tb at a particular rate, so the higher the mRNA higher the protein production but faster the degradation so it will quickly so it will kind of pull to a sort of intermediate states. It is like a negative feedback, okay. So this is how the final network looks like. **"Professor – student conversation ends"**

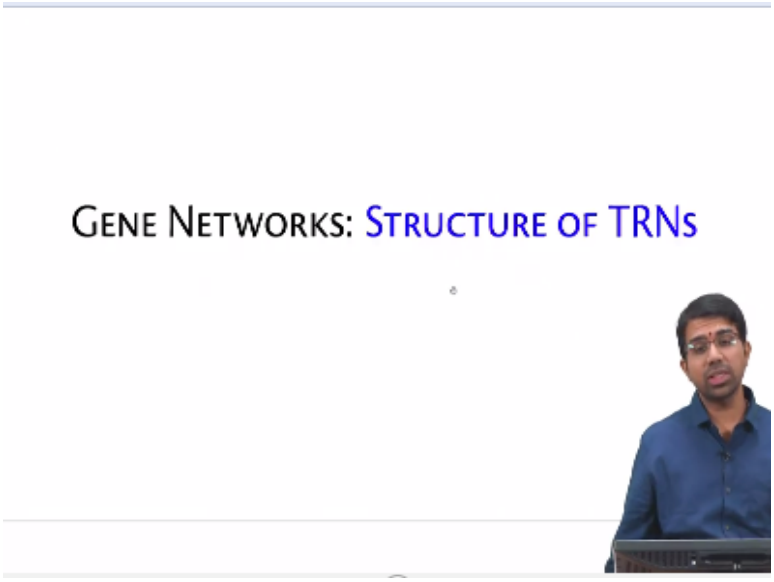
So gene 1 inhibits gene 2 that is because you project this, right and gene 4 activates gene 2, it is actually through a complex mechanism or whatever but still you just projected to this gene network. You want to finally build these gene networks and study them. So this ending with - sign inhibiting and ending with normal RO it like activating, that is in somewhat in tune with SBJ that we discussed few classes, few lectures ago.

No, no so every protein every gene is we are assuming is corresponding to one protein, you may have complexes in that case you will have a little more complexity, right. Here you have complex 3, 4 right but that is actually the complex 3, 4 is going to have a function, so together 3 and 4 have this function so you basically have these edges, right so 3 and 4 are you may want to actually split this by putting 1 more intermediate nodes and connecting it, right.

Because if you remove 3 then the action of 4 is also going to be affected, these are not truly independent adjust. But you will always find some sort of such approximations that exists in any of these networks. You cannot really embed the entire interaction truth into a sort of network formalism.

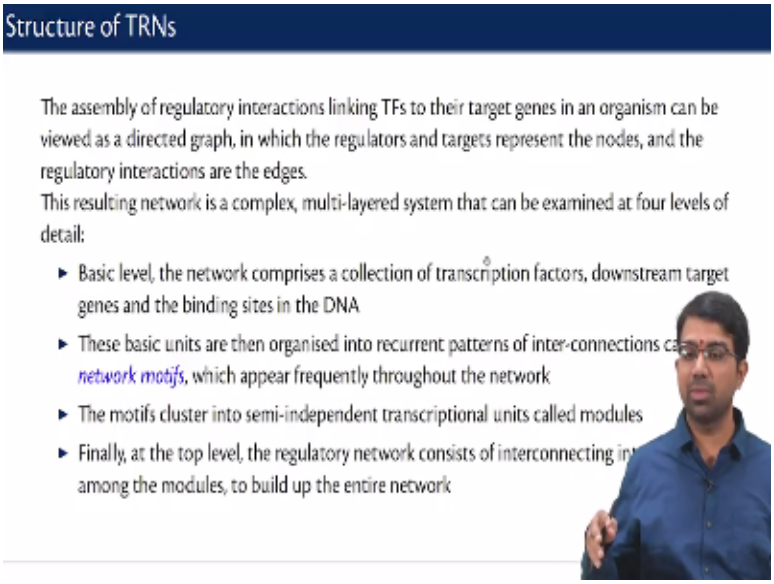
"Professor – student conversation starts" Gene will not be the total number of nodes, right? Fair, but it is the information that you encode right, what is it that, that you want to use it for finally? There are no rules, right. There are no rules, there are no-- so you have to make sensible assumptions if some of your assumptions are too simplifying to, you know detrimental to the modelling exercise, fix them and move on. **"Professor – student conversation ends"**

(Refer Slide Time: 14:21)



So how do this transcriptional regulatory networks look like. What is the structure of these transcriptional regulatory networks?

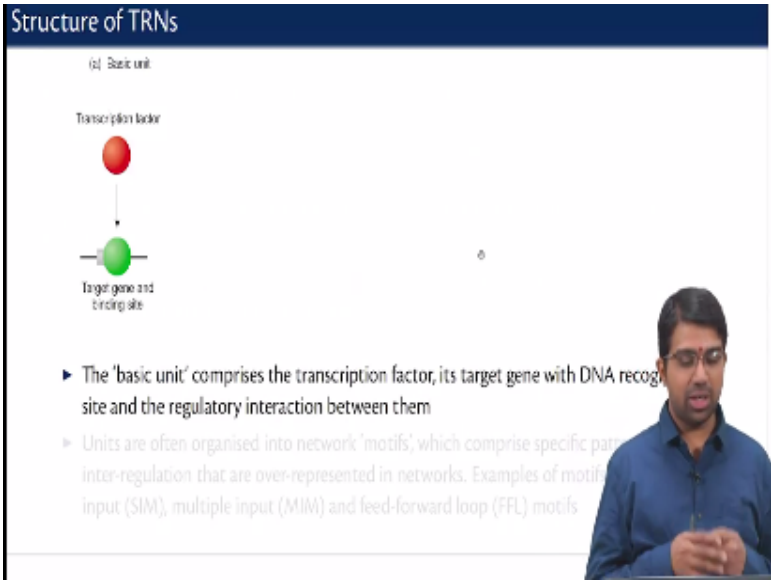
(Refer Slide Time: 14:32)



So obviously task them all into a directed graph where you know every node you know either a regulator or is being regulated a target, right. So the-- and it is essentially a multi-layered network and you will see how this maps back to what we studied in the previous classes about motifs and communities and so on. At the basic level you have edges, right. You have different transcription factors there and there downstream targets.

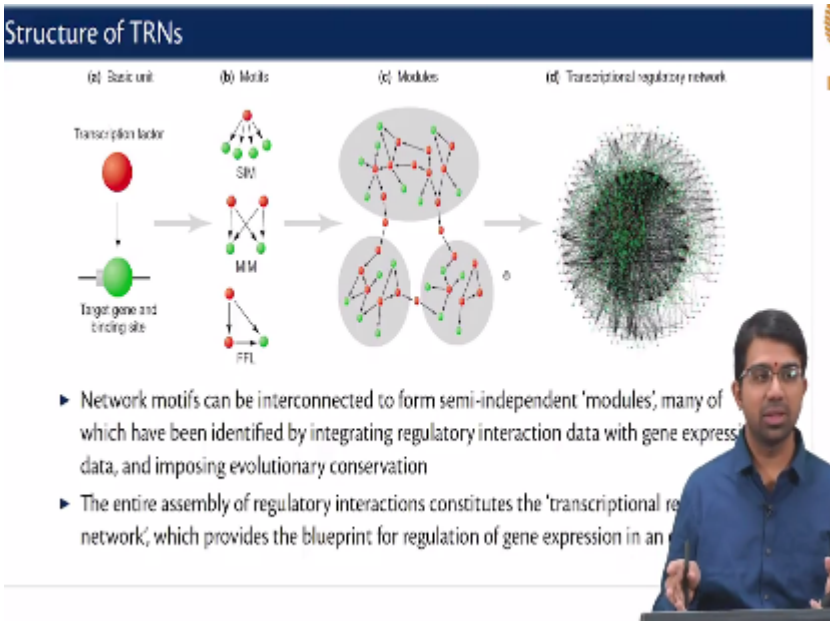
And then they are organized into motives which make sense, right you have a few edges that form motifs. The motifs themselves cluster into independent transcriptional units, right or different kinds of modules, right. Then, finally at the top level you have a really complex and regulatory network with a particular distribution of these motifs the interactions and so on. Finally, orchestrating this complex regulatory behaviour that you find in any living organism.

(Refer Slide Time: 15:35)



So how does this look? This is your edge, so you have a transcription factor that signals that basically affects a signal binding side and then you have motifs, right so I have feed-forward loop, multiple inputs, single input and so on, right.

(Refer Slide Time: 15:52)



And then these motifs are organized into more complex modules and finally into a really complex network, right. This-- the diagram that you are more use to seeing but if you bisect this you will be able to find such interesting modules and finally some motifs. Of course allying analyzing all of these is individual regulatory interactions. And those are the biological becomes very important.

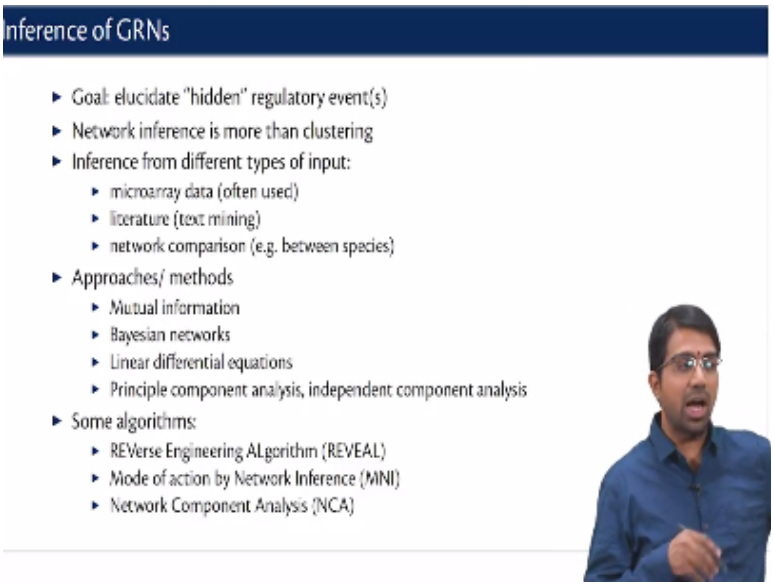
Some biologist done lot of work trying to identify these edges but this is something a biologist may not be able to see your analyze, right. So this is where the different behaviours emerge. These are this is the final reductionist view, right you have single interactions. All these single interactions together make up for a very complex network which cannot be understood just in terms of these simple interactions.

Note, go back to the first class where we talked about the seven blind men and the elephant, right. Because if you only look at this part of the network you may not be able to identify many of these interesting cross talks and behaviours. And this becomes very, very crucial in any network the momentous network, crosstalk and interactions cross interactions become very crucial.

But more so in a biological network where the networks are really large and you have such very interesting emergent properties, right that you do not find in a smaller network or when you have

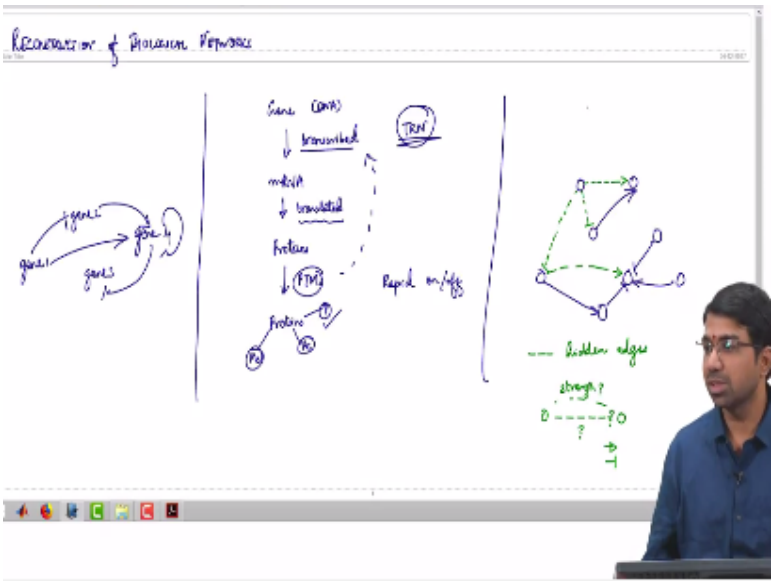
such isolated networks. So this is one more reason to say why we need to look at a system picture to study these complex networks.

(Refer Slide Time: 17:29)



Inferring gene regulatory network is a very big field of studying itself. One could have a course on inferring gene regulatory networks. Usual goal in inference is elucidate what we call hidden regulatory events is just another way of saying--

(Refer Slide Time: 17:44)



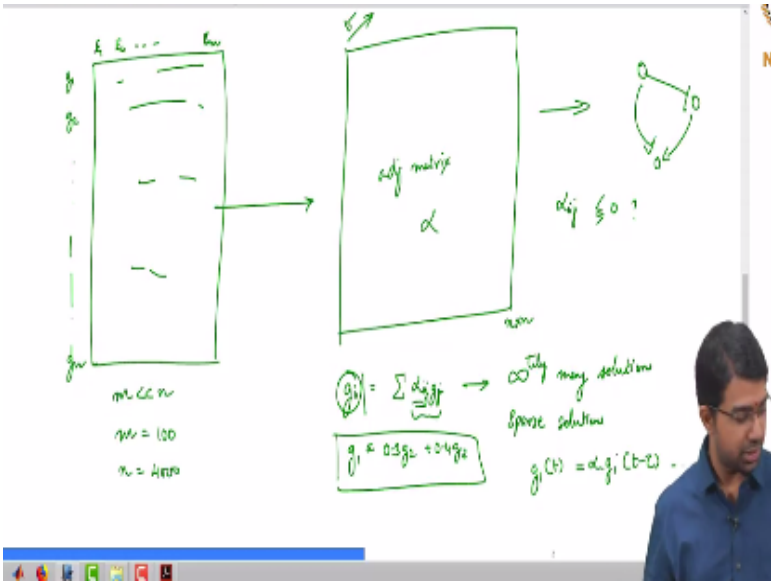
Let us say these are all genes, more importantly also may have such hidden events, right, hidden edges, right. So these edges may basically not exist or you may not know them ahead of time, these maybe experimentally elucidated interactions, but then you may further infer based on how

the network behaves and so on that these edges must also exist. So this is the process of regulatory network inference and this is a tricky process, right.

And there is a lot of study by say chemical engineers, electrical engineers, biologist from different fields of engineering and science but the trick is, you have two nodes that is fine. First thing you have to infer as I was telling you is, is there an edge then what is it look like? Is it a + or is it a – and what is the strength. It is becoming very difficult to infer some of these things. There are many simplifying models but you know there are many interesting ways to determine some of these edges.

We will look at how—for example it is easier to construct protein interaction networks, but one common task is to construct a gene regulatory network, right.

(Refer Slide Time: 19:49)



A typical thing you will have is a matrix. So this is going to be g1, g2, gn and this is let us say expression 1, expression condition 2, expression condition n. It is a very classic matrix that you will get, right a gene expression matrix. From this you need to infer essentially an n cross n adjacent, adjacent symmetrics, right where you know let us call the alpha so your alpha ij you have to first find out whether it is < or > 0 or = 0, right.

First of all, whether does it exists then it is positive or negative and what is the number exactly, okay. So one simplifying way that has been studied in the past is you basically write out gi the expression level of the ith gene as some weighted function of, right. So you just say that every gene the level of every gene is some in this case of very simplified linear function of other genes.

So you can have any, so this is the matrix that you want to infer then you can say it is equal to alpha ij*gj, right. So the expression level of a gene in a particular condition and so on. So how do you infer this? Obviously, you might have infinitely many solutions. So one way is to look for sparse solutions and so on. **"Professor – student conversation starts"** Can you imagine the problem with sparse solutions?

A sparse solution would say that basically the function of one gene is influenced by only few genes but you know that there are hubs in these networks and so on. So that may not be entirely accounted for when you go in for a sparse solution. So how do you infer this becomes quite a tricky challenge, okay. But it is typically a data driven inference. So you have a lot of data and how do you infer the interactions. This basically finally translates into this right.

(()) (22:55) It is always, you see, you usually work with so little data, right so it is invariably very, very underdetermined. So gene 1 will have (()) (23:09) in the matrix. Yeah. So sometimes and n is typically maximum. So typically, so let us usually a very reasonable number would be even for a say equally m would be a 100, this is already asking for a lot, n will be 4,000. Okay. What are m, mn?

So n is the number of genes and m is the number of conditions. Have gene expression different conditions. And sir what is gi, that value is from this matrix or adjacent symmetric? So gi could be the level of expression of or this could be even you know in some conditions or the average expression of gene I across the m conditions, something like that. But then you can say that it is related to in some linear fashion to other genes to the expression values of other genes.

So you could say something like g1 is 0.3, g2+0.4 g3. (()) (24:26) Something like-- this is a very simplified module; you can have many more complex models but this is a very fundamental

simplistic model. There are many other algorithms I will not get into details even we have not even looked at this in great detail but this is essentially one way to go about an assuming how the genes interact, right. We assume that the level of every gene is some linear combination of the levels of some other genes. **"Professor – student conversation ends"**

We basically assuming the entire network is hidden, I do not know any alpha here, I do not know any alpha here, so I solve this equation, okay. This is going to be a very difficult equation to solve so you will have you need to have access, you could have another temporal access, right and then you could even say g1 of t is some g1 of t-tau and so on so you can have, you can have any number of complexity and this is a very intense field of research, I am not going to get into the details.

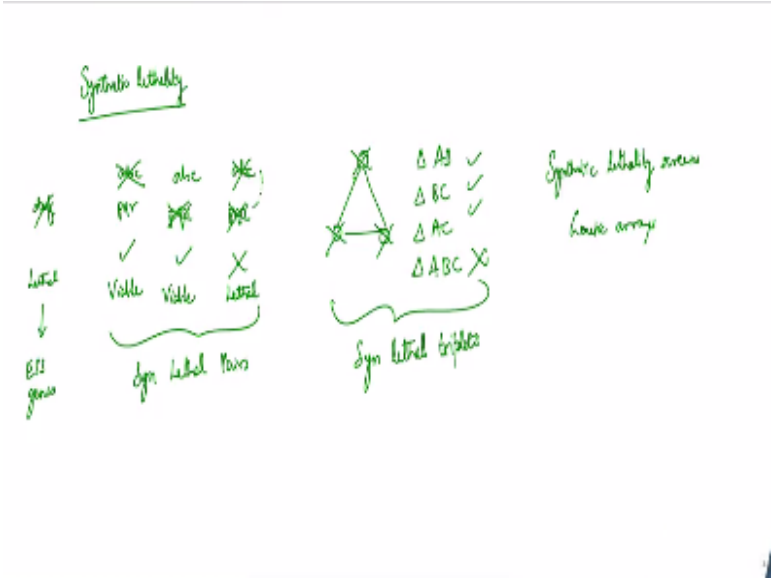
But basically know that there are many algorithms to reconstruct networks and this would be useful for your final project and so on. If you want to because abounded gene expression data, right you have gene expression mRNAs data for several organisms today from different types of databases. There is a gene expression (()) (25:52) which is now essentially the only big database are used to be stand for microarray database that is sort of now shutdown it is deprecated.

So the gene expression (()) (26:02) is one source for huge number of such gene expression datasets. So you can infer you can start with clustering the data, right. So clustering might give you some insights into what kind of genes to work together and so on. But that would be the simplest form of gene expression data analysis. But then you may want to go ahead try to use network data or text mining or comparing networks across species, right.

You may have some known networks in some species. Can you compare your species with that? and identify there are going to be possible adjust that is going to be conserved from that species to yours and so on. There are many other approaches that based on Mutual information networks, Bayesian networks, Linear differential equations and then of course you seeing several dimensionality reduction methods and so on to find you know what is the effect of one gene by the others and what are the top five genes affecting by gene something like that.

You can start building those kinds of analysis. So there are some algorithms called REVEAL, MNI, NCA and so on which are popularly used. The other form other interesting set of interactions is something known as Synthetic lethality.

(Refer Slide Time: 27:18)



So let us consider a cell which has two genes ABC and PQR. When you delete ABC nothing happens in the cell. When you delete PQR the cell still grows. When you delete both ABC and PQR so this means that there is some interaction genetic interaction between these two genes, right. So maybe they were both producing the same metabolite if one was there one if ABC was there the cell could still grow.

Instead if when you hit ABC, PQR could compensate for it or vice versa if you hit PQR compensated for it. But when you hit both the cell can no longer survive. You can also have higher levels of such interactions. So you remove this nothing happens, you remove this nothing happens, when you remove this also the cell dies. Right, you remove any pair, you remove AB-BC or AC still survives. When you remove ABC all these genes a cell dies, okay.

These are higher levels of synthetic lethality. This is a very useful tool to study interactions, genetic interactions in different organisms you have typically you have some synthetic lethality screens basically big genetic arrays to experimentally study, so they called synthetic genetic

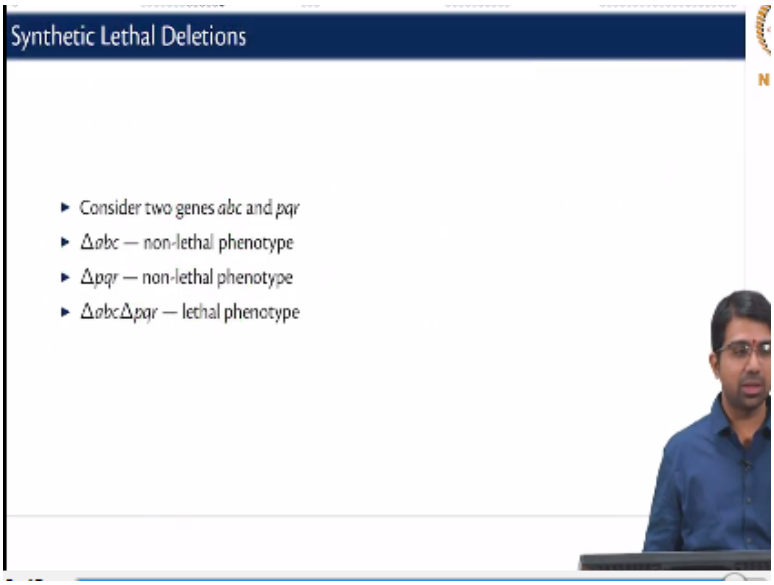
arrays and so on to experimentally study lethality. You can also predict these based on the gene network or the metabolic network and so on.

So in many of these cases we may not actually worry about finding the alphas as such we will stop finding if alpha is not non-zero, right. First step is to find out alpha as non-zero and maybe possible find the sign of alpha, finding the value of alpha is a lot more tricky. **"Professor – student conversation starts"** So lethality only exists for (()) (29:35)? There are single arrays that could be lethality you could have another gene like you know DEF which when you delete the organism will die already, right. **"Professor – student conversation ends"**

So these are essential genes. These are synthetic lethal pairs. And one would call these synthetic lethal triplets. So a lot of studies, we will study some of these later in the course as to how these are studied using metabolic networks. It is quite easy to study these using metabolic networks. Wherever you can predict the phenotype you can start and the phenotype on a given perturbation so you remove A see what happens, if remove B see what happens; you remove A B and C and see what happens. Okay.

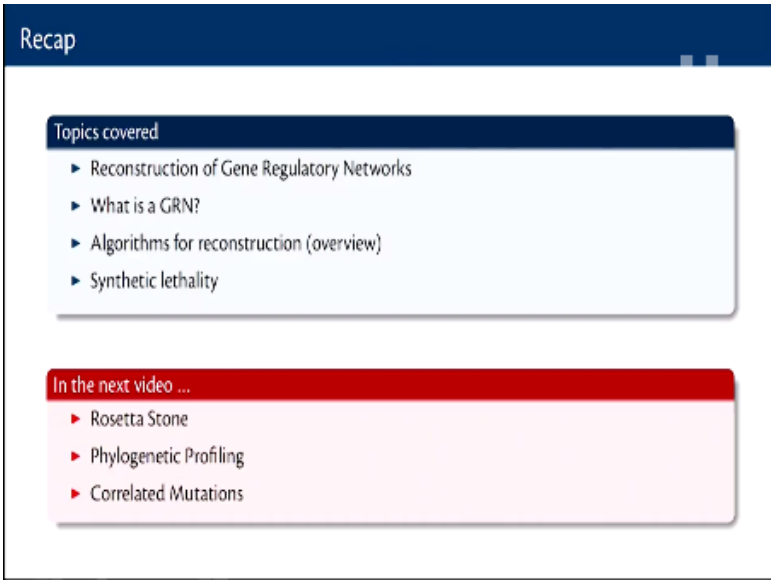
Using this kind of a methodology you can actually identify synthetic cells in any organisms. So all these can help you to form hypothesis which you can later try to validate through other kinds of experiment. So this is what it just explains.

(Refer Slide Time: 30:41)



So get non-lethal phenotypes when you delete a single gene, but when you delete both genes you get a lethal phenotype, this is called a synthetic lethal because you synthetically super impose the two deletions, right only then you get a lethal phenotype.

(Refer Slide Time: 30:56)



So I hope you had a good overview of how one builds and reconstructs networks particularly gene regulatory networks. And we also looked at the concept of synthetic lethality. This is the topic that we will revisit in little more detail when we look at metabolic networks because there are very interesting algorithms to identify synthetic lethal in metabolic networks.

In the next video, we will look at ways to reconstruct protein-protein networks the kind of networks you find in string and database and so on. And methods such as Rosetta Stone, Phylogenetic Profiling and Correlated Mutations which enable you to predict these functional associations.