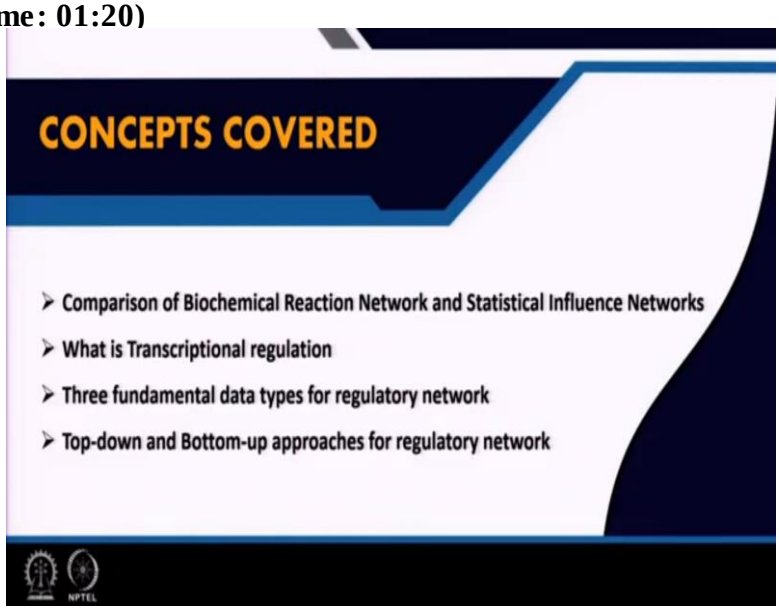
Metabolic Engineering
Prof. Amit Ghosh
School of Energy Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 14
Regulatory Networks

Today we will start a new topic that is regulatory network. So far you have learned a little bit about the metabolic networks. And the metabolic networks assume that all the metabolites are actually present inside the cell without regulation. But if you put regulation, not all the metabolites are synthesizing inside the cell and that is why the regulation of metabolism is very crucial and the regulatory networks are very important for metabolism. So, in the metabolism which you want to understand the metabolism or the metabolic network, you have to actually understand the regulation. The regulation is actually driving the metabolism.

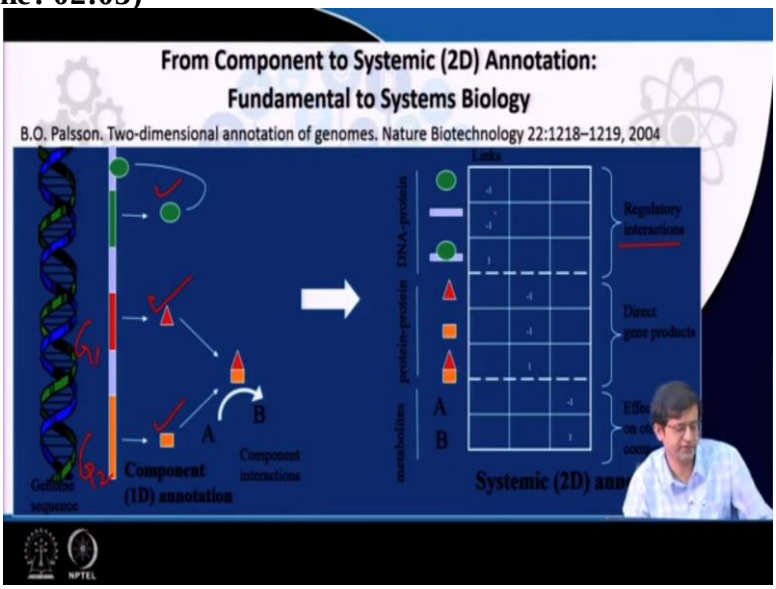
(Refer Slide Time: 01:20)



So, understanding the metabolic regulatory network you need to understand things like what is the difference between the biochemical reaction network and statistical influence network. So, metabolic network sometime also known as biochemical reaction network and statistical influence networks is actually the regulatory network. So, we will compare these 2 networks how it is different and then what is transcriptional regulation. So, we will learn the basic concept in transcriptional regulation and there are 3 fundamental data types for regulatory network.

And mainly we have top down and bottom up approaches for regulatory networks. What is top down and what is bottom up approaches? That also we are going to learn today.

(Refer Slide Time: 02:03)



So, this is the slide in which we have seen in my previous lecture where you have the annotation like given a genome sequence because of the genomics you have readily available genome sequence and genome sequence you can annotate based on the open reading frame and using the open reading frame you identify, you can see that this you can identify different component. So, this is one component, this is another component and this is another component.

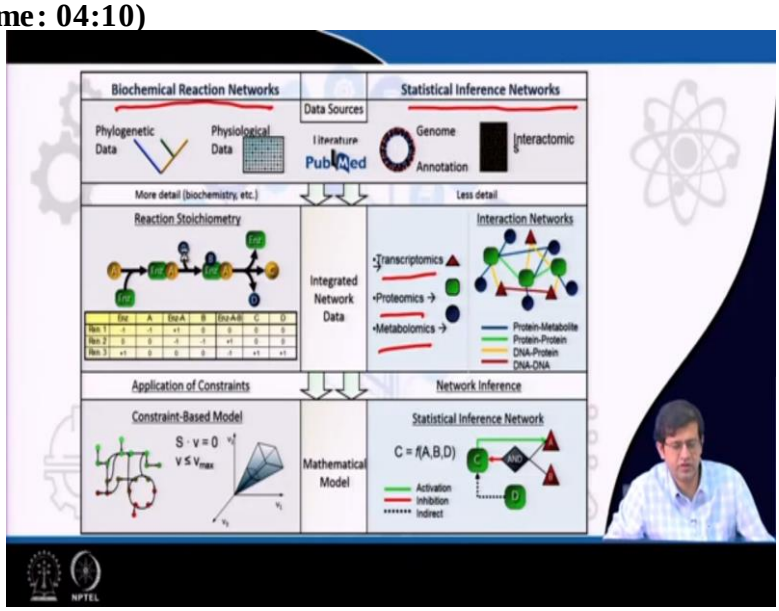
So, this component, you can identify and then this is known as one dimension annotation and then you can see that the protein which is generated from the gene is actually binding the DNA. DNA double strand and it might be a transcription factor that is why it is binding the protein DNA interactions are happening. And then we have it may happen that the 2 protein can interact each other and they become a subunit of a given protein and which can catalyze a reaction.

So, where it can be the 2 subunit this is coming from 2 different gene that is a gene 1 and gene 2 and this form 2 subunit and it is catalyzing a reaction from A to B. So, this becomes the component of interaction and that you can represent it in a matrix. The matrix form that we learned in the previous class where the stoichiometric matrix is basically the metabolite. We have metabolite A, B and the rows are basically the reaction.

So, and also you can have protein-protein interaction that can also be stored in a matrix and also the regulatory interaction also can be stored in a matrix. So, in mathematically you can

store different information like protein-protein, protein-DNA, metabolite which are participating in the reaction also you can store in a stoichiometric matrix.

(Refer Slide Time: 04:10)



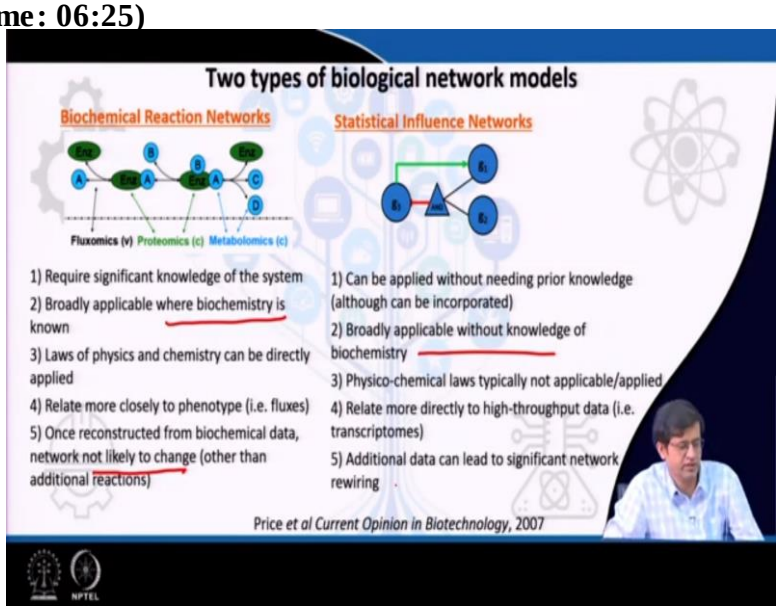
So, there are 2 kinds of networking in biological networks, broadly classified as a biochemical reaction network and this statistical influence network. So, biochemical reaction network is basically the metabolic network which is basically the biochemistry of the cell, every cell has a particular biochemistry and that biochemistry has come up because of lot of experiments have been done over the last several years and you get the biochemical reaction network completely on the biochemistry of the cell.

But the statistical inference network are actually we get because of the transcriptomics, proteomics, metabolomics data. So, we have this high throughput data available transcriptomics, proteomics, metabolomics data and then you try to infer the type of interaction the protein-metabolite interaction, protein-protein interaction, protein-DNA interaction and DNA-DNA interaction. So, this interaction you would be able to get from the high throughput data which are readily available nowadays.

So, these are basically the interactomics data and using that data you try to infer some network the statistical inference network where you can represent as a Boolean network as a function of A, B, D it either it can activate a gene. So, here you can see that these 2 component A and B is actually represented by a AND gate and it is actually activating some gene that is C. So, C becomes a function of A, B, C, D. The A, B, D and then you get the product.

This way you can build the network based on the high throughput data. Similarly, the reaction stoichiometry which I have already explained, we have the reaction in the every enzyme corresponding we catalyze one reaction, the reaction 1, 2, 3 like that you can actually represent depending on the biochemistry of the cell. So, this 2 different the reaction stoichiometry and the interaction network based on transcriptomics, proteomics data, metabolomics data that helps you to build the regulatory networks also.

(Refer Slide Time: 06:25)



So, if you make a difference between the biochemical reaction and statistical influence network. The statistical influence network is basically you have the statistical system level inference that you get the statistical inference at the system level that you can build based on the data. And so, biochemical reaction one of the first difference is that there is a significant knowledge of the system.

If you have a significant knowledge of the system then you can construct the biochemical reaction network. And on the other hand, the statistical influence network is basically you do not need much prior knowledge about the system, you can build the network without knowing much about the system. The broadly applicable in this case it can be broadly applicable where biochemistry is known. Biochemistry is known then you can actually make a biochemical reaction network otherwise not.

And the biochemistry you know that it takes a lot of time to actually know the biochemistry of the cell. And on the other hand, the statistical inference network is basically without the

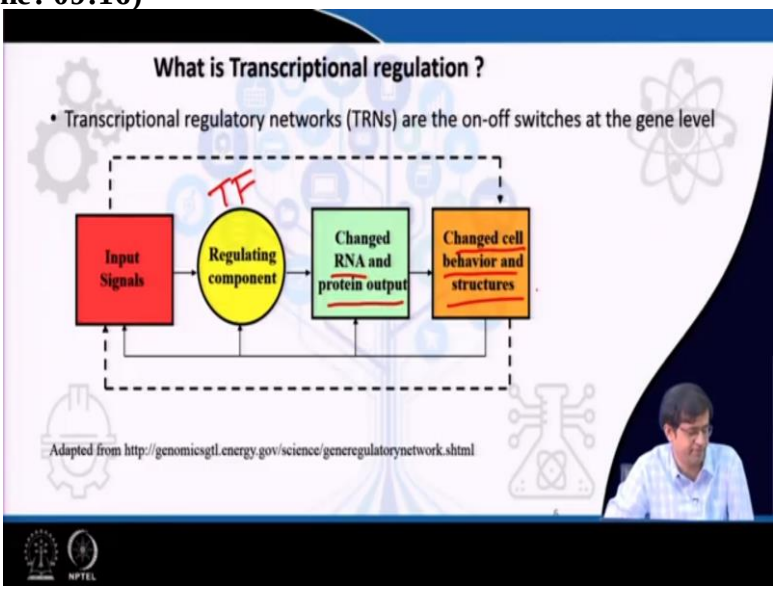
knowledge of biochemistry. So, you do not need to know the biochemistry of the cell. Whereas in biochemical reaction network the laws of physics and chemistry can be applied. Whereas in the other statistical networks, we have the physico chemical laws typically not applicable.

So, these are the main differences related more closely to phenotype. We say a biochemical reaction are closely related to the phenotype that is fluxes because from the biochemical reaction network, you can calculate the fluxes whether you apply regulation or not but still at any condition, you can calculate the flux. Relate more directly to high throughput data, the statistical influence network are more directly to high throughput data.

Once reconstructed from biochemical data, network not likely to change. So, in the biochemical reaction network, it is not going to change this is important difference that the biochemistry of the cell is not changing with time. So, this is the major difference between biochemical reaction networks and statistical influence network where the as you add more data the statistical influence network then the wiring diagram changes, it is actually data driven.

Whereas, the biochemical reaction is not going to change is almost rock solid. So, this is the major difference between these 2 kinds of network which are present in biology. One is the biochemical reaction network another is the statistical inference network.

(Refer Slide Time: 09:16)

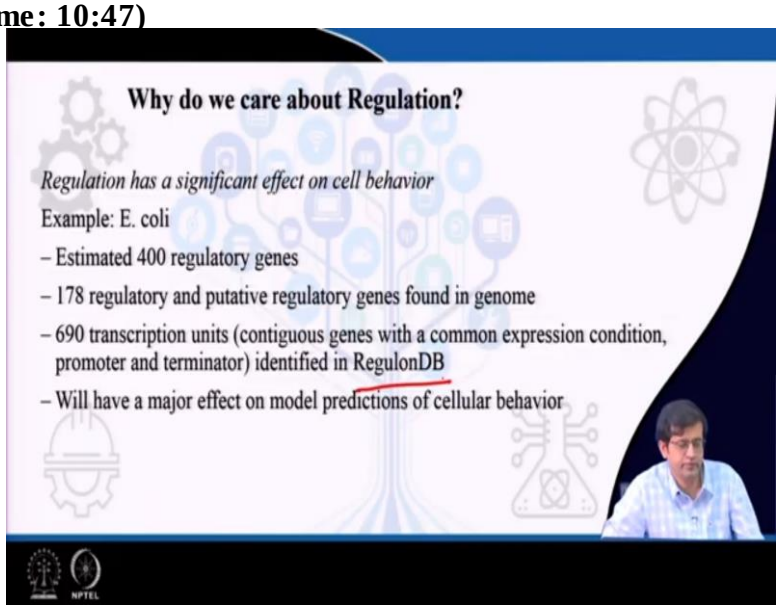


Then now I just want to explain like what is the transcriptional regulation? The transcriptional regulation is basically you have some input signal and then based on the input signal the regulatory component get activated, like transcription factors. So, transcription factor, TF get that is the regulatory component which get activated because of the input signal and the transcription factor, the activated transmitter either it can repressed or activate the gene for expression.

So, this is how you get your RNA protein output increases because of the regulatory component and then ultimately you see the change behaviour of this cell, cell behaviour changes and also the structure of the cell changes because of the regulation. That is why transcriptional regulatory networks are very important and they act like a on and off switches at the gene level. So, either the gene will be expressed or not expressed depend on the transcriptional regulatory network.

And that is very important for metabolism as well. Because the metabolism without the regulation it may not be accurate. You may construct the metabolic network but if you do not apply the regulatory network, the transcriptional regulatory network then your prediction or the network is incomplete.

(Refer Slide Time: 10:47)



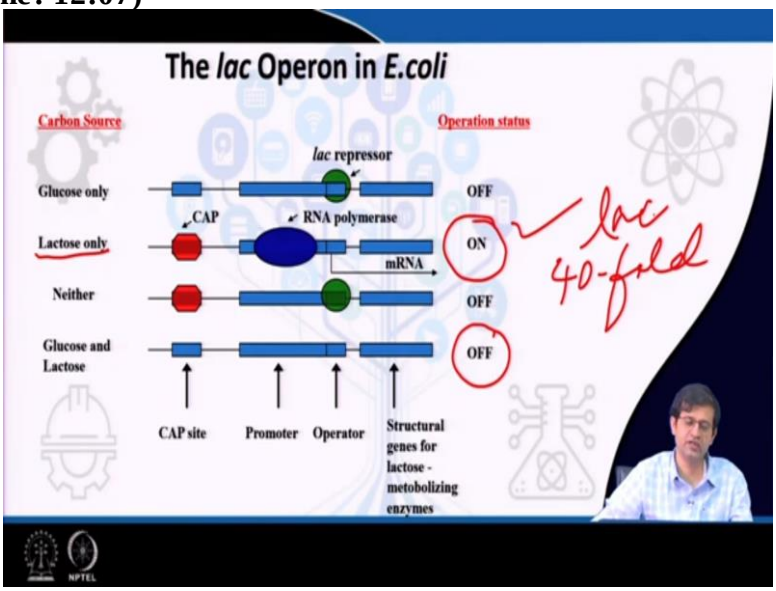
So, what do you care about the regulation? Why you want to learn about regulatory regulation because the regulation if you see in E.coli, it is estimated that around 400 regulatory genes are present, 400 genes are present in E.coli itself. And there are 178

regulatory and putative regulatory gene found in genome. So, these are regulatory and putative regulatory genes.

And then we have 690 transcriptional units that is a continuous contiguous gene which are with a common expression condition promoter and terminator identified in Regulon database. So, we have various databases for example, the Regulon database can give you more data about the transcriptional regulation. So, how many transcription units are there that you can identify from different and also it will help in prediction of the cellular behaviour.

Once you add the regulatory component then the prediction of the cellular behaviour will be more accurate because it is much more close to the in-vivo result. So, the cell behaviour can be more, much more manipulated when you include the regulatory component.

(Refer Slide Time: 12:07)



So, here I have shown the lac Operon in E.coli. The lac Operon in E.coli consists of 3 structural gene that is lac A, lac Z, lac Y and involve in lactose utilization, the operon is regulated by lactose and glucose signal mediated by 2 DNA binding protein, the lac repressor and CAP respectively, lac repressor binds DNA only in the absence of lactose whereas CAP bind only in the absence of glucose.

So, this we have 2 transcription binding protein and that has 2 different role basically, the one binds in the absence of glucose that the CAP protein actually binding in absence of glucose. And the other DNA binding protein that is that the lac repressor also bind in absence of

lactose. So, this you can understand how it is happening like when glucose is present, only then there is no lactose.

So that is why the lac repressor binds the promoter region of the DNA and then when lactose is only present that is the only time the lac repressor is absent. So, the here you can see the lac repressor is moving away because lactose is present, whenever lactose is present then the lac repressor will not bind then it can recruit the RNA polymerase in the promoter region and that actually help for the expression of the gene that is the mRNA is formed.

And since there is no glucose the CAP protein is not binding. So, now, we have neither of them, so, neither of glucose or lactose is present then the lac repressor is still binding. So, that is why there is no expression of the gene. And in the other case, we have glucose and lactose are present. So, when glucose and lactose are present then what happened? Then the CAP is not binding and also the lac repressor is also not binding.

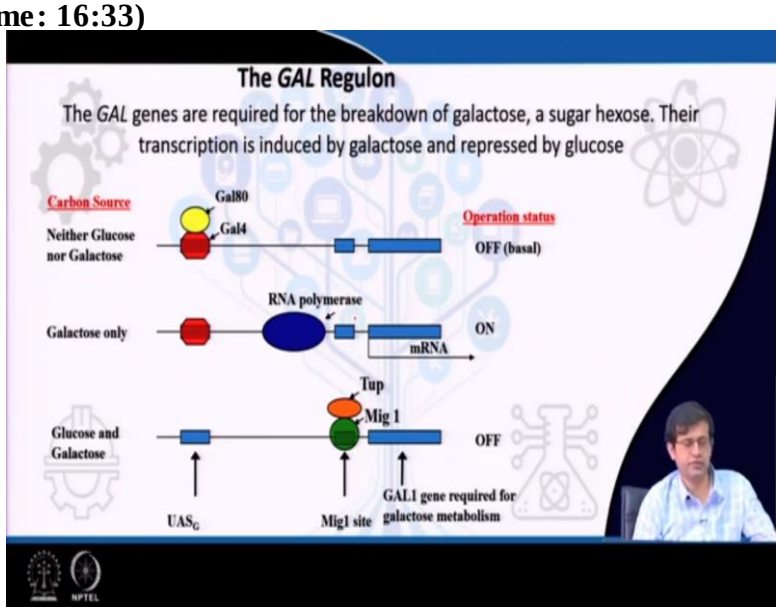
Since there is no binding and none of the transcription factor is binding then also we do not have expression of the gene. Because RNA polymerase is not binding. So, the mRNA level is 0. So, this switch is actually, you can see it is almost like a switch where the gene is expressed only in one condition that is on condition where there is only lactose. So, this lac Operon is activated only when the lactose is present. And if lactose and glucose both are present then also you can see it is off.

Because the CAP protein is not binding, the cap protein is only binding when there is the absence of glucose. So, the CAP binds only in the absence of glucose and also the lac repressor binds DNA only in the absence of lactose. So, whenever the lactose is absent the lac repressor is binding. So, this is making the regulation off where the expression level is 0. The RNA polymerase binds weakly to the promoter and the operation is transcribed at a low level and whenever the lactose is present, only the lactose is present then only the lac repressor is moving away.

And it is recruiting RNA polymerase to the promoter region and the expression level of the lac gene is increased by 40 fold. So, in this case, the lac gene is expressed only in this case, we have the expression of lac gene by 40 fold. So, this is the lac Operon how it is regulate

this we have already read in the book but this is just to give you an overview of how the lac operon is working in E.coli.

(Refer Slide Time: 16:33)



Similarly, for GAL regulon, GAL genes are required for the breakdown of galactose. So, some cells are actually consumed galactose for those GAL genes are required. It is a sugar molecules sugar hexose molecule and their transcription is induced by galactose and replaced by glucose. So, when the cell does not have glucose or galactose what happened that GAL is not activated. So, since the Gal80 is actually binding Gal4 there it is repressed and it is not activated.

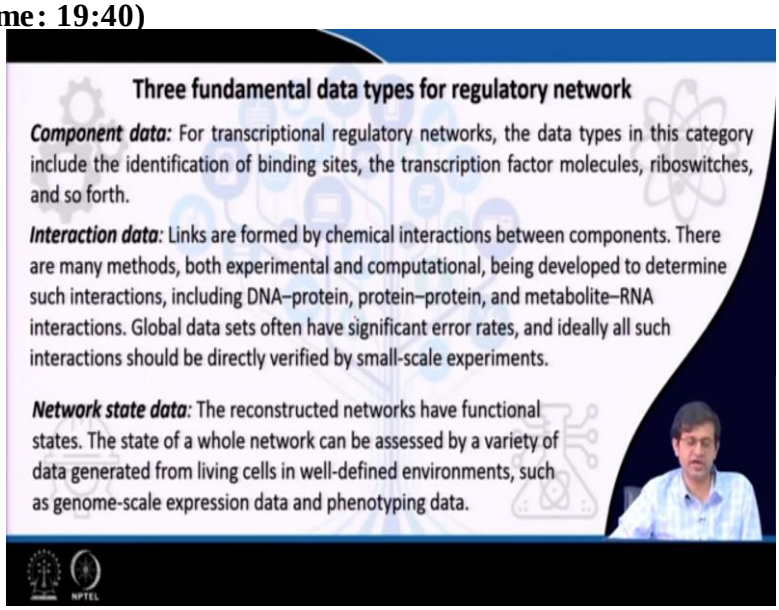
That is why you do not see any production of the GAL gene is not expressed in when both glucose and galactose is not present in the medium. So, now when galactose is only present, so that time we can see that the mRNA is expressed and the Gal genes are transcribed and we get the expression and that is when where we can see that whenever galactose is present then the Gal80 interaction between Gal4 and Gal80 is removed.

And it can now recruit the Gal4 can recruit the RNA polymerase because Gal4 is activated in this case where you see the expression of the mRNA. And in the next case, what happened when we have the glucose and galactose what happened because whenever you have glucose and galactose Gal4 is activated. But, the repression of because Mig 1 binds the promoter region and it actually repress the gene and that is why you do not find any expression of the Gal gene.

And Gal gene required for galactose metabolism. So, if galactose is present and glucose is absent then Gal80 inhibition is released, Gal80 is released and then Gal4 recruit RNA polymerase. So, the RNA polymerase Gal80 is removed from Gal4 then only the RNA polymerase recruited to the promoter region and you see the expression of the gene otherwise not.

So, this is again a regulation of expression that the genes are expressed depending on the metabolite present in the cell. So, if galactose is present then only you see that the Gal gene is expressed and the Gal gene is actually required for the breakdown of the galactose. So, if the cell is having galactose then only the Gal genes are expressed and you can see that they metabolite galactose that is the hexose sugar hexose for metabolism.

(Refer Slide Time: 19:40)



So, there now I come to 3 fundamental data type for regulatory network that is a component data, interaction data and network state data. So, these are the 3 types of regulatory fundamental data types for regulatory network. For transcriptional regulatory network the data types in this category include the identification of binding site. So, in previous lectures also I told that you have to identify the component and the component can be a transcription factor or riboswitches or the binding sites of the promoter region.

So, DNA binding sites we can identify from the component data, this is one category of data which require the characterization of the component and then we have the interaction data, the interaction data can be protein-DNA interaction data, Protein-Protein interaction data or metabolite-RNA interaction data. There are both experimental and computational methods to

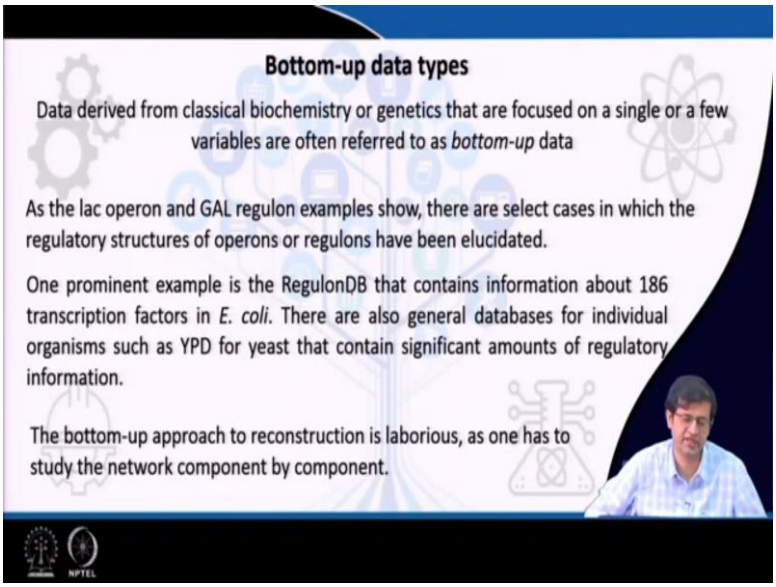
identify the interaction data that happens between DNA-protein, protein-protein and metabolite-RNA interaction.

The global data sets often have significant error rate and these kinds of data, the experimental data or the computational data may have a lot of error rate which you have to identify before or ideally, you should be verified using small scale experiments. So, this interaction data need to be verified through some experiment. And the third data is basically the network state data, the reconstructed metabolic network or any reconstructed networks have different functional state that is have different phenotype.

The state of the whole network can be assessed by a variety of data generated from living cell in a well defined environment such as a genome scale expression data or phenotyping data. The whole network can be assessed by a variety of data generated depending on the environment and these are basically genome scale expression data like the micro expression, the micro array expression data that you can actually measure for the entire cell at a different condition.

Like suppose you have a particular condition the cell is growing by that particular expression, environmental condition, you can calculate the expression or you can measure the expression of the gene using micro array and those also you can measure the phenotyping data is basically you can measure the growth rate and so, at that particular network state. These data are required for making the regulatory network and component data, interaction data and the network state data. These are the 3 types of data required for regulatory network.

(Refer Slide Time: 22:27)



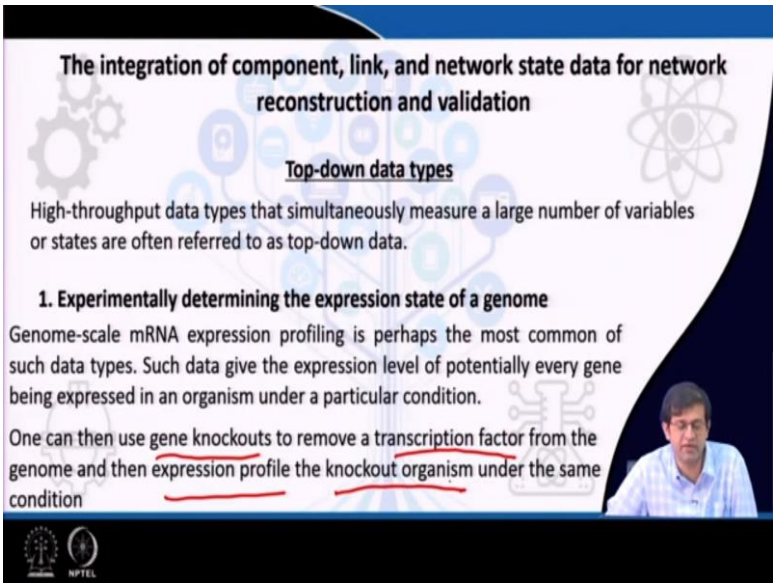
Now, I will describe the bottom up data types. What are bottom up data types? The data derived from classical biochemistry or genetics that are focused on a single or a few variables are often referred to as bottom up data. So, it can be the variables that is smaller component of the network can be identified using the bottom up data types, for example which I expressed the lac Operon or the Gal Regulon.

These are the example where the bottom up data types are basically where you have individually look for a particular component and identify the regulation. One prominent example is the regulon database that contains information about 186 transcription factors in E.coli. So, each of the transcription factor is actually characterized for E.coli and has been stored in a Regulon database. There are also general database for individual organisms such as YPD for yeast that contain significant amount of regulatory information.

The bottom up approach to reconstruction is laborious and one has to study the network, component by component. So, in this network construction, you are actually component by component you are identifying and that is why it is very laborious, it is very time consuming because each component you have to identify individually. This approach, the identification of individual component is very important because at the component level you can identify experimentally and then most of the data are available in different data for yeast and E.coli.

You can search this organism and find out the component the exact function of the component step by step you can identify.

(Refer Slide Time: 24:27)



Now, we have another approach that is top down data types. The top down data type, the bottom up approach is basically identifying the individual component and the top down data types is basically high throughput and this is because of the high throughput technology in biology. And these top down data types are now increasing day by day. And the integration of component, link a network state required for network reconstruction and validation.

So, we have to integrate the component, link and the network state together for the network reconstruction and also for the validation. First you build a network and then you have to validate and for that we have the top down data types that is the high throughput data that simultaneously measure a large number of variables or states are often referred to as top down data.

In bottom up data you have seen that only one variable or 2 variables, few variables are considered. But in top down data types we have large number of variables that is large number of transcription factor; large numbers of states are often referred in the top down data type. And there are ways you can measure that experimentally and number one is basically the experimentally determining the expression state of a genome. And how do you do that? The genome scale mRNA expression profiling is perhaps the most common data types.

Such data give the expression level of potentially every gene being expressed in an organism under a particular condition. So, genome scale expression analysis or profiling using microarray expression data analyzers where you can get each and every gene, the expression

level of each and every gene in the organism under a particular condition and this is really helpful.

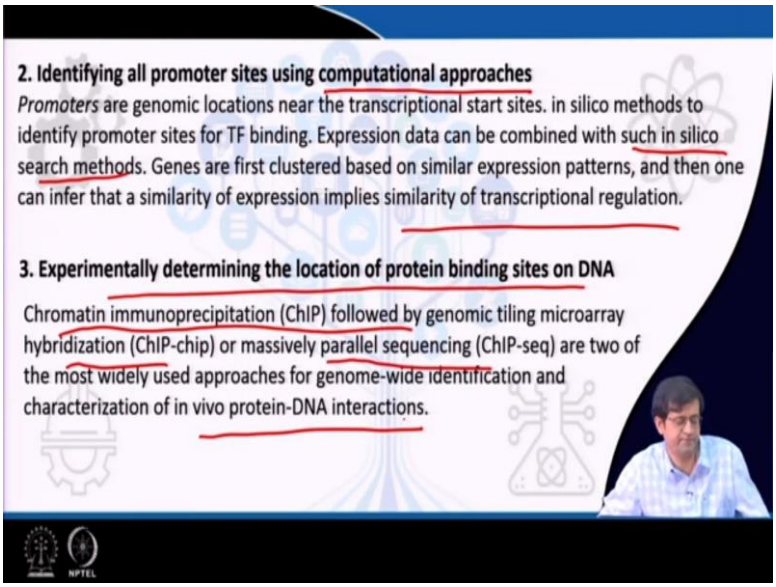
Because, suppose you have, you want to understand whether the transcription factor playing a role or not or the transcription factor is binding the DNA or not, experimentally, you can verify by knocking off one of the gene. So, you can gene one of the gene and then knock off one of the gene and then in order to remove the transcription factor, suppose you want to remove one transcription factor and see how the genome scale expression is changing.

So, the expression profile you can calculate it for the knockout organism and also for the wild type organism and then you can take a difference. So, this is very useful in metabolic engineering. So, in the metabolic engineering once you knock off a gene, you try to calculate the expression level for the entire organism at the genome scale level, genome scale mRNA expressions have been calculated and then they compared with the knockout strain and also with the wild type.

And this gives an enormous idea of how the transcription factor is actually playing the role, what is the importance of the transcription factor you can actually able to figure out from the expression data. So, this study the experimentally determining the expression state of a genome and the genome scale level become very crucial in metabolic engineering you will see that many papers, many articles is actually use this concept for actually understanding the influence of different regulatory elements.

Like transcription factor, how it is actually influencing the expression of the gene and that you can do it very easily nowadays because of the genome scale mRNA expression date profiling.

(Refer Slide Time: 28:03)



And then we have other 2 methods that is the identifying all promoter using computational approach. So, how do you actually identify the promoter? That you do not have to do an experiment, computationally also you can do it. The promoters are genomic location near the transcriptional start site. So, how do you identify the promoter because every gene is actually starting with a promoter?

And in the last class, I told that to you once you have identified the open reading frame that is the start codon and the stop codon. So, before the start codon you have the promoter region, it can be 1 kb, 2 kb, 3 kb based on the gene you are actually expressing. In silico methods are applied to identify the promoter site for transcription or binding. So, transcription factors are binding the promoter site that is why you have to identify the promoter.

The expression data can be combined with such in silico search method as well, you can combine the in silico data for identifying the promoter region. Genes are first clustered based on similar expression pattern and then one can infer that similarity of expression profile gives rise to similarity of transcriptional regulation. So, you can cluster these expression profiles in one and that can give rise to a similarity of transcriptional regulation.

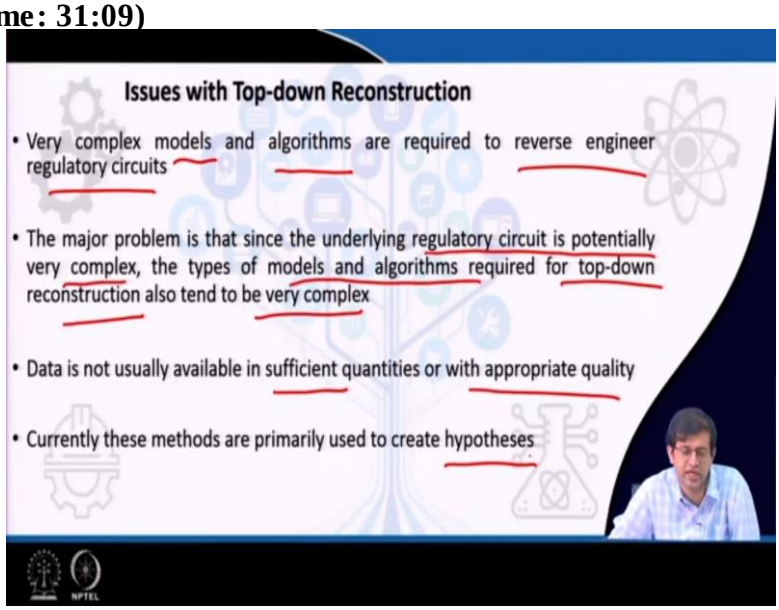
So, clustering methods are actually applied where you calculate the similar expression profile under some particular condition. So, these computational methods are also very powerful to identify the promoter size and also the binding regions of the transcriptional factor binding are also calculated. And lastly, the experiment will determine the location of protein DNA

protein binding site on DNA. So, in this case I told about the computational approaches and this is the experimentally you can determine the protein DNA binding sites.

And how do you do that? There are experimental methods like chromatin immunoprecipitation followed by microarray hybridization technique. These are very well known high throughput techniques which are available today. Where you can combine the immunopercipitation, ChiP, Chromatin immunoprecipitation along with the microarray hybridization also you can massively parallel sequencing like, ChiP seq are also applied nowadays.

These are the 2 methods most widely applied approach for genome wide identification and characterization of in vivo protein DNA interaction. So, in vivo protein DNA interactions are also identified by these 2 techniques that is ChiP chip or ChiP seq. So, these techniques are also very popular where you identify the protein DNA binding because the protein the transcription factor binding the DNA is very crucial for the expression of some genes it may be metabolic genes or many other genes we can understand in a detail.

(Refer Slide Time: 31:09)



So, there are issues with top down approach because where the model is very complex the algorithms are required to reverse engineer the regulatory circuit. Because the top down approach you get a lot of data, a lot of protein-DNA interaction data that you have to actually build a model ultimately because you have a lot of data then you also do not know what information you have to bring out from the data.

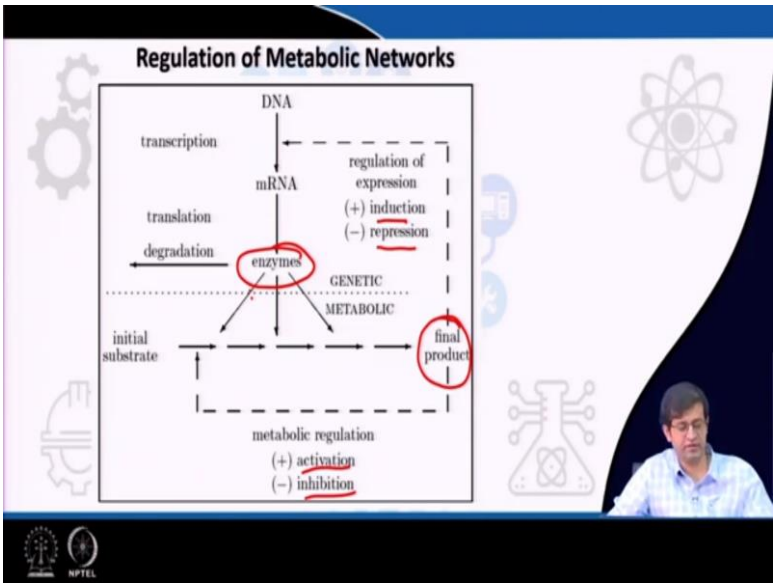
So, these high throughput data, big data, you know already that how do you analyze the data for that you have to build the model, so, building the model and then applying the algorithm. So, are required to reverse engineer the regulatory circuit. The regulatory circuit which you get in terms of protein DNA interaction data that you have to do reverse engineering. For that, you have to build the model and also you have to develop algorithms to understand the data in better way.

See major problem is that, since the underlying regulatory circuit is potentially very complex, the regulatory circuit or the genetic circuit which I have discussed already is potentially very complex and the types of model and the algorithm required from the top down reconstruction also tend to be very complex. So, both the data is complex and also the model and algorithm are also very complex.

So that is why this is a major hurdle or major problem in understanding the top down reconstruction, the data is not easily available in sufficient quantity or with appropriate quality. So, the one is the quantity of the data and also the quality of the data. Sometime is not available in that amount. So, the quality of data need to be improved also need more data to actually build this kind of model.

Currently, this method are primarily used to create hypotheses or once use the data you try to many journals publish high throughput data on top down reconstruction and they build different hypotheses, they build different model and algorithm and keep on updating. So, more you have will see in future also this kind of model coming up where they try to model the high throughput data that is the protein-DNA interaction data or the mRNA expression data at the genome scale level.

(Refer Slide Time: 33:37)

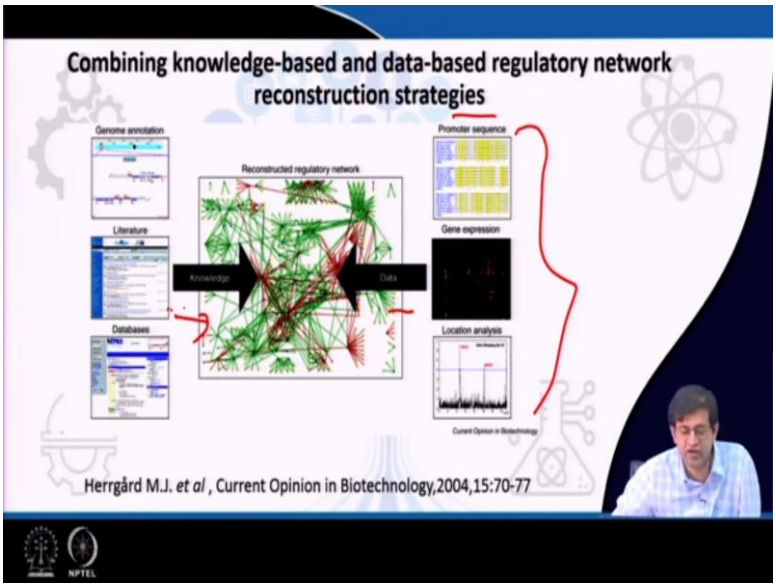


So, the regulation of the metabolic network once you know the metabolic regulatory network then you can add in the metabolic network then you know that how many metabolic genes are actually formed inside the cell because the metabolic genes are formed inside the cell will be crucial for the metabolic networks. So, in the last class also I have told that the metabolites which are form in the reaction and either they can activate the gene or inhibit the gene.

So, the enzyme can be inhibited or activated depending on the metabolite which is binding the enzyme. And also the metabolite can also do repression in the lac operon we have seen that the metabolites are actually repressing the gene and also it can induce the gene also. So, both way we can actually see the product, the final product which is formed through the reaction can also either it can repress the gene or do induction of the gene or it can do activation or inhibition of the enzyme.

So, this way the regulation you have to be very careful that is the wiring diagram of life while you will be able to actually draw a proper regulatory network which can be applied in the metabolism.

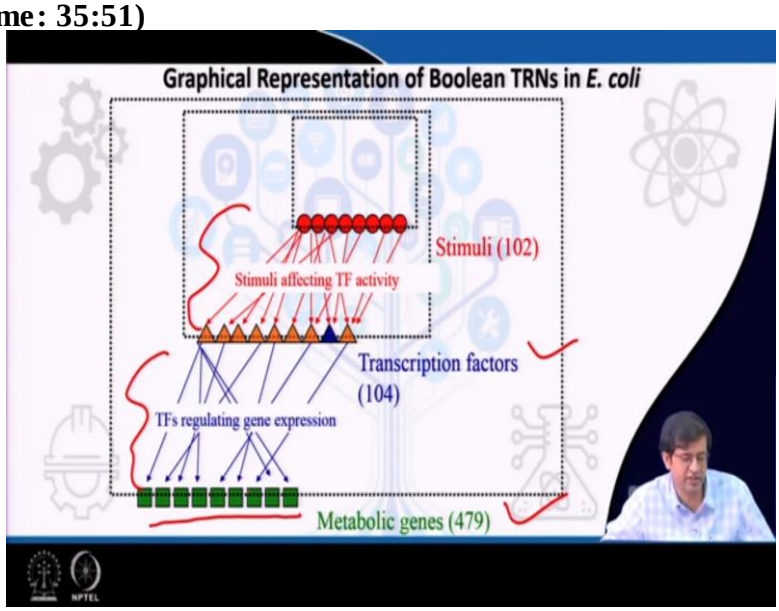
(Refer Slide Time: 34:56)



So, this is the overall, how you can combine the knowledge based and data based regulatory network. The data based network or data based regulatory network is basically from the high throughput data where you have the gene expression data, promoter sequence data and the location analyses. It is high throughput data you can use to build the regulatory network also you can have the genome annotation data.

The database literature that also can be this is the knowledge based from the knowledge which is already available from the literature from other regulatory network that you can use to build the complete regulatory network. So, one hand you can use the data and other hand you can use the already existing knowledge. If you can combine you can build a much more accurate regulatory network.

(Refer Slide Time: 35:51)

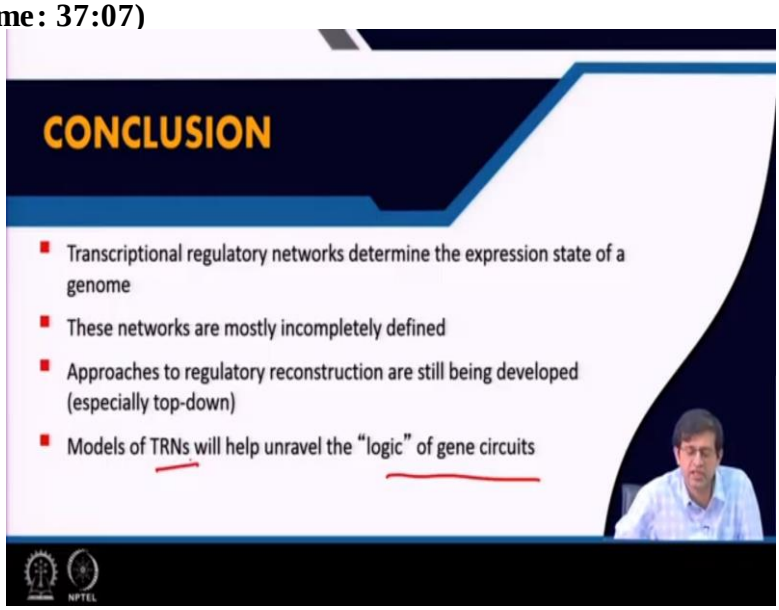


So, here you can see the cascade of network, in the bottom you see the metabolic genes, the metabolic genes are actually controlled by regulatory genes. So, the transcription factor which are actually regulating the expression of the gene are shown in the green color and the transcription factor is shown in the blue color and this transcription factor is actually get activated because of this signaling, the stimuli.

So, this part is basically the signaling network and then this is the regulatory network and that is why you ultimately get the metabolic genes. So, this kind of interaction or the cascade of networks are present inside the cell, while the transcription factors are activated and then we see the metabolic genes are expressed. So, this graphical representation of the Boolean transcription network, so, these are forming a Boolean network.

The transcription factor which are including the metabolic genes are represented by a Boolean network and are already available for E.coli that can be used for understanding the or you can incorporate the regulatory network in the metabolic network and understand the cell regulation.

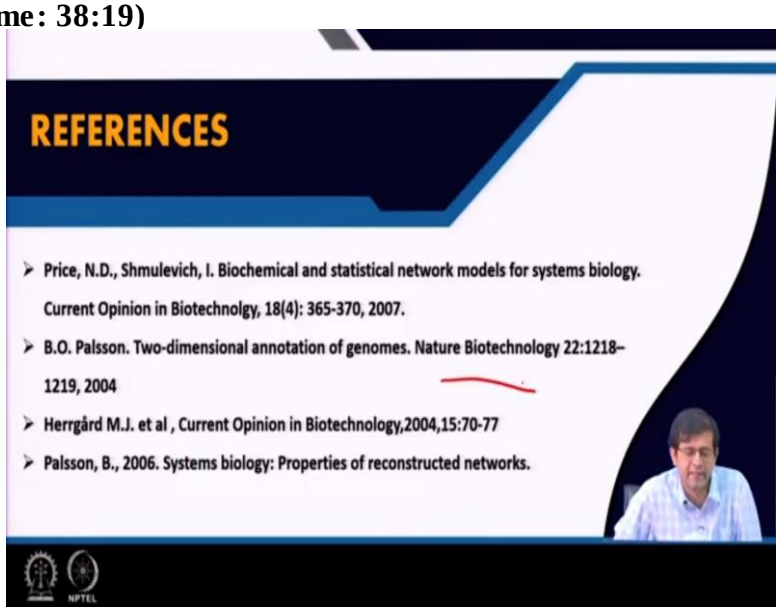
(Refer Slide Time: 37:07)



So, in conclusion, what do you see the transcriptional regulatory network determine the expression state of a genome. So, using the transcription regulatory network, you will be able to determine the expression of the total expression of the entire genome and these networks are mostly incomplete because as you add more data then you can have a different interpretation. So, more, high throughput data, more data available then the network will be much more appropriate.

So, approaches to regulatory reconstruction are still being developed, especially the top down. And the top down approaches a lot of work is going on and how you model the transcription regulatory network will help in unraveling the genetic circuit. So, the logic of genetic circuit can be understood through transcription regulatory network. The more number of transcription regulatory network you know, better you will know the genetic circuit and that ultimately help in metabolic engineering. The approach you will be learning to develop the regulatory network is very crucial.

(Refer Slide Time: 38:19)



So, the references for this lecture is basically you can look for Biochemical and Statistical network model for systems biology which is published in Current Opinion in Biotechnology as also you can read the Nature Biotechnology paper for the 2 dimensional annotation. You can follow these references. Thank you. Thank you for listening.