Introduction to Proteogenomics

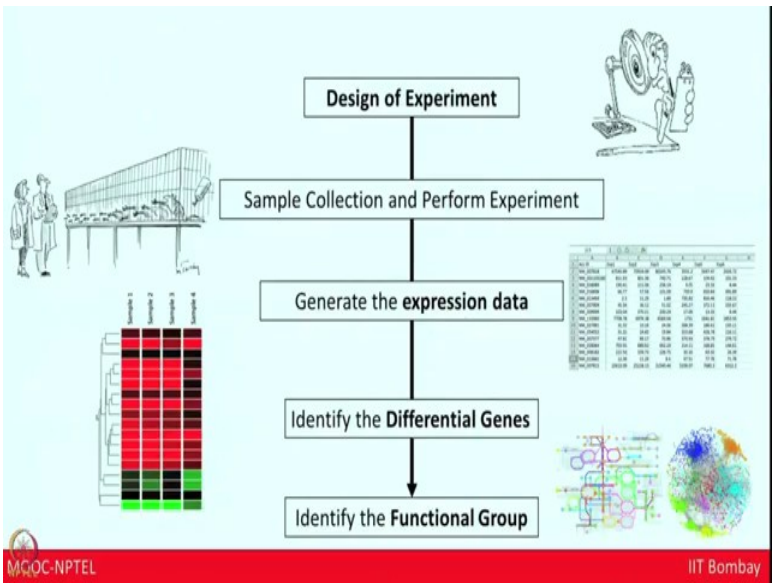
Dr. Sanjeeva Srivastava
Mr. Deeptarup Biswas
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 14
Pathway Enrichment and Network Analysis

Welcome to MOOC course on Introduction to Proteogenomics. In today lecture Deeptarup Biswas a PhD scholar from proteomics laboratory, IIT Bombay will give you a brief idea about pathway enrichment and network analysis. As they are already aware that these next generation sequencing and mass spectrometry based technologies can provide you large OMIC datasets in a very short time, but how to get meaningful information from the big data. Therefore, doing further analysis using pathway enrichment and network analysis becomes very crucial to address different biological questions.

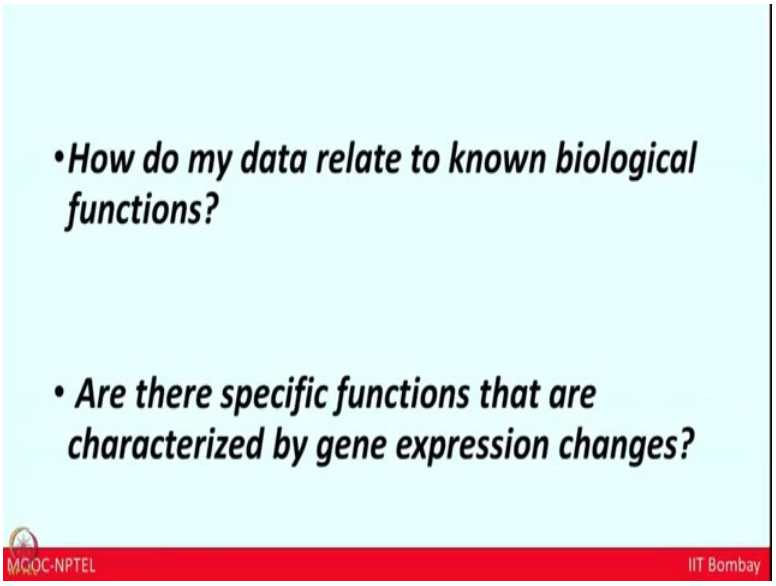
We have learnt a lot about different type of data analysis strategies and how to get significant candidate lists, whether it is genes or proteins, but until we relate these lists and expression values to obtain the biological insight the whole list or the information remains pretty incomplete. Today, he will discuss about how your data set can be used to generate the functional groups, protein-protein interaction modules and pathway enrichment. He will also discuss about multiple online available tools and how these tools can be used to relate your data set with biological function. So, let us welcome Mr. Deeptarup Biswas for today's session.

(Refer Slide Time: 01:56)



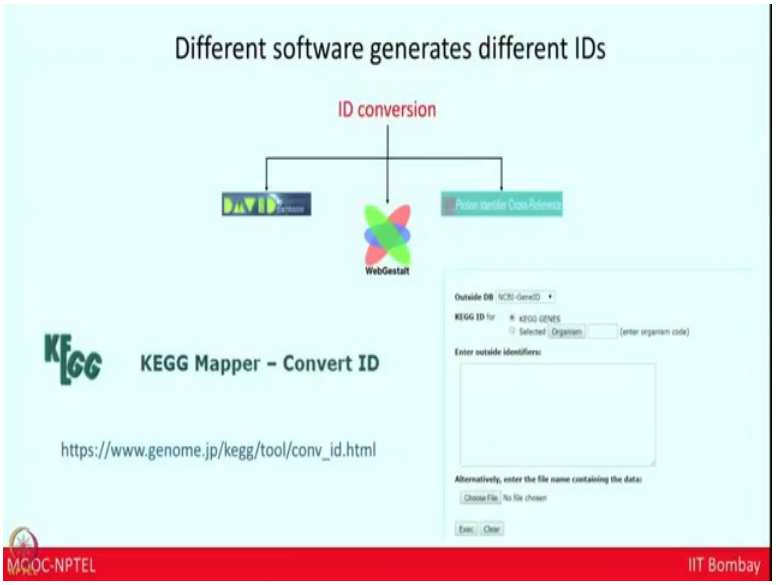
Till now, we have learned a lot about how to design an experiment? What are the conditions? What are the parameters that need to be taken into account for sample collection and performing an experiment. After everything, we have learned a lot about statistical power, primary analysis and secondary analysis to generate the expression data set. Now, the main important thing which is coming that we got a very good pattern of differential gene regulation.

(Refer Slide Time: 02:29)



So, the questions come what next? I want to start with two important question; how do my data relate to known biological function are there specific function that are characterized by gene expression changes. After the secondary analysis, what I feel the most important things to do is the tertiary analysis; that means, identify the functional group. This functional group identification is based on different pathway enrichment, network analysis and PPI modules that is protein-protein interaction modules. So, in the workflow what I have added is the last one after the identification of the differential gene is identify the functional group.

(Refer Slide Time: 03:11)



Different kind of softwares that generate different IDs, if we are using any proteome discoverer, commercial software or Trans proteome pipeline, they will give you different kind of IDs in the protein identification. But, to start and a tertiary analysis, we have to get multiple ID and that is possible only though ID conversion. So, this is a very basic thing, but still I want to take a little time and want to tell you how this ID conversion can be done. So, there are mainly three important platform which we can take into consideration the first thing is David, David is of multiple ID conversion tool apart from this it can help in also different types of annotation and enrichment studies.

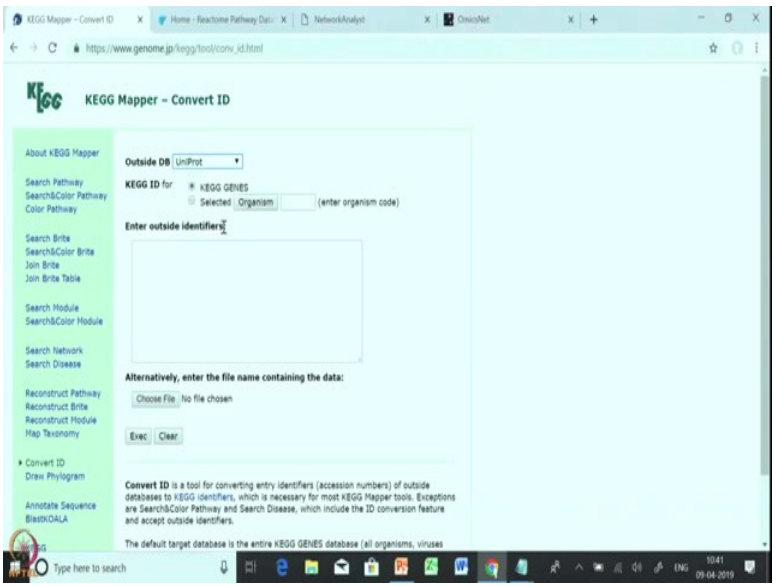
Next is WebGestalt, I think Dr. Bing Zhang has already told a lot about WebGestalt and this platform can also be used for different kind of ID generation. The next is protein identifier cross reference, this is another platform where we can upload our ID and we can get a multiple ID converted from this software. So, David, WebGestalt and protein identifier cross

reference is a very simple tool where we just need to put the ID in the list of ID and we can select what are the what list of ID we will get as a conversion.

But there is an important tool that is KEGG Mapper converter ID that is a very important tool, because like other platform we cannot put any other id in the KEGG Mapper we have to have, we have to get the KEGG Gene ID from the KEGG Mapper and then only we can put it in the identifier identification toolbox and we can get the KEGG mapping pathway. So, I will show you a glimpse, like how from a test data set we can approach to KEGG Mapper and we can convert the ID to KEGG Gene and after that we can put those KEGG converted ID into the KEGG Mapper to get the pathway.

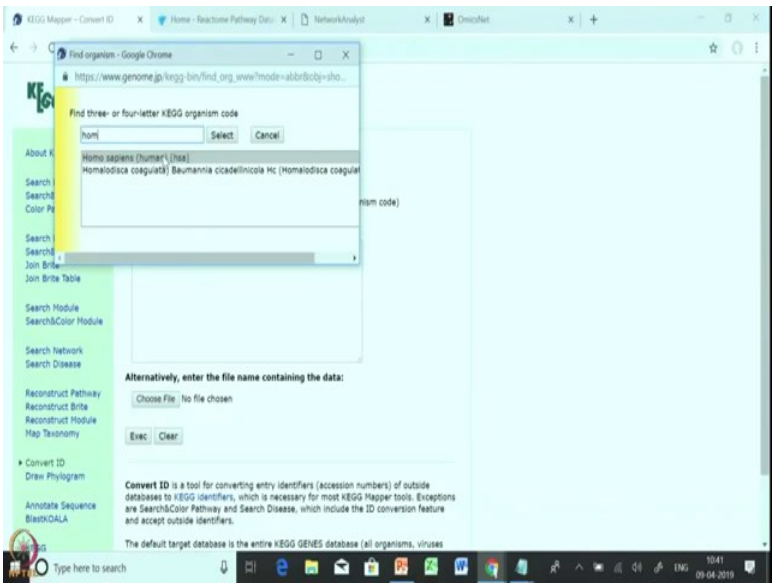
So, I have already shared the text one the text one text file you have to open the text file there is a list of gene that is mainly a GBM repository data set a processed file which I have taken and we will copy paste that list into the KEGG convertor ID. So, I have already shared the link in the slide. So, please go to the KEGG ID converter and copy paste the link.

(Refer Slide Time: 05:53)



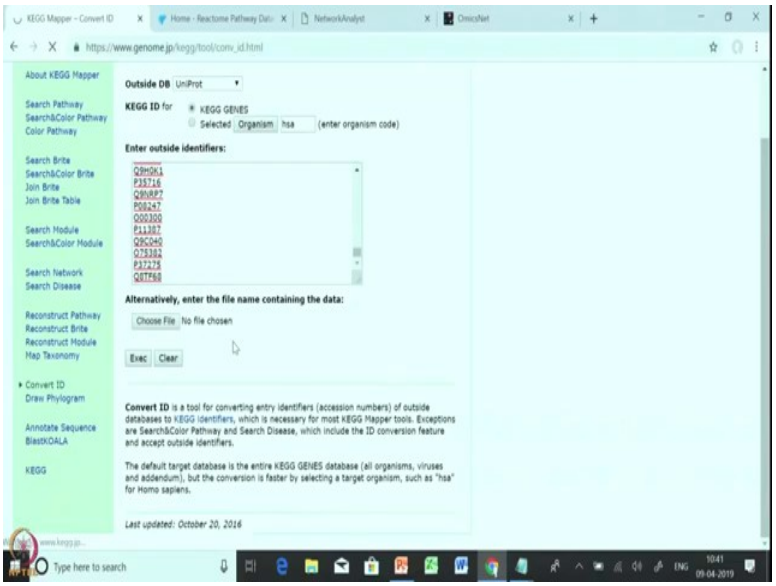
So, this is the home page of KEGG Mapper converter ID, where in the first out outside DB you can choose what is the different NCBI gene ID or NCBI protein ID or uniprot ID you are putting. So, as I have given a list of uniprot ID. So, I will be selecting uniprot ID here, after that the important thing is the what is the organism?

(Refer Slide Time: 06:13)



So, if when we are clicking this so, we have to write the name of the organism and as I know I have taken the file from the *Homo sapiens* depository. So, I will be selecting the *Homo sapiens* and I will select the tab and. So, it is showing here HSA. Here, in the enter out box outside identifier, we will copy paste the list of the uniprot gene and then we will select we will click on the execution tab.

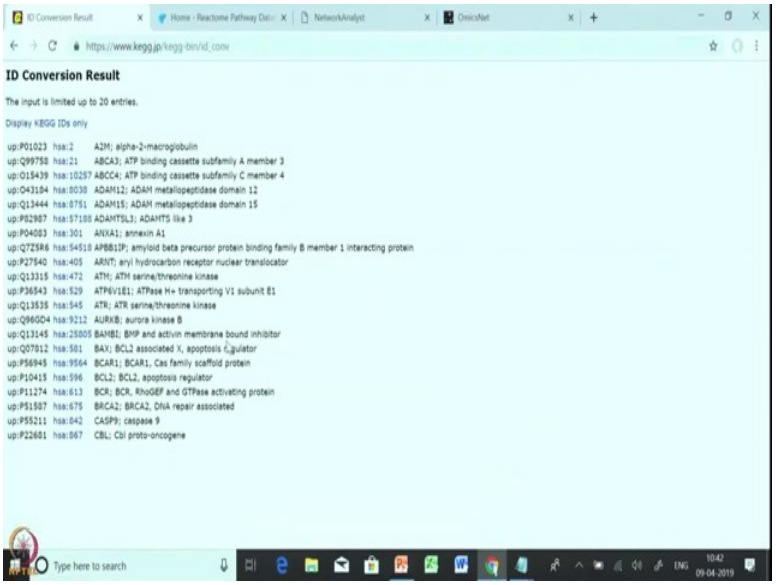
(Refer Slide Time: 06:38)



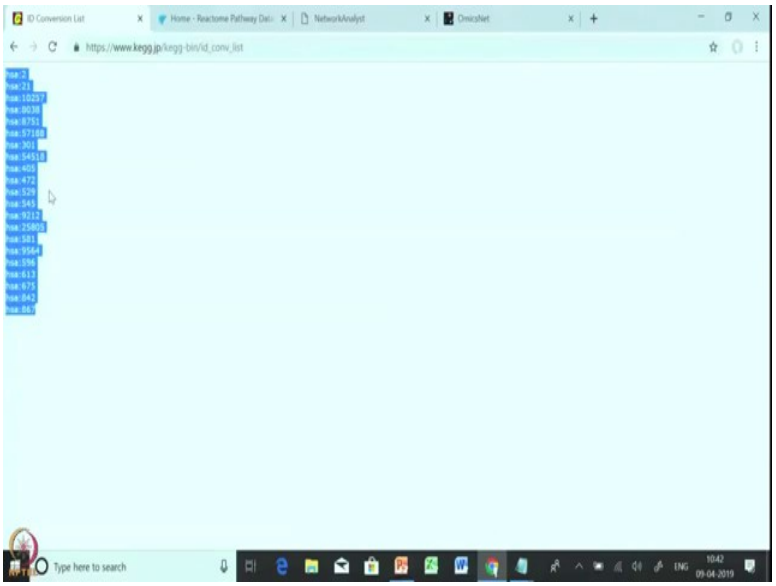
So, it will take a little time and it will give you the complete converted ID from uniprot to KEGG ID. So, as you can see the conversion result is giving you the name of the uniprot list.

So, we will be copying only the KEGG ID. So, we will click here display KEGG ID only, where we will get the name of the KEGG IDs.

(Refer Slide Time: 06:53)

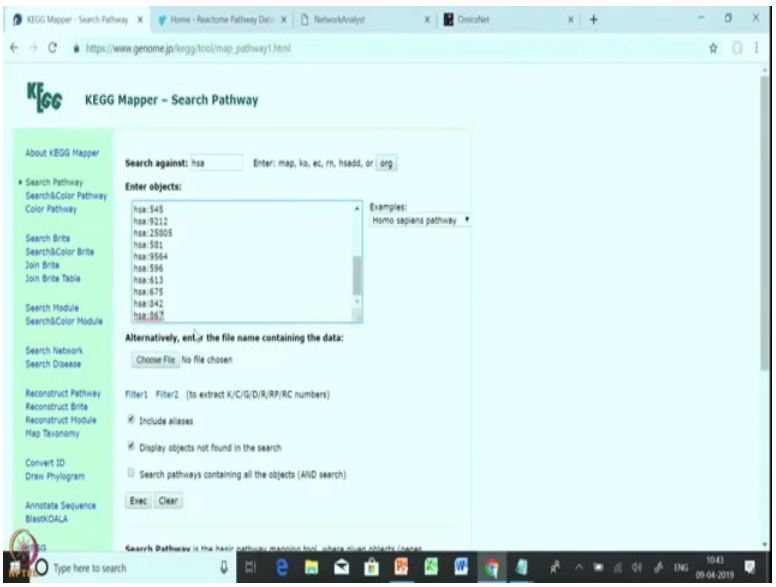


(Refer Slide Time: 07:06)



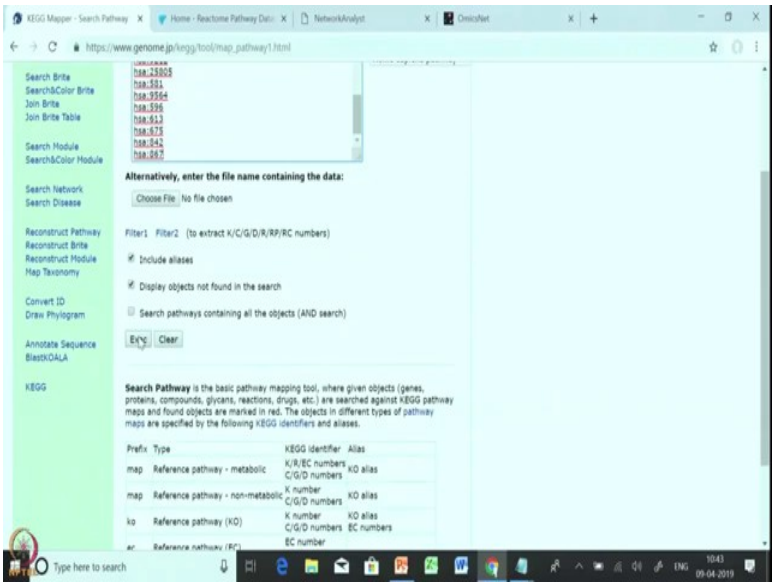
So, we will copy paste the complete KEGG ID from here. So, we have to come to the home page back again, there is a option of search pathway.

(Refer Slide Time: 07:16)



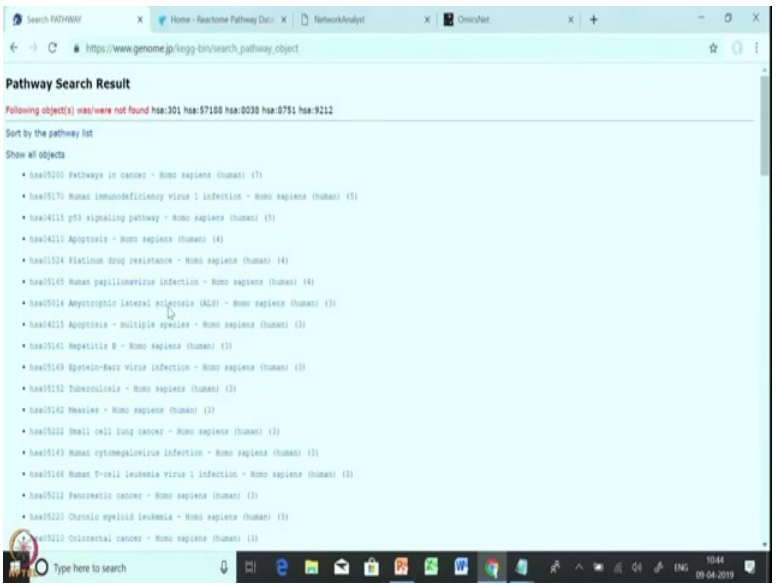
So, we have to click the options search pathway here and the search pathway dialog box will open. So, here we have to paste the KEGG ID which we have already converted and here we have to select the Homo sapiens pathway.

(Refer Slide Time: 07:47)



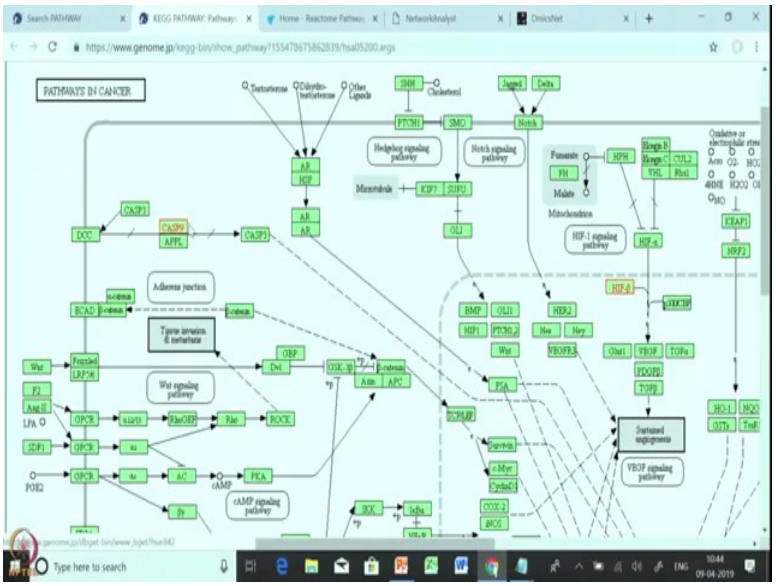
So, after selecting the Homo sapiens pathway, we have to go down and click the execution tab.

(Refer Slide Time: 07:48)



So, after clicking the execution tab you can see the KEGG Mapper has already generated a complete profile of what are the different kinds of pathways are there and how many paths, how many hits are there in each pathway.

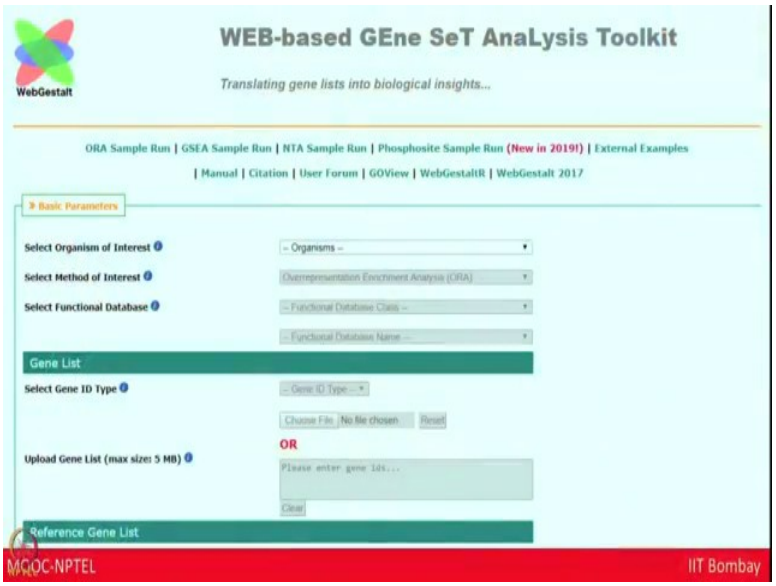
(Refer Slide Time: 08:07)



So, if we click into each of this pathway, it will redirect you to the complete to the pathway and you will found that what are the proteins that are present and they are highlighted with a yellow color and the red font.

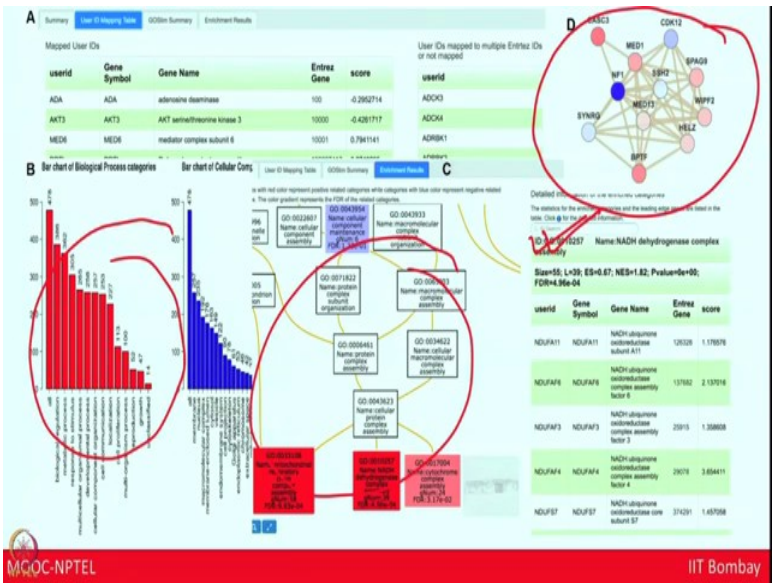
So, this is the glimpse of how you will use KEGG Mapper and how you will convert your uniprot ID into a KEGG ID. Next thing what I want to show you is a WebGestalt. I think Dr. Bing Zhang has already talked a lot about WebGestalt and I feel this is one of the best software in omics platform where it is giving a complete downloadable data downloadable image from your data sets. So, ORA sample run, GSEA sample run and NTA sample run are three important platform that WebGestalt is providing apart from this in 2019.

(Refer Slide Time: 08:54)



They have also included phosphosite sample run into this software. So, as Dr. Bing Zhang has already give a very good glimpse of this software.

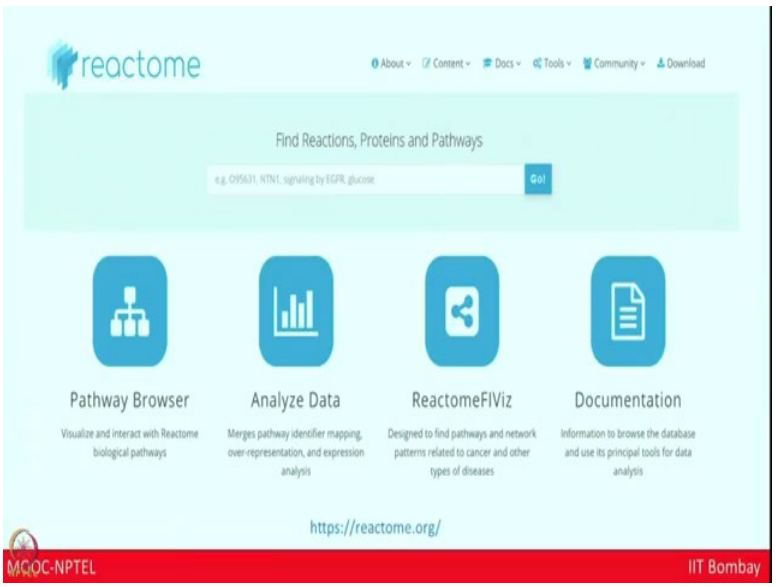
(Refer Slide Time: 09:05)



So, I will I want to show you the different kind of images, downloadable images it is providing. So, in the left you can see it is providing a complete list of different kind of classification like starting from biological pathways, cell components and so on. In the middle there is a complete list of gene ontology, different kind of gene ontology from your data set, what they are coming and how they are linked.

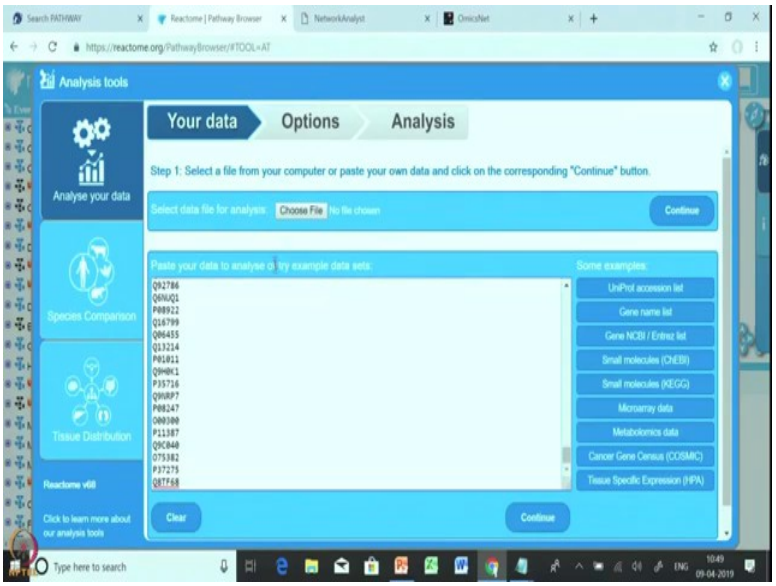
Apart from this whatever I was talking about WebGestalt gene conversion. So, they are also providing a complete conversion of your ID into different kinds of gene name and entrance gene. Apart from this it is also giving a come glimpse of the network whatever you are getting through PPI interaction protein-protein interaction module.

(Refer Slide Time: 09:54)



So, now I will be talking about a very new, but a widely used software that is reactome. So, reactome is nothing as our database, which can help you to link your proteins, link your candidates with a different kind of pathway. This database, this software is so much robust and dynamic that it will not only end up with giving a single pathway rather than it will give you a different kind of sub pathway and sub network and followed by single single reaction.

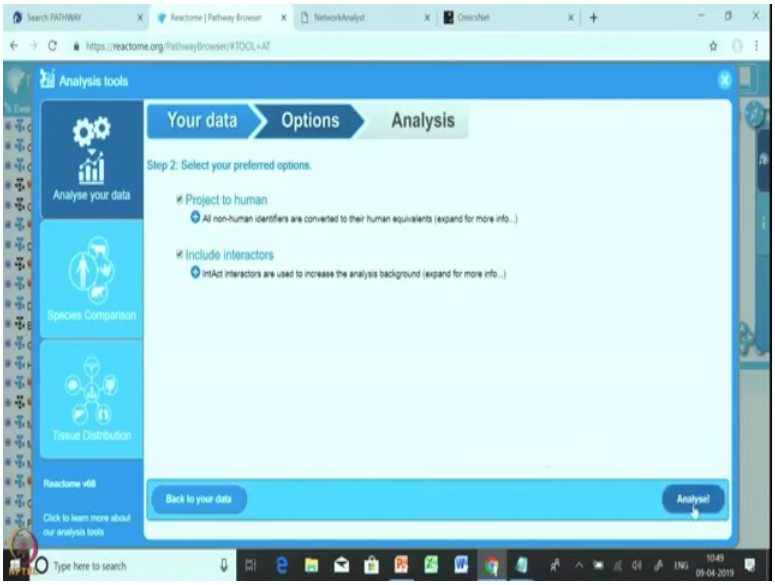
(Refer Slide Time: 10:29)



So, after clicking the analyzed data, the another window is opening which is asking for to submit your data. So, here we can submit the data in two way; first choosing a file, where we

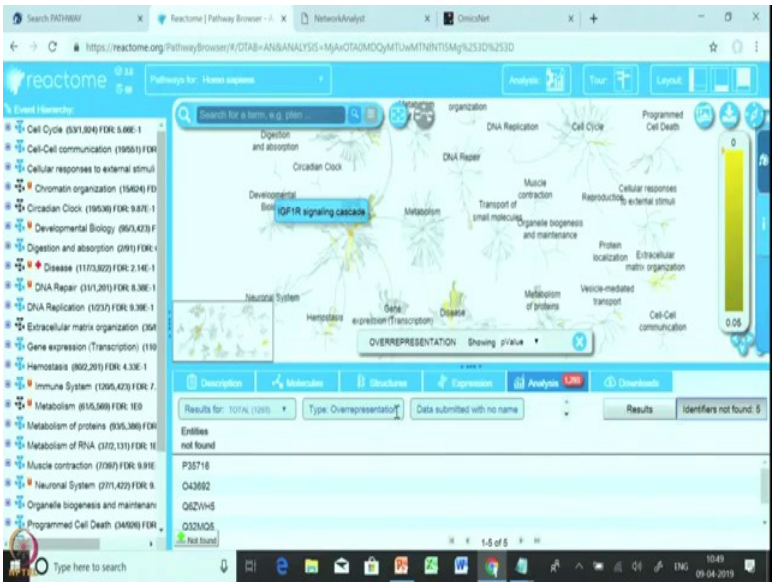
can choose a text file with the name of the candidates. Apart from this we can just simply go to the box and copy paste our candidate gene. After that we have to select the continue.

(Refer Slide Time: 10:55)



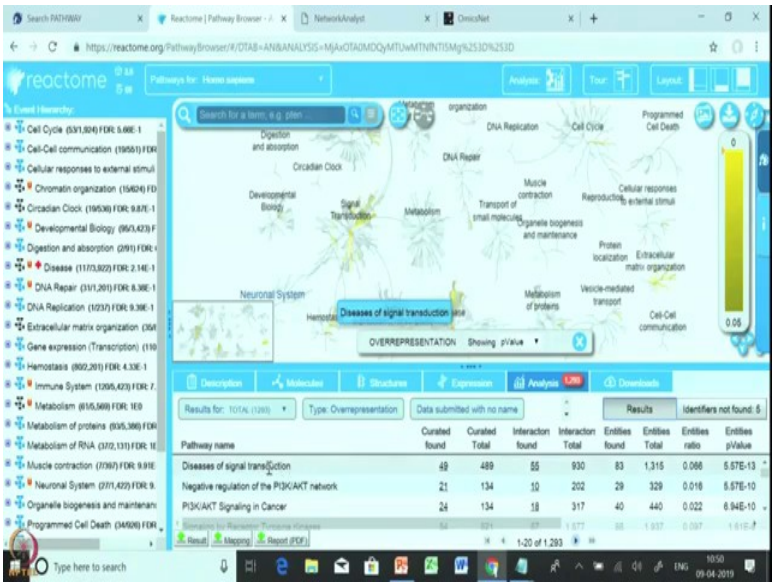
So, here a two option; the first option is project to human and the second option is include interactors. Include interactors means; what are the proteins that is interacting with your current candidates or what are the chemical compounds that mainly are metabolites that is in including different kinds of drugs that can be link with your candidates will be also shown.

(Refer Slide Time: 11:24)



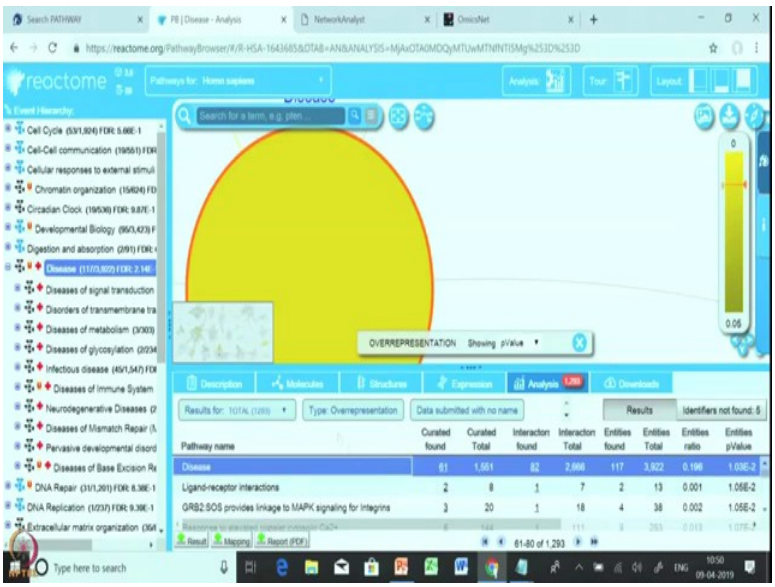
So, we will be clicking here and we will start analyzing the data. So, as you can see there is a complete list of different kind of pathways they have given and here they have also given what are the identifiers that are not found, that may be due to the upgradation of the databases.

(Refer Slide Time: 11:40)

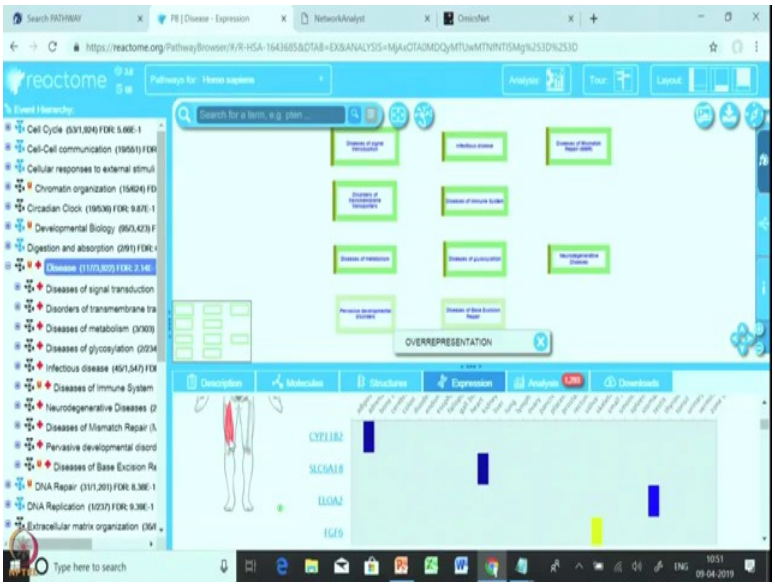


So, now if we want to check the top pathway which has coming that is there to the disease of signal transduction. So, when we will be clicking this that pathway we can find, the reactome database will show the complete details of the pathway. And when we will zoom in the pathway, we will found we will find that this has given a complete glimpse what are the different sub networks that are present.

(Refer Slide Time: 11:48)

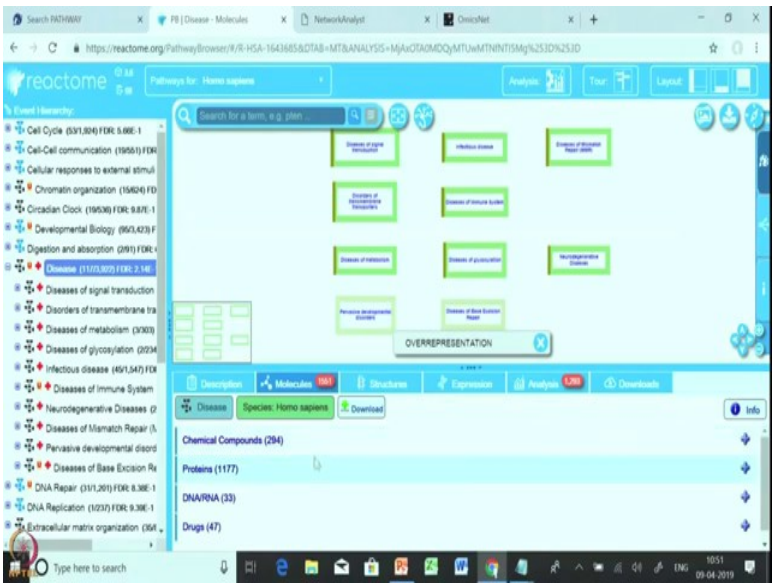


(Refer Slide Time: 12:09)



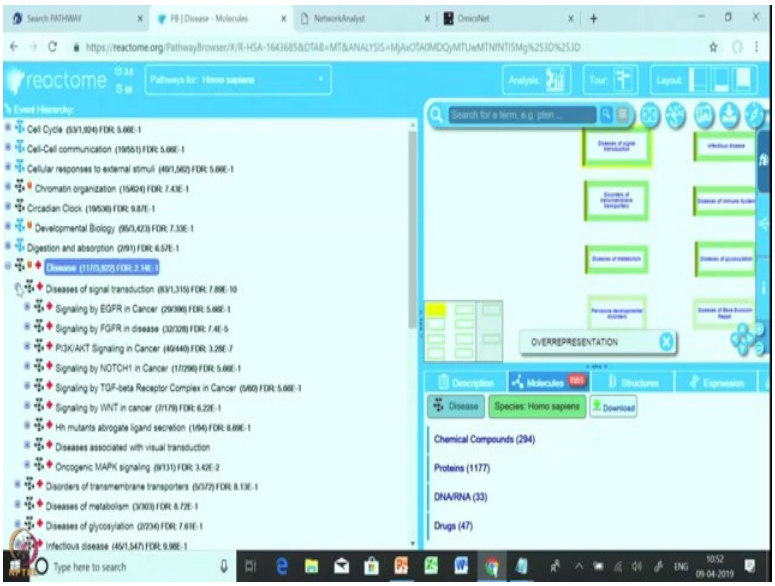
So, apart from this, if we come to the expression, the expression is nothing, but whatever the different candidates that are present in the pathway and what are their expressions throughout in different tissues are present here. So, if we select different kinds of tissues from here and we can find that what is the expression level of that candidate in the tissues.

(Refer Slide Time: 12:34)



If we go to the molecule tab over here, we will find that a couple of options are already available; that means, chemical compound, proteins, DNA, RNA and drugs. This says what are the different kind of molecules, that are present in this pathway of which chemical compounds are mainly the metabolites, different kinds of proteins, DNA and RNA and different kind of drugs.

(Refer Slide Time: 12:58)

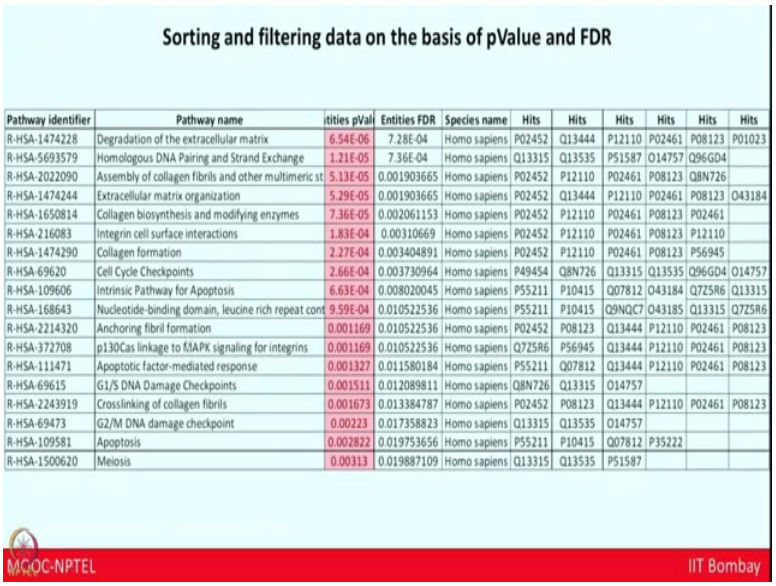


If I am selecting one disease and we can found there are different kind of sub pathways that was coming and in this sub pathway there are two important symbols that is 1 is U and 1 is

plus (+). So, if we keep the pointer over here we can find that the U says that this is the updated databases and the plus is it is related to a disease. So, likewise reactome gives a lot of information about your data set to large aspect.

So, now, the important thing is like downloading the result file. So, here is a tab of downloading the result file, where clicking this one will download the result in .CSV format. So, the .CSV format will have all the datasets and all the complete data file of the analysis. So, from there we have to select certain criteria like P value or FDR, which are already given as you can see. So, on the basis of that we have to select and we have to filter the complete analysis.

(Refer Slide Time: 14:01)

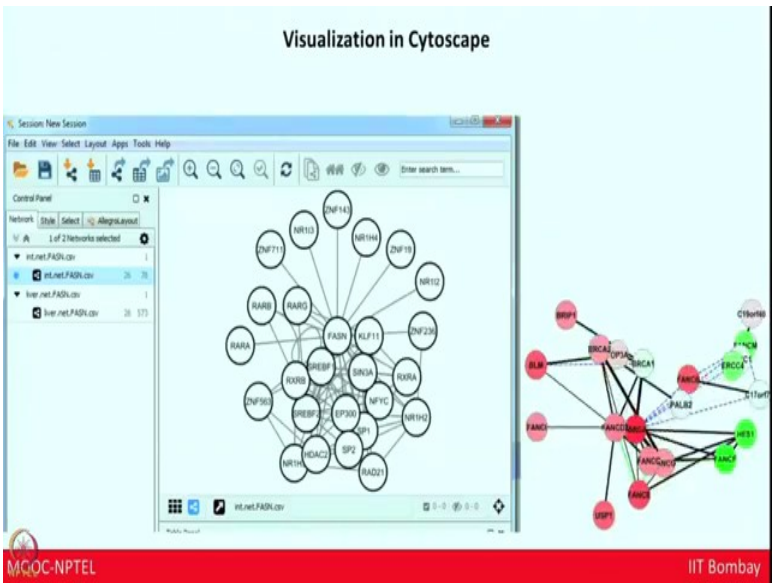


Pathway identifier	Pathway name	itities pVali	Entities FDR	Species name	Hits	Hits	Hits	Hits	Hits	Hits
R-HSA-1474228	Degradation of the extracellular matrix	6.54E-06	7.28E-04	Homo sapiens	P02452	Q13444	P12110	P02461	P08123	P01023
R-HSA-5693579	Homologous DNA Pairing and Strand Exchange	1.21E-05	7.36E-04	Homo sapiens	Q13315	Q13535	P51587	O14757	Q96GD4	
R-HSA-2022090	Assembly of collagen fibrils and other multimeric st	5.13E-05	0.001903665	Homo sapiens	P02452	P12110	P02461	P08123	Q8N726	
R-HSA-1474244	Extracellular matrix organization	5.29E-05	0.001903665	Homo sapiens	P02452	Q13444	P12110	P02461	P08123	O43184
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	7.36E-05	0.002061153	Homo sapiens	P02452	P12110	P02461	P08123	P02461	
R-HSA-216083	Integrin cell surface interactions	1.83E-04	0.00310669	Homo sapiens	P02452	P12110	P02461	P08123	P12110	
R-HSA-1474290	Collagen formation	2.27E-04	0.003404891	Homo sapiens	P02452	P12110	P02461	P08123	P56945	
R-HSA-69620	Cell Cycle Checkpoints	2.66E-04	0.003730964	Homo sapiens	P49454	Q8N726	Q13315	Q13535	Q96GD4	O14757
R-HSA-109606	Intrinsic Pathway for Apoptosis	6.63E-04	0.008020045	Homo sapiens	P55211	P10415	Q07812	O43184	Q7Z5R6	Q13315
R-HSA-168643	Nucleotide-binding domain, leucine rich repeat cont	9.59E-04	0.010522536	Homo sapiens	P55211	P10415	Q9NQC7	O43185	Q13315	Q7Z5R6
R-HSA-2214320	Anchoring fibril formation	0.001169	0.010522536	Homo sapiens	P02452	P08123	Q13444	P12110	P02461	P08123
R-HSA-372708	p130Cas linkage to MAPK signaling for integrins	0.001169	0.010522536	Homo sapiens	Q7Z5R6	P56945	Q13444	P12110	P02461	P08123
R-HSA-111471	Apoptotic factor-mediated response	0.001327	0.011580184	Homo sapiens	P55211	Q07812	Q13444	P12110	P02461	P08123
R-HSA-69615	G1/S DNA Damage Checkpoints	0.001511	0.012089811	Homo sapiens	Q8N726	Q13315	O14757			
R-HSA-2243919	Crosslinking of collagen fibrils	0.001673	0.013384787	Homo sapiens	P02452	P08123	Q13444	P12110	P02461	P08123
R-HSA-69473	G2/M DNA damage checkpoint	0.00223	0.017358823	Homo sapiens	Q13315	Q13535	O14757			
R-HSA-109581	Apoptosis	0.002822	0.019753656	Homo sapiens	P55211	P10415	Q07812	P35222		
R-HSA-1500620	Meiosis	0.00313	0.019887109	Homo sapiens	Q13315	Q13535	P51587			

So, after sorting and filtering a data on the basis of P value and FDR, I have found some top pathways, which I will be taking for the next part of the analysis. So, as you can see this is the table I have taken from the result file and you can find the pathway identifier. These are nothing, but the unique identifier ID of each pathway in reactome these are the pathway name, this is the P value, this is the FDR species name and these are the hits; that means, these what are the proteins from your sample data is matching with this pathway.

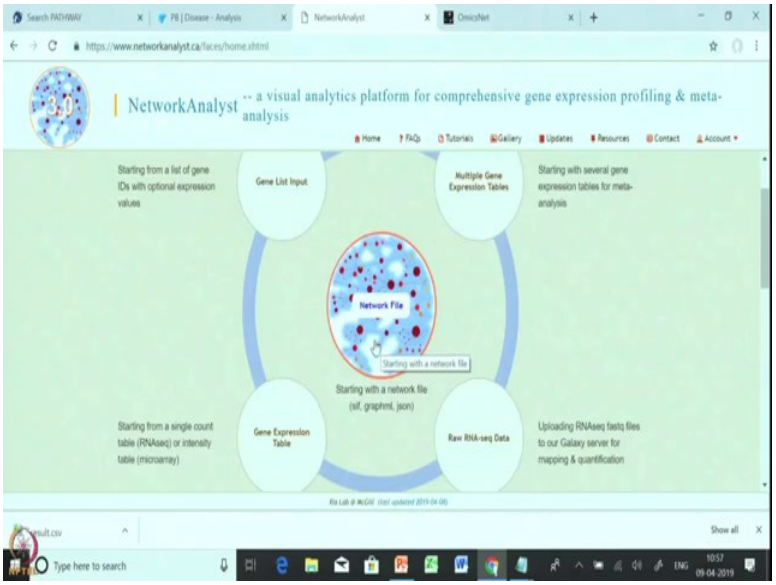
So, this data we can put into any kind of protein-protein interaction module and from there we can take the .zip file or the JSON file and we can put it directly into the Cytoscape and check what is the visualization is coming.

(Refer Slide Time: 14:46)



So, as Cytoscape can here is having different kind of plugins, we can generate different kinds of visualization network, but apart from this today, I will I want to show you a very robust visualization, a visual analytic platform for comprehensive gene expression, profiling and meta-analysis, that is network analyst.

(Refer Slide Time: 14:59)

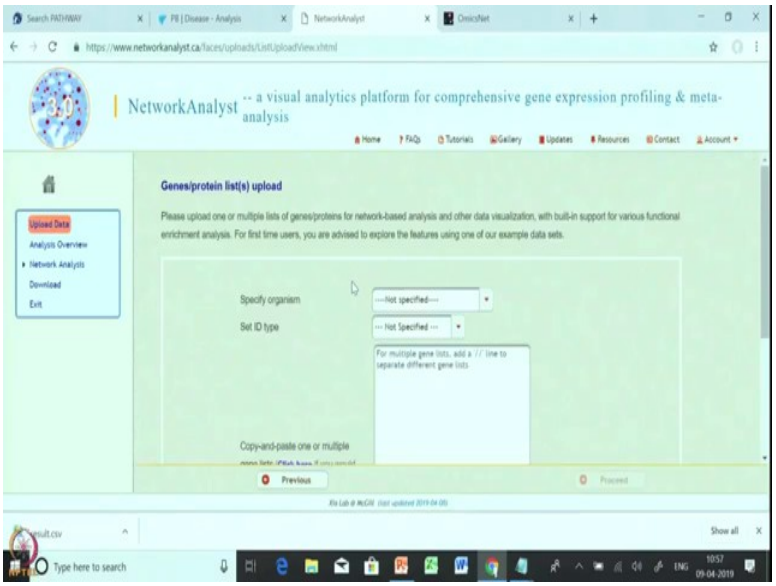


So, network analyst will give you different kind of visualization platform where you can do single analysis and multiple analysis, even multi gene expression analysis. So, now, we will go to the next hands on that how the data that we have already generated from the pathway,

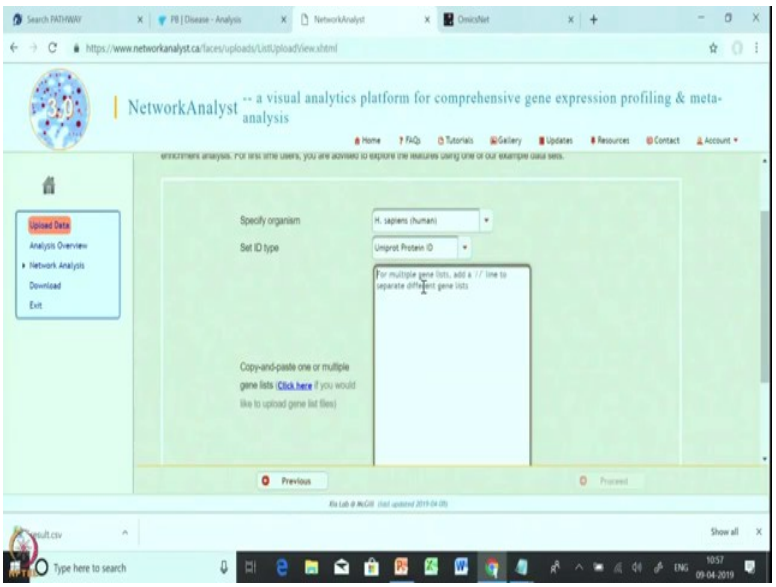
we can link in to a network platform to generate a network analysis. This is the homepage of network analyst where there are four top platforms.

First is a gene list input, where when you are having a normal gene list with p-value or fold change we can use this one. This one is the multi gene expression tables where multiple genes with different expressions can be checked. This is a gene expression table where micro RNA microarray and RNA sequence data can be done and this is the raw RNA sequence data from where we can take the sequencing data and we can start with. This is the network file where analyzed file of .SIF or JSON can be directly incorporated and we can get the visualization.

(Refer Slide Time: 16:18)

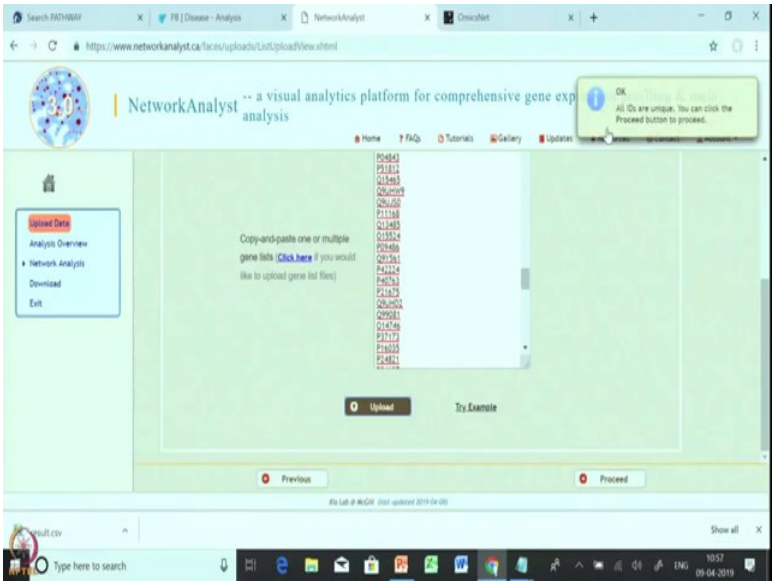


(Refer Slide Time: 16:22)



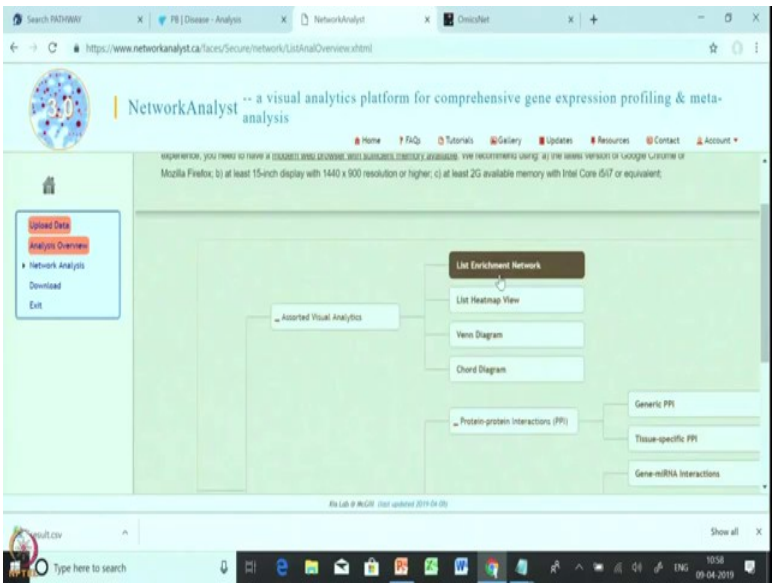
After clicking the gene list input we will be having this home page, where we have to select the organism that is *Homo sapiens* we have to select the ID, as we are taking the IDs from the same text one file. So, we know that this is a uniprot ID, then we have to copy paste the ID names.

(Refer Slide Time: 16:43)



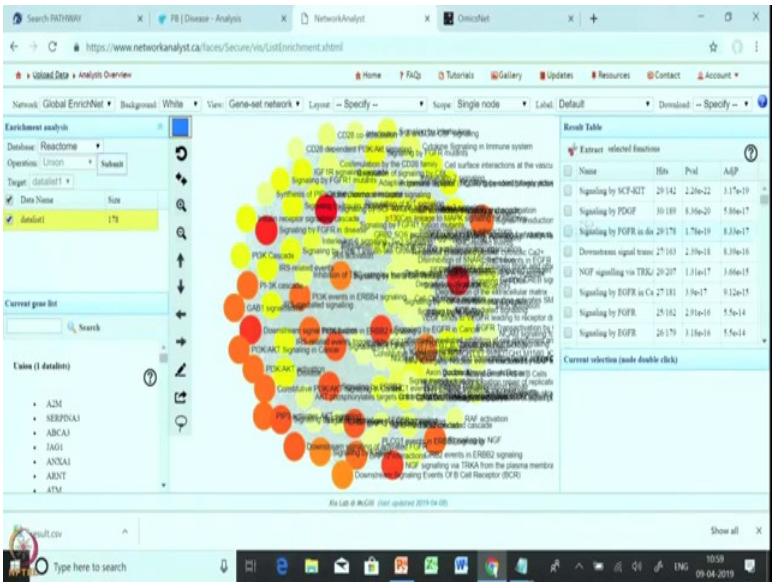
After this, we have to select the upload here and if we are getting any kind of duplicates or errors. So, it will be shown over here. If everything is fine we have to select the proceed option.

(Refer Slide Time: 16:54)



So, first we will check the list enrichment network, that is a pathway enrichment network visualization platform, where after clicking this you can see there is a complete network interaction modules, generated from different kinds of pathway.

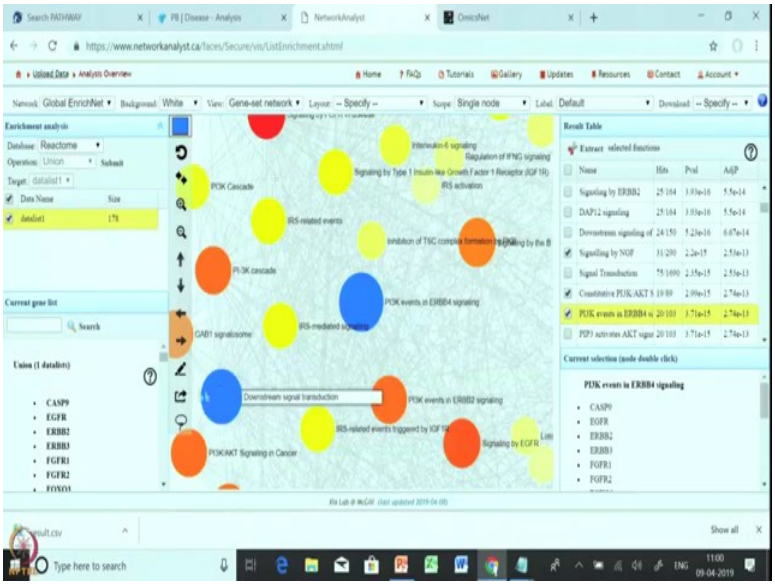
(Refer Slide Time: 17:04)



So, in the left we can select the reactome database. After this we have to change the background color to white and submit the database. After submission you can see there is already a huge number of pathways are already there and we really do not need this many pathways as it in making the complete network very much complex. So, now, we will go to

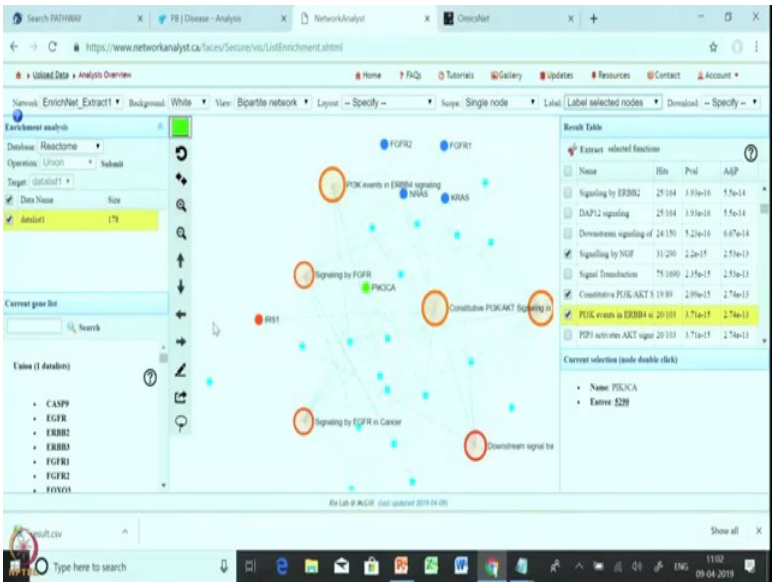
our table sort table that we have generated from the reactome and we have all the pathways that are present and we have taken on the basis of significant p-value.

(Refer Slide Time: 17:56)



And we will select those pathway which is already present there like this one downloading signaling, Matrix signaling of EGFR, signaling of FGFR, signaling of NGF and like this way we have to select some of the pathways from here and we have to come to and we have to extract this pathways.

(Refer Slide Time: 18:18)



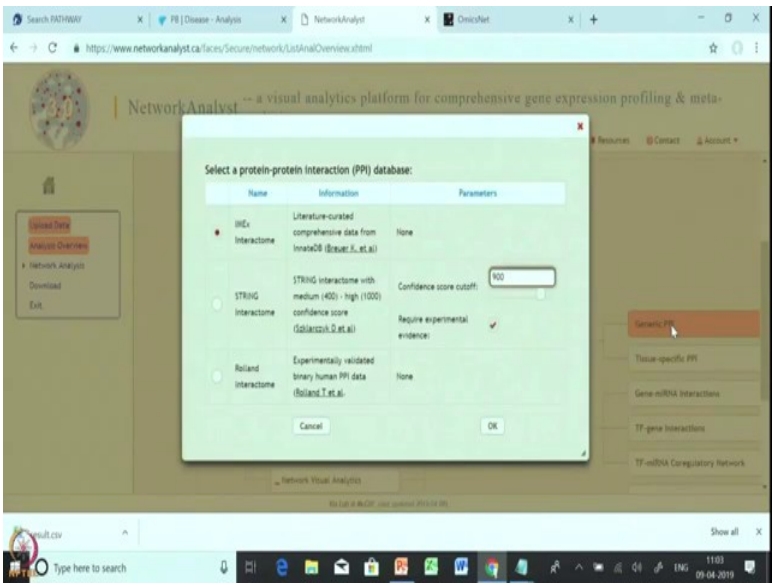
So, after extracting these pathways we can see there are only few pathways which we have extracted. So, just give you the glimpse, I have selected some few pathways, but your data set may have different top pathways that you can take into account. So, after this there is option of view and from here we will be selecting the bipartite network that will not only give you the name of the pathway, but also it will give you the name of the proteins.

So, as you can see before, whatever pathways we have selected are already present there and apart from this whatever the proteins that you have submitted in your data set is already available now. Now, if we select each pathway each proteins from like this and there is option of label and label the selected nodes. So, already these pathways, these proteins are already labeled. So, by this way we can select different proteins, different candidates, according to our data set and we can select those and highlight those protein.

Even we can change the color of each protein, like if I want to show that this protein is up regulated so, I can put in red color whereas, this protein is down regulated so, I can put green color and here again, I have to select the label selected protein and it will show that IRS1 is up regulated one whereas, the PINCI1 is the down regulated one. So, there is a lot of thing that can be done in this list of bipartite network and network analysis. So, now, we know how to generate a very good pathway enrichment model. So, the same way we can go for the protein-protein interaction model.

So, to get the protein-protein interaction model, we have already uploaded our data set in network analysis. We will be choosing the Generic PPI. So, there is another very good platform, that is a tissue specific PPI like if someone is working in brain or someone is working with kidney. So, there are already this kind of tissues are already available in the database and they can check.

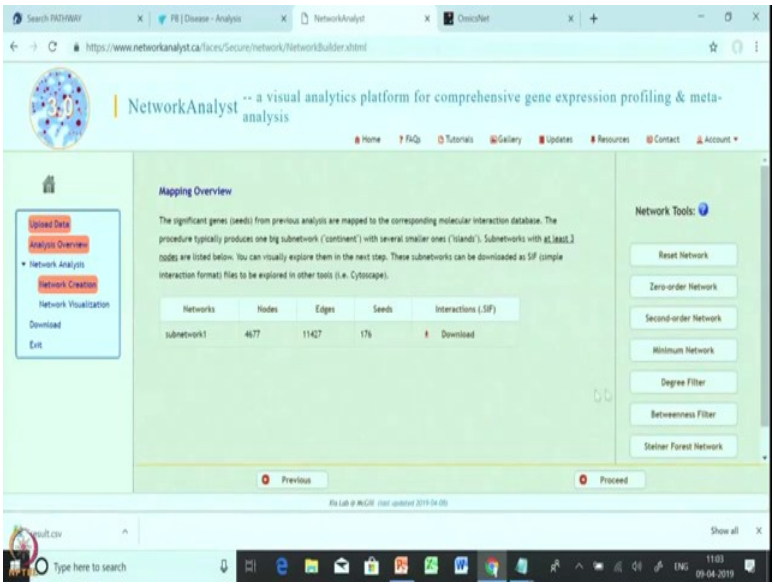
(Refer Slide Time: 20:46)



But as I just want to give you the glimpse so, I will be choosing the generic PPI, where three names of the databases are already there. So, this three are the PPI that is protein-protein interaction database one is IMEx interactome, STRING interactome and Rolland interactome.

So, people generally use STRING, but IMEx interactome will give you a very big profile of different kind of interactors that are present, which they mainly update their database from the curated literature.

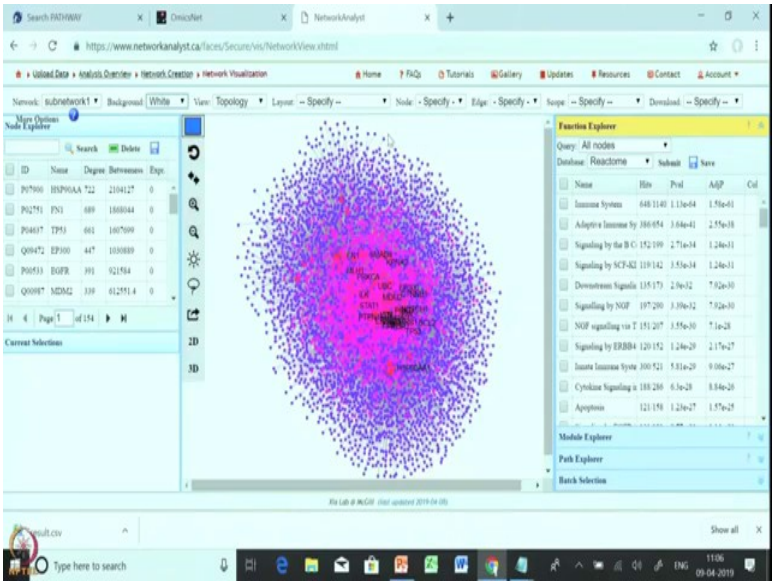
(Refer Slide Time: 21:21)



So, if I am choosing the IMEx interactome and we can see like there is a one sub network with 4677 nodes 11427 edges 176 seeds. So, from here you can download the .sif file of the interactions and we can upload it again into the network for future use.

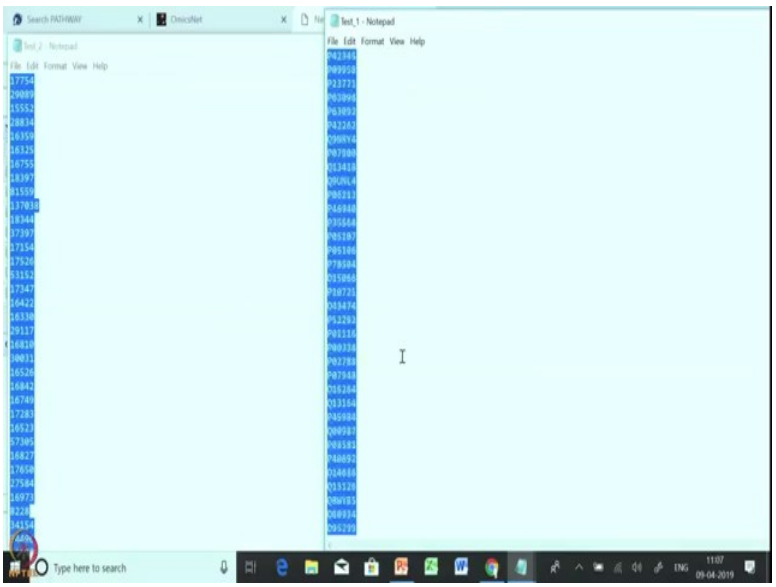
So, now we will proceed and we will found the database, the software has generated a complete protein-protein introduction module, which is a big; which is a big module and now, I will show you how to make this module small or informative and how to decrease this complexity of the network.

(Refer Slide Time: 22:04)



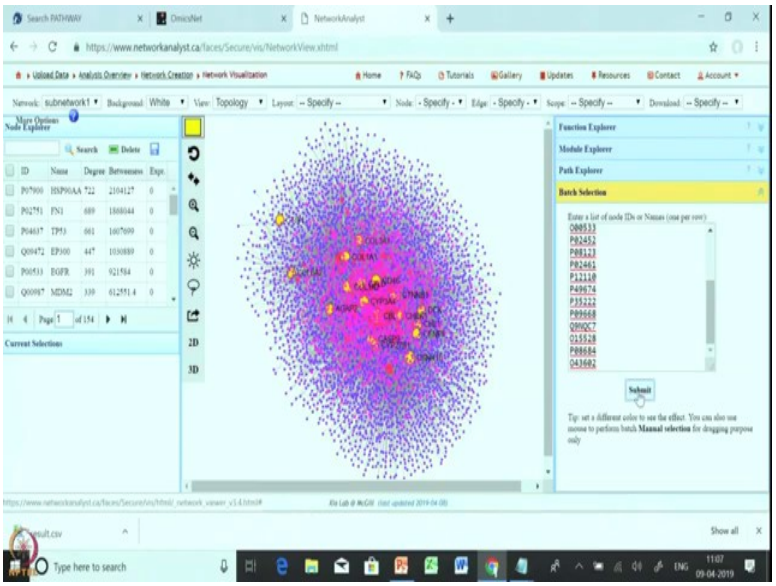
So, as you can see the software has generated the complete protein-protein interaction module, which is really very complex. So, first we have to select what is the database we want to choose. So, let us go with reactome database and after that we will be changing the color to white.

(Refer Slide Time: 22:28)



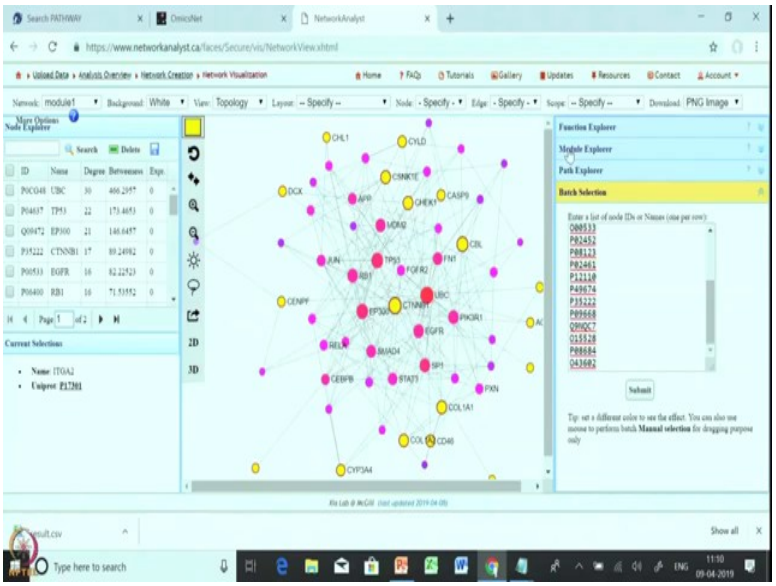
Now, I have already given you a in the text one file. So, now there is an option of batch selection, which says that whatever what are the proteins that we are interested in, we can copy paste those protein accession ID and after that we have to click the submit. So, after submission we can see there is a highlighted candidates that we can found in this complex network.

(Refer Slide Time: 22:50)



So, now, as we are very much interested with these candidates, we will select an extract those candidate from this complex network. So, after selection and extracting the candidates we found that these are the proteins that we have selected and is present in this network.

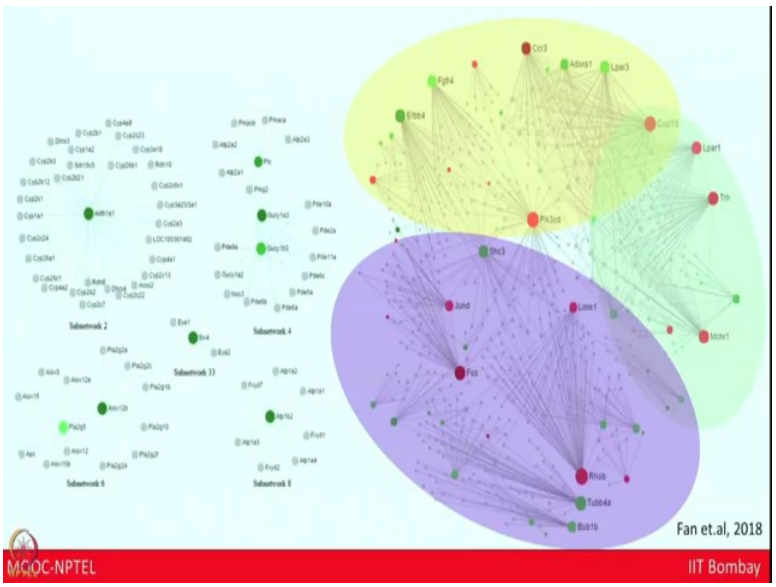
(Refer Slide Time: 23:08)



Apart from this, these proteins which are there are the top much interactors that is coming in this protein protein interaction modules. So, like this we have to make some adjustment, to make this network visually interpretable. So, that can be also done from their given layout which are different kind of layouts are already available. But, for reducing the overlapping, I will choose this one to reduce the overlap and as you can see the complex network has got some clarity.

So, after adjusting little manually, we can download this one with as a PNG image and we can save the file as a PNG or JPG. So, now we know. So, now, we know how from a data set we can generate the pathway enrichment model, protein-protein interaction model. So, after this I will show you an example of a recent paper that got published in nature scientific by Fan et al in 2018.

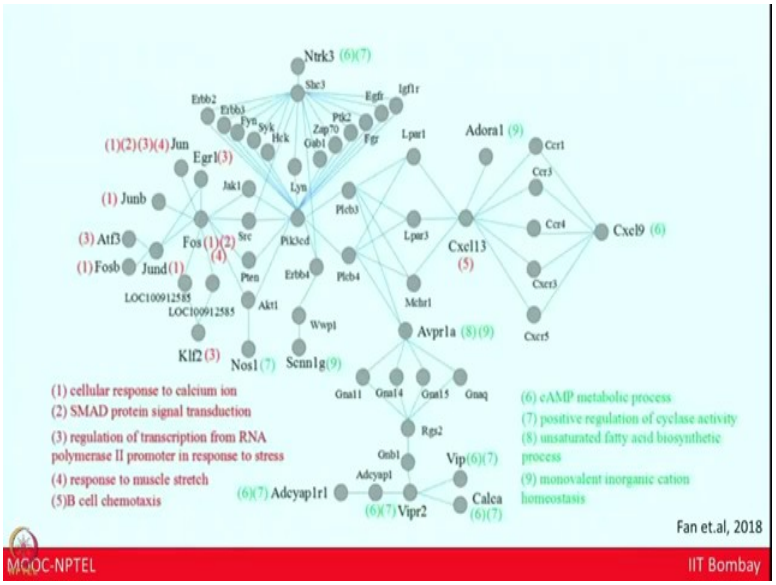
(Refer Slide Time: 24:29)



So, they have given a very good they have used this network analysis software to large extent and they have given how this network visualization platform can be used to produce this kind of network analysis images.

So, in the first one you can see, they have differentiated this protein, can the differentiated the protein candidate list in up regulated and down regulated manner. And, they have also generated different sub networks like this four sub networks they have generated and on the basis of different pathway of protein protein interaction. So, now, if I want to check like what are the different clusters that is coming in my protein-protein introduction module in terms of pathway that can also be done, checking the curated sorted reactome list that we have generated.

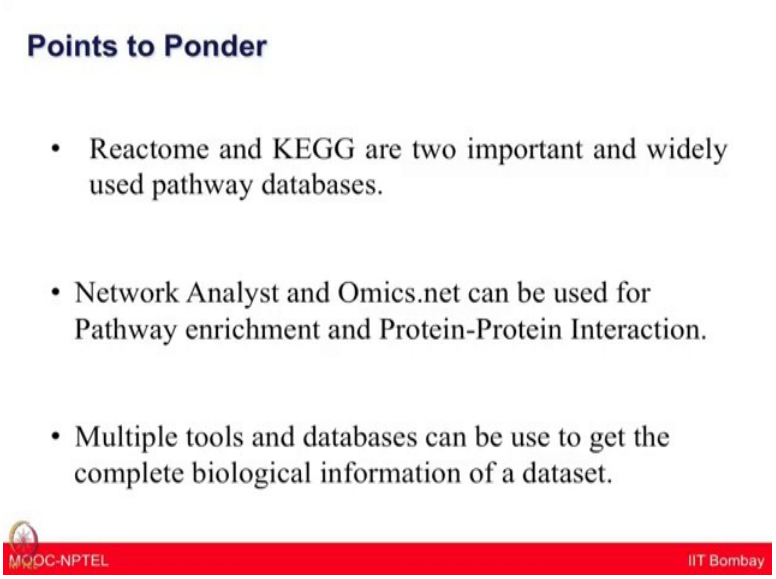
(Refer Slide Time: 25:23)



Apart from this, they have given a complete view of what are the proteins present and whether the single candidate is present in multiple pathway or single pathway with this help of the, with the help of this diagram.

So, that is all thank you.

(Refer Slide Time: 25:41)



I hope today's session was informative for you all, where you got an idea how tertiary analysis plays an important role in data analysis. As Deep mention to you a data set can be used in different ways to extract the biological information. However, there are many

parameters that need to be considered to obtain the meaningful data sets. He also showed you that how combination of different software tools can be used to obtain a very good interpretation of your data and what strategies need to be used to obtain more meaningful information.

I will suggest you to practice the tools mentioned today. You can download data set and explore these kind of tools by yourself. In the next supplementary video, we will talk about the case studies in cancer.

Thank you.