

Interactomics: Basics and Applications
Prof. Sanjeeva Srivastava
Prof. Deeptarup Biswas
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 57
Pathway Enrichment and Data Analysis

Hello students. I hope you know from the previous sessions you got motivated and you started playing with certain online tools available for you to start analyzing your own data set and more importantly how best to plot your data and start you know visualizing the data start presenting data and then start making the biological sense of the data.

So, once you have analyzed your data you end up getting a shorter list you know from the very big large big data set which was the starting point from the NGS or mass spec or microarrays. And now you know you have narrowed down to a shorter list of the most significant proteins.

But now you would like to put them together into the biological context that what are their roles these proteins belongs to which pathways, they interact with which other proteins, are they part of you know the a network or the given physiological pathway where that makes more sense right.

So, then once you start getting that idea then only probably you will have you know the follow up experiments plan, can now start thinking about a drug which might be able to inhibit and control a given pathway right. So, many time these clues only comes when you start your big data set, they start looking at the most significant list narrowing it down and then further you process after looking at the pathways and networks, they come up with a hypothesis and then look at what can be the actionable hypothesis to build the next set of experiments.

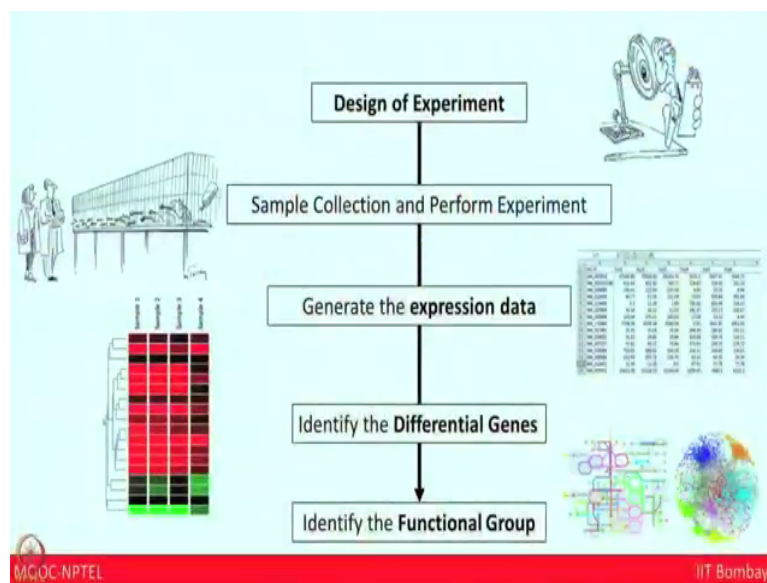
So, there are many online tools which are currently available which are really good, some of them are you know developed from lots of resources and funds provided by the initial

governments, but with mandate to make it publicly available. You can start utilizing these resources and start looking at various example data set and see whether you are getting comfortable in doing these kind of analysis.

So, today my research scholar Deeptarup Biswas, he will walk you through a different steps involved in looking at the pathway enrichment analysis and network analysis using various tools. So, let us start today's lecture.

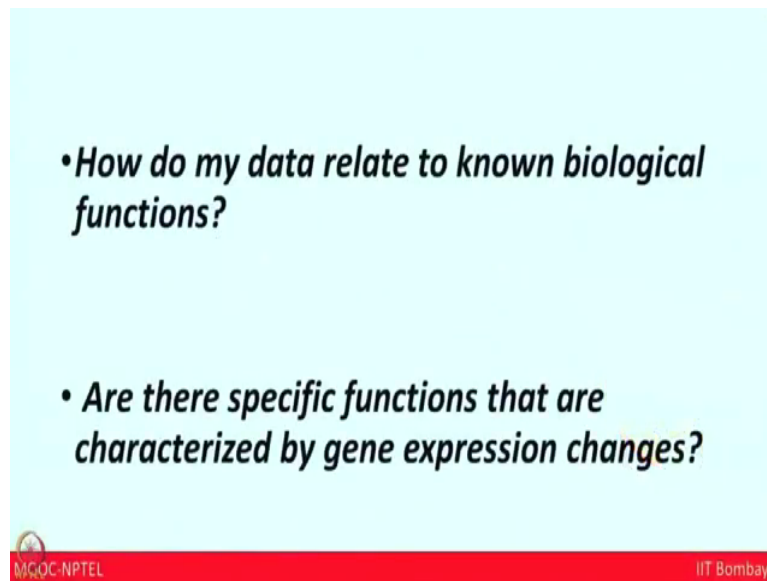
Till now we have learned a lot about statistical power primary analysis and secondary analysis to generate the expression data set.

(Refer Slide Time: 02:37)



Now, the main important thing which is coming that we got a very good pattern of differential gene regulation. So, the questions come what next?

(Refer Slide Time: 02:48)



- ***How do my data relate to known biological functions?***
- ***Are there specific functions that are characterized by gene expression changes?***

NPTEL IIT Bombay

I want to start with two important question, how do my data delayed 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 function 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: 03:30)

Different software generates different IDs

ID conversion

WebGestalt

Protein Identifier Cross-Reference

KEGG Mapper - Convert ID

https://www.genome.jp/kegg/tool/conv_id.html

Outside DB: NCBI GeneID

KEGG ID for: ☒ KEGG GENES ☐ Selected (enter organism code)

Enter outside identifiers:

Alternatively, enter the file name containing the data:

No file chosen

MOOC-NPTEL IIT Bombay

Different kinds of software's that generate different IDs. If we are using any proteome discover, commercial software or trans proteome pipeline they will give you different kinds of IDs in the protein identification. But we to start a tertiary analysis we have to get multiple ID and that is possible only through 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 taken into consideration; the first thing is David. David is a multiple ID conversion tool, apart from this it can help in also different types of annotation and enrichment studies. Next is 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 identify cross reference is a very simple tool where we just need to put the ID, 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 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.

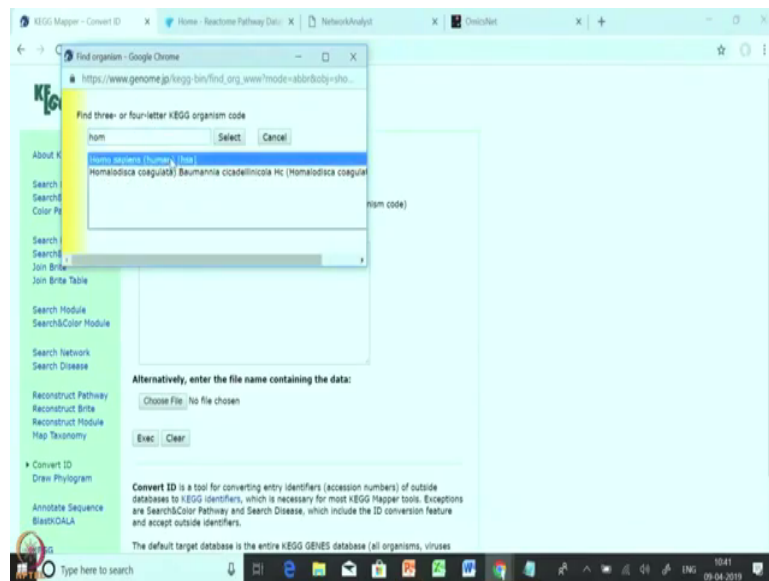
So, I will show you a glimpse like how from a test say data set we can approach to KEGG Mapper and we can convert the ID to KEGG gene and after that we can put those KEGG converted ID into the KEGG Mapper to get the pathway. So, I have already shared the test 1 text file, you have to open the text file 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 converter 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: 06:06)

The screenshot shows the KEGG Mapper - Convert ID web interface. The browser address bar displays https://www.genome.jp/kegg/tool/conv_id.html. The page has a light blue header with the KEGG logo and the title "KEGG Mapper - Convert ID". On the left, there is a green sidebar with a list of navigation links: "About KEGG Mapper", "Search Pathway", "Search&Color Pathway", "Color Pathway", "Search Brite", "Search&Color Brite", "Join Brite", "Join Brite Table", "Search Module", "Search&Color Module", "Search Network", "Search Disease", "Reconstruct Pathway", "Reconstruct Brite", "Reconstruct Module", "Map Taxonomy", "Convert ID", "Draw Phylogram", "Annotate Sequence", and "BlastKOALA". The main content area is white and contains the following elements: a dropdown menu for "Outside DB" set to "UniProt"; a section "KEGG ID for" with radio buttons for "KEGG GENES", "Selected", "Organism", and "hse" (with a note "(enter organism code)"); a text input field labeled "Enter outside identifiers:" containing a list of UniProt IDs: Q9H0K1, P35716, Q9H0P7, P00347, O00300, P11367, Q9C040, Q75382, P37275, and Q17756; an "Alternatively, enter the file name containing the data:" section with a "Choose File" button and the text "No file chosen"; and "Exec" and "Clear" buttons. At the bottom, there is a paragraph explaining that "Convert ID" is a tool for converting entry identifiers (accession numbers) of outside databases to KEGG identifiers, and a note stating "The default target database is the entire KEGG GENES database (all organisms, viruses)". The Windows taskbar at the bottom shows the date and time as 10:41 on 09-04-2019.

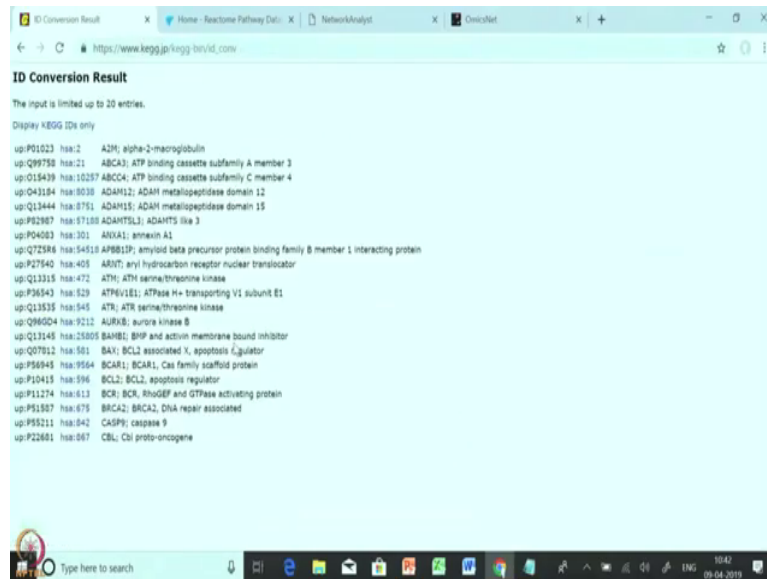
So, this is the homepage of KEGG Mapper convert ID where in the first outside DB you can choose what is the different NCBI gene ID or NCBI protein ID or UniProt ID you are putting. So, as I have given a list of UniProt ID, so I will be selecting UniProt ID here. After that the important thing is the what is the organism?

(Refer Slide Time: 06:57)



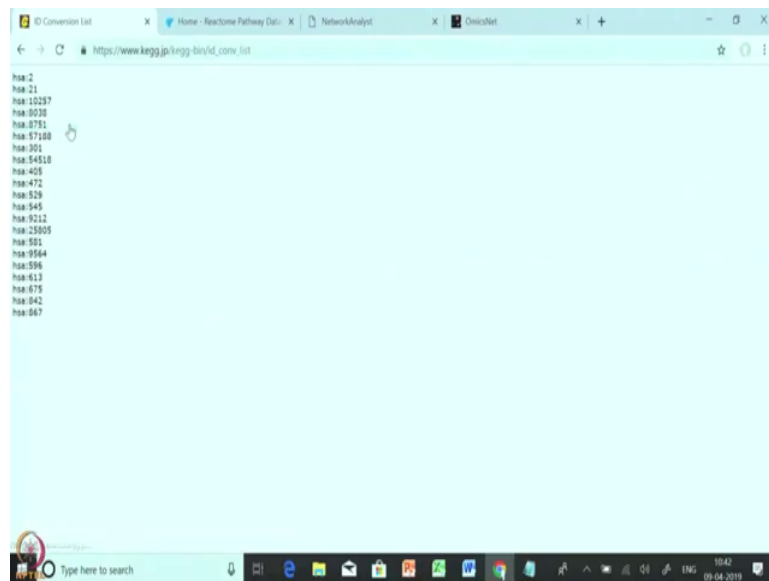
So, if when we are clicking this, so we have to write the name of the organism and as I know I have taken the file from the Homo sapiens repository. So, I will be selecting the Homo sapiens and I will select the tab and so it is showing here hsa. Here in the inter outbox outside identifier we will copy paste the list of the UniProt gene and then we will select we will click on the execution tab. So, it will take a little time and it will give you the complete converted ID from UniProt to KEGG ID.

(Refer Slide Time: 07:06)



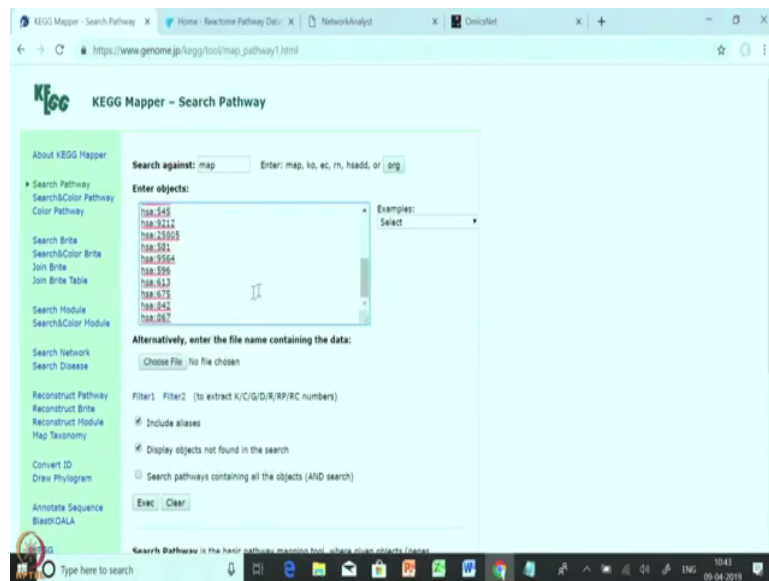
So, as you can see the conversion result is giving you the name of the UniProt list. So, we will be copying only the KEGG ID. So, we will click here display KEGG ID only where we will get the name of the KEGG IDs.

(Refer Slide Time: 07:19)



So, we will copy paste the complete KEGG ID from here.

(Refer Slide Time: 07:30)



So, we have to come to the homepage back again there is our option of search pathway. So, we have to click the option search pathway here and the search pathway dialogue box will open.

(Refer Slide Time: 07:51)

The screenshot shows the KEGG Mapper - Search Pathway web interface. The browser address bar displays https://www.genome.jp/kegg/tool/map_pathway1.html. The page has a light blue background with a green sidebar on the left containing navigation links: About KEGG Mapper, Search Pathway (selected), Search&Color Pathway, Color Pathway, Search Brite, Search&Color Brite, Join Brite, Join Brite Table, Search Module, Search&Color Module, Search Network, Search Disease, Reconstruct Pathway, Reconstruct Brite, Reconstruct Module, Map Taxonomy, Convert ID, Draw Phylogram, Annotate Sequence, and BlastKOALA. The main content area is titled "KEGG Mapper - Search Pathway". It features a "Search against:" dropdown menu set to "hsa" with a hint "Enter: map, kb, ec, rn, hadd, or .org". Below this is an "Enter objects:" text area containing the text "hsa:7167 hsa:GPI cpd:C00118 ALDOA 1.2.1.12 C00236". To the right of the text area is a dropdown menu labeled "Examples:" with "Homo sapiens pathway" selected. Below the text area is a section titled "Alternatively, enter the file name containing the data:" with a "Choose File:" button and the text "No file chosen". Further down are filter options: "Filter1: Filter2: (to extract K/C/G/D/R/PP/RC numbers)", "Include aliases" (checked), "Display objects not found in the search" (checked), and "Search pathways containing all the objects (AND search)" (unchecked). At the bottom of the main area are "Exec" and "Clear" buttons. The Windows taskbar at the bottom shows the search bar, task view button, and several application icons. The system tray on the right shows the date and time as 10:43 on 09-04-2019.

So, here we have to paste the KEGG ID which we have already converted and here we have to select the Homo sapiens pathway. So, after selecting the Homo sapiens pathway we have to go down and click the execution tab.

(Refer Slide Time: 08:01)

KEGG Mapper - Search Pathway

Home - Reactome Pathway Data | NetworkAnalyst | OmicsNet

https://www.genome.jp/kegg/tool/map_pathway1.html

Search Brite
Search&Color Brite
Join Brite
Join Brite Table
Search Module
Search&Color Module
Search Network
Search Disease
Reconstruct Pathway
Reconstruct Brite
Reconstruct Module
Map Taxonomy
Convert ID
Draw Phylogram
Annotate Sequence
BlastKOALA

KEGG

Alternatively, enter the file name containing the data:
Choose File: No file chosen

Filter1: Filter2 (to extract K/C/G/D/R/RP/RC numbers)
☒ Include aliases
☒ Display objects not found in the search
☐ Search pathways containing all the objects (AND search)

Search

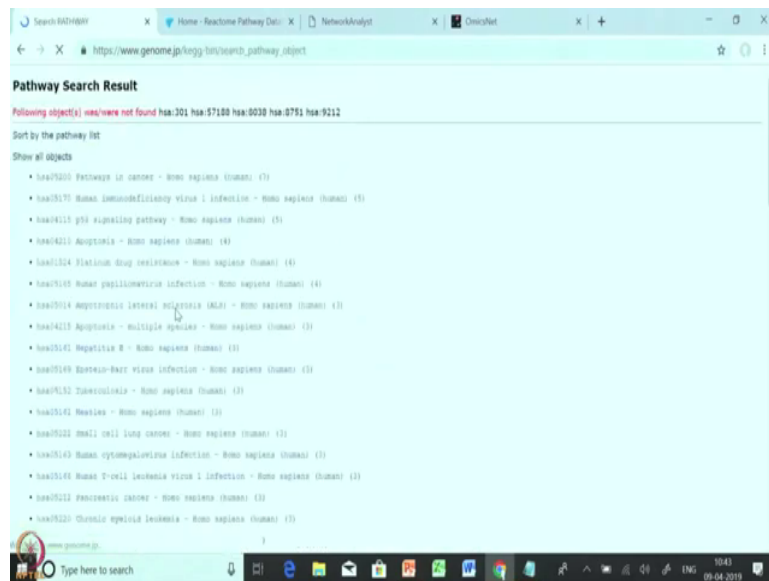
Search Pathway is the basic pathway mapping tool, where given objects (genes, proteins, compounds, glycans, reactions, drugs, etc.) are searched against KEGG pathway maps and found objects are marked in red. The objects in different types of pathway maps are specified by the following KEGG identifiers and aliases.

Prefix	Type	KEGG Identifier	Alias
map	Reference pathway - metabolic	K/R/EC numbers C/G/D numbers	KO alias
map	Reference pathway - non-metabolic	K number C/G/D numbers	KO alias
ko	Reference pathway (KO)	K number C/G/D numbers	EC numbers
ec	Reference pathway (EC)	EC number	

Type here to search

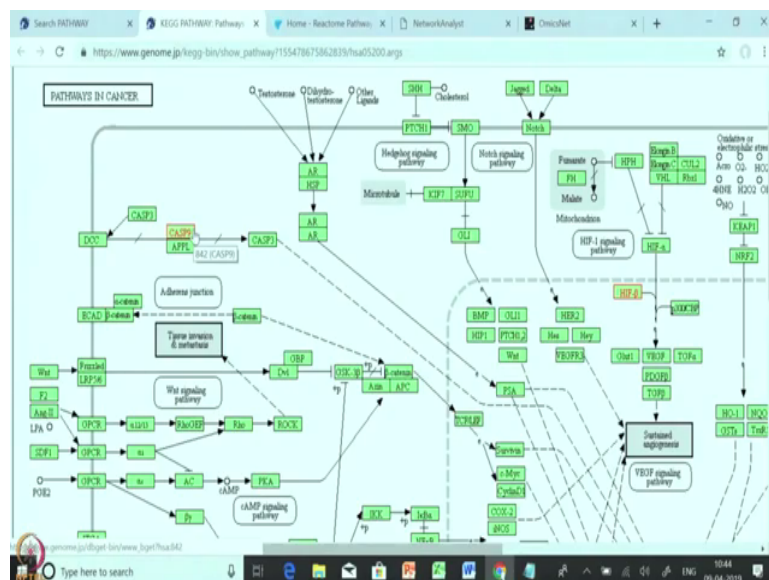
09-04-2019

(Refer Slide Time: 08:02)



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.

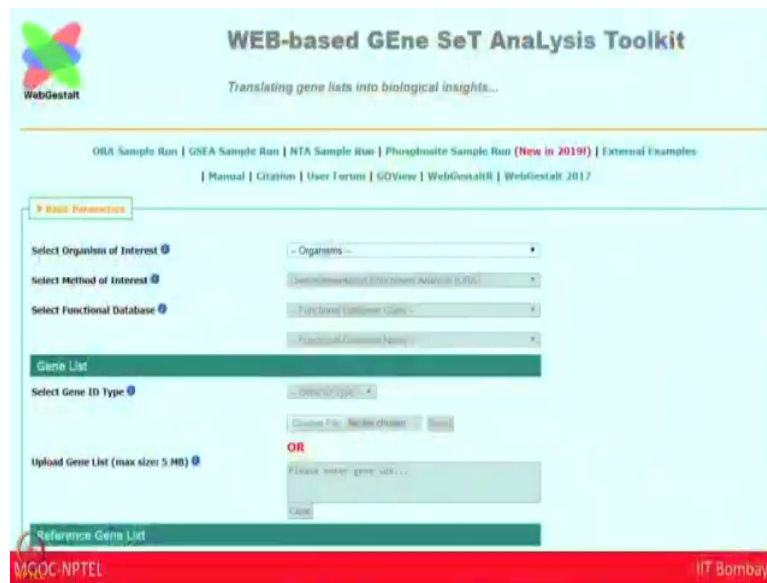
(Refer Slide Time: 08:21)



So, if we click into each of these 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 the yellow color and the red font.

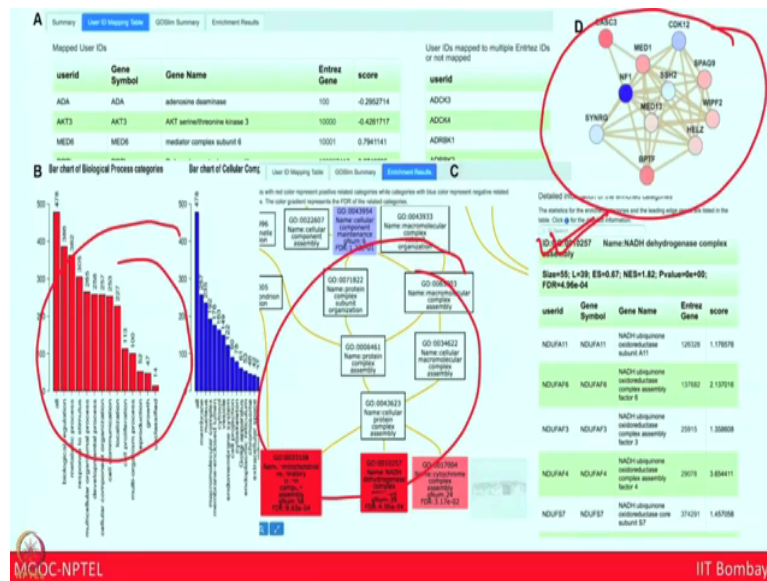
So, this is the glimpse of how you will use KEGG Mapper and how you will convert your UniProt ID into a KEGG ID. Next thing what I want to show you is a WebGestalt and I feel this is one of the best software in omics platform where it is giving a complete downloadable data, downloadable image from your data sets. So, over a sample run GSEA sample run and NTA sample run are three important platform that WebGestalt is providing. Apart from this in 2019 they have also included phosphosite sample run into this software.

(Refer Slide Time: 09:04)



The screenshot displays the WebGestalt web application interface. At the top left is the WebGestalt logo, a colorful four-leaf clover. The main title is "WEB-based GENE SeT Analysis Toolkit" in bold black text, with the tagline "Translating gene lists into biological insights..." below it. A navigation bar contains links: "ORA Sample Run", "GSEA Sample Run", "NTA Sample Run", "Phosphosite Sample Run (New in 2019)", "External Examples", "Manual", "Citation", "User Forum", "GDocView", "WebGestaltR", and "WebGestalt 2017". Below the navigation bar is a "Basic Parameters" section with several dropdown menus: "Select Organism of Interest" (set to "Organisms"), "Select Method of Interest" (set to "Overrepresentation Enrichment Analysis (ORA)"), "Select Functional Database" (set to "Functional Categories"), and "Select Gene ID Type" (set to "Gene ID Type"). There are also buttons for "Choose File" and "Select" next to the "Select Gene ID Type" dropdown. Below these is a section for "Gene List" with a text input field labeled "Upload Gene List (max size: 5 MB)" and a "Submit" button. A red "OR" label is placed between the "Select Gene ID Type" section and the "Upload Gene List" section. At the bottom left, there is a "Reference Gene List" section with a "Submit" button. The footer of the page is red and contains the text "MGCC NPTEL" on the left and "IIT Bombay" on the right.

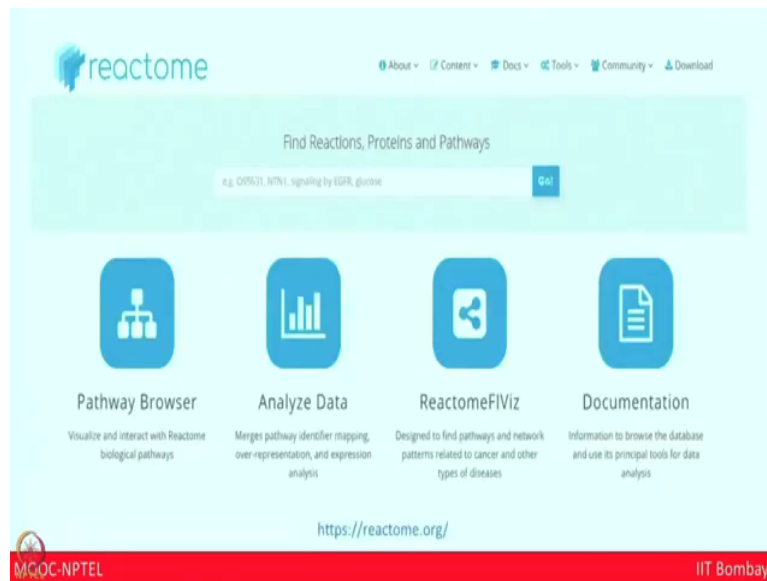
(Refer Slide Time: 09:08)



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 kinds of gene ontology from your data set what they are coming and how they are linked.

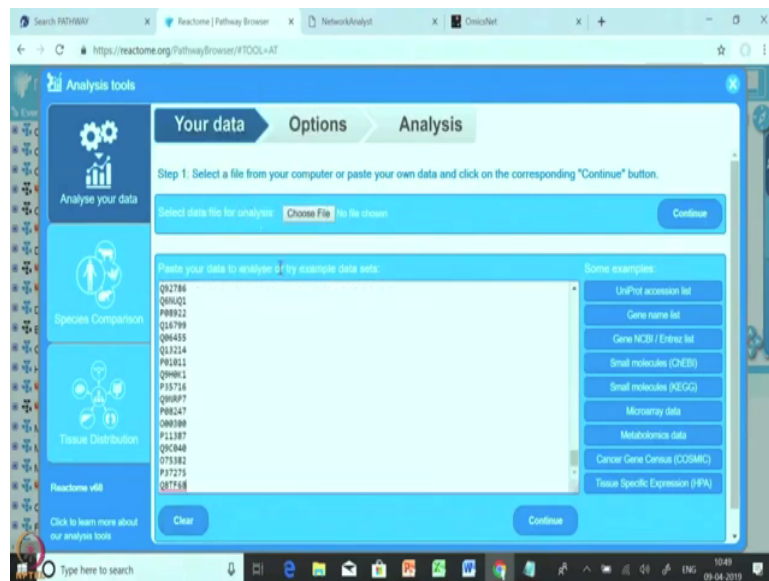
Apart from these whatever I was talking about WebGestalt gene conversion, so they are also providing a complete conversion of your ID into different kinds of gene name and entrance gene. Apart from this it is also giving a glimpse of the network whatever you are getting through PPI interaction Protein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

(Refer Slide Time: 09:57)



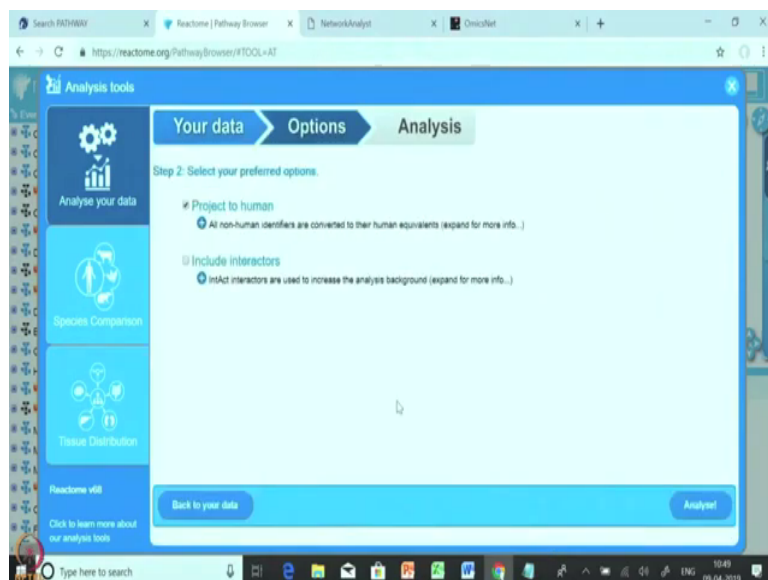
So, now I will be talking about a very new, but widely used software that is reactome. So, reactome is nothing a database which can help you to link your proteins, link your candidates with a different kind of pathway. This database this software is so much robust and dynamic, that it will not only end 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: 10:32)



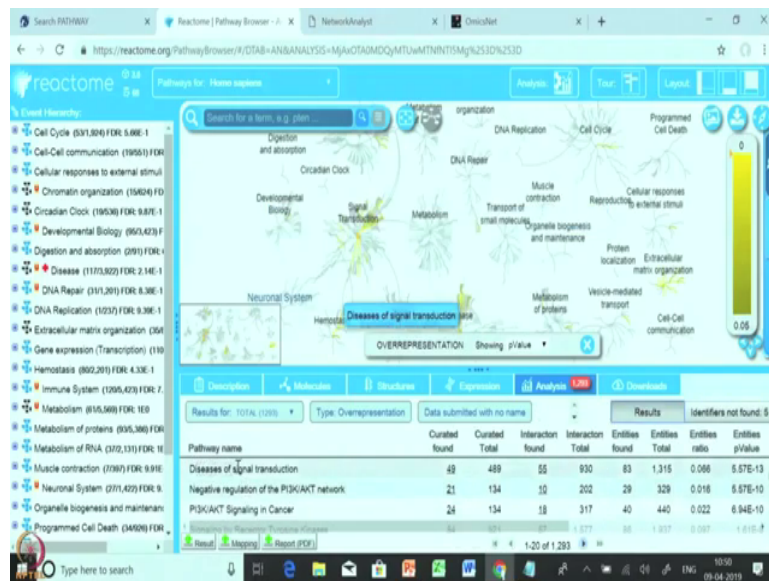
So, after clicking the analyze data the another window is opening which is asking for to submit your data. So, here we can submit the data in two way; first choosing a file where we can choose a text file with the name of the candidates apart from this we can just simply go to the box and copy paste our candidates gene.

(Refer Slide Time: 10:58)



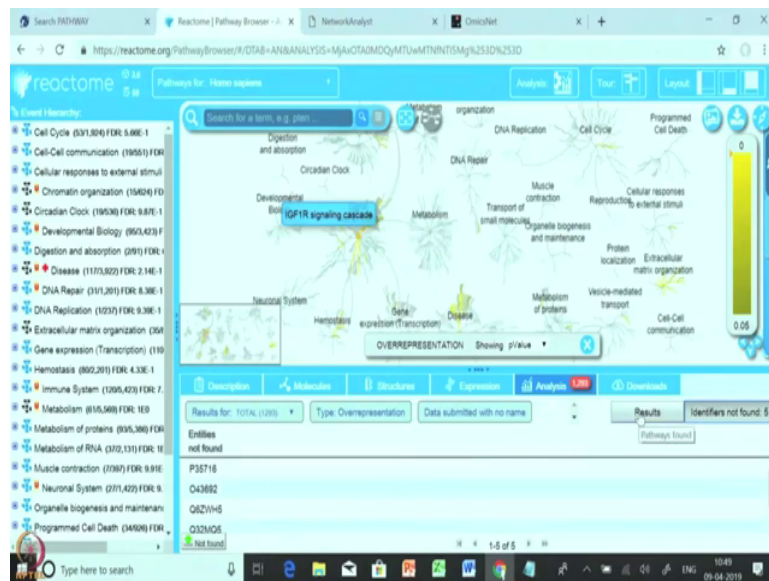
After that we have to select the continue. So, here are two option; the first option is project to human and the second option include interactors. Include interactors means what are the proteins that is interacting with your candidates or what are the chemical compounds that mainly are metabolites that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

(Refer Slide Time: 11:28)



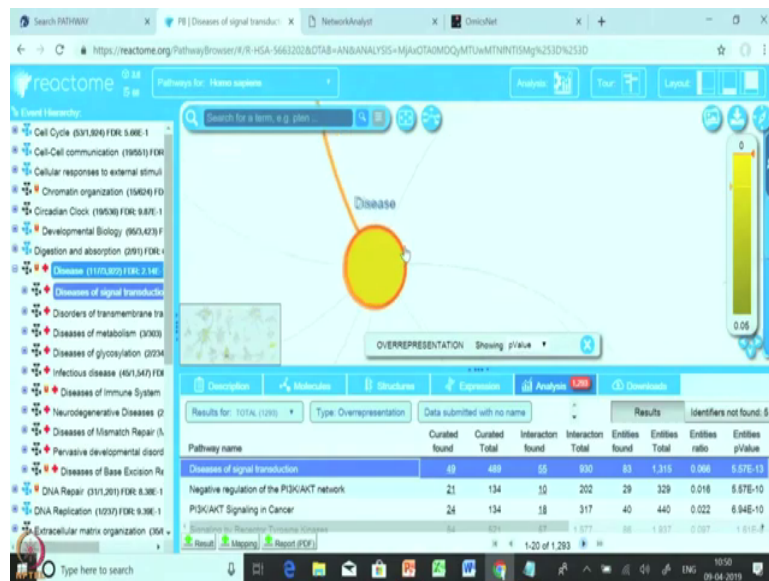
So, we will be clicking here and we will start analyzing the data. So, as you can see there is a complete list of different kind of pathways they have given.

(Refer Slide Time: 11:37)



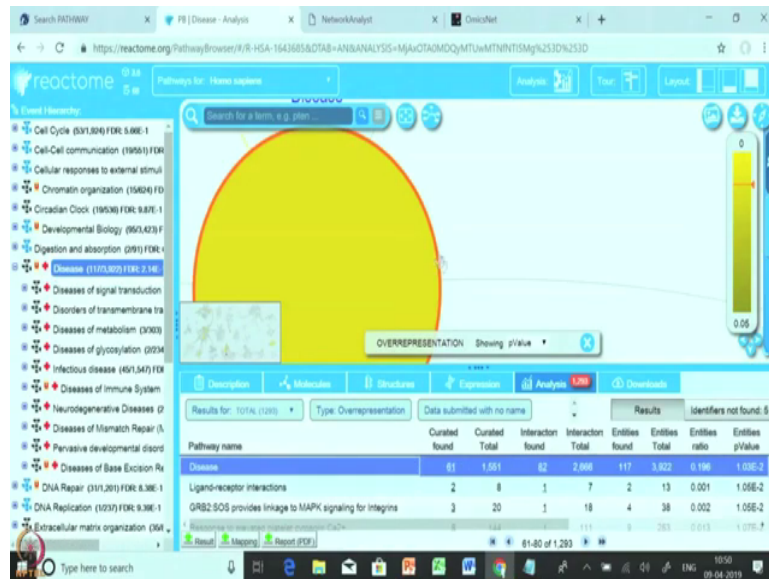
And here they have also given what are the identifiers that are not found that may be due to the upgradation of the databases. So, now, if we want to check the top power pathway which has come in that is related to disease of signal transduction.

(Refer Slide Time: 11:50)

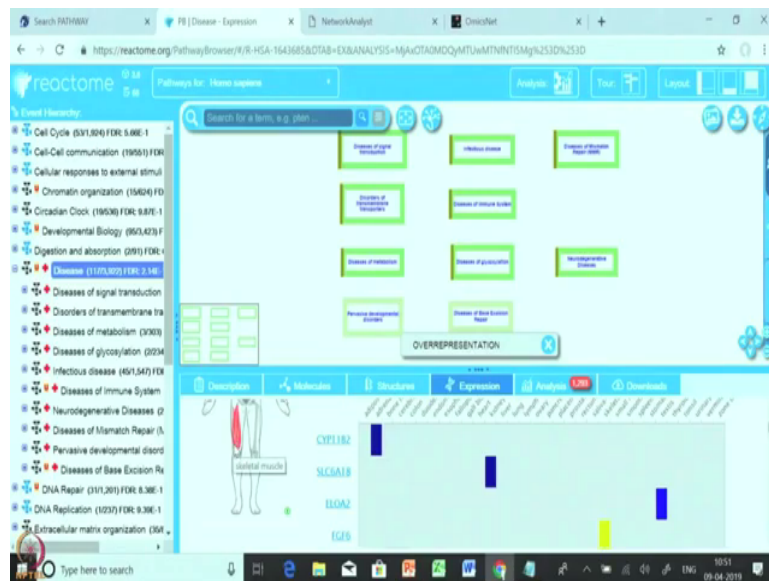


So, when we will be clicking this that pathway we can find, the reactome database will show the complete details of the pathway and when we will zoom in the pathway, we will find that this has given a complete glimpse what are the different sub networks that are present.

(Refer Slide Time: 12:02)

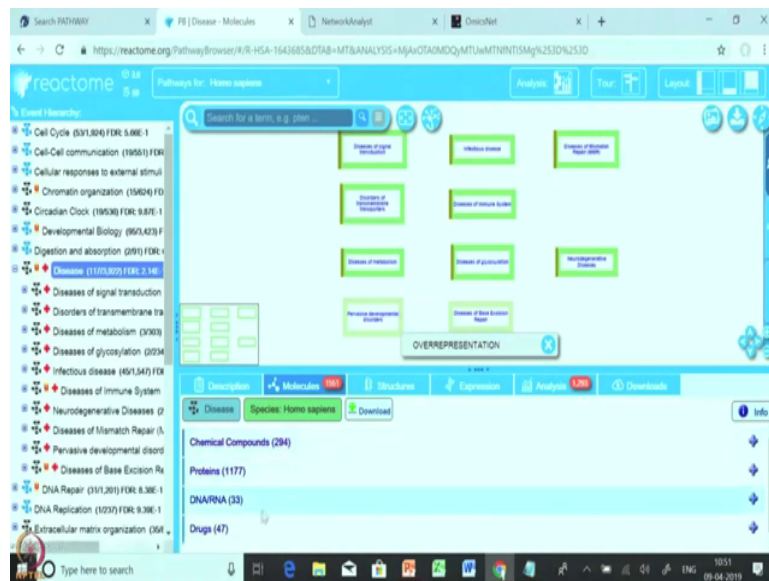


(Refer Slide Time: 12:11)



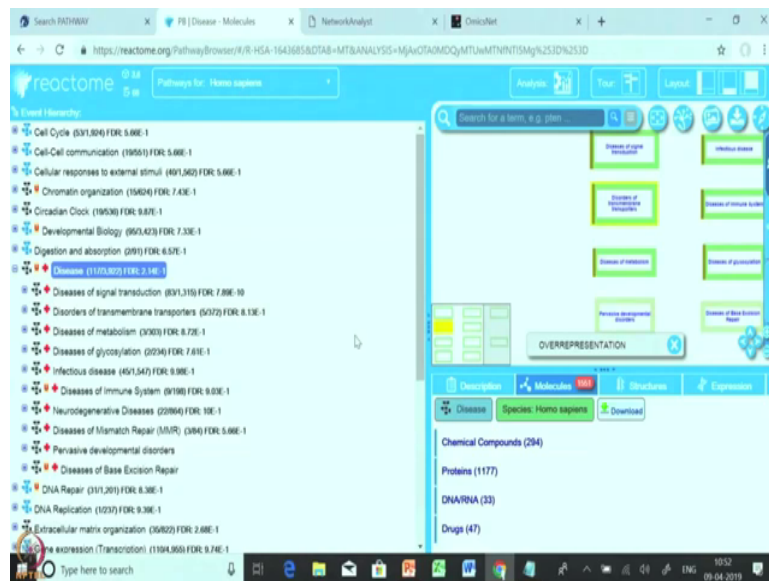
So, apart from this if we come to the expression, the expression is nothing but whatever the different candidates that are present in the pathway and what are their expression throughout in different tissues are present here. So, if we select different kinds of tissues from here and we can find that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

(Refer Slide Time: 12:35)



If we go to the molecule tab over here we will find there are a couple of options are already available; that means, chemical compound proteins DNA, RNA and drugs. This says what are the different kinds 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 kinds of drugs.

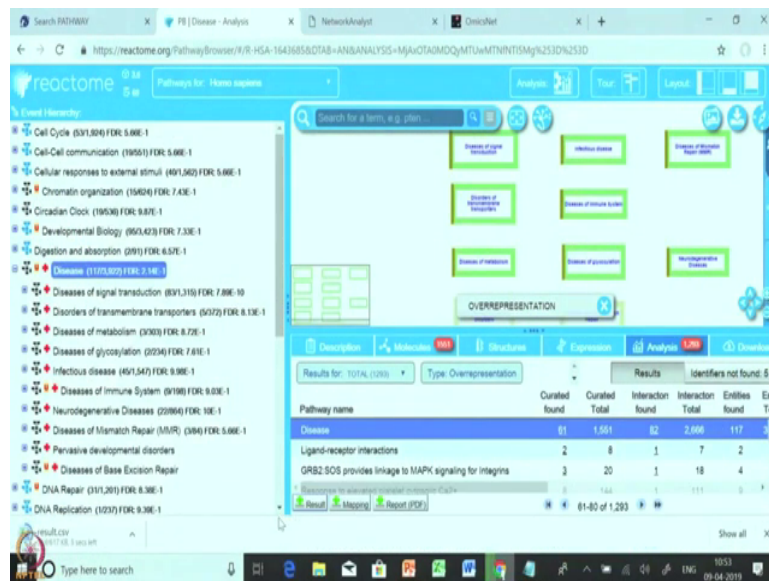
(Refer Slide Time: 13:00)



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

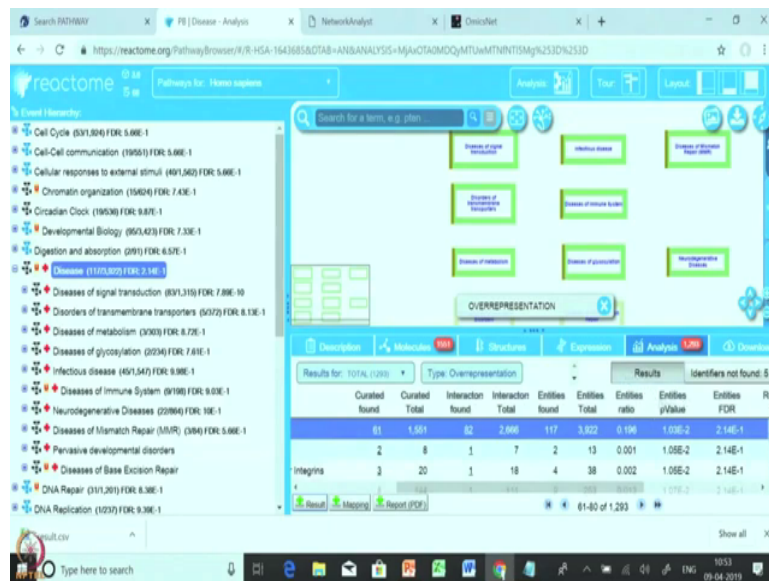
So, likewise reactome gives a lot of information about your data set to large aspect. So, now, the important thing is like downloading the result file.

(Refer Slide Time: 13:39)



So, here is a tab of downloading the result file where clicking this one will download the result in dot CSV format. So, the dot CSV format will have all the data sets and all the complete data file of the analysis. So, from there we have to select certain criteria like pValue or FDR which are already given as you can see.

(Refer Slide Time: 13:58)



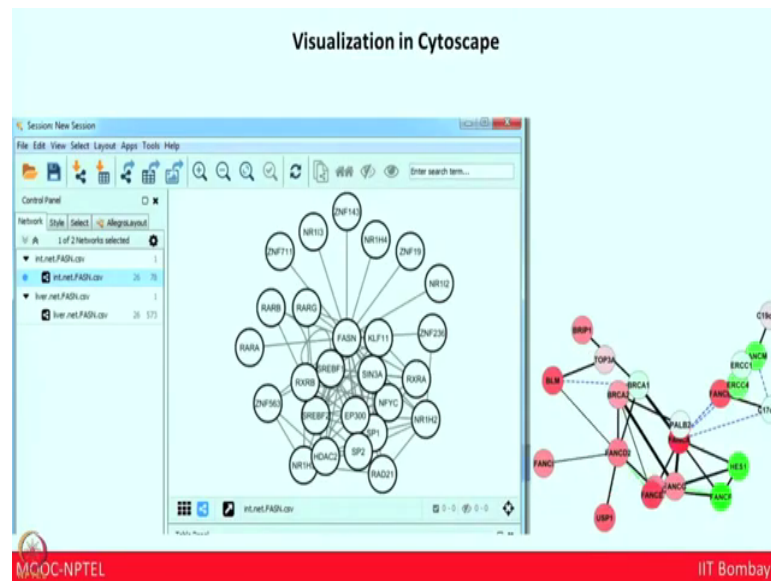
So, on the basis of that we have to select and we have to filter the complete analysis.

(Refer Slide Time: 14:03)

Sorting and filtering data on the basis of pValue and FDR											
Pathway identifier	Pathway name	titles pVal	Entities FDR	Species name	Hits	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	Q14757	Q96G04		
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	Q96G04	Q14757	
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	Q9NQ7	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	Q14757				
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	Q14757				
R-HSA-109581	Apoptosis	0.002822	0.019753656	Homo sapiens	P55211	P10415	Q07812	Q35222			
R-HSA-1500620	Meiosis	0.00313	0.019887109	Homo sapiens	Q13315	Q13535	P51587				

So, after sorting and filtering a data on the basis of pValue 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 pValue, 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.

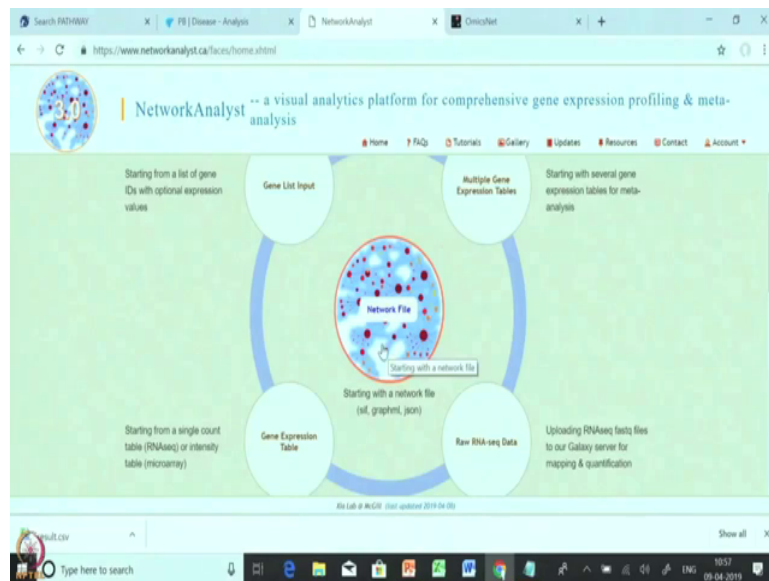
(Refer Slide Time: 14:48)



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

So, as cytoscape can is having different kind of plugins we can generate different kind of visualization network.

(Refer Slide Time: 15:01)

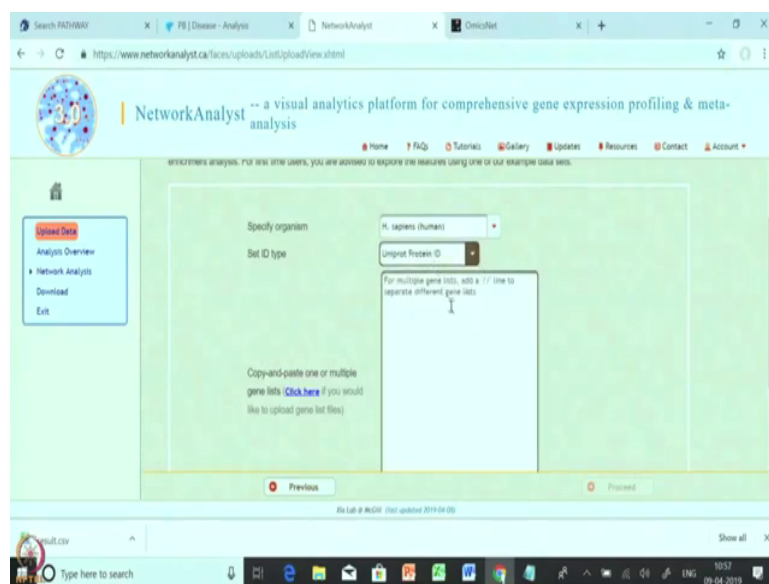


But apart from this today I will I want to show you a very robust visualization of visual analytics platform for comprehensive gene expression profiling and meta analysis that is network analyst. So, network analysts will give you different kind of visualization platform where, you can do single analysis 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, but 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 gene with different expressions can be checked.

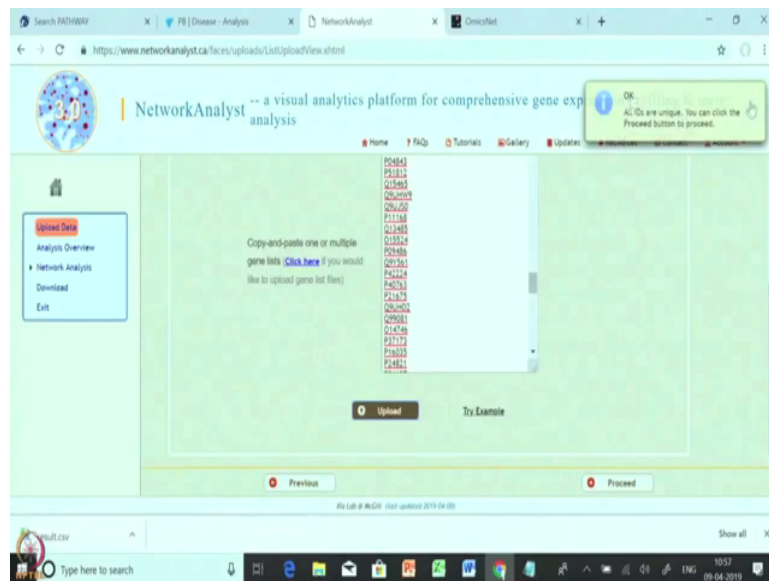
This is a gene expression table where micro RNA micro array and RNA sequence data can be done and this is the raw RNA sequence data from where we can take the sequencing data and we can start with. This is the network file where analyzed file of dot c or JSN can be directly incorporated and we can get the visualization.

(Refer Slide Time: 16:20)



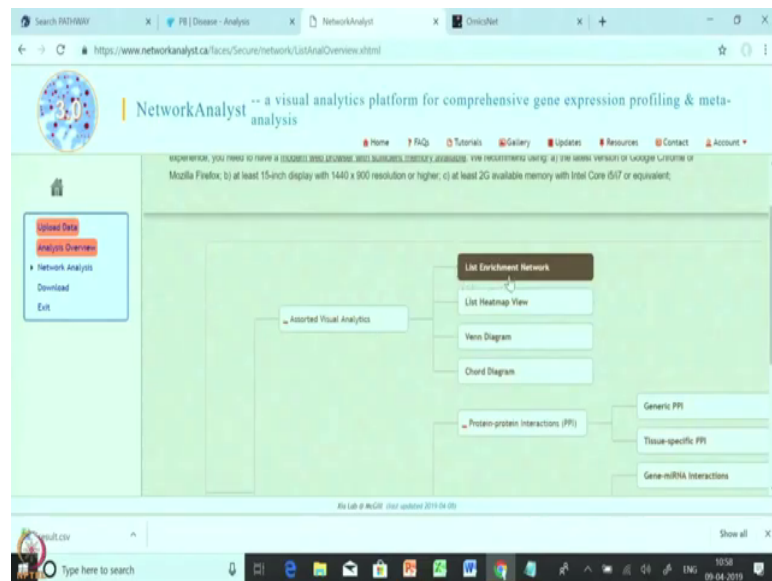
After clicking the gene list input we will be having this homepage where we have to select the organism that is homosapiens, we have to select the ID as we are taking the IDs from the same paste 1 file. So, we know that this is a UniProt ID.

(Refer Slide Time: 16:39)



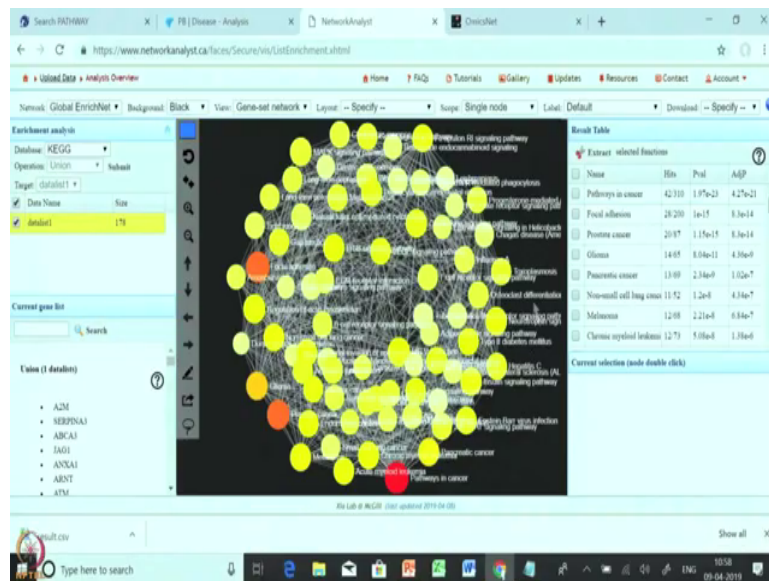
Then we have to copy paste the ID names after this we have to select that upload here and if we are getting any kind of duplicates or errors.

(Refer Slide Time: 16:56)



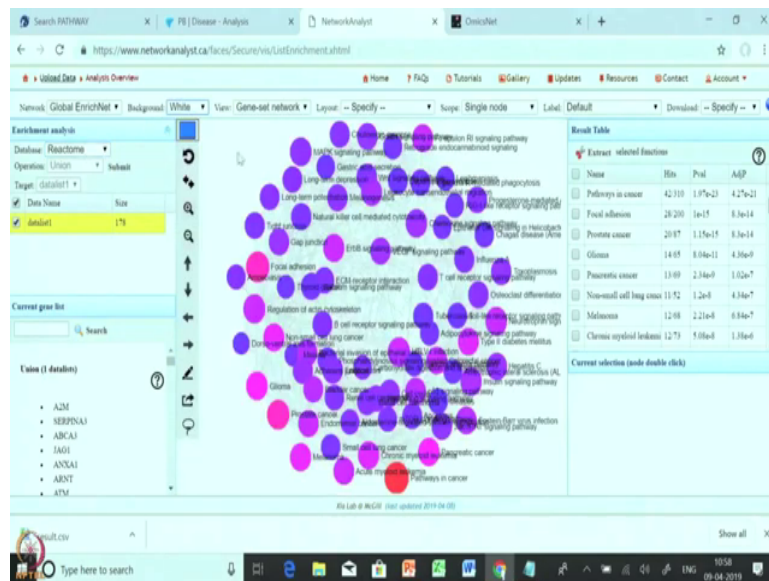
So, it will be shown over here, if everything is fine we have to select the proceed option.

(Refer Slide Time: 17:05)



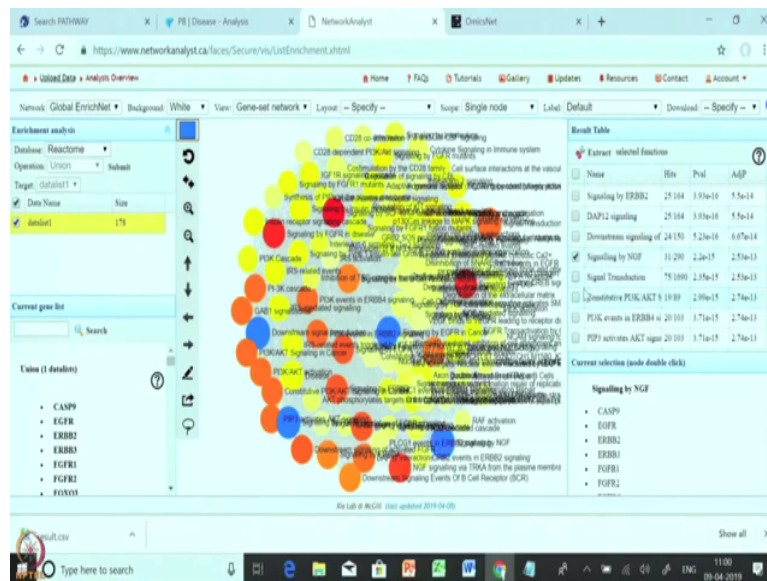
So, first we will check the list enrichment network that is a pathway enrichment network visualization platform, where after clicking this you can see there is a complete network interaction module generated from different kinds of pathway.

(Refer Slide Time: 17:21)



So, in the left we can select the reaction database, after this we have to change the background color to white and submit the database.

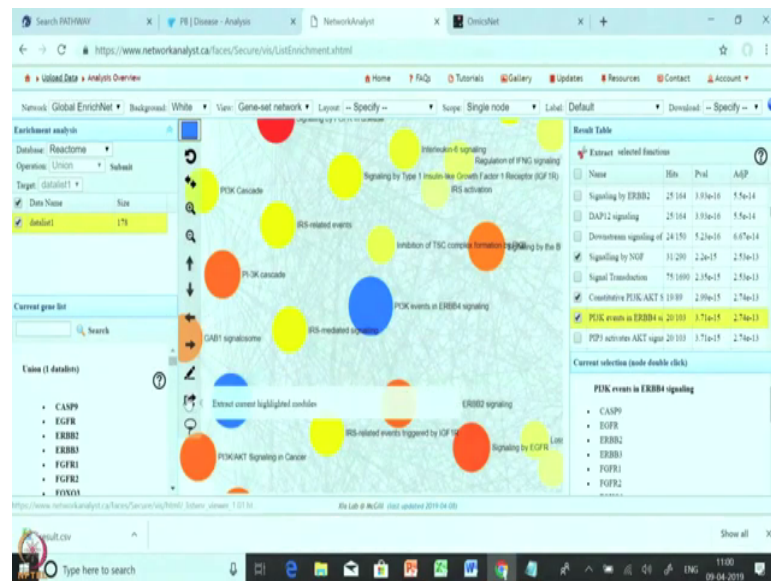
(Refer Slide Time: 17:27)



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.

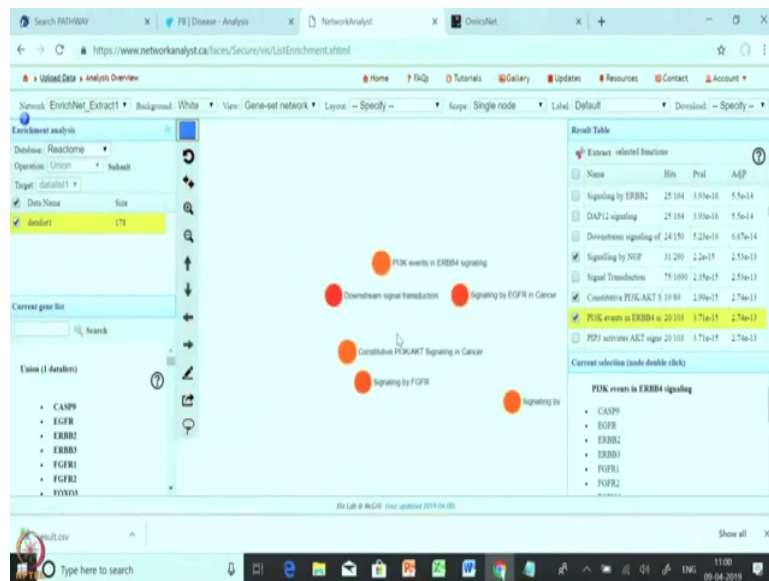
So, now we will go to our table sorted table that we have generated from the reactome and we have all the pathways that are present and we have taken on the basis of significant pValue. And we will select those pathway which is already present there like this one downloading signaling matrix, signaling of EGFR, signaling of FGR FGFR, signaling of NGF and like this way we have to select some of the pathways from here.

(Refer Slide Time: 18:08)



And we have to come to and we have to extract these pathways.

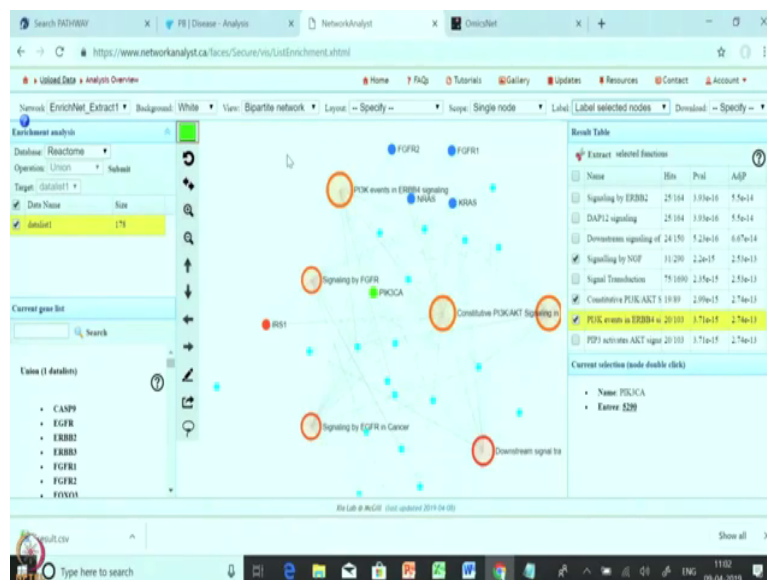
(Refer Slide Time: 18:18)



So, after extracting these pathways we can see there are only few pathways which we have extracted.

To 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.

(Refer Slide Time: 18:43)



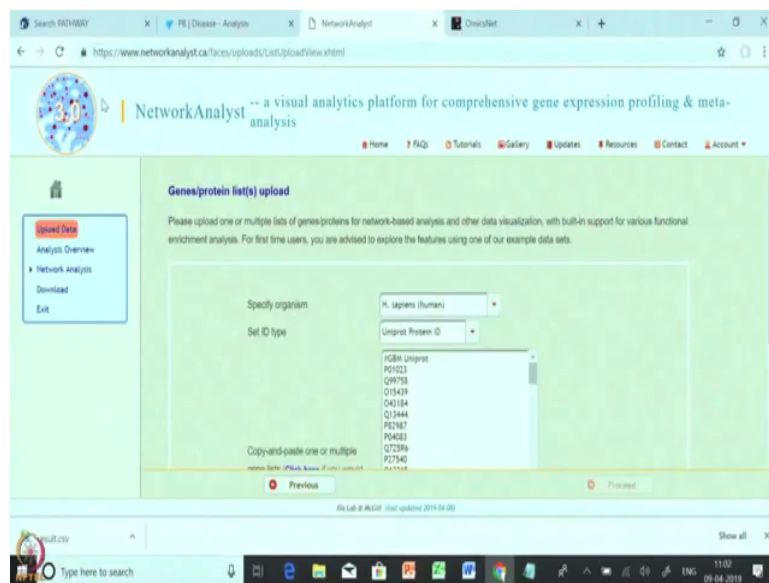
And from here we will be selecting the bipartite network that will not only give you the name of the pathway, but also it will give you the name of the proteins. So, as you can see the whatever pathways we have selected are already present there and apart from this whatever the proteins that you have submitted in your data set is already available now. Now, if we select each pathway each proteins from like this and there is a option of label and label the selected nodes. So, already these pathways these proteins are already labelled.

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 upregulated, so I can put in red color whereas, this protein is down regulated. So, I can green color and here again I have to select the label selected protein and it will show that IRS1 is upregulated one whereas, the PNCIO

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.

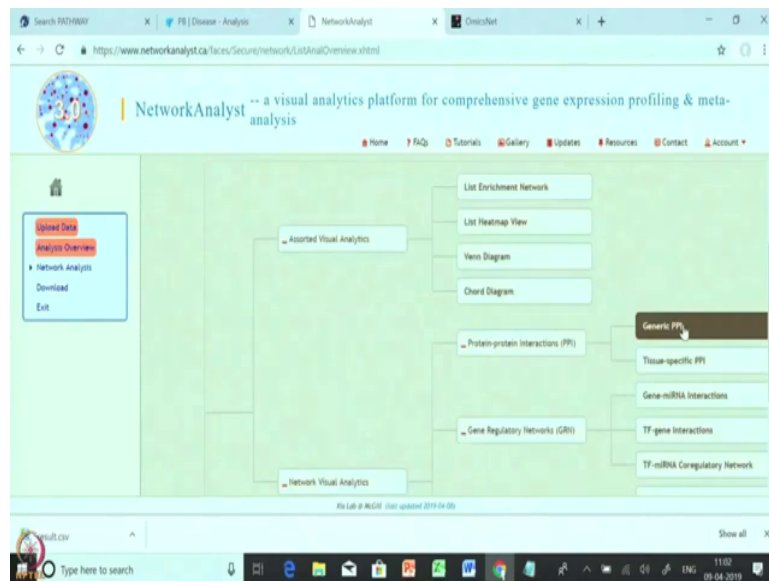
So, now, we know how to generate a very good pathway enrichment model. So, the same way we can go for the protein protein interaction model.

(Refer Slide Time: 20:19)



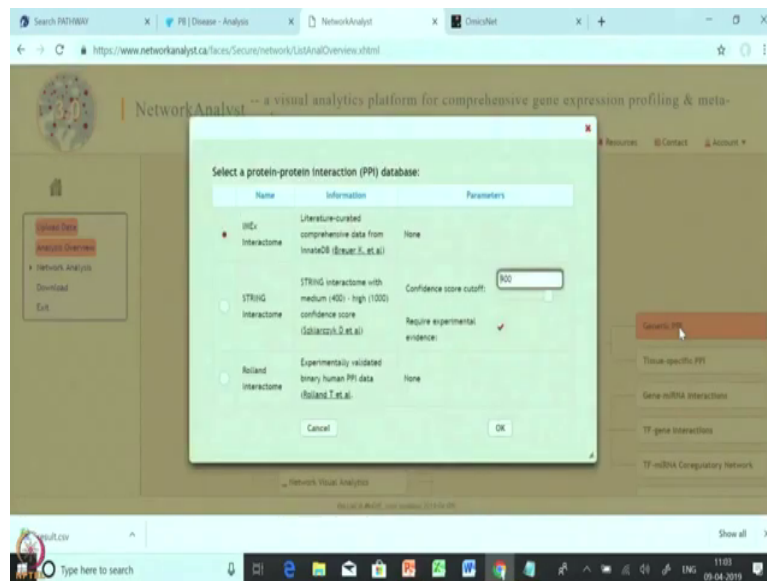
So, to get the protein protein interaction model we have already uploaded our data set in network analyst, we will be choosing the generic PPI.

(Refer Slide Time: 20:25)



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 these kind of tissues are already available in the database and they can check, but as I just want to give you the glimpse.

(Refer Slide Time: 20:47)



So, I will be choosing the generic PPI where three names of the databases are already there. So, these 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: 21:23)

The screenshot displays the NetworkAnalyst web application interface. The browser's address bar shows the URL: <https://www.networkanalyst.ca/ncsu/Secure/network/NetworkBuilder.html>. The page header includes the NetworkAnalyst logo and the tagline "a visual analytics platform for comprehensive gene expression profiling & meta-analysis". A navigation menu at the top contains links for Home, FAQs, Tutorials, Gallery, Updates, Resources, Contact, and Account.

The main content area is titled "Mapping Overview" and contains a descriptive paragraph: "The significant genes (seeds) from previous analysis are mapped to the corresponding molecular interaction database. The procedure typically produces one big subnetwork ('continent') with several smaller ones ('islands'). Subnetworks with at least 1000 nodes are listed below. You can visually explore them in the next step. These subnetworks can be downloaded as SIF (simple interaction format) files to be explored in other tools (i.e. Cytoscape)."

Below the text is a table with the following data:

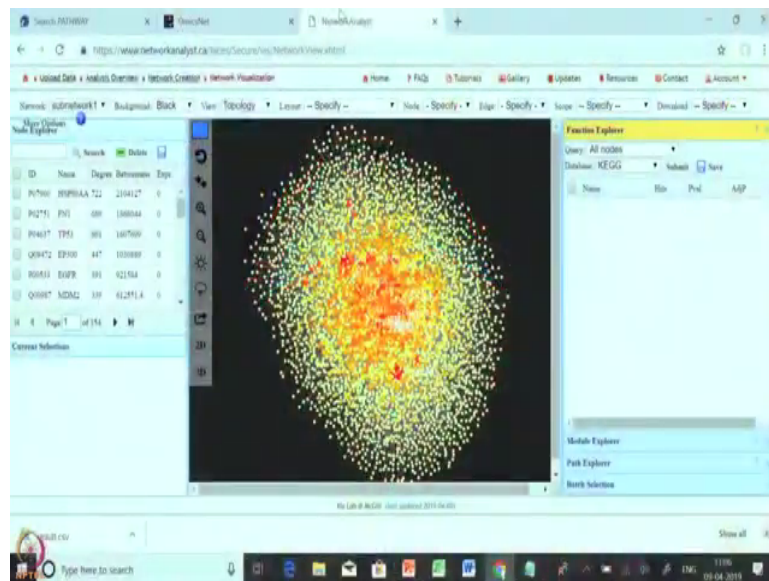
Networks	Nodes	Edges	Seeds	Interactions (.SIF)
subnetwork1	4677	11427	176	Download

On the left side, there is a sidebar with a "Network Analysis" menu containing options: Upload Data, Analysis Overview, Network Creator, Network Visualization, Download, and Exit. On the right side, there is a "Network Tools" panel with buttons for: Reset Network, Zero-order Network, Second-order Network, Minimum Network, Degree Filter, Betweenness Filter, and Steiner Forest Network. At the bottom of the main content area, there are "Previous" and "Proceed" buttons.

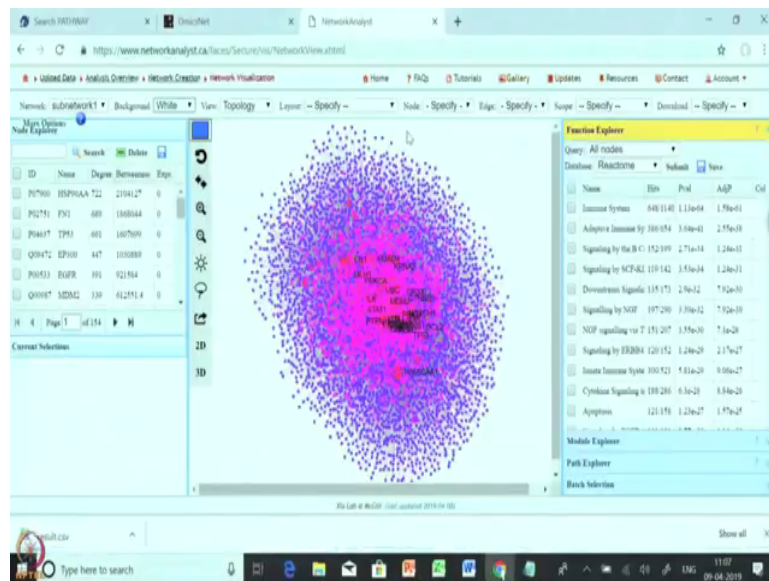
So, if I am choosing the IMEx interaction and we can see like there is a one sub sub network with 4677 nodes, 114 11427 edges, 176 seats. So, from here you can download the dot c file of the interactions and we can upload it again into the network for future use.

So, now, we will proceed and we will found the data the software has generated a complete protein protein interaction 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: 22:06)

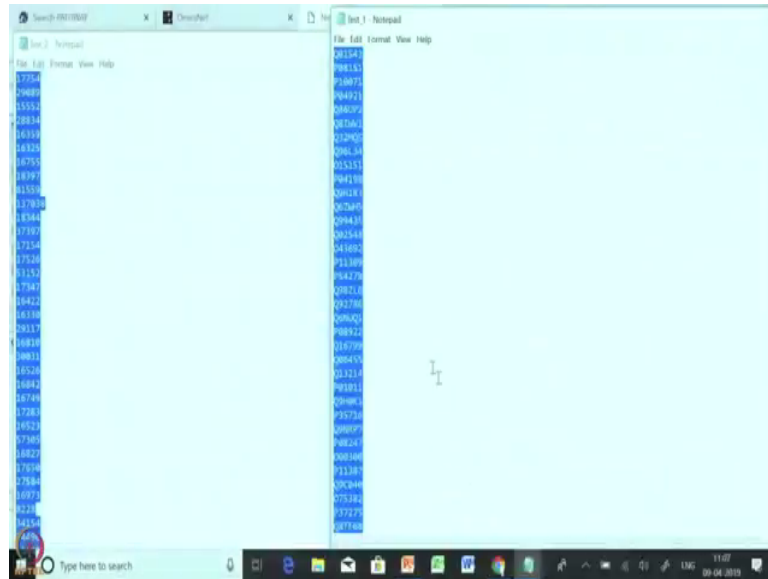


(Refer Slide Time: 22:24)



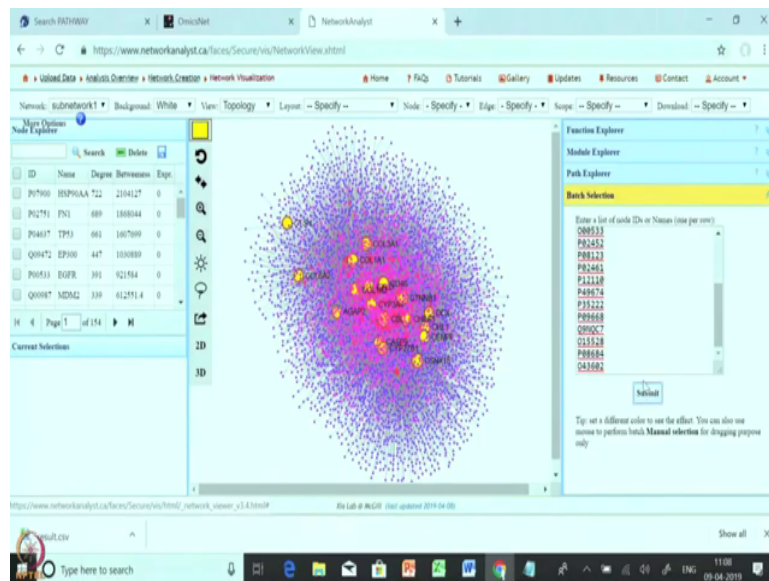
So, as you can see the software has generated the complete protein protein interaction module this is really very complex. So, first we have to select what is the database we want to choose. So, let us go with reactome database and after that we will be changing the color to white. Now, I have already given you in the test 1 file.

(Refer Slide Time: 22:30)



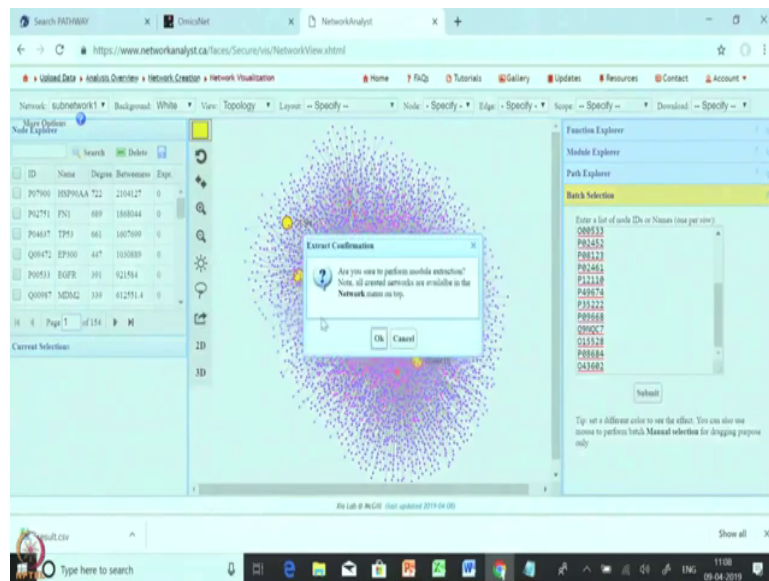
So, now there is a option of batch selection which says that whatever what are the proteins that we are interested in, we can copy paste those protein at (Refer Time: 22:47) ID.

(Refer Slide Time: 22:39)



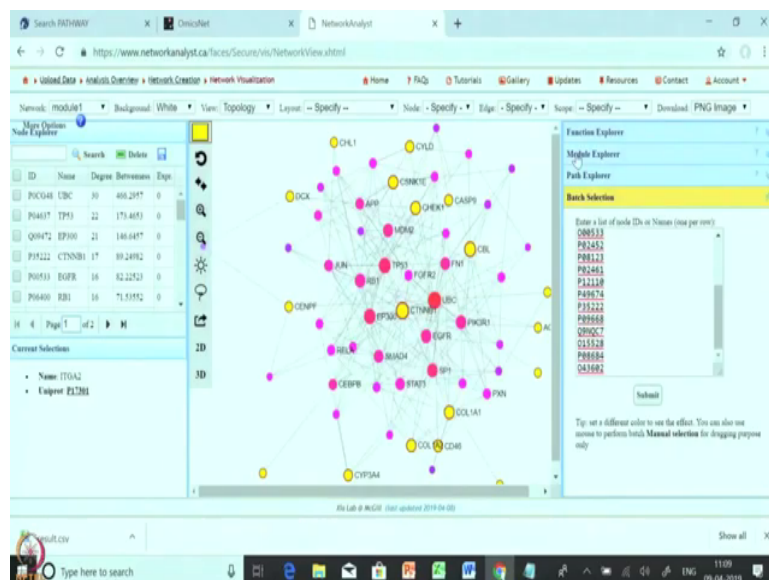
And after that we have to click the submit. So, after submission we can see there is a highlighted candidates that we can found in this complex network.

(Refer Slide Time: 23:06)



So, now as we are very much interested with this candidates, we will select and extract those candidate from this complex network.

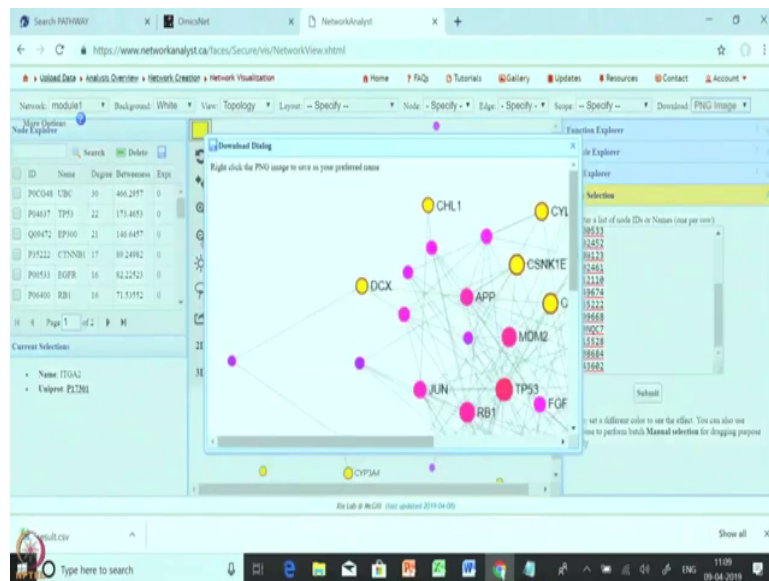
(Refer Slide Time: 23:09)



So, after selection and extracting the candidates we found that these are the proteins that we have selected and is present in this network. Apart from these this proteins which are there are the top most interactors that is coming in this protein protein interaction modules. So, like these we have to make some adjustment to make this network visually interpretable.

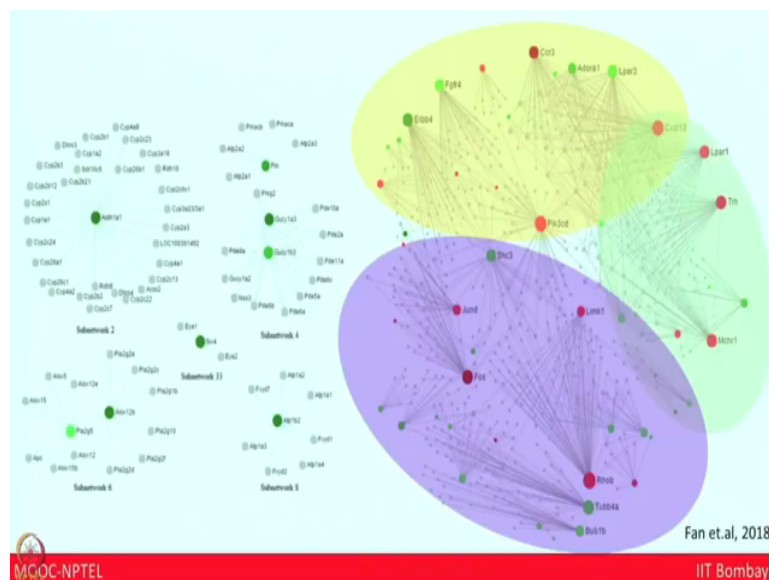
So, that can be also done from their given layout which a different kind of layouts are already available, but for reducing the overlapping I will choose this one to reduce the overlap and as you can see the complex network has got some clarity. So, after adjusting little manually we can download this one with as a PNG Image and we can save the file as a PNG or JPG.

(Refer Slide Time: 24:00)



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 a Fan et al in 2018.

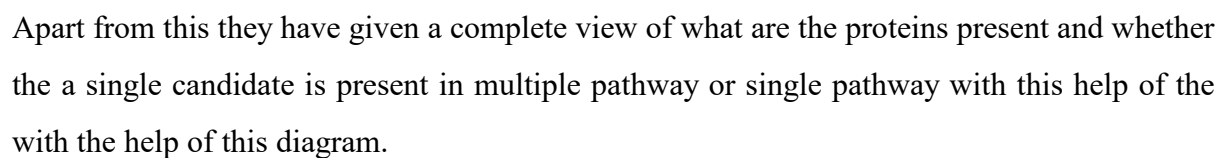
(Refer Slide Time: 24:31)



So, they have given a very good they have used this network analysis software to a 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 differentiated the protein candidate list is upregulated and downregulated manner. And they have also generated different sub networks like these four sub networks they have generated and on the basis of different pathway or protein protein interaction.

So, now, if I want to check like what are the different clusters that is coming in my protein protein interaction module in terms of pathway that can also be done checking the curated sorted reactive list that we have generated.



(Refer Slide Time: 25:42)

Points to Ponder

- Reactome and KEGG are two important and widely used pathway databases.
- Network Analyst and Omics.net can be used for Pathway enrichment and Protein-Protein Interaction.
- Multiple tools and databases can be use to get the complete biological information of a dataset.



MOOC-NPTEL

IIT Bombay

As you have got some sense today that all the experiments that are defined are based on certain hypothesis and therefore, the data analysis and interpretation becomes really crucial. What is very important that you have to be very unbiased when you are starting on a big data field you have to be very unbiased you have to start from the big data big table and look at what is most significant changes happening.

Looking at the statistical value, looking at the p values and then thinking about you know various threshold which you have set it up to the very high stringency filters which you applied to obtain a much shorter list which is the most confident candidates, the genes or proteins which you would like to take forward.

Now, based on these then you would like to make some actionable hypothesis and then you would like to do a follow up experiment to test out is my hypothesis working. If not then you

will look for you know other proteins or other set of you know genes on the same list and look at is there any alternate hypothesis which might be more effective right.

So, you should start with very unbiased way looking at the data at the same time you should also be on top of literature and what is has already been published in the (Refer Time: 27:10) you know if you go to the various publication of that question what people have already published.

So, you would like to also make some sort of a strong foundation from the publication already available and look at the data independently then start trying to map the things together are there certain parts of the published report is also getting mapped in your you know unknown data set.

And if that is the case then of course, you are more confident that people have reported these pathways in these you know set of the proteins which are very significant in their publications. And we also see that among our top 100 proteins 50 of them are coming, but then there are 50 more which are totally new and unknown proteins and what those proteins are.

Now, then you know your more curiosity will refer there how to tell these, out how to ensure that this protein what we are finding our top list they are the real proteins right. So, the pathway analysis and the network analysis theory tries to give you the much more comprehensive picture which is very close to the biological question which you wanted to add.

So, its good idea for you to get familiar with these software tools, with these bioinformatics aspect of analysis. So, that you know you can start looking at your data in a very very different way which is otherwise not possible just by looking at your excel sheet and just looking at proteins and gene lists in the isolation.

We will continue more of these discussion in the next lecture, till then.

Thank you.