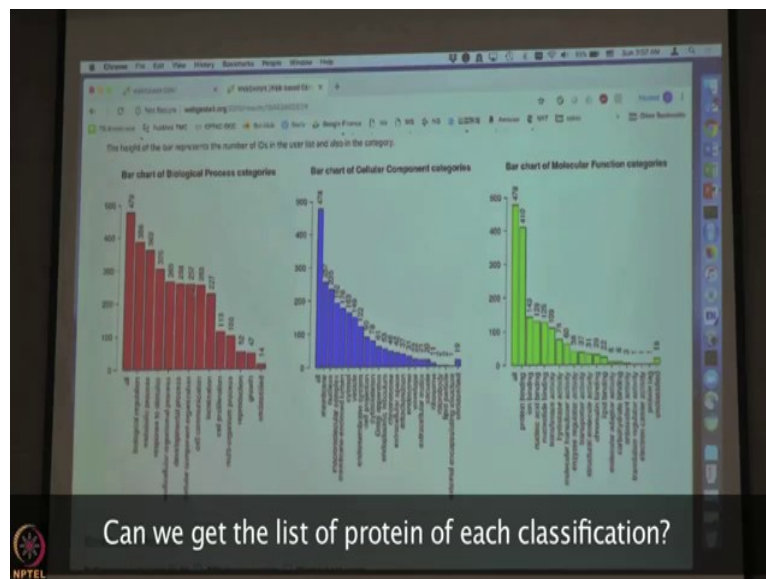


Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. Bing Zhang
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay
Baylor College of Medicine

Lecture - 50
WebGestalt II

Welcome to MOOC course on Introduction to Proteogenomics. Welcome to the hands on session of WebGestalt. In today's session Dr. Bing Zhang will teach you about how the results and job summary can give you useful information. He will also show you how the enrichment analysis can be visualized in different forms. He will discuss about different types of network-based methods like direct neighbor based approach, module based approach or diffusion based approach. So, let us welcome Dr. Bing Zhang for today's hands on session on WebGestalt.

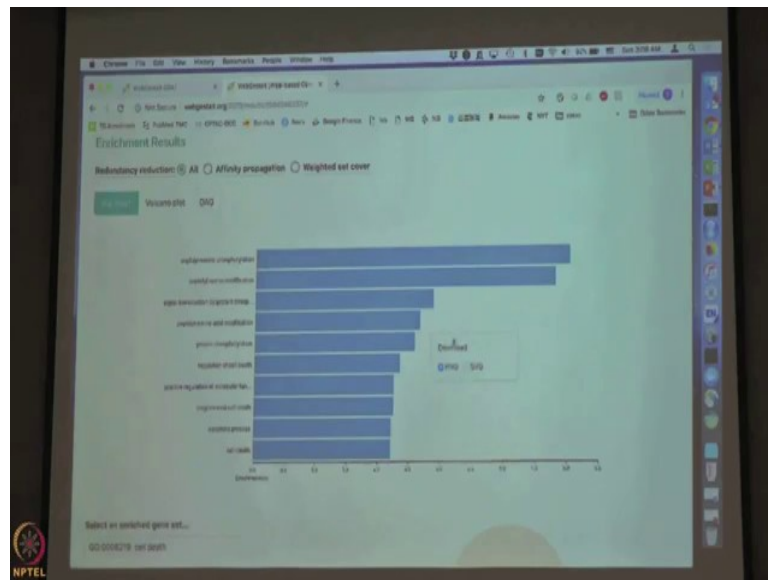
(Refer Slide Time: 01:07)



Course name summary. So, this is not first, this is not enrichment analysis. Do not use this to report as your enrichment analysis. This is just the simply a classification of the genes you submitted based on some pre-selected the biological process, cellular component and the molecular function categories, give you idea of how many if example, how many genes are related to biological regulation metabolic process or response to stimuli. So, these are the high level categories, yeah.

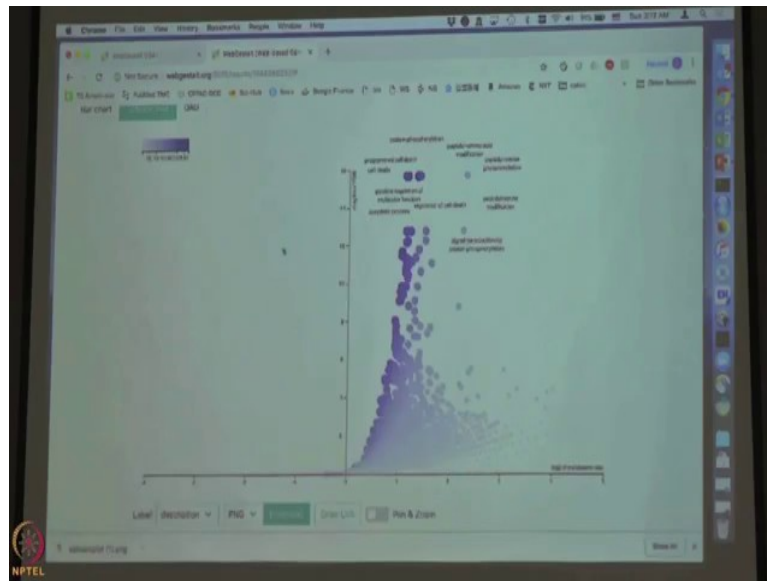
Well, again that is I think you have very good representative of the users we got to this request from a lot of users. So, yeah that is to do list at yeah, we would do that in the 2019 release. And it is just basically similar to the pie chart you typically say I mean to classify genes and then the important part is this enrichment results.

(Refer Slide Time: 02:15)



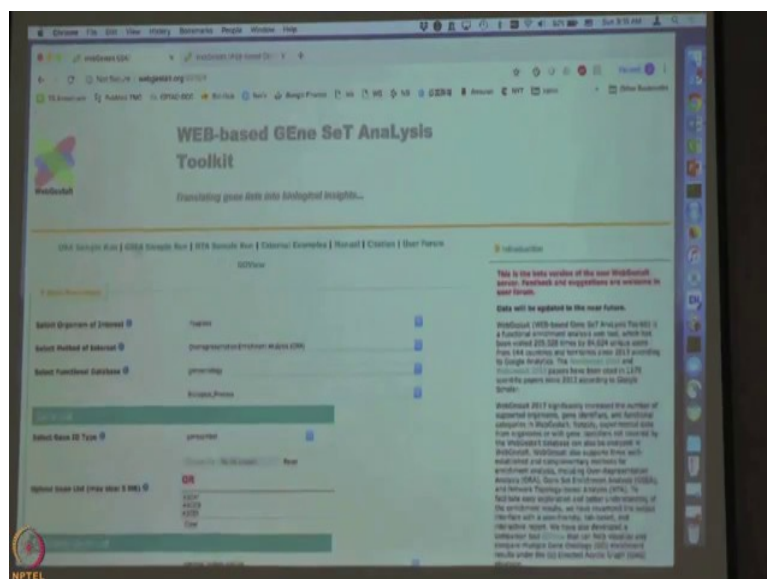
And next do not do the redundancy reduction yet and then the default view is a bar chart, so basically this shows you the enriched the categories. Because we choose top 10 options that is why we get ten categories here, right and on the y axis you have the enrichment ratio from the fishers analysis. And then you can this one can be downloaded if you right click and then you can download this as PNG or SVG for your presentation or for and even for publication I think the quality is good enough.

(Refer Slide Time: 03:02)



Or you want to maybe visualize in another way you can visualize all your results you know volcano plot this and then you see this as the GO terms highlighted it is the top 10 categories. You also have the option to change it, you change can change the label from the gene certain name that is the gene ontology which you do not understand what it is to something is more descriptive, make the description of the gene ontology terms. If you do not have this volcano plot then you are still using the old version probably.

(Refer Slide Time: 03:38)



So, you have to go to the website and if you the old website is like this and you have to use the WebGestalt 2019 beta version or you just put 2019 in your URL. Yeah, next this is a completely new feature. So, and the programmer also did a very nice job as you can see because these labels are crowded and sometimes you will not be able to see, right, but the good thing is you can actually use your mouse to move these around, to move this to the right place that you want it to be. And the of course now, it is difficult to see where it is; where you can show the link and then best way you can rearrange them in the way that it can be used for publication.

Student: How are we going to draw the link?

Click on the join link

Yeah and the.

Student: What is there in the x axis? What is there on the y axis?

Yeah, the axis is a log two of enrichment ratio. So, best thing this is you have expected the number of genes, right or if you do random sampling and then if you have an enrichment. What is the enrichment ratio this is this and then the y axis is FDR in the log scale. But after you move this around you can see now you best may have a very nice view of every cell. You can keep the link or you can remove the link to get a clean view and then you can download these plots that is very.

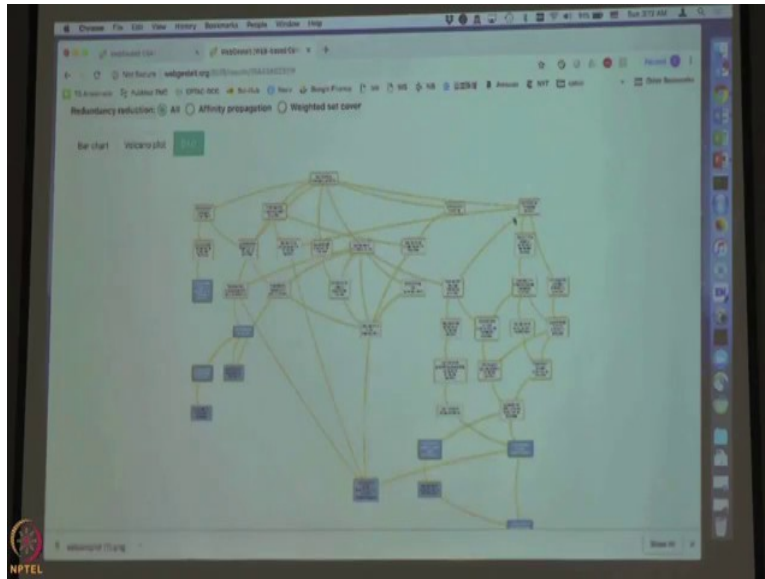
Student: Sir, if on the y axis it is FDR so...

Yeah minus log FDR.

Student: Minus log FDR.

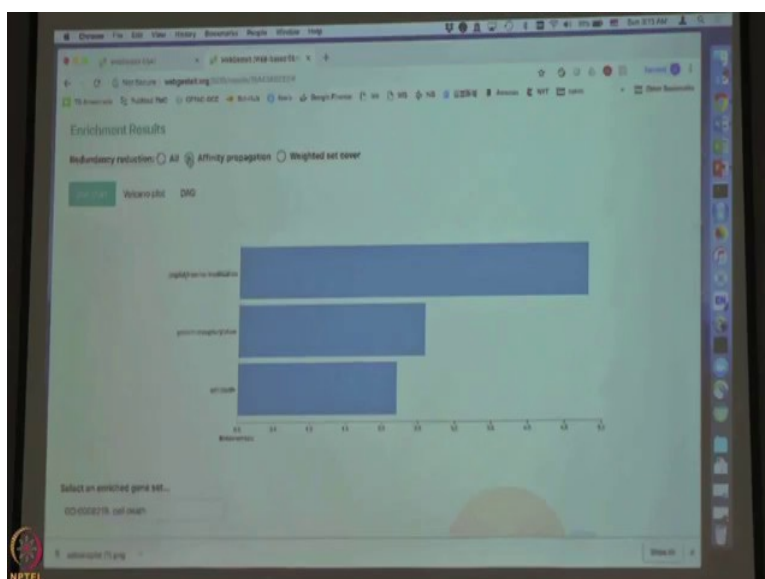
Yeah, yeah. So, the smaller end up with the higher, yeah.

(Refer Slide Time: 05:50)



So, this is a volcano view and then you can also do the data view because these genes are organizing directed the technique graph in gene ontology, you can actually see this as well, but this is a verbal in the old version as well. And from idea of this and sometimes if you end up with too many categories, some of these are quite redundant as you can see in go, right, the; it has this hierarchical structure I mean the parent term and the child term. So, you want to simplify this and the for example, you can do the affinity propagation based a simplification. Let us look at this; if you. Let us do a pie-chart you originally have 10 significant terms.

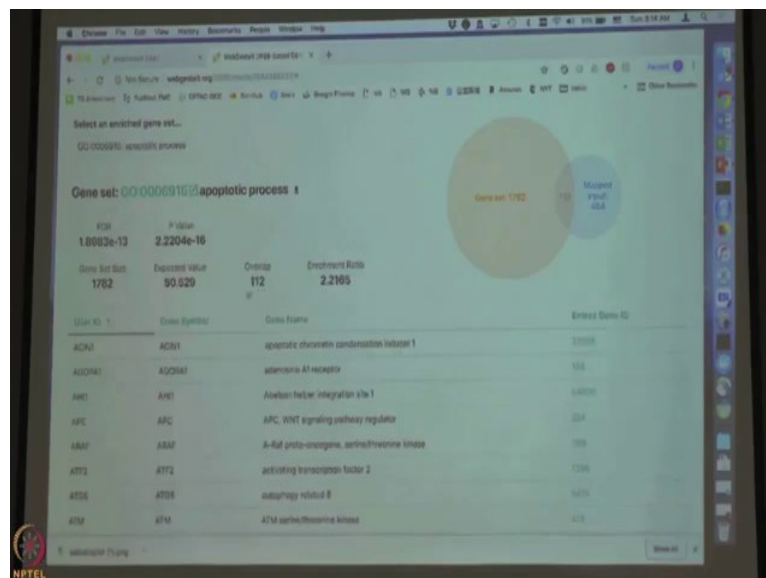
(Refer Slide Time: 06:41)



But if you do the affinity propagation you end up with 3; that means, this will groups 10 terms into basically 3 clusters and only pick one from each as a representative that will simplify your interpretation. And the another algorithm the implemented is the base set which is safe cover this end up with 4. Either of this is useful for you to simplify your photo I mean it is not a problem at all, but you can imagine when you have 200 and this would be a very useful feature to look at.

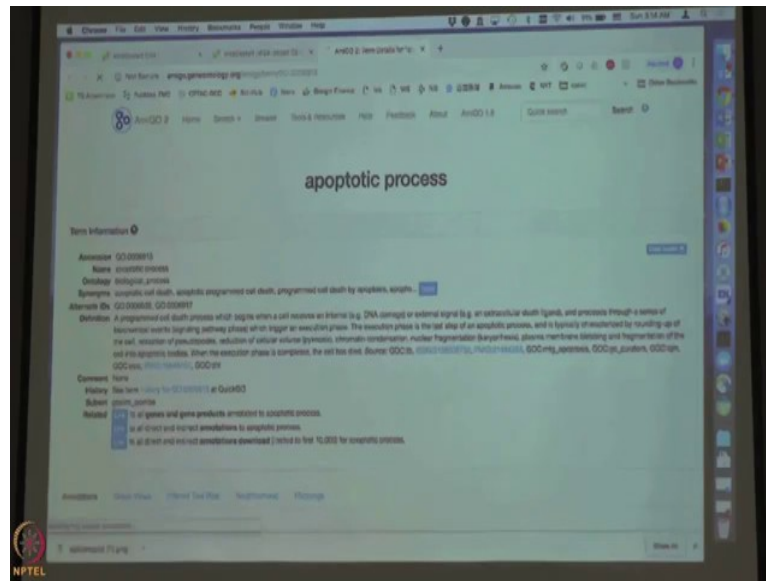
And, now we have the overview of the result and you can download and save, but of course, you want to understand why this is enriched, right. You want to look at the detailed result you can click on any of this bar that which should show you the detailed results for that pi at the bottom part of this. For example, this use a apoptotic process.

(Refer Slide Time: 07:30)



And then you can click on this, it is linked to the database amigo.

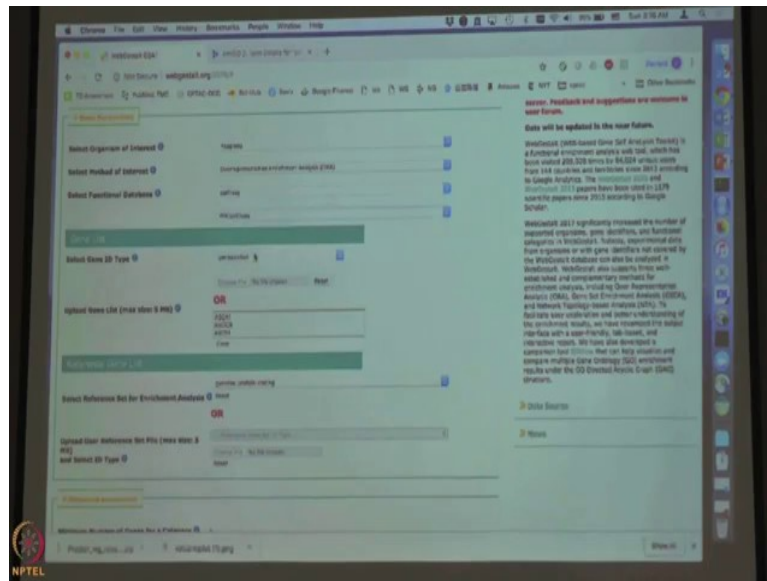
(Refer Slide Time: 07:46)



So, this will give you a description of what is a apoptotic process and then here you have the FDR result, the P value before the adjustment and then you have the gene set size. You have the expected value and the overlap; overlap the number of genes and enrichment ratio. So, basically the enrichment ratio is overlapped divided by the expected and then you have also very easily understand about one diagram to help you to understand this result. And then all the genes in overlapping genes here in this table you can cross all the genes and you can also sort them in different ways.

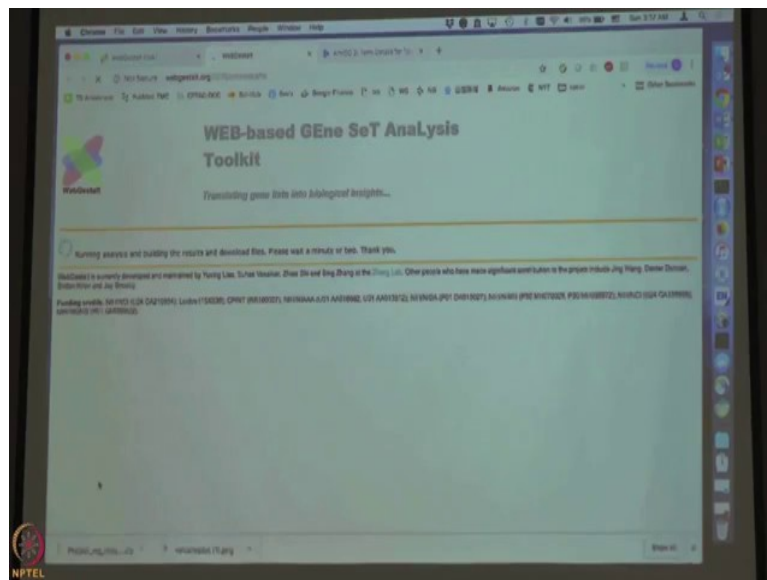
So, yeah and then let us see or you did this analysis and you want to share with your colleagues or if you want to share with your supervisor, he just does not have time to run this analysis. So, and then you can click on the result download. So, this will save the result has a zip file. If you open this zip file it will include the html file and other files and if you click on that html file that will bestly reproduce this exact result. So, it also provides a very easy way to share your results with others. So, yeah, it is very simple like this, but you can explore this in many different ways.

(Refer Slide Time: 09:22)



And you can go back and then change the let us say we did the analysis against the biological process, right. Now, we want to do pathway necessary, we do change the function of database to a pathway type and then we use a wiki pathway for the analysis and maybe we change the cutoff from the top 10 to FDR 0.05, so and then you can submit.

(Refer Slide Time: 09:56)

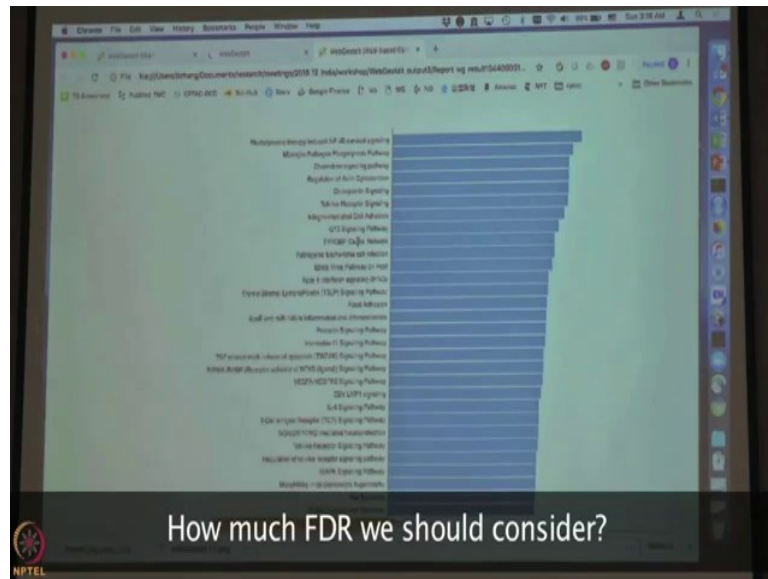


So, basically this will compare your gene list base all the wiki pathways. Let us see what we get. I am using wiki pathway as a demo.

Student: Oh.

Yeah, you can choose Reactome if you want or ok, but the yeah. I am using wiki pathway.

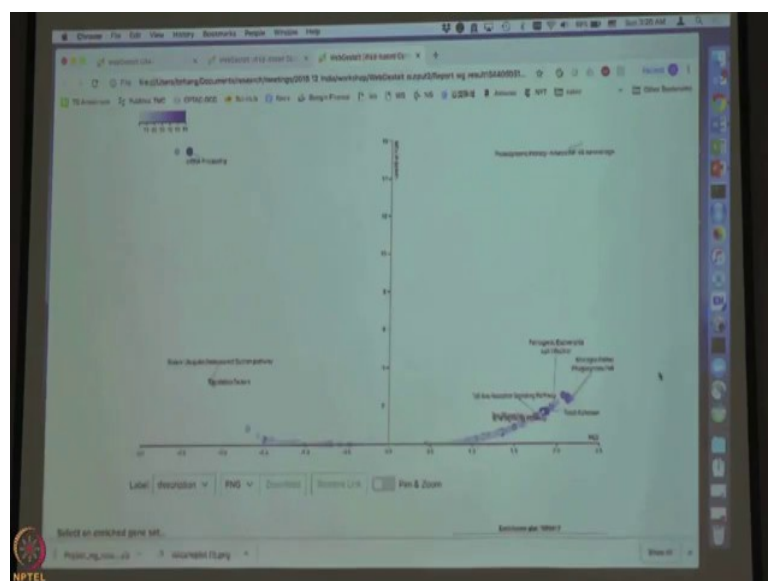
(Refer Slide Time: 10:16)



FDR, yeah, I think you can do 0.05 for example, yeah. It is up.

Yes, yeah. Pathway I mean you can do up to a point. Some people use 25 percent, 0.25 that is a not just I think you cannot go, yeah. But I think you need you go with 0.05 or 0.01. This is a result I got. So, basically it is very similar to the gene ontology analysis, but of course, we do not have the deck because I this is not gene ontology.

(Refer Slide Time: 11:09)

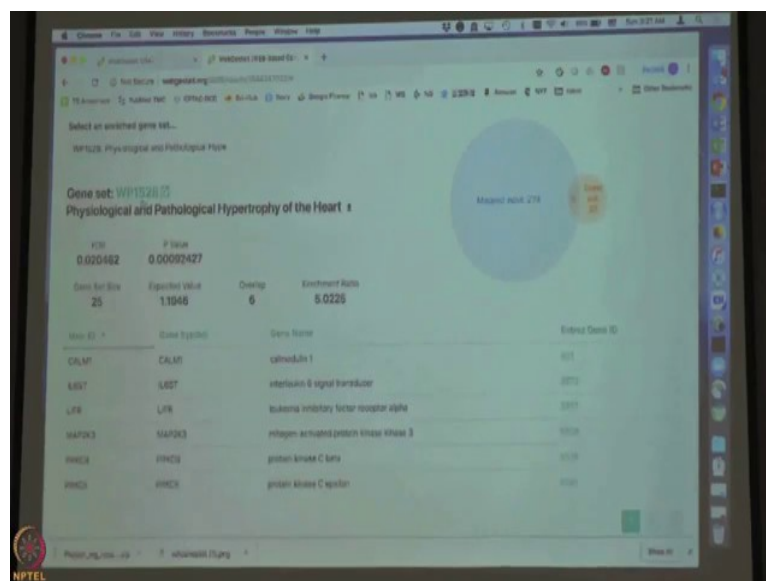


So, there is that directly the statistical graph our structure anymore, but you still have the volcano plot and then you can still put the description here and now you can see because we have a lot of enriched categories. So, it is more useful to apply the finished propagation. For example, you have a lot, right and this difficult to go through and then use affinity propagation of which it will gave you or reduce the sets that can represent, personal highlights the representative ones and yeah that will make your volcano plot easier to look at as well.

And another difference between the this analysis and the gene ontology analysis is for gene ontology and they just put genes together, but without defining the relationship between genes within that gene etcetera, but here and we can have the pathway map.

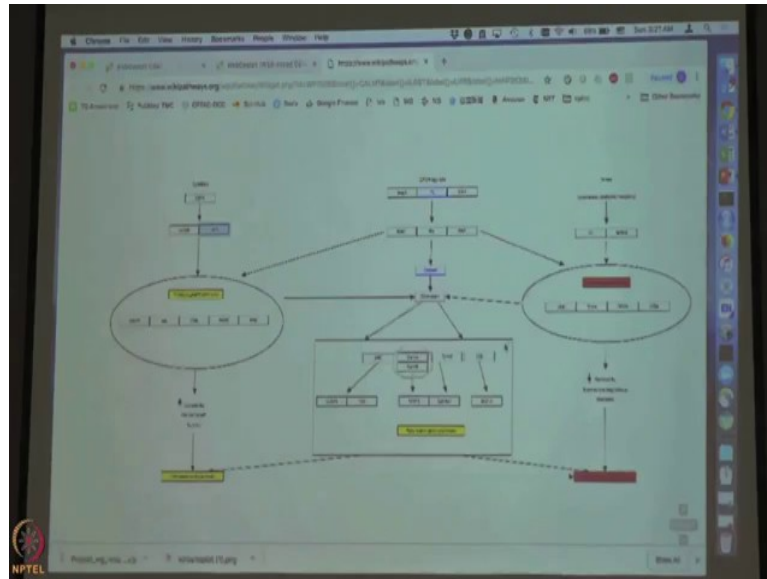
Yeah, this is the result.

(Refer Slide Time: 12:14)



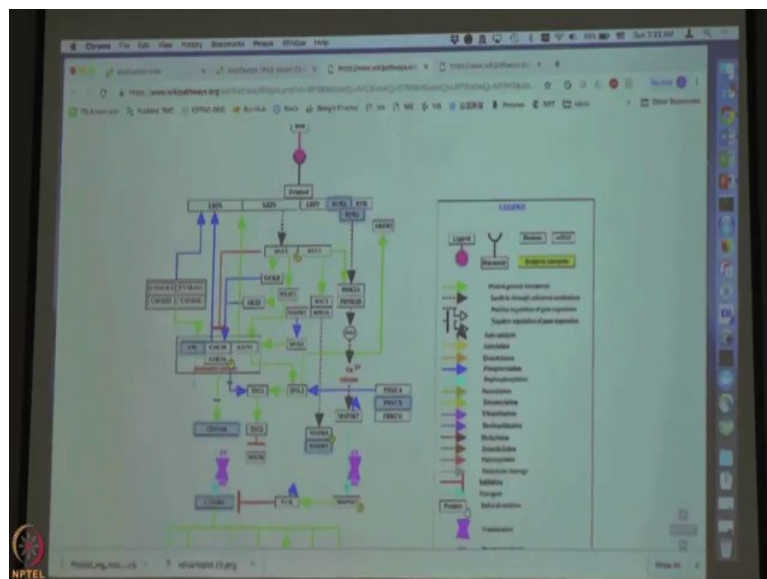
So, but if you click on one of these pathways and then if you click on the pathway ID it will also show you the overlapping genes here, there are 6 genes here, right, and then you have the pathway map and the genes will be highlighted in the map.

(Refer Slide Time: 12:30)



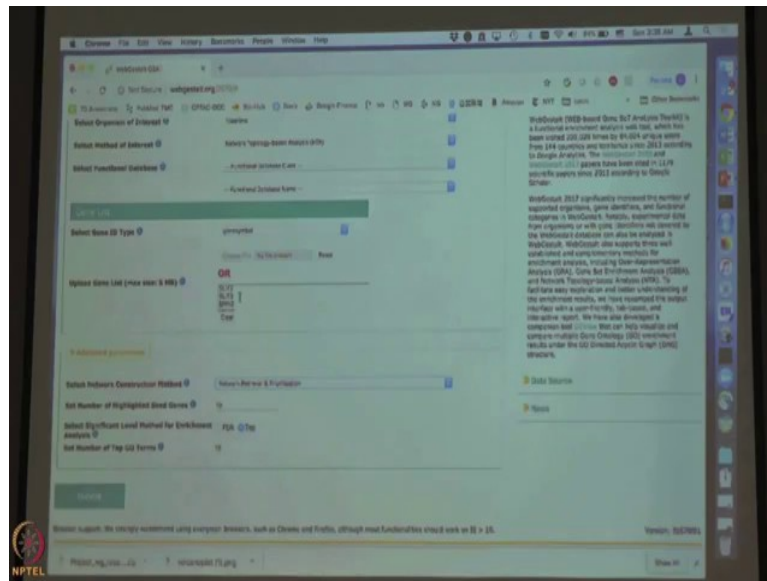
I do not know why only one is highlighted here. Let us try another one, yeah.

(Refer Slide Time: 12:46)



So, the hierarchy depends on like because sometimes one node might have represent a complex or something like that there might be multiple genes in one node. But this provides a very either way that you have not only have the pathway identified, but also you know where your genes are located on the pathway. And this is highlighting function is available in all of the pathway analysis like KEGG , Reactome, wiki pathway.

(Refer Slide Time: 13:24)



So, that is a first example. And I think you can change the parameters in different ways for example, we only look at the gene ontology biological process and wiki pathway and the using a specific cut off. But you are free to change those and the do your own analysis. But in the next example and I want to show you how you can do the gene set enrichment analysis in WebGestalt, and the specifically what we have here is a pre rank based analysis, we do not do the differential analysis in the system, rather you do you choose your statistical methods and you do your differential analysis and after that you get a statistic or P value that will allow you to rank all the genes.

In this case for the ORA as I said, let us say if ORA and the input is very simple it is just a list of genes. So, that is best when you to your differential analysis or correlation analysis and you identify 100-200 genes and then you copy pasted here or you put that in a text file and upload here. But if you want to do GSEA, let us see let us click on this GSEA sample run. And what you need to have is something like this, you have everything has a value associated with that gene.

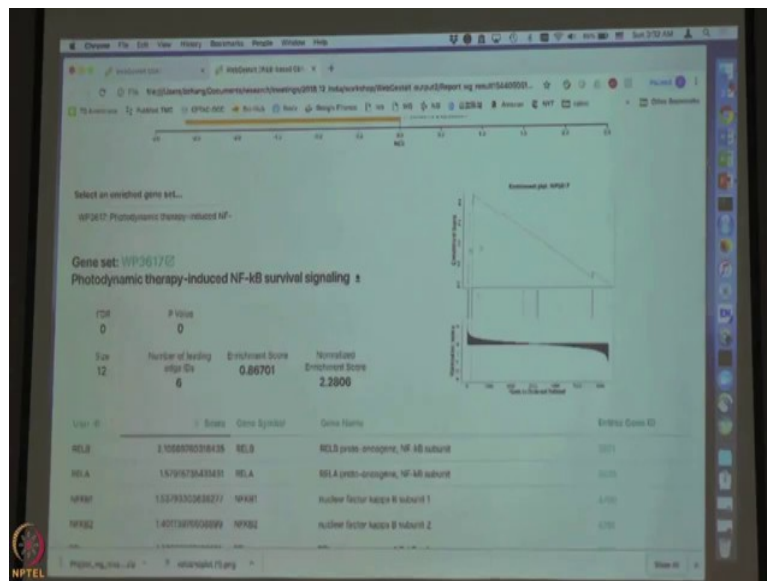
And in this case, for example this could be t statistic or minus P value something like that. So, that will allow you to rank the genes. But notice this in this case you do not set any cutoff, if you do a proteomic study and then you identified 5000 proteins and then you have a t statistic for order 5000 you need to put all every single protein in this. But in this format one branching of one protein and the value this and then the system will sort the genes and the

identify the location of each gene for specific genes that where the genes are located in the ranked list.

So, yeah because of the time let us just do the GSEA sample run. I think that would be easier than getting the file and upload. But when you prepare your file is basically same as this just a file text file with two columns remember for ORA it is just a one column or just the gene list. GSEA always you have to have a gene and a value from your statistical test.

For example, you can calculate the correlation of the genes to drug sensitivity and then for each of those you have a person's correlation and then you can use that to rank all the genes. So, this is a, if you succeed in getting the result, I mean the top part will be very similar to the ORA analysis basically you still get the bar chart or volcano plot or if you do the code you get the tag of your achieve, you enrich the terms.

(Refer Slide Time: 16:56)



The major difference is here and rather than the Venn diagram that shows you the overlapping genes between your uploaded gene list and genes enrich the genes. For example, let us see if we are interested in this focal adhesion, we click on this, and then it is, you let us say this is a ranked list, this is the data you this is your input. So, best mean now because every gene came by the value, then the system was able to rank all the genes from the largest value to the lowest value, right.

Here you can see this is ranked the new symmetric, this is the number one gene, this is a last gene basically from the positive values to the negative values. And now, for the focal adhesion genes this part best we represent, but where are the focal adhesion genes located on this rank the list. As we can see there is a tendency for these genes to be located on this part rather than this part, basically that means, there are a lot more up regulated focal adhesion genes than the I mean the very few down regulated focal adhesion genes. That is why it is enriched at this part.

Also it has this scoring plot enrichment score plot and this part is the leading edge, meaning these are the genes that give you the enrichment signal and these are the genes that are listed in the a table at the bottom yeah a total of 53 genes out of the 128 genes that are in the leading edge that is giving you the enrichment signal. So, basically if you think about the Venn diagram and this you can probably can help you better understand the difference between these two approach.

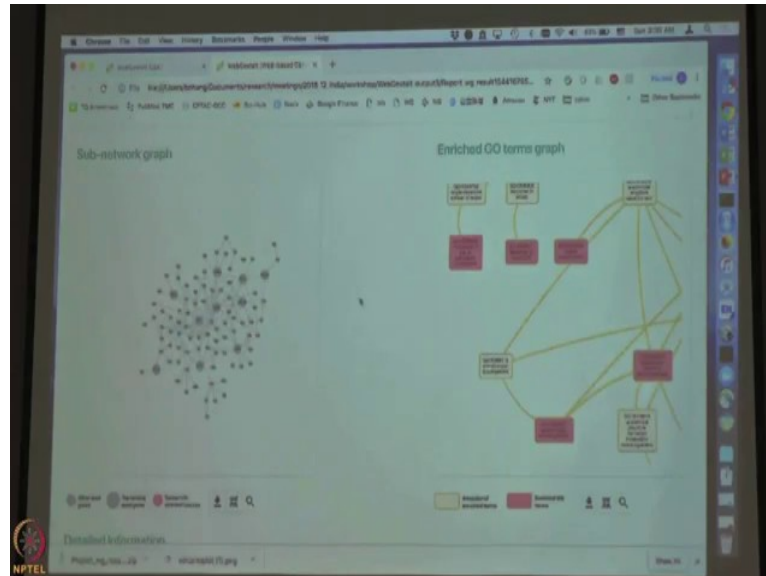
Well, approach you have a cut off and then it becomes two sets and then you do the overlap, and then you use a fisher test or hyper geometric test to check the enrichment of the overlap, but here you are not stating a cut off then you get all the venues for all the genes and then you are testing their enrichment in the rank the list. So, I think this is again I mean you can also at the top there is a download button and then you can also save and download and save and then share your results.

So, finally, and let us take a look at the NTA sample run. So, the NTA sample run the input is also very simple it is just an again a list of genes you are put it down let us say you have only list of differentially expressed genes or some genes you can just put here. The idea is to say I mean if I identify let us say we do association study we identified 300 genes that are potentially interesting which one should I test first, right, do I experiment first. And this will help you to answer that question.

So, in there are two options why is to do network expansion, but the other is network retrieval and the prioritization, this prioritization is when you have a lot of genes you want to pick the top and then you do this. But this can also be used let us say I can show you another example later let us say you have one gene, but you do not know the function of this gene, then you can just type the name of the gene here and then you can do the same analysis that will help

you to retrieve the neighborhood of that gene and then to go enrichment analysis to help you to predict the function for that gene. Let us look at this first.

(Refer Slide Time: 21:29)



I think the analysis probably will take very long because of there are a lot of people submitting job at the same time, but this is a result and if you do the analysis later and you succeeded in getting the result. It is not this, and at the top you have the job summary and the here you have the sub network.

So, yeah if we go back to here we see basically this is our input gene list, right and the network we chose the PPI power grid. So, this is a protein-protein interaction database. So, best thing we are mapping these genes to a protein-protein interaction network and then first the way we retrieved all the genes that are included in the network, and then you get a sub network like this. And then you notice that in the top 10 genes because of your parameter is and you want to get the highlights the 10 genes top 10 genes, right. Then these are the 10 genes that are highlighted and then you can zoom in and then it will tell you and for example, the FM1 and ERM and TGF β 1.

I mean these genes the top genes based on the network diffusion analysis that means, these are probably the hub genes or the have more connections to the other genes in this network than the genes that was not selected as for highlighting. On the right side and basically this is an enrichment gene ontology analysis for the network and it will tell you what has enriched the functions for these genes and you can click on this. For example, this is response to one

being and then that those genes will be highlighted. So, basically it provides you a way to retrieve the network, and also understand what are the major functions associated with the network and then you can also visualize and which genes in my network has that function.

(Refer Slide Time: 23:56)

Seeds in the sub-network			Ranked Seeds		Enriched GO categories (Top 10 categories)						
Gene Symbol	Gene Symbol	Random Walk Probability	GO ID	GO Name	C	Q	Raw P-value	Adjusted P-value	Signif. Probability		
THBS1	FHLN1	3.53E-03	GO:0001225	engorgement	377	26	0	0	»		
THBS2	THBS1	3.51E-03	GO:0001285	blood vessel development	537	38	0	0	»		
MMP2	FHLN2	3.28E-03	GO:0001944	vasculature development	557	36	0	0	»		
COL1A1	ELN	3.22E-03	GO:0008028	movement of cell or subcellular component	1817	63	0	0	»		
ELN	SMN	3.14E-03	GO:0001125	cell adhesion	1470	52	0	0	»		
COL7	FHLN2	3.14E-03	GO:0007278	multicellular organism development	4239	83	0	0	»		
VCAN	THBS3	3.13E-03	GO:0001225	engorgement	2147	76	0	0	»		
MMP9	MMP11	3.12E-03									
COL4											


And at the bottom you have all the genes you submitted and their ranked net FNI is number one and then THBSY is number two. But basically you can have your complete list here. We are here. this is a gene ontology enrichment analysis result for the network you have I mean. So, this is the typical gene ontology over representation analysis reported without similar to the ORA analysis, but focusing on the genes in the network and the using the gene ontology biological process for evaluation, yeah.

So, I these are the stream, I mean major functions for the WebGestalt. So, again the ORA and the NTA the input are just gene list it is very easy, you can just copy and paste one is to two here. But make sure you understand the parameter setting and get the right parameters and then the result is very simple to understand I think you, and the you can download the figures for very easily from the interface.

(Refer Slide Time: 25:25)

Points to Ponder

- WebGstalt gives us data of gene enrichment, gene ontology, PPI (protein-protein interaction) module and pathway analysis.
r-based approach
Module-based approach
- In pathway analysis we learned that we can choose different functional database name like KEGG, Reactome and WikiPathway. The filtering of pathway analysed data can be done with 0.05 FDR.




MOOC-NPTEL IIT Bombay

(Refer Slide Time: 25:35)

Points to Ponder

- There are three different types of Network-based methods like
Direct neighbor-based approach
Module-based approach
Diffusion-based approach



MOOC-NPTEL IIT Bombay

I hope today's session was useful to you where you got an idea about its exploration of WebGestalt result. So, we learned that WebGestalt gives you data of gene enrichment, gene ontology, protein-protein interaction modules and pathway analysis. In pathway analysis we learnt that we can choose different functional database such as KEGG, Reactome and wiki pathway. The filtering of pathway analyze data can be done with astringent criteria, such as 0.05 false discovery rate.

I hope you appreciate that resources like WebGestalt are so precious open access available to the community where from your complex mass spectrometry or even NGS dataset you can try to now further get an idea about what is the best biological sense of a data, how to really get the best biology out of that the complex mass spectra, how best now you can try to address the biological question which you originally wanted to start with. So, these resources are highly valuable.

Of course these short lectures and hands on sessions may not be able to provide you all the information, but more and more you make yourself familiar by using these tools. You will then appreciate that the same data set which you have obtained from these high through put technologies can now give you some very novel insight and probably the, right answer to the biological question which you wanted to address.

In the next lecture, Dr. Bing Zhang will teach you about network analysis.

Thank you.