

Proteomics: Principles And Techniques
Prof. Sanjeeva Srivasatava
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Module No. # 32

Lecture No. # 32

Microarray work - flow: Data analysis

Welcome to the proteomics course, in today's lecture, we will talk about Microarray Workflow, especially focused on Data Analysis. This is a continuation of our previous lectures where we talked about different strategies involved in the performing microarray experiments, how to acquire good images. And now we would like discuss image analysis, the microarrays have become integral part of clinical and drug discovery process, they have been used extensively to find differential gene expression in variety of samples.


The microarrays have been used for biomarker discovery, finding genes to correlate the disease progression is studying about effects of various drugs and toxin in a field know as toxicogenomics, testing the target selectivity prognostic test, disease sub class determination in clinical diagnosis and many other applications. The data analysis becomes very crucial to make sense out of massive amount of data, which is generated by using microarray based experiments.

There are many commercial software as well as free software available, which can be used to analyze microarray data sets. However any single software package may not answer all the questions related to a fundamental genomic or proteomics based questions.

(Refer Slide Time: 02:21)

Lecture outline

- Microarray data analysis
- Discussion on microarray data analysis
 - Normalization
 - Supervised or unsupervised
 - Analytical methods
 - Hierarchical clustering
 - Self-organizing maps
 - Principal-components analysis




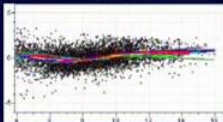
IIT Bombay 2 Proteomics Course NPTEL

So, in today's lecture we will talk about microarray data analysis, we will have a discussion on microarray data analysis to cover various type of concepts, such as normalization, supervised or unsupervised analysis, different type of analytical methods, such as hierarchical clustering, self organizing maps and principle component analysis. To elaborate clarify the analysis; I will have a discussion with mister Pankaj from spinco biotechnology. He will be representing molecular devices and we will have discussion on acuity software to give you a demonstration on the software operation for data analysis.

(Refer Slide Time: 03:25)

Normalization

- Microarray experiments are performed on multiple chips
- To compare multiple microarray measurements, data need to be normalized
- Normalization is performed so that:
 - Data from a single experiment are as accurate as possible (correcting unbalanced PMTs)
 - Data from different experiments can be compared to each other



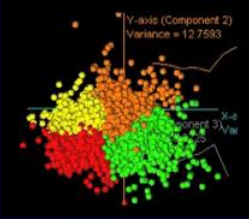
IIT Bombay 3 Proteomics Course NPTEL


We will discuss about various basic concepts including normalization, so in the microarray experiments, people use different type chips in different experiments, so to compare multiple microarray measurements data need to be normalized. The normalization is performed, so that data from a single experiment are as accurate as possible also correcting for the unbalanced PMT's. The data from different experiments can be compared to each other, so it can be performed by adjusting various type of parameters as well as using the expression level of housekeeping genes, we will discuss this in more detail, while looking at the software demonstration.

(Refer Slide Time: 04:15)

Principle Component Analysis

- PCA works by finding “supergenes” that: explain the most variance in the sample are orthogonal to each other



 NPTEL

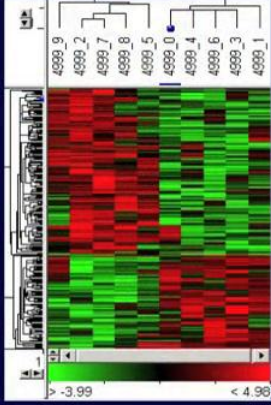
IIT Bombay 4 Proteomics Course NPTEL

Principle component analysis; principle component is a linear combination of optimally weighted absurd variables, to test whether the protein expression is consistent throughout multiple sample from same experimental group; are their protein outlets spots mismatched or true proteins to analyze all of these type of variations principle component analysis is performed. The PCA works by finding supergenes that explain the most variance in the sample are orthogonal to each other.

(Refer Slide Time: 05:06)

Clustering

- **Hierarchical:** genes are placed in a hierarchical relationship to each other, as in taxonomy
- **Non-Hierarchical:** genes are placed in clusters that do not necessarily have any relationship to each other



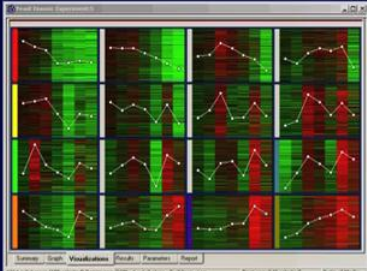
IIT Bombay 5 Proteomics Course NPTEL

Clustering, after analyzing the microarray set, you would like to cluster your data to find out the patterns of the type of question, which you have asked, whether your controls and treatments they fall into different clusters. There are different type of clustering, broadly hierarchical and non hierarchical; the hierarchical clustering involves where the gens are placed in an hierarchical relationships to each other as in the taxonomy. The non hierarchical clustering involves, where genes are placed in clusters that do not necessarily have any relationship to each other.

(Refer Slide Time: 06:00)

Self-Organizing Map

- Replicate Dye-Swap microarrays can be quickly inspected for quality using a self-organizing map.



IIT Bombay 6 Proteomics Course NPTEL

Self organizing maps, in microarray experiment it is important that you perform dye swap experiments to avoid any effects of cy 3 or cy 5 labeling, so that is no bias of labeling in the control and treatment groups. They replicate dye swap microarrays can be quickly inspected for the quality by using self organizing maps, such as one shown here in this slide then. There are different type of supervise approaches to determine genes that fit a predetermined pattern or unsupervised patterns, to characterize the components of a data set without (()) input or knowledge of training signal.

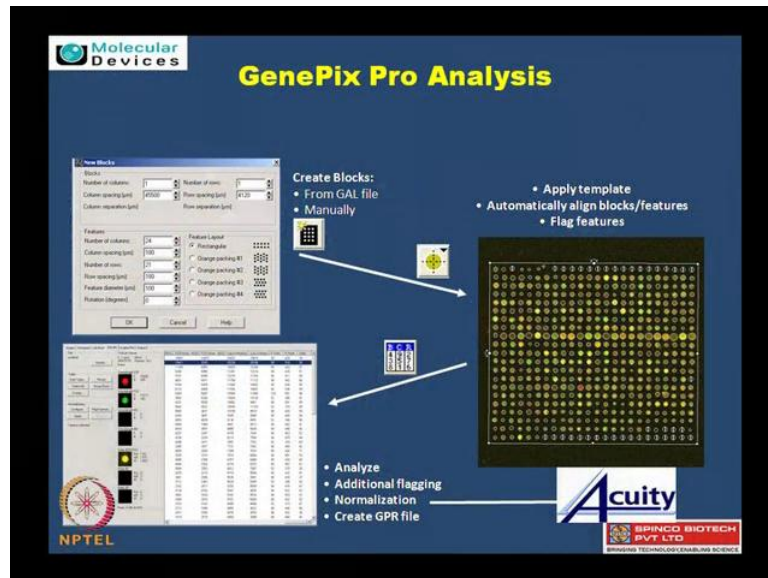
So, we will try to cover various concepts involved in data analysis and provide you the demonstration of the software, while involving discussion with Mr Pankaj from spinco. So, let us have a discussion and then we will conclude our today's lecture, this is pleasure introduce Mr Pankaj Khanna, manager application support from spinco biotech private limited. Today Pankaj will talk to us about acuity softwarem which is used for analysis of microarray data, the software is for molecular devices and spinco is distributor for the same.

In the last lecture we discussed about, how to scan the slides microarray slide by using genepix pro software and once the data was acquired, now next step and next challenge is how to obtain some meaningful biological information from that data. So, there are various software commercially available acuity is one among them; and to know more about how to operate the acuity software and how actually one can analyze microarray data, I have invited Mr Pankaj for this discussion, hello Pankaj welcome to this lecture.

Thank you, Dr Shrivastava.

So, in the previous lecture, we actually discussed about various type of parameters, which are used to acquire very good microarray image by using genepix pro software, can you just give us an overview of that whole process; how in a nut shell so that we are briefed about the same; and then we are ready for analysis with the acuity software.

(Refer Slide Time: 08:25)



Sure, so let us quickly go through with a genepix pro what we have done. Ok. So, once you are ready with the slide, usually people put on and hardware parameters are been selected, once that is being done the image is scanned and stored in the tif format. So based on the laser type 1, 2 or 3, 4 you get 24 bit maximum image resolution possible; and once you are ready with the tif image, you perform little bit basics of analysis in genepix pro. Say for example, aligning of a different features in the form of gal file which you have seen.

Once we have done alignment you go for the results, were the background corrections and all other things would be calculated and then given to the results different column tabs. So once you are ready with these results, this can be saved in the form of GPR file which stands for genepix result file, that is we just briefed as dot GPR file.

So, let us go through as we are seeing here in this slide that the first one is getting the image, getting the alignment done and once the alignment is done, the result tab after doing the result tabbing head, you get the different column details in the form of different stats possible.

(Refer Slide Time: 09:38)

Molecular Devices

GenePix Pro Analysis

Measuring Tools

- Detailed calculations from specified regions
 - Intensity profiles (lines, poly-lines, and wide lines)
 - Histograms (rectangles, polygons, and ellipses)
- Statistics and pixel information
- Quick measurements and channel comparison
 - No gridding or analysis needed

Background Subtraction

- Local
- Global
- User-defined
- Negative controls
- Morphological

Data Normalization

- Ratio Based
- Global
- Norm. feature
- Wavelength Based

Configure Normalization

Ratio-based
Normalize the data in each image so that the mean of the **Ratio of Medians** of **all of the features** is equal to **1**

User-defined
Normalize the data in each image by the following factors:

Wavelength: E35
Wavelength: S32

OK Normalize Cancel Help

NPTEL

SPINCO BIOTECH PVT LTD
BRIDGING TECHNOLOGY AND SCIENCE

So, once you have the results in place in different formats sometimes you require to have a measuring tool but, usually all commercial and even the academic software, gives the GAL file details so you really do need to do manually but, in case if you want to do, you can do it. (()). And then I guess, there are various parameter one need to look for by performing good scanning and acquiring a data, including the background subtraction and how to normalize the data can just elaborate on this.

Sure, so in the result tab immediately, what you see is a window which gives you configure which can configure different type of normalizations. So, there are different kind of actually background subtractions one can perform, so as you see in the image so if this is my spot in the yellow, which has a periphery ending in black. And the surrounding area, which are surrounded by white is can be calculated for the local background correction.

The local background correction is immediately near the feature, which is the area which would not have fluorescence should be coming in, which comes is just because of the background, that is called as a local background. And we also have a global, so any other place hole in the chip, where the spot is not present the different backgrounds levels can be calculated at a different specific position. Now, this can be used to calculate for the global background corrections as you have elaborated in the last lecture user defined

ones, say for example, you have a positive control, you have a normal control, you have also got a shape control, morphologically different ones.

So, you calculate them as features and allow the acuity in the configuration to allow, which one to go for, you also have a negative control, which totally gives only the negative background in the same area of defined other types so. I guess we discussed the need for these controls, how important those are and now I think we can see a cheer like, when we are acquiring these images, how each of the positive and negative control features, play a crucial role in the analysis process. So, after background subtraction, I think next important thing will be the normalization, may be you can just explain on that.

Yes so, important factor is normalization, because we do microarray experiments chip to chip basis, experiment to experiment basis, what happens there is knowing to that fact that different time points are being used to do the experiment. There are different ways fair in the variance can come in, so you want to avoid maximum possible variations apart from biology. So, these all can be handled by the way of normalization, so normalization helps to balance the chip variation across the chips as well as within the chips, within the chips we do because we are using at least two lasers at a time. Ok 532 and 635, so you want to correct for them, that both intensity should match the ratio of 1, so that the difference contributed knowing to the fact of the laser powers and the (()) stability does not come into play of biology. So, there are different ways of doing data normalization and the best suggested ones are certain ratio based normalization on the mean or the median values; which is actually a continuous type, which does not change the shape of the data, the meaning is that this is being escalated or collected but, nothing is lost in the form. Ok.

The other way of normalization is lowest normalization, wherein you really change the data structure, so there some extremes can be removed for the data balance to be made, which is actually little less preferred. So, major preferred ones are ratio based which involves global and normalization factor and the wavelength base correction, which can be done over where. Now, maybe you can just brief, so the analysis aspect of the genepix pro, before we just move onto sharing the data for acuity.

(Refer Slide Time: 13:36)

The screenshot displays the GenePix Pro Analysis software interface. At the top left is the Molecular Devices logo. The main title is "GenePix Pro Analysis". On the left, there is a "Normalization" section with a "Flag Features" button circled in red. Below this, the "Feature Flagging" section lists: "Boolean Queries", "Include in Normalization", and "Sort through data using quantitated results". In the center, there are two plots: a scatter plot with a regression line and a histogram. On the right, a "Flag Features" dialog box is open, showing a list of features with columns for "Name", "Value", and "Status". Below the list are checkboxes for "Flag", "Include in normalization", and "Exclude from normalization". At the bottom right, the "Scatter Plot/Histogram" section lists: "Immediate visualization of results", "Plot any Data Type vs. another", "Plot any Data Type on Histogram", "Regression Lines and standard deviations", and "Fully linked to Image and Results tabs". Logos for NPTEL and SPINCO BIOTECH PVT LTD are visible at the bottom.

Correct. So, the very important thing here is the flagging of the spots, meaning is as we that few spots are could be controls, so you do not want to take them for the further analysis what you do is you flag them as present, absent not to be calculated. So, this can be done by the flagging features, you can also give some boolean queries basically on the requirements what you want to avoid, so that the spots of the requirement go for the further analysis.

And once you include the normalizations or you do not include the normalizations, you can save the GPR or GPR result file, which involves the basic thing, which is required to correct for the images. So, once you have in hand all these things you can check for the Q C's in the form of scatter plots, histograms and so also immediate visualizations, in the form of data versus different intensity plots. So, which gives one an availability that fine I have Q C'd my data spots are looking good, all good spots are went in and we tried to avoid some kind of physical variations, which happen and this now can be saved as a GPR file which can be further do the analysis.

I guess, one thing is one need to unsure that, the data which is going to be further analyzed for any biological significance, should be cleaned, it should be a quality controlled check and all the control parameters are in place. And once we have verified all of those things at this stage, then only that data is actually ready for the next level of analysis. The more better you do Q C. Correct.

The more better biological results you do expect, so very rightly said that yes Q C most important one need to spend little bit of time on that. Specially, we need to talk about high throughput analysis, which in the case in microarrays he will be talking about 20,000, 30,000 in fact during very high density. During very yes yes. Arrays are available you are talking like, so many data points and attained in that one file, so until and unless you are very sure about the overall good quality of the experiment, I think otherwise you will be analyzing the very data. So, these high throughput platforms provide us an opportunity to analyze very, very large data set in a very short time. At a same thing, what is very important here that one need to ensure that the data quality is good, because if it is not good, I think it is better to just leave that chip a side and move onto repeating the whole experiment at once; because doing all the corrections and all the things will not help to really do the further analysis, until and unless you are starting with a very good slide to began with.

Very true totally agreed sir.

Yeah so, I guess now once we have Q C'd this slide, we are ready for the saving the data for the next analysis . So, maybe you can explain that.

(Refer Slide Time: 16:28)

Molecular Devices

Saving data to Acuity

- Three ways:
 - Click “Save to Acuity” button
 - Select “File.../Save to Acuity”
 - Press “CTRL+Q”

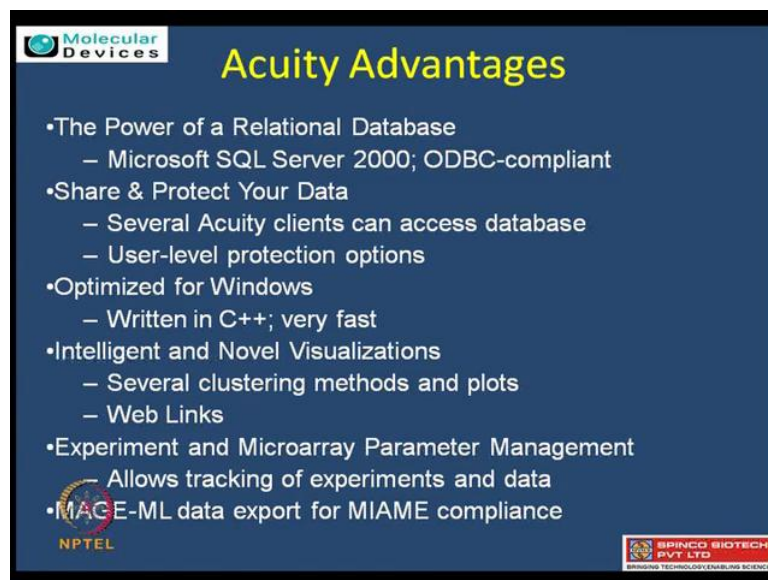
The screenshot shows the GenePix Pro software interface with the 'File' menu open, highlighting the 'Save to Acuity' option. The interface includes various toolbars and a data table with columns for 'Flag', 'Normalize', 'Autoflag', 'Block', 'Column', 'Row', and 'Name'. A logo for 'SPINCO BIOTECH PVT LTD' is visible in the bottom right corner.

So, the basic work flow involves that you do first level analysis, that is genepix pro involving the QC and then immediately the genepix protein gives you a direct compatibility with the acuity. There is a button on the side, which allows to say that just

save the data to acuity, and immediately the data is exported inside the acuity. So, they import the data based on the export from the genepix directly and not only this, acuity can also work as a standalone, so there are ways to import the data in the text format or of the different format which it understands. So, this is how the acuity can be used for the further analysis now.

So, I think we hearing about acuity, now so maybe we should talk little bit more, what acuity software can do, what are its major features. So, maybe you can explain just few points about acuity before, we move onto the knowing the details for the acuity software and what we can do with the analysis.

(Refer Slide Time: 17:30)



Molecular Devices

Acuity Advantages

- The Power of a Relational Database
 - Microsoft SQL Server 2000; ODBC-compliant
- Share & Protect Your Data
 - Several Acuity clients can access database
 - User-level protection options
- Optimized for Windows
 - Written in C++; very fast
- Intelligent and Novel Visualizations
 - Several clustering methods and plots
 - Web Links
- Experiment and Microarray Parameter Management
 - Allows tracking of experiments and data
- MAGE-ML data export for MIAME compliance

NPTEL **SPINCO BIOTECH PVT LTD**

Sure to began with acuity is band formatting software, so it gives you a power that whatever basic analysis you have done, through GPR can be now further taken for the analysis acuity advantages. So, let us quickly look at few of the acuity advantages, it is actually client server relational database understanding, so we give MS SQL 2,000 with this which allows you to save the data in the form of servers.

So, this gives the power that this can be your data warehouse, meaning all the important attach files save any file life TIF image, JPEG image, GPS that is setting files all can be stored with the result files, which allows one to again look back whenever you required to. And so also it is optimized for the windows, it is return in C plus plus, which is

actually gives a very fast power for it, so it can work very fast and give the results saving your time and so also allowing one to look at more different statistical possibilities.

Intelligent in the form of novel visualizations, we do have like different kind of clustering is possible, we have scattering available for you graph scattering coming in. So, this gives one an opportunity to analyze the data visually to quickly understand, what is happening at the biological of us experiment and the microarray parameter management.

So, many scientist want to give a different parameters and allow one software to sort or understand the biology based on that, which is we call it as a parameter files, actually this is this is being MDT files for us, which you can import and manage your all parameters, within the experiment, so that you group them and do the analysis accordingly.

There is a MAGE-M L data export, what happens is as we discussed different QC formats, so this particular MAGE-M L is based on the MIAME requirements, which says what all is required and one how to do for the microarray experiment, this has a direct export capability of that, so this gives one a very good opportunity not only from the data to the export or different levels.

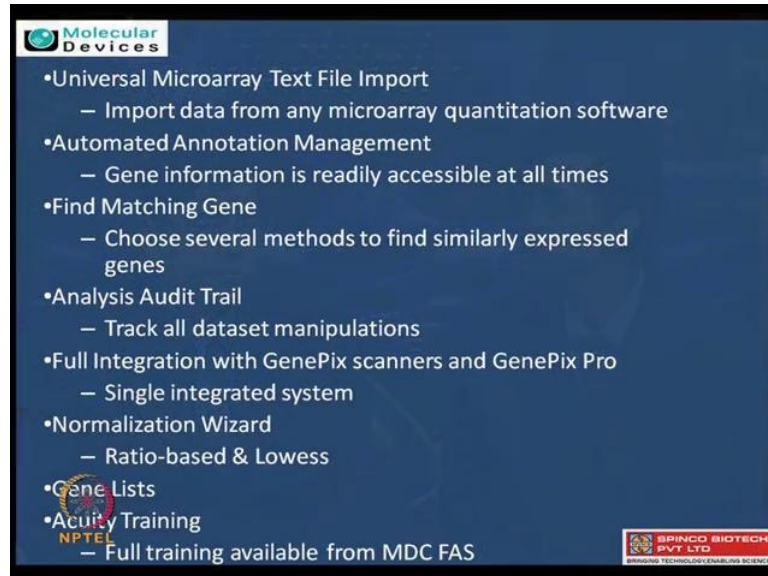
So, I guess last two point, which you mentioned, one is the tracking the data based on the experiments like the thing that is very important, that is also like depending upon the need of the experiment one need to in fact track and make software, learn your experiment so that...

True. One can actually apply the same knowledge for the various slides throughout to track that data set. Yes. Now, the second point which you mentioned about MIAMI (()), I think that is very important, because one need to do all the quality control checks and overall data analysis in the very uniformed guidelines provided; so one has to adhere to those quality control checks.

True so, another important factor is that as I said acuity can be a standalone analysis system, so not only the data coming from GPR only can be analyzed, so we are not restricted it to only genepix, it can also take other format even in the form of text format, wherein you need give an information, what each column means in; and then again you

can perform the same statistics. So, there is an automation management also possible with this, so you have number of slides coming in every time

(Refer Slide Time: 20:50)



The image is a screenshot of a presentation slide with a dark blue background. At the top left, there is a logo for 'Molecular Devices' with a green circular icon. The slide lists several features in white text:

- Universal Microarray Text File Import
 - Import data from any microarray quantitation software
- Automated Annotation Management
 - Gene information is readily accessible at all times
- Find Matching Gene
 - Choose several methods to find similarly expressed genes
- Analysis Audit Trail
 - Track all dataset manipulations
- Full Integration with GenePix scanners and GenePix Pro
 - Single integrated system
- Normalization Wizard
 - Ratio-based & Lowess
- Gene Lists
- Acuity Training
 - Full training available from MDC FAS

At the bottom right, there is a small red and white logo for 'SPINCO BIOTECH PVT LTD'.

So, we do experiment add on to some more so there is a possibility that you can add to your present experiment itself. (()) Which gives a very good opportunity, that you need not repeat over and over to understand what is happening. So, find matching genes, the best possible application of expression profiling is differential expression but, sometimes you also need to know the matching of genes at the level of tissues. So, even that can be handled very effectively here analysis audit trails, the meaning is that we can look at what all analysis is being done as in the case of genepix that logging, will be happening to understand what happens to each one.

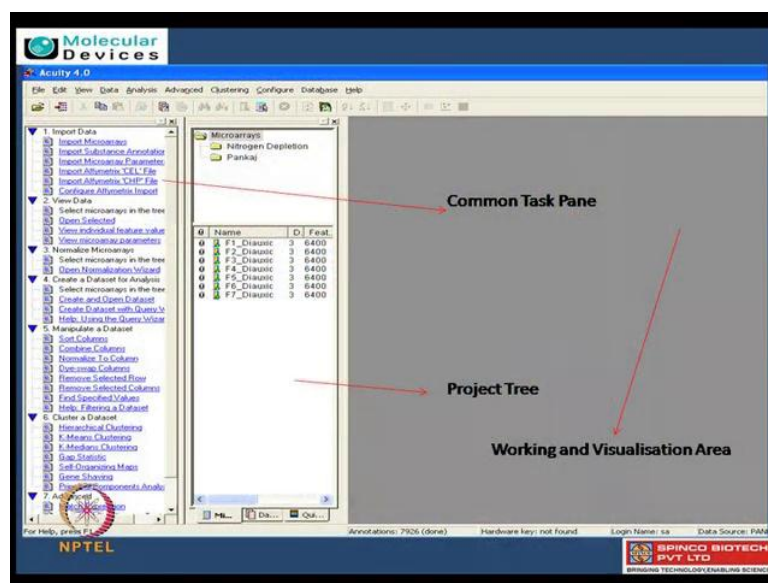
And you can always correct for it and look back one, what has done, so sharing becomes very important in that so full integration with genepix scanner and genepix pro, which allows the users of genepix pro to immediately store the data and start doing the tertiary type of statistical analysis. But this software is also compactable with other scanners and other platforms.

Yes as it can take up any text file. So, basically whatever if understand that it is coming from 530 to 635, tell that it is coming from this wavelength. (()) And still you can do the statistics. Sure so, regardless of what platform is being used it is just the wavelength and

text file which matters here. True, true and we give training at the level of different stages also, so that one can make become friendly with the software's too.

(()) it will be useful if you can demonstrate us about the acuity software, so that one can actually learn that how the data obtained can be transformed into the meaningful biological information and also that statistical significance of that data. But, may be you can just first share the software interface, so that we are familiar with the windows and all the keys over there. Sure. Before we switch to the real demonstration.

(Refer Slide Time: 22:47)



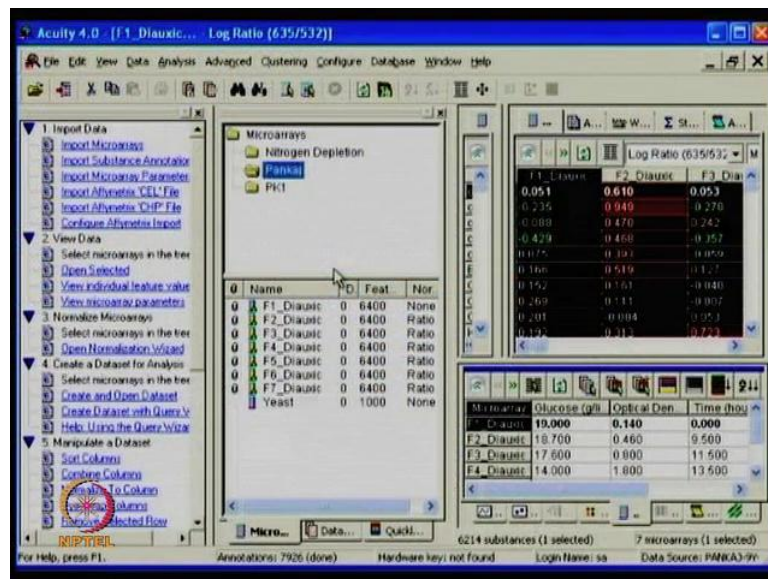
Sure, so what you are looking at as a G U I interface of an acuity, which is first diverted on top in the form of any typical file based drop down list, which has a various functions and then towards your extreme left you will be able to see a common task pane. So, this common task pane actually gives basic steps, which one has to do one by one in a flow, so that you end up with a biological information, the idea is it starts with a import of the data and end with the statistics and visualization, how one want to do. So, in this fashion common task pane actually a very good tool, for any new business as well as for the mature or the advance uses to understand, what one can do with the microarrays.

So, I think it this just guides you the step, why like how you can walk through the entire process. True, so it just gives you from the input to the analysis a stepwise that, what all you can do and what you want to do in and towards the middle, what you see is a microarray root directory, which houses all your data in a different formats. So, this is a

warehouse point on the top in the folder base arrays and on bottom, it shows individually the each one slide by slide. (()). And towards your extreme left you are seeing a area which is a working and visualization area, where you do or output different task, what you have done towards the common task pane or towards the advanced one, so this is a basic user interface of acuity. I guess now, we can actually look at the real software and the data demonstration, so that we are very familiar with the stepwise analysis.

Sure. So, pankaj in the last lecture, when we talked about genepix pro, then you showed one yeast slide, how to scan that ease slide by using the genepix pro software, I guess now it will be good if you can use the same slide what we scanned in the last lecture. And see how we can analyze that here, so this slide actually was used for looking at the glucose response in the various time point, from 0 to 20.5 hours in east and I think it will be interesting to see what type of trends, we observe in various time point with the glucose utilization here. Sure. So, please use that slide and... Sure. We can look the demo here, let us walk through the data.

(Refer Slide Time: 25:03)



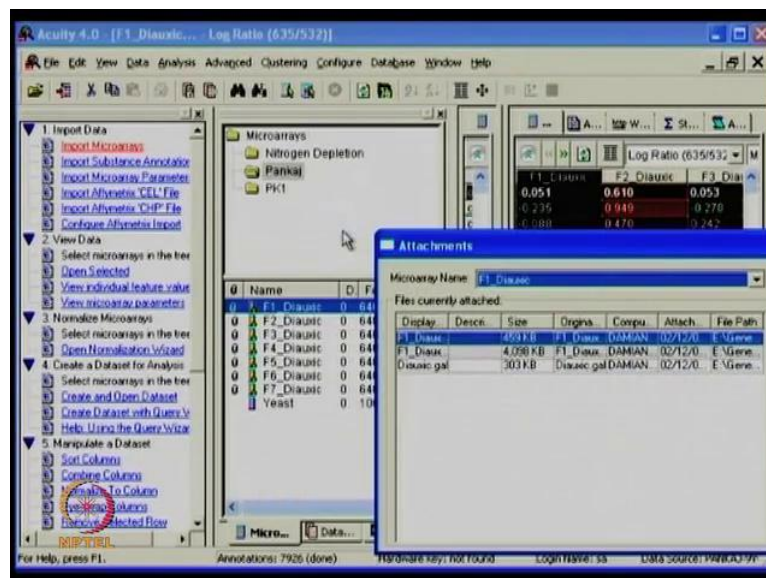
So, as we describe so we see a first import data tags, so basically this allows one to take the data from the microarray database and store in the form of GPR file and allow that to be (()) by the acuity software. We also have opportunity in the form of text file import, where you will define what is available for what, so in this fashion I here, I have defined

the microarrays the folder, where it says about the training and it says, what all different slides are available to us to design.

So, here we see seven different kind of time points collected and these are individual slides which one has run with both cy 3 and cy 5. So, this is ratio based image, which we are going to see in and we have also got an yeast chip in another file, which is come from the text to show you even that can also be imported. So first you have to individually scan each of the images, make one folder where you group all of this and then use that whole data set for the combined analysis.

(()) So, you can have all the GPR file stored or one by one. (()) From the genepix pro and just import the data in the form of import microarray data file, so it it just goes on looking for the GPR file. And this now can be imported and ready for your analysis, so once the data is in this will be displayed, in the down in the form of all over, which you have kept in for the analysis. So, here the folders are as we discussed, that this can be used for a data warehouse as such, so this little (()) kind of image here, tells you that there is some files are attached.

(Refer Slide Time: 26:53)



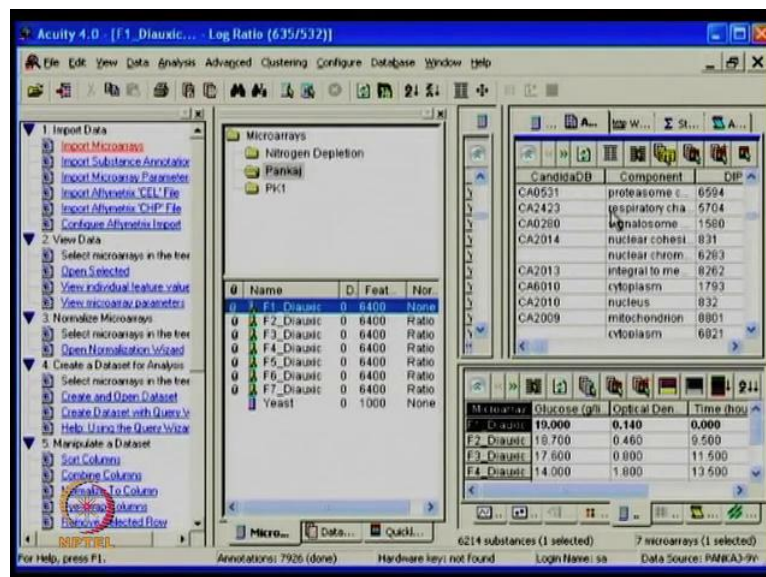
You can view them you can see them that what all somebody has attached, so if I see view all attachments, I will be able to see one has attached to it the important files coming from the GPS in the form; so also a GAL file and the image. So, it gave me an complete opportunity to look at it is good to emphasize here, if you have a GIF file this

particular one can also behave a partial visualization tool as in case of genepix. In case you want look how the spot has behaved, so in this fashion it, it can be any any file can be attached to it.

And since you have the gal file is also know, the general list so at any time point if you are identify any spot which is looking interesting, you know what that gene is try aligning with the gal map. Correct.

And another, this is a good point you raised, because very important next step is a substance annotation; the substance here I mean is a each spot, which could be the feature, which is again could be a RNA or a gene or a protein. So, that is why we called as a substrate annotation, very few people extensively trace them.

(Refer Slide Time: 28:19)



So, here I show you in the form of tab data in the annotation one can look at what all different information can be seen for each tab wise. So you have the same annotation tab being given for the substance I d and then because it is each database of candidized being attached here, component different functions even at the level of enzyme commission numbers. So, at different annotations which people try to get in, which also you can import in the form of text delimiter renamed to dot SDT 5, which allows one to take all the annotation information possible, so another very important thing the parameter five.

So, as we have said already that it is very, very important for one scientist to look all the parameters are grouped accordingly.

This can be made again in the text delimited form and can be renamed to dot MDT 5 and this again can be imported to look at all the parameters are visible in the form in the down window here to look at. In a few seconds more you will understand what each window means but, as in this case, you can just switch over to different type and you can just go to parameter five and look at what all details I have. So, maybe you can just briefly explain each of the tabs. Sure.

So that students are clear about what they can infer from each of these windows. Sure, so here in the working area which we have defined it is again splitted into two, so which allows one on the top to the level of different data visualization methods. So, what all different features are there and what all different arrays, say for example, I only opened one array it shows me only one array and it tells me what I am looking at. Correct.

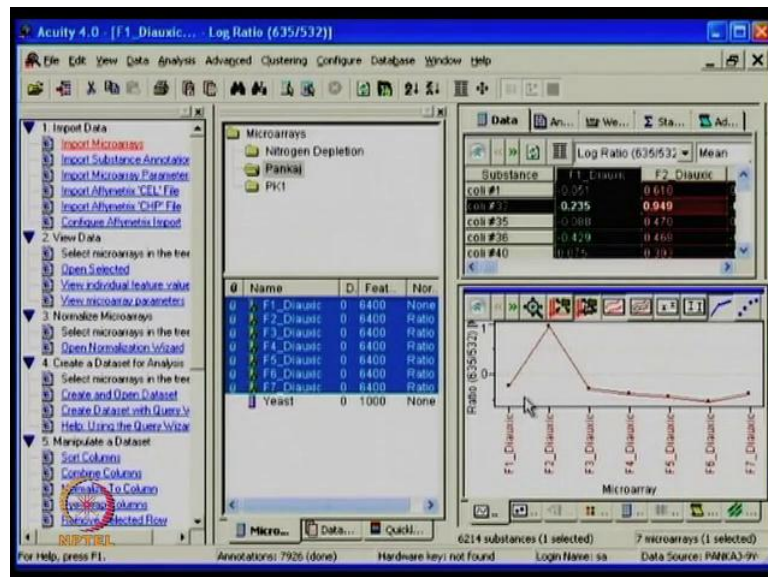
I am looking at log ratio data. Yeah in the similar fashion, I can look at any other 1 because it is GPR import what. All data you have. g p pix all the data you have to watch for. You can look at the background signal individual intensities but, majorly used either log ratios for specially expression analysis but, if you use protein or single wave length base, we can look at only one wave length base one.

So, you can always control what you are watching, apart from this the other tab include that annotation, which gives you that what all different one is tracing at the level of annotation teams, in the form of different data bases information on the genes or how the protein is behaving or even the localization.

So, where it is being localized all that can be traced and the other one to that it also gives little bit of other details in the form of statistics, warehouses and few are the auto scripting capability, which advance uses sometimes want to use. But, this one particular statistic one allow you to see what all you want to see which we see in detail, in bottom what you essentially see is how the data is looked at, many a times what happens let us say quickly for example, I want to go to one data and I would like to see how the first base is looking at.

So, What I am going to do is look at each particular spot and look at the profile of affect so, because it is only 1 it is showing you 1 dot point. If I in keep on including my more and more arrays the data bolts start increasing and immediately one feature profiling how it has performed can immediately seen. Can we select now couple of arrays and align.

(Refer Slide Time: 31:24)

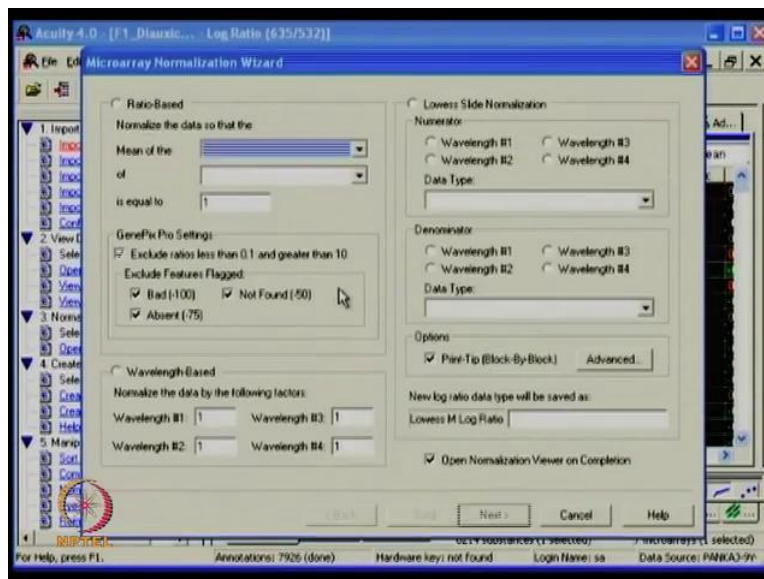


So, the way to select here is just hold the shift button and if you want to select only one one more or if select r to the last all can get select that shown. Then right click and click on open selected, what it does is it is allow you to open all the images here. So, it is given you all whatever calculated one you want to display and if you click on that each profile, now can be seen here with refresh button if you keep it will be able to see all. So, in the down, if I just click on the profile button based on what I have selected, I can look at the different profiles how it has behaved. So, in this fashion immediately I know my parameter file that each one is what is it and I see that this is all normal average in one of the case it went up.

So similarly, different features can be individually analyzed and then checked at how the behavior is. See, you can actually look at the trend for the same gene, during through whole time course analysis. Correct. So, you can also trace little bit of working on the data before going into this. Correct. Let me explain an important factor here, what does each images mean. Sure. Actually, if you carefully look may not be very clear, that this is little purplish in color and the other down one's are little reddish in color.

Yeah The purple one means that the data is not normalized. The red color means that the data is normalized and little dark green color what you are seeing tells me that, I have a jpeg image which I can see down here. So, it allows me a connectivity of what is happening just by visualization. If you want say that, the data is not normalized it is a easy process of doing it, so you once imported the data on that you can just click, right click and look at a normalization wizard.

(Refer Slide Time: 33:31)



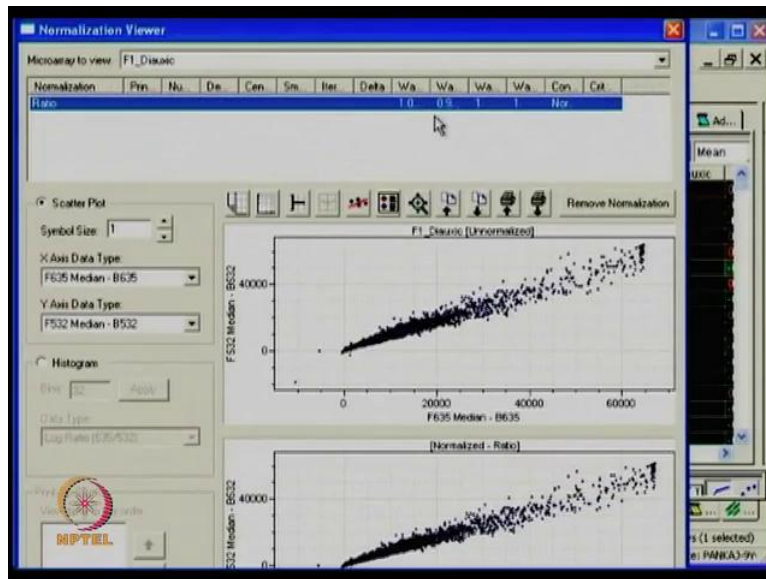
This normalization wizard allow one to chose different kind of normalization process, which we have discussed earlier, that it could be a ratio based or lowest normalization base, it is continuous and it is discontinuous type. So, one can select but, one has to remember the way one has been normalized all my time point has to be normalization. Same way. Yeah. So, you cannot cross difference in the form of a different normalizations and compare them.

So, it you are looking at little bit of a different (()). Now, because until and unless normalized in the same scheme, you cannot compare those across, why because you are going to compare different points here. Correct. So, as other one's have being analyze in the form of ratio base, I am going to select the one, you have an opportunity to select different types. I am going to select ratio of medians. which is the most preferred and if you see again just the next button with all the flagging. which is said if they are flag please do not allow for the calculations. And then you just click next, it allows that it is

available for me it is being done and I can say finish, which allows that fine I have finished my normalization.

So, if you carefully look back the spot there with the purplish will change to red, so which allows one to understand, that yes all my images are now being kind of normalization in similar fashion. So, I guess we are dealing here with lot of data set, so it takes some time for processing the whole thing.

(Refer Slide Time: 34:59)



So, it describes how many flags were there, so 6,400 at a time were analyzed and which you are see after doing the normalization and before normalization, how the data looks like and after normalization, how the data is looking like, so you can look at in different way, so what we have done is we have corrected at the level of background. (())

And I am trying to display across, how this an x y is being scattered together before and after normalization. Once you are satisfied with this (Refer Slide Time: 35:28), now if I look back, now this has change to same color of reddish to from the purple. Red. Which tells me this is also. Or a normalized.

So, you can also look at the if say want to reconfirm, which way I have done the normalization, I can always go back and look at normalization viewer which allows one to say it is ratio based. And you can look back what kind of this one is being done for using the normalization process, so I can cross check once again how fun is going about.

So, once you have all the data being normalized after the import and you have all the places in the form of annotation and the parameter file ready for you, these are few ways which which ways you can look at the data the best. So, I think before moving forward it is important to ensure that normalization was done properly and one need to look at each slide. Yes. So, as discuss whether it is very essential to have same normalization process done for all. Yeah

And it is not a thumb rule that which one is more preferable, one can chose anything but, make sure all your different slides are been handled in a similar way. And you can do different ways, get the data and do analyzes in the different normalization process also. (()). So, one have then opportunity to even correct back, so because you have rise the point say for example, I want to remove this normalization and put some other normalization. I can just click here remove normalization it removes normalization.(()) It allows go back to the raw data.(()) Renormalization in a different chance.

May be you be you want to do now lowest normalization Sure. And check back all in that format. And you can also select multiple, in the similar way in the in a one shot itself, you can do normalization in a single one. So, this is what I prefer, so usually you do not mess around be different kind of normalization. This is either select all are removal, because I have imported when to show you how the process is being done here and then immediately one would like to see how my data looks like. Yeah

The one way to look is the numbers which is little combustion, other way people like its color if you carefully observe the coloring scheme is going here. Red, black and green. May be you can tell that it is a conventional things for the microarray, people always represent in these colors, so what each on these color code means. Correct, so the black color is towards one, the meaning is when I am looking at the data you expect that when green and red channel both are giving same color, same intensities it becomes blackish in color.

If they are up regulated, people put them towards the red color and if it is down regulated minus sign will be given and that would become down regulated. The idea with this is which laser is being used what, so in context we have shown you are using which kind ratio means to check so usually it is case over control what people report for. Yeah

So, in this fashion conventionally you can see the colors but, here there are many one's they are many, which we have open of six. So, I want to see inner cell what is happening acuity allows you to do it, by seeing you can do an auto fit color, so quickly the numbers have gone only colors are shown to tell you how each particular substance or gene has behave acrossed your samples. You can look this is being attach, I have just split tables, so that I column look back at the annotation also, I can have just split table available put an annotation file here, so that I keep looking at what I am interested in. So, if there are it is up enough opportunity to for you to play around, so how you want to look and customize your view.

So, I think these type of heat maps away gives you feel about, what type of genes across each time point has shown, the variation or modulation in the expression profile looking at the color itself like for the first one. I can just easily say it is going down as we are moving across toward the 20th hour. True and so, I think by looking at this type data one can visually actually, get the feel about so expression changes across the different time point. Very true.

So, this gives you a immediate visualization tool to understand, what is happening and get a rough idea and this is mind you just the neat raw data. (()) that have not perform just the normalization. Yeah And we are just seeing how they are behaving; so it gives you rough profile I have some biology, which is going for this particular design of experiment. So, with here on if I want to go back to numbers or auto fit my data, I can select appropriate I can select appropriate one auto fit or data, so it is says it just fitted based on that, so it again show you the number back.

So, we have got the data imported now, we have done the normalizations, we are trying to see how they have behaved. Very important thing which sometimes people like is in the form of like able to move the data sorting up and down but, before doing that acuity tells you that first you make a dataset. The meaning of dataset is this is just looking at the raw data and I want to extract the data and allow to keep in a dataset here towards the down.

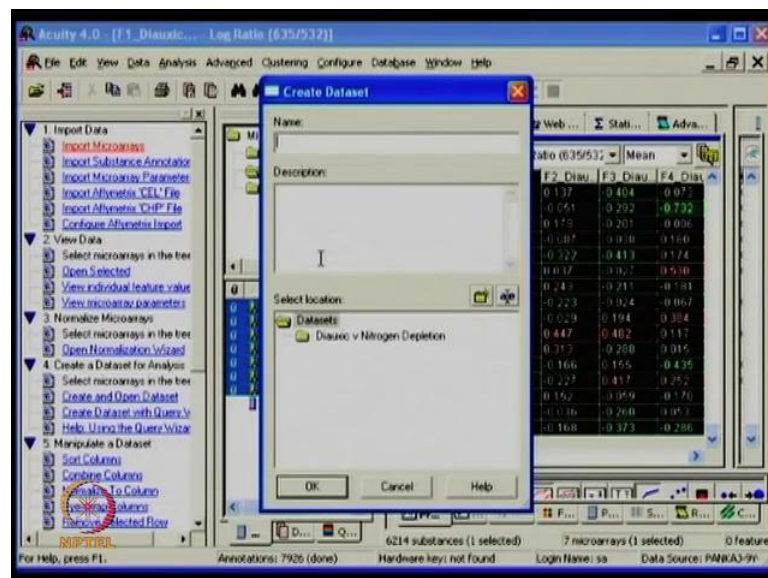
So, there you are available with all the different kind of things, so now there are two ways of doing things in the dataset, so one thing is take all the features and sometimes people say I want to have my criteria's defined, such that my visualization makes more

senses to me. So, maybe one can actually be stringent at this stage itself and say I want only very, very biological significant ones so. True. Define the free values. True. And then just sort the data based on that. True. So, people can do which are changes up and down with a range of so and so, which is twofold up regulated twofold down regulated.

So, usually differential expressions data is being logged, the meaning of log to the base 2 is essentially log to the base 2, value 1 is equal to twofold change. So, you are talking of log to the base 2 means you are talking of fourfold change, which really become significant you can filter based on different parameters and generate the data sets. So, essentially you are reducing the numbers, so you make more sense in the form of visualizations.

Otherwise also you can take all the data and you can do it, let us quickly see how we can do that particular job. Sure. So, I can select my data again holding the shift, I can click and I want it will me what you want to do it says you can have different kind of opportunities but, you can create datasets from selections, so this allows that whatever I have selected to create a dataset from that.

(Refer Slide Time: 42:15)

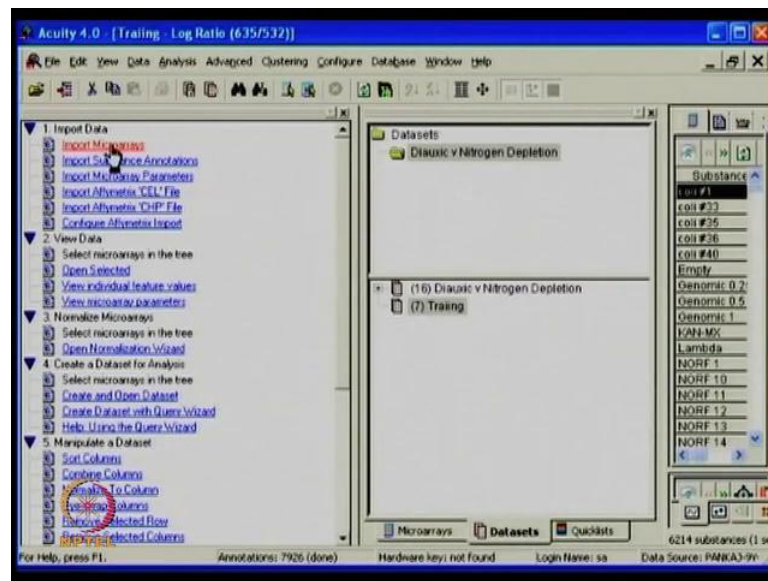


Once, I click that it is says here, where you want to keep in the folder so before hand I can generate, my own folders I can define my studies. And I can place them say for example, I am going to give it a name called training and it creates all the data from that.

So, here you go from 7 microarrays you have got and you have got all the data Exactly yeah.

Now, this is the one to get all complete data you can also have other ways as you suggested at p value importance coming or value is coming in; you can define a criteria of doing that, the way to do that again look at a common task pane, we have done the three steps and now after doing normalization, I can click and create an open dataset.

(Refer Slide Time: 43:11)



If for actually you move just technically good idea, you just refresh again sort of the task stepwise. So the... First was the import data. Correct. So, if you see first was the import data then importing that, substance annotation file which is the dot SDT file. Which gives all the annotations, which you have made a tab delimited and name it to dot SDT, then the third one is microarray parameters, where it is being stored in the form of tab delimited named to dot MDT file. So, which allows one to trace all the details, the other three you have seen is based on the affymetrix, which is actually a cel file. So, there image is actually a dat file, which they convert to the intensity level cel file.

That can also be handled but, just it does a RMA analysis, where they are moved little bit further for the different kind of analysis but, it does give opportunity, even to look at affymetrix outcome data. And then you have a viewing data, so as we have seen you can auto fit the data you can look at the color being coded, you can view the data in a

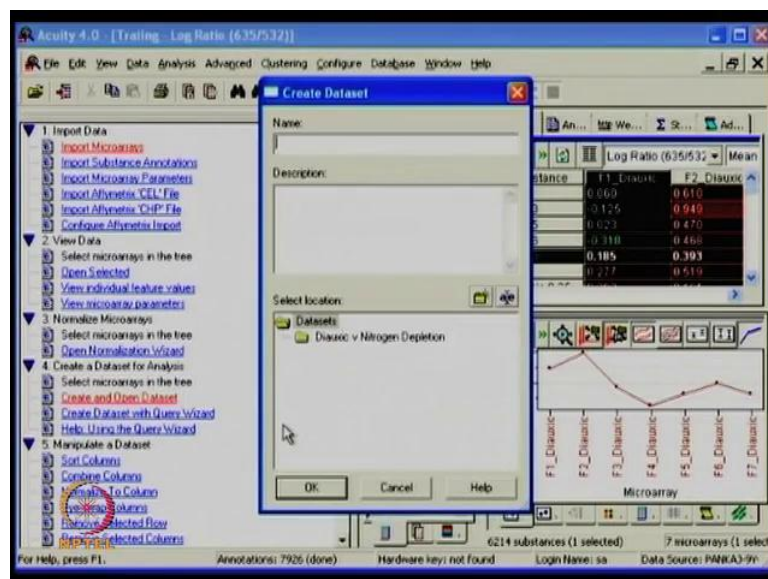
different form, you can individually looked up look at them or look with the numbers and comprise them and see all the arrays how the behavior is happening.

You can also do look at the profiling at the down, based on what you have seen earlier you can look at any of the things click on the profile; and you will be able to see how this has behaved acrossed. So, apart from that next step involves the normalization result, so I need to make sure that, all my chips have been normalized similarly.

(())

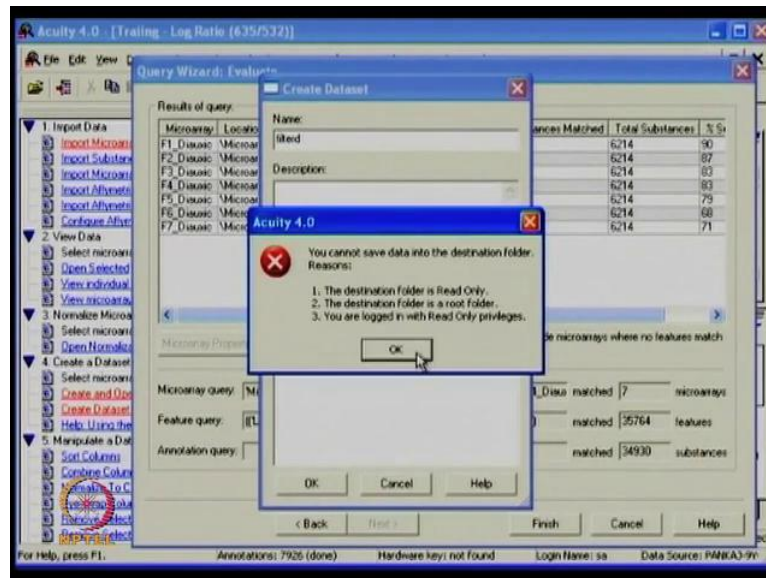
Or essentially, when I import the data I prefer it non normalized and all I select and select one base of normalization. I can create same level of experiment with different may be different normalization (()) and do the further analysis down there. So, once we have finished the normalization method, we can look at a query which is creating a datasets, so as we described one simple way is write at the data select the chips and then create a data or you can come here and say create a dataset from the open data sets. So, how to go for this one, if I click on create and open datasets.

(Refer Slide Time: 45:08)



Yeah. You can see similar way it has popped up the window, where it shows in the dataset the mother folder and then the child of it what all can be generated now. And then I can create a dataset here in this form, which we can do directly there very other important factor is creating a dataset from the query.

(Refer Slide Time: 45:28)



The meaning is that you can define the different criteria from the design of the experiment to little bit of more details of statistics to import, what I want to import inside. Say for example, quickly here I can define, which experiment I want to do type or based on the parameters as such I can say that, I want to import the data in the form of only one particular parameter, this is little little tricky, because if you have same name multiple times the data will be imported twice. So, you need to make sure your folder is right and then you have given only one copy to avoid that. So, I can select a folder and then I can say, I am into this particular folder and I want to import all the data from that.

So, what happens is it selects again in the same shift fashion, it creates a query you can add an add query and it can be created in this fashion. So, once you have done you want to say which parameter you want to select. I can just quickly do this for you that, I want to take a ratio based one, I will just log ratios and I can create a parameter of less than or equal to 0.3, 0.33 is log ratio changes something like two.

Yeah Add to list and greater than 2 greater, than say for example, 2. Again I can add a query, so it if I select both of them, I can create an OR or and (()) if I create OR you can apply that any feature, which is this OR that you select that for my import. So, I will be able to just add an query on this to the final result, then you have the data available for you coming up; you can even select the database basis of annotation and where you filter only with a mitochondrial specific.

Depends on what questions you are asking, so quickly look at only the, what you are saying, say it says how many of them are there in that, so it reduce the numbers. Sure. So, in this fashion I can import a limited number of database as well, I can give a name to it like filter and it will be important. Ok So, you cannot create files on the root directory. You need to create your folder and do the job, so it can do the job in this fashion. So, once you are ready with your dataset, what you want to do sometime is based on your experiment it gives you an opportunity what all you can do, number 1 you can sort the columns meaning f 1 to f 7, I want to make f 7 first.

I can use a sorting by that and so also I can do a different ways of combining columns one, why you need to combine columns, say for example, I have given three technical replicates for each. And then biological replicates of three, so I can just take an average of that combine the columns and it will take an average of all of these, and you are ready with the data to go ahead further analysis part. So, many a times you can also go for normalize to column, the meaning is I am having the 0 to maximum, so I want all my data to get normalized to first column. This feature is used when you are essentially using a time point, so all will be standardized or normalized to one column which you have defined as a 0th one.

And very important one, other one is the dye-swap, usually we know that cy 3 and cy 5, cy 3 is little less in size cy 5 is little bigger; so there is a variation in the incorporation of that, to take care usually people do at least one dye-swap experiment to accommodate the variations happening because of the dye. So, what you can do is, you can apply a dye-swap I can quickly show how the dye-swap works. Dye-swap just changes the way you look at so you have taken a ratios of 1 by another wavelength, so it is just reverses it so it is just minus x, so minus will become plus plus will become minus.

Sure. So it just changes the dye-swap, which will take care in the form of combining the data and avoiding the variability happening due to the dye-swap effect. We also talked about dye-swap and I was talking to them about dyes technology and how one need to use in fact labeling with different dyes and reverse dye-swapping, so that there is no dye bias in the analysis. True.

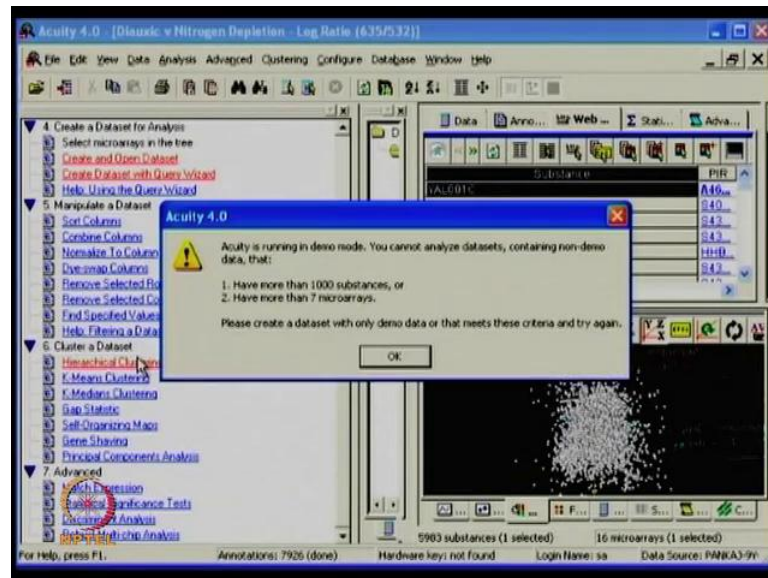
So, same can be accommodated here at the microarray to look at I think, so there is nice actually to go ahead with. And then many times which you find that the few rows, that I

want to remove I can remove few rows, I can select few of the columns, I can remove say example some Q C has not passed, so I just remove that this can allow me to do a different ways. And then once you have done that you can go and do a clustering, which which a visualization method.

So, technically the clustering is divided into two types hierarchical and nonhierarchical; hierarchical means that in the starting you have only one start point and then all other feature are attached to that. Other one is nonhierarchical type, where each group behaves independently of each other, so there are k means, some k medians and the particularly one the people use hierarchical when they do not know where to start with.

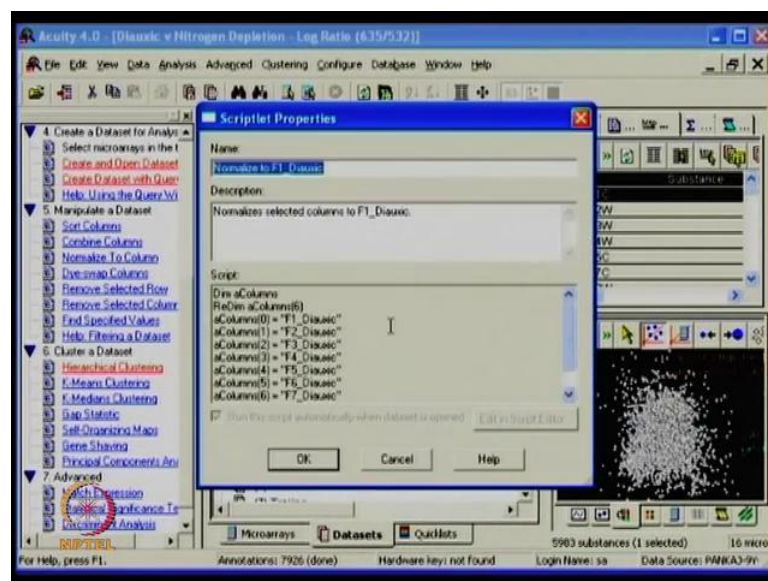
So, they start with hierarchical, when they do not know how many groups could happen and how results I am expecting. And once you do have the results and idea you can cut down, because hierarchical is little ram consuming, it will take little bit of more time. Whereas, because you can imagine all has to be linked to one and there are different ways of doing it center based, the best like coefficient correlation based or distance metric based and lastly binary based. So, binary based is usually used for only C G S kind of analysis where it is present absent type. Whereas the earlier two ones are extensively used in the microarray data, the Pearson's correlation with centric are being used, for the median or the mean type variation. So, when you quickly do it we will be able to see, what we can do say for example, I want to do on my filtered data which I have already done one of the clustering's, say I have got the filtering, I have got the log ratios.

(Refer Slide Time: 52:12)



And another important thing is whichever is something bluish in color, that particular one you have selected to work with. So this filter data... Yeah so, this is actually a filtered data and allow to work with and when I hit a kind of clustering possibility it say you need to create a quick data set .Ok. The meaning of that one is, you have the data already in hand but, now what you need to do is look at a third tab, which allows one to look in the form, which you want to create the data set; for which you want to analyze going little lengthier.

(Refer Slide Time: 52:45)

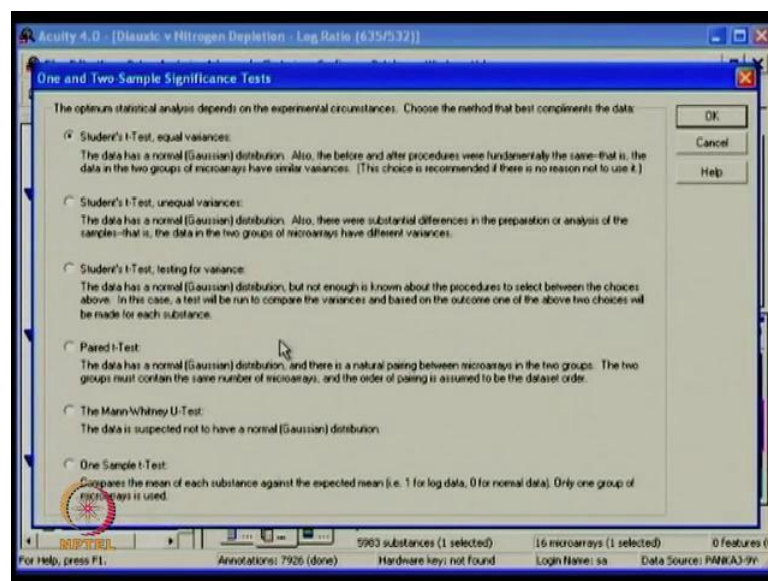


Yeah. So, here I have normalized the data and normalized the data to f 1, so I can see the skip what is happening; so I can just look at the one which I have performed at the level of different processing, so that I can look at the data. So, here when I look at some which I have done at the level of 4 into 3, clusters, soms is something like non hierarchical type so they are individually being blocked. So, all different genes behave differently and based on the profile, they are made into one group, so because I have given 4 into 3, I will be able to see 4 into 3, so 4 number of columns and 3 rows.

So, independently that is being divided, so one can have any numbers. The idea come from the hierarchical but, as you see here, immediately the clustering gives an immediate response, that this was very low in this once, again it went down and it you are able to see it again up. So in different time code it is a different expression.

Correct, so based on the behaviors you can see that this is being grouped up with, so there are different ways which people prefer but, usually preferred ones, are the candal ones, for the hierarchical type after the Pearsons correlation and further this one after soms, people usually use (()) squares using some of the non hierarchical type. So, k means and k median is a better ones to use with, so this actually gives you an opportunity how the data is being visualized. So once you have visualized the data, so what you can do is you can look at a statistics.

(Refer Slide Time: 54:24)

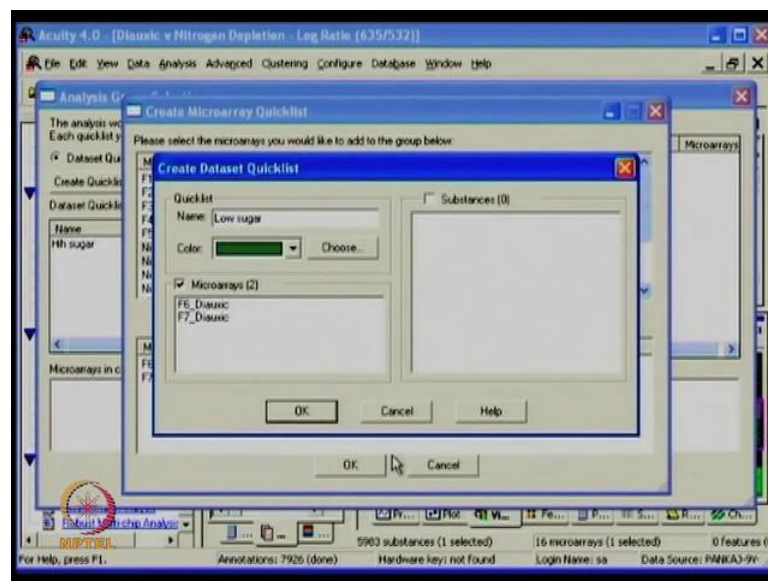


Statistics in the sense, if you click on any of the ones which is something like statistical significant test, it tells you I have different options to go ahead with. Different type of test one can perform. Different type of test one can perform, like student t test and the reason is, when it is a equal variance with Gaussian normal bell shaped curve, you prefer this test. When the variance is not equal you want to go with the second type, where there is a small modification of the again student t test but, without normal variance happening.

And there are student t test, where you do not what actually variance means to and you prefer paired test, when the sample is coming from the same origin; specially in the form of cancers, where the when the cancer is being removed surgically people remove normal samples, so that will be In the same time. Yeah. Same patient. Same biological patient

In the original pair, so this helps this particular one will allow one to select based on what background you have and the other one is Mann Whitney test, which actually people use where there is a no normal Gaussian; say for example, few times it is only standalone ones. So, up and down that is it so you want to select this parameter for that, so based on right kind of design one will select the different kind of statistics available.

(Refer Slide Time: 55:38)



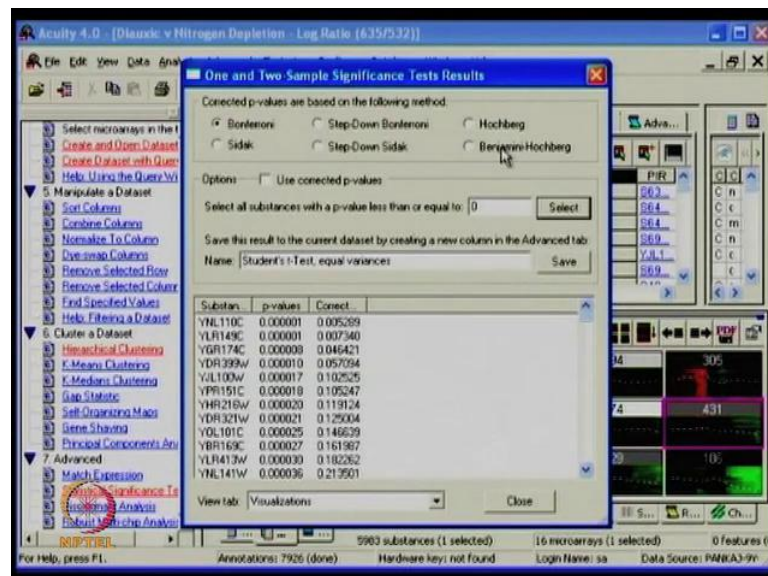
And you perform the analysis data. So, one need to look at their experimental design and then select the statistical parameter for the analysis. Correct, so in this fashion, what

the it does is it tells you that I have got different kind of data sets, which you want to select with; so I select them and I say group them according to like 0's and 1's like all cases and all controls.

And then I perform the analysis. So, let us quickly do that this data is based on something like different kind of components, functions, so I want to understand the differences between the different functions or I can quickly say you can quickly create a data set; so I can create a data set say which one you want to create with, so I said create a data set from all these. Different time points we can (() Different so, basically I know that it is high glucose at the beginning, so I can group them all as very high.

Because, this is a time point although it is decreasing more likely as select only once in only 19. Sure. And then compare with very low time and I want to see, how they are getting different changes. Overall changes. Overall changes is happening in the low and high level change, add to quick list data and it is available tool and then the name I give is high sugar. So it is available for me now and I can also create a one with low sugar and I am going to say I am going to compare this. These two groups I can differentially color them and I can see how the things are happening.

(Refer Slide Time: 57:24)



And you are ready with the different things a very important thing is for the multiple testing corrections, it also gives an opportunity of correcting for the different multiple

correction type. Say for example, Bonferroni which is being used Hochberg, Benjamini Hochberg; so they are like different ways of correcting for the multiple test corrections.

So, you can apply different ones and look more preferably it is Bonferroni, which is more stringent type Hochberg and Benjamin is little lenient on the multiple correction type. So, you select them and see still there is huge number of things, which is getting significant p value changes; so these genes become really important for me, I can say I can create the image store them and look back what these genes are what those functions are.

So, this fashion one can visualize look at differential gene with different kind of statistics with different kind of multiple corrections and look at the data up so in this fashion acuity (()) understanding the data analysis. I would like to see that there are so many parameters and options, we have here for Q C'ing the data and analyzing it further for obtaining some meaningful information. I guess there is no end to doing all of these analysis till one really feels confident about that whole process has performed well. So I think I will finish here, so thank you for giving a very useful demonstration on this software and at least giving a glimpse of the entire workflow, how different type of process are involved. I am sure there is lot more can be explained and lot more can be done here.

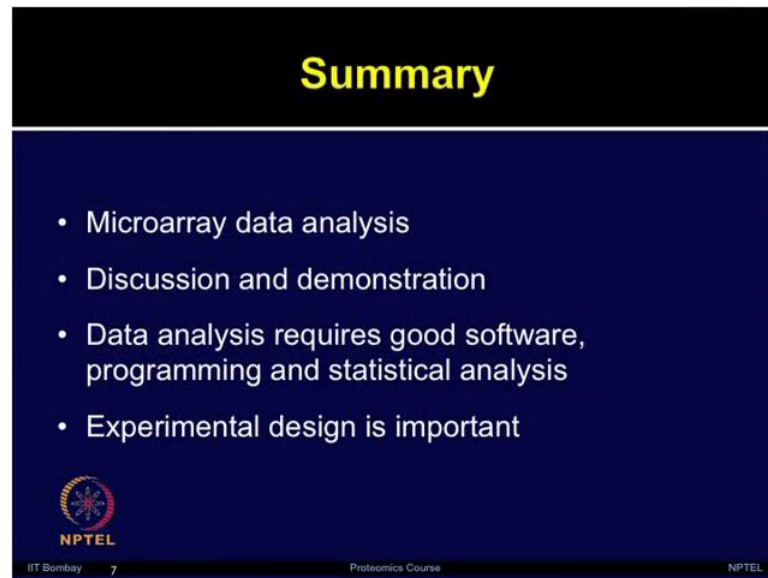
But, just due to the overall time and ah this lecture, I think we should finish here, at least student have got the glimpse of the overall process involved in the analysis and how one can look at a stringent way different types of statistical test one need to perform and then different type of filtering can be done to obtain a different type of floor change. And different ratios one need to obtain and further one need to look at the trends for each of those and which can be color coded and presented in different ways. So, thank you very much Pankaj for being here and giving a very useful demonstration on acuity software for micro data analysis, thank you.

Thank you, Dr Srivastava.

Thank you


Pleasure

(Refer Slide Time: 59:46)



Summary

- Microarray data analysis
- Discussion and demonstration
- Data analysis requires good software, programming and statistical analysis
- Experimental design is important

 NPTEL

IIT Bombay 7 Proteomics Course NPTEL

So, today we had discussion about micro array data analysis you also seen demonstration of acuity software, to learn about various type of parameters involved, in performing these microarray data analysis. You must have got a glimpse of how complex this process is and one really need to put several hours of hard work to obtain any meaningful biological information. The data analysis not only require good software it requires programming and very good statistical analysis.

Experimental design is very important in these microarray based experiments, if you are putting garbage in, so you should expect garbage out. In that case you have to first plan your experiments, your controls all of those things very carefully before starting any microarray experiment. The software tool can help in analysis but, it is more important to have a good understanding of both the biology involved as well as the analytical techniques involved; rather than totally relying on one software on a software you should also think about the biological context, look at the controls.

And then after careful biological as well as analytical analysis you can probably obtain some meaningful information, from these data set. So, microarray experiments generate high throughput data they provide you thousands of features information in a very short time. But, it becomes very challenging to analyze the data, especially when you have to compare various slides from different experiments and you have to normalize, you have to treat them equally, so that you can compare all slides on same platform. So, careful

image processing and data analysis becomes very crucial in microarray based experiments, thank you.