Interactomics: Protein Arrays & Label Free Biosensors
Professor Sanjeeva Srivastava
MOOC NPTEL Course
Indian Institute of Technology Bombay
Module 7
Lecture No 35
MicroArray Data Analysis: Part I

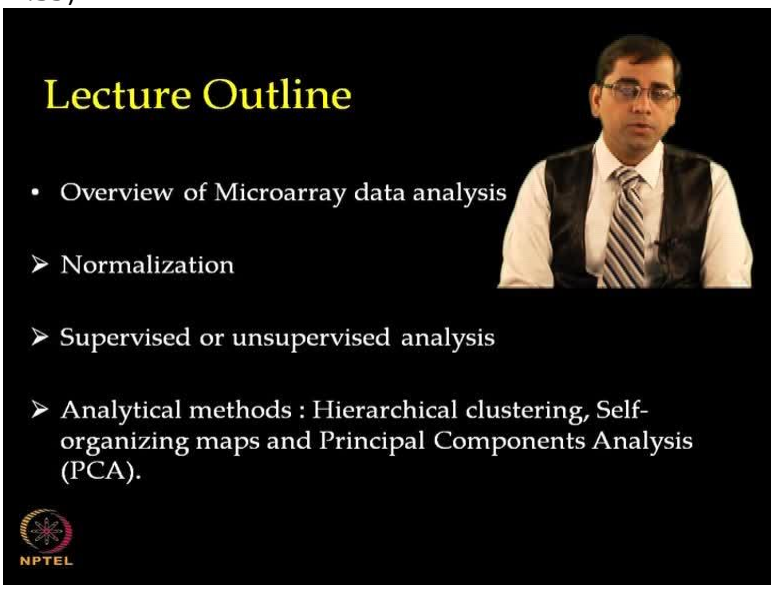
Welcome to mooc interactomics course. In today's lecture we will talk about microarray workflow with focus on data analysis. This is in continuation to our previous lecture where we talked about various strategies involved in performing microarray experiments, various parameters to be taken into consideration for image acquisition and the user interface for image processing.

Microarrays have become integral part of clinical and drug discovery processes. They have been used extensively to find differential gene expression in variety of samples. Microarrays have been used for biomarker discovery finding genes to correlate the disease progression, studying about effects of various drugs and toxins in a field known as toxicogenomics.

Testing the target selectivity, prognostic test, disease subclass determination, in clinical diagnosis and many other applications. Data analysis is crucial to make sense out of the humongous data generated using microarray based experiments.

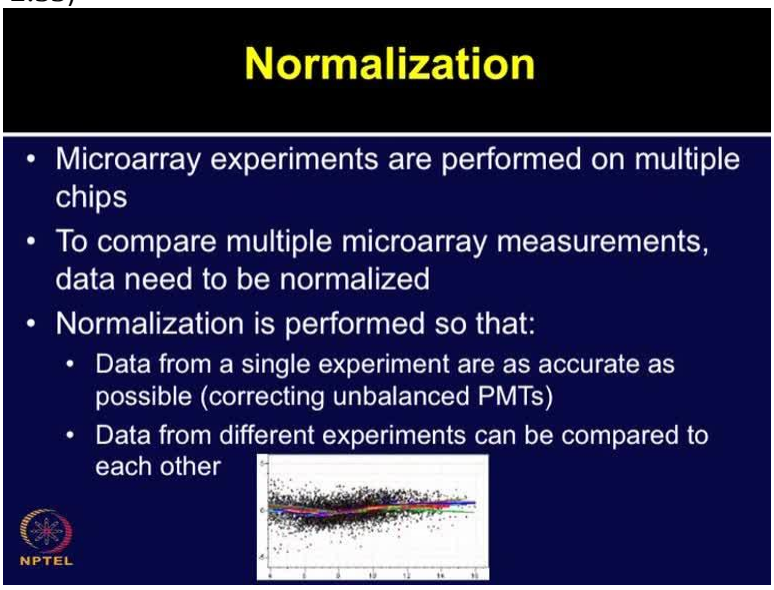
There are many commercial as well as free software available which can be used to analyze microarray data set. However no single package answers all the questions related to a functional genomics or proteomics question.

(Refer Slide Time: 1:59)



In today's lecture we will talk about microarray data analysis to cover various type of concepts such as normalization, supervised or unsupervised analysis, different types of analytical methods such as hierarchical clustering, self-organizing maps and principal component analysis. But before we move on to very advanced modules, let us discuss some basic concepts involved in the data analysis.

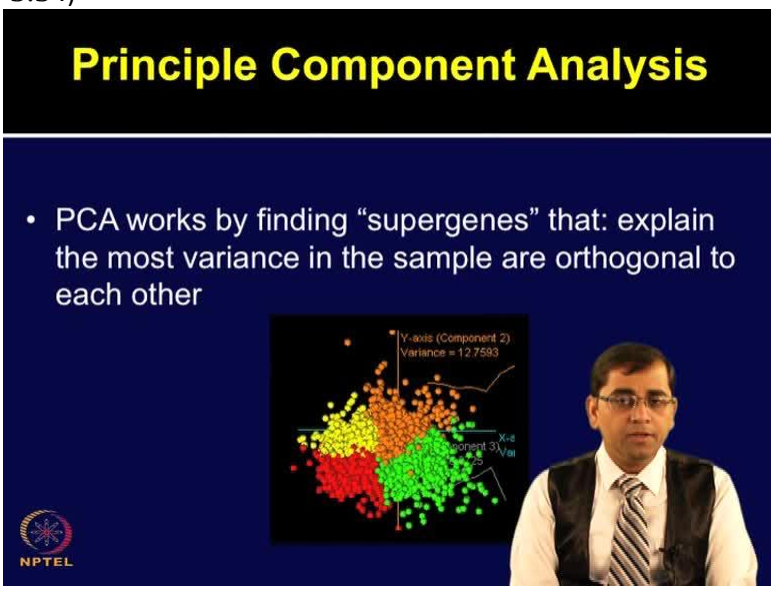
(Refer Slide Time: 2:33)



Normalization. In microarray experiments there are many variables involved. Normalization is performed for reducing bias resulting from variables like printing related issues, dye bias or day today variations. Some commonly used algorithms for normalization are quantile normalization, variance stabilizing normalization, cyclic loess and robust linear model normalization. Normalized data can be used for comparative analysis so, whiles to ensure that comparisons are unbiased.

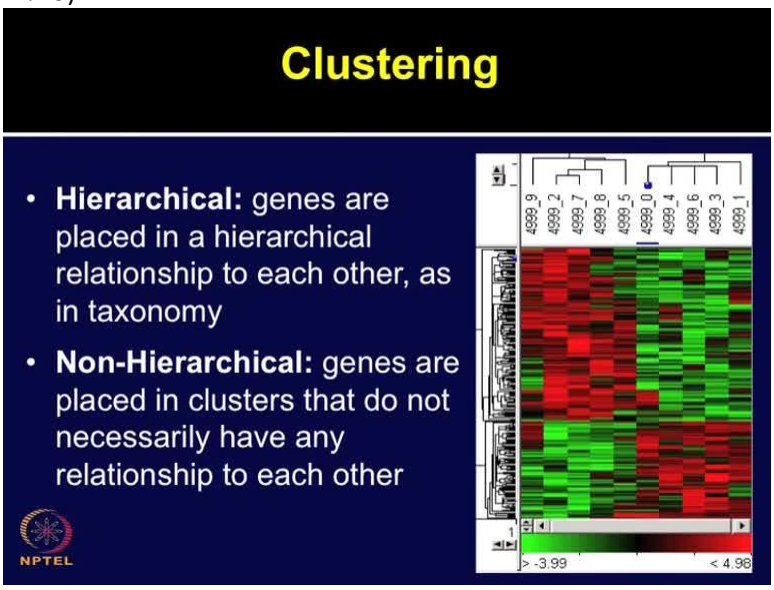
Principal component analysis or PCA. Principle component is the linear combination of optimally weighed optional variables to test whether the protein expression is consistent throughout multiple samples from the same experimental group.

(Refer Slide Time: 3:34)



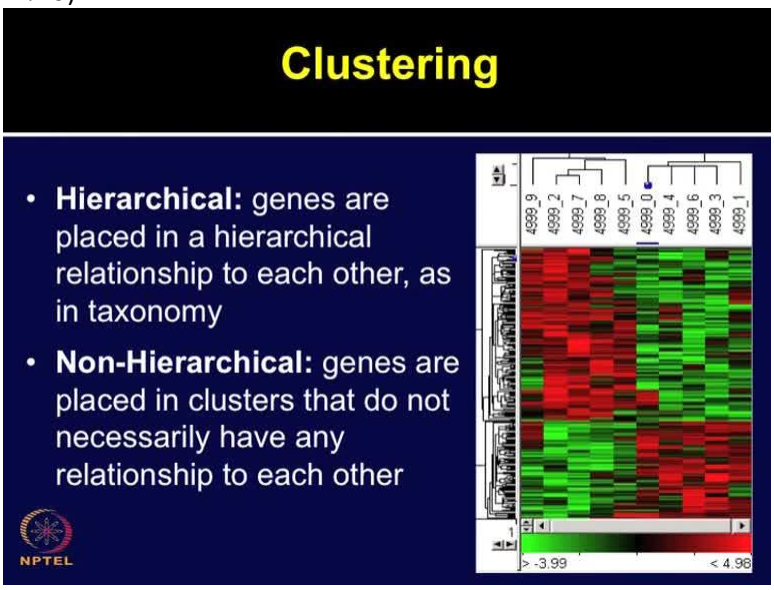
They can be used to identify the protein outliers, miss match spots etc. The PCA works by finding supergenes that explains the most variance in the sample are orthogonal to each other. Clustering. After analyzing the microarray dataset you would like to cluster a data to find out the patterns in which (())(4:01) are segregating. You would like to know whether your controls and treatments can fall into different clusters.

(Refer Slide Time: 4:10)



There are different types of clustering broadly, hierarchal and non-hierarchal clustering. The hieratical clustering involves where genes are placed in a hierarchical relationship to each other as in the taxonomy. The non-hierarchal clustering involves where genes placed in clusters they do not necessary have any relationship to each other.

(Refer Slide Time: 4:40)



Self-organizing maps. In microarray experiment it is important that you perform dye-swap experiments to avoid any effect of si3 or si5 dye labeling so that there is no bias for labeling in control and the treatment groups. To replicate dye-swap, microarrays can be quickly inspected for quality by using a self-organizing map such as the one shown here in this slide.

There are different types of supervised approaches to determine if the data fits in predetermined pattern or unsupervised pattern to characterize the components of a data set without a prior input or knowledge of training set. The data can also be subjected to dimensionality direction techniques like correspondence analysis or CA. CA provides a list of differentially expressed proteins that are statistically significant.

Support vector machine or SVM is another recursive feature elimination model which can be used to generate a list of classify proteins to differentiate between any two (())(5:57). We will try to cover few concepts involved in data analysis and provide you the demonstration of the software in discussion section.

Mr. Pankaj from Spinco biotech will talk about the basic user interface of acuity software from molecular devices which is used for analysis of microarray data. He would also demonstrate the software operation for the data analysis. Prof. Sanjeeva Srivastava: This is pleasure to introduce a Mr. Pankaj Khanna, Manager, Application Support from Spinco biotech private limited. Today Pankaj will talk to us about uhh acuity software uhh which is use for analysis of microarray data. This software is from molecular devices and Spinco is distributor for the same.

In the last lecture we discussed about how to scan the slides uhh microarray slide by using GenePix pro software and once the data were 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 one can actually analyze microarray data I have invited Mr. Pankaj for this discussion. Hello Pankaj. Welcome to this lecture.

Mr. Pankaj Khanna: Thank you Dr. Srivastava.

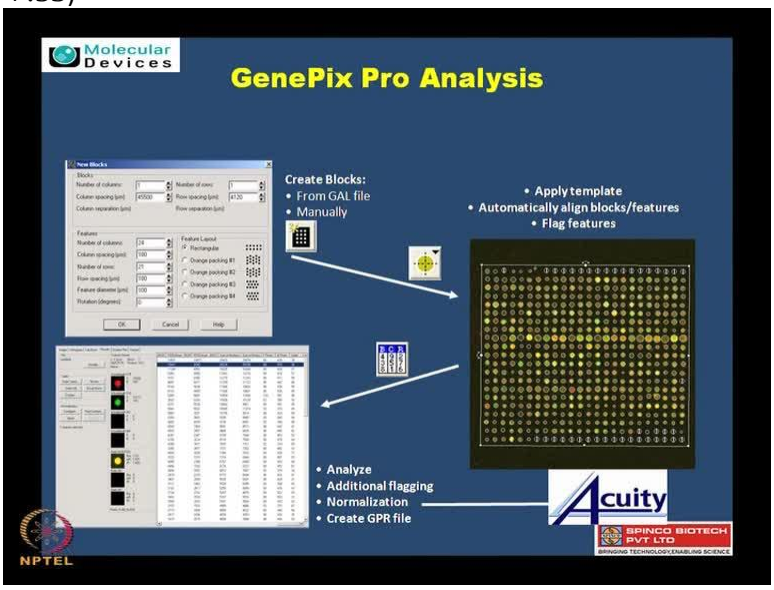
Prof. Sanjeeva Srivastava: So, in the previous lecture uhh we actually uhh discussed about various type of parameters which are used uhh to acquire very good microarray image by using

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.

Mr. Pankaj Khanna: Sure. So, let us quickly go through with the GenePix pro what we have done.

Prof. Sanjeeva Srivastava: Okay.

(Refer Slide Time: 7:53)



Mr. Pankaj Khanna: So, once you are ready with the slide usually people put on and I had the hardware parameters are being selected. Once that is being done the image is being scanned and stored in the TIFF format. So, based on the laser type one two or three or four you get 24 bit maximum image resolution possible and once you are ready with the TIFF image you perform little bit basics of analysis in GenePix pro say for example aligning of different features in the form of GAL file which we have seen.

Prof. Sanjeeva Srivastava: Right.

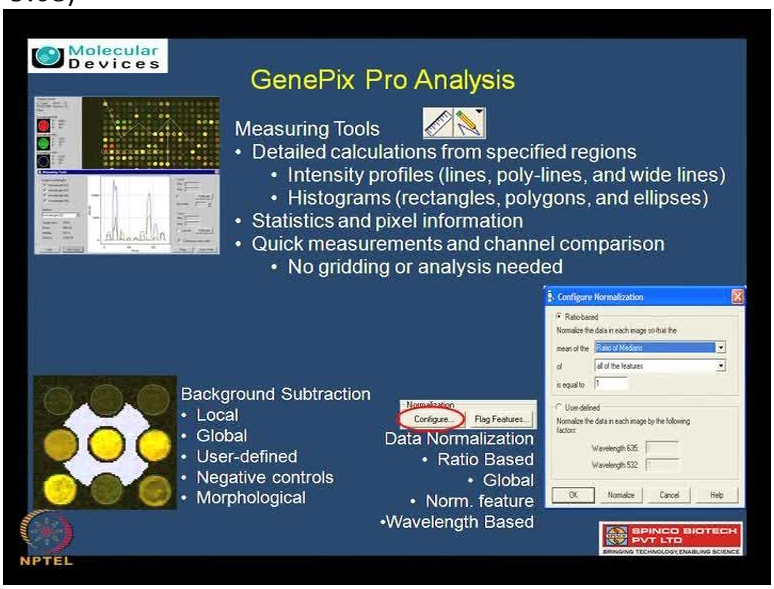
Mr. Pankaj Khanna: Once we have done alignment you go for the results where the background corrections and all other things will be calculated and then given to the results different column

tabs. So, once you are ready with these results these can be saved in the form of GPR file which stands for GenePix result file. That is why it is just briefed as dot GPR files.

Prof. Sanjeeva Srivastava: okay.

Mr. Pankaj Khanna: So, let us go through as we have seen here in the slide that the first one is getting the image, getting the alignment and and once the alignment is done the result tab after doing the result tab (())(9:01) you get the different column details in the form of different stats possible.

(Refer Slide Time: 9:08)



So, once you have the result in place in different formats sometimes you require to have a measuring tool but usually its all commercial and even the academic software gives the GAL file details so, you really uhh no need to do manually but in case if you want to do you can do it and then.

Prof. Sanjeeva Srivastava: It is like I guess there are various parameter one need to look for uhh while performing a good scanning and acquiring the data including the background subtraction and uhh how to normalize the data right? Can you just elaborate on these?

Mr. Pankaj Khanna: 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 subtraction one can perform. So, as you see in the image 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.

So the local background correction is immediately near the feature which is the area which would not have any fluorescent 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 whole in the chip where the spot is not present the different backgrounds levels can be calculated at different specific positions. Now this can be used to calculate for the global background corrections.

As you have elaborated in the last lecture user define once 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 gains only the negative background in the same area of defined other types.

Prof. Sanjeeva Srivastava: I guess we discussed the need for these controls right? How important those are and how I think we can see each here like when we are acquiring these images how each of the positive and negative control features play crucial role in the analysis process. So, after background subtraction I think next important thing will be the normalization right? May be you can just explain on that?

Mr. Pankaj Khanna: Yes. So, important factor is normalization because we do microarray experiments chip to chip basis, experiment to experiment basis, what happens there is owing to the 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 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 532 and 635. So you want to correct for them that both intensity should match the ratio of one so that the difference is contributed owing to the fact of the laser powers and then (fore) forcibility does not come into

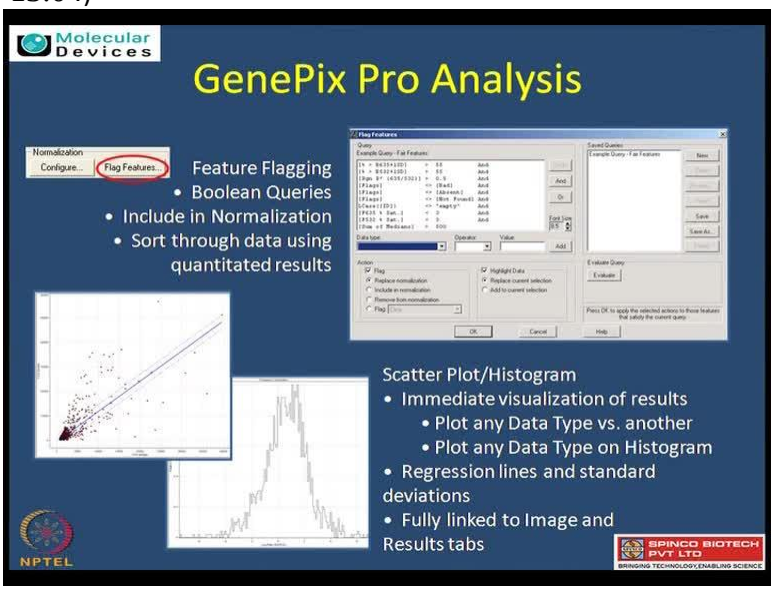
play of biology. So there are different ways of doing a data normalization and the best suggested ones are (())(12:21) ratio based normalization on the mean or the median values which is actually a continuous type which does not change the shape of the data. The meaning is that this is being as collected or collected but nothing is loss in the form.

Prof. Sanjeeva Srivastava: Okay.

Mr. Pankaj Khanna: The other way of normalization is lowest normalization where in you really change the data structure. So there some extreme scan we remove for 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 based correction which we can be done over with.

Prof. Sanjeeva Srivastava: Okay now uhh may be you can just brief us on the analysis aspect uhh of the GenePix pro before we just move on to saving the data for acuity.

(Refer Slide Time: 13:04)



Mr. Pankaj Khanna: Correct. So the very important thing here is the flagging of the spots. Meaning is as we know that few spots 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, these can done by the flagging features. You can also give some Boolean queries basically all the requirements what you want to avoi. So that the spots of the requirement go for the further

analysis. And once you include the normalization or you do not include the normalization you can save the GPR or GPR result file which involves the basic things which is required to correct further images.

So, once you have in hand all these things you can check for the QCs in 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 find I have QCed my data, spots are looking good, all good spots are maintained and we try to avoid some kind of physical variations which happen and this now can be saved as GPR file which can be further do the analysis.

Prof. Sanjeeva Srivastava: I guess one thing is one need to ensure that the data which is going to be further analyzed for any biological significance should clean. There should be a quality control 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.

Mr. Pankaj Khanna: The more better you do QC the more better biological results you do expect. So, very rightly said that yes QC is most important one need to spend little bit of time.

Prof. Sanjeeva Srivastava: Specially when we talk about high throughput analysis which is the case in microarrays here we are talking about 20000, 30000, in fact now there are very very…

Mr. Pankaj Khanna: Yes yes.

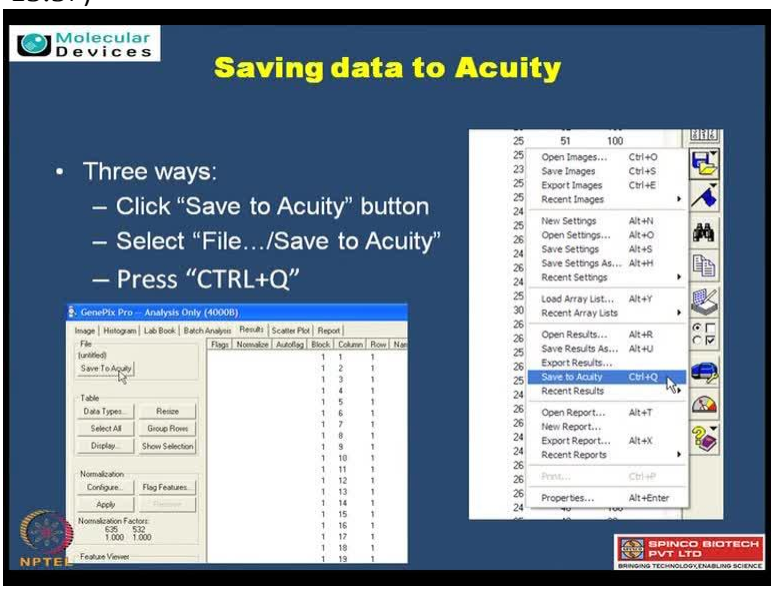
Prof. Sanjeeva Srivastava: …high densities are is available right? You are talking like so many data point generated in that one file. So until unless you are very sure about the overall good quality of the experiment I think otherwise you will be analyzing the very wrong data. So, uhh these high throughput platforms provide us an opportunity to analyze uhh very very large data sets in a very short time.

At the same time 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 aside and move on to repeating whole experiment at once because doing all the corrections and all the things will not help to really do the further analysis until unless you are starting with a very good slide to begin with.

Mr. Pankaj Khanna: Very true, totally agree sir.

Prof. Sanjeeva Srivastava: Ya. So I guess now once we have QCed this uhh slides we are ready for the saving the data for the next analysis right? So, may be you can explain that.

(Refer Slide Time: 15:57)

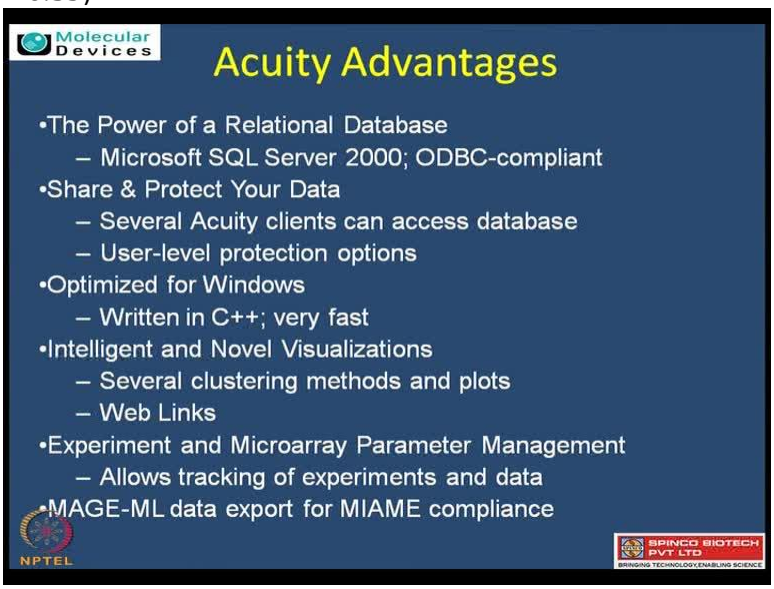


Mr. Pankaj Khanna: so, the basic workflow involves that you do first level analysis that is GenePix pro involving the QCs and then immediately the GenePix pro gives you a direct compatibility with the acuity. There is a button on the side which allows to say that just save the data to acuity and immediately 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 stand-alone. So there are ways to import the data in the text format or the different format which it understands. So, this is how the acuity can be used for the further analysis now.

Prof. Sanjeeva Srivastava: Okay so I think we are hearing about acuity now so maybe we should a talked little bit more what acuity software can do, what are its major features. So maybe you can explain a just few uhh points about acuity before we move on to the knowing the details for the acuity software and what we can do with analysis.

(Refer Slide Time: 16:59)



Mr. Pankaj Khanna: Sure. To begin with acuity is a bioinformatics software. So it gives you a power that whatever basic analysis you have done though 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 data based understanding. So we give MSSQL 2000 with this which allows you to save the data in the form of servers.

Prof. Sanjeeva Srivastava:Okay

Mr. Pankaj Khanna: So, this gives the power that this can be your data warehouse meaning all the important attached files, save any file like TIFF image, JEPG image, GPS that is setting file, all can be stored with the result files which allows one to again look back whenever you require to. And so also it is optimized for the windows it is written 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 normal visualizations we do have like different kind of clusterings possible, we have scattering available for you, scattering graphs coming in. So this gives one an opportunity to analyze the data visually to quickly understand what is happening in the biological. Experiment and the microarray parameter management. So, many scientist want to gives 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 input and manage your all parameters within the experiment so, that you group them and do the analysis accordingly.

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

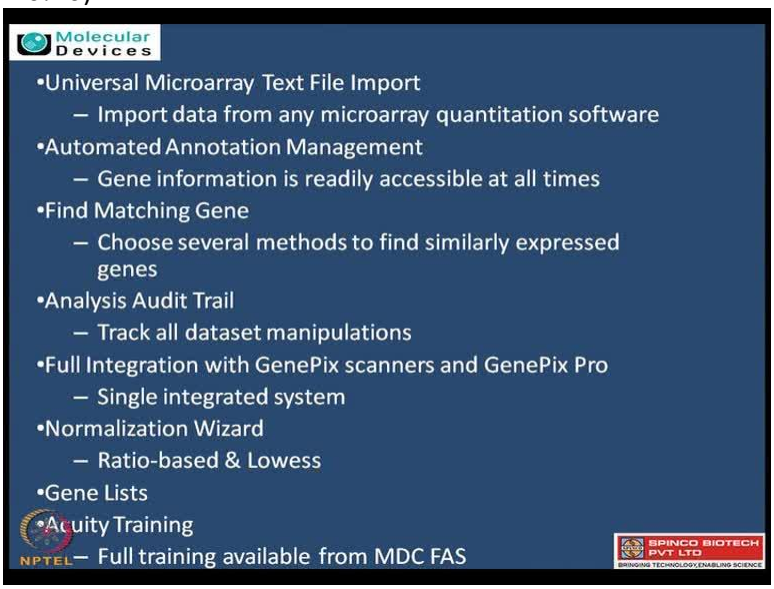
Prof. Sanjeeva Srivastava: So, I guess last two point which you mentioned uhh one is the tracking the data base on the experiments. I think that is very important uhh that is also like depending upon the need of the experiment one need to in fact track and uhh make software learn your experiment so that one can actually apply the same knowledge for the various slides throughout to track that data set.

Now second point which you mentioned about MIAME compliance I think that is very important because one need to do all the quality control checks and overall data analysis with the very uniformed guidelines provided. So, one has to adhere to those quality control checks.

Mr. Pankaj Khanna: True so, another very 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 analyze. So we have not restrictedt it to only GenePix. It can also take other format even in the form of text format where in you need to give an information what each column means is and then again you can perform the same statistics.

So, there is an automation management also possible with this. So, you have number slides coming in every time.

(Refer Slide Time: 20:20)



So, you 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 an 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 tissues.

So, even that can be handled very effectively here. Analysis audit trails the meaning is that you 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.

Prof. Sanjeeva Srivastava: Sure.

Mr. Pankaj Khanna: So, the 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 level statistical analysis.

Prof. Sanjeeva Srivastava: But this software is also compatible with other scanners and other platforms?

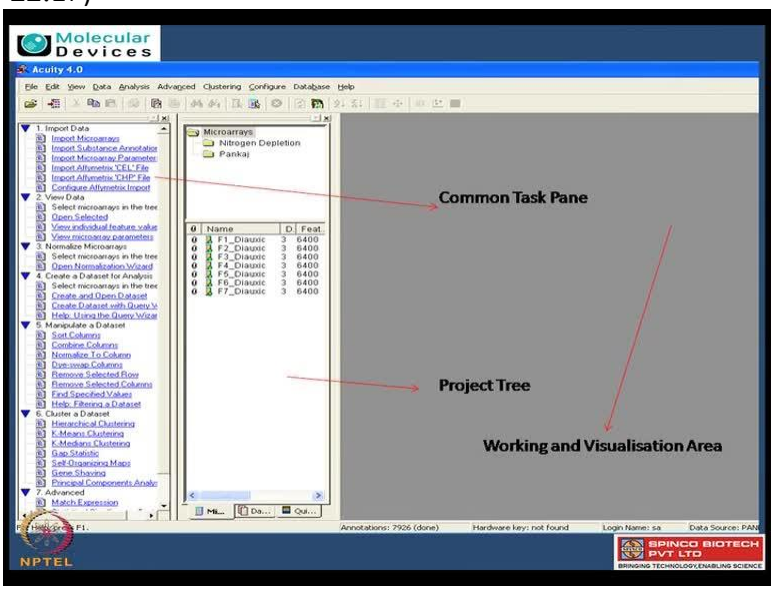
Mr. Pankaj Khanna: Yes as it can take up any text file so, basically whatever if I will understand that okay it is coming from 532 or 635 tell that it is coming this wavelength and still you can do the statistics.

Prof. Sanjeeva Srivastava: Sure. So, regardless of what platform is being used is just the wavelength and text file which matters here.

Mr. Pankaj Khanna: True true. And with give training at the level of difference stages also so that one can become friendly with the software.

Prof. Sanjeeva Srivastava: okay so, it will be useful if you can uhh demonstrate us about the acuity software uhh so that one can actually learn that how uhh the data obtained can be transformed into the meaningful biological information and also the statistical significance of that data uhh 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 before we switch to the real demonstration.

(Refer Slide Time: 22:17)



Mr. Pankaj Khanna: Sure. So, what you are looking at as GUI interface of an acuity which is first divided on top in the form of any typical file based dropped down list which has various functions and then towards your extreme left you will able to a 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 the biological information.

The idea is it start with the import of the data and end with the statistics and visualization how one want to look. So, in this passion common task pane actually a very good tool for any new beginners as well as for the mature or the advance users to understand what one can do with the microarrays.

Prof. Sanjeeva Srivastava : So, I think it just guides you the stepwise like how you can uhh walk through the entire process.

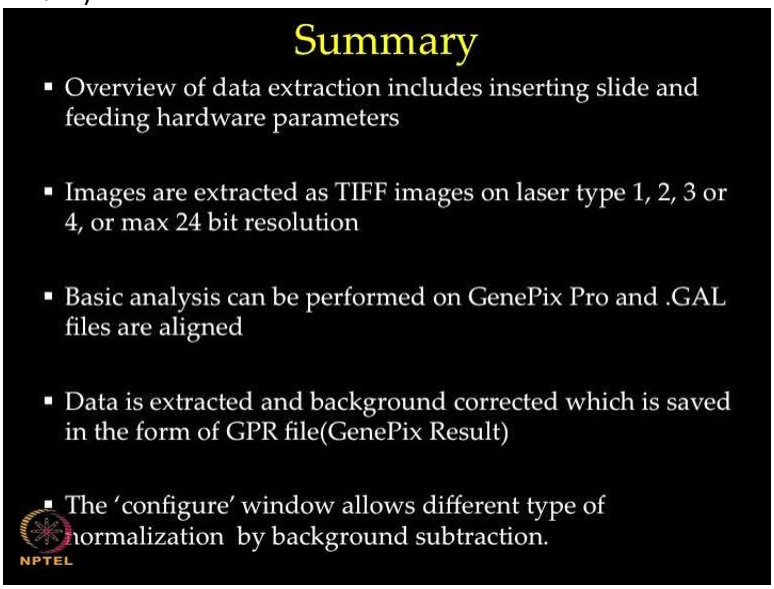
Mr. Pankaj Khanna: True. So, It just gives you right from the input to the analysis step wise that what all you can do and what you want to do in. And towards the middle what you see is a microarray route directory which houses all your data in a different formats. So, this is a warehouse point on the top in the folder based arrays and on bottom it shows individually the each one slide by slide.

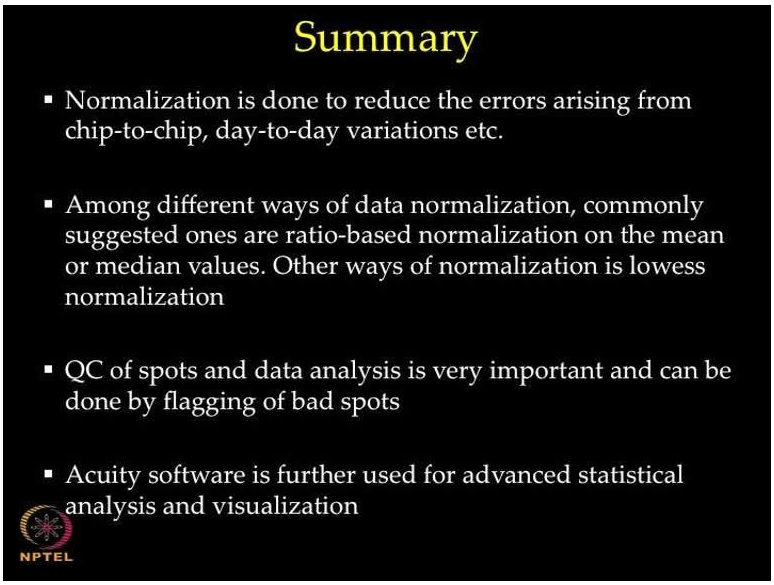
Prof. Sanjeeva Srivastava: Right.

Mr. Pankaj Khanna: And towards your extreme left you are saving 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 advancement. So, this is the basic user interface of acuity.

Prof. Sanjeeva Srivastava: From this lecture you got in inside into the detail user interface of acuity software. You have got the demonstration of various features and controls that the software allows you to control. Well there are many software available for the data analysis but the basic workflow remains the same. We will continue demonstration of how real data is actually analyzed further using this software. Thank you.

(Refer Slide Time: 24:14)





References

- Anna Pursiheimo, L.T.A. (2015). Statistical Analysis of Protein Microarray Data: A Case Study in Type 1 Diabetes Research. J. Proteomics Bioinform. 8.
- Bengtsson, A., and Bengtsson, H. (2006). Microarray image analysis: background estimation using quantile and morphological filters. BMC Bioinformatics 7, 96.
- Chandra, H., Reddy, P.J., and Srivastava, S. (2011). Protein microarrays and novel detection platforms. Expert Rev. Proteomics 8, 61–79.

References

- DeLuca, D.S., Marina, O., Ray, S., Zhang, G.L., Wu, C.J., and Brusic, V. (2011). Data processing and analysis for protein microarrays. Methods Mol. Biol. Clifton NJ 723, 337–347.
- Gupta, S., Manubhai, K.P., Kulkarni, V., and Srivastava, S. (2016). An overview of innovations and industrial solutions in Protein Microarray Technology. Proteomics 16, 1297–1308.
- Rydén, P., Andersson, H., Landfors, M., Näslund, L., Hartmanová, B., Noppa, L., and Sjöstedt, A. (2006). Evaluation of microarray data normalization procedures using spike-in experiments. BMC Bioinformatics 7, 300.

References

- Syed, P., Gupta, S., Choudhary, S., Pandala, N.G., Atak, A., Richharia, A., K P, M., Zhu, H., Epari, S., Noronha, S.B., et al. (2015). Autoantibody Profiling of Glioma Serum Samples to Identify Biomarkers Using Human Proteome Arrays. Sci. Rep. 5, 13895.

- Wellhausen, R., Seitz, H., Wellhausen, R., and Seitz, H. (2012). Facing Current Quantification Challenges in Protein Microarrays, Facing Current Quantification Challenges in Protein Microarrays. BioMed Res. Int. BioMed Res. Int. 2012, 2012, e831347.