LECTURE-32

MICROARRAY WORK-FLOW: DATA ANALYSIS

TRANSCRIPT

Welcome to the proteomics course. In today's lecture we will talk about microarray work flow, especially focus on data analysis. This is in continuation to our previous lectures where we talked about different strategies involved in performing microarray experiments, how to acquire good images and now today we would like to 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, studying about effects of various drugs and toxins in a field known as toxicogenomics, testing the target selectivity, prognostics test, disease subclass 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 proteomic-based question. So in today's lecture, we will talk about microarray data analysis. We will have a discussion on Microarray data analysis to cover various types of concepts such as Normalization, supervised or unsupervised analysis, different types of analytical methods such as Hierarchical clustering, self-organizing maps and principal-components analysis. To elaborate and clarify the analysis, I will have a discussion with Mr. Pankaj from Spinco Biotechnology. He will be representing molecular devices and we will have a discussion on Acuity software to give you a demonstration on the software operation for data analysis.

We will discuss about various basic concepts including Normalization. In the microarray experiments people use different types of chips in different experiments to compare multiple microarray measurements, data needs to be normalized. The normalization is performed so that the data from the single experiment are as accurate as possible, also correcting further the unbalanced PMTs. 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.

Principle component analysis. Principle component is the linear combination of optimally weighted optional variables to test whether the protein expression is consistent throughout multiple samples from same experimental group. Are there protein outliers, spots mismatched or not to proteins, to realize all of this type of variations, principal components analysis is performed. 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 like to cluster data to find out the patterns of the type of question, which was asked? Your control or treatment fall into different clusters. There are different types of clustering, broadly hierarchical and non-hierarchical. The hierarchical clustering involves where genes are placed in a hierarchical relationship to each other, as in the taxonomy. The non-hierarchical clustering involves where genes placed in clusters that do not necessarily have any relationship to each other.

Self-organizing Map. In microarray experiment it is important that you perform dye-swap experiments to avoid any effects of Cy3 and Cy5 labeling so that there is no bias for labeling in the control and treatment groups. To replicate dye swap microarrays can be quickly inspected for quality by using a self-organizing map such as one shown here in this slide.

Then there are different type of supervised approaches to determine chains that fit a predetermined pattern or unsupervised patterns to characterize the components of a

data set without a prior input or knowledge of training signal. We will try to cover various concepts involved in data analysis and provide you the demonstration of the software while discussion with Mr. Pankaj from Spinco. So let us have a discussion and then we will conclude our today's lecture.

Prof. Srivastava: This is pleasure to introduce Mr. Pankaj Khanna, Manager-Application Support from Spinco Biotech Pvt Ltd. Today Pankaj will talk to us about acuity software which is used for analysis of microarray data. The software is from 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 with the data was acquired; now the next step or next challenge is how to obtain some meaningful biological information from that data. 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.

Speaker A: Prof. Sanjeeva Srivastava; Speaker B: Mr. Pankaj Khanna

A: Hello Pankaj, welcome to this lecture.

B: Thankyou, Dr. Srivastava.

A: So, in the previous lecture, we discussed about various types of parameters which are used to acquire a good microarray image using pix Pro software. Can you give us an overview of that whole process how, in a nutshell, so that we are informed about the same and then we are ready for the analysis with the Acuity software?

B: Sure, so let us quickly go through GenePix Pro, what we have done.

A: OK

B: Once you are ready with the slide, insert the slide and hardware parameters are selected. Once that is being done and image is being scanned as stored in the TIFF format so based on the laser type 1, 2, 3 or 4, you get 24 bit maximum image resolution

possible. Once you are ready with the TIFF image, you can perform basic analysis in GenePix Pro. See, for example, aligning of a different feature in the form of .GAL files which we are seeing. Once you have done alignment, you go for the results where 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's why it is just briefed as .GPR file. So let's go through, as we are seeing here in this slide, that the first one is getting the image, getting the alignment done. Once the alignment is done, the result tab after doing the result tab being hit, you get the different column details in the form of different stats possible. 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 give the GAL file details so you really don't need to do manually. But Incase if you want to do, you can do it.

A: yes. I guess there are various parameters that one need to look for while performing good scanning and acquiring data including the background subtraction and how to normalize the data, right? Can you just elaborate on these?

B: So in the result tab, immediately what you see is a window which gives you 'configure', which can configure different type of normalization. So there are different kinds 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 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 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 whole in the chip where the spot is not present, the different background levels can be calculated different specific positions. Now this can be used to calculate for the global background corrections. As we already did in our 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.

A: I guess we discussed the need for these controls, right? How important those are and now I think we can see it clearly where acquiring these images how each of the positive and negative control features play a crucial role in the analysis process. So after the background subtraction process, the next important thing will be the normalization process, right? So maybe you can just explain on that.

B: 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 have been used to do the experiments, there are different ways where 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 chip as well as within the chip. Within the chip would 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 difference is contributed owing to the fact of the laser powers and the flurophore stability doesn't come into play of biology. So in different ways of doing data normalization and the best suggested ones are ratio-based normalization on the mean or median values which is actually continuous type which doesn't change the shape of the data. The meaning is that this is being escalated or corrected; but nothing is lost in the form. The other way of normalization is lowest normalization where in you really change the data structure. So there are some extreme variance can be removed, 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 can be done over with.

A: okay, so maybe you can just brief us on the analysis aspect of the genepix pro before we just move onto saving the data for acuity.

B: Correct. So the very important thing here is the flagging of the spots, meaning is as we know that a few spots could be controlled so you don't 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 all the requirements what you want to avoid, so that spots of the requirement go for the further analysis and once you include the normalizations or you don't include the normalizations, you can save the GPR or GPR result file which involves the basic things which is required to correct for the images. So once you have in hand all these things you can check for the QCs in the form of scatter plots, histograms and show also immediate visualization in the form of data vs different intensity plots so which give one the availability that fine, I have QCed my data, spots are looking good, all happen good spots are maintained and we try to avoid some kind of physical variations which and this now can be saved as .GPR file which can be further do the analysis.

A: there is one thing, one needs to ensure that the data which needs to be further analyzed for any biological significance, should be cleaned. It should be quality control checked and all the control parameters are in place and once we have verified all those things at this stage, then only that data is ready for next level of analysis.

B: the more better you do QC, the more better biological results you do expect. Yes, QC is most important; one needs to spend little bit of time on that.

A: especially when you talk about highthrough put analysis which in the case of microarrays and here you are talking about 20,000 - 30,000 intensities are available, right. You are talking of so many data points in that one file so until and unless, you are very sure about the overall good quality of the experiment, I think otherwise you'll be in a way analyzing the wrong data. So these high throughput platforms provide us an opportunity to analyze a really large data set in a short time. At the same time, what is very important is that one need to ensure that the data quality is good because if it is not good, it is better to just leave the chip aside and move on to repeating the whole experiment at once because doing the correction 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 begin with.

B: very true. Totally agreed.

A: So I guess now once we have QC'ed this slide, we are ready for saving the data for the next analysis, right? So maybe you can explain that.

B: 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 Genepix directly and not only this acuity can actually work as a standalone 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.

A: okay, so I think we are 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 on the knowing details of acuity software and what we can do with the analysis.

B: Sure. To begin with acuity is bioinformatics software so it gives you a power that whatever basic analysis you have done through GPR can now be further taken for the analysis.

Acuity advantages. So lets us quickly look at few of the acuity advantages. It is actually client server relational database understanding so we give MS SQL 2000 with this which that it allows you to save the data in the form of the server so this gives power that this can be a data warehouse meaning all the important attach files so if any file like .TIFF image, .JPEG image, GPS that is setting file, all can be stored with the result file which allows one to again look back whenever you to require. It is also optimized for windows and it is written in C++ which actually gives a very fast power for it so it can

work very fast and gives the results saving your time and allowing one to look at more different statistical possibilities. Intelligent in the form of novel visualizations. We do have like different kinds of clustering 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 at the biological levels. Experiment and microarray parameter management. So many scientists want to give a different parameter and allow one software to sort or understand the biology based on that which is we call it as parameter files. Actually this is being MDT file for us which you can import or 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 discuss different QC formats, so this particular MAGE-ML is based on the MIAME requirements which says what all is required and how to do a microarray experiment. This has a direct export capability of that so this gives one a very good opportunity not only for the data to the export at different levels.

A: so I guess, last two points which you mentioned that, one is the tracking the database on the experiment, right, I think that's very important. That's also like depending upon the need of the experiment, one need to in fact track and make software learn your experiment so that one can actually apply the same knowledge for the various slides throughout to track that data set and the second point which you mention about MIAME compliance, I think that's very important because one needs to do all the quality control checks and all the data analysis with the uniform guidelines provided. So one has to adhere to the quality control checks.

B: 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 analyzed so we are not restricted it to only genepix, it can also take other formats even in the form of text format where in you need to give an information what each column means and then again you can perform the same statistics so there is an automation management also possible with this so you have a number of slides coming in every time. 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 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 level of tissues. So, even that can be handled very effectively here. Analysis audit trail, 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 at what one has done. Sharing becomes very important in that. So full integration with GenePix scanners and GenePix Pro which allows the user of genepix Pro to immediately store the data and start doing the tertiary level statistical analysis.

A: But is this software also compatible with other scanners and other platforms.

B: Yes. It can take up any text file so basically whatever, if it is coming from 532 or 635, tell that it is coming from this wavelength and still you can do the statistics.

A: Sure. Regardless of whatever the platform is being used, it is the text file and the wavelength which matters here.

B: True. True. And we'd give training at the level of different stages also so that one can become friendly with the software.

A: so I think it will be useful if you can demonstrate us about the acuity software so that one can actually learn how the data obtained can be transformed into the meaningful biological information and also the statistical significance of that data. But maybe you can just first share the software interface so that we are familiar with windows and keys over there before we switch to the real demonstration.

B: Sure. So what you are looking at as a GOI interface of an acuity which is first divided on top in the form of any file-based typical 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 the biological information. The idea is that it starts with the

import of the data and end with the statistics and visualization how one want to look. So in this common task pane, actually very good tool for any new beginner as well as for the mature or advanced users to understand what one can do with the microarrays.

A: so I think this guides you step wise like how you can walk through the entire process.

B: true. So it just gives you right from import to analysis step wise process 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 based array and on the bottom it shows individually the each one slide by slide. And towards your extreme left you are seeing an area which is a working and visualization area where you do or output different task you have done towards the common task pane or towards the advanced ones. So this is a basic user interface of acuity.

A: So I guess now we can look at the real software and the data demonstration so that we are very familiar with the step-wise analysis.

B: sure.

A: So, Pankaj, in the last lecture when we talked about genepix pro then you showed one yeast slide, how to scan that yeast slide by using genepix pro software. I guess 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 at various time points from 0 to 20.5 hours in yeast. And I think it will be interesting to see what type of trends we observe in various time points with the glucose utilization here. So please use that slide and we can look the demo here.

B: Sure, Let's walk through the data. So as we describe, we see 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 understand 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 in 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 kinds of time points collected and these are individual slides where one has run with both Cy5 and Cy3. So this is ratio based image which we are going to see in and we have also got yeast another file which has come from the text to show you even that can also be imported.

A: so first you have to usually scan each of those images, make a folder where you group all of this and whole data set for combined analysis.

B- Right. 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 just goes on looking for the GPR file and this now can be imported and ready for your analysis. And once the data is in, this will be displayed in the down in the form of folder which you have kept for the analysis. So here, the folders as we discussed that this can be used for the data warehouse as such. So this little pit kind of image here tells you that there is some files are attached. You can view them, you can see them but what all somebody has attached so if I see view all attachments, I will be able to see what one has attached to it the important files coming from the GPS in the form show also a GAL file and image. So it gave me a complete opportunity to look at. It is good to emphasize here, if you have a .GIF file also this particular one can also behave as a partial visualization tool as in case of genepix. In case you want to look how the spot has behaved. In this fashion, it can be any file can be attached to it.

A: So in GAL file, we also know the gene has a list so at any time point if you identify any spot which is looking interesting, you know what the gene is.

B: right, aligning with the GAL map. Correct and another..this is good point that you raised because a very important next step is a substance annotation. This substance here I mean is a each spot, which could be a feature which again could be a RNA or a gene or a protein. So this is why it is called a substance annotation and many few people extensively trace them so here I'll show you in the form of tab- data in the annotation, one can look at what all different information can be seen for each, tabwise,

so you have that same annotation tab being given for the substance ID and then because it is each database of candidates being attached here, component, different functions even on the level of enzyme commission numbers so there are different annotations which people try to get in which also you can import in the form of text delimited renamed to .SDT file which allows one to take all the annotation information possible. So, another very important thing the parameter file so as you have said already that it is very important for one scientist to look at all parameters to group accordingly this can be made in the text delimited form and can be renamed to .MDT file and this again can be imported to look at all the parameter are visible in the form in the down window here to look at. In few seconds more you will understand what each window means but as it is case you can just switch over different types and you can just go to parameter file and look at what all details I have obtained.

A: so maybe you can just briefly explain each of the tabs, so that students are clear about, what is that they can infer from each of these windows.

B: sure. So here in the working area which we have defined again splitted into two 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 array, say for example, I only open one array it shows me only one array and tells me what am looking at. I am looking at log ratio data, in a similar fashion, I can look at any other one because it is a .GPR import, whatever data you have to watch for. You can look at the background signal individual intensities but majorly used in the log-ratios for specially expression analysis but if we use for protein or single wavelength base, we can look at only wavelength base one. So you can always control what you are watching.

Apart from this, the other tab include annotation which gives you that what all different one is tracing at the level of annotation base in the form of different databases, information on genes, how protein is behaving or even the localization so where it is being localized, all that can be traced. Other than to that, it also gives little bit of other details in the form of statistics, warehouses and few of the auto scripting capability

which advanced users sometimes want to use but this one particular statistic one allow to see what all you want to see, in detail. In bottom essentially what you see is how the data is looked at, many a times, let's see quickly for example, I want to go in my 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 it, so because it is only one, it is showing you one dot point and if I keep on including more and more arrays, the databot starts increasing and immediately one feature profiling how it has performed can be immediately seen.

A: can we select couple of arrays for alignment..

B: right, the way to select here is, just hold the shift button and if u wants to select only one, 'one more' or if u select all to the last, all can get selected. Then right click and click on open selected, what it does is, it allows you to open all the images here. So, its given you whatever calculated one you want to display and if you click on that each profile now can be seen here. Refresh button if u 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 different profiles how it has behaved. So in this fashion, immediately, I know my parameter file that each one is what it is and I see that okay this is all normal average, in one of the case, it went up. so similarly, different features can be individually analyzed and checked at how the behavior is.

A: so you can actually look at the trend for the same gene during the whole time course analysis.

B: so you can also trace little bit of working on the data before going into this, let me explain an important factor here, what does each images mean. Actually if you careful look, may not be very clear that this is little purplish in color and the other down ones are little reddish in color. The purple one means that the data is not normalized. The red color means that the data is normalized and the little dot green color what you are seeing tells me that I have .JPEG image which I can see down here so it allows me a connectivity of what is happening just by visualization. If u want, say that the data is not

normalized, it's a easy process of doing it so you once imported the data on that u can just click right click look at the normalization wizard. This normalization wizard allows one to choose different kinds of normalization process which we have discussed earlier that could be a ratio based or log normalization based. 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 points has to be normalized same way. So you cannot cross different in the form of different normalization and compare them. So you are looking at little bit of a different balance here.

A: Because until and unless you normalize in the same scheme you cannot compare those across slides because you are going to compare different time points here.

B: correct. As other ones are being analyzed in the form of ratio based I am going to select the one you have an opportunity to select different types I am going to select the ratio of medians which is the most preferred and just the next button which all the flagging that is set if they are flagged, please don't allow for the calculation. So right, and then you just click next and it allows that okay its available for me. Its done and I can say finish. So fine, I have finished my normalization. So if you carefully look back the spot there the purplish will change to red which allows one to understand that yes, all my images are being kind of normalized in a similar fashion.

A: So I guess we are dealing here with lot of data set that's why its taking some time for processing the whole thing.

B: so it describes how many flags were there; 6400 at a time were analyzed for normalization. See before and after doing the normalization how the data looks. So, you can look at it at a different way so what we done is we have corrected at the level of background and I am trying to display across how these and X and Y and is being scattered together before and after normalization. Once you are satisfied with this, now when I look back the same colour has changed from reddish to purple which tells me okay we have, right, all is normalized. Say if I 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 if we look back what kind of this one being done for using the normalization process so I can crosscheck once again how one 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 files ready for you and there are few ways with which you can look at the data.

A: so I think before moving forward its important to show that normalization was done properly and one needs to look at each slide carefully

B: so as discussed it is very essential to have same normalization for all and it is not a thumb rule that which one is more preferable. One can choose anything but make sure all your different slides are being handled in a similar way and you can do different ways and get the data and do analyze the different normalization processes also and so one has an opportunity to even correct that so because you have raised a point say for example, I want to remove this normalization and put some other normalization I can just click here remove normalization and it removes normalization and then it allows you to go back to the raw data and then you can renormalize. Renormalize in a different sense, maybe you want to try lowest normalization and check back all in that format and also select multiple in a similar way and you know in one shot itself you can do normalization in a same color bar. So this is what I prefer and usually you don't mess around with different kind of normalization. Either select all or remove all because I have imported one to show you how the process is being done here and then immediately one would like to see how my data looks like. The one way to look is the number which is cumbersome, other way people like is color. If you carefully observe the coloring scheme is going here, red, black and green.

A: yes, so maybe you can tell the exact conventional things for the microarray? People always represent these colors so what each of this color codes mean.

B: so the black color is towards Y. the meaning is so when I am looking at the data, you expect that when green and red channels both are giving the same color same intensities, it becomes blackish in color. If they are up-regulated, people put them

towards the red color and if it down-regulated minus sign will be given and that will 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 of ratio needs to check basically it is case over control what people report for, right, so in this fashion conventionally you can see the colors but here they are many ones. There are many which we have open now, 6. So we want to, in nutshell what is happening so acuity allows you to do it by seeing you can do an autofit color so quickly the numbers have gone and colors are there to tell you how each particular substance or gene has behaved across your sample. You can look this is being attached I have just split tables and look back at annotations also so I can have just split table available. Put an annotation file here so that I keep looking at what I am interested in. so there Is enough opportunity for you to play around and how you want to look and customize your view.

A: so I think these type of heat maps right away gives you a feel about what type of genes across each time points has shown the modulation and variation in the expression profile, looking at the color itself like for the first one, I can really say that it is going down as we are moving across toward the 20^{th} hour.

B: True.

A: So I think by looking at the type of data one can visually actually get the feel about expression changes across the various time points.

B: Very true. So this gives you an immediate visualization tool to understand what is happening and get a rough idea and this is, mind you, just the neat raw data. You have now performed just the normalization and we are seeing how they are behaving. So it gives you a rough profile, okay, 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 autofit my data, I can select appropriate one autofit all data so it says it just fitted based on that. so it again shows you the number back so we have got the data imported now we have done the normalization, we are trying to see how they have behaved. Very important thing we sometime people like is in the form of like able to move the data sorting up and

down but before that acuity tells you that first you make a data set. The meaning of data set is this is just looking at the raw data and I want to extract the data and allow to keep in a data set here towards down so there you are available with all kind of different things. So now, there are two ways of doing things in a data set, so one thing is take all the features and sometimes people say I want to have my criteria defined such that my visualization make more sense to me.

A: One can actually be stringent at this stage itself and say I want only very little biological significant ones to define P values and then just sort the data based on that.

B: So people can do okay which are changes up and down with a range of so and so two fold upregulated and two fold downregulated using differential expression data is been logged. The meaning of log to the base two is essentially log to the base two value one is equal to two fold change. Talking of log to the base two means you are talking about fourfold change really become significant you can filter based on various parameter and generate the data set. So, essentially you are reducing the numbers so you make more sense in the form of visualizations otherwise also you can also take all the data and you can do it. Lets quickly see how we can do that particular job. So I can select my data again holding the shift and I can click, it will tell me what you want to do, it says you can have different kind of opportunities but you can create data sets from selections so this allows whatever I have selected, do create a data set from that. Once I click that, it says where you want to keep in the folder so beforehand 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 we go, from the seven microarrays we have got, we have all the data available. This is the one to get all complete data you can also have other ways, as you suggested, P values importance coming or log values importance coming in define a criteria of doing that. The way to do that again look at the common task pane we have done the three steps and now after doing normalization, I can click create and open data sets.

A: actually it is a good idea you just refresh again, sort of the task step wise, so the first one import data.

B: Correct, so if you see first was import data then importing the substance annotation file which is the .SDT file gives all the annotations which you have made the tab delimited and name it to .SDT then the third one is microarray parameters where it is being stored in the form of tab delimited named to .MDT file which allows one to trace all the details. The other three you are seeing is based on the Affymetrix which is actually a .CEL file so there image is actually a .DAT file which they convert to industry level .CEL file that can also be handled but just it doesn't RMA analysis where they move little bit further for a different analysis but it does give opportunity even to look at affymetrix outcome data and then you have a viewing data and then as we have seen you can autofit the data, you can look at the color being coded, you can view the data in a different format, you can individually look at them or look at the numbers, 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 profile and you will be able to see how this has behaved across. So apart from that, next step involves the normalization wizard so I need to make sure that all my chips are being 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 wavelength and different normalization per se and further analysis down there. So once, we have finished the normalization method, we can look at a query which is creating a data sets. As we described one simple way is right at the data, select the chips and create the data or you can come here and say create data sets from the open data sets. So how to go for this one. If I click on create and open data sets, right, so you can see similar way it has popped up the window where it shows in the data set the mother folder then the childs of it what all can be generated now.

I can create a dataset here in this form which we can do directly there. Other important factor is creating the dataset from the Query the set the meaning is that you can define different criteria right 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 microarray experiment I want to do type and based on the parameters as such I can say that okay I want to import the data in the form of only one particular parameter. This is little tricky because if you have same name, multiple times, the data will be imported twice. 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 this particular data from that. So what happens is it selects again in the same shift fashion, it creates a Query and you can add a 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 so I can just quickly do this for you that I want to take a ratio based one, okay I'll just take log ratios and I can create a parameter of less than or equal to 0.33. 0.33 is log ratio change is something like 2 and add to list and greater than 2. Again I can add a query so If I select both of them I can create an or/and if I create or, you can apply that any feature which is this or that you select for my importance. I will be able to add a query on this to add a final wizard and then you have the data available for your final coming up. You can even select the database basis of annotation filter only which are mitochondrial specific. Depends on what questions you are asking so quickly look at only what you are saying so it says how many of them are there in that, it reduces the number. So in this fashion I can import unlimited number of database as well, right, I can give a name to it like filter. So you cannot create files on the root directory, you need to create a folder and do the job, so it can do the job in this fashion. So once you are ready with your data set what you want to sometime Is based on the experiment it gives you an opportunity what all you can do. Number 1 you can sort the column meaning f1 to f7, I want to make f7 first sorting by that so I can do a different ways of combining columns, one why you need to combine column, for example I have given three technical duplicates for each and so I can take an average of that in the column and it will take an average of all of this and u are ready with the data to go ahead for the analysis part. So many a times you can also go for the normalize two column, the meaning is I am having a zero 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 zero at one. And very important one is the dye swap. We know that Cy3 and Cy5. Cy3 is little less and cy5 is a little bigger so there is variation in the incorporation of that. To take care people usually do at least one dye swap experiment to accommodate the variation happening because of the dye. 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 the ratio of one by another wavelength so it just reverses it, so minus "X", will become plus, plus will become minus so it just changes the dye swap which will take care combining the data and variations happening due to the dye swap effect.

A: You also talked about dye swap when I was talking to them about DIGE technology and how one needs to use in fact labeling with different dye and reverse dye swapping so that there is no bias in the analysis.

B: true. So same can be accommodated here at the microarray to look at, and it is a very nice part actually to go ahead with and then many a times you find the few rows which I want to remove, I can remove few rows, I can select columns and remove, for example, some QC has not passed so I can remove that, which can allow me to do a different ways, and once you have done that you can go and do the clustering which visualization method. Technically clustering is divided into two types: hierarchical and non hierarchical. Hierarchical means that in the starting you have only one start point and then all other features are attached to that. Other one is non-hierarchical type where each group behaves independently of each other so there are K-means, SOMs, K-medians and the particularly one where people use hierarchical where you don't know where to start with. So they start with hierarchical when they don't know how many groups could happen and how many results I am expecting. Once you do have the results and idea, you can cut down because hierarchical is little RAM consuming, it takes little bit of more time because you can imagine all has to be linked to one and there are different ways of doing it, one is centre based, the best being coeffient correlation based or distance metric based or binary based. So binary based is usually

used for only CGS kind of analysis where it is present-absent type whereas the earlier two ones are extensively used in the microarray data. The pearsons correlation with centric ones are being used for the median or the mean type variation so when you quickly do it, you 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 clusterings, say I have got the filtering, I have got the log ratios and another important thing is whichever is something like bluish in color that particular one you have selected to work with. Yeah so this actually a filtered data, allowed to work with and when kind of clustering possibility, it says you need to create a quick data set. The meaning of that one is you have but now what you need to do is you need to look at third tab which allows one to look in the form which you want to create dataset for which you want to analyze. So here I have normalized the data to f1 so I can see the script what is happening just look at the one which we have performed at the level of different processing so that I can look at the data so here when I look at SOMs, at the level of 4x3 something like non-hierarchical type so it is individually being blocked so all different genes behave differently based on the profile they are made into one group so because I have given 4x3 I'll be able to see 4X3; so 4 number of columns and 3 number of rows they are being divided so one can have any number. The idea come from the hierarchical but as you see here immediately the clustering gives an immediate response that okay, this is very low in these ones, again it went down and you are able to see again up.

A: In different time, it'll be coding different expression.

B: correct. So based on the behaviors, you can see that okay this is being grouped up well so there are different ways which people prefer but usually preferred ones are the Kendal ones for the hierarchical type after the pearson's correlation with centric once are used. and for the this one after soms(?) people usually use Euclidian squares, using the some of the non hierarchical types K means and K medians is a better ones to use 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 the statistics. Statistics in the sense if you click on any of the ones which is something like statistical test. I have

different options to go ahead with different types of test one can perform like student t-test and the reason is when it equal variance with gausian normal bell shaped curve, you prefer this test. When the variance is not equal, you want to go with the second type where a small modification of student t-test but without normal variance happening and there are student t-test where you don't know what actually variance means to and you prefer pair test where the sample is coming from the same origin especially in the form of cancers where the cancer being removed surgically, people remove some normal sample and same biological patients give the origin of that. So this helps this particular one to allow one to select based on which background you have and other one is ManWhittney test which people use which actually people use where there is no normal Gaussian say for example few times it is only standalone ones, ups and down. 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 and you perform the analysis data.

A: So you need to look at the experimental design and then select their right statistical parameters for further analysis.

B: correct. So in this fashion, what it does is, it tells you that okay, I have got different kind of data sets which you want to select with so I select them and say group them according to o and 1s..like all cases and all controls and then I perform the analysis. So let's quickly do that..so this data is based on something like different kinds of components, functions so I want to understand the differences between different functions or I can quickly say, you can quickly create a dataset, so I can create a dataset, says which one you want to create with, so okay create a dataset from all these different time point so basically I know that it is high glucose in the beginning so I can group them all as very high..because this is a time point although this is decreasing. More likely I will select only once with 90 and then compare with very low type to see how they are getting differently expressed? To see overall changes happening in the low and high level, I have to quickly analyze this data available to me. the name i give is high sugar so it is available for me now. I can also create a one with low sugar and I am going to compare these two groups. I can differentially color them and see how the

things are happening. hit ok and you are ready with the different things. other important thing is multiple T-test correction, it also gives an opportunity for different correction for the multiple type, say for example, Bonferroni being used, Hodgeberg, Benjamin ddifferent ways of correcting for the multiple test correction. You can apply different ones and work preferably Bonferroni, which is more stringent type. Hodgeberg and Benjamin little lenient on multiple correction type. You select them and see there are still many which are getting significant P values so these genes become really important for me, I can see I can create the image, store them and look back what those genes are, what those functions are. So in this fashion one can visualize, look at the different genes and kind of statistics with different kind of multiple corrections and look at the data. So in this fashion Acuity helps in understanding the data analysis.

A: okay, it was very nice to see that there are so many parameters and options which we have here for QCing the data and analysing it further for obtaining some meaningful information. i guess there is no end to doing all of this 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 of this software and at least giving a glimpse of the entire workflow, how different types of processes 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 this lecture, I think I should just finish here. At least students have got the glimpse of the process involved analysis and how one can look at it in a stringent way test one need to perform and then different type of filtering can be done to obtain different type of fold change and different type of ratios one need to obtain. Further one needs to look at the trend of each of those 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 microdata analysis.

B: thank you, Dr.Srivastava

A: thank you.

So today we had discussion about microarray data analysis. you also seen demonstration of acuity software to learn about various type of parameters involved in performing 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 but also requires programming and very good statistical analysis. experimental design is very important in these microarray based experiments if you are putting garbage in then 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 tools 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 completely relying on one software or a you should also think about the biological context, look at the control and then after careful biological as well as analytical analysis, you can probably obtain some meaningful information from these datasets. so microarray experiments generate high throughput data. They provide you thousands of features information in a short time but it becomes very challenging to analyse the data specially when you have to compare various slides from different experiments and you have to normalize them equally so that you can compare all the slides on same platform. So careful image processing and data analysis becomes very crucial in microarray based experiments. Thank you!