HANDOUT

LECTURE-32

MICROARRAY WORKFLOW: DATA ANALYSIS

Slide 1

In today's lecture we will talk about Microarray work flow: especially focus on Data Analysis. This is in continuation of our previous lectures where we talked about different stages involved in the 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 sets 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 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 type 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.

SLIDE 2

We will discuss about various basic concepts.

- Normalization: In the microarray experiments people use different types of chips for different types of experiments so 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.
- Experiments 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.

SLIDE 3

Principle component analysis:

- Principle component is the linear combination of optimal weighted optional variables to test whether the protein expression is consistent throughout multiple sample from same experimental group.
- Are there protein outliers, spots mismatched or not to proteins, to realize all this type of variations, principal components analysis is performed.
- The PCA works by finding superchains that explains the most variance in the sample are orthogonal to each other.

SLIDE 4

Clustering.

- After realizing the microarray data analysis, you like to cluster data to a pattern, If possible.

- Your control or treatment falls 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.

SLIDE 5

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 using a self-organizing map such as one shown here in this slide.

SLIDE 6

Then there is one different type of approaches 1) a supervised approach to determine chains that fit a predetermined pattern or b) an unsupervised patterns to characterize the components of a data set without a prior input or knowledge of training signal.

A summary of the discussion with application specialist, Mr Pankaj from Spinco which covers various concepts involved in data analysis along with a demonstration of the software, is provided here.

Discussion with Application Specialist, Mr. Pankaj Khanna From Spinco Biotech(a distributor for Molecular Devices Products in India).

- 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.

Overview of data acquisition using Genepix Pro

- Once you are ready with the slide, usually people put on and hardware parameters are been selected, the 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 and once you are ready with the TIFF image, you perform basic analysis in GenePix Pro.
- See, for example, aligning of a different feature in the form of .GAL 5 which you are seeing. Then 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, or a .GPR file.
- So lets 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.
- 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. 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. 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 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.

- Normalization: Because we do microarray experiments on a chip to chip or experiment to experiment basis, variance can come in. To address this, normailization is necessary
- 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 for 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 spots which can be removed from 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 based correction which can be done over with.
- 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 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 normalization or you don't include the normalization, 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 visualize in the form of data vs different intensity plots.

- Given the immense amount of data generated, quality control is extremely important. The basic workflow involves a first level analysis, where the GenePix Pro software does some QC checks.
- 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 imported inside the acuity

Acuity software:

- This 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.
- 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 it so 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 look back whenever to require to and also it is optimized for windows.
- It is written in C++ , so it can work very fast and giving you 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 visually to quickly understand what is happening at the biological levels.
- Experiment and microarray parameter management: So many scientist 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.

- 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 application 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 ones 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.

Acuity Software Demo

- What you are looking at as a GOI interface of an acuity which is first divided on top in the form of a 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.

- 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.

Steps in Acuity workflow

- 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.
- 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.
- 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 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.
- Substance annotation: A very important next step is a substance annotation. This substance here I mean is 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 you have said already that it is very 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 switch over different types and just go to parameter file and look at what all details.
- So here in the working area, which we have defined again, split 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 shows 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.
- 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 the gene 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 CEs 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 performing be immediately seen.
- 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

signally different features can be individually analyzed and checked at how the behavior is.

- 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 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 logos 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.
- 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.
- There is a description of how many flags were there; 6400 at a time for normalization. See after doing the normalization and before the normalization how the data looks like and after normalization how the data is looking like. 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 is 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.
- 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. 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 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.
- 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 and that's that my visualization make more sense to me.

- So people can do okay in detail up and down with a range of so and so two fold up-regulated and two fold down-regulated 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.