Previously on Cbio:

Last lesson we discussed the usage of E-M algorithm for motif discovery.

2 General overview

Microarray is a technology developed for gene expression analysis. It takes advantage of the base-pairing property which is the making of hydrogen bindings of single stranded DNA with its reverse complement. As a double stranded molecule, DNA reaches its most stable condition.

The idea is to 'fetch' a specific sequence by placing it’s reverse complement as 'bate'. We’ll refer the reverse complement sequence as probe. Note that we will be 'fetching’ a DNA sequence despite of the fact that:

- Gene expression = amount of transcribed mRNA.
- We’ll deal with that issue later in this scribe.

Microarrays are large scale array of such probes that enables us to identify a large amount of DNA sequences from a certain genome. Here we will discuss 4 aspects of microarray technology:

1. Techniques for extracting DNA that reflect gene expression in a cell.
2. Techniques for building a microarray.
3. Hybridizing extracted DNA with microarray probes (one-dye vs. two-dye)
4. Problematic issues regarding microarray technology.

We will conclude with different approaches on planning a microarray experiment.

3 Techniques for extracting DNA

Our goal is to extract mRNA from a given group of cells (for which we like to learn of gene expression) and modify it, so when placing on an array of probes, conclusions can be made on amount of mRNA in the cells for each gene. The process is best explained using the flow chart in figure 1. In general, cDNA is made out of the mRNA using reverse transcription, amplified and labeled.

4 Techniques for building a microarray

4.1 Array of cDNA

When trying to learn about gene expression, we often do not focus on one or several genes, but wish to witness a certain expression pattern, or find the up and down regulated genes. Thus, our set of probes must contain all cDNA for a specific organism we examine.
On a microscope glass, we place dots, each containing a different sequence from the cDNA (many oligonucleotides for the certain sequence for that dot). There dots are ‘printed’ using a robot. The advantage here is that all gene sequence is included. The disadvantage is for long sequences, partial hybridization might occur and cause experimental noise. For cDNA microarray we will use two-dye strategy that will be explained later, to minimize this noise.

4.2 Artificially synthesized probe

When looking for a specific gene, an oligonucleotide can be synthesized and be placed on the array as a probe. However these synthesized probes are 50-70 bp long, so the gene must have a sub-sequence that long that differentiates it from other genes, and maintains the property of random hybridization (which is explained in the next section).

4.3 Synthesizing oligonucleotides on the microscope glass - Photolithography

Affimetrix was the company to exploit the technology for creating electrical circles, photolithography, for microarray purposes.

‘Photolithography is a process used in semiconductor device fabrication to transfer a pattern from a photomask (also called reticle) to the surface of a substrate’ (Wikipedia).

In our case the pattern is the different probes (where they are to be located on the microarray) and their oligonucleotide sequence. The surface of a substrate is the microscope glass on which we print the probes. The big advantage here is that our microarray can be extremely dense.

However the probes created are only 25 bp long, which is a big disadvantage because there is a good probability a random sequence identical to 70% of the probe will appear in the genome, and hybridize with the probe misleading the researcher.
Affimetrix tries to solve this issue:

1. For each gene, many 25bp probes are placed. Each probe represents a different sub-sequence of the gene.

2. Each ‘real’ probe (Perfect Match probe), is adjacent to an additional probe (MisMatch probe) that is identical to the real except for the middle nucleotide. A non specific hybridization is expected to occur in both probes, while a specific one, will occur in the Perfect Match probe only. Different association affinity of the sequence depending on it’s location on the glass will have no effect here as the probes are adjacent.

5 Hybridizing extracted DNA with microarray probes (one-dye vs. two-dye)

At this point we possess labeled DNA representing the genes expressed in a certain group of cells (under certain conditions) and a microarray containing single stranded DNA probes.

In order for the labeled DNA to associate with the probes, at least part of it should be as a single stranded molecule. Therefore, when putting labeled DNA on the array, the temperature is such that part of the dsDNA opens into ssDNA molecules. Then they are able to associate with reverse complement sequences of the probes. Note that a too high temperature will cause all DNA to appear as ssDNA and even in the case of affinity to a probe, dsDNA molecule of prob-labeled DNA will not be made. When putting labeled DNA on the array, an association or a disassociation may occur.

At that point we are facing two strategies for hybridizing labeled DNA with array probes:

1. **One dye:** The extracted labeled DNA will associate with the probes, then the array will be washed and then scanned with a special scanner. The scanner will fire futhons on the array surface and read the frequency of the light returned from the fluorescent labeling. This will indicate the content of the extract. Here association depends on base pair match between the labeled DNA and the probe.

2. **Two dye:** We use two extracts of (differently) labeled DNA. Both extracts are put on the array. The array is then washed and the scanner now reads the ratio between the two dyes. The basis of this process is allowing sequences from both extracts to compete over a probe. That is, assuming both dyes are chemically alike (in the matter of association property), if a sequence dominates in one extract over another, we will see the ratio shifted towards that extract. The advantage of this strategy is that the process of extracting and labeling DNA was similar in both extracts. This neutralizes part of the noise which was as a result of experiment conditions.

Figure 2 displays the % of binding depending on concentration according to Michaelis-Menten equation. For each strategy, the expression is to be measured at a different stage.

5.1 One-dye scanning (the Affimetrix way)

Although two-dye strategy ensures elimination of at least some experimental noise, Affimetrix claim, that due their addition of the ‘MisMatch’ probe and placement of many sub-sequences for each gene, even one-dye can give reasonable results.

The scanning process is of the following: The scanner uses a laser which screens the array area ad reads the frequencies. The labeled associated molecules return the pick of the frequency. The resolution of the scanner must be identical to array printing resolution. The scanner produces a JPEG file, where each pixel indicates the intensity of the signal for that pixel on the array. Next, the exact position of the spot (the associated sequence) is determined using the average intensity and differentiating between spot intensity and background intensity.

5.2 Two-dye scanning

The scanning is similar to one-dye except for the fact it is done twice while the scanner reads a different frequency each time (the two extracts are differently labeled). Data on each spot is then outputted including variance, average, median and ratio between the two dyes.
6 Problematic issues regarding microarray technology

6.1 Array printing:

Microscopic differences in the amount of DNA in each probe can effect signal intensity to not truly reflect differences in labeled extracted sequences.

When printing is done manually many errors can occur, from DNA amount to probe sequence. When printing is done using robots many of these noise factors can be eliminated.

6.2 Two-dye hybridization:

As discussed before a ratio is returned as an indicator of DNA content in one extract in comparison to the other. Is this ratio representative as we wish it to be? Source for trouble: 1. Different growth of initial cells from which cDNA is extracted will cause a certain ratio as a result of some factors different from what we’d like to check. 2. Different amounts of extracted DNA for the two groups of cells will cause shifting of the ratio not necessarily caused by different expression pattern. 3. ‘Stronger’ labeling of one of the extracts will cause shift in the ratio towards that dye, distorting the results. 4. The signal read by the scanner for each dye is not a direct representation of gene expression and thus, so is the ratio. Due to all problems mentioned, a quality control for two dye arrays had to be applied:

1. If the same amount of labeled DNA was extracted from both groups of cells, the total signal (sum of all hybridized sequences) for each one of the dyes should be the same. The signals will differ for probes representing genes with varying expression for the two extracts.

2. When plotting a graph for \( Cy3 \) dye vs. \( Cy5 \) dye for each probe, we expect the distribution as appears in figure3. But in fact, from real microarray data we receive a plot that looks something like figure4. Therefore we will usually calculate the log \( Cy5/Cy3 \) vs. \( Cy5 + Cy3 \), and expect to receive the graph in figure5.
6.3 Dealing with local phenomena on the microscope glass:

When analyzing scanner output we must make sure that on each window the average results will be similar. That way we can eliminate shifting of results deriving from local distinctions on glass surface.

6.4 Affimetrix is also facing a problem:

The problems described above are solved by Affimetrix on the electronic engineering level. However they face another problem:

As described earlier Affimetrix places on their array a MisMatch probe in adjacency to a Perfect Match probe. When receiving output from the scanner, given that for some sequence Perfect Match signal values are \( x_1, \ldots, x_k \) and MisMatch signal values are \( y_1, \ldots, y_k \), the score returned for that sequence is

\[
\frac{1}{k} \sum_{i=1}^{k} (x_i - y_i)
\]

Why is that problematic?

1. If a signal for a certain probe gave very different results, out of the scale of others, it will effect dramatically the total score when in fact it is far from representing the data.

2. It is possible that the MisMatch sequence is in fact a Perfect Match for another existing sequence. In that case, the difference between Perfect Match signal to MisMatch signal is misleading in regards to validating specific hybridization.
3. The result number returned for each sequence is not in direct correlation with the amount of mRNA in the initial extract.

Solution techniques:
1. Looking for coloration in adjacent probes. If a certain probe is found incompatible with the other data, multiply its signal by some constant or ignore it.
2. Measuring signaling for a certain gene on several chips (in two-dye) can give pretty credible results regarding the ratio of gene expression in both extracts.
3. Software packages such as dChip and rma for probe-level and high-level analysis of Affimetrix gene expression microarrays attempt to solve these problems.

6.5 Different results formats:

One-dye:
Affimetrix returns values that represent ‘intensity’. They are a function of the original signal intensity returned by the scanner, that in it’s turn, reflects the gene expression measurements. Rma and dChip return values that represent \( \log(\text{intensity}) \). Note that we are unable to compare Affimetrix results and rma/dChip results directly.

Two-dye:
The expression usually reported is \( \log(Cy5/Cy3) \).
However many are interested in a term called fold change which stands for:

\[
R = \begin{cases} 
  Cy5/Cy3 & \text{if } Cy5 > Cy3 \\
  -Cy3/Cy5 & \text{else}
\end{cases}
\]

6.6 Additional thing that can go wrong:

Hybridization and scanning at different times might result in additional noise, for example a high correlation was discovered between the quality of hybridization (resulting in quality of scanned results) and the ozone level.

7 Planning an experiment

Lets say you want to estimate gene expression over a set of conditions (changing temperature, change over time, etc.). Working in one dye will be very restricting because no kind of control group can be checked on the same array. In two dye you can use two types of designs:
Figure 6: Star design

1. Star design - figure 6 shows the basic concept of the design. $S_0$ is used as a control (for example a mixed extract of all checked cell extracts), while $S_1, \ldots, S_5$ represent the extracts under 5 different conditions. 5 two-dye arrays are made while in each array, $Cy_5$ labels an extract under some condition and $Cy_3$ labels the control.

Advantages:

$S_0$ is a strong control. One control is used for all extracts, and placed on the same microscope glass with each extract. Thus noise, resulting from surface differences is minimized.

Disadvantages:

(a) When gene expression pattern for the control is very different from that of the checked extracts, while the difference between the checked extracts are more subtle, it will be hard for us to calculate it in a precise way.

(b) If a certain gene is not expressed in the control cells, it will reflect a noise on the other extracts.

Solutions:

Regarding the second disadvantage, we can simply choose a control that does express (in some extent) the genes we are interested in.

As a general solution we might want to use a ‘pool’ technique, which means averaging the RNA of all extracts to create the control.

2. Graph design - figure 7 shows the basic concept of the design. Here there is not one control, but every two extracts (those connected in the graph) are checked together.

Advantages:

(a) No external control is used which can cause certain noise.

(b) Pairs of extracts can be compared directly.

(c) One can choose the form of the graph, for example extracts that differ in one condition will be connected and compared directly while others will be compared in a roundabout way.

Disadvantages:

One has to find a correct way to analyze results to reflect difference between connected extracts as well as unconnected extracts.

Usually the final results are returned in an Excel file. It holds the following data expression: $\log (x_{ij} / c_i)$ when $x_{ij}$ is the expression of gene $i$ in the extract $j$, and $c_i$ is a control factor for gene $i$ which is usually the geometric average for all extracts (a virtual pool).
Why log? Generally in biology, the noise is multiple and not additive, so it will look normal on a log scale. The total noise includes technology noise also, which is an additive one.