CSC2427 - Algorithms in Molecular Biology Wednesday March 1, 2006

Microarrays

Lecturer: Timothy Hughes Scribe: Desiree Tillo

Outline

- Microarray experiments
- Normalization
- Different types of microarrays
- Other applications besides expression profiling
- Clustering and interpretation
- Suggested reading

Traditional approaches to analyze gene expression have been a time consuming process, with different experiments required for each individual gene. With the use of microarray technology, it is now possible to analyze thousands of genes at once. The main principle underlying this approach is that there is a single, identical chemistry for all 4 base pairs of DNA. A microarray works by exploiting the ability of complementary strands of DNA (or RNA) to bind specifically, or *hybridize* to each other under appropriate conditions.

Microarray Experiments

One type of microarray, the cDNA microarray, contains DNA from each open reading frame (ORF) from the organism of interest spotted on glass microscope slides. These probes are used to detect cDNA (complementary DNA), which is DNA synthesized from a mature, fully spliced mRNA transcript. A robot spots DNA probes onto a 25 x 75 mm glass slides coated with poly-lysine (poly-lysine creates a positively charged matrix to allow the negatively charged DNA to adhere to the slide). Each of these spots are about 100-150 microns in diameter and are 200-300 microns apart, allowing approximately 20,000 spots to be easily placed on one slide.

Once the array is made, two samples of *labeled* cDNA (experiment and control samples, see Figure 1) are pooled together and are hybridized to the probes on the array. Some of the common ways to label a nucleic acid include (See Figure 2):

1. **Random priming of double stranded DNA -** Primers are short segments of artificially synthesized single-stranded DNA (ssDNA) oligonucleotides, which are complementary to a segment of a DNA molecule. A primer is necessary in order to create a double stranded DNA from a single stranded template. In order to create a labeled strand of DNA from double stranded DNA (dsDNA), there are three main steps:

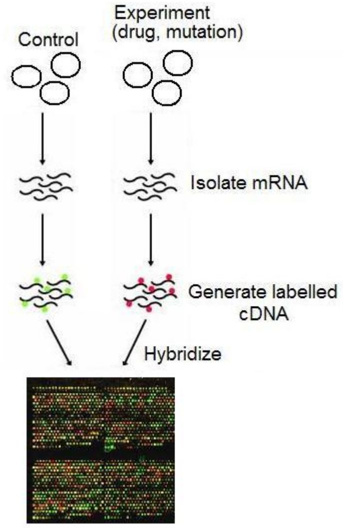

Figure 1: A cDNA microarray experiment. Each spot on the microarray contains DNA corresponding to an open reading frame of the organism of interest. Labeled cDNAs (green - control sample, red - experiment) are pooled and hybridized to the array. The fluorescent patterns on the array allow the determination of differences in gene expression.

(a) The double-stranded template is heated to $94 - 96^{o}$C in order to break the hydrogen bonds connecting the two strands of dsDNA, thereby separating the strands and creating ssDNA.

(b) The temperature is lowered allowing the the primer can attach or anneal itself to the newly separated DNA strands.

(c) DNA polymerase present in the reaction mixture then copies the DNA strands. Starting at the annealed primer, it works its way along the DNA strand incorporating nucleotides that are complementary to the template strand. These nucleotides are labeled either with a radioisotope or a fluorescent molecule (fluor). The end result is a new complementary strand of labeled DNA.

In cases when there is a complex mixture of different DNA templates, instead of designing primers specific to each template, random octamers are used. These are short primers made up of every combination of 8 nucleotides ($4^8 = 65,536$ different combinations). Each octamer can bind (anneal) to any section of DNA and a labeled cDNA pool can be made with minimal effort.

2. **Poly-T primed cDNA synthesis -** Messenger RNA (mRNA) transcripts have a stretch of adenines on the 3' end of the transcript (poly-A tail). A labeled complementary strand of DNA can be created by using a poly-T primer (complementary to the poly-A tail) and labeled nucleotides.

3. **Direct labeling -** Here, a combination of enzymatic and chemical steps are performed to modify and add the label directly to specific nucleotides present in the target strand.

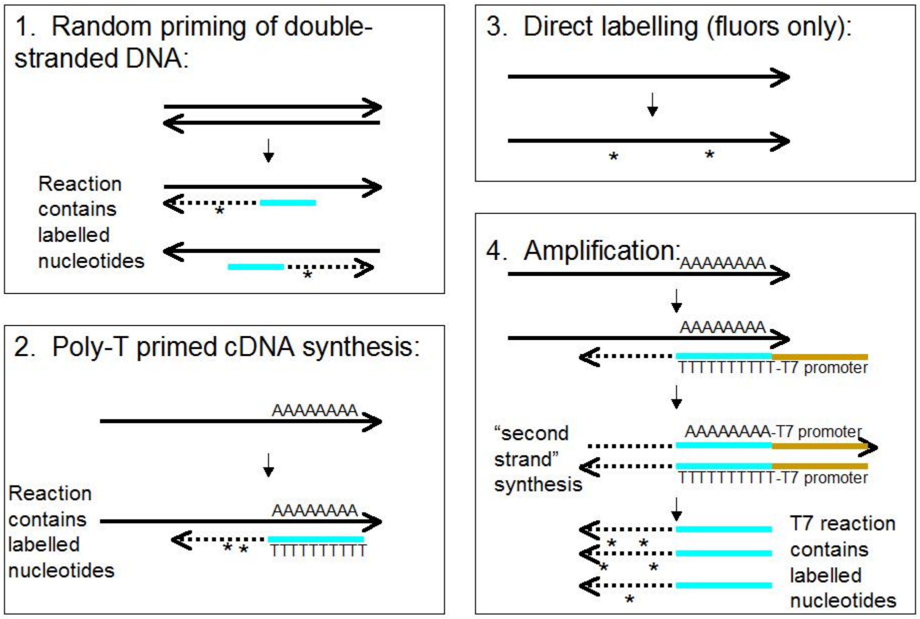
1. Random priming of double-stranded DNA:
Reaction contains labelled nucleotides
*
*
2. Poly-T primed cDNA synthesis:
AAAAAAAA
Reaction contains labelled nucleotides
AAAAAAAA
TTTTTTTTTT
* *
3. Direct labelling (fluors only):
* *
4. Amplification:
AAAAAAAA
AAAAAAAA
TTTTTTTTTT-T7 promoter
"second strand" synthesis
AAAAAAAA-T7 promoter
TTTTTTTTTT-T7 promoter
T7 reaction contains labelled nucleotides
* *
* *
*

Figure 2: Common methods for "labeling" nucleic acids (see text for details)

4. **Amplification -** Certain microarrays require a lot of cDNA so that fluorescence can be detected after hybridization. The addition of the T7 promoter to the poly-T primer allows approximately 100 copies of cDNA to be generated from a single mRNA template using the polymerase from the T7 bacteriophage (a virus that infects bacteria).

Common "labels" for nucleic acids for microarray experiments are the fluorescent dyes Cy3 and Cy5. A green laser (523nm) excites Cy3 and emitted energy is detected by an emission filter that detects emitted energy at 557-592nm. Red lasers (635nm) excite Cy5 and emission is detected by a filter that detects at 650-690nm. The array is scanned twice, once per channel used. The primary data consists of two grayscale TIFF files. The image is then processed and normalized. Processing includes lining up the spots on the grid, flagging bad spots, subtracting background, and quantitating the fluorescence of each spot on the array. Raw data from the scanner can be displayed in an spreadsheet, and contains information about each spot including the identity of the sequence, its location on the array, and most importantly, the intensity and the normalized log ratio, which is utilized to determine the relative expression of the genes represented on the array.

Normalization

Figure 3 shows the data from a single microarray experiment plotted in a scatter plot. In the scatter plot in the left panel, each point corresponds to a single spot on the array, and this plot shows the differences in expression of each gene (the ratio of drug and no drug, or red and green, $\log_{10}\left(\frac{red}{green}\right)$) and by how much, given by the $\log_{10}(fluorescence\ intensity)$.

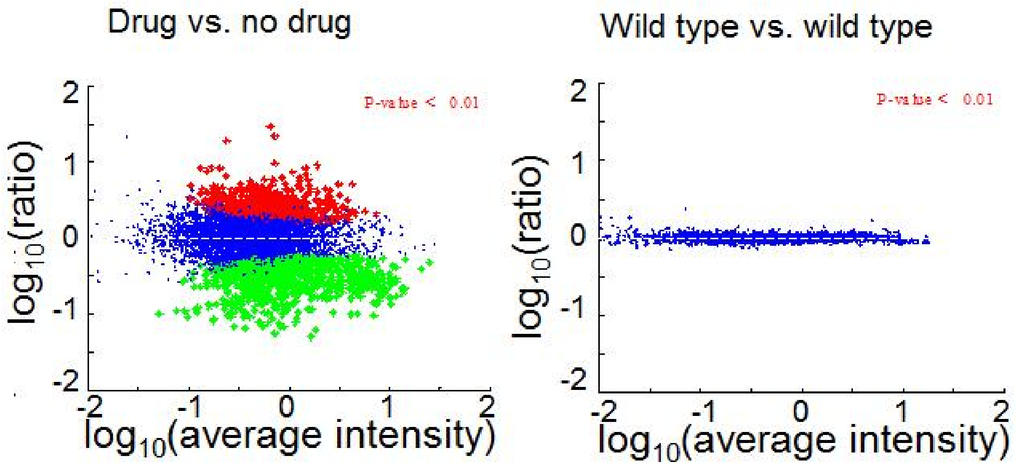

Figure 3: Microarray data from two experiments.

The scatter plot on the right shows the differences in expression of identical cell cultures (in this case, yeast). Since the cultures being compared are identical, there are no considerable differences in expression between them. The differences in intensity in this case are due to "biological noise", which is due to different genes being expressed at different levels, which normally occurs in the cell. Sometimes wild-type vs. wild type comparisons take a curved shape, similar to a "banana" or "jumping whale", which may be due to effects that arise from variation in the microarray technology or the experimental methods used. These can include unequal amounts of starting RNA, differences in labeling efficiencies of the fluorescent dyes used, and systematic biases in the measured expression levels. It is for these reasons that it is important to normalize the data in order to discern whether the results are due to actual biologically important differences between the cDNA samples detected.

One of the most common normalization techniques for microarray data is "lowess" or "loess smoothing", which is derived from the term **lo**cally **we**ighted **s**catter plot **s**mooth, using locally weighted linear regression to normalize the data. This process is considered local because a smoothed value is calculated using data points defined within a given window, or span (usually some percentage of the total data points on the plot).

Other methods used to process and normalize microarray data include:

- High Pass Spatial Detrending, which removes the local dependencies of each of the intensity values due to their physical positions on the array (See: O. Shai, Q. Morris, B.J. Frey (2003)).
- Variance Stablizing Normalization (VSN), which involves derivation of a parametric family of transformations of the form $y = \text{arcsinh}(ax + b)$ from the measureed fluorescence intensities (See: Huber *et al.*, 2002).

Other Types of Microarrays

Up until now, we have been discussing cDNA microarrays. Another type of array is the oligonucleotide array. Instead of being "spotted" onto glass slides, short DNA sequences, or oligonucleotides (oligos) are synthesized directly onto the glass slide via a number of different methods, listed below:

Photolithographic arrays (Affymetrix) - Probes on these arrays are synthesized using photolithography. Briefly, the slide is coated with a light-sensitive chemical compound that prevents the formation of a bond between the slide and the first nucleotide of the DNA probe being created. Chromium masks are then used to either block or transmit light onto specific locations on the surface of the slide. A solution containing one of either thymine, adenine, cytosine and guanine is flooded over the slide, and a chemical bond is formed in areas of the array that are deprotected by the mask (exposed to light). This process is repeated 100 times in order to synthesize probes that are 25 nucleotides long. This method allows for high probe density on a slide. Since these arrays are known as the industry standard, the protocols for their use are well developed. Unfortunately, not all probes work well, sample preparation requires amplification, and only single color hybridizations can only be performed (i.e. - you have to buy two chips to do a comparative hybridization experiment).

Inkjet arrays (Agilent) - Sequences on these arrays are synthesized using an ink-jet printer head, but instead of applying ink, it applies the phosphoramidites (a modified form the base) to create the oligos. The sequences on these arrays can be customized, are longer (60-mers), the data obtained using these slides correlates well with cDNA microarrays. The disadvantages of these arrays are that they are available only from single supplier (although it may soon be possible to make your own synthesizer), and density is currently limited to approximately 45000 spots/arrray.

"Maskless arrays" (Nimblegen) - Instead of the masks used to construct Affymetrix arrays, this synthesis procedure makes use of small mirrors to deflect or allow light to pass through. These mirrors are controlled by a computer which reflects or deflects a desired pattern of UV light, allowing coupling of nucleotides in the unprotected areas. These arrays allow the user to specify their desired sequences, and the synthesis process also allows for high probe density (like the photolithogaphic arrays). However, in order to bypass intellectual property issues (Affymetrix), hybridizations have to be done in Iceland, where the Nimblegen headquarters are located.

Other applications besides expression profiling

Aside from looking at differences in mRNA levels under different experimental conditions, microarrays can be used to answer a number of other important biological questions. These include:

- DNA Copy number - Differences in copy number of certain genes (such as what happens in cancers) can be detected on microarrays by measuring differences in fluorescent signals, which are proportional to both decreases and increases in copy number.

- Genotyping - Microarrays can be used to detect single base differences (single nucleotide polymorphisms) in different individuals. This array contains probes that match single nucleotide variants of a particular gene of interest.

- Protein-DNA associations - Binding sites of DNA binding proteins (for example, transcription factors) can be identified. One way in which this is done is by crosslinking (fixing) all of the DNA binding proteins in the nucleus with formaldehyde, isolating a particular protein-DNA complex using an antibody specific to a protein of interest, followed by identification of the binding sites by isolating the DNA bound to that protein and hybridizing it to an array containing intergenic sequence.

- Molecular "barcoding" - Each probe on these arrays is a unique sequence or barcode which is a unique short sequence that allows the identification of different strains or species present in a complex population (See: Giaever *et al.*, 2002).

- Protein arrays - Instead of DNA, these arrays contain different proteins on each spot. These allow the identification of protein-protein interactions, which are important in many biological processes.

- Transformation arrays - These arrays contain cells expressing a defined cDNA (instead of poly-lysine, these arrays are coated with gelatin). These arrays are useful in screening drug targets or cDNAs that cause a phenotype of interest.

Clustering and interpretation

A common procedure using microarrays is to conduct several experiments across the same genes then measuring gene expression during each trial. For example, gene expression can be measured across time points, different patients, different mutants, etc. Data from multiple microarray experiments can be compared via scatterplots and clustering analysis. For two experiments, a scatterplot of the log ratios can be used to compare the behaviour of two genes. If two genes interact or are part of the same biological pathway, their behaviour across experiments will be similar.

Cluster analysis is another way to interpret expression data. One widely used algorithm (Eisen *et al.,* 2002) groups genes based on the similarity of their expression levels across experiments. Steps involved in clustering are as follows:

- Remove experiments and transcripts falling below a cut-off P-value and ratio threshold

- Cluster experiments and transcripts:

  - Create a gene expression similarity score matrix for each pair of genes by calculating the correlation coefficient between them.
  - Compute a dendrogram that assembles all of elements in this matrix into a single tree. This particular approach uses an agglomerative hierarchical clustering algorithm to group the elements in the matrix:

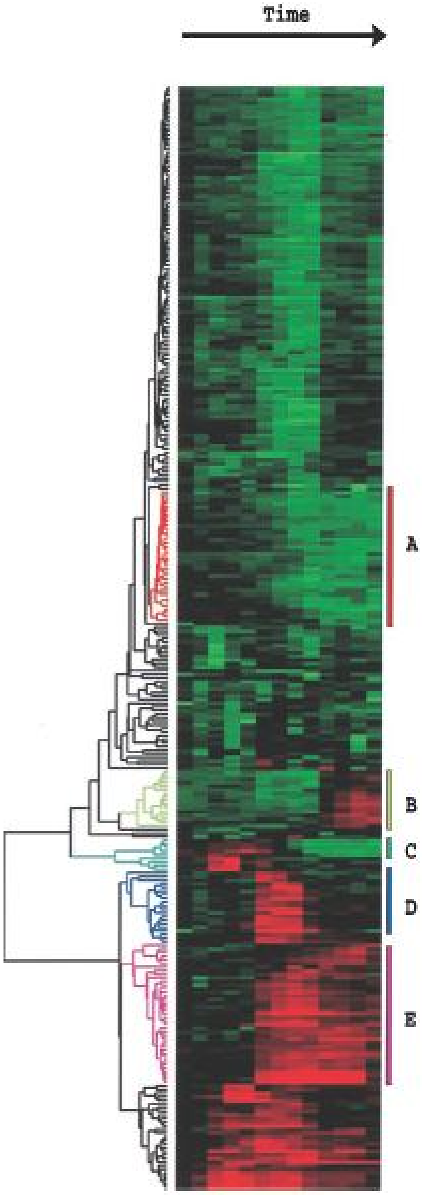

Figure 4: A sample clustergram (from Eisen *et al.*, 2002) which shows the expression data from time course experiments of primary human fibroblasts following serum stimulation. Genes of similar function (A, B, C, D, E) tend to cluster together.

- Given a n by n similarity matrix (n genes), scan the matrix to identify the highest value (represents the most similar pairs of genes).
- Join these two genes with a node, and compute a gene expression profile for the node by averaging observation for these two joined elements.
- Update the matrix by replacing with this calculated value replacing the joined elements.
- Repeat $n - 1$ times until only one element remains.

The dendrogram is then used to order the genes and experiments in a data table, and this result is displayed graphically in a "clustergram". Each cell in this data table is coloured based on its fluorescence ratio (log ratios of 0 - black; postively increasing log ratios - red with increasing intensity; negatively increasing log ratios - green with increasing intensity). An example of one of these clustergrams is shown in Figure 4.

Clustergrams allow easy interpretation of expression data. If the expression values between genes are are highly correlated, can be hypothesized that these genes are co-regulated and possibly functionally related (this can be useful in assigning function to genes whose functions are unknown). It is also possible when comparing experiments, to identify genes which are differentially expressed in different situations.

Other types of clustering algorithms (K-means, for example) can also be used to analyze microarray data.

Suggested reading

Eisen MB *et al.*, 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A.* 95(25):14863-8.

Hartigan JA. Clustering Algorithms, Wiley, New York and London (1975).
No longer in print, is available on CD

Jain AK *et al.*, 1999. "Data Clustering: a review". ACM Computing Surveys 31(3).

Hegde P *et al.*, 2000. A concise guide to cDNA microarray analysis. *Biotechniques.* 29(3):548-50, 552-4, 556

Sherlock G, 2000. Analysis of large-scale gene expression data. *Curr Opin Immunol.* 12(2):201-5.