Microarrays Sequencing

Section 4

Chih Long Liu

Oct. 14th, 2003

Outline

- Logistics
  - Final Project Deadlines
- RNA1 Topics
  - Traditional RNA analysis
  - Microarrays
    - Types and Construction
    - Overview and Terminology
    - Usage and Analysis
  - Sequencing

Final Project Deadlines

- **Project ideas due Tues. 10/21/03**
  - 1-2 paragraph description
  - Team members listed (if any)
  - Please submit to your section TF
- **Proposal due Tues. 11/4/03**
  - Length: about 1 page
  - Should include description on overall goals, planned approach, and any progress
  - Please submit to your section TF

Final Project Deadlines

- **Project due Tues. 12/2**
  - Includes BOTH written report and powerpoint presentation
  - 1MB email limit on file attachments
  - 5% penalty per late day
- **Project presentations 12/2, 12/9, & 12/16**
  - (Problem Set 5 is due 12/9, so plan carefully)

Important Terminology

- **Nucleic acid hybridization**
  - The binding of complementary nucleic acid strands (A pairs with T/U, G to C). Two hybridized strands are said to form a duplex.
- **Nucleic acid denaturation (melting)**
  - The opposite of hybridization; when two complementary sequences come apart into single strands. This can be accomplished by heating, extremes of pH, or reducing the salt concentration.
- **Melting temperature (T_m)**
  - The temperature, under a standard condition, at which a double stranded DNA molecule denatures.
  - Major factors in determining T_m are the length of the duplex and the GC content (GC base pairs are more stable than ATs).

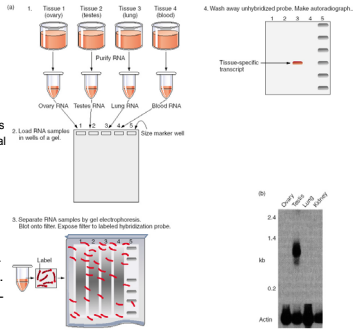
Important Terminology

- **Nucleic acid probe**
  - A short single-stranded nucleic acid sequence whose sequence allows it to hybridize to a sequence of interest. It will also be "labeled" in some way, e.g. have a radioactive or fluorescent attachment, so it can be detected.
  - Probe design involves optimization of melting temperature, secondary structure, and probe-probe sequence similarity.
- **cDNA (complementary DNA)**
  - Created from an enzyme called reverse transcriptase, which can copy RNA molecules into DNA. cDNA sequence is the reverse complement of the original RNA template.
  - cDNA is commonly used to make probes that represent an RNA sample.

Traditional RNA Analysis

Northern Blots

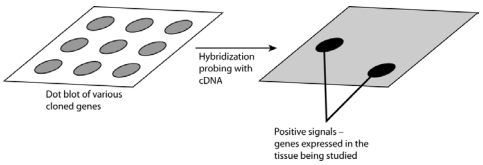
- Used to detect the presence of a particular RNA transcript
- RNA purification
  - Chemically extracted with phenol/chloroform from homogenized cells and tissues
  - mRNA transcripts (<5% of total RNA) have a poly-A tail and can be isolated with a poly-dT matrix
- An RNA sample is separated by size on a gel, transferred to a membrane and then allowed to hybridize to a radioactively-labeled nucleic acid 'probe'.
- Northern blots can be semi-quantitative but aren't very precise



From Hartwell L. H. et al, *Genetics: From Genes to Genomes* (2000) p. 362

Traditional RNA Analysis

Dot Blots



- **Microarray precursor**
  - **Note – it doesn't inform you of RNA size**

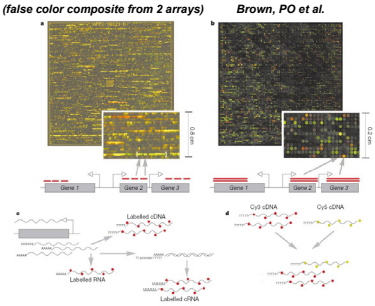
Key Principle of all hybridization techniques:

Over a certain nucleic acid concentration range, the amount of nucleic acid which is hybridized is proportional to its concentration in the hybridization solution.

DNA Microarrays

Affymetrix
(false color composite from 2 arrays)

Spotted
Brown, PO et al.

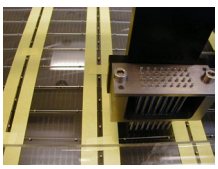


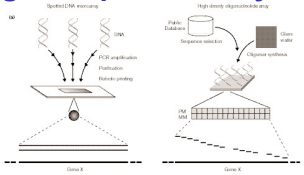
- Microarrays were initially developed to enable genome-scale gene expression analysis.
- The utility of microarrays lie in their ability to perform thousands of simultaneous measurements of a nucleic acid sample.
- Two major classes of DNA microarrays are ***high density oligonucleotide arrays*** (Affymetrix) and ***spotted arrays*** (developed by PO Brown et al.).

From Lockhart D.J. & Winzeler E.A., "Genomics, gene expression and DNA arrays". *Nature*. Vol. 405, no. 6788, 15 June 2000, p. 828.

Microarray Construction

Ways of getting DNA put on arrays





From Harrington et al., *Current Opinion in Microbiology*, 3(3): 285-91.

Spotted microarray (Brown, PO et al.) **Affymetrix microarray**

- **Spotted arrays**
  - Whole-gene or fragments of DNA physically deposited on glass slides
  - Easily customizable with a wide variety of different kinds of DNAs (next slide)
  - Robotic contact printing (shown) or piezoelectric printing (like today's inkjet printers)
- **High density oligonucleotide arrays (Affymetrix)**
  - Chemically synthesized *in situ* on glass wafers using lithographic processes
  - Short oligonucleotides (25-mers) tiled across gene
  - Paired PM (Perfect Match) and MM (MisMatches) format

Microarray Construction

What kind of DNA to use?

- **Regions of the genome**
  - ORFs, or Open Reading Frames: regions that are actually translated into proteins. Most common type used in DNA microarrays, and is used to measure global gene expression via mRNA transcript abundances
  - Intergenic regions: regions between different genes.
    - These regions are most useful in "Chip^{2nd} studies (Chromatin immunoprecipitation of protein-DNA compexes put on DNA Chips, or microarrays).
    - These studies examine protein binding (e.g. transcription factors, histone tail-binding proteins) to regulatory and non-transcribed regions of genes
  - ESTs, or Expressed Sequence Tags: transcribed sequences which are converted into DNA and partially sequenced. Most are of unknown function.
  - Clone libraries: Assortments of DNA fragments collected by a variety of means.
    - Sequences can be inserted into bacteria or viruses. If a sequence turns out to be interesting, the clone harboring that sequence can be grown up to produce enough to study.
    - A common clone library is a cDNA library, where each clone contains a single cDNA reverse transcribed from an RNA in an RNA sample of interest.
  - Tiling arrays: oligonucleotide arrays which contain oligonucleotides corresponding to sequences spaced at short intervals across the region of interest in the genome

Microarray Construction

What kind of DNA to use?

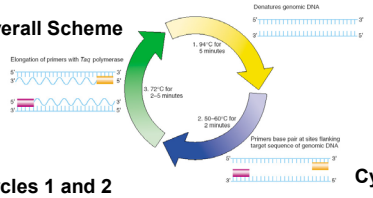
- **PCR products vs oligonucleotides**
  - PCR products
    - Derived from cells/tissues as starting template
    - The first microarrays consisted of all ORFs in the yeast genome, spotted as PCR products
    - Tend to be long (0.5-3 kb or longer)
    - Produce more stable duplexes (hybridized strands)
    - Averages signal across the entire gene
    - Low initial cost but variable quality
  - Oligonucleotides
    - Chemically synthesized
    - Tend to be short (25-70 bases)
    - Duplexes less stable (hence Affymetrix's PM/MM system)
    - Can target specific regions of a gene and test splice variants
    - High initial cost but frequently high quality

PCR (for non-biologists)

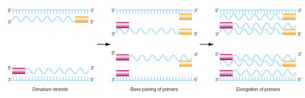
- **The Polymerase Chain Reaction (PCR)**
  - This Nobel-prize winning enzymatic reaction can make a large amount of DNA from a very small amount of starting material.
  - In principle only a single molecule is needed.
  - By designing 'primers' which surround a region of interest, the region between the primers can be copied exponentially.
  - Sets of primers can be designed to copy any region of the genome in quantities large enough to spot on a microarray.

PCR (for non-biologists)

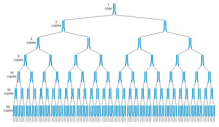
Overall Scheme



Cycles 1 and 2



Cycles 1-7, etc.

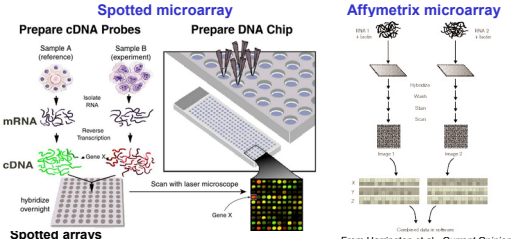


From Hartwell L. H. et al, *Genetics: From Genes to Genomes* (2000) pp. 294-5

Using Microarrays

for gene expression profiling

Spotted microarray | **Affymetrix microarray**



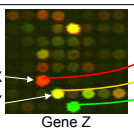
From Harrington et al., *Current Opinion in Microbiology*, 3(3): 285-91.

- **Spotted arrays**
  - A competitive hybridization of two RNA samples
  - cDNA Probe is labeled with fluorescent dyes (direct incorporation or amino-allyl coupling)
    - (typically cyanine 3 = reference, and cyanine 5 = experiment)
  - Data obtained is measured as a ratio of one color to the other (e.g. Red/Green ratio) and provides *relative* abundance information
- **High density oligonucleotide arrays (Affymetrix)**
  - A non-competitive hybridization of one single RNA sample (per array)
  - Data obtained is measured as absolute intensity units if it passes PM/MM and provides *absolute* abundance information

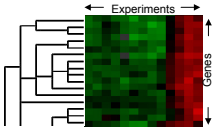
Analysis of Raw Microarray Data

Raw spotted microarray image → Signal intensity analysis → Color-balance normalization

Net signal intensity → Red/Green ratio calculation → Log transformation



Cy3	Cy5	$\frac{Cy5}{Cy3}$	$\log_2\left(\frac{Cy5}{Cy3}\right)$	
200	10000	50.00	5.64	X
4800	4800	1.00	0.00	Y
9000	300	0.03	-4.91	Z



Clustering preview (Section 6)

Repressed (downregulated)

Gene Expression

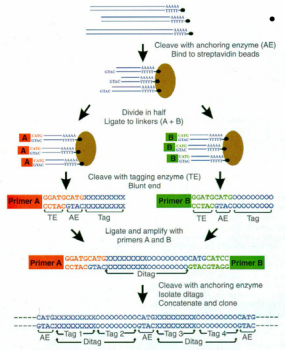
Induced (upregulated)

Interpreting Microarray Data

Spotted Microarrays

- Only measures relative abundances
  - Each spot on an array has a different hybridization efficiency; only self-comparisons are valid
  - Normalization is necessary to compare identical spots across separate arrays
- Cross-hybridization
  - Similarity between sequences or high abundances of other transcripts with low but significant affinities for the spot can hybridize to the 'wrong' spot.
  - Washing disrupts duplexes of given stabilities depending on stringency of the conditions used.
  - However, washing cannot completely eliminate cross-hybridization, especially when stabilities of specific and non-specific duplexes can overlap.
    - (consider definitions of T_m and denaturation and what the properties of a stringent wash might be)
  - Intelligent oligonucleotide design for spotted oligonucleotide arrays can greatly reduce cross-hybridization

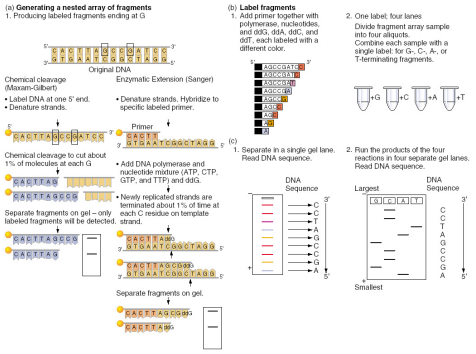
SAGE



From Velculescu, et.al, "Serial Analysis of Gene Expression", *Science* 270: 484-487 (1995).

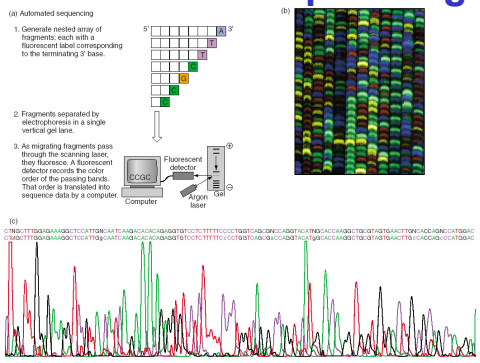
- **Serial Analysis of Gene Expression**
  - Employs sequencing to quantify RNA abundance on a global scale
  - RNA samples are processed to produce small sequence tags (10-14 bp) from each one
  - Tags are concatenated along with tags from other RNAs.
  - These stretches of tags are sequenced and the number of tags from each transcript is counted
  - Genome sequences enable identification of tag sequence to the corresponding gene
  - This method is digital (discrete counting events), whereas hybridization methods are analog (continuous response to changes in RNA concentration)
  - SAGE is currently more labor-intensive than microarrays

DNA Sequencing Methods



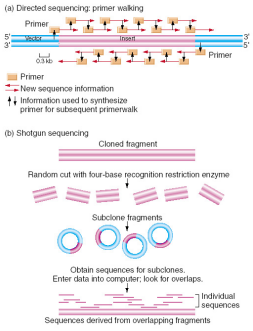
From Hartwell L. H. et al, *Genetics: From Genes to Genomes* (2000) p. 288

Automated Sequencing



From Hartwell L. H. et al, *Genetics: From Genes to Genomes* (2000) p. 292

Directed and Shotgun Sequencing



- **Directed Sequencing**
  - Involves "primer walking"
  - Slow and laborious, but more reliable
- **Shotgun Sequencing**
  - Randomly cut region being sequence
  - Reassemble region into contig via sequence alignment of overlaps
  - Much more rapid but less reliable
  - Requires several-fold coverage of region

Large sequencing projects usually use both methods

From Hartwell L. H. et al, *Genetics: From Genes to Genomes* (2000) p. 293

Next Week

- Population Genetics

Acknowledgement / References

Harrington et al., "Monitoring gene expression using DNA microarrays". *Current Opinion in Microbiology*, 3(3): 285-91.

Hartwell L. H. ed. et al. *Genetics: From Genes to Genomes* (2000). McGraw-Hill Companies, Inc.

Lockhart D.J. & Winzeler E.A., "Genomics, gene expression and DNA arrays". *Nature* 405 (6788): 827-836.

Velculescu, et.al, "Serial Analysis of Gene Expression", *Science* 270: 484-487 (1995).

This handout includes material written by Suzanne Komili, Yonatan Grad, Doug Selinger, and Zhou Zhu.

This handout also includes material from the laboratory of Pat Brown, Dept. of Biochemistry, Stanford University.