Introduction to microarrays (1)

Basic concepts
Software
Annotations
Databases
Presentation
Content

• Introduction
• Production and use of microarrays
• From Images to expression matrices
• Microarray bioinformatics
  – Software for the analysis of microarray data
  – Annotations and annotations databases
  – More microarrays databases.
Introduction
Some history

- Molecular biology has many techniques to measure RNA, DNA, proteins or metabolites.
  - Northern blot, differential display, SAGE
  - Southern blot: [similar to microarray]
- What characterizes the post genomic era is not what can be measured but the number of simultaneous measurements that can be performed.
A CSI approach to gene selection

• A crime has been committed (tumour)
• You’re CSI – Horatio Fisher- and want to find who’s responsible for this.
• Let’s see how you would act …
  – In the old times,
  – In the microarray age,
  – In the next generation age.
In the old times …

- You would chase the “Usual Suspects” and make an in deep interrogation.
  - If guilty you might make them talk,
  - But if not you might miss the bad guy.

- That is looking at specific genes may yield great or awful results.
In the microarray age...

• You have the census of most people and their fingerprints.
  – If you find a fingerprint in your database that is clean enough you may find the bad guy.
    • What about bad prints?
    • What about those who are not censed.
    • And those no-fingerprints?

• That is you may look at all known genes but you
  – do it Indirectly and noisly
  – miss genes/forms that are uncensed.
In the NGS epoch (now)

• If the crime scene had had cameras you would have directly known who the criminal was.

• **Sequencing allows you to access everything**
  - Known and unknown forms are sequenced.
  - The technique is less noisy and the resolution higher.
Microrrays represent a *paradigm shift*

With the same resources we obtain an image of lower resolution but wider scope.
So, What is a microarray?

- An experimental format,
- based on the synthesis or attachment of probes, which represent genes (or proteins, or metabolites ...),
- on a solid substrate (glass, plastic, silica ...),
- Intended to be exposed to the target molecules (the sample).
How does it work?

- The level of hybridization between specific probes and target molecules is generally indicated by means of fluorescence and is measured by image analysis.
- The measure obtained indicates the level of expression of the gene corresponding to the probe in the test sample.
Types of microarrays

Protens

- Tissues
- DNA
  - CGH arrays
  - SNP arrays
- RNA (expression)
  - Two color (or cDNA)
    - e.g. Agilent
  - One color (or Affymetrix or oligonucleótidos)
    - GeneChip® Affymetrix
    - Illumina bead arrays
Microarray applications

- Study of genes that are differentially expressed between various conditions (Healthy / sick, mutant / wild treated / untreated)
- Molecular classification of complex diseases
- Identify sets of genes characterizing a disease (signature or "signature")
- Predicting the response to treatment
- Detection of mutations and single polymorphisms (SNP)

But also
- Circadian clock analysis,
- Plant defence mechanisms,
- Environmental stress responses,
- Fruit ripening,
Production and use of microarrays
Expression microarrays

• There are many types of microarrays
• They rely on similar principles but
• the details of its operation change from one to another case
• Here we focus on expression arrays
  - 2-color arrays (spotted)
  - Oligonucleotide arrays (in situ synthesized).
Two colour microarrays (spotted)

- Chip design and production
- Sample Preparation
- Hybridization
- Scanning the chip
- Image analysis
General overview of the process

To visualize an animation go to: http://www.bio.davidson.edu/courses/genomics/chip/chip.html
Oligo microarrays synthesized *in situ*

- More advanced design than 2 colors
- Rely on technologies developed for microelectronics
- Some distinctive features
  - Not based on competitive hybridization: each chip containing samples from a single type (aka "1 color")
  - Probes are synthesized directly on the chip instead of in vitro synthesized and then attached to slides
  - Each gene is represented by a group of short probes rather than a single long probe
**Probesets, probes, PM & MM**

- A set of probes is used to measure mRNA level of a single gene.
- Each group (probeset) consists of multiple pairs of cells (probe cells)
  - with millions of copies of a 25bp oligo.
- Pairs consist of
  - a Perfect Match (PM) which coincides exactly with a portion of the gene
  - A Mismatch (MM) identical to PM except in the central nucleotide replaced with its complementary
Probesets are made of “Probe pairs” which represent different parts of same gene (1 gene = 1 probeset).

Probes are selected to be specific to the gene they represent and to have good hybridization properties.
Probe pairs = PM and MM
An attempt to detect noise
Proces overview (Affy)
Comparison between two types

Ramaswamy S, Golub T R JCO 2002;20:1932-1941
## Comparison between two types

<table>
<thead>
<tr>
<th>cDNA Microarrays</th>
<th>Oligonucleotide Microarrays</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ADVANTAGES</strong></td>
<td><strong>ADVANTAGES</strong></td>
</tr>
<tr>
<td>Cheaper (not anymore)</td>
<td>Quick and robotic manufacturing</td>
</tr>
<tr>
<td>Flexibility in experimental design</td>
<td>High Reproducibility</td>
</tr>
<tr>
<td>High signal intensity (long secs)</td>
<td>High specificity (short sequences)</td>
</tr>
<tr>
<td></td>
<td>Use many probes / gene</td>
</tr>
<tr>
<td><strong>DISADVANTAGES</strong></td>
<td><strong>DISADVANTAGES</strong></td>
</tr>
<tr>
<td>Low reproducibility</td>
<td>Requires more specialized equipment</td>
</tr>
<tr>
<td>Cross-hybridization (low specificity)</td>
<td>Expensive</td>
</tr>
<tr>
<td>Need more manual handling (possibility of contamination)</td>
<td>Less flexible (genes on the chip cannot be selected)</td>
</tr>
</tbody>
</table>
From image to expression matrix

- **Raw data**: Array scans, 3 slides, 100 rows, 100 columns
- **Quantitation matrices**: 3 files, 10,000 columns, $K$ measures
- **Gene expression data matrix**: 1 matrix, 10,000 columns, 3 columns (one per slide)
Expression measures (cDNA)

Gene expression is measured from intensity measures as the (corrected) relative intensity of one dye vs. the (corrected) relative intensity of the other.

\[ M = \frac{R_g}{G_g}, \text{ or } M_{\text{CORR}} = \frac{R_g - bgR_g}{G_g - bgG_g} \]

Background correction may be needed, or not, according to the array quality.
Example: gene expression data

Gene expression data on 6348 genes for 16 samples.

Gene expression level of gene $i$ in mRNA sample $j$

$$
M = \begin{cases}
\log\left( \frac{T_{SRBI}}{C^*} \right) \\
\log\left( \frac{C_{FVB}}{C^*} \right)
\end{cases}
$$
Expression measures (affy)

• Obtaining expression measures for affymetrix arrays is less straightforward.
  – Background correction and normalization is required.
  – There are multiple PM and MM values which must be integrated into one single expression value.

• The resulting summarized values are absolute expression measures which are more difficult to interpret than relative expression values.

\[
\text{Avg. Diff} = \frac{1}{|A|} \sum_{j \in A} (PM_j - MM_j)
\]
**Example: absolute expression values**

<table>
<thead>
<tr>
<th>Gene</th>
<th>C01-001.CEL</th>
<th>C02-001.CEL</th>
<th>C03-001.CEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1415670_at</td>
<td>8.954387</td>
<td>9.088924</td>
<td>8.833863</td>
</tr>
<tr>
<td>1415671_at</td>
<td>10.700876</td>
<td>10.639307</td>
<td>10.610953</td>
</tr>
<tr>
<td>1415672_at</td>
<td>10.377266</td>
<td>10.510106</td>
<td>10.461701</td>
</tr>
<tr>
<td>1415673_at</td>
<td>7.320335</td>
<td>7.252635</td>
<td>7.112313</td>
</tr>
<tr>
<td>1415674_a_at</td>
<td>8.381129</td>
<td>8.332256</td>
<td>8.393718</td>
</tr>
<tr>
<td>1415675_at</td>
<td>8.120937</td>
<td>8.082713</td>
<td>8.051514</td>
</tr>
<tr>
<td>1415676_a_at</td>
<td>10.322229</td>
<td>10.287371</td>
<td>10.282812</td>
</tr>
<tr>
<td>1415677_at</td>
<td>9.038344</td>
<td>8.979641</td>
<td>8.905711</td>
</tr>
</tbody>
</table>
Software for microarray data analysis
Which software for the analysis?

Microarray experiments generate huge quantities of data which have to be.
  Stored, managed, visualized, processed …
Many options available. However…
No tool satisfies all user’s needs.
Trade-off. A tool must be.
  Powerful but user friendly.
  Complete but without too many options,
  Flexible but easy to start with and go further.
Available, to date, well documented but affordable.
We picked up some options…

Many tools

Free / Commercial

[R, BRB, MeV, dChip…] / [Partek, GeneSpring, Ingenuity]

Downloadable / On-line

[R, BRB, MeV…] / [Gepas,…]

Standalone / As part of suites

[BRB, dChip] / [MeV (TM4), OntoTools]

A survey of free microarray data analysis tools:

http://chagall.med.cornell.edu/I2MT/MA-tools.pdf
# Open source analysis tools

<table>
<thead>
<tr>
<th>Programa</th>
<th>😊</th>
<th>😟</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R/Bioconductor</strong></td>
<td>Powerful, flexible, updated</td>
<td>Console-based, hard to use</td>
</tr>
<tr>
<td></td>
<td>Unix/Windows/Mac</td>
<td></td>
</tr>
<tr>
<td><strong>BRB tools</strong></td>
<td>Excel based</td>
<td>Hard to recover from error</td>
</tr>
<tr>
<td></td>
<td>User-friendly</td>
<td>Hard to extend</td>
</tr>
<tr>
<td><strong>dChip</strong></td>
<td>Expresión &amp; SNP’s</td>
<td>Windows</td>
</tr>
<tr>
<td></td>
<td>User-friendly</td>
<td>Les options</td>
</tr>
<tr>
<td><strong>GEPAS</strong></td>
<td>Web-based,</td>
<td>Web-based</td>
</tr>
<tr>
<td></td>
<td>Many options</td>
<td>A little rigid</td>
</tr>
<tr>
<td></td>
<td>Good tutorials</td>
<td></td>
</tr>
</tbody>
</table>

...
GEPAS (currently Babelomics)

Herrero et al., 2003, 2004; Vaquerizas et al., 2005 NAR; Montaner et al., 2006 NAR; Al-Shahrour et al., 2005, 2006 NAR; 2005 Bioinformatics
So, what you need is “R”?

R is an open-source system for statistical computation and graphics. It consists of.

A language.
A run-time environment with.
  Graphics, a debugger, and.
  Access to certain system functions,

It can be used.

  Interactively, through a command language.
  Or running programs stored in script files.
R and Microarrays

R is a popular tool between statisticians. Once they started to work with microarrays they continued using it.

- To perform the analysis.
- To implement new tools.

This gave rise very fast to lots of free R-based software to analyze microarrays.

The Bioconductor project groups many of these (but not all) developments.
The Bioconductor project

http://www.bioconductor.org

Open source and open development software project for the analysis and comprehension of genomic data.

Most early developments as R packages.

Extensive documentation and training material from short courses.

Has reached some stability but still evolving !!!

→ what is now a standard may not be so in a future.
Some pro’s & con’s

• Powerful,
• Used by statisticians
• Easy to extend
  – Creating add-on packages
  – Many already available
• Freely available
• Unix, windows & Mac
• Lot of documentation

• Not very easy to learn
• Command-based
• Documentation sometimes cryptic
• Memory intensive
  – Worst in windows
• Slow at times

We believe the effort is worth the pity!!!

• If you “just want to do statistical analysis”
  ➔ Easy to find alternatives
• If you intend to do microarray data analysis
  ➔ Probably one of best options
BRB-ArrayTools


Selected Features of BRB-ArrayTools

Multivariate permutation tests for class comparison to control false discovery proportion with any specified confidence level
SAM
Find Gene Ontology groups and signaling pathways that are differentially expressed
Survival analysis
Analysis of variance
Class prediction models (7) with prediction error estimated by LOOCV, k-fold CV or .632 bootstrap, and permutation analysis of cross-validated error rate
DLDA, SVM, CCP, Nearest Neighbor, Nearest Centroid, Shrunken Centroids, Random Forests
Clustering tools for class discovery with reproducibility statistics on clusters
Built in access to Eisen’s Cluster and Treeview
Visualization tools including rotating 3D principal components plot exportable to Powerpoint with rotation controls
Import of Affy CEL files and apply RMA probe processing and quantile normalization
Extensible via R plug-in feature
Links genes to annotations in genomic databases
Tutorials and datasets
Some pro’s & con’s

• Much easier to learn
• User friendly (Excel interface)
• Freely available
• Good, uniform, documentation

• Less powerful than "raw" R and Bioconductor,
• Difficult to extend
  – If you miss something it's hard to get it
• Only available for Windows
Gene Annotations in Genomics Experiments
Biological preliminaries

- Every cell in the human body contains the entire human genome: 3.3 Gb or ~30K genes.
- The investigation of gene expression is meaningful because different cells, in different environments, doing different jobs express different genes.
- To-do list to create platforms for gene expression analysis:
  - Define what a gene is.
  - Identify genes in a sea of genomic DNA where <3% of DNA is contained in genes.
  - Design and implement probes that will effectively assay expression of ALL (most? many?) genes simultaneously.
  - Cross-reference these probes.
Cell Biology, Gene Expression and Microarray analysis

DNA
RNA
Protein
Gene: Protein coding unit of genomic DNA with an mRNA intermediate.

DNA Probe

mRNA

START protein coding

5' UTR AAAAAA 3' UTR

Genomic DNA

3.3 Gb

Sequence is a Necessity
From Genomic DNA to mRNA Transcripts

Genomic

mRNA variant 1

mRNA variant 2

5’ EST  →  ←  3’ EST

5’ EST  →  ←  3’ EST

5’ EST  →  ←  3’ EST

5’ EST  →  ←  3’ EST
Sequence and Gene databases

- Probes have to be mapped to databases.
- These may be either
  - gene or sequence databases
- Sequence databases:
  - From which sequence has the probe been synthesized?
- Gene databases
  - Which gene is the probe intended to interrogate
NCBI-Entrez
Entrez Gene

• Entrez database was constructed to replace the widely known and used LocusLink database in the year 2004.

• Entrez Gene integrates information from LocusLink and from genes annotated on Reference Sequences from completely sequenced genomes.

RefSeq

- The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.
- Similar to a review article, a RefSeq is a synthesis of information integrated across multiple sources at a given time. RefSeqs provide a foundation for uniting sequence data with genetic and functional information.
- They are generated to provide reference standards for multiple purposes ranging from genome annotation to reporting locations of sequence variation in medical records.

UniGene

- UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters.
- Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.
- In addition to sequences of well-characterized genes, hundreds of thousands novel expressed sequence tag (EST) sequences have been included. Consequently, the collection may be of use to the community as a resource for gene discovery.

Databases for microarrays
Microarray Data in a Nutshell

Lots of data to be managed before and after the experiment.

Data to be stored before the experiment.
- Description of the *array* and the *sample*.
- Direct access to all the cDNA and gene sequences, annotations, and physical DNA resources.

Data to be stored after the experiment
- Raw Data - scanned images.
- Gene Expression Matrix - Relative expression levels observed on various sites on the array.

Hence we can see that *database software capable of dealing with larger volumes of numeric and image data is required.*
Why Databases?

Tailored to datatype

Tailored to the Scientists

Intuitive ways to query the data

  - Diagrams, forms, point and click, text etc.

Support for efficient answering of queries.

  - Query optimisation, indexes, compact physical storage.
Gene Expression Databases Require Integration

- There are many different types of data presenting numerous relationships.
- There are a number of Databases with lots of information.
- Experiments need to be compared because the experiments are very difficult to perform and very expensive.
- Solution: Make all the databases talk the same language.
- XML was the choice of data interchange format.
Existing Microarray Databases

Several gene expression databases exist: Both commercial and non-commercial.

Most focus on either a particular technology or a particular organism or both.

Commercial databases:

- Rosetta Inpharmatics and Genelogic, the specifics of their internal structure is not available for internal scrutiny due to their proprietary nature.

Some non-commercial efforts to design more general databases merit particular mention.

We will discuss few of the most promising ones

- ArrayExpress - EBI
- The Gene expression Omnibus (GEO) - NLM
- The Standford microarray Database
- ExpressDB - Harvard
- Genex - NCGR
<table>
<thead>
<tr>
<th>Database</th>
<th>Organization</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAD</td>
<td>Stanford University/University of California at Berkeley, University of California at San Francisco (UCSF)</td>
<td>local installation</td>
</tr>
<tr>
<td>ArrayExpress</td>
<td>European Bioinformatics Institute (EBI)</td>
<td>public data deposition and public queries (coming soon)</td>
</tr>
<tr>
<td>ChipDB</td>
<td>Whitehead Institute for Biomedical Research/MIT Centre for Genome Research</td>
<td>public queries</td>
</tr>
<tr>
<td>Dragon</td>
<td>Johns Hopkins University</td>
<td>public queries</td>
</tr>
<tr>
<td>ExpressDB</td>
<td>Harvard University</td>
<td>public queries of E. coli and yeast data</td>
</tr>
<tr>
<td>GeneX</td>
<td>NCGG</td>
<td>local installation, public data deposition, and public queries of E. coli and yeast data</td>
</tr>
<tr>
<td>GeneDirector</td>
<td>BioDiscovery</td>
<td>local installation</td>
</tr>
<tr>
<td>GeNet</td>
<td>Silicon Genetics</td>
<td>local installation, public data deposition, and public queries</td>
</tr>
<tr>
<td>GEO</td>
<td>National Center for Biotechnology Information (NCBI)</td>
<td>public data deposition and public queries</td>
</tr>
<tr>
<td>GXD</td>
<td>The Jackson Laboratory</td>
<td>public data deposition and public queries of mouse data (coming soon)</td>
</tr>
<tr>
<td>mAdb</td>
<td>National Cancer Institute (NCI)</td>
<td>local installation</td>
</tr>
<tr>
<td>mAdbSQL</td>
<td>The University of Manchester</td>
<td>local installation</td>
</tr>
<tr>
<td>NOMAD</td>
<td>UCSF</td>
<td>local installation</td>
</tr>
<tr>
<td>RAD</td>
<td>University of Pennsylvania</td>
<td>public queries</td>
</tr>
<tr>
<td>Expression Connection</td>
<td>Stanford University/Saccharomyces Genome Database</td>
<td>public queries of yeast data</td>
</tr>
<tr>
<td>SMD</td>
<td>Stanford University</td>
<td>local installation and public queries</td>
</tr>
<tr>
<td>YMGV</td>
<td>Ecole Normale Superieure</td>
<td>public queries of yeast data</td>
</tr>
</tbody>
</table>
ArrayExpress

Public repository of microarray based gene expression data. Implemented in Oracle at EBI.

Contains:

- several curated gene expression datasets
- possible introduction of an image server to archive raw image data associated with the experiments.

Accepts submissions in MAGE-ML format via a web-based data annotation/submission tool called MIAMExpress.

A demo version of MIAMExpress is available at:
http://industry.ebi.ac.uk/~parkinso/subtool/subtype.html

Provides a simple web-based query interface and is directly linked to the Expression Profiler data analysis tool which allows expression data clustering and other types of data exploration directly through the web.
Gene Express Omnibus

The Gene Expression Omnibus is a gene expression database hosted at the National library of Medicine.

It supports four basic data elements:
- Platform (the physical reagents used to generate the data)
- Sample (information about the mRNA being used)
- Submitter (the person and organisation submitting the data)
- Series (the relationship among the samples).

It allows downloading of entire datasets, it has not ability to query the relationships.

Data are entered as tab delimited ASCII records, with a number of columns that depend on the kind of array selected.

Supports Serial Analysis of Gene Expression (SAGE) data.
ExpressDB
ExpressDB is a relational database containing yeast and E.coli RNA expression data.
It has been conceived as an example on how to manage that kind of data.
It allows web-querying or SQL-querying.
It is linked to an integrated database for functional genomics called Biomolecule Interaction Growth and Expression Database (BIGED).
BIGED is intended to support and integrate RNA expression data with other kinds of functional genomics data.
Survey of existing microarray systems

- A comparison of microarray databases

- [http://mybio.wikia.com/wiki/Microarray_databases](http://mybio.wikia.com/wiki/Microarray_databases)
The Microarray Gene Expression Database Group (MGED)

History and Future:
Founded at a meeting in November, 1999 in Cambridge, UK.
In May 2000 and March 2001: development of recommendations for microarray data annotations (MAIME, MAML).

MGED 2\textsuperscript{nd} meeting:
- establishment of a steering committee consisting of representatives of many of the worlds leading microarray laboratories and companies

MGED 4\textsuperscript{th} meeting in 2002:
- MAIME 1.0 will be published
- MAML/GEML and object models will be accepted by the OMG
- concrete ontology and data normalization recommendations will be published.

Information can be obtained from http://www.mged.org
The Microarray Gene Expression Database Group (MGED)

**Goals:**

- Facilitate the adoption of standards for DNA-array experiment annotation and data representation.
- Introduce standard experimental controls and data normalization methods.
- Establish gene expression data repositories.
- Allow comparison of gene expression data from different sources.
MGED Working Groups

Goals:

- MIAME: Experiment description and data representation standards - Alvis Brazma

- MAGE: Introduce standard experimental controls and data normalization methods - Paul Spellman. This group includes the MAGE-OM and MAGE-ML development.

- OWG: Microarray data standards, annotations, ontologies and databases - Chris Stoeckert

- NWG: Standards for normalization of microarray data and cross-platform comparison - Gavin Sherlock
References

URL:

- Tutorial on Information Management for Genome Level Bioinformatics, Paton and Goble, at VLDB 2001:
  http://www.dia.uniroma3.it/~vldbproc/#tutEuropea

- European Molecular Biology Network http://www.embnet.org/

- Univ. Manchester site (with relational version of Microarray data representation, and links to other sites)
  http://www.bioinf.man.ac.uk

- Database textbook with absolutely no bioinformatics coverage For Microarray Data
  http://linkage.rockefeller.edu/wli/microarray/