DNA Chip Technology

Benedikt Brors

Dept. Intelligent Bioinformatics Systems

German Cancer Research Center
Why DNA Chips?

- Functional genomics: get information about genes that is unavailable from sequence
- Understand how cells/organisms react to external stimuli
- Understand gene regulation networks
- Determine what makes the difference between healthy and diseased tissue
- Simply do 15,000 Northern Blots at a time
Comparison Northern blot ↔ DNA array

Northern Blot

library of arrayed DNA probes (known sequence)

DNA array
Functional Genomics

- There may be 100,000 different transcripts in human cells (± 50,000)
- We only have sound information on ≈ 12,000 genes
- All cells have the same genome, but there are more than 200 cell types in a single organism
- Gene expression determines the cell type (neuron, lymphocyte, fibroblast etc.) and directs development of an organism (by spatial/temporal patterns)
- DNA chip technology promises to solve such unanswered questions
Basic Biology

- Genes contain construction information
- All structure and function is made up by proteins
- mRNA is sort of ‘working copy’, containing design of one protein
- mRNA is transferred to cytoplasm where protein is made
More Schematically ...

Chromosome Genome → mRNA Transcriptome → Proteins Proteome

DNA chips measure transcript levels
DNA Chip Technology

- **Array**: Small glass slide, contains 100s to 10,000s of DNA fragments (‘spots’) on few cm$^2$

- Each DNA fragment will bind specifically a complementary DNA/RNA: ‘Hybridization’

- ‘Active’ (transcribed) genes can be extracted from cells/tissues, labeled and hybridized to the array ⇒ ‘active’ genes will light up on the array
DNA Chip Technology 2

Sample cells → mRNA → cDNA

Control cells → mRNA → cDNA
DNA Chip Technology 3
DNA Chip Technology 4

- Chip is read out by video camera

- Digitized image is analyzed by image analysis software

- Result: list of numbers

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>spot1</td>
<td>1,346.2</td>
<td>1,575.8</td>
</tr>
<tr>
<td>spot2</td>
<td>100,326.1</td>
<td>30,872.0</td>
</tr>
<tr>
<td>spot3</td>
<td>987.1</td>
<td>177.2</td>
</tr>
<tr>
<td>spot4</td>
<td>(…)</td>
<td>(…)</td>
</tr>
</tbody>
</table>

- N.B. the second column, ‘G’, is missing for one-color experiments
Competing Technologies

- Two systems: printed/‘spotted’ chips and on-chip synthesis

- For (A) mostly long DNA strands (500–3000 nt)

- For (B) only oligonucleotides (≤ 25 nt)
Printed/Spotted Chips

- Any DNA fragment can be put on the chip. Often, cDNA from libraries is used; to get a suitable amount for spotting, mass PCR has to be used. However, though performed less frequently, synthetic oligonucleotides can be used as well.

- For cDNA chips, a new PCR reaction has to be performed for every batch of chips, these are quite different from each other with respect to the amount of DNA bound in a certain spot.

- DNA fragments are transferred to the chip either by a spotting robot that transfers nanoliter quantities of liquid, or by an ink-jet like device.
• Spotted or printed chips are usually hybridized with two differently labeled mRNA preparations (i.e. their cDNA representation). By competitively hybridizing with two targets, the DNA amount in a single spot becomes less important (but not irrelevant!!).
On-chip synthesis of oligonucleotides

- Oligonucleotides are built up on the chip surface by sequentially elongating the growing chain with a single nucleotide. To determine the sequence of the final oligonucleotides on each position of the chip, a process called photolithography is used.

- As chemical yield of the stepwise elongation is limited, oligonucleotides can’t be grown to more than 25 nt length.
Photolithography

Lamp

Mask

Array
Light activated oligo synthesis

![Diagram showing light activated oligo synthesis]

**Light sensitive protection**

nucleoside

Light activation
On-chip synthesis of oligonucleotides

- Oligonucleotides are built up on the chip surface by sequentially elongating the growing chain with a single nucleotide. To determine the sequence of the final oligonucleotides on each position of the chip, a process called *photolithography* is used.

- As chemical yield of the stepwise elongation is limited, oligonucleotides can’t be grown to more than 25 nt length.

- Hybridization to short oligos is quite unspecific, thus a number of them has to be used to probe for a single gene (usually 12–25).

- Frequently, cross-hybridization occurs. To eliminate this effect, hybridization is compared with that of an oligo that bears a single mismatch.
cDNA vs. Oligo Chips

• long DNA strands are more specific than oligos:
  ★ cDNA chips: 1 (2,3 identical) spots per gene
  ★ oligo chips: many oligos per gene

• Oligo chips by on-chip synthesis: Affymetrix GeneChip™:
  ★ Single-color readout
  ★ approx. 20 oligos per gene
  ★ mismatched control for every oligo
  ★ sophisticated weighting and averaging over 20 oligo pairs
  ★ much of the information is proprietary
References


Measures of expression

- For cDNA chips, mostly the *ratio of expressions* is used:

\[
\text{ratio}_i = \frac{R_i}{G_i}
\]

- The logarithm of the ratios is symmetric around ratio=1 (no change with respect to control condition):

\[
\log\text{ratio}_i = \log \frac{R_i}{G_i} = \log R_i - \log G_i
\]

- Logratios to different bases of the logarithm (2, e, 10) are identical up to a constant factor:

\[
\log_2(x) = \log_{10}(x) \cdot \log_2(10)
\]
Distribution of ratios/logratios

**Distribution of ratios**

**Distribution of logratios**

- Frequency on the y-axis for ratios.
- Frequency on the y-axis for logratios.
- X-axis for ratios from 0 to 50.
- X-axis for logratios from -4 to 4.

- Frequency values for ratios range from 0 to 2500.
- Frequency values for logratios range from 0 to 400.
Measures of expression 2

- Ratios are independent of absolute signal intensity, i.e. $R_i/G_i = 20/10 = 2$ will give the same ratio as $R_i/G_i = 20,000/10,000$. Sometimes, values of $M$ and $A$ are used:

\[
M = \log \frac{R_i}{G_i} \quad \text{(logratio)}
\]

\[
A = \frac{1}{2} \log (R_i \cdot G_i) \quad \text{(average expression)}
\]

- For Affymetrix-type arrays, the signal intensities of the whole probe set have to be aggregated first. Affymetrix software (MAS) uses trimmed means:

\[
\text{AvgDiff} = \frac{1}{|A|} \sum_{j \in A} (PM_j - MM_j), \quad A \subset N
\]
Questions asked to microarray data: three case studies
Alizadeh et al.: Lymphoma

- Gene expression profiling of Diffuse Large B-Cell Lymphoma (DLBCL)
- Lymphoma is a blood cancer where *peripheral* blood cells degenerate and divide without control
- DLBCL is an aggressive form of this disease, originating from B-lymphocytes. Overall 5-year survival is about 40%.
- Current clinical risk factors are not sufficient.
Alizadeh et al.: Methods

- A special cDNA chip was used, the *Lymphochip*

- spotted cDNA array of approximately 17,000 clones related to Lymphocytes

- 42 samples of DLBCL were analyzed, plus additional samples of normal B cells and of related diseases

- mRNA from these samples was competitively hybridized against control mRNA, stemming from a pool of lymphoma cell line mRNA preparations

- Data were analyzed by clustering
Alizadeh et al.: Results 1
Alizadeh et al.: Results 2
Van’t Veer et al.: Breast cancer

- looks for prognostic markers in breast cancer
- two classes of patients: those with distant metastasis (other than in breast) within 5 years, and those without (also had negative lymph node status)
- In statistical thinking, this is a *classification* problem: given a set of *variables*, can we train a *classifier* such that it predicts for any new sample the *class* as correctly as possible?
Van’t Veer et al.: Methods

- A custom-made 25,000-clone chip was used; each feature contained a unique 60-mer oligonucleotide. This oligo was transferred to the chip by ink-jet likr printing.

- The chips were hybridized competitively; the reference mRNA was obtained from a pool of patient mRNA (98 patients in total).

- Only data from certain genes (231) were used; finding out informative genes is called feature selection in machine learning.

- A home-made ad hoc classification method was used (no details given here). You can do better with established classification methods (tought later in this course).

- The model was validated by cross validation and by an independent test set.
Cross-validation

class A       class B

1 2 3 4 5 6 7 8 9 10

Training set  Test set  run 1

1 2 3 4 5 6 7 8 9 10

Training set  Test set  run 2

Training ...  Test set  ... set  run 3
Van’t Veer et al.: Results 2

a

Sporadic breast tumours
patients <55 years
tumour size <5 cm
lymph node negative (LN0)

Prognosis reporter genes

Distant metastases
<5 years

No distant metastases
>5 years

b

Correlation to average
good prognosis profile

Metastases

Tumours
Shipp et al.: Lymphoma again

- published in *Nat. Med.* 8:68–74

- Same lymphoma (DLBCL) as in the study of Alizadeh et al. was investigated

- Samples from 58 patients with DLBCL were subjected to gene expression analysis

- Affymetrix chip was used (6,800 probe sets)

- A classification (supervised) approach was taken

- Results were compared with those of Alizadeh et al.
Shipp et al.: Results 1
Possible extension: Regression

- This was treated as a classification problem, i.e. there were distinct *classes* (cured vs. fatal) as *response variables*

- One could also use a *continuous* response variable: e.g. survival time, or the probability of being cured

- Fitting a model that predicts a continuous response is called *regression* in statistics (methods to be discussed later)