Lecture 2
Microarray Data Analysis

Bioinformatics Data Analysis and Tools

Purpose of lecture

- Introduce HTP gene expression and array comparative genomics hybridization (aCGH) data and analysis techniques
- Later BDAT lectures on data mining and clustering lectures will use microarray and aCGH data

Content

- Justification
- cDNA arrays
- Short oligonucleotide arrays (Affymetrix)
- Serial analysis of gene expression (SAGE)
- mRNA abundance and function
- Comparing expression profiles
- Eisen dataset
- Array CGH

A gene codes for a protein

DNA makes mRNA makes Protein

- If you want to measure gene activity, you should measure the protein concentration
- There are now protein chips, but the technique is in its infancy
- As a widely used alternative, researchers have developed ways to get an idea about the mRNA concentrations in a cell
- They have developed high throughput (HTP) techniques to measure (relative) mRNA concentrations

DNA makes mRNA makes Protein

Translation happens within the ribosome
DNA makes mRNA makes Protein

- How good a model is measuring mRNA levels for the concentration of the protein product?
- Competition of mRNA to get onto the ribosome is still not well understood
  - Ribosomes can be very busy, so you get a “waiting list” of mRNAs
  - This leads to time delays and a non-linear relation between mRNA and corresponding protein concentrations

Ribosome structure

- In the nucleolus, ribosomal RNA is transcribed, processed, and assembled with ribosomal proteins to produce ribosomal subunits
- At least 40 ribosomes must be made every second in a yeast cell with a 90-min generation time (Tollervey et al. 1991). On average, this represents the nuclear import of 1100 ribosomal proteins every second and the export of 80 ribosomal subunits out of the nucleus every second. Thus, a significant fraction of nuclear trafficking is used in the production of ribosomes.
- Ribosomes are made of a small (‘2’) in Figure) and a large subunit (‘1’ in Figure)

Genomics and transcriptome

- Following genome sequencing and annotation, the second major branch of genomics is analysis of the transcriptome
- The transcriptome is defined as the complete set of transcripts and their relative levels of expression in particular cells or tissues under defined conditions

‘The analysis of gene expression data is going to be a very important issue in 21st century statistics because of the clinical implications’

High-throughput measuring of gene expression data

Many different technologies, including
- High-density nylon membrane arrays
- cDNA arrays (Brown/Botstein)
- Short oligonucleotide arrays (Affymetrix)
- Serial analysis of gene expression (SAGE)
- Long oligo arrays (Agilent)
- Fibre optic arrays (Illumina)

Biological background

Transcription

DNA

GTATCCTC
CATTAGGAG

mRNA
GUAUCCC

Translation

Translation happens within the ribosome
Idea: measure the amount of mRNA to see which genes are being expressed in (used by) the cell. Measuring protein directly might be better, but is currently harder (see earlier slides).

Reverse transcription

Clone cDNA strands, complementary to the mRNA

<table>
<thead>
<tr>
<th>mRNA</th>
<th>GUAAUCCUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>cDNA</td>
<td>TTAGGAG</td>
</tr>
</tbody>
</table>

Transcriptome datasets

- cDNA microarrays
- Oligonucleotide arrays
  - Most suitable for contrasting expression levels across tissues and treatments of chosen subset of genome
- Serial analysis of gene expression (SAGE)
  - Relies on counting sequence tags to estimate absolute transcript levels, but less suited to replication

What is a microarray

- Slide or membrane with numerous probes that represent various genes of some biological species.
- Probes are either oligo-nucleotides that range in length from 25 to 60 bases, or cDNA clones with length from a hundred to several thousand bases.
- The array type corresponds to a list of reference genes on the microarray with annotations. For example: (1) 22K Agilent oligo array, and (2) NIA 15K cDNA membrane array. Many individual users want to add their own array types to the list.

What happens to a microarray

- Microarrays are hybridized with labeled cDNA synthesized from a mRNA-sample of some tissue.
- The intensity of label (radioactive or fluorescent) of each spot on a microarray indicates the expression of each gene.
  - One-dye arrays (usually with radioactive label) show the absolute expression level of each gene.
  - Two-dye arrays (fluorescent label only) can indicate relative expression level of the same gene in two samples that are labelled with different colours and mixed before hybridization. One of these samples can be a universal reference which helps to compare samples that were hybridized on different arrays.
Universal reference

- Universal reference is a mixture of cDNA that represents (almost) all genes of a species, while their relative abundance is standardized.
- Universal reference is synthesized from mRNA of various tissues.
- Universal reference can be used as a second sample for hybridization on 2-dye microarrays. Then all other samples become comparable via the universal reference.

cDNA microarrays

cDNA microarrays are used to compare the genetic expression in two samples of cells.

**PRINT**
- cDNA from one gene on each spot

**SAMPLES**
- cDNA labeled red/green with fluorescent dyes
- e.g. treatment / control
- normal / tumor tissue

**HYBRIDIZE**
- Add equal amounts of labeled cDNA samples to microarray.

**SCAN**
- Laser
- Detector measures ratio of induced fluorescence of two samples (Cy3 and Cy5 scanned separately (dye channels))

Sample is spread evenly over microarray, specific cDNAs then hybridize with their counterparts on the array, after which the sample is rinsed off to only leave hybridized sample.

cDNA microarray experiments

mRNA levels compared in many different contexts

- Different tissues, same organism (brain versus liver)
- Same tissue, same organism (treatment vs. control, tumor vs. non-tumor)
- Same tissue, different organisms (wildtype vs. knock-out, transgenic, or mutant)
- Time course experiments (effect of treatment, development)
- Other special designs (e.g. to detect spatial patterns).
Replication

- An independent repeat of an experiment.
- In practice it is impossible to achieve absolute independence of replicates. For example, the same researcher often does all the replicates, but the results may differ in the hands of another person.
- But it is very important to reduce dependency between replicates to a minimum. For example, it is much better to take replicate samples from different animals (these are called biological replicates) than from the same animal (these would be technical replicates), unless you are interested in a particular animal.
- If sample preparation requires multiple steps, it is best if samples are separated from the very beginning, rather than from some intermediate step. Each replication may have several subreplications (=technical replications).

Some statistical questions

- Planning of experiments:
  - Design, sample size
  - Selection of genes relevant to any given analysis
- Image analysis:
  - Addressing, segmenting, quantifying
  - Quality: of images, of spots, of (log) ratios
- Normalisation: within and between slides
- Biological analysis:
  - Which genes are (relatively) up/down regulated?
  - Assigning p-values to tests/confidence to results.
  - Analysis of time course, factorial and other special experiments & much more
- Discrimination and allocation of samples
- Clustering, classification: of samples, of genes

Some bioinformatic questions

- Connecting spots to databases, e.g. to sequence, structure, and pathway databases
- Discovering short sequences regulating sets of genes: direct and inverse methods
- Relating expression profiles to structure and function, e.g. protein localisation, co-expression, etc.
- Identifying novel biochemical or signalling pathways, ..........and much more.

What types of things can go wrong?

- Spot size variances
- Dye labeling efficiency differences (performing dye swap experiments and/or improving dye labeling protocols help)
- Positional biases (can be due to print tip, spot drying time dependencies, hybridizations not being uniform, etc.)
- Plate biases
- Variance in background (dye sticking to the array, dust, hairs, defects in the array coating, etc.)
- Scanner non-linearities
- Sample biases (e.g. contamination of DNA in your RNA sample, sample handling, storage, and preparation protocol variances)

...with automatically scanning the microarrays
Does one size fit all?

Segmentation: limitation of the fixed circle method

Inside the boundary is spot (foreground), outside is not – Background pixels are those immediately surrounding circle/segment boundary

Identify differentially expressed genes

When calculating relative expression levels, one loses sense of absolute concentrations (numbers) of cDNA molecules

O n p genes for n slides: p is O(10,000), n is O(10-100), but growing

<table>
<thead>
<tr>
<th>Slides</th>
<th>slide 1</th>
<th>slide 2</th>
<th>slide 3</th>
<th>slide 4</th>
<th>slide 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
<td>0.30</td>
<td>0.80</td>
<td>1.51</td>
<td>0.90</td>
</tr>
<tr>
<td>2</td>
<td>-0.10</td>
<td>0.49</td>
<td>0.24</td>
<td>0.06</td>
<td>0.46</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
<td>0.74</td>
<td>0.04</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>4</td>
<td>-0.45</td>
<td>-1.03</td>
<td>-0.79</td>
<td>0.56</td>
<td>-0.32</td>
</tr>
<tr>
<td>5</td>
<td>-0.06</td>
<td>1.06</td>
<td>1.35</td>
<td>1.09</td>
<td>-1.09</td>
</tr>
</tbody>
</table>

Gene expression level of gene 5 in slide 4

\[ \log_2 \left( \frac{\text{Red intensity}}{\text{Green intensity}} \right) \]

These values are conventionally displayed on a red (>0) yellow (0) green (<0) scale.

Quantification of expression

For each spot on the slide we calculate

\[ \text{Red intensity} = R_{fg} - R_{bg} \]

\[ \text{Green intensity} = G_{fg} - G_{bg} \]

and combine them in the log (base 2) ratio

\[ \log_2 \left( \frac{\text{Red intensity}}{\text{Green intensity}} \right) \]

The red/green ratios can be spatially biased

Top 2.5% of ratios red, bottom 2.5% of ratios green
The red/green ratios can be intensity-biased if one dye is under-incorporated relative to the other. The red/green ratios can be intensity-biased if one dye is under-incorporated relative to the other.

\[ M = \log_2 R/G = \log_2 R - \log_2 G \]

\[ A = \log_2 (\sqrt{R \times G}) = \frac{\log_2 R + \log_2 G}{2} \]

Values should scatter about zero.

How we “fix” the previous dye bias problem: Normalisation

• Normalise using housekeeping genes that are supposed to be present in constant concentrations
  • Shift data to M=0 level for selected housekeeping genes
  • Problem: which genes to select?
• Dye swapping (flipping), taking average value (normal and flipped)
• LOWESS (LOcally WEighted Scatterplot smoothing) normalisation. Also called LOESS transformation.
  • Calculate smooth curve m(A) through data points and take M – m(A) as normalised values

Normalization: how we “fix” the previous problem
Loess transformation (Yang et al., 2001)

Normalising: before

The curved line becomes the new zero line

Normalising: after

Table:

<table>
<thead>
<tr>
<th>Red</th>
<th>Green</th>
<th>Diff</th>
<th>R(G/R)</th>
<th>Log2R</th>
<th>Norm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>16500</td>
<td>15104</td>
<td>-1396</td>
<td>0.915</td>
<td>-0.128</td>
<td>-0.048</td>
</tr>
<tr>
<td>357</td>
<td>158</td>
<td>-199</td>
<td>0.443</td>
<td>-1.175</td>
<td>-1.095</td>
</tr>
<tr>
<td>8250</td>
<td>8025</td>
<td>-225</td>
<td>0.973</td>
<td>-0.039</td>
<td>0.040</td>
</tr>
<tr>
<td>978</td>
<td>836</td>
<td>-142</td>
<td>0.855</td>
<td>-0.226</td>
<td>-0.146</td>
</tr>
<tr>
<td>65</td>
<td>89</td>
<td>24</td>
<td>1.369</td>
<td>0.453</td>
<td>0.533</td>
</tr>
<tr>
<td>684</td>
<td>1368</td>
<td>539</td>
<td>2.000</td>
<td>1.000</td>
<td>1.080</td>
</tr>
<tr>
<td>13772</td>
<td>11209</td>
<td>-2563</td>
<td>0.814</td>
<td>-0.297</td>
<td>-0.217</td>
</tr>
<tr>
<td>856</td>
<td>731</td>
<td>-125</td>
<td>0.854</td>
<td>-0.228</td>
<td>-0.148</td>
</tr>
</tbody>
</table>
Analysis of Variance (ANOVA) approach

- ANOVA is a robust statistical procedure
- Partitions sources of variation, e.g. whether variation in gene expression is less in subset of data than in total data set
- Requires moderate levels of replication (4-10 replicates of each treatment)
- But no reference sample needed
- Expression judged according to statistical significance instead of by adopting arbitrary thresholds

Contributions to measured gene expression level

\[ Y_{ijkg} = \mu + A_i + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijkg} \]

All these noise effects (grey, blue) are taken into account to discern the best possible signal (yellow)

Analysis of Variance (ANOVA) approach has two steps

- Raw fluorescence data is log-transformed and arrays and dye channels are normalised with respect to one another. You get normalised expression levels where dye and array effects are eliminated
- A second model is fit to normalised expression levels associated with each individual gene

Analysis of Variance (ANOVA) approach

- Advantage: design does not need reference samples
- Concern: treatments should be randomised and all single differences between treatments should be covered

E.g., if male kidney and female liver are contrasted on one set, and female kidney and male liver on another, we cannot state whether gender or tissue type is responsible for expression differences observed
Analysis of Variance (ANOVA) experimental microarray setups

• Loop design of experiments possible: A-B, B-C, C-D, and D-A
• Flipping of dyes (dye swap) to filter artifacts due to preferential labeling
  • Repeating hybridization on two-dye microarrays with the same samples but swapped fluorescent labels.
  • For example, sample A is labeled with Cy3 (green) and sample B with Cy5 (red) in the first array, but sample A is labeled with Cy5 and sample B with Cy3 in the second array.
• Dye swap is used to remove technical colour bias in some genes. Dye swap is a technical replication (=subreplication).
• Completely or partially randomised designs

Oligonucleotide arrays

• Affymetrix GeneChip
• No cDNA library but 25-mer oligonucleotides
• Oligomers designed by computer program to represent known or predicted open reading frames (ORFs)

Oligonucleotide arrays

• Up to 25 oligos designed for each exon, expression is only inferred if hybridization occurs with (almost) all of them
• Each oligo printed on chip adjacent to (single base pair) mismatch oligo
• Match/mismatch oligos used to calculate signal intensity and then expression level

ATGCCCTGGGCGTTAGAAGCTTTAC
ATGCCCTGGGCGTGAAAAGCTTTAC
  • But: not everybody agrees with Affymetrix mismatch strategy: is it biologically relevant?

SAGE

• SAGE = Serial Analysis of Gene Expression
• Based on serial sequencing of 10 to 14-bp tags that are unique to each and every gene
• SAGE is a method to determine absolute abundance of every transcript expressed in a population of cells
• Because SAGE does not require a preexisting clone (such as on a normal microarray), it can be used to identify and quantitate new genes as well as known genes.
SAGE

- A short sequence tag (10-14bp) contains sufficient information to uniquely identify a transcript provided that the tag is obtained from a unique position within each transcript;
- Sequence tags can be linked together to form long serial molecules (strings) that can be cloned and sequenced; and
- Counting the number of times a particular tag is observed in the string provides the expression level of the corresponding transcript.
- A list of each unique tag and its abundance in the population is assembled
- An elegant series of molecular biology manipulations is developed for this

Some of the steps of SAGE

1. Trap RNAs with beads
2. Convert the RNA into cDNA
3. Make a cut in each cDNA so that there is a broken end sticking out
4. Attach a "docking module" to this end; here a new enzyme can dock, reach down the molecule, and cut off a short tag
5. Combine two tags into a unit, a di-tag
6. Make billions of copies of the di-tags (using a method called PCR)
7. Remove the modules and glue the di-tags together into long concatamers
8. Put the concatamers into bacteria and copy them millions of times
9. Pick the best concatamers and sequence them
10. Use software to identify how many different cDNAs there are, and count them;
11. Match the sequence of each tag to the gene that produced the RNA.

Trap RNA with beads

- Unlike other molecules, most messenger RNAs end with a long string of "As" (A stands for the nucleotide adenine.)
- This allows researchers to trap them. Adenine forms very strong chemical bonds with another nucleotide, thymine (T). A molecule that consists of 20 or so Ts acts like a chemical bait to capture RNAs.
- Researchers coat microscopic, magnetic beads with "TTTTT" tails hanging out.
- When the contents of cells are washed past the beads, the RNA molecules will be trapped. A magnet is used to withdraw the bead and the RNAs out of the "soup".

Concatamer

Example of a concatemer:

```
ATCTGAGTTGCAGCACTTCCCGCTACAATCTGAGTTCTAGGACGAGG ...
```

A computer program generates a list of tags and tells how many times each one has been found in the cell:

<table>
<thead>
<tr>
<th>Tag Sequence</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATCTGAGTTGC</td>
<td>1075</td>
</tr>
<tr>
<td>GAGCGACATT</td>
<td>125</td>
</tr>
<tr>
<td>TCCTCCCTACA</td>
<td>112</td>
</tr>
<tr>
<td>TAGACGGGC</td>
<td>92</td>
</tr>
<tr>
<td>GAGATGCGGC</td>
<td>91</td>
</tr>
<tr>
<td>TAGCCAGAT</td>
<td>83</td>
</tr>
<tr>
<td>GAGGTTGTTA</td>
<td>80</td>
</tr>
<tr>
<td>GGTATTTTTC</td>
<td>66</td>
</tr>
<tr>
<td>TAGCTTCTCA</td>
<td>66</td>
</tr>
<tr>
<td>TCCCTTACAT</td>
<td>66</td>
</tr>
<tr>
<td>GGATACACAT</td>
<td>55</td>
</tr>
<tr>
<td>AKGTTCGCG</td>
<td>54</td>
</tr>
<tr>
<td>CAGACCGCCG</td>
<td>50</td>
</tr>
<tr>
<td>GAGGCGGGCCG</td>
<td>48</td>
</tr>
</tbody>
</table>

The next step is to identify the RNA and the gene that produced each of the tags:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Sequence</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATATTGCTCA</td>
<td>5</td>
<td>translation elongation factor 1 gamma</td>
</tr>
<tr>
<td>AAGGTCGGAA</td>
<td>2</td>
<td>T-complex protein 1, z-subunit</td>
</tr>
<tr>
<td>ACCCTCCCTC</td>
<td>1</td>
<td>no match</td>
</tr>
<tr>
<td>GCTTGTITTA</td>
<td>61</td>
<td>rps12 mRNA fragment for ribosomal proteins</td>
</tr>
<tr>
<td>GAGGACAGATC</td>
<td>45</td>
<td>subunit 5.2-A extension protein</td>
</tr>
<tr>
<td>CCCTGCTTGGG</td>
<td>9</td>
<td>SFI protein (SFI gene)</td>
</tr>
<tr>
<td>TTTTGTGTAA</td>
<td>90</td>
<td>NASH dehydrogenase 3 (ND3) gene</td>
</tr>
<tr>
<td>GCAAACCCGCG</td>
<td>63</td>
<td>rpl.23</td>
</tr>
<tr>
<td>GGAGCCCGCC</td>
<td>45</td>
<td>ribosomal protein L18a</td>
</tr>
<tr>
<td>GGGCGCCGCA</td>
<td>34</td>
<td>ribosomal protein S31</td>
</tr>
<tr>
<td>GCCGAACTTG</td>
<td>30</td>
<td>ribosomal protein S5 homolog (Mrj155D)</td>
</tr>
<tr>
<td>TACGGACGCG</td>
<td>4</td>
<td>hscDNA GM412730</td>
</tr>
</tbody>
</table>

SAGE issues

- At least 50,000 tags are required per sample to approach saturation, the point where each expressed gene (e.g. human cell) is represented at least twice (and on average 10 times)
- Expensive: SAGE costs about $5000 per sample
- Too expensive to do replicated comparisons as is done with microarrays
SAGE quantitative comparison

• A tag present in 4 copies in one sample of 50,000 tags, and in 2 copies in another sample, may be twofold expressed but is not going to be significant
• Even 20 to 10 tags might not be statistically significant given the large numbers of comparisons
• Often, 10-fold over- or under-expression is taken as threshold

SAGE quantitative comparison

• A great advantage of SAGE is that the method is unbiased by experimental conditions
• Direct comparison of data sets is possible
• Data produced by different groups can be pooled
• Web-based tools for performing comparisons of samples all over the world exist (e.g. SAGEnet and xProfiler)

Transcript abundance in typical eukaryotic cell

as measured by SAGE

• <100 transcripts account for 20% of total mRNA population, each being present in between 100 and 1000 copies per cell
• These encode ribosomal proteins and other core elements of transcription and translation machinery, histones and further taxon-specific genes

General, basic and most important cellular mechanisms

Transcript abundance in typical eukaryotic cell (2)

• Several hundred intermediate-frequency transcripts, each making 10 to 100 copies, make up for a further 30% of mRNA
• These code for housekeeping enzymes, cytoskeletal components and some unusually abundant cell-type specific proteins

Pretty basic housekeeping things

Transcript abundance in typical eukaryotic cell (3)

• Further 50% of mRNA is made up of tens of thousands low-abundance transcripts (<10), some of which may be expressed at less than one copy per cell (on average)
• Most of these genes are tissue-specific or induced only under particular conditions

Specific or special purpose products

Transcript abundance in typical eukaryotic cell (4)

Get some feel for the numbers (can be a factor 2 off but order of magnitude about right)


If

• ~80 transcripts * ~400 copies = 32,000 (20%)
• ~600 transcripts * ~75 copies = 45,000 (30%)
• 25,000 transcripts * ~3 copies = 75,000 (50%)
• Then Total = 150,000 mRNA molecules
Transcript abundance in typical eukaryotic cell (5)

- This means that most of the transcripts in a cell population contribute less than 0.01% of the total mRNA.
- Say 1/3 of higher eukaryote genome is expressed in given tissue, then about 10,000 different tags should be detectable.
- Taking into account that half the transcriptome is relatively abundant, at least 50,000 different tags should be sequenced to approach saturation (so to get at least 10 copies per transcript on average).

SAGE analysis of yeast (Velculesco et al., 1997)

Some statistical research stimulated by microarray data analysis

- Experimental design: Churchill & Kerr
- Image analysis: Zuzan & West, ….
- Data visualization: Cart et al
- Estimation: Ideker et al, ….
- Multiple testing: Westfall & Young, Storey, ….
- Discriminant analysis: Golub et al.,….
- Clustering: Hastie & Tibshirani, Van der Laan, Fridlyand & Dudoit, ….
- Empirical Bayes: Efron et al, Newton et al,…..
- Multiplicative models: Li & Wong
- Multivariate analysis: Alter et al
- Genetic networks: D'Haezeleer et al and more

How do we assess microarray data

\[ z = \frac{M - \mu}{\sigma}, \text{ where } \mu \text{ is mean and } \sigma \text{ is standard deviation.} \]
This leads to zero mean and unit standard deviation

- If M normally distributed, then probability that \( z \) lies outside range -1.96 < \( z \) < 1.96 is 5%.
- There is evidence that log(R/G) ration are normally distributed. Therefore, R/G is said to be lognormally distributed.
Example 1
Breast tumor classification

Dutch Cancer Institute (NKI)
Prediction of clinical outcome of breast cancer
DNA microarray experiment
117 patients
25000 genes

78 sporadic breast tumors
70 prognostic markers genes

Validation set: 2 out of 19 incorrect

Is there work to do?
• What is the minimum number of genes required in these classification models (to avoid chance classification)
• What is the maximum number of genes (avoid overfitting)
• What is the relation to the number of samples that must be measured?
• Rule of thumb: minimal number of events per variable (EPV)>10
  – NKI study ~35 tumors (events) in each group → 35/10=3.5 genes should maximally have been selected (70 were selected in the breast cancer study) → overfitting? Is the classification model adequate?

Example 2
Apo AI experiment (Callow et al 2000, LBNL)

Goal. To identify genes with altered expression in the livers of Apo AI knock-out mice (T) compared to inbred C57Bl/6 control mice (C). Apo-lipoproteins are involved in lipid transport.
• 8 treatment mice and 8 control mice
• 16 hybridizations: liver mRNA from each of the 16 mice (T_i, C_i) is labelled with Cy5, while pooled liver mRNA from the control mice (C*) is labelled with Cy3.
• Probes: ~ 6,000 cDNAs (genes), including 200 related to lipid metabolism.

Example 3
Leukemia experiments (Golub et al 1999, WI)

Goal. To identify genes which are differentially expressed in acute lymphoblastic leukemia (ALL) tumours in comparison with acute myeloid leukemia (AML) tumours.
• 38 tumour samples: 27 ALL, 11 AML.
• Data from Affymetrix chips, some pre-processing.
• Originally 6,817 genes; 3,051 after reduction.

Data therefore 3,051 × 38 array of expression values.

Acute lymphoblastic leukemia (ALL) is the most common malignancy in children 2-5 years in age, representing nearly one third of all pediatric cancers.
Acute Myeloid Leukemia (AML) is the most common form of myeloid leukemia in adults (chronic lymphocytic leukemia is the most common form of leukemia in adults overall). In contrast, acute myeloid leukemia is an uncommon variant of leukemia in children. The median age at diagnosis of acute myeloid leukemia is 65 years of age.
Genome-Wide Cluster Analysis

Eisen dataset

- Eisen et al., PNAS 1998
- S. cerevisiae (baker’s yeast)
  - all genes (~ 6200) on a single array
  - measured during several processes
- human fibroblasts
  - 8600 human transcripts on array
  - measured at 12 time points during serum stimulation

The Eisen Data

- 79 measurements for yeast data
- collected at various time points during
  - diauxic shift (shutting down genes for metabolizing sugars, activating those for metabolizing ethanol)
  - mitotic cell division cycle
  - sporulation
  - temperature shock
  - reducing shock

The Data

- each measurement represents
  \[ \log(\text{Red}/\text{Green}_i) \]
  where red is the test expression level, and green is the reference level for gene \( G \) in the \( i \)th experiment
- the expression profile of a gene is the vector of measurements across all experiments \([G_1, G_n]\)

The Data

- \( m \) genes measured in \( n \) experiments:
  \[
  \begin{array}{c}
  g_{1,1} \ldots g_{1,n} \\
  g_{2,1} \ldots g_{2,n} \\
  \vdots \\
  g_{m,1} \ldots g_{m,n}
  \end{array}
  \]
  Vector for 1 gene

The Task

Gene Similarity Metric

- to determine the similarity of two genes \( X, Y \)
  \[
  S(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{X_i - X_{\text{offset}}}{\Phi_X} \right) \left( \frac{Y_i - Y_{\text{offset}}}{\Phi_Y} \right)
  \]
- \( \Phi_G = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (G_i - G_{\text{offset}})^2} \)
  \( G_{\text{offset}} \) and \( Y_{\text{offset}} \) are the mean values over the expression levels \( X \) and \( Y \), respectively.

This is called ‘correlation coefficient with centering’
**Gene Similarity Metric**

- since there is an assumed reference state (the gene’s expression level didn’t change), $G_{dist}$ is set to 0 for all genes

\[
S(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{\sqrt{\sum_{i=1}^{n} X_i^2}} \left| \frac{Y_i}{\sqrt{\sum_{i=1}^{n} Y_i^2}} \right|
\]

\[
p = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \quad \text{Basic correlation coefficient}
\]

**Similarity measures for expression profiles**

- $S(X, Y) = \sum (X_i - \mu_i)(Y_i - \nu_i)/((\sum (X_i - \mu_i)^2)(\sum (Y_i - \nu_i)^2))$  
  Correlation coefficient with centering

- $S(X, Y) = \sum X_i Y_i/((\sum X_i^2)(\sum Y_i^2))$  
  Correlation coefficient (without centering)

- $S(X, Y) = (\sum (X_i - \mu_i)^2)^{1/2}$  
  Euclidean distance

- $S(X, Y) = \sum |X_i - Y_i|$  
  Manhattan (City-block) distance

\(\Sigma\) is the summation over \(i = 1, n\)
\(\mu_i\) is the mean value of \(X_i, X_2, \ldots, X_n\)

**Eisen et al. cDNA array results**

- redundant representations of genes cluster together
  - but individual genes can be distinguished from related genes by subtle differences in expression
- genes of similar function cluster together
  - e.g. 126 genes strongly down-regulated in response to stress

**Eisen et al. Results**

- 126 genes down-regulated in response to stress
  - 112 of the genes encode ribosomal and other proteins related to translation
  - agrees with previously known result that yeast responds to favorable growth conditions by increasing the production of ribosomes

**Partitional Clustering**

- divide instances into disjoint clusters
  - flat vs. tree structure
- key issues
  - how many clusters should there be?
  - how should clusters be represented?
Partitional Clustering from a Hierarchical Clustering

we can always generate a partitional clustering from a hierarchical clustering by “cutting” the tree at some level

$K$-Means Clustering

- assume our instances are represented by vectors of real values
- put $k$ cluster centers in same space as instances
- now iteratively move cluster centers

$K$-Means Clustering

- each iteration involves two steps:
  - assignment of instances to clusters
  - re-computation of the means

$K$-Means Clustering

- in $k$-means clustering, instances are assigned to one and only one cluster
- can do “soft” $k$-means clustering via Expectation Maximization (EM) algorithm
  - each cluster represented by a normal distribution
  - E step: determine how likely is it that each cluster “generated” each instance
  - M step: move cluster centers to maximize likelihood of instances

Array-CGH (Comparative Genomics Hybridisation)

- New microarray-based method to determine local chromosomal copy numbers
- Gives an idea how often pieces of DNA are copied
- This is very important for studying cancers, which have been shown to often correlate with copy events!
- Also referred to as ‘a-CGH’
Tumor Cell

Chromosomes of tumor cell:

Example of a-CGH Tumor

Clones/Chromosomes ⇒

a-CGH vs. Expression

a-CGH
- DNA
  - In Nucleus
  - Same for every cell
- DNA on slide
- Measure Copy Number Variation

Expression
- RNA
  - In Cytoplasm
  - Different per cell
- cDNA on slide
- Measure Gene Expression

CGH Data

Clones/Chromosomes ⇒

Algorithms for Smoothing Array CGH data

Kees Jong (VU, CS and Mathematics)
Elena Marchiori (VU, CS)
Aad van der Vaart (VU, Mathematics)
Gerrit Meijer (VUMC)
Bauke Ylstra (VUMC)
Marjan Weiss (VUMC)

Naïve Smoothing
Discrete Smoothing

Copy numbers are integers

Why Smoothing?

- Noise reduction
- Detection of Loss, Normal, Gain, Amplification
- Breakpoint analysis

Recurrent (over tumors) aberrations may indicate:
- an oncogene or
- a tumor suppressor gene

Is Smoothing Easy?

- Measurements are relative to a reference sample
- Printing, labeling and hybridization may be uneven
- Tumor sample is inhomogeneous
  - do expect only few levels
  - vertical scale is relative

Problem Formalization

A smoothing can be described by
- a number of breakpoints
- corresponding levels

A fitness function scores each smoothing according to fitness to the data

An algorithm finds the smoothing with the highest fitness score.

Breakpoint Detection

- Identify possibly damaged genes:
  - These genes will not be expressed anymore
- Identify recurrent breakpoint locations:
  - Indicates fragile pieces of the chromosome
- Accuracy is important:
  - Important genes may be located in a region with (recurrent) breakpoints
**Fitness Function**

We assume that data are a realization of a Gaussian noise process and use the maximum likelihood criterion adjusted with a penalization term for taking into account model complexity.

We could use better models given insight in tumor pathogenesis.

**Fitness Function (2)**

CGH values: $x_1, \ldots, x_n$

breakpoints: $0 < y_1 < \ldots < y_N < x_N$

levels: $\mu_1, \ldots, \mu_N$

error variances: $\sigma_1^2, \ldots, \sigma_N^2$

likelihood:

$$
\prod_{i=1}^{y_1} \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x_i - \mu_1}{\sigma_1} \right)^2} \ldots \prod_{i=y_{N-1}+1}^{\infty} \frac{1}{\sigma_N \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x_i - \mu_N}{\sigma_N} \right)^2}
$$

**Fitness Function (3)**

Maximum likelihood estimators of $\mu$ and $\sigma^2$ can be found explicitly.

Need to add a penalty to log likelihood to control number $N$ of breakpoints

$$
f(y_1, \ldots, y_N) = \sum_{i=1}^{N+1} (y_{i+1} - y_i) \log \sigma_i + \lambda N
$$

**Algorithms**

Maximizing Fitness is computationally hard

Use genetic algorithm + local search to find approximation to the optimum

**Algorithms: Local Search**

choose $N$ breakpoints at random

while (improvement)

- randomly select a breakpoint
- move the breakpoint one position to left or to the right
Genetic Algorithm

A genetic algorithm is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Generally, a genetic algorithm starts with a population of candidate solutions (called individuals, creatures, or phenotypes). Solution candidates may be represented by binary strings, numbers, or any other suitable data structure. The process of applying genetic algorithms includes:

1. Selection of individuals (or parents) from the current population for reproduction. The selection is usually proportional to fitness (solution quality), which is determined by a fitness function. The fitness function assigns each individual a value that reflects how well the problem has been solved by that individual.
2. Reproduction (also called crossover or genetic crossover) is where a new population is created using the selected individuals. This is done by combining the genetic information (solution candidates) of the selected individuals to form candidate solutions for the next generation.
3. Mutation is a process where random changes are made to the genetic information within the new population. This is done to maintain genetic diversity among the solutions in the new population and to avoid premature convergence.
4. The new population formed in the previous steps replaces the current population.

This process is repeated until a termination condition is satisfied. The termination condition is defined by the problem to be solved, the solution algorithm, or a user-defined number of generations.

Conclusion

• Breakpoint identification as model fitting to search for most-likely-fit model given the data
• Genetic algorithms + local search perform well
• Results comparable to those produced by hand by the local expert
• Future work:
  – Analyse the relationship between Chromosomal aberrations and Gene Expression

Comparison to Expert

Algorithm vs. Expert

Breakpoint Detection

• Identify possibly damaged genes:
  – These genes will not be expressed anymore
• Identify recurrent breakpoint locations:
  – Indicates fragile pieces of the chromosome
• Accuracy is important:
  – Important genes may be located in a region with (recurrent) breakpoints