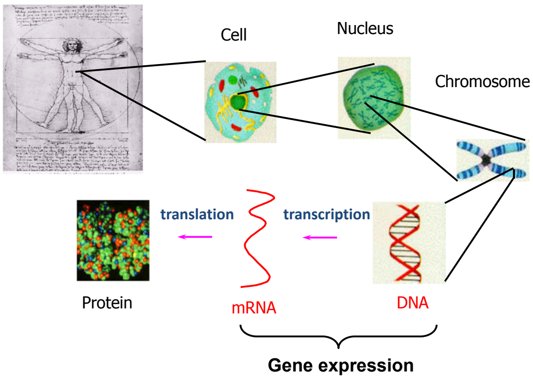
COSC 348: Computing for Bioinformatics

Lecture 16:
Microarray data analysis: introduction

Lubica Benuskova

http://www.cs.otago.ac.nz/cosc348/

Microarrays measure gene expression



Cell

Nucleus

Chromosome

translation

transcription

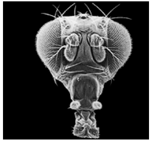
Protein

mRNA

DNA

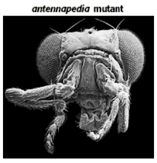
Gene expression

wild-type *drosophila*



Gene Expression

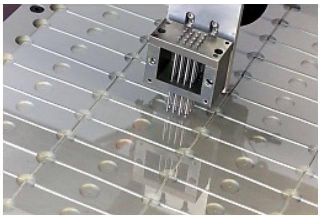
antennapedia mutant



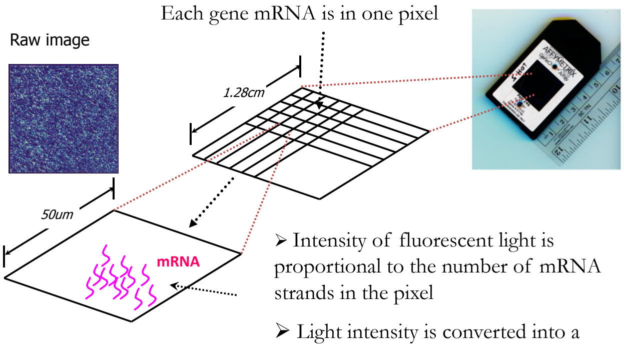
- Cells are different because of **differential gene expression** *(i.e. in different body organs different genes are expressed)*.
- Gene is expressed by transcribing DNA into **many copies** of mRNA.
- mRNAs are then **translated** into protein molecules.
- **Microarrays measure the level of mRNA (i.e. concentration of mRNA), and thus the level of gene expression.**

Principle of microarrays

- mRNA levels are proportional to the rate (or level) of gene expression (*how many copies of mRNA are produced by a gene*)
- mRNA is isolated from cells and labeled with a fluorescent dye.
- **Level of mRNA is proportional to intensity of fluorescent light emission, which is measured.**



Affymetrix microarrays



Each gene mRNA is in one pixel

Raw image

1.28cm

50um

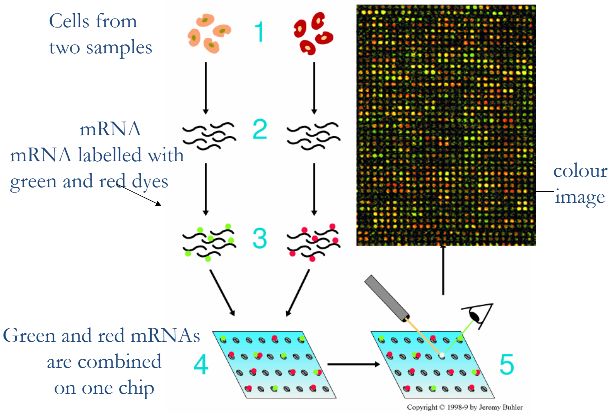
mRNA

➢ Intensity of fluorescent light is proportional to the number of mRNA strands in the pixel

➢ Light intensity is converted into a number by a special reading device

http://www.affymetrix.com

Spotted or cDNA microarrays



Cells from two samples

1

mRNA

mRNA labelled with green and red dyes

2

3

colour image

Green and red mRNAs are combined on one chip

4

5

Copyright © 1998-9 by Jeremy Buhler

Image analysis in spotted arrays

- The 2 fluorescence images are overlaid, with the colours combined (red and green make yellow in the RGB scheme).
- **Yellow spots** indicate genes which were equally expressing in both conditions/ cell groups (i.e. **both genotypes**)
- **Green spots** indicate genes which were only expressing in the control condition (e.g., **wild-type genotype**).
- **Red spots** indicate genes which were only expressing in the treatment condition (e.g., **mutant genotype**).

Fold changes in spotted arrays

- Differential expression at a spot is often reported as a fold change:

$$\textit{Fold change} = \frac{\textit{red intensity}}{\textit{green intensity}}$$

- In spotted arrays too, the light intensity is converted into a numerical value (fold change) by a special equipment.
- Often $\log_2$ scale is used:

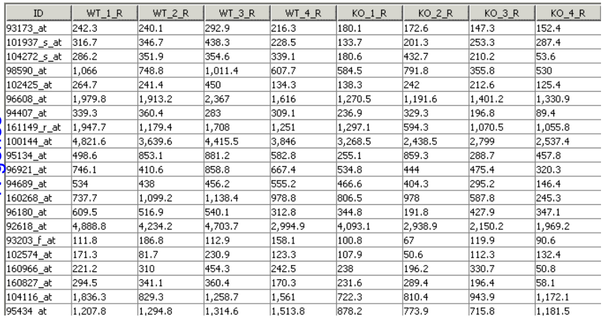
$$\log_2(\textit{Fold change}) = \log_2\left(\frac{\textit{red intensity}}{\textit{green intensity}}\right) =$$

$$= \log_2(\textit{red intensity}) - \log_2(\textit{green intensity})$$

Microarray data – N x M matrix of numbers

M samples / subjects

N genes



ID	WT_1_R	WT_2_R	WT_3_R	WT_4_R	KO_1_R	KO_2_R	KO_3_R	KO_4_R
93173_at	242.3	240.1	292.9	216.3	180.1	172.6	147.3	152.4
101937_s_at	316.7	346.7	438.3	228.5	133.7	201.3	253.3	287.4
104272_s_at	286.2	351.9	354.6	339.1	180.6	432.7	210.2	53.6
98590_at	1,066	748.8	1,011.4	607.7	584.5	791.8	355.8	530
102425_at	264.7	241.4	450	134.3	138.3	242	212.6	125.4
96608_at	1,979.8	1,913.2	2,367	1,616	1,270.5	1,191.6	1,401.2	1,330.9
94407_at	339.3	360.4	283	309.1	236.9	329.3	196.8	89.4
161149_r_at	1,947.7	1,179.4	1,708	1,251	1,297.1	594.3	1,070.5	1,055.8
100144_at	4,821.6	3,639.6	4,415.5	3,846	3,268.5	2,438.5	2,799	2,537.4
95134_at	498.6	853.1	881.2	582.8	255.1	859.3	288.7	457.8
96921_at	746.1	410.6	858.8	667.4	534.8	444	475.4	320.3
94689_at	534	438	456.2	555.2	466.6	404.3	295.2	146.4
160268_at	737.7	1,099.2	1,138.4	978.8	806.5	978	587.8	245.3
96180_at	609.5	516.9	540.1	312.8	344.8	191.8	427.9	347.1
92618_at	4,888.8	4,234.2	4,703.7	2,994.9	4,093.1	2,938.9	2,150.2	1,969.2
93203_f_at	111.8	186.8	112.9	158.1	100.8	67	119.9	90.6
102574_at	171.3	81.7	230.9	123.3	107.9	50.6	112.3	132.4
160966_at	221.2	310	454.3	242.5	238	196.2	330.7	50.8
160827_at	294.5	341.1	360.4	170.3	231.6	289.4	196.4	58.1
104116_at	1,836.3	829.3	1,258.7	1,561	722.3	810.4	943.9	1,172.1
95434_at	1,207.8	1,294.8	1,314.6	1,513.8	878.2	773.9	715.8	1,181.5

WT = wild-type (i.e. all genes present in the genome);
KO = gene knock-out (one gene is removed/silenced)

Matrix description

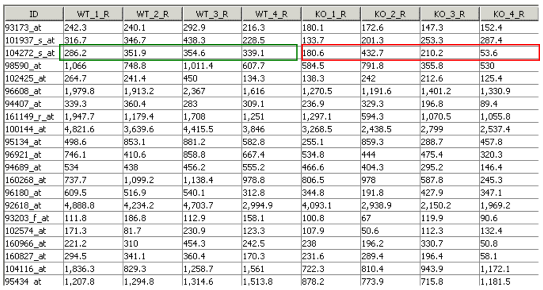
- Microarray data can be viewed as an N×M matrix:
  - Each of the N rows represents a gene
  - Each of the M columns represents a sample (e.g., patient, animal, etc.)
  - Each matrix pixel represents the *expression level* of a gene. It can be either an absolute value (e.g. Affymetrix GeneChip) or a relative expression ratio (e.g. spotted microarrays).
  - A row is referred to as the "*expression profile of the gene*".
  - A column is referred to as the "*expression profile of the sample*".

Microarray data mining challenges

- too few records (samples, animals, patients), usually < 100
- too many columns (genes), usually 1,000 < # < 10,000
- for exploration, a large set of all relevant genes is desired
- for diagnostics or identification of therapeutic targets, the smallest set of genes is needed
- model needs to be explainable to biologists

Differential gene expression analysis

- The Experiment measures gene expression in rats:
  - Two groups: (WT: wild-type rat, KO: gene knock-out rat)
  - Question: Which genes are affected by the treatment? How significant is the effect? **We compare each pair of genes**.



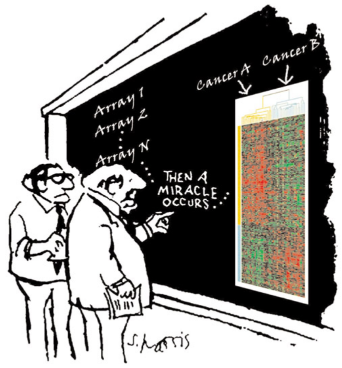
ID	WT_1_R	WT_2_R	WT_3_R	WT_4_R	KO_1_R	KO_2_R	KO_3_R	KO_4_R
93173_at	242.3	240.1	292.9	216.3	180.1	172.6	147.3	152.4
101937_s_at	316.7	346.7	438.3	228.5	133.7	201.3	253.3	287.4
104272_s_at	286.2	351.9	354.6	339.1	180.6	432.7	210.2	53.6
98590_at	1,066	748.8	1,011.4	607.7	584.5	791.8	355.8	530
102425_at	264.7	241.4	450	134.3	138.3	242	212.6	125.4
96608_at	1,979.8	1,913.2	2,367	1,616	1,270.5	1,191.6	1,401.2	1,330.9
94407_at	339.3	360.4	283	309.1	236.9	329.3	196.8	89.4
161149_r_at	1,947.7	1,179.4	1,708	1,251	1,297.1	594.3	1,070.5	1,055.8
100144_at	4,821.6	3,639.6	4,415.5	3,846	3,268.5	2,438.5	2,799	2,537.4
95134_at	498.6	853.1	881.2	582.8	255.1	859.3	288.7	457.8
96921_at	746.1	410.6	858.8	667.4	534.8	444	475.4	320.3
94689_at	534	438	456.2	555.2	466.6	404.3	295.2	146.4
160268_at	737.7	1,099.2	1,138.4	978.8	806.5	978	587.8	245.3
96180_at	609.5	516.9	540.1	312.8	344.8	191.8	427.9	347.1
92618_at	4,888.8	4,234.2	4,703.7	2,994.9	4,093.1	2,938.9	2,150.2	1,969.2
93203_f_at	111.8	186.8	112.9	158.1	100.8	67	119.9	90.6
102574_at	171.3	81.7	230.9	123.3	107.9	50.6	112.3	132.4
160966_at	221.2	310	454.3	242.5	238	196.2	330.7	50.8
160827_at	294.5	341.1	360.4	170.3	231.6	289.4	196.4	58.1
104116_at	1,836.3	829.3	1,258.7	1,561	722.3	810.4	943.9	1,172.1
95434_at	1,207.8	1,294.8	1,314.6	1,513.8	878.2	773.9	715.8	1,181.5

Experiments and questions

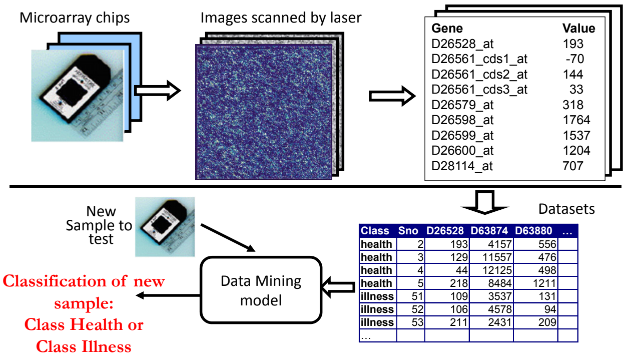
- Two-condition comparison against some form of control:
  1. Gene knock-out against wild-type (KO vs. WT)
  2. Subjects with a disease vs. healthy subjects
  3. Treated subjects vs. untreated subjects
  4. Etc
- Question of interest in these experiments are:
  1. which genes are influenced by the missing gene?
  2. which genes are responsible for the disease?
  3. which genes are influenced by the administered drug ?
  4. Etc.

Goals of a Microarray Experiment

1. **Find the genes** that change expression between experimental and control samples
2. **Find patterns:** Groups of biologically related genes that change expression together.
3. **Classify new samples** based on a gene expression profile.



Goal: develop a model to classify a new sample



Gene	Value
D26528_at	193
D26561_cds1_at	-70
D26561_cds2_at	144
D26561_cds3_at	33
D26579_at	318
D26598_at	1764
D26599_at	1537
D26600_at	1204
D28114_at	707

Class	Sno	D26528	D63874	D63880	…
health	2	193	4157	556	
health	3	129	11557	476	
health	4	44	12125	498	
health	5	218	8484	1211	
illness	51	109	3537	131	
illness	52	106	4578	94	
illness	53	211	2431	209	
…					

Classification of new sample: Class Health or Class Illness

Steps in microarray data analysis

Raw Data Image

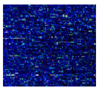
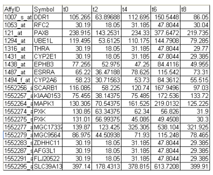


Image Analysis of chip-cell intensity

mRNA level = matrix of numbers



AffyID	Symbol	t0	t2	t4	t6	t8
1007_s_at	DDR1	105.265	63.89688	112.695	150.5448	86.05
1053_at	RFC2	30.19	18.05	31.185	47.8044	30.04
121_at	PAX8	238.915	143.2531	234.33	377.6472	219.735
1294_at	UBE1L	119.495	53.6125	110.175	144.7908	79.285
1316_at	THRA	30.19	18.05	31.185	47.8044	29.77
1431_at	CYP2E1	30.19	18.05	31.185	47.8044	29.385
1438_at	EPHB3	77.255	52.975	47.25	84.4116	49.955
1487_at	ESRRA	65.22	36.47188	78.625	115.542	73.31
1494_f_at	CYP2A6	58.23	30.71563	53.73	84.3612	55.515
1552256_a	SCARB1	116.085	58.225	120.74	167.9496	97.03
1552257_a	KIAA0153	75.455	38.14375	75.485	172.536	133.72
1552264_a	MAPK1	130.305	70.54375	161.525	219.0132	125.225
1552274_a	PXK	130.85	63.34375	62.34	56.826	31.9
1552275_s	PXK	131.01	56.99375	45.085	49.4508	30.3
1552277_a	MGC17337	139.87	123.425	325.305	538.104	321.925
1552279_a	MGC9564	86.975	44.50938	71.93	115.248	78.465
1552283_s	ZDHHC11	30.19	18.05	31.185	47.8044	29.385
1552287_s	AFG3L1	30.19	18.05	31.185	47.8044	29.385
1552291_a	FLJ20522	30.19	18.05	31.185	47.8044	29.385
1552295_a	SLC39A13	397.14	178.4313	378.815	613.7208	399.91

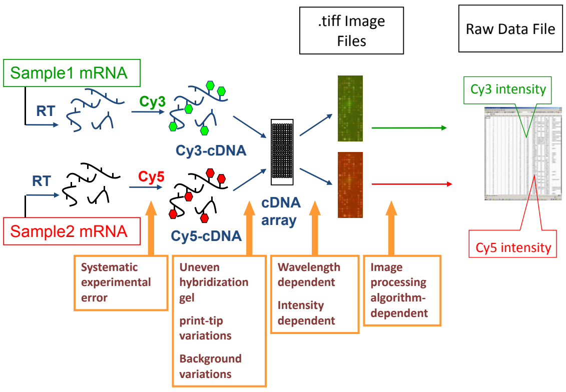
Normalisation and removal of non-biological variation

Statistical and computational analysis

Of interest to us

Functional analysis (identify affected processes and pathways)

Microarray data are very noisy



Removal of noise: thresholding & filtering

- *Thresholding*: Removing bad intensity spots is an important process of quality control. For example, the scanner has a measurement limit below which intensity values cannot be trusted. Values below the cut-off point are usually removed (filtered) from the data because they are likely to be artifacts.
  - Typically, the lowest intensity value of reliable data is 100–200 for Affymetrix data and 100–1000 for cDNA microarray data.
- *Filtering*: remove genes with insufficient variation between two conditions:
  - e.g. MaximumValue – MinimumValue < Δ (usually 500)
  - MaximumValue / MinimumValue < r (usually 5)
- Input for further processing is a matrix of numbers.

Normalisation

- Gene expressions can differ by an order of magnitude.
- Normalisation is needed for gene selection, clustering and classification models.
- *Normalisation:* mathematical transformation of values of gene expression from the interval $[m_{min}, m_{max}]$ → $[m'_{min}, m'_{max}]$, either
  - Linearly
  - Logarithmically
  - to Mean = 0, Std. Dev = 1
  - other
- **Whatever the method: normalise each gene row separately!**

Linear normalisation

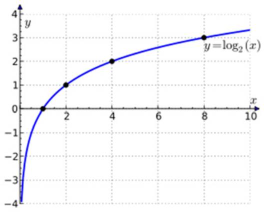
- *Linear Normalisation:* Let *m'* be the new normalised value of gene expression / mRNA level:

$$m' = \frac{m - m_{\min}}{m_{\max} - m_{\min}}$$

- This equation transforms values of gene expressions from the interval $[m_{min}, m_{max}]$ → [0, 1] **uniformly**.
  - When $m = m_{min}$, then $m' = 0$
  - When $m = m_{max}$, then $m' = 1$

Logarithmic normalisation

- If the data have a huge value span like from 10^2 to 10^4, then it's more suitable to use a logarithm, e.g. $\boldsymbol{m' = \log m}$. (Either $\log_{10}$ or $\log_2$).
- This equation transforms values of gene expressions from the interval $[m_{min}, m_{max}]$ → $[m'_{min}, m'_{max}]$ **nonuniformly**.



Other equations for normalisation:
http:\\people.revoledu.com\kardi\tutorial\Similarity\Normalization.html

What's next after normalisation:

- Gene Selection
  - find genes, which would be the best predictors (of disease, treatment outcome, etc.)
- Clustering (Unsupervised, no class labels)
  - Exploration and finding patterns
  - find new biological classes of genes / refine existing ones
- Classification (Supervised, needs class labels)
  - identify disease and its genetic profile
  - predict outcome / select best treatment
- Functional / ontology analysis

Potential applications of microarrays

- Biological and medical discovery
  - discovery of putative functions of genes
  - finding and refining biological pathways
  - new and better molecular diagnostics / "personalised" medicine
  - appropriate treatment for genetic signatures
  - potential genetic targets for new therapies