BIOMENG 261

TISSUE AND BIOMOLECULAR ENGINEERING

Module I: Reaction kinetics and systems biology

Oliver Maclaren
oliver.maclaren@auckland.ac.nz

MODULE OVERVIEW

Reaction kinetics and systems biology (*Oliver Maclaren*)
[12 lectures/3 tutorials/2 labs]

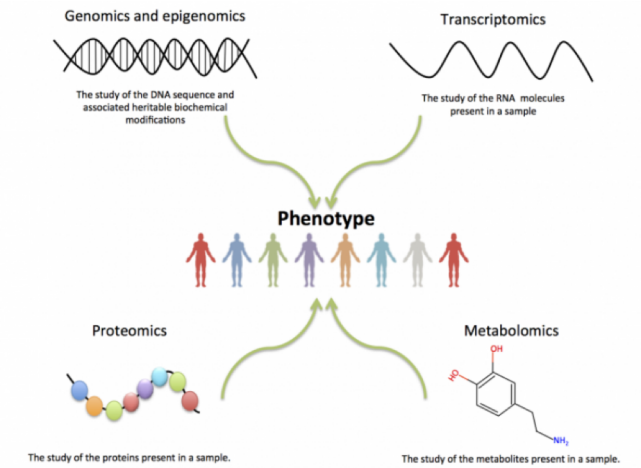
1. *Basic principles: modelling with reaction kinetics* [6 lectures]
   Physical principles: conservation, directional and constitutive. Reaction modelling. Mass action. Enzyme kinetics. Enzyme regulation. Mathematical/graphical tools for analysis and fitting.
2. *Systems biology I: overview, signalling and metabolic systems* [3 lectures]
   Overview of systems biology. Modelling signalling systems using reaction kinetics. Introduction to parameter estimation. Modelling metabolic systems using reaction kinetics. Flux balance analysis and constraint-based methods.
3. *Systems biology II: genetic systems* [3 lectures]
   Modelling genes and gene regulation using reaction kinetics. Gene regulatory networks, transcriptomics and analysis of microarray data.

LECTURE 12: GENE EXPRESSION DATA AND GRNS

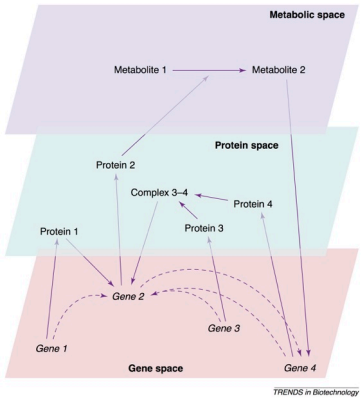
- *Larger systems*
  - Gene space and gene regulatory networks (GRNs)
- *Brief overview of microarray data*
  - Experiment types
  - Data organisation and expression matrices
- *Analysis types*
  - Clustering, distance matrices and dendrograms
  - Control analysis and regulatory matrices

Note: there are many images stolen from the internet in what follows...

MUCH LARGER SYSTEMS - 'OMICS'



GENE SPACE



See: Brazhnik et al. (2002) 'Gene networks - how to put the function in genomics' (on Canvas)

TRANSCRIPTOMICS

- A subfield of *functional genomics*
- Functional genomics: study of how genes and intergenic regions contribute to biological function
- The focus is on *gene expression*
- In particular, via *measuring mRNA* (the transcripts)

See: Lowe et al. (2017) 'Transcriptomics technologies' (on Canvas)

EXPRESSION ANALYSIS

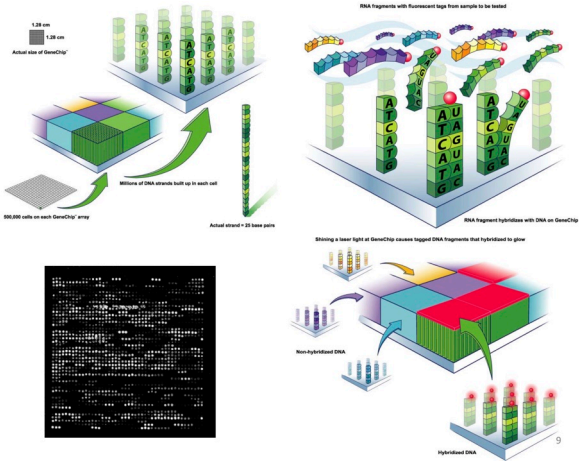
- *Microarrays*
- Mature technology
- Relatively well-established data analysis methods
- *RNA-seq*
- Newer technology, rapidly overtaking microarrays
- Less standardisation of analysis methods
- Much more computationally/storage intensive

But: *microarrays still relevant and useful*: we will consider these (easier and better understood)

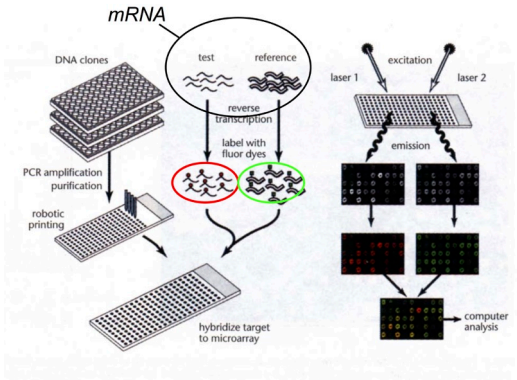
MICROARRAYS

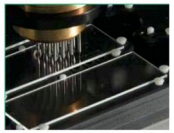
For video intros: see e.g. https://youtu.be/0ATUjAxNf6U or https://youtu.be/VNsThMNjKhM

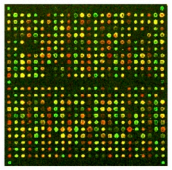
MICROARRAYS



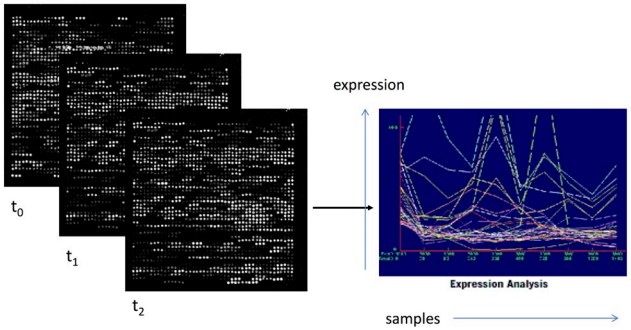
MICROARRAYS: COMPARATIVE EXPRESSION



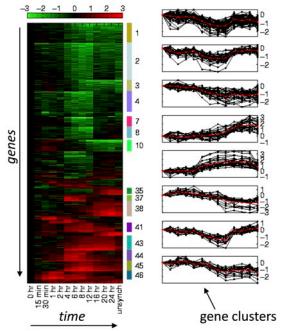




MICROARRAYS: TIME SERIES



MICROARRAYS: RELATIVE EXPRESSION OVER TIME



green – *down* regulation
red – *up* regulation

DATA ANALYSIS: STATISTICAL/MACHINE LEARNING

Clustering, unsupervised and supervised learning etc. see:

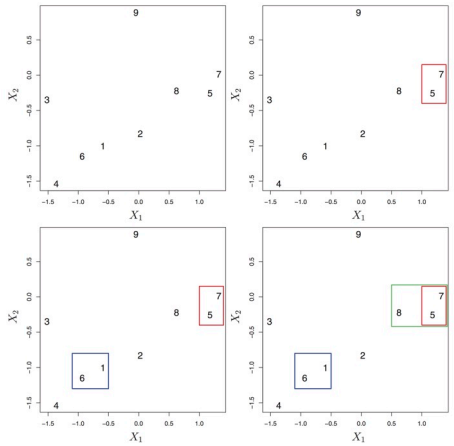
- James et al. 'An Introduction to Statistical Learning'
  - Available at: http://www-bcf.usc.edu/~gareth/ISL/
- Hastie et. al 'Elements of Statistical Learning: Data Mining, Inference and Prediction'
  - Available at: http://web.stanford.edu/~hastie/ElemStatLearn/

CLUSTERING

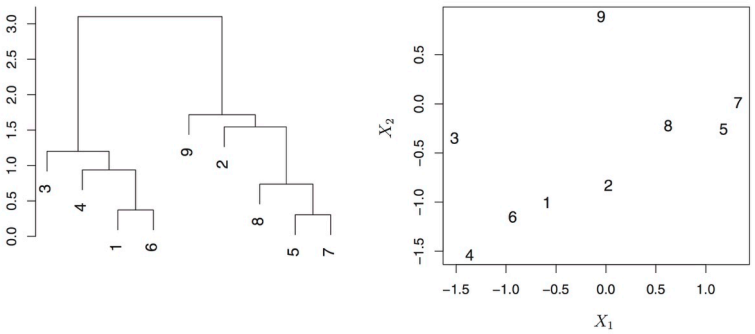
- An *unsupervised learning* method for *pattern discovery*
- Two popular algorithms are
  - K-means
  - Hierarchical clustering

See James et al. Chapter 10 for detailed algorithms. We will look at *hierarchical clustering* here.

HIERARCHICAL CLUSTERING EXAMPLE (JAMES ET AL.)



HIERARCHICAL CLUSTERING: DENDROGRAMS



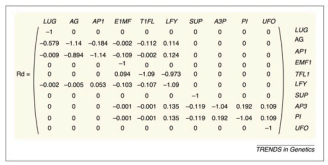
PERTURBATION APPROACH FOR INFERRING REGULATORY MATRICES/NETWORKS

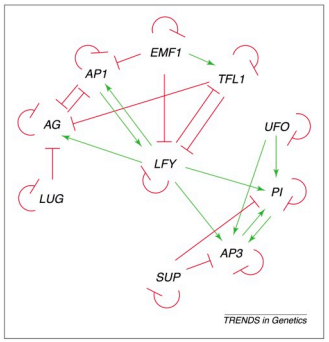
- Perturb *transcription rates* for *each gene in turn*
- Measure changes in *steady-state expression levels* for all genes (including self)
- Gives indication of underlying *regulatory network*
- Summarise in *regulatory strength matrix or network diagram*.

See de la Fuente et al. (2002) 'Linking the genes' (on Canvas)

PERTURBATION APPROACH FOR INFERRING REGULATORY MATRICES/NETWORKS

Example: flower morphogenesis (see de la Fuente et al. 2002 for details):





EXAMPLE PAST QUESTIONS

(b) DNA microarrays are used are used to simultaneously measure the expression of many different genes in a sample. Explain briefly the difference between *time series* and *comparative* microarray measurements.

(4 marks)

(c) In a series of experiments, the amounts of mRNA for different genes are perturbed and the changes in mRNA of all genes of interest (namely A, B and C) are measured. A set of 'co-control coefficients' was calculated and organised into the 'Regulatory Strength' matrix, *Rd*, given below

$$Rd = \begin{array}{c} \begin{matrix} A & B & C \end{matrix} \\ \begin{bmatrix} -0.5 & 0 & 0 \\ 0.1 & 0 & 1.2 \\ 0.8 & -0.1 & 1 \end{bmatrix} \end{array}$$

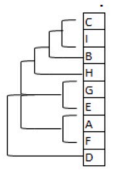
Using the information in *Rd* sketch a qualitative regulatory network showing how each gene regulates the expression of all genes (including itself). Use arrows (→) to show positive regulation and blunt arrows (—⊣) to indicate negative regulation.

(5 marks)

EXAMPLE PAST QUESTIONS

5)

a) In a microarray experiment, the expression of 8 genes was measured as a function of time and the data were analysed to create the following dendrogram:



State if the following statements are TRUE, FALSE or INDETERMINATE.

i) Genes C and I show similar expression patterns at the different time points.

(1 mark)

ii) The Euclidean distance between genes C and I is less than the Euclidean distance between genes I and B.

(1 mark)

Biomeng 261 Lecture 12

- Large genetic systems
  - 'Gene space' & GRNs (genetic regulatory networks)
  - Expression analysis methods

~~Takeaways~~ Upshot: know/interpret/draw

- matrix types
  - ↳ Gene expression matrix } main general format
    - ↳ gives { distance matrix ① for clustering
      regulatory strength matrix ② for control analysis
- dendrograms
  - ↳ summarise distance matrix / clustering algorithm ①
- regulatory networks
  - ↳ summarise regulatory strength matrix ② (control analysis).

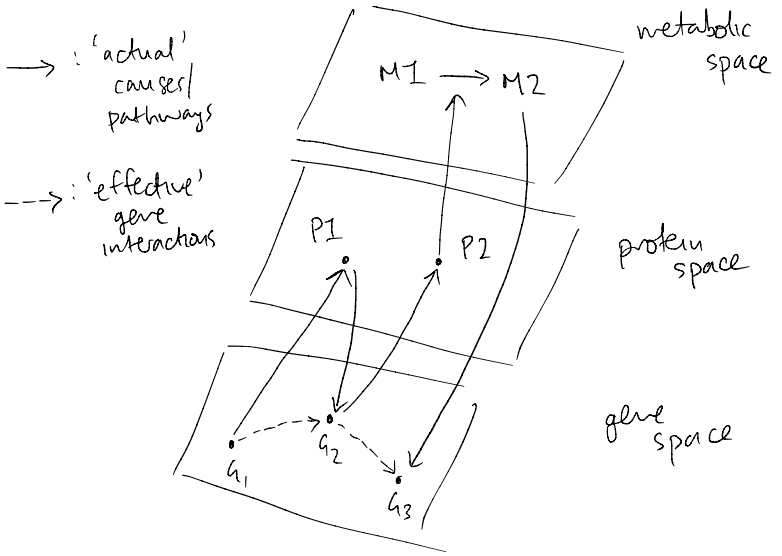
Setting: even more complex networks than last time

→ eg 1000s of genes (or more!)

(Background):

Gene Space

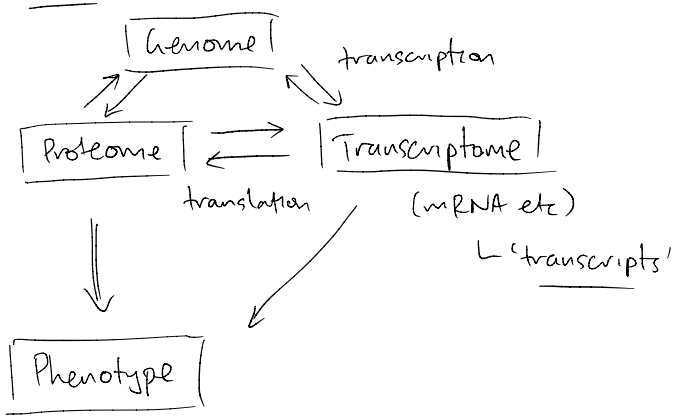
- A way of 'projecting all the action' down into interactions between genes



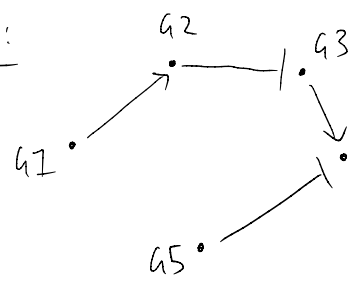
(Background)

Gene Regulatory Networks

From:



To:


G2
G3
G1
G4
G5

gene's eye view

ie a Gene Regulatory Network (GRN)

→ : positive
⊣ : negative

Transcriptomics

- A subfield of 'functional genomics'
  - ie the study of how genes & intergenic (between gene) regions contribute to biological function

- Transcriptomics focuses on gene expression levels
  - In particular as measured via the levels of the transcripts (mRNA) associated with genes
    - ↳ mRNA easier to measure etc that proteins, but see 'proteomics'

Ideas:
- does expression go up or down 'under treatment'?
- do groups of genes go up/down together?

Expression Analysis

Two key approaches:

- **Microarrays**
  - uses (known) 'probes' (eg cDNA) — need to know roughly what looking for
  - samples 'hybridise' if complementary to probes
  - amplify & quantify via qPCR
    - PCR: polymerase chain reaction (see lab lectures)
- **RNA-seq**
  - direct sequencing of transcripts — don't need pre-chosen 'probes'
  - 'next gen', high-throughput sequencing

---

* we will discuss **microarrays** *
  - well understood
  - more mature & easier to analyse
  - still used & useful
  - ... but RNA-seq overtaking!

(see Lowe et al. 2017 'Transcriptomics technologies')

Microarrays: (more) background (see slides & videos etc)

Idea: want to measure mRNA levels (hence gene expression)

- Microarrays can measure **1000s** of mRNA levels at a time
  - ↳ ie **1000s of genes at a time**

→ consist of a grid of 'spots' containing cDNA (complementary DNA)

→ mRNA samples preferentially bind to ('hybridise' with) corresponding cDNA

↳ 'reverse transcription'! (DNA ← mRNA)

Also:
- use fluorescent tags to distinguish different samples
- amplify levels of products via PCR/qPCR.

Preprocessing

Preprocessing is crucial

BUT: we won't go into here

Typical considerations:

- artifacts
- absolute vs relative expression levels
- log transformations

etc

---

usually work with

log(relative expression)

From now: take measure as 'given' & just use numbers.

→ Vinod knows more about experimental side!

Data & Experiment Types

We consider **two types** of experiment

1. **Perturbation** (or **comparative**)
   - effect of treatment
   - different tissue types etc

   } multiple sample 'types', same 'time'

   — eg treated vs untreated
   — cancer or not etc.

2. **Time Series**
   - same cell/tissue etc studied over time

   } one sample type, multiple times

Idea:
- time has natural order
  - continuous or categorical ordinal
- treatment/class not necessarily ordered
  - cancer or not etc

BUT essentially

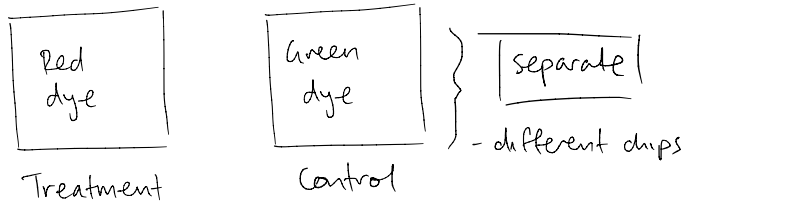
- just different independent vars, same basic ideas
- also, often want to combine, eg compare two time series from different treatments at same times

Perturbation / comparative

- Have both a treatment & control (& at single time)

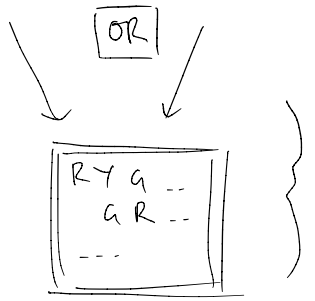
Variations:

(A).



or

(B).



Competitive (typical)
- same chip for both treatment & control
- compete to hybridise with probes

where:

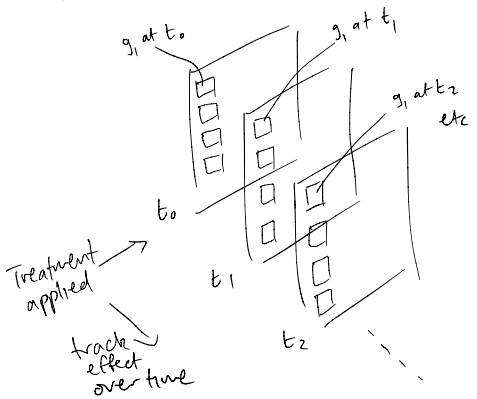
direction of treatment relative to control

Who wins?
↑ : red
↓ : green
— : yellow

resulting colour on chip.

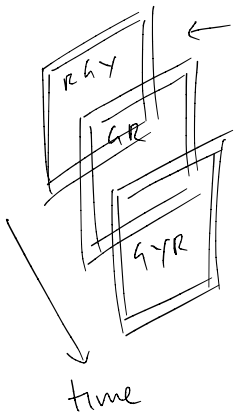
ie red: expression ↑ for treatment
green: expression ↓ for treatment
(rel. to control).

Time Series



- treat one sample type at t_0 (say)
- new chip for each time
- track resulting expression patterns over time.

Combined:



← each array has both treatment & control (or some other comparison)

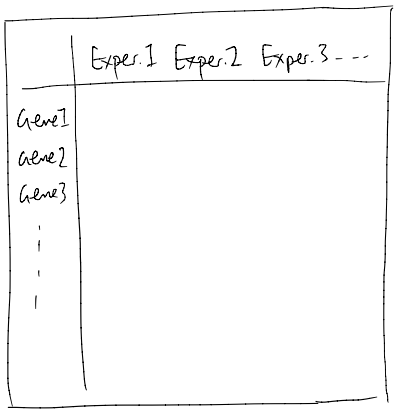
+

track over time

↘ (new array per time)

Data format : **Gene expression matrices**

→ A general format that we can put both time series & perturbation/comparative into:



Eg experiment 1 : time t_0, control

experiment 2 : time t_1, control

⋮

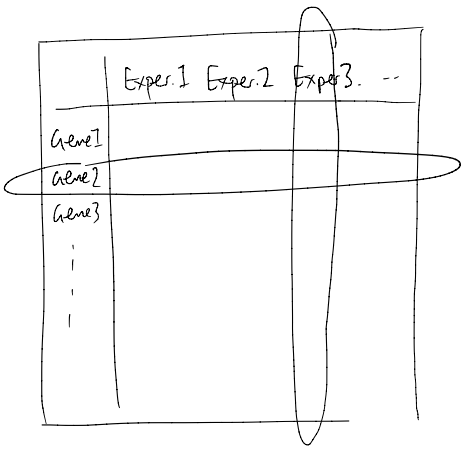
experiment n+1 : time t_0, treatment

experiment n+2 : time t_1, treatment etc.

(Warning : many 'data array' formats in stat./ eg R, Python etc are the transpose of above : genes are col, exp. are row)

Gene expression matrices

Column: a particular experiment, ie an array (gene chip/microarray) of all genes.

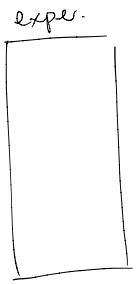


row: a gene expression profile over experiments for a given gene.

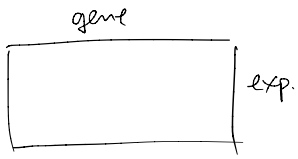
Typically : # genes >> experiments

unknowns data

Here:



→ in stats often use transpose format



Questions / Problems

Q: which genes 'cause' difference between control & treatment / two samples for comparison?

→ many more genes than experiments

→ overfitting issues if try to explain via single genes!

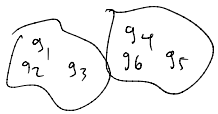
eg one difference (cancer or not) → 1000 possible explanations!

→ problem for naive/traditional stat. inference!

└ motivation for many modern stat/ML methods...

One way to tackle:

- focus on sets/clusters of genes

eg 

→ smaller number of these groups

→ also relates to idea that genes work together & co-express

Typical analyses

1. Clustering: similarity analysis (eg over time)

2. Perturbation: linear control analysis based (SS).

---

1. **Clustering**

- a form of **unsupervised learning** for pattern discovery

- need a notion of 'distance' however
  └ user input (see eg prev. lectures & below)

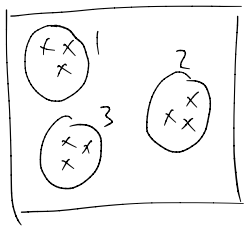
For fun (not examinable):

Defining idea of 'distance' & related:

distance	**metric**	**dissimilarity**
i) $d(x,y) \geq 0$	i–iii) &	i–ii) &
ii) $d(x,y) = d(y,x)$	iv) $d(x,y) = 0$ iff $x = y$	iii)*: $d(x,y)$ increases monotonically as x & y more 'dissimilar' (subjective)
iii) $d(x,x) = 0$	v) $d(x,y) + d(y,z) \geq d(x,z)$	

(Terminology)

Unsupervised? Note on 'learning types'

Supervised	Unsupervised
$X \longrightarrow Y$	X ↻
learns function $X \to Y$	pattern discovery (in X)
Train: supervisor gives examples	Eg Find clusters
→ Happy } labels given	
→ Sad	→ group 'similar'
Test: predict on new (unseen) set	→ no labels given
→ Happy ✗	→ do need a distance/ similarity measure
→ Happy ✓	→ In general harder to evaluate (some methods exist --)

Example: Clustering expression profiles

→ across time &/or experiments (ie profiles of genes: )

Expression matrix:

	Experiment 1	2	3	-----	8
gene A	-1	-2	2	---	2
gene B	-1	-2	-1	---	
⋮					
gene I	-1	-2	0	----	

Q: which genes have 'similar' profiles?

→ define distance ... eg Euclidean (since easy)

Distance between two profiles A, B :

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (y_i^{(A)} - y_i^{(B)})^2} \quad \text{(Euclidean)}$$

(square root of sum of squared differences)

where
- n : number of experiments
- $y_i^{(A)}$: expression level of gene A in experiment i

Example

A	1	-1	2	0
B	1	-1	-1	2

two gene profiles.

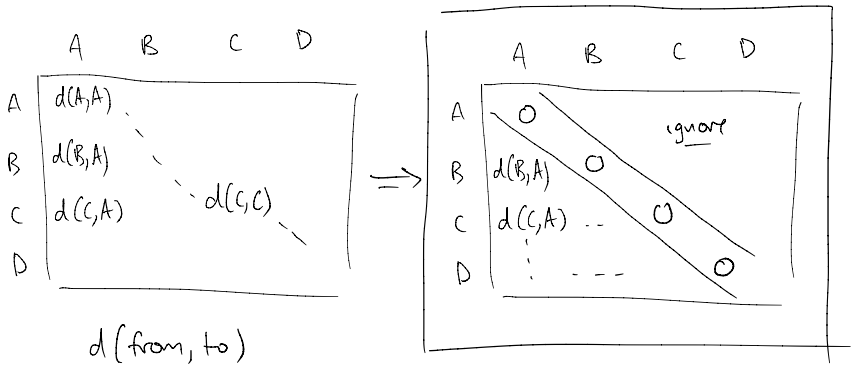
Squared diffs:	0^2	0^2	3^2	2^2	Sum squares!
ie	0	0	9	4	13

so $\boxed{d(A,B) = \sqrt{13}}$

Distance matrices

- we can summarise all pairwise differences between profiles in a **distance matrix**

(note : this isn't an expression matrix)



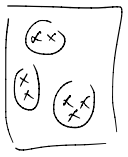
d(from, to)
= d(row, col)

Note : - symmetric, so don't need explicit upper part
ie $d(B,A) = d(A,B)$

- diagonals are zero.
$d(A,A) = 0$ etc.

Distance-based clustering

Goal: group together if 'close' eg:



etc.

Two popular algorithms

- K-means
- Hierarchical

Here: will consider hierarchical

Pseudocode (Hierarchical)

Begin with n observations

Find all pairwise distances

For $i = n, n-1, \ldots 2$:

Find smallest distance

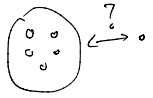
'Fuse' or 'group' together (so n obs → n-1 obs)

Recompute distances to 'fused' cluster *
(see below)

* Distance to cluster?

- multiple types
  - min
  - max
  - average
  - etc...

} 'linkage type'

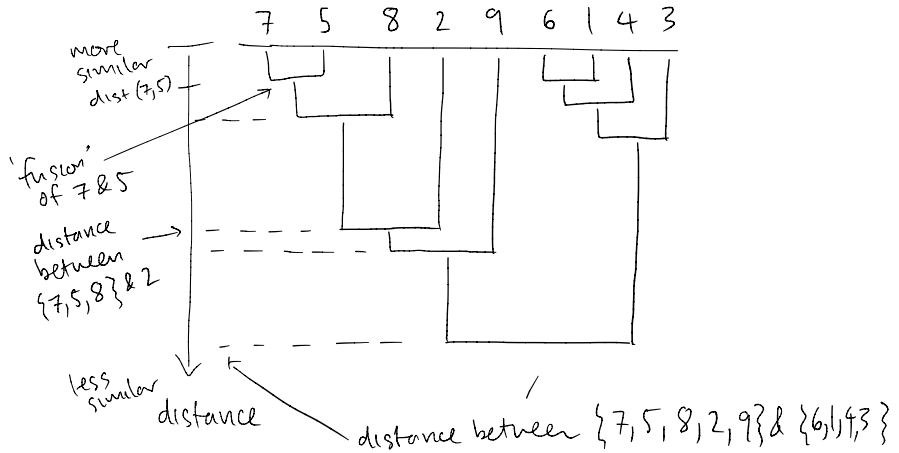


Dendrograms

- summarise results of hierarchical clustering

- Indicate distance between various clusters in 'tree-style' diagram

Example:

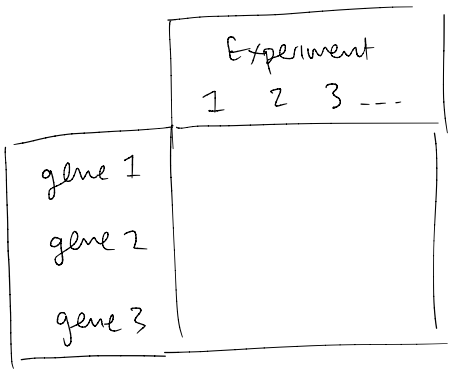


- height (y axis) of 'fusion' indicates distance
- for any two clusters, can find where they first fused

Other analysis type.

2. 'Control' analysis via individual perturbations

Consider gene expression matrix again:



Goal: how do genes 'affect' other genes?

Can we deduce 'control network'?

Effect of perturbations

→ Make each experiment a perturbation of a single gene

- Increase transcription rates of each gene in turn
- measure the change in steady state concentrations (expressions of mRNA) for all genes (incl. self).

→ get (co-) 'control coefficients' (eg +1.5 or −0.5)

→ summarise in a regulatory strength matrix

see: de la Fuente (2002) 'Linking the genes' for details.

Effect of perturbations of gene transcription rates on ss. concentrations/expression levels

Result:

- Can summarise in **regulatory strength matrix** R_d
- can use to draw potential gene regulatory networks

R_d =

Experiment: rate ↑ of..	gene 1	gene 2	...
Δgene 1 level	−1	0.5	...
Δgene 2 level	−0.2	1	
⋮	⋮	⋮	⋱

↑ normalised change in expression level

Example

$$R_d = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} -0.5 & 0 & 0 \\ 0.1 & 0 & 1.2 \\ 0.8 & -0.1 & 1 \end{bmatrix} \end{array}$$

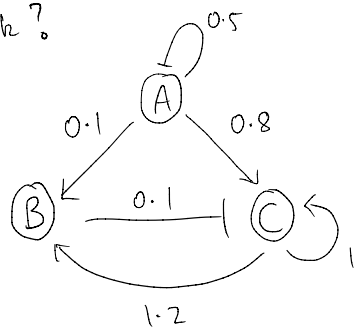
rate increase (↙ A B C)

↑ s.s. conc. change

Interp: effect of.. on [ ]

Q: Network?

A:



⊣ negative
← positive

negative auto (self) regulation: A

positive auto regulation: C