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*)
[11-12 lectures/3 tutorials/2 labs]

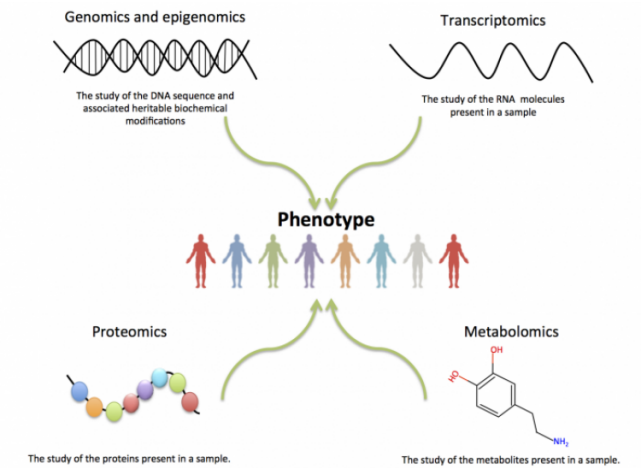
1. *Basic principles: modelling with reaction kinetics* [5-6 lectures]
   Physical principles: conservation, directional and constitutive. Reaction modelling. Mass action. Enzyme kinetics. Enzyme regulation. Mathematical/graphical tools for analysis and fitting.
2. *Systems biology I: signalling and 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

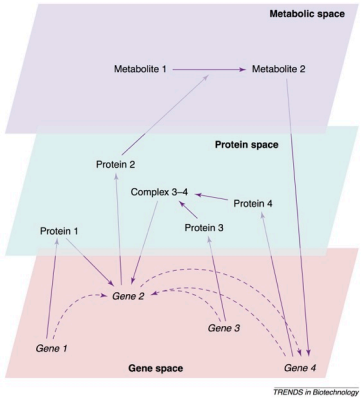
- *Larger systems*
  - Gene space and gene regulatory networks (GRNs)
- *Brief overview of microarray data*
  - Experiment types
  - Data organisation and expression matrices
- *Analysis types*
  - Clustering, distance matrices and dendrograms
  - Control analysis and regulatory matrices

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

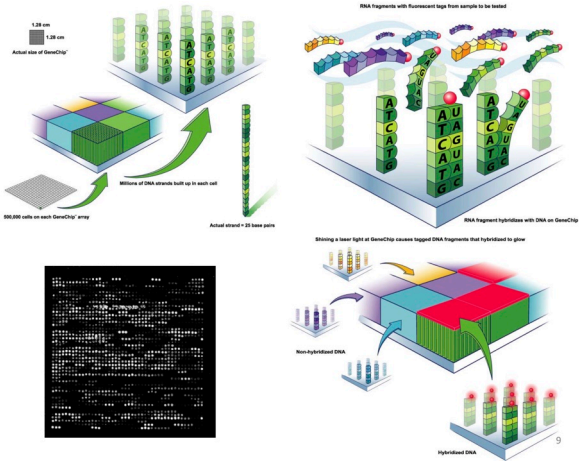
- *Microarrays*
- Mature technology
- Relatively well-established data analysis methods
- *RNA-seq*
- Newer technology, rapidly overtaking microarrays
- Less standardisation of analysis methods
- Much more computationally/storage intensive

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

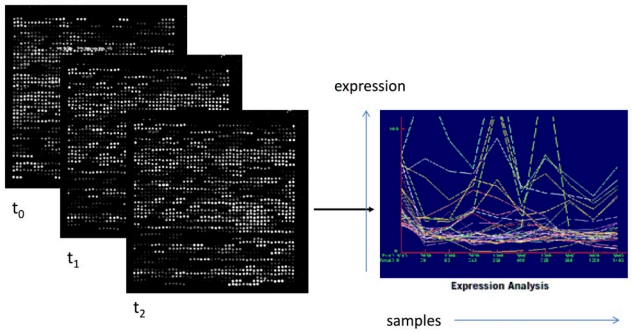
MICROARRAYS

For a video intro: see e.g. http://www.youtube.com/watch?v=VNsThMNjKhM

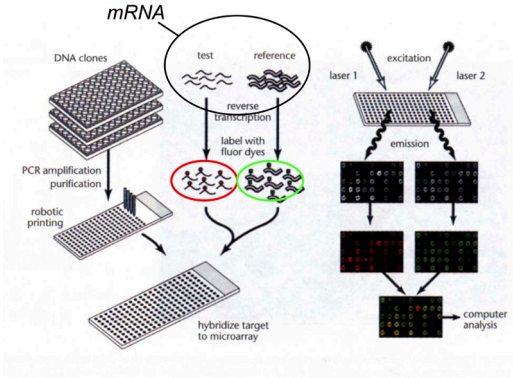
MICROARRAYS

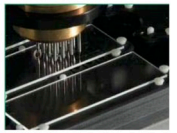


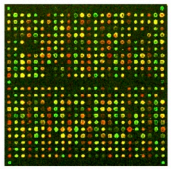
MICROARRAYS: TIME SERIES



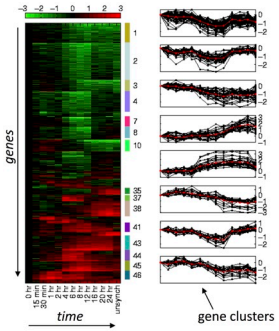
MICROARRAYS: COMPARATIVE EXPRESSION







MICROARRAYS: RELATIVE EXPRESSION OVER TIME



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

DATA ANALYSIS: STATISTICAL/MACHINE LEARNING

Clustering, unsupervised and supervised learning etc. see:

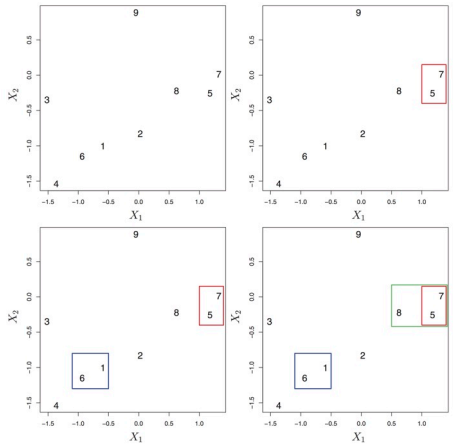
- James et al. 'An Introduction to Statistical Learning'
  - Available at: http://www-bcf.usc.edu/~gareth/ISL/
- Hastie et. al 'Elements of Statistical Learning: Data Mining, Inference and Prediction'
  - Available at: http://web.stanford.edu/~hastie/ElemStatLearn/

CLUSTERING

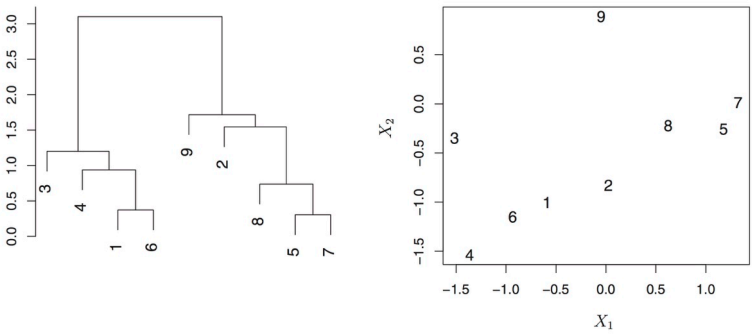
- An *unsupervised learning* method for *pattern discovery*
- Two popular algorithms are
  - K-means
  - Hierarchical clustering

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



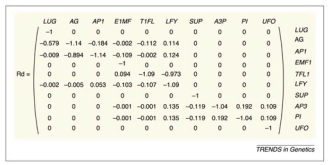
PERTURBATION APPROACH FOR INFERRING REGULATORY MATRICES/NETWORKS

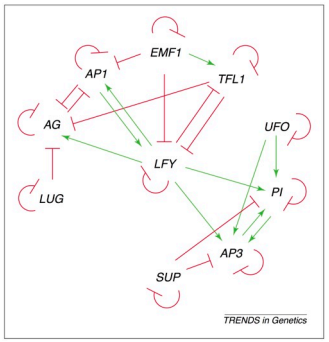
- Perturb *transcription rates* for *each gene in turn*
- Measure changes in *steady-state expression levels* for all genes (including self)
- Gives indication of underlying *regulatory network*
- Summarise in *regulatory strength matrix or network diagram*.

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

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

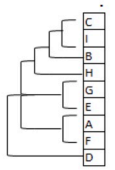
Using the information in *Rd* sketch a qualitative regulatory network showing how each gene regulates the expression of all genes (including itself). Use arrows (→) to show positive regulation and blunt arrows (⊣) to indicate negative regulation.

(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: know & interpret:

- matrix types
  - ↳ Gene expression matrix } main
    - ↳ gives { distance matrix ①
      regulatory strength matrix ②
- dendrograms
  - ↳ summarise distance matrix ① / clustering algorithm
- regulatory networks
  - ↳ summarise regulatory strength matrix. ②

Background: Gene space

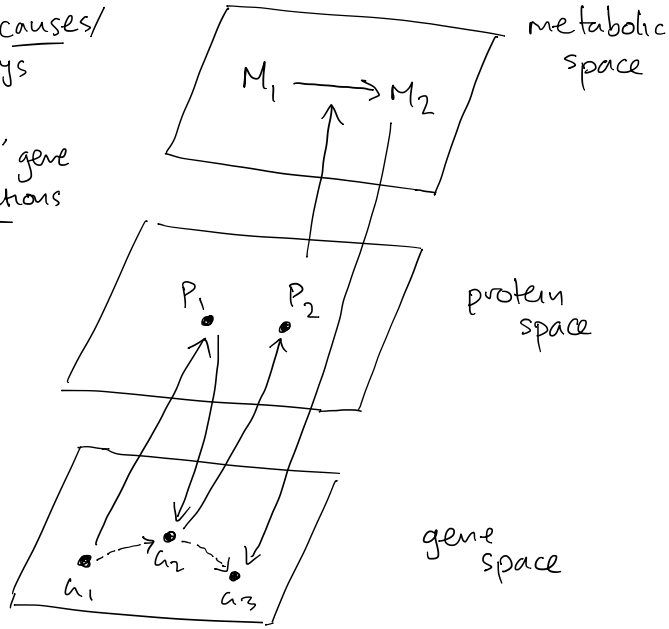
- A way of 'projecting all the action' into interactions between genes

---

Arrow types:

→ 'actual' causes/ pathways

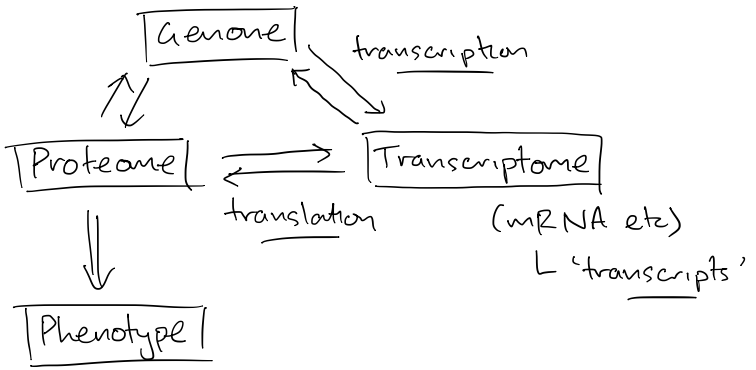
⇢ 'effective' gene interactions



(Background)

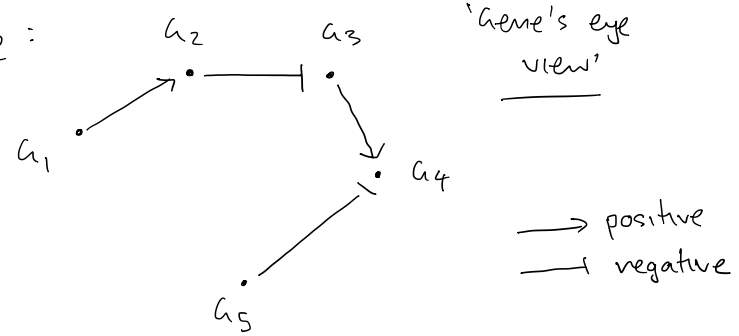
Leads to : Gene Regulatory Networks

From :



To :

'Gene's eye view'



"Gene regulatory network" (GRN)

Warning: we also loosely used this term for 'genetic regulation' models in gene. usually: gene's eye view.

(Background)

Transcriptomics ('omics': as a whole)

- subfield of 'functional genomics'

  [The study of how genes & intergenic regions contribute to biological function]

- Transcriptomics focuses on **gene expression levels** as measured by the levels of the **transcripts** (mRNA) associated with genes

  ↳ mRNA typically easier to measure etc than proteins, but see 'proteomics'!

Key idea

under 'treatment':

- does expression go up or down?
- do groups of genes go up or down together?

Expression Analysis

Two key approaches:

- **Microoarrays**
  - uses 'probes' (eg cDNA) (complementary DNA)
  - samples 'hybridise' to probes complementary to probes
  - amplify & quantify using qPCR (PCR: polymerase chain reaction)
- **RNA-seq**
  - direct sequencing of transcripts
  - 'next gen' high-throughput sequencing

We will focus on **microarrays**!

→ well understood
→ more mature, easier to analyse
→ still used / still useful
---- but RNA-seq is 'rapidly overtaking'

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

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

- Microoarrays 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 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.

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'

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

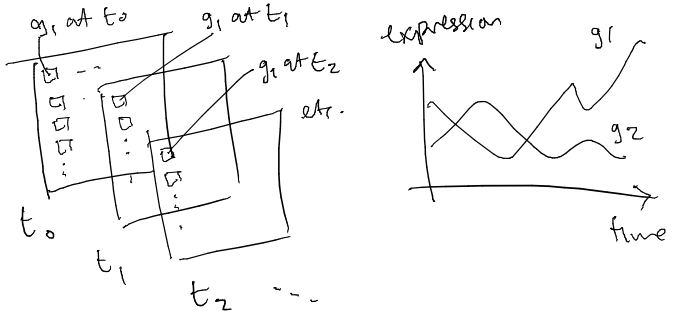
   } one sample type, multiple times

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

BUT essentially

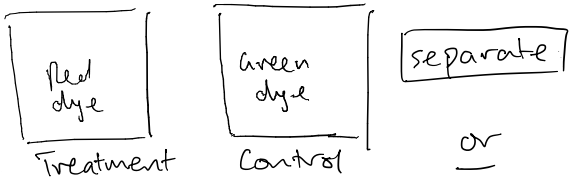
- just different independent vars, same basic ideas
- also, often want to combine, eg compare two time series from different treatments at same times

Time Series



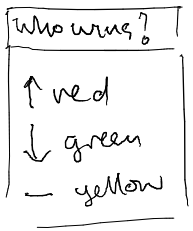
new chip for each time

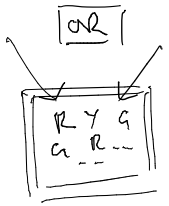
Perturbation/comparative



separate

or





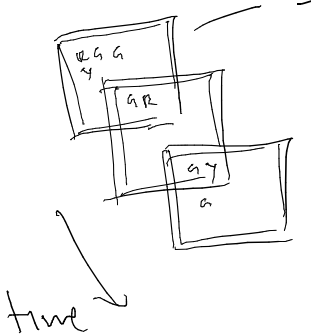
competitive

- same chip for both treatment & control
- compete to hybridize

usually

↳ direction of treatment relative to control

Combined



each array has both treatment & control

+

track over time (new array per time)

time

Idea: genes that co-express (correlated expression) are potentially co-regulated.

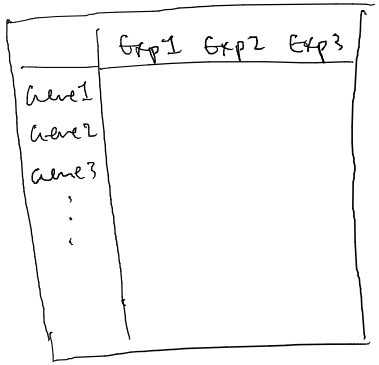
Two key goals

- try to cluster (group) 'similar'
- gene-gene effects
  ↳ perturb each gene and see effects

But first →

Gene expression matrices

A general format that we can put both time series & perturb./comparative into.



Exp: experiment (not expression)

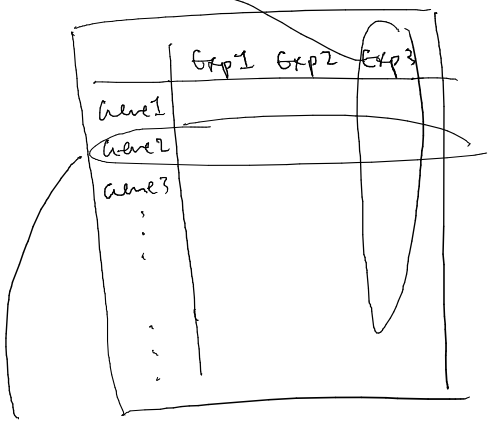
→ eg exp1 : time t_0, control } experiment 1.
exp2 : time t_1, control
⋮
expn : time t_0, treatment
expn+1: time t_1, treatment

etc.

(warning: actually the transpose of many 'data array' formats eg R dataframes)

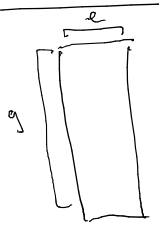
Gene expression matrices

Column: an experiment, ie an array (gene chip/ microarray)



Row: a gene expression profile for a given gene

Typically: #genes >> #experiments



- many more genes than experiments
- problem for statistical inference!
  ⇒ focus on smaller sets or clusters of genes

Some typical analyses

1. Clustering: similarity analyses (eg over time) ←
2. Perturbators: linear control based analyses

---

1. **Clustering**
   - a form of **unsupervised learning** for pattern discovery
   - Need notion of 'distance', however.
     - Recall prev. lectures.

Note: more precise distinctions

distance
i) $d(x,y) \geq 0$
ii) $d(x,y) = d(y,x)$
iii) $d(x,x) = 0$

metric
i–iii) &
iv) $d(x,y) = 0$ iff $x = y$
v) $d(x,y) + d(y,z) \geq d(x,z)$

dissimilarity
i–ii) &
iii)* $d(x,y)$ increases monotonically as x & y 'more dissimilar' (subjective)

(Terminology)

Unsupervised? Note on 'learning types'

supervised

$X \longrightarrow Y$

learn function $X \to Y$

↳ Train: supervisor examples

 → happy (labels given)

 → sad

Test: predict on new set

 → happy ✗

 → happy ✓

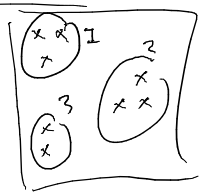
'reasonably straightforward' to evaluate/validate.

unsupervised

X ↻

Pattern discovery in X

Eg: Final clusters



group 'similar'

→ no labels given

→ goal: discover labels/clusters etc

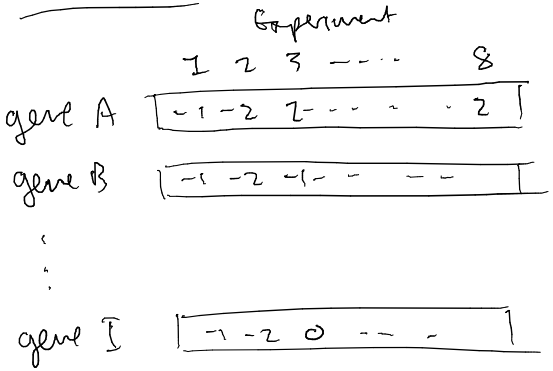
→ do need 'distance'/similarity $d(x_1, x_2)$

→ Can do some eval/valid. eg via consistency/stability on training/test sets
↳ but in general, harder to 'validate'

Example : Clustering expression profiles

→ across time/exp. (profiles)

Expression matrix



which genes have 'similar' profiles?

Distance? Here: Euclidean (for simplicity)

→

Distance between two profiles

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

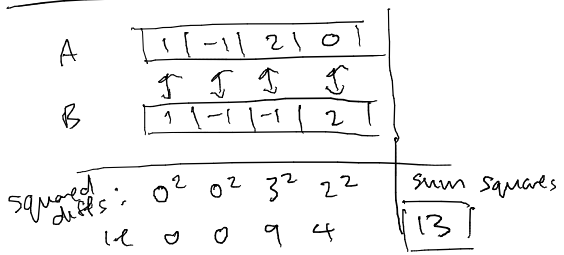
Euclidean: square root of sum of squared diff.

where

n : number of experiments

$y_i^{(A)}$: expression level of gene A in experiment i

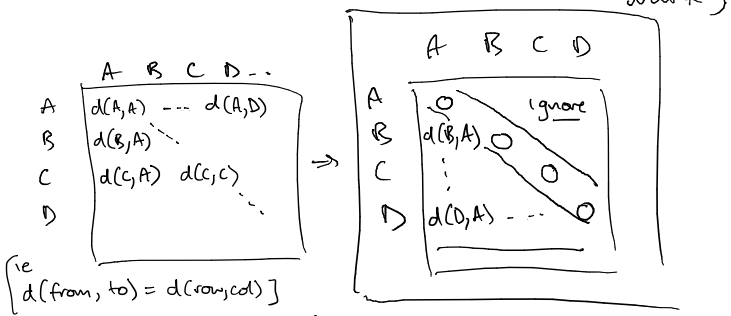
Example:



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

Distance matrices

We can summarise all pairwise differences in a **distance matrix** (note: is not an expression matrix)



[ie d(from, to) = d(row, col)]

Note:
- symmetric so don't need upper part of $d(B,A) = d(A,B)$
- diagonals are zero

Distance-based clustering

Goal: group together if close.

Two popular algorithms:

- K-means
- Hierarchical

see over → for pseudocode

Here: consider hierarchical

→

Hierarchical clustering

Begin with n observations

Find pairwise distances

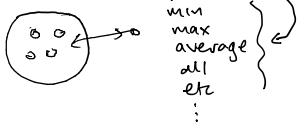
For $i = n, n-1, \ldots 2$:

Find smallest distance

'Fuse' together (so: $n \to n-1$ observ.)

Recompute distances to 'fused' cluster

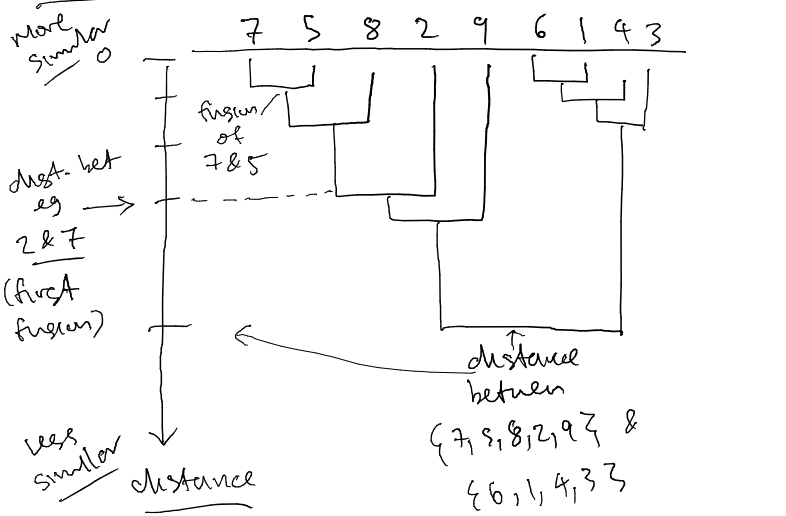
↳ need notion of distance to a cluster ('linkage' type)



Dendrograms

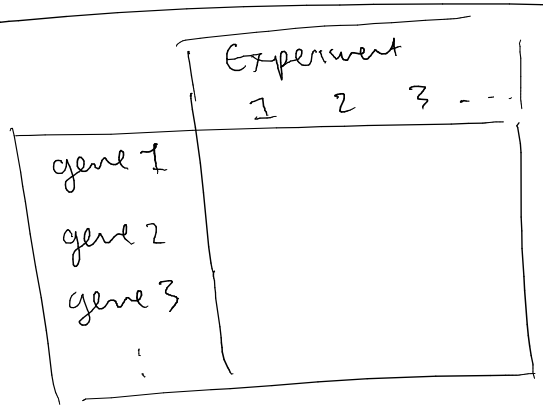
- Summarise a hierarchical clustering algorithm

- Indicate distance between various clusters in 'tree'-style diagram
- height of 'fusion' indicates distance
- for any two observations, can find where they first fused



2. 'Control' analyses via individual perturbations

Consider gene expression matrix again:



Goal: how do genes 'affect' other genes?

Can we deduce 'network'?

⇒

Effect of perturbations

→ Make each experiment a perturbation of a single gene process

↳ we increase the transcription rates of each gene in turn.

↳ measure the change in steady state concentrations of all genes

↳ get (co-) 'control coeff'

↳ summarise regulatory strengths matrix

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

Effect of perturbations of gene transcription rates on S.S. concentrations

Upshot

- Can summarise in regulatory strengths matrix / Rd
- Can use to draw potential gene regulatory network

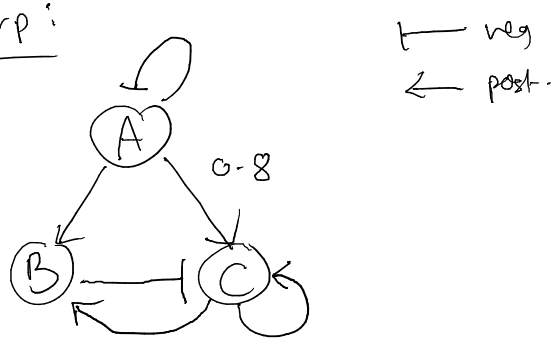
Example

rate inverse

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

SS. conc change

Network Interp:



negative autoreg : A

positive autoreg : C

|Rd|=

Normalised change in exp. level	Experiment: Rate ↑ of gene 1	gene 2	...
Δ Gene 1 level	-1	0.5	-
Δ Gene 2 level	-0.2	1	-
⋮	⋮	⋮	⋱