Bioeng 261 Microarrays

- Basic biology (very brief)
- Data / Experiment types
- Basic analysis methods

Upshot: Know & interpret:

- matrix types
  - gene expression matrix } many
  - ↳ gives { distance matrix / regulatory strength matrix
- dendrograms
  - ↳ summarise distance matrix / clustering alg.
- regulatory networks
  - ↳ summarise regulatory strength matrices

Background Biology (see slides & video link)

Idea

Want to measure mRNA levels (gene expressions)

- microarrays can measure 1000s of mRNA levels at a time
  - ↳ i.e. 1000s of genes at a time
  - ↳ a grid of 'spots' containing cDNA (complementary DNA)

'reverse transcription' (get DNA from RNA)

↳ 'Hybridisation'
↳ mRNA preferentially binds to corresp. cDNA

---

also → use fluorescent tags to distinguish samples
→ amplify via PCR/qPCR

Preprocessing

Preprocessing is crucial

BUT: we won't go into

Qs: - artifacts
- absolute vs relative expression
- log transformations

etc.

Typically work with

$\boxed{\log(\text{relative expression})}$

→ see slides for an example conversion

→ from now, take as given

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

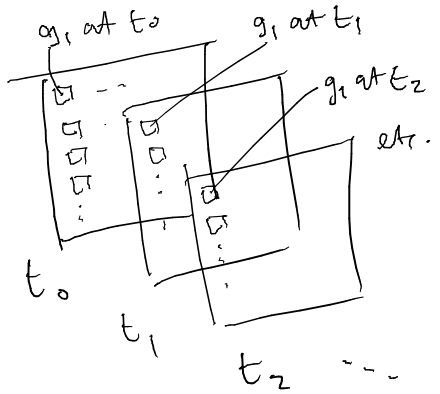
   } one sample type, multiple times

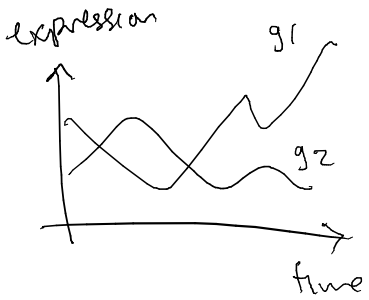
---

idea: • time has natural order (continuous or categ. ordinal)
• treatment/class not nec. ordered
- has cancer, doesn't, etc.

BUT essentially
⇒ just different independent vars, same basic ideas
⇒ also, often want to combine:
eg compare two time series from diff. treatments at same times

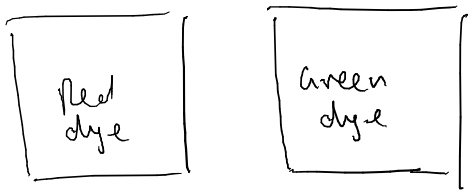
Time series





new chip for each time

Perturb./comparative



(separate)

treatment I control

(more commonly)

Who wins?

↑ red
↓ green
— yellow

↳ direction of treatment rel. to control.

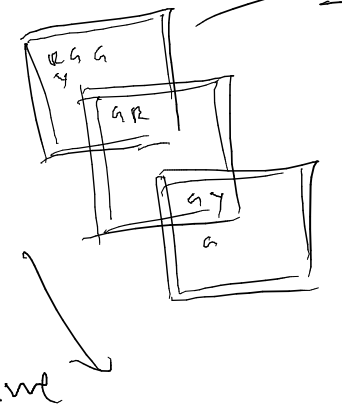
OR

Competitive

↳ [same chip] for both treatment & control

↳ compete to hybridize

Combined



each array has both treatment & control

track over time.

(new array per time)

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

Two key goals

- try to cluster 'similar' &/OR
- 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.

	Exp1	Exp2	Exp3
Gene1			
Gene2			
Gene3			
⋮			

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.

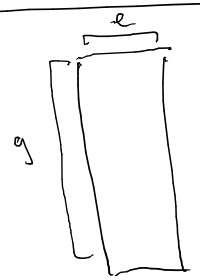
Gene expression matrices

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

	Exp1	Exp2	Exp3
Gene1			
Gene2			
Gene3			
⋮			

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 analysis
   (eg over time)

2. Perturbations based: linear control analysis

---

1. Clustering

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

Note: more precise distinctions

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

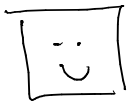
Unsupervised? Note on 'learning types'

Supervised

$X \longrightarrow Y$

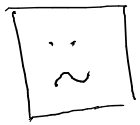
learn function $X \to Y$

↳ Train: supervisor examples


→ happy
(labels given)


→ sad

Test: predict on new set


→ happy ✗


→ happy ✓

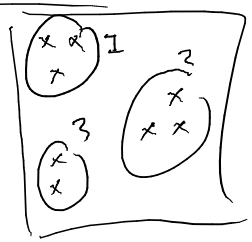
'Reasonably straightforward' to evaluate/validate.

unsupervised

$X \circlearrowleft$

Pattern discovery in X

Eg: Find 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

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

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:

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

squared diffs: 0^2 0^2 3^2 2^2

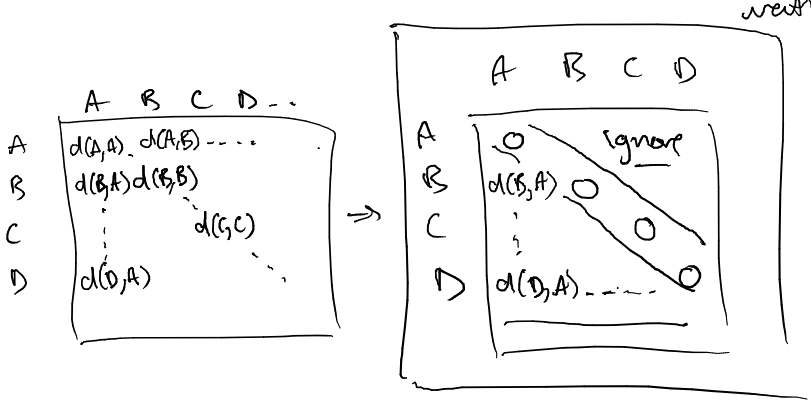
i.e. 0 0 9 4

sum squares: $\boxed{13}$

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

Distance matrices

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



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

Example: slides

Distance-based clustering

Goal: group together if close.

Two popular algorithms

- K-means
- Hierarchical

see slides for pseudocode

Here: consider hierarchical

→

Hierarchical clustering

Begin with n observations

Find pairwise distances

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

Find smallest distance

'Fuse' together

Recompute distances to 'fused' cluster

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

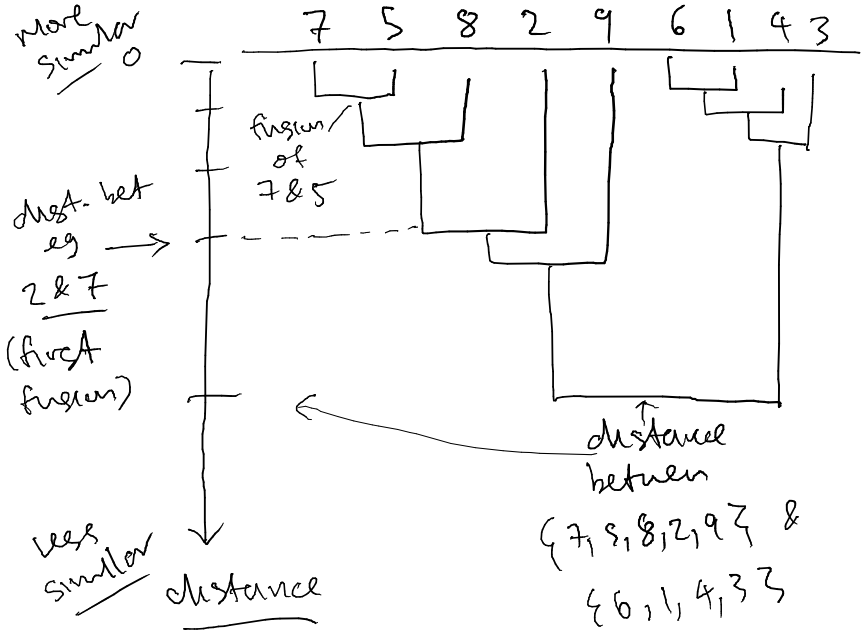
min
max
average
all
etc.

pics: see slides

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:

	Experiment 1	2	3	...
gene 1				
gene 2				
gene 3				
⋮				

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 $\boxed{\text{rates}}$ of each gene in turn.

↳ measure the change in steady state $\boxed{\text{concentrations}}$ of all genes

↳ get (co-) 'control coeff'

↳ summarise $\boxed{\text{regulatory strengths}}$ 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 $\boxed{\text{regulatory strengths matrix}}$ $|Rd|$
- Can use to draw potential gene regulatory network

$|Rd| =$

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	...
⋮	⋮	⋮	⋱

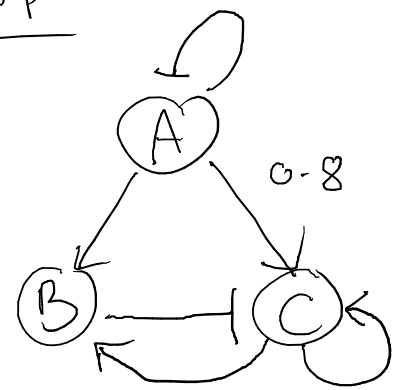
Example

rate increase

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

ss. conc change

Network Interp:



⊢ reg

← posit.

negative autoreg : A

positive autoreg : C