For written notes on this lecture, please read chapter 14 of *The Practical Bioinformatician*.

CS2220: Introduction to Computational Biology
Unit 2: Gene expression analysis

Li Xiaoli

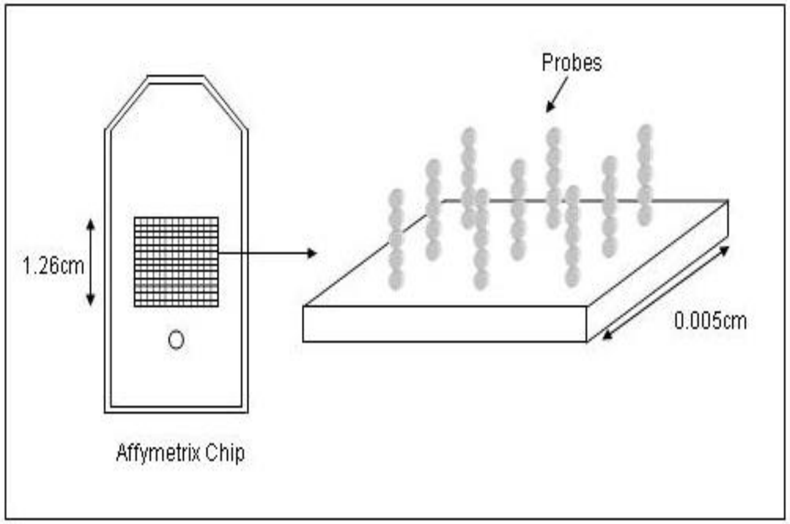
25 August 2016

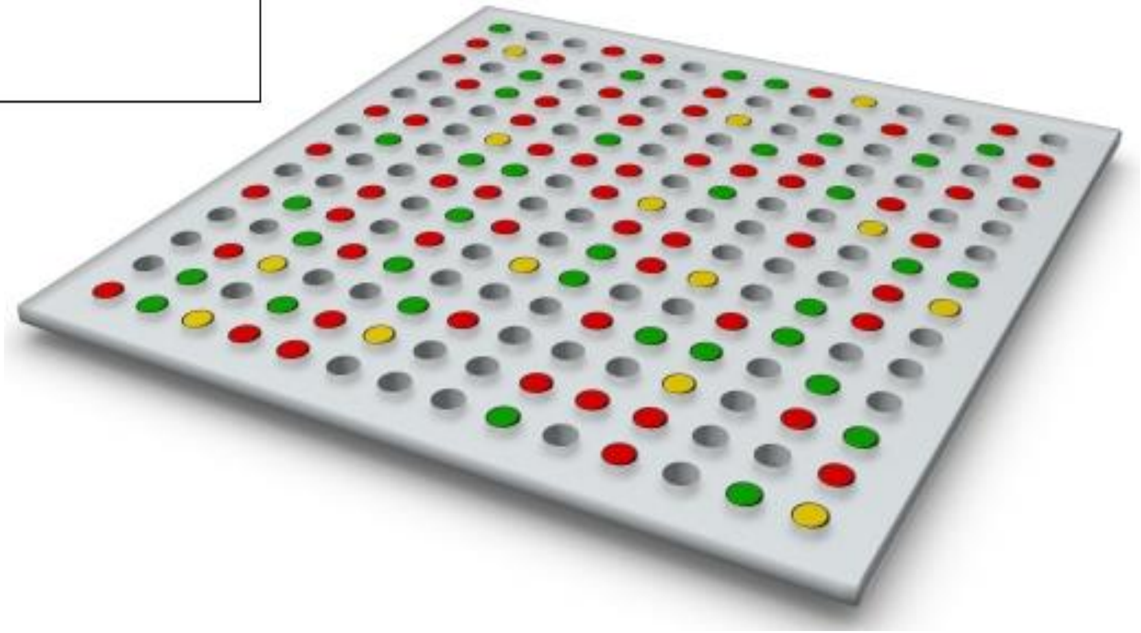
NUS
National University of Singapore

Plan

- **Microarray background**
- **Gene expression profile clustering**
- **Some standard clustering methods**

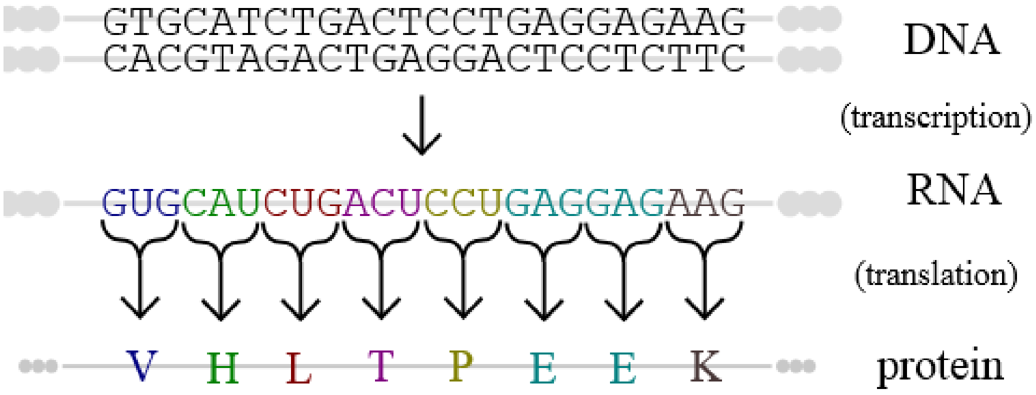
Background on microarrays





What is a microarray?

- **Gene expression is the process by which info from a gene is used in the synthesis of a functional gene products, e.g. functional RNA, proteins**



- **Genes are expressed by being transcribed into RNA, and this transcript may then be translated into protein**
  http://en.wikipedia.org/wiki/Gene_expression

What is a microarray?

- Contain large number of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers
- Detect what genes are being expressed in a cell of a tissue sample
- Measure expression of thousands of genes simultaneously

Good intro videos on microarrays

- **Short Video (1-3 min each)**
  - http://www.youtube.com/watch?v=_6ZMEZK-alM
  - http://www.youtube.com/watch?v=VNsThMNjKhM
  - http://www.youtube.com/watch?v=SNbt--d14P4

- **Long Video (25 min)**
  - http://www.youtube.com/watch?v=0Hj3f7vQFZU

Wet-lab experiments

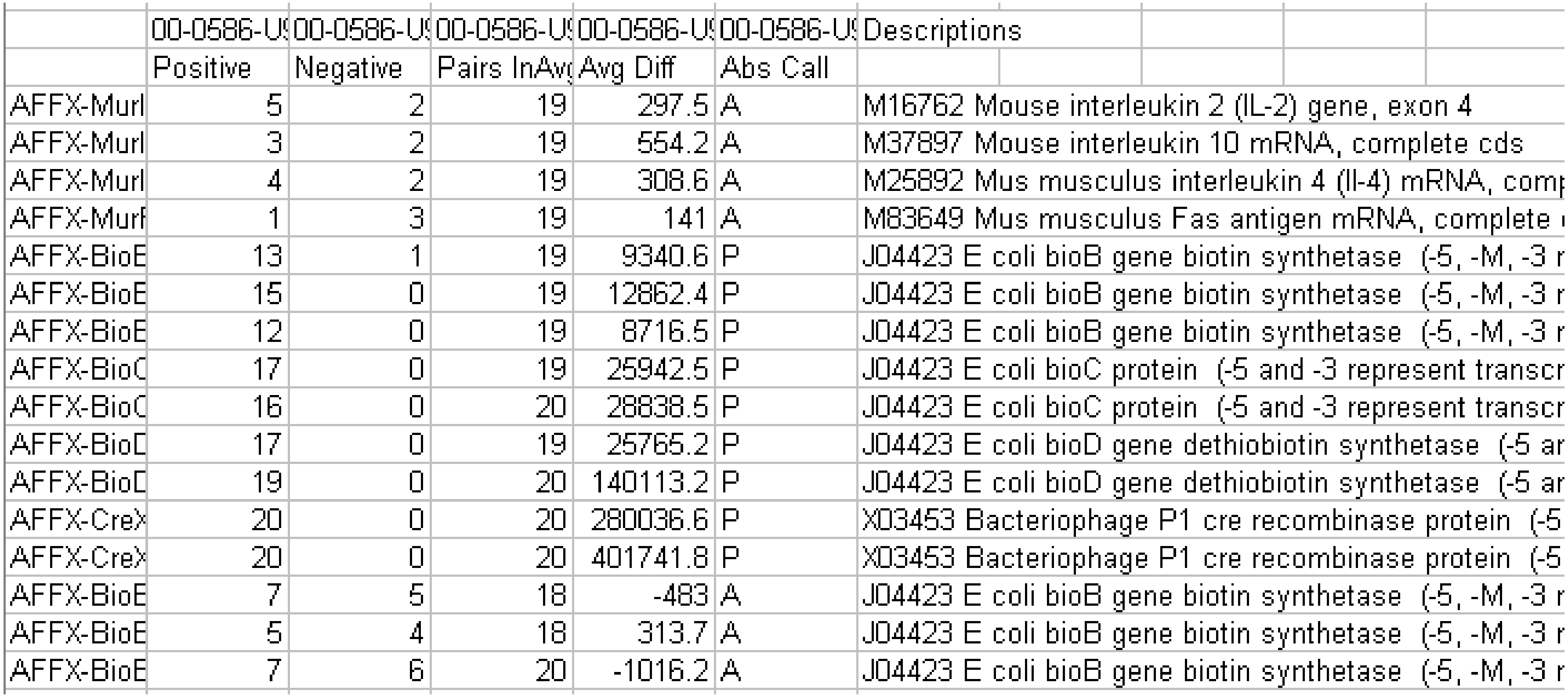
- **Key idea: If a gene is expressed, then it generates mRNA. When we produce cDNA from mRNA, cDNA and DNA will anneal and bind together**

According to base pairing rules (A with T and C with G), *hydrogen bonds* bind the bases of the two separate polynucleotide strands (DNA, cDNA) together

How to do Wet Lab experiments

http://www.bio.davidson.edu/Courses/genomics/chip/chip.html

Sample Affymetrix GeneChip data (U95A)



	00-0586-U9	00-0586-U9	00-0586-U9	00-0586-U9	00-0586-U9	Descriptions
	Positive	Negative	Pairs InAvg	Avg Diff	Abs Call	
AFFX-MurI	5	2	19	297.5	A	M16762 Mouse interleukin 2 (IL-2) gene, exon 4
AFFX-MurI	3	2	19	554.2	A	M37897 Mouse interleukin 10 mRNA, complete cds
AFFX-MurI	4	2	19	308.6	A	M25892 Mus musculus interleukin 4 (Il-4) mRNA, comp
AFFX-MurF	1	3	19	141	A	M83649 Mus musculus Fas antigen mRNA, complete
AFFX-BioB	13	1	19	9340.6	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	15	0	19	12862.4	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	12	0	19	8716.5	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC protein (-5 and -3 represent transcr
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC protein (-5 and -3 represent transcr
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD gene dethiobiotin synthetase (-5 ar
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD gene dethiobiotin synthetase (-5 ar
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacteriophage P1 cre recombinase protein (-5
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacteriophage P1 cre recombinase protein (-5
AFFX-BioB	7	5	18	-483	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	5	4	18	313.7	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	7	6	20	-1016.2	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r

The impt field is "Avg Diff", which gives the expression level of the gene. The "Abs Call" field is also impt, which tells whether the corresponding number in the "Avg Diff" field is reliable or not. "P" means present and thus the number is reliable. "A" and "M" tell you the number is unreliable and should be ignored.

http://yfgdb.princeton.edu/Affymetrix_Empirical.txt

Some biological knowledge on gene expression regulation

- **Regulation of gene expression refers to the control of the amount and timing of appearance of the functional product of a gene**

- **Control of expression is vital to allow a cell to produce the gene products it needs when it needs them; in turn this gives cells the flexibility to adapt to a variable environment, external signals, damage to the cell**

The patchy colours of a tortoiseshell cat are the result of different levels of expression of pigmentation genes in different areas of the skin.

Gene types depending on how they are regulated

- **A constitutive gene continually transcribes to mRNA**
- **A housekeeping gene is typically a constitutive gene that is transcribed at a relatively constant level**
  - A housekeeping gene's products are typically needed for maintenance of the cell
- **A facultative/ inducible gene is a gene only transcribed when needed as opposed to a constitutive gene**
  - Its expression is either responsive to environmental change or dependent on the position in the cell cycle

Example of real gene expression data

- http://nemates.org/uky/520/Lab/lab10/yeastall_public.txt

- **Exercise: store the whole gene expression data into a excel file to understand more**

Type of gene expression datasets

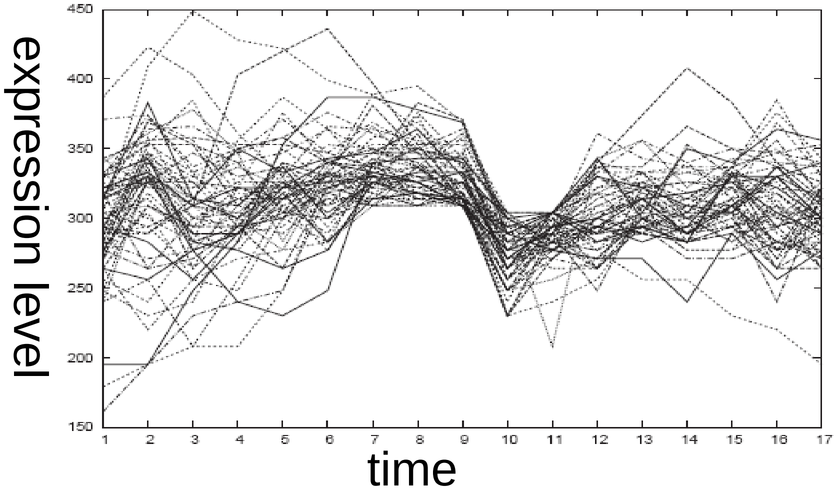
- Gene-Conditions or **Gene-Sample** (**numeric** or discretized)

1000 - 100,000 columns

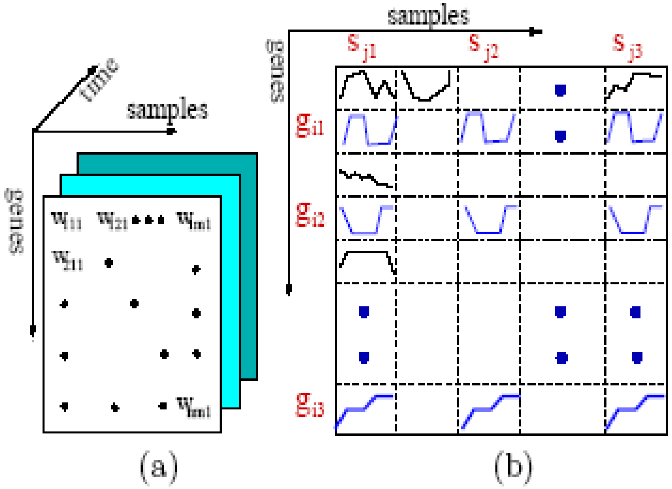
100-500 rows

	Class	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	.....	
Sample1	Cancer	0.12	-1.3	1.7	1.0	-3.2	0.78	-0.12		
Sample2	Cancer							1.3		
.										
	~Cancer									
SampleN	~Cancer									

- Gene-Time (different genes)



- Gene-Sample-Time



Type of gene expression datasets

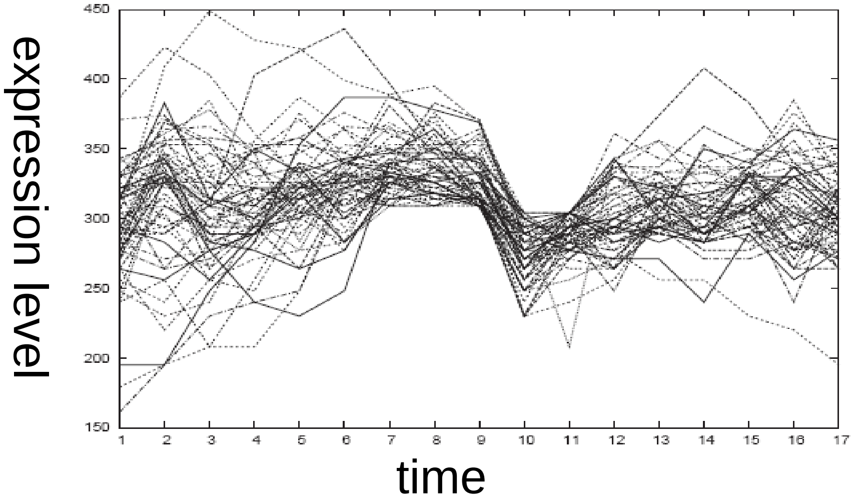
- Gene-Conditions or **Gene-Sample** (numeric or **discretized**)

1000 - 100,000 columns

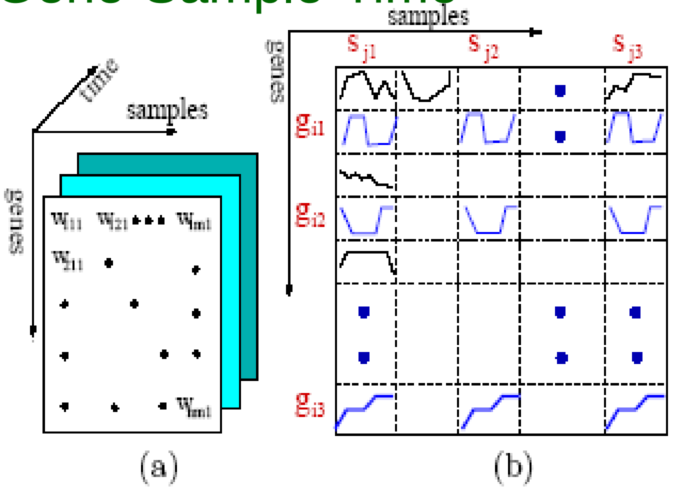
100-500 rows

	Class	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	.....	
Sample1	Cancer	1	0	1	1	1	0	0		
Sample2	Cancer							1		
.										
	~Cancer									
SampleN	~Cancer									

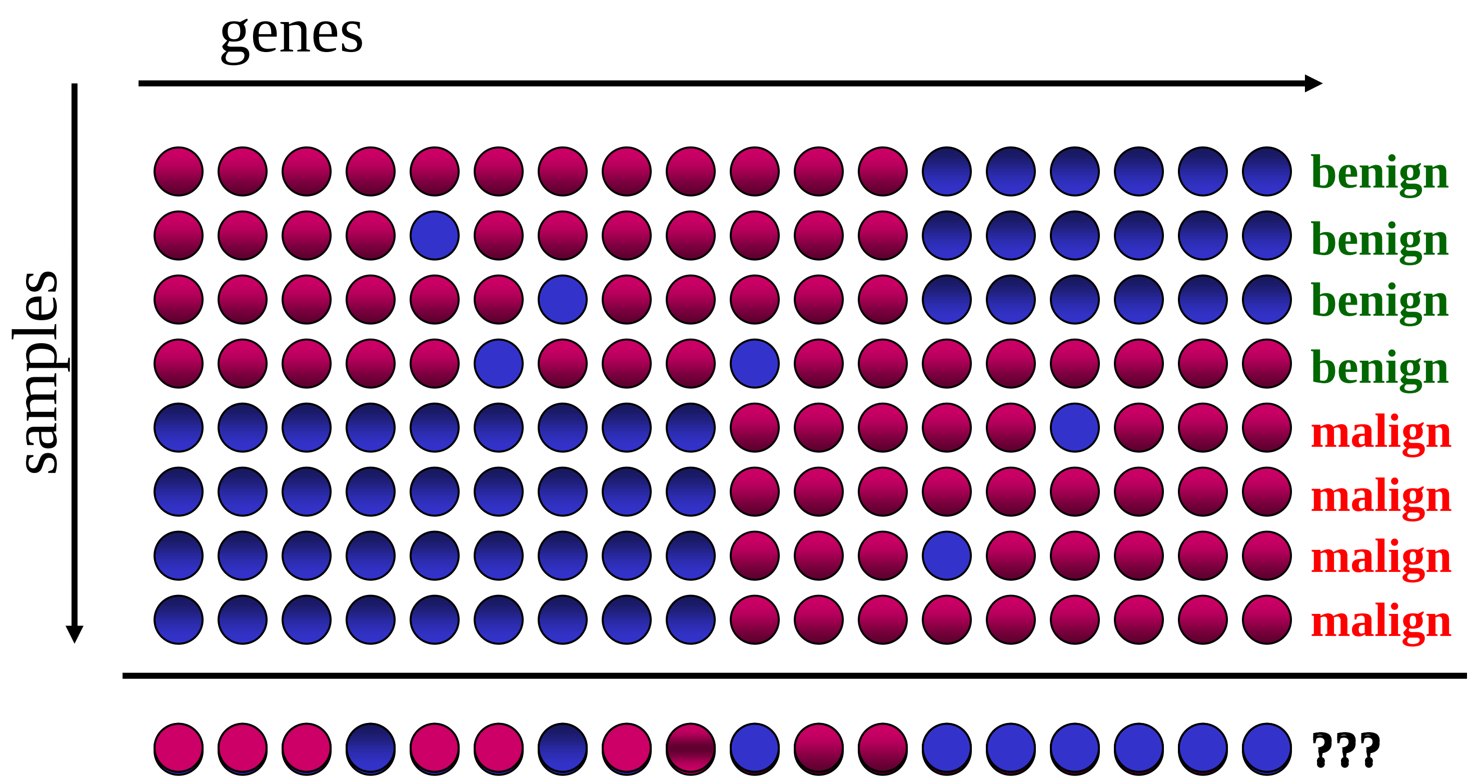
- Gene-Time



- Gene-Sample-Time

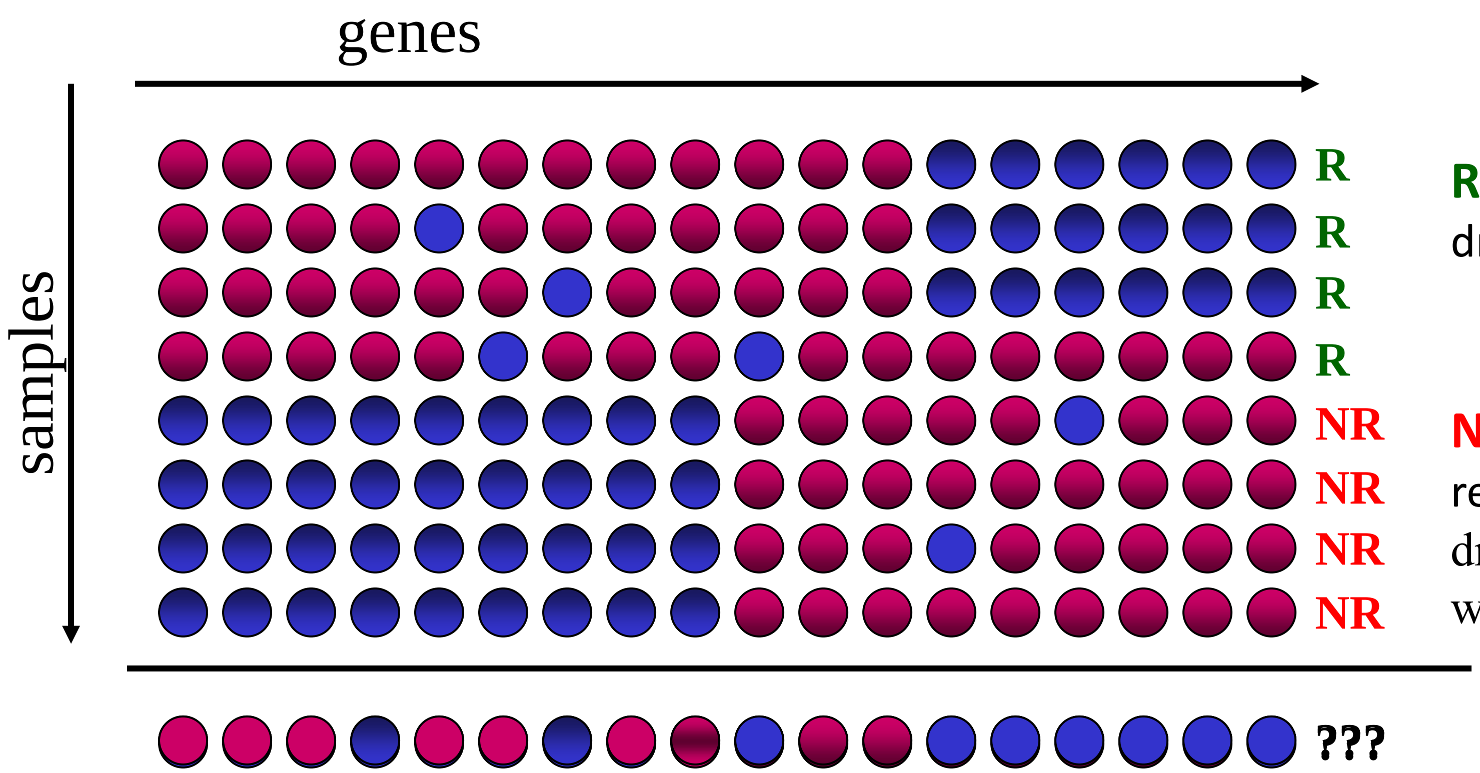


Application: Disease diagnosis



Gene expression data to perform diagnostic task

Application: Treatment prognosis

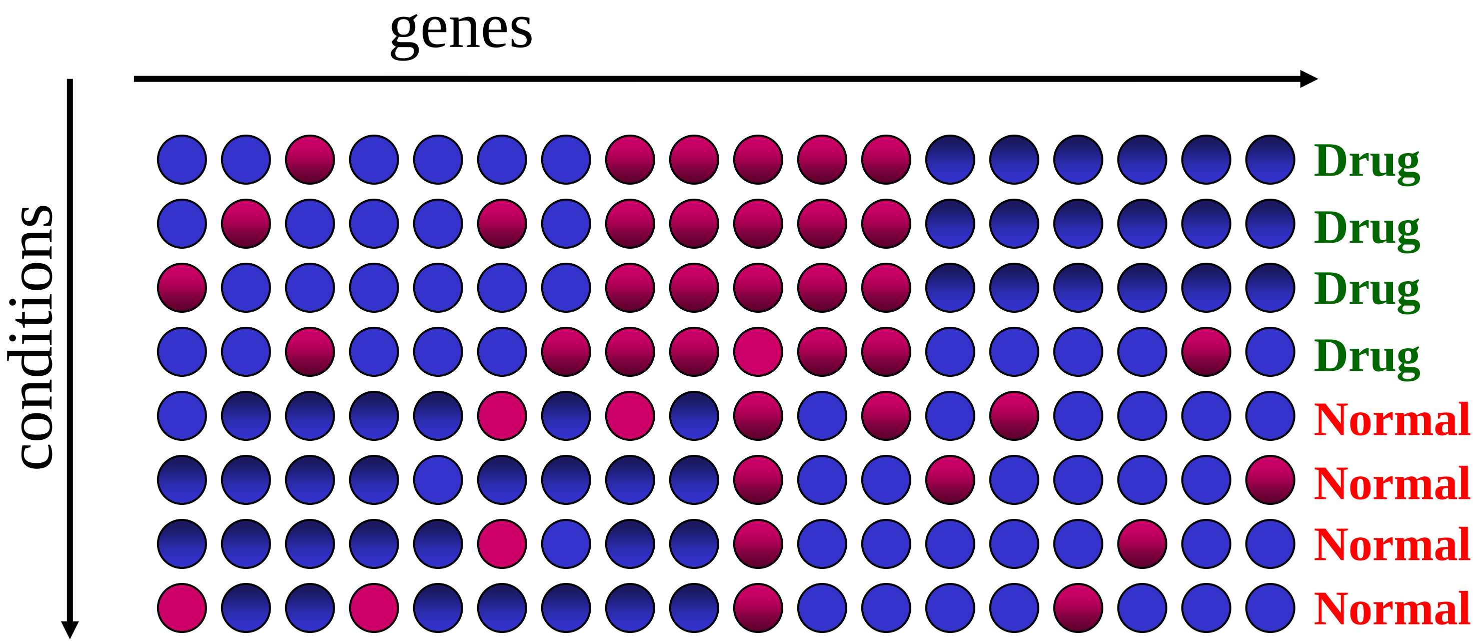


R: Responder, drug is working

NR: Non-responder, drug is not working

Identify the biomarkers of people who will benefit from continued used of the drug. We can thus predict the treatment outcomes, e.g. working or not-working or should we give a patient the treatment?

Application: Drug action detection



Normal: The control tissues

Drug: The same tissue after injecting the drug

Which group of genes are the drug affecting on?

With drugs, which the gene expression values have big changes?

Gene expression profile clustering

- **Novel Disease Subtype Discovery**

Childhood acute lymphoblastic leukemia (ALL)

- **Existing known subtypes in 2000:**
  - T-ALL,
  - E2A-PBX,
  - TEL-AML,
  - BCR-ABL,
  - MLL genome rearrangements,
  - Hyperdiploid>50

Type of gene expression datasets

- **Gene-Sample** (numeric)

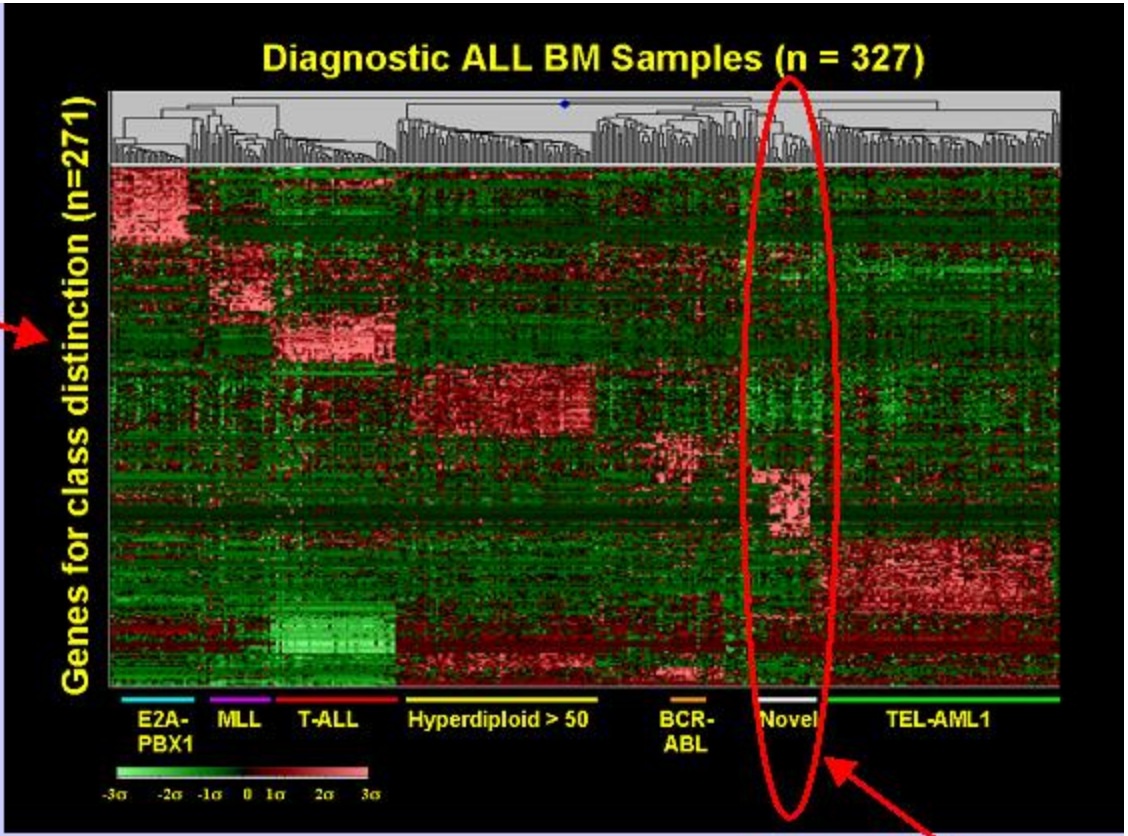
100-500 Samples /columns

1000 - 100,000 rows/ genes

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	.....
Gene 1	0.12	0.34	-0.23	-0.34	0.28	0.11	0.23	
Gene 2								
.								
Gene N								

Is there a new subtype?

- **Hierarchical clustering of gene expression profiles reveals a novel subtype of childhood ALL**



Genes selected by χ2

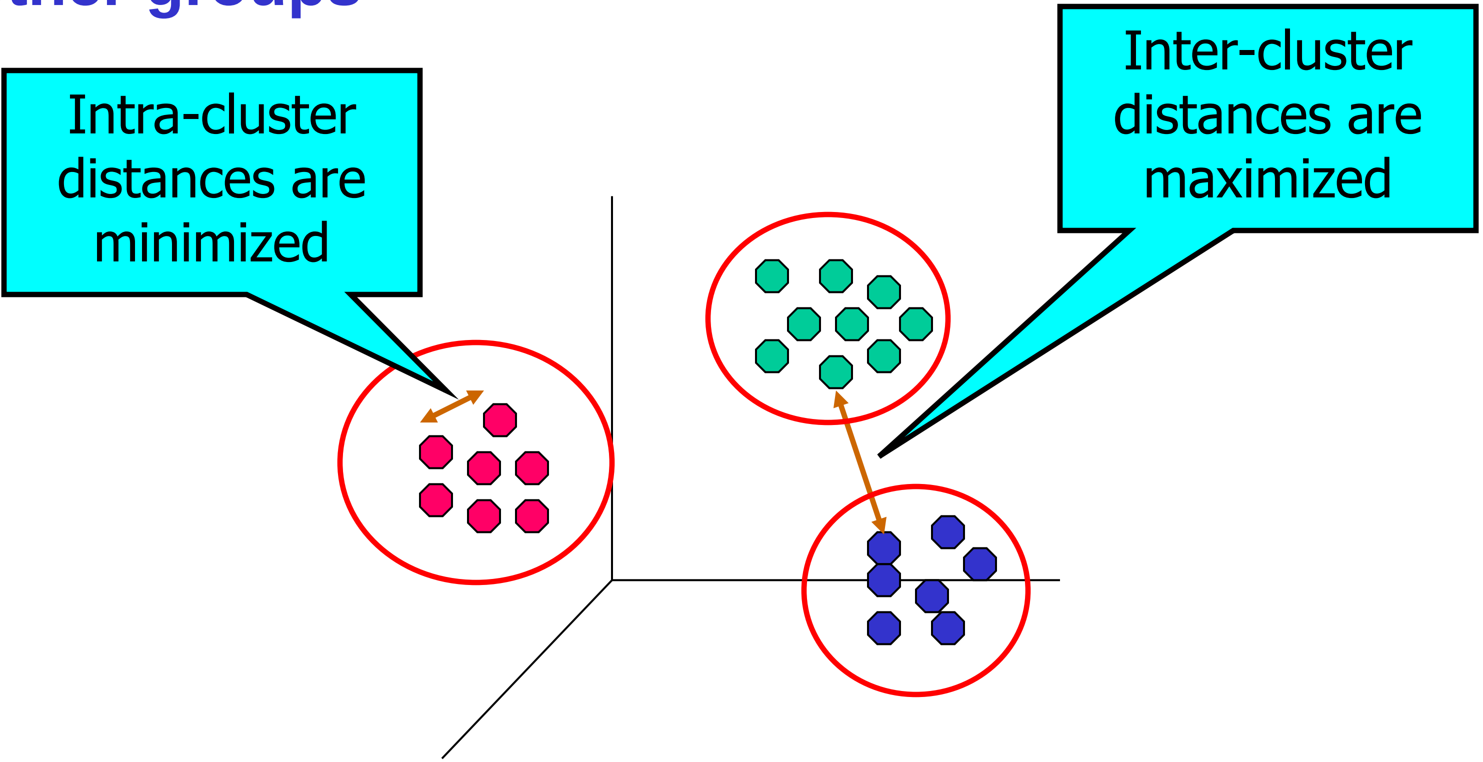
New subtype discovered

Clustering methods

- **K-means**
- **Hierarchical Clustering**

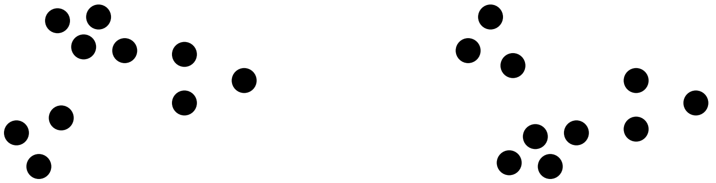
What is cluster analysis?

- **Finding groups of objects such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups**

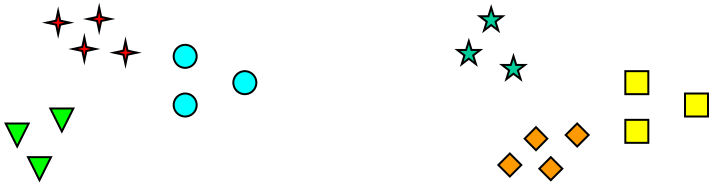

Intra-cluster distances are minimized

Inter-cluster distances are maximized

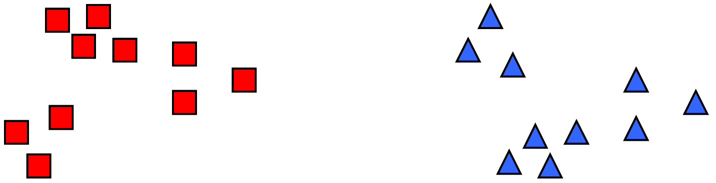
Notion of a cluster can be ambiguous



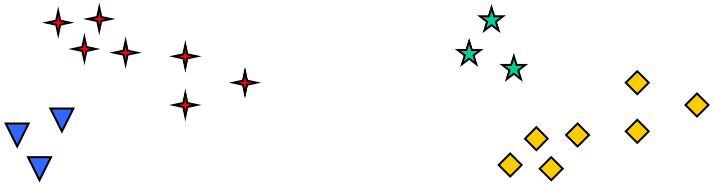
How many clusters?



Six Clusters



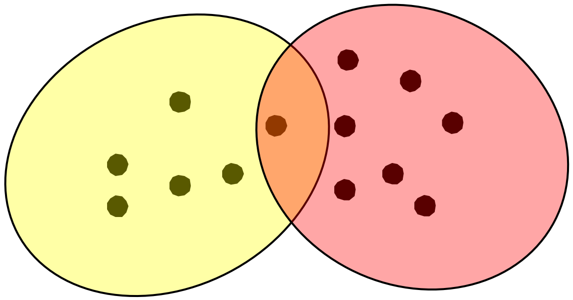
Two Clusters



Four Clusters

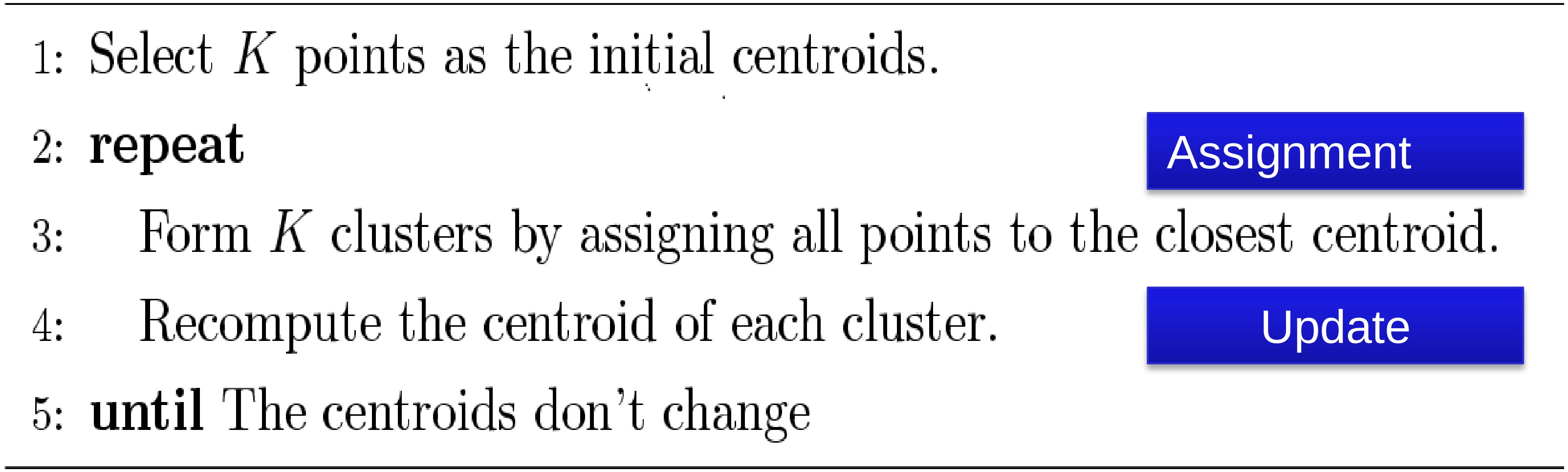
We use colors to represent the clustering results/groups

We could also have



K-means clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

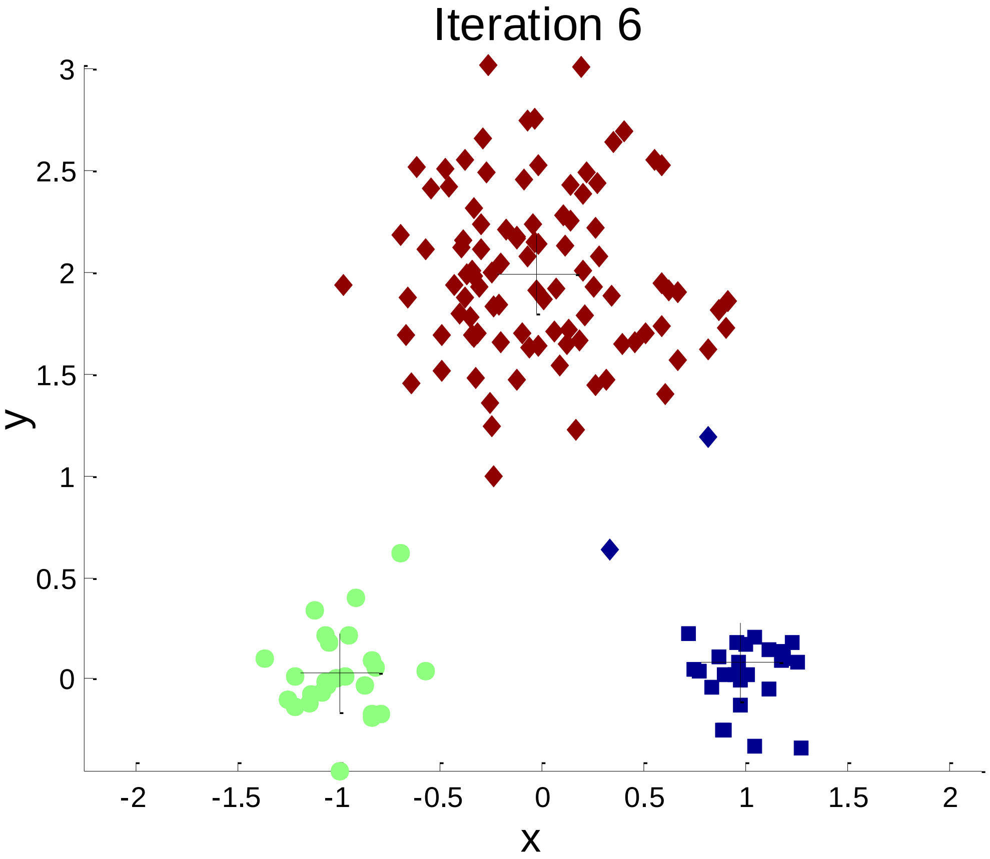


```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

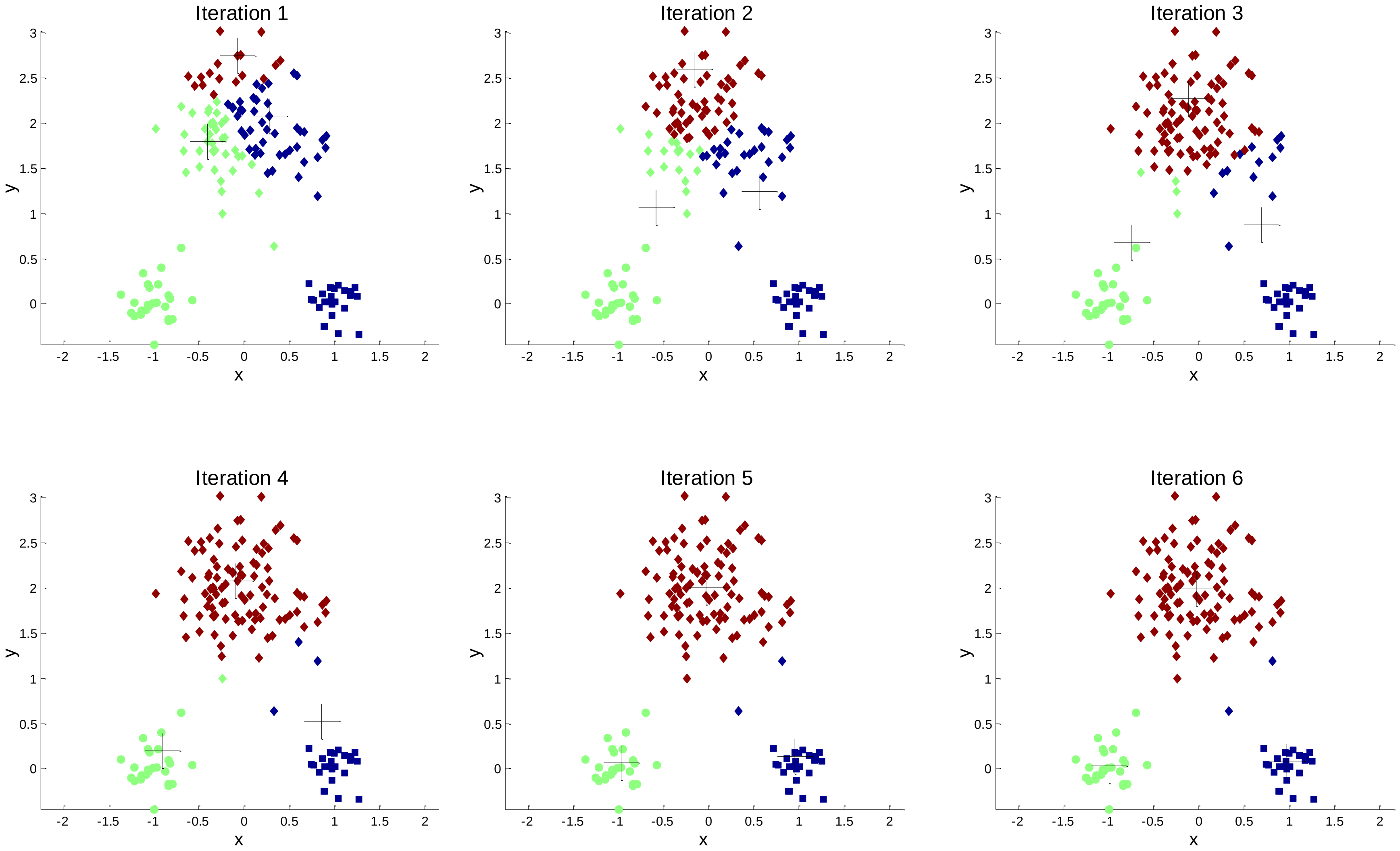
Assignment

Update

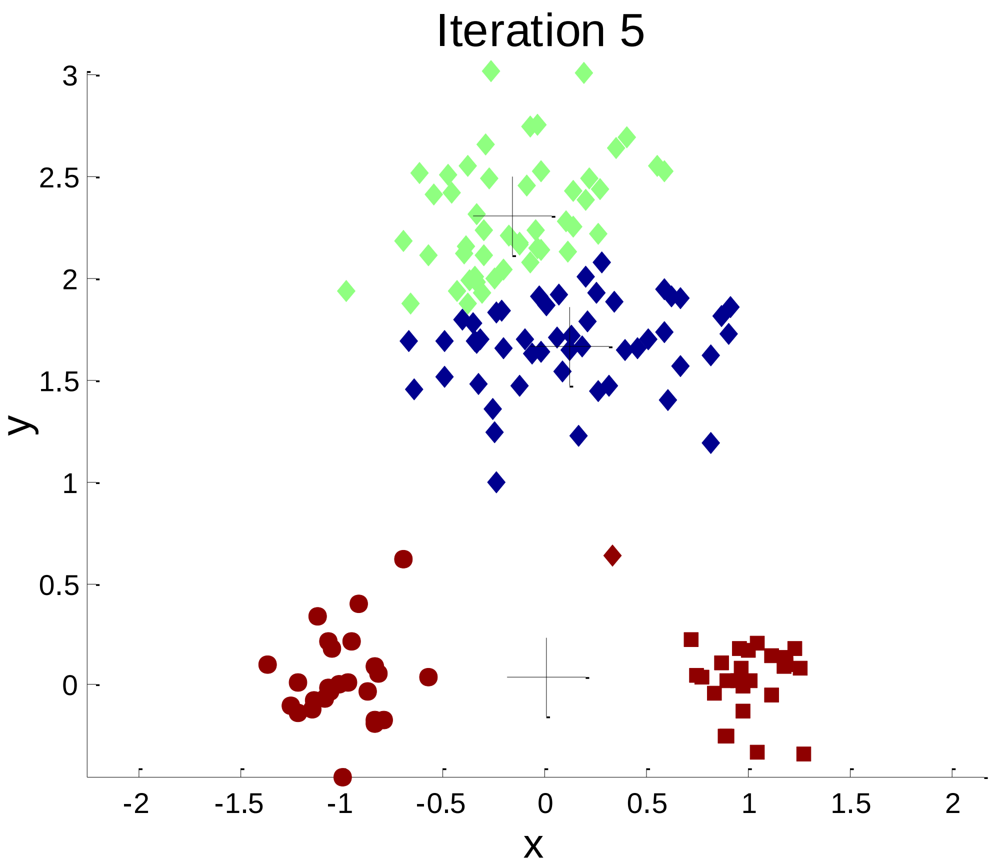
K-means clustering illustration



K-means clustering illustration



Importance of choosing initial centroids



Hierarchical clustering

- **Two main types of hierarchical clustering**
  - Agglomerative:
    - **Start with the points as individual clusters**
    - **At each step, merge the closest pair of clusters until only one cluster (or k clusters) left**
  - Divisive:
    - **Start with one, all-inclusive cluster**
    - **At each step, split a cluster until each cluster contains a point (or there are k clusters)**
- **Traditional hierarchical algorithms use a similarity or distance matrix**
  - Merge or split one cluster at a time

Agglomerative clustering algo

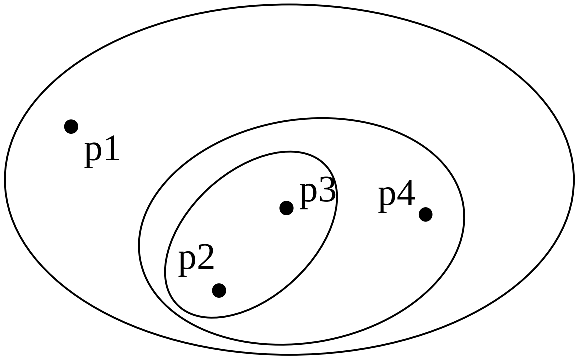
- **More popular hierarchical clustering technique**
- **Basic algorithm is straightforward**
  - Compute the proximity matrix
  - Let each data point be a cluster
  - Repeat
  - Merge the two closest clusters
  - Update the proximity matrix
  - Until only a single cluster remains

Merge

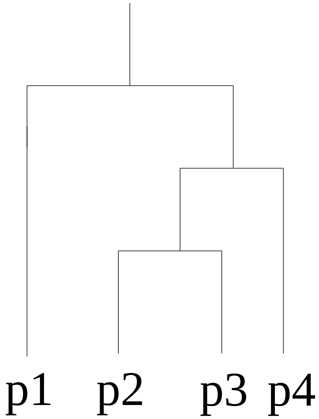
Update

- **Key operation is computation of the proximity of two clusters**
  - Different approaches to defining the distance / similarity betw clusters

Visualization of agglomerative hierarchical clustering

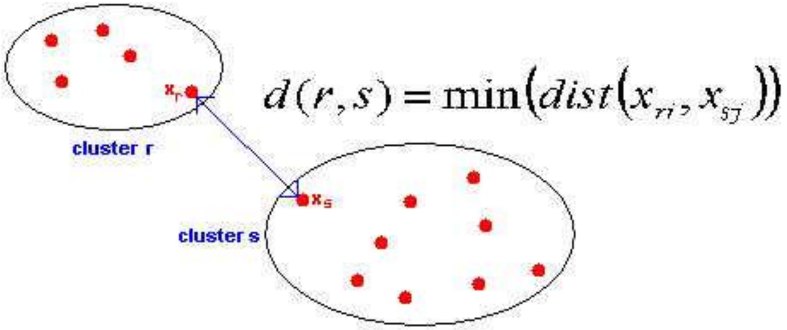


Traditional Hierarchical Clustering

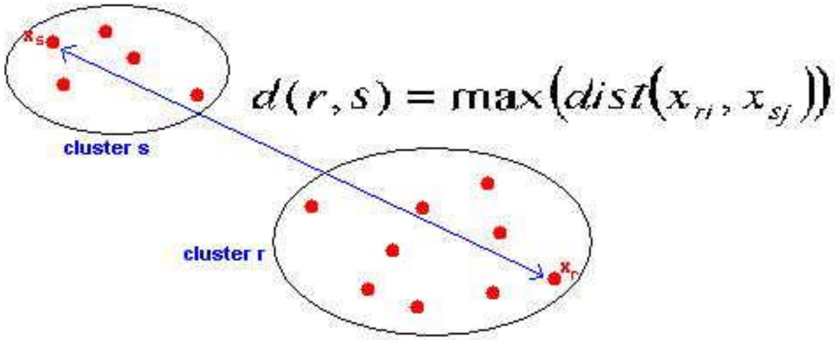


Traditional Dendrogram

Single, complete, & average linkage



$$d(r,s) = \min(dist(x_{ri}, x_{sj}))$$



$$d(r,s) = \max(dist(x_{ri}, x_{sj}))$$

Single linkage defines distance betw two clusters as min distance betw them

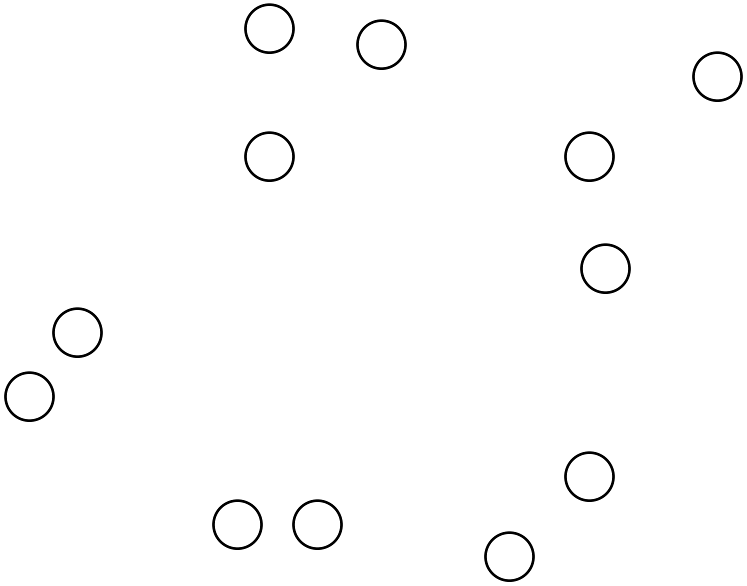
Complete linkage defines distance betw two clusters as max distance betw them

Exercise: Give definition of "average linkage"

Image source: UCL Microcore Website

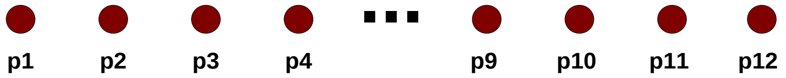
Simulation: Starting situation

- **Start with clusters of individual points and a proximity matrix**



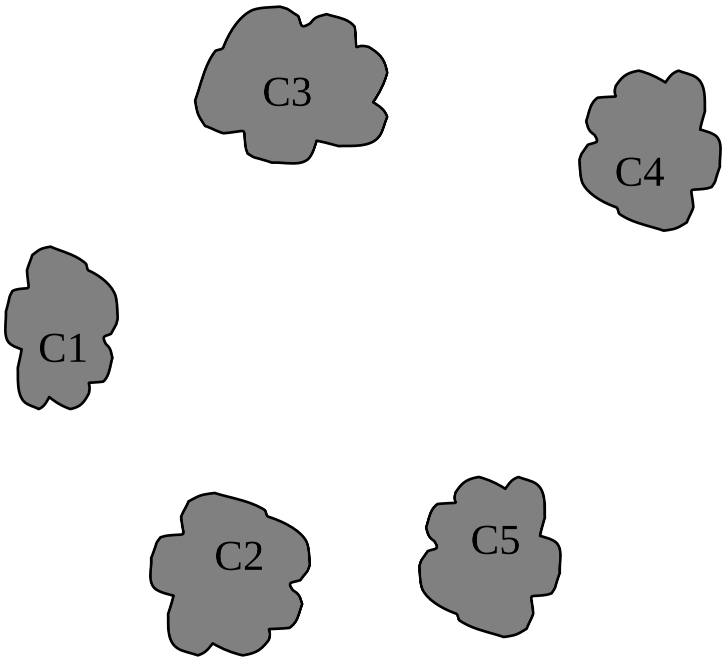
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



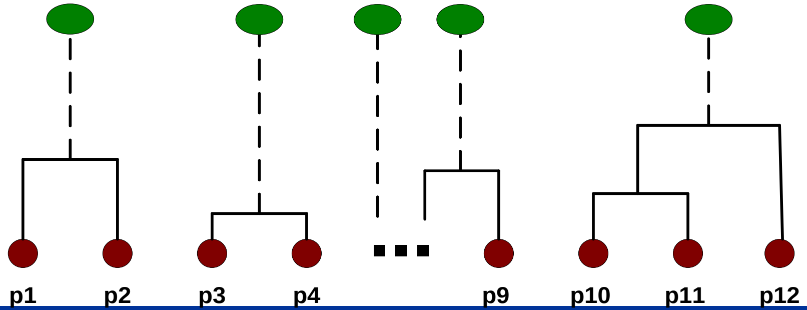
Intermediate situation

- **After some merging steps, we have some clusters**



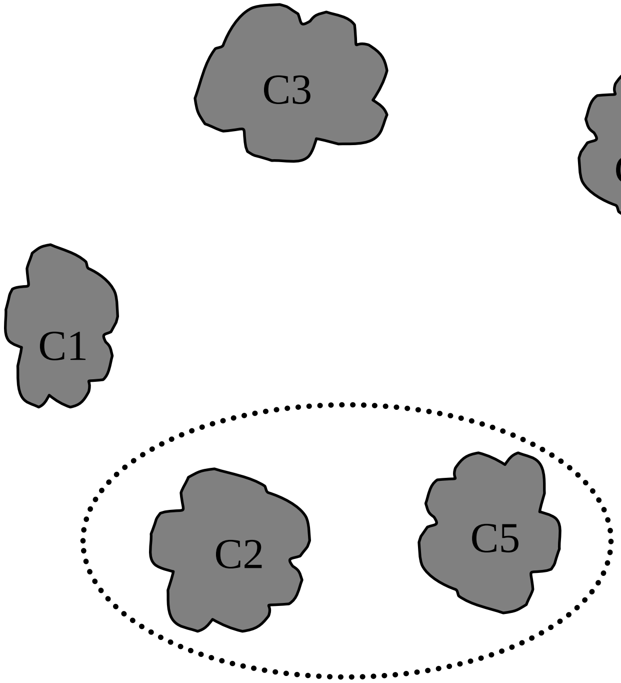
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



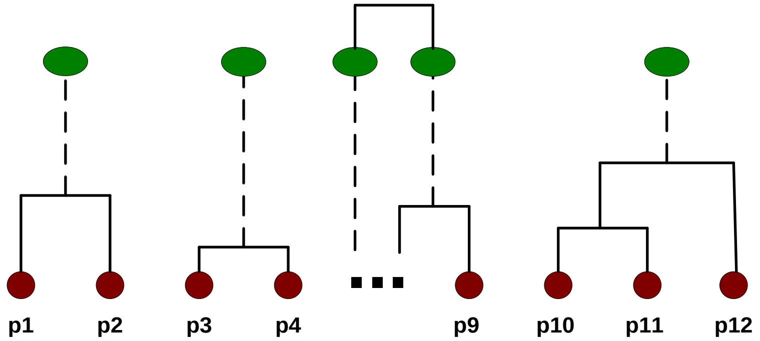
Intermediate situation

- **We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.**



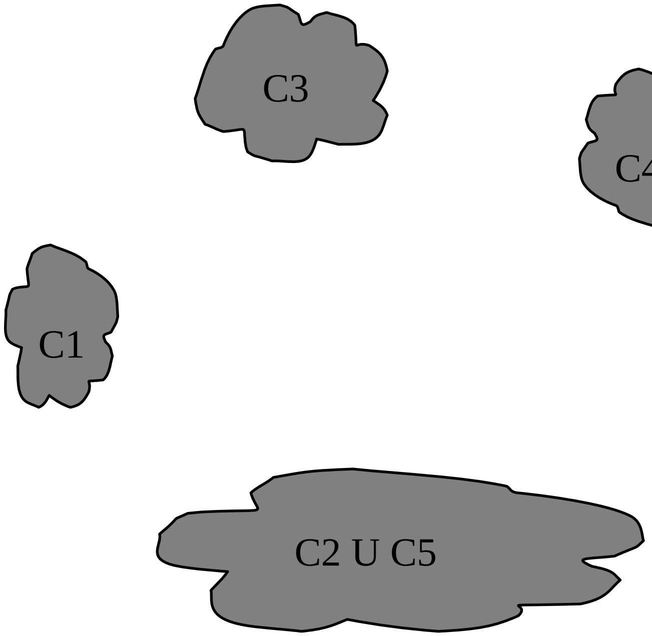
	C1	C2	C3	C4	C5
C1		▨			▨
C2	▨	▨	▨	▨	▨
C3		▨			▨
C4		▨			▨
C5	▨	▨	▨	▨	▨

Proximity Matrix



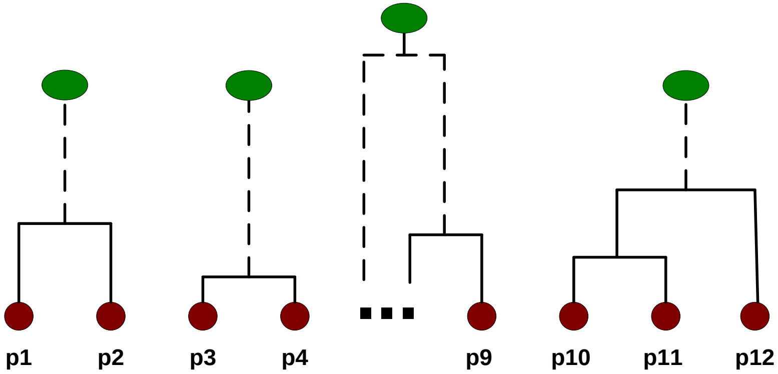
After merging

- The question is “How do we update the proximity matrix?”

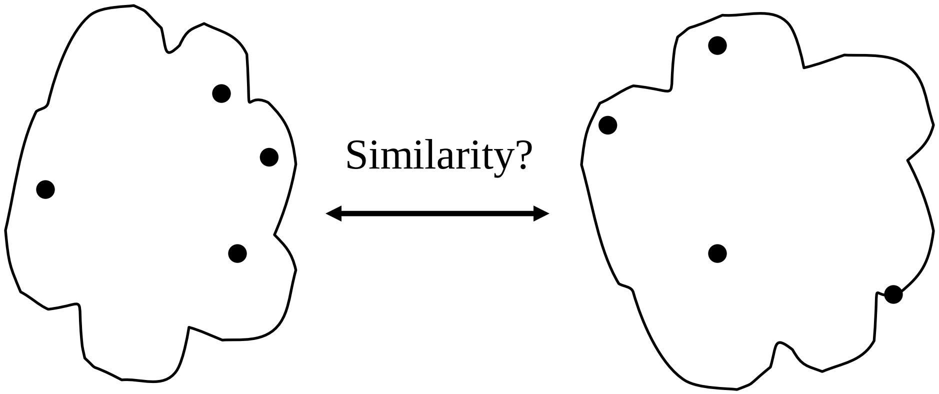


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to define inter-cluster similarity

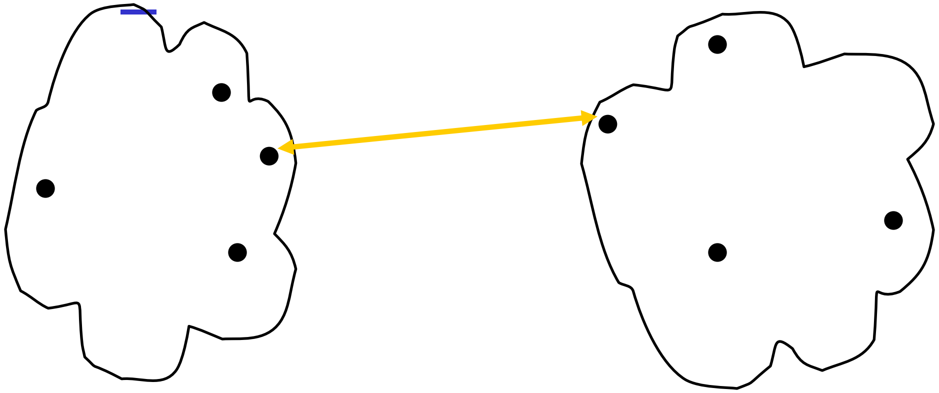


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

How to define inter-cluster similarity

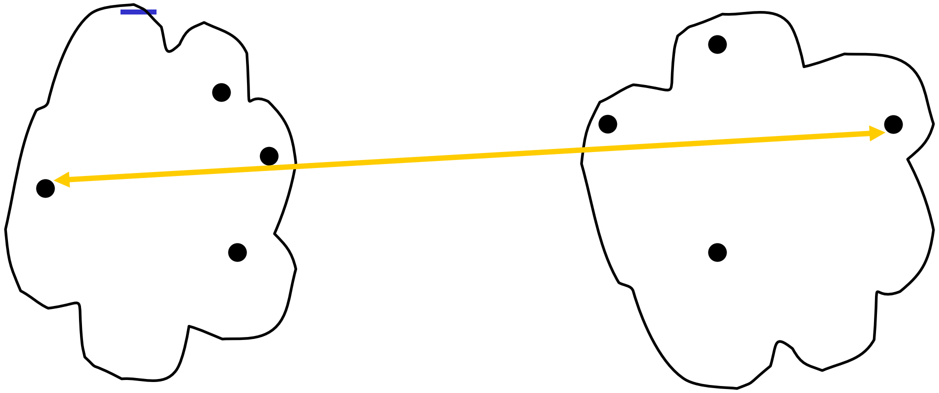


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

How to define inter-cluster similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						

.

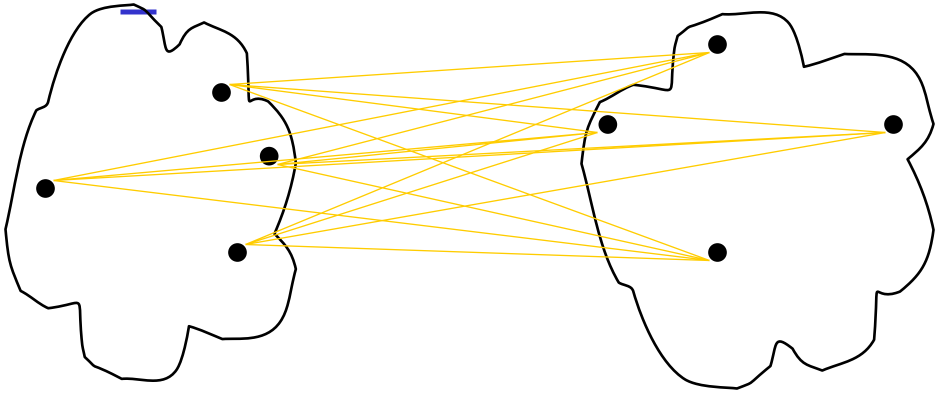
.

.

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

How to define inter-cluster similarity

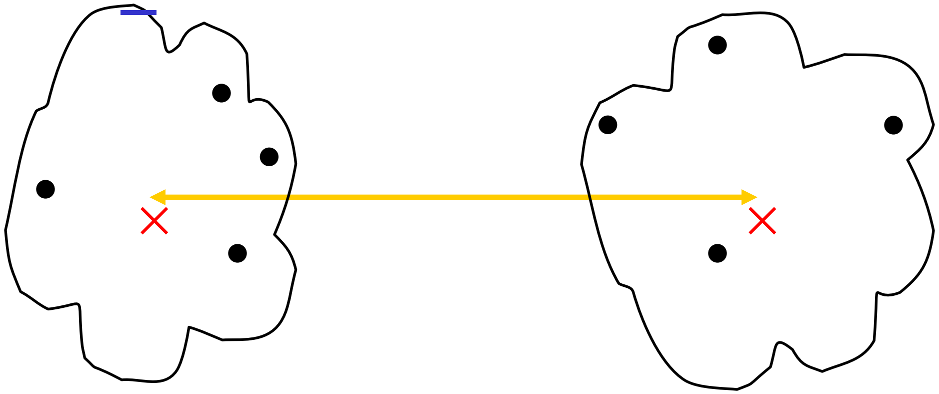


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

How to define inter-cluster similarity



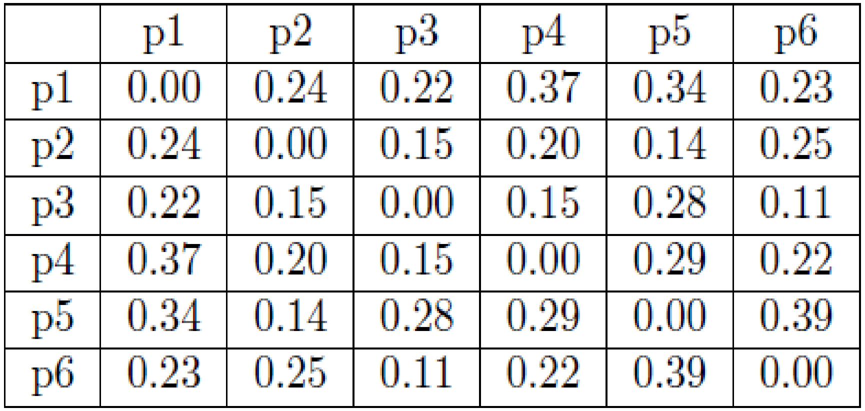
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- Min
- Max
- Group average
- Distance between centroids

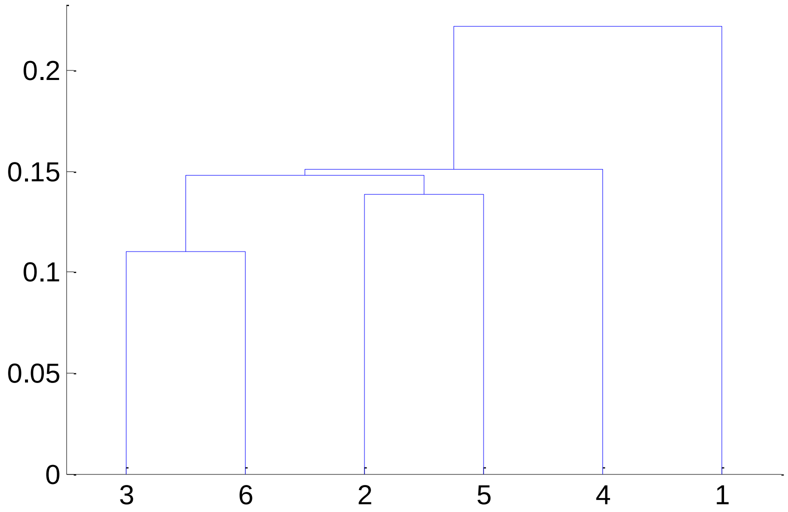
Cluster similarity: Min or single link

- **Similarity of two clusters is based on the two most similar (closest) points in the different clusters**
  - Determined by one pair of points, i.e., by one link in the proximity graph

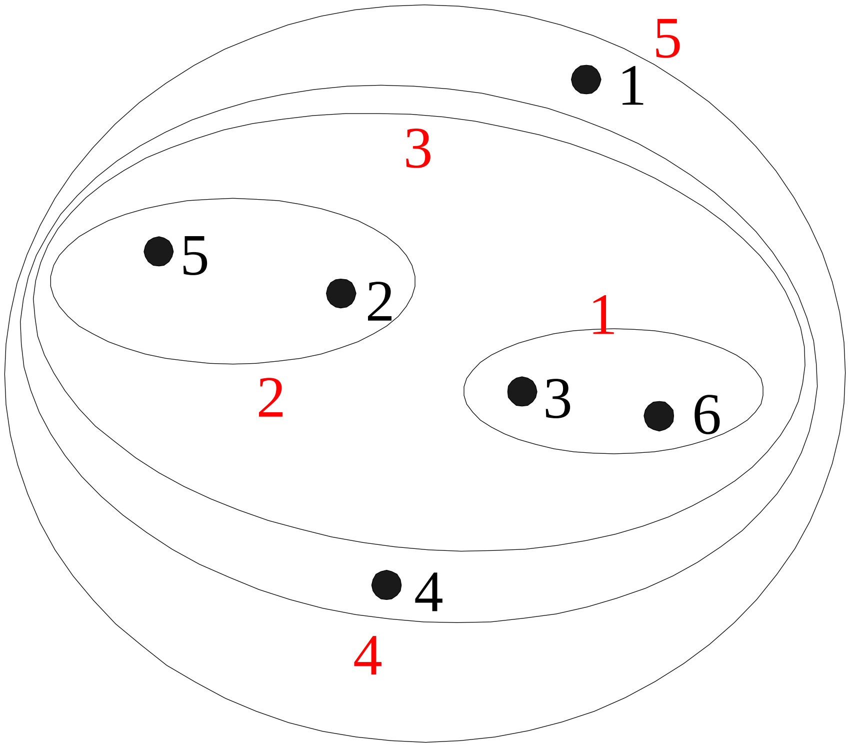


	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

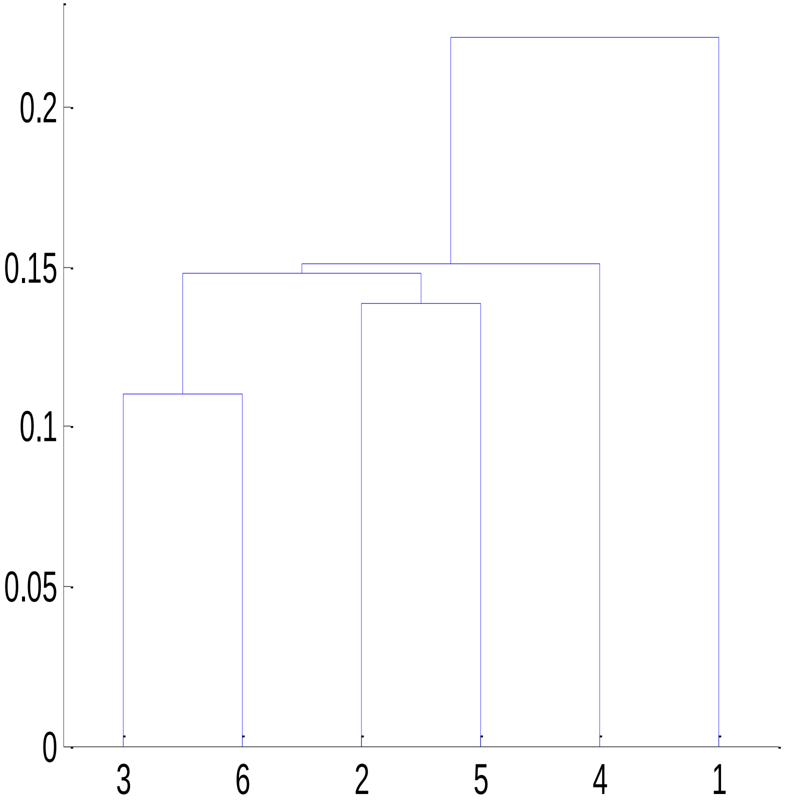
Table 8.4. Euclidean distance matrix for 6 points.



Hierarchical clustering: Min



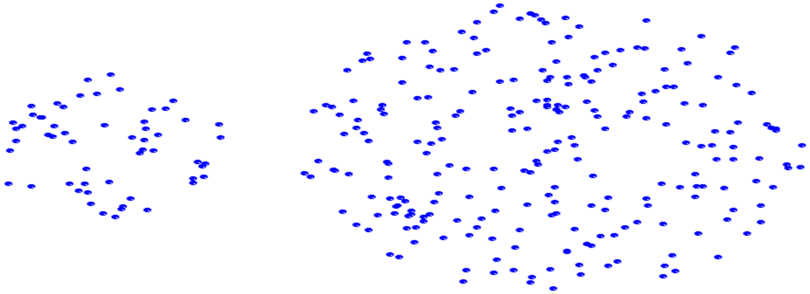
Single Link Clustering



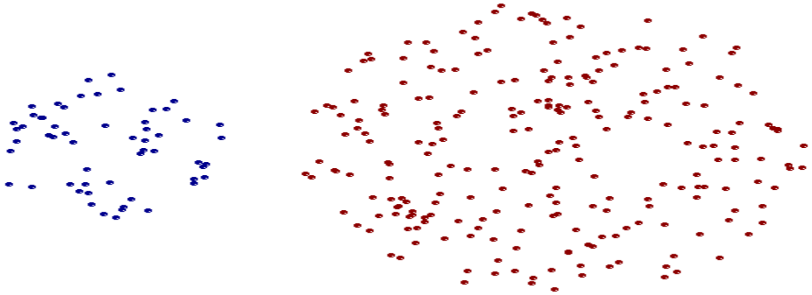
Single Link Dendrogram

Strength of Min

- **Can handle non-elliptical shapes**



Original Points

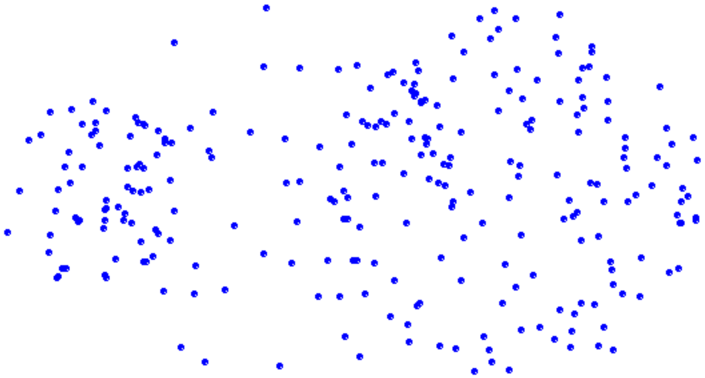


Two Clusters

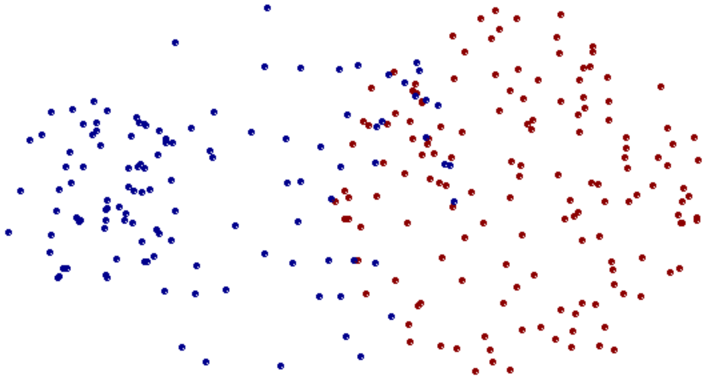
The algo likely to merge the points within same clusters if they are clearly separated

Limitations of Min

- **Sensitive to noise and outliers: cc**



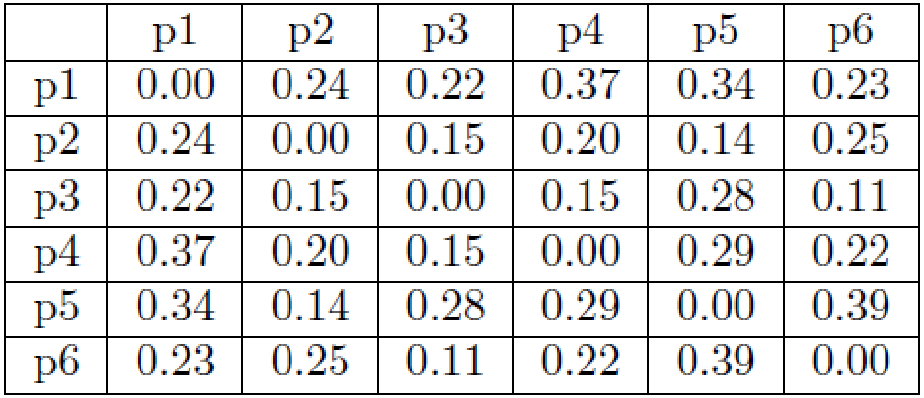
Original Points



Two Clusters

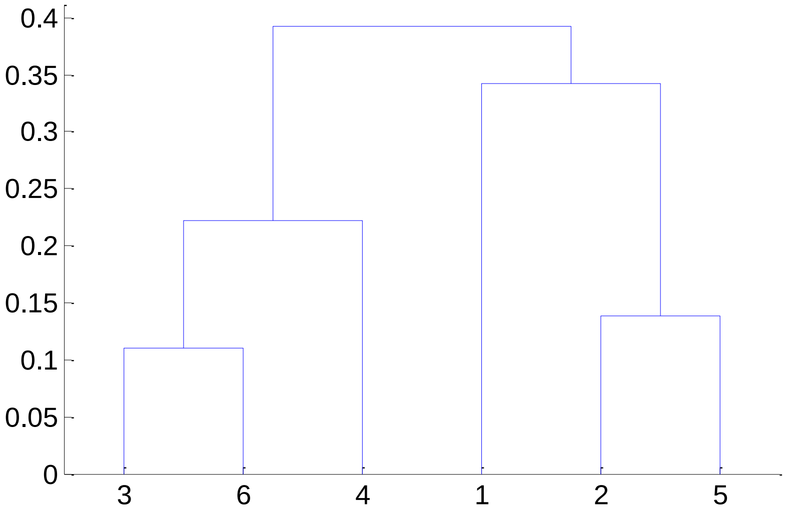
Cluster similarity: Max or complete linkage

- **Similarity of two clusters is based on the two least similar (most distant) points in the different clusters**
  - Determined by all pairs of points in the two clusters

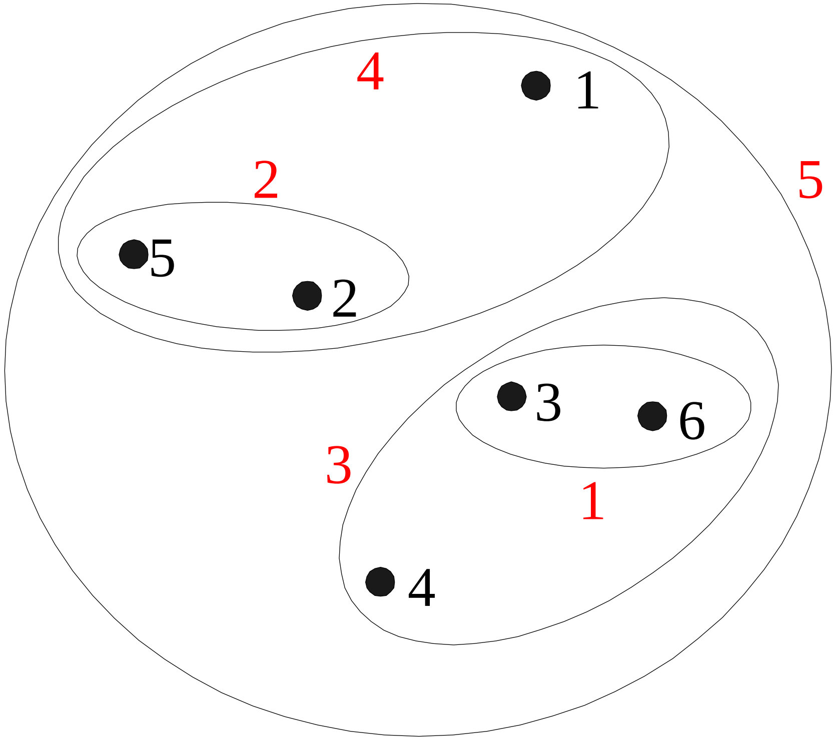


	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

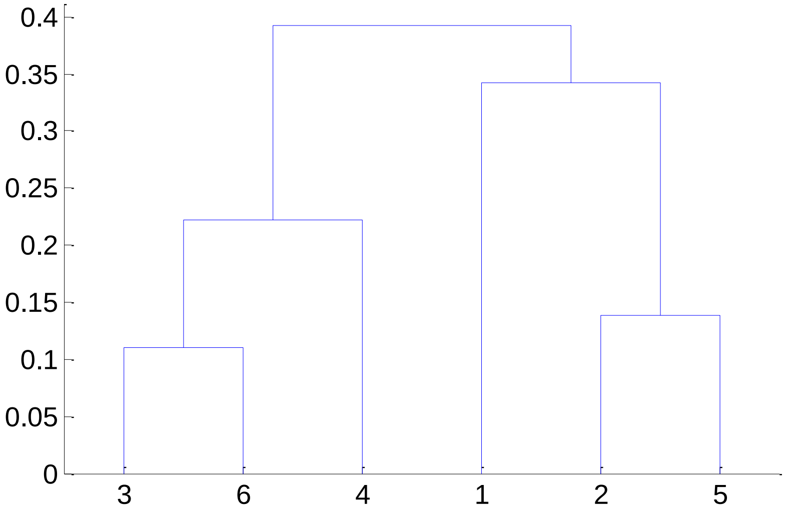
Table 8.4. Euclidean distance matrix for 6 points.



Hierarchical clustering: Max



Nested Clusters

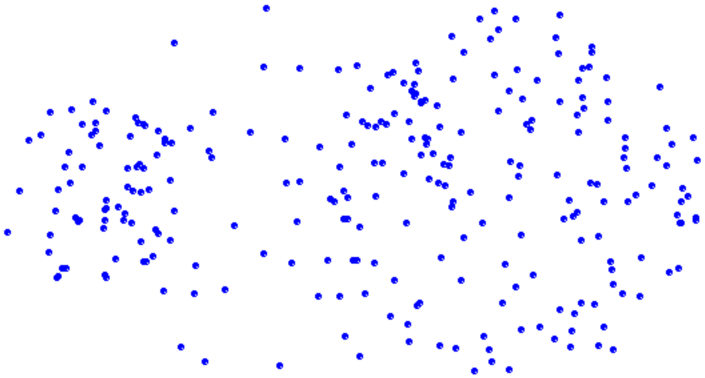


Dendrogram

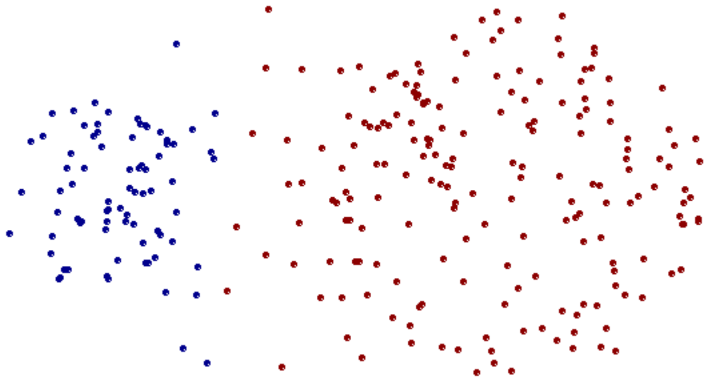
Note we still want to merge two most similar clusters each time. However, we define the distance between clusters based on MAX

Strength of Max

- Distance is based on **most distant points** in the different clusters



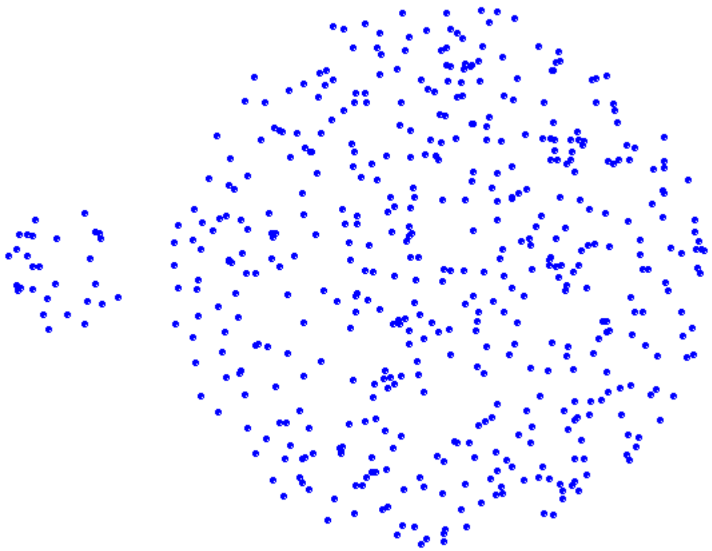
Original Points



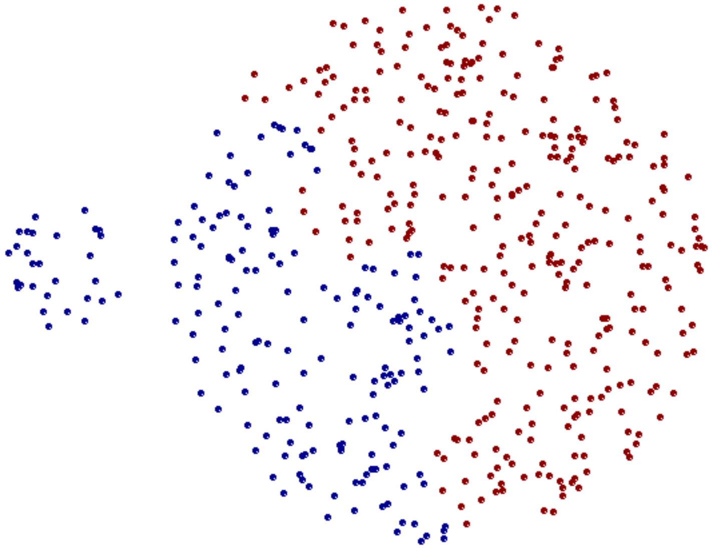
Two Clusters

- Less susceptible to noise and outliers

Limitations of Max



Original Points



Two Clusters

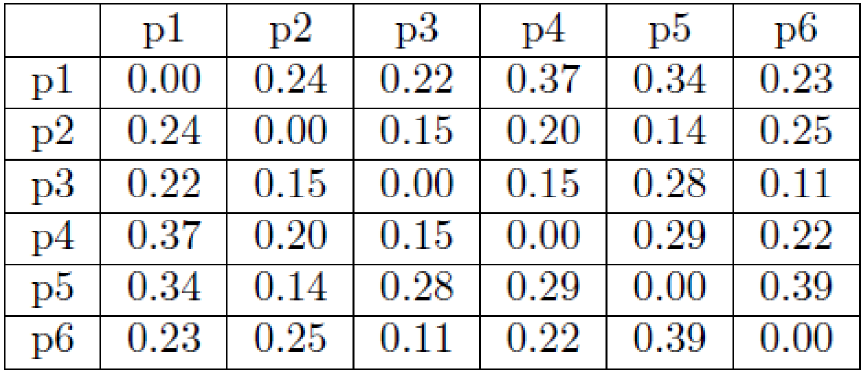
- **Tends to break large clusters**
  – Too big, so they are far away
- **Biased towards globular clusters**

Cluster similarity: Group average

- **Proximity of two clusters is the average of pairwise proximity between points in the two clusters**

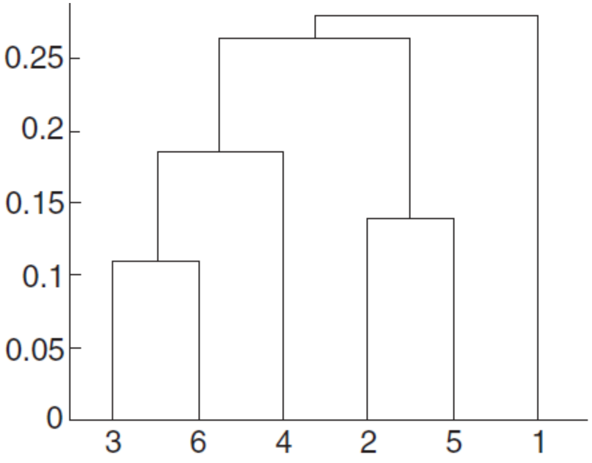
$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\displaystyle\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

- **Need to use average connectivity for scalability since total proximity favors large clusters**

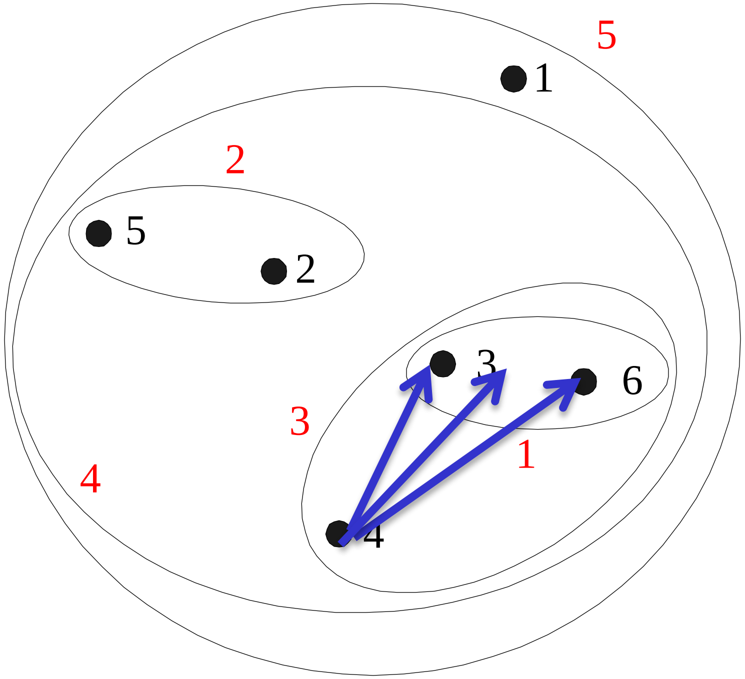


	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

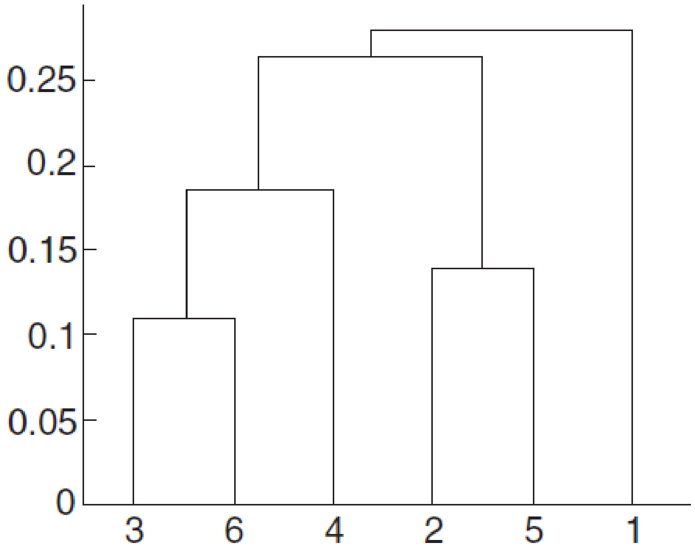
Table 8.4. Euclidean distance matrix for 6 points.



Hierarchical clustering: Group average



Group Average Clustering

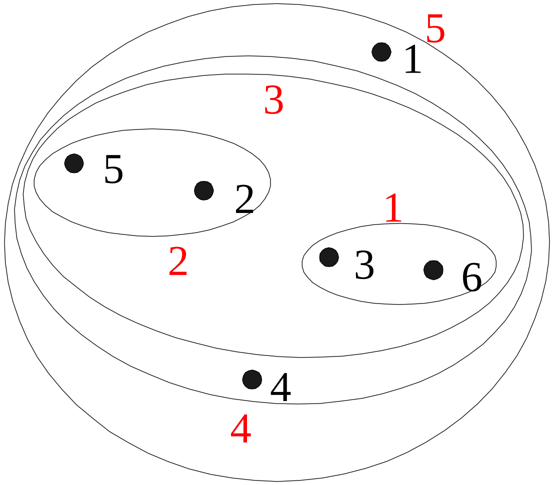


Group Average Dendrogram

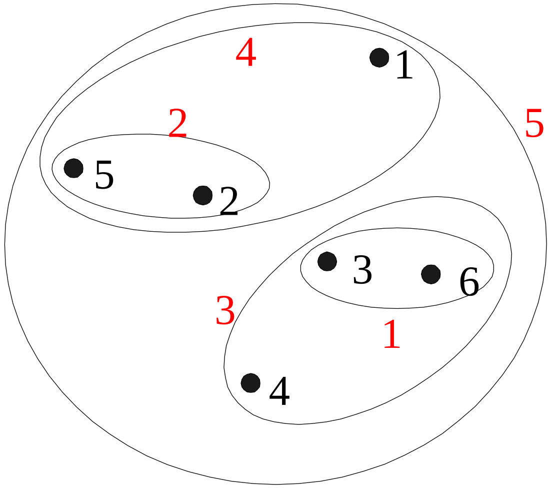
Hierarchical clustering: Group average

- **Compromise between Single and Complete Link**
- **Strengths**
  - Less susceptible to noise and outliers
- **Limitations**
  - Biased towards globular clusters

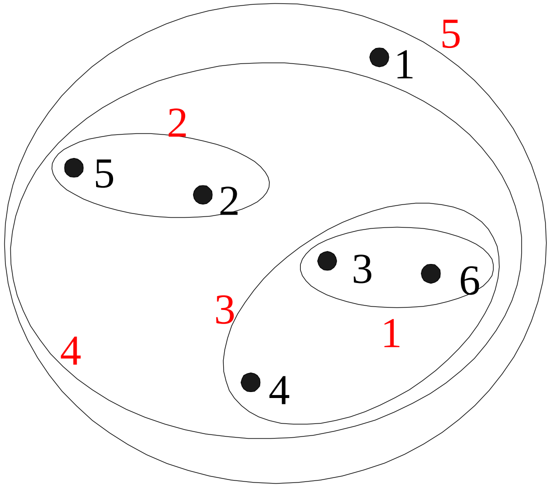
Hierarchical clustering: Comparison



Min



Max



Group average

Hierarchical clustering: Time & space requirements

- **O(N^2) space since it uses the proximity matrix**
  - – N is the number of points
- **O(N^3) time in many cases**
  - – There are N steps and at each step the size, N^2, proximity matrix must be updated and searched
  - – Complexity can be reduced to O(N^2 log(N)) time for some approaches

Bi-clustering in gene expression datasets

- What happens if the similarity does not exist for all the attributes?
- More advanced clustering techniques: Bi-clustering, i.e. cluster both rows and columns simultaneously
- http://www.powershow.com/view/11b05a-ZTg4N/Biclustering_in_Gene_Expression_Datasets_powerpoint_ppt_presentation
- Slide 1 - 7

Thank You

Contact: xlli@i2r.a-star.edu.sg if you have questions

For written notes on this lecture, please read chapter 14 of *The Practical Bioinformatician*.

CS2220: Introduction to Computational Biology
Unit 2: Gene Expression Analysis

Li Xiaoli

1 September 2016

NUS

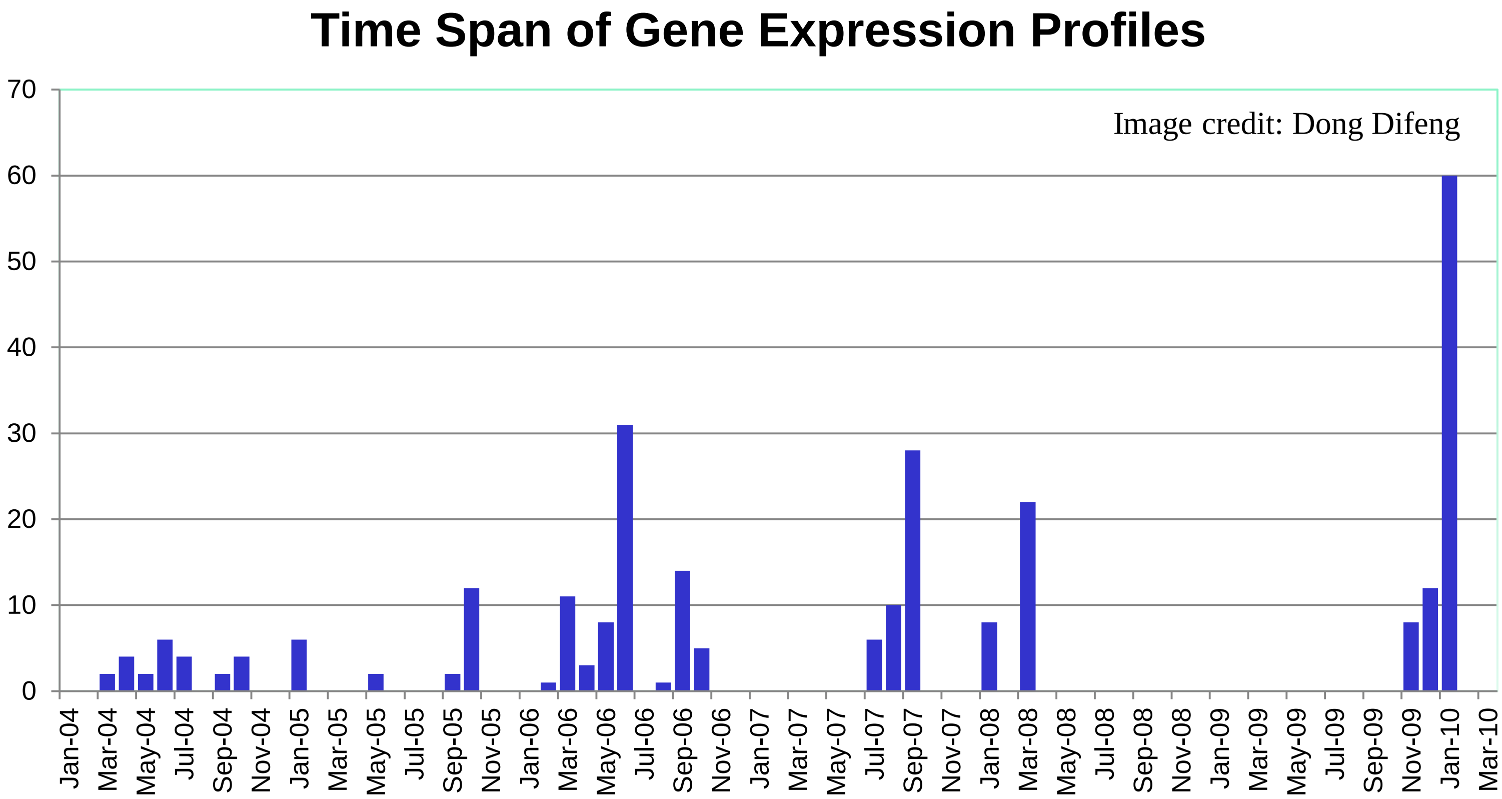
National University of Singapore

Plan

- **Normalization**
- **Computing similarity/distance between two gene expression profiles**
- **Gene expression profile classification**
- **Gene interaction prediction**
- **Simple introduction of Gene Ontology**

Normalization

Sometimes, a gene expression study may involve batches of data collected over a long period of time…



In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

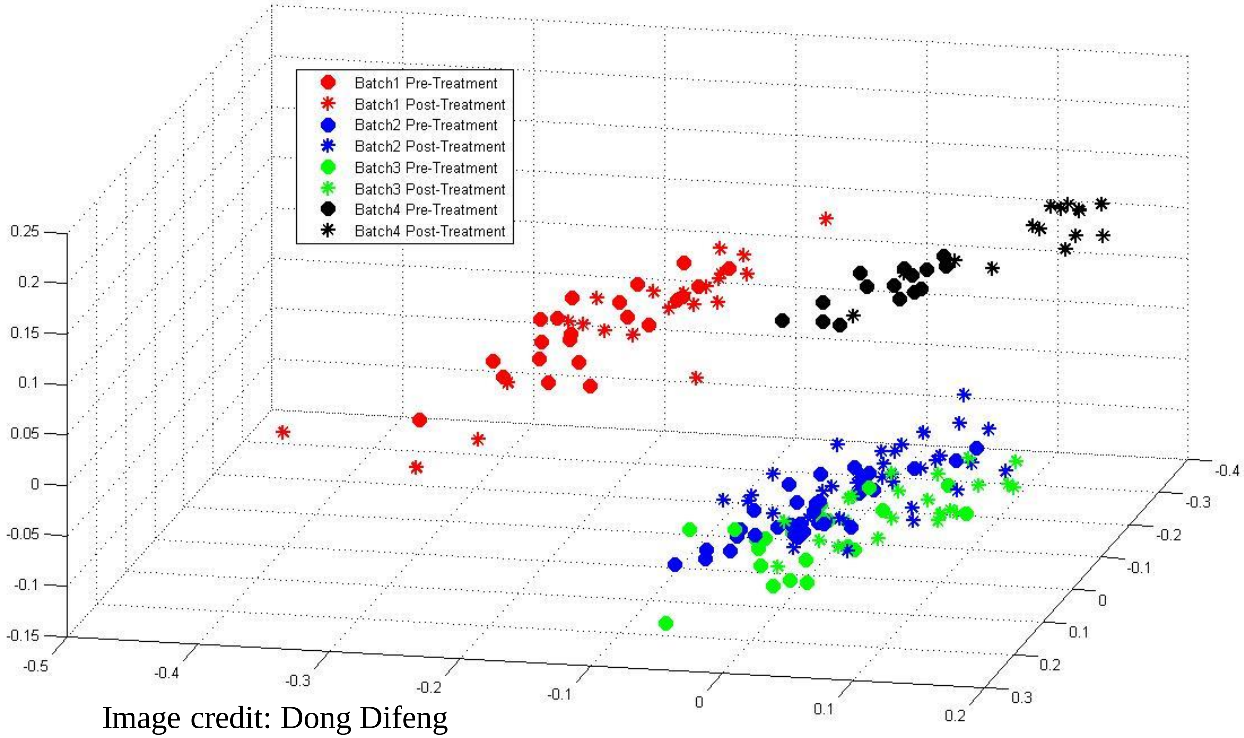


Image credit: Dong Difeng

⇒ **Need normalization to correct for batch effect**

Approaches to Normalization

- **Aim of normalization:**

  Reduce variance w/o increasing bias

- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

- **Xform data so that distribution of probe intensities is same on all arrays**
  - E.g., $(x - \mu) / \sigma$

- **Quantile normalization**

Quantile normalization

- **Given *n arrays of length p, form X of size p × n where each array is a column***
- **Sort each column of *X to give* X_{sort}**
- **Take means across rows of X_{sort} *and assign this* mean to each element in the row to get X'_{sort}**
- **Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as X**

- **Implemented in some microarray s/w, e.g., EXPANDER**

Can you perform quantite normalization?

Array 1, 2, …, *n*

Gene 1, 2, …, *p*

	1	2	…	*n*
1	0.8	0.7		
2				
3				
…..				
P				

Sort each column to give X_{sort}
Take means across rows of X_{sort} *and assign this* mean to each element in the row to get X'_{sort}
Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as *X*

Exercise

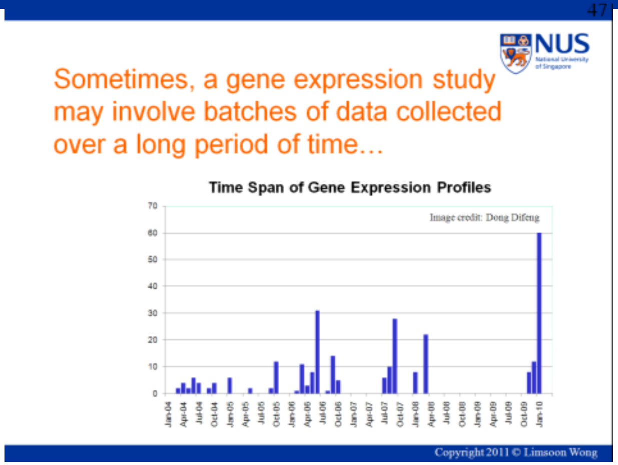
- http://en.wikipedia.org/wiki/Quantile_normalization

- Arrays 1 to 3, genes A to D

	Array 1	Array 2	Array 3
A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

How to perform quantile normalization?
Rank->Average-> Replace (same order)

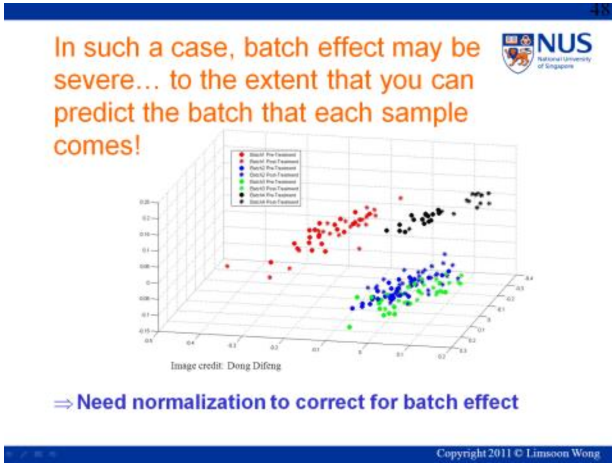
After quantile normalization



Sometimes, a gene expression study may involve batches of data collected over a long period of time…

Time Span of Gene Expression Profiles

Image credit: Dong Difeng



In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

Image credit: Dong Difeng

⇒ Need normalization to correct for batch effect

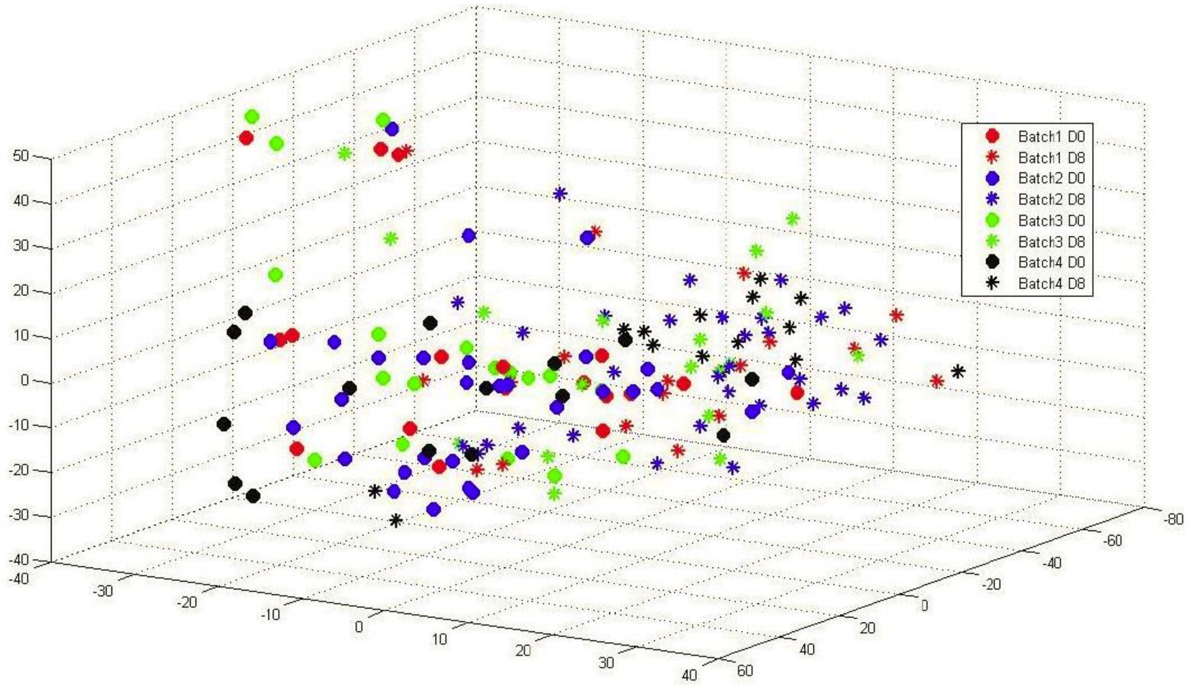


Figure 3.6: GEPs after the batch effects removing.

References

- E.-J. Yeoh et al., “Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling”, *Cancer Cell*, 1:133--143, 2002
- H. Liu, J. Li, L. Wong. Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data. *Bioinformatics*, 21(16):3377--3384, 2005.
- L.D. Miller et al., “Optimal gene expression analysis by microarrays”, *Cancer Cell* 2:353--361, 2002
- J. Li, L. Wong, “Techniques for Analysis of Gene Expression”, *The Practical Bioinformatician*, Chapter 14, pages 319—346, WSPC, 2004
- B. Bolstad et al. “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”. *Bioinformatics*, 19:185–193. 2003

Quantile normalization in statistics

- QN is a technique for making two distributions identical in statistical properties
- To quantile normalize two or more distributions to each other, we sort, then set to the average of the distributions
- The highest value in all cases becomes the mean of the highest values; the second highest value becomes the mean of the second highest values, and so on
- Quantile normalization is frequently used in microarray data analysis

Quantile normalization (rank array)

- **Arrays 1 to 3, genes A to D**

	Array 1	Array 2	Array 3
A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

- **For each column determine a rank from lowest to highest and assign number i-iv**

A	iv	iii	i
B	i	i	ii
C	ii	iii	iii
D	iii	ii	iv

These rank values are set aside to use later. We will convert the ranks into actual values

Quantile normalization
(average genes' rank values across array)

Smallest Values

- Go back to the first set of data. Rearrange that first set of column values so each column is in order going lowest to highest value

A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

A	2	1	3
B	3	2	4
C	4	4	6
D	5	4	8

Largest Values

- Now find the mean for each row to determine the values for the ranks

A (2 1 3)/3 = 2.00 = rank i

B (3 2 4)/3 = 3.00 = rank ii

C (4 4 6)/3 = 4.67 = rank iii

D (5 4 8)/3 = 5.67 = rank iv

Quantile normalization (average genes' rank values across array)

Smallest Values

- **Go back to the first set of data. Rearrange that first set of column values so each column is in order going lowest to highest value. The result is:**

A	**5**	**4**	**3**
B	**2**	**1**	**4**
C	**3**	**4**	**6**
D	**4**	**2**	**8**

A	2	1	3
B	3	2	4
C	4	4	6
D	5	4	8

Largest Values

- **Now find the mean for each row to determine the ranks**

A (2 1 3)/3 = 2.00 = rank i

B (3 2 4)/3 = 3.00 = rank ii

C (4 4 6)/3 = 4.67 = rank iii

D (5 4 8)/3 = 5.67 = rank iv

Quantile Normalization (explanation)

- **Go back to the first set of data. Rearrange that first set of column values so each column is in order going lowest to highest value. The result is:**

A	**5**	**4**	**3**
B	**2**	**1**	**4**
C	**3**	**4**	**6**
D	**4**	**2**	**8**

→

A	2	1	3
B	3	2	4
C	4	4	6
D	5	4	8

- **Now find the mean for each row to determine the ranks**

A (2 1 3)/3 = 2.00 = rank i — Average of the smallest

B (3 2 4)/3 = 3.00 = rank ii — Average of the second smallest

C (4 4 6)/3 = 4.67 = rank iii — Average of the second largest

D (5 4 8)/3 = 5.67 = rank iv — Average of the largest

Quantile Normalization (Replace)

2.00 = rank i, 3.00 = rank ii , 4.67 = rank iii , 5.67 = rank iv

- **Now take the ranking order and substitute in new values**

A	**iv**	**iii**	**i**
B	**i**	**i**	**ii**
C	**ii**	**iii**	**iii**
D	**iii**	**ii**	**iv**

A	5.67	4.67	2.00
B	2.00	2.00	3.00
C	3.00	4.67	4.67
D	4.67	3.00	5.67

Original Data

A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

Compute similarity/distance between two gene expression profiles

Cosine similarity

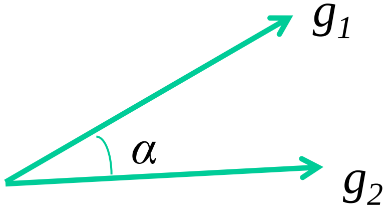
- If g_1 and g_2 are two gene profile vectors, then

$$\cos(g_1, g_2) = (g_1 \bullet g_2) / ||g_1|| \, ||g_2|| ,$$

where • indicates vector dot product and || *g*|| is the length of vector *g*.

- It is a measure of the cosine of the angle between the two vectors.

- Example:



g_1 = 3 2 0 5 0 0 0 2 0 0

g_2 = 1 0 0 0 0 0 0 1 0 2

$g_1 \bullet g_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||g_1||$ = $(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5}$ = $(42)^{0.5}$ = 6.4807

$||g_2||$ = $(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5}$ = $(6)^{0.5}$ = 2.4495

$$\cos(g_1, g_2) = 5/(6.4807*2.4495) = 0.3150$$

Pearson correlation coefficient

- **In statistics, the Pearson correlation coefficient (typically denoted by r) is a measure of the correlation (linear dependence) between two variables X and Y**
- **The values of r are between -1 and +1 inclusive**
- **It is widely used in the sciences as a measure of the strength of linear dependence between two variables**
- **In our case, variables are genes, we measure the correlation between their expression profiles**

Example

- X= (X1, X2, X3) = (0.03, 0.08, 1.83)
- Y= (Y1, Y2, Y3) = (0.01, 0.09, 2.12)
- Z= (Z1, Z2, Z3) = (2.51,0.10, 0.01)

- r(X,Y)=?
- r(X, Z)=?

X,Y, Z could be very high dimension vectors!!!

Formula - Pearson's correlation coefficient

- **Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:**

$$r = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

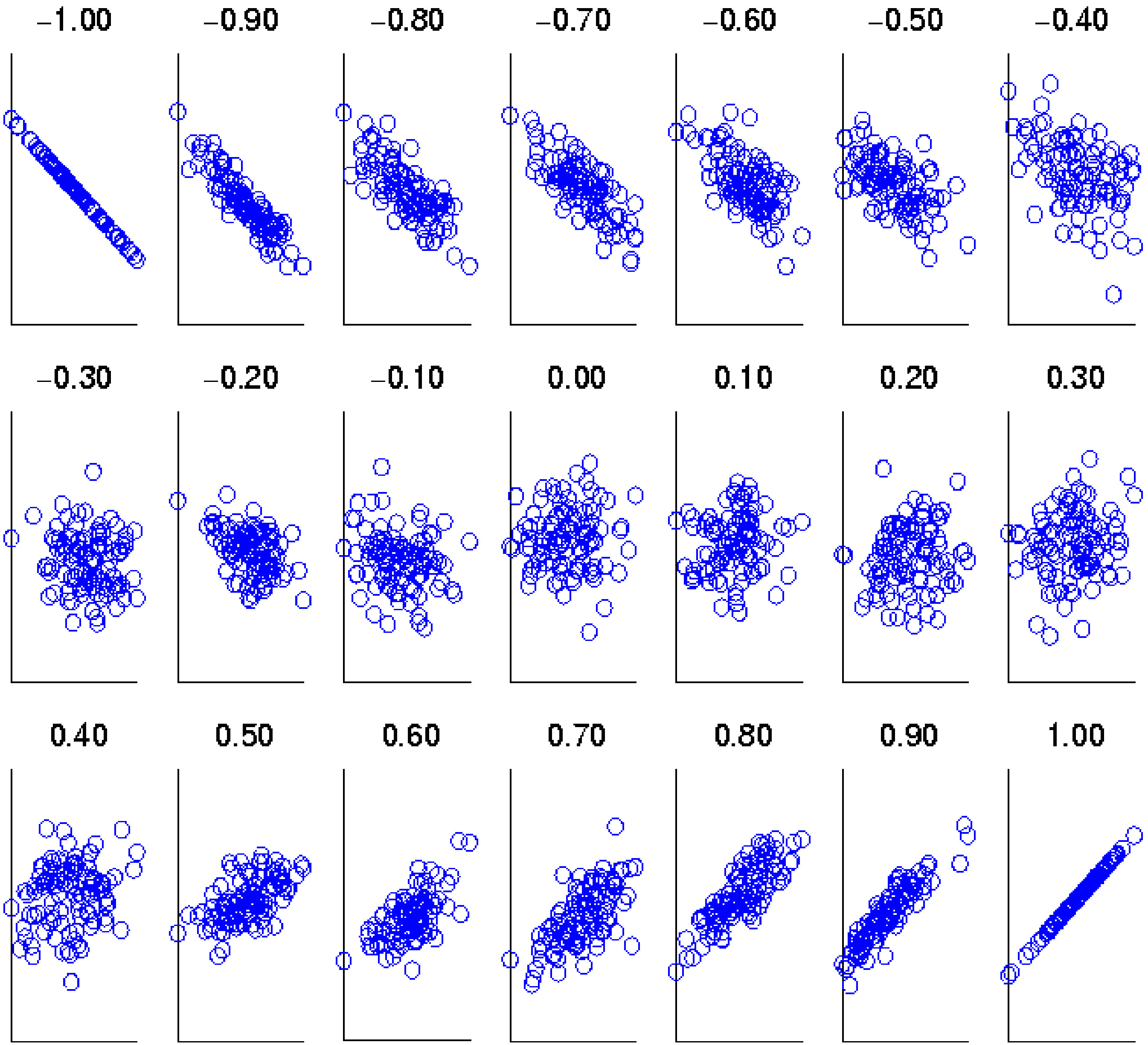
$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}.$$

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Easy to compute

Example: Visually Evaluating Correlation

Scatter plots showing the correlation from
–1 to 1.



1. Scatter plots illustrating correlations from -1 to 1.

An example to compute Pearson's correlation coefficient

- I will show an example to compute Pearson's correlation coefficient using Excel in Tutorial

- You can replace the numbers in the excel file to check how the values affect the PCC results

-

Euclidean distance

- **Euclidean Distance between two n-dimensional vectors (objects) p and q**

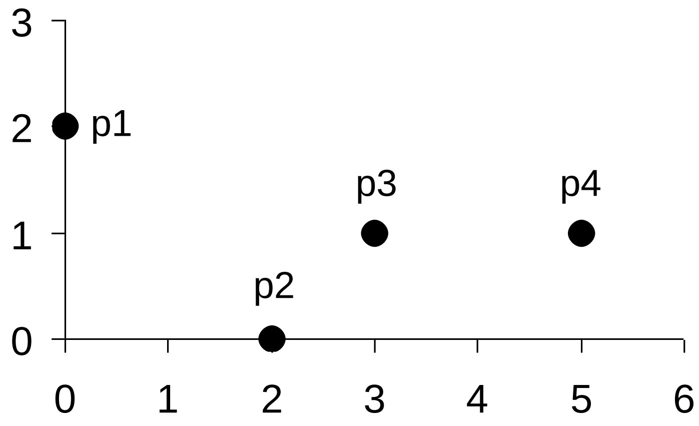
$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

where **p={**p_1 , p_2 , p_k , …, p_n }, **q={**q_1 , q_2 , q_k , …, q_n }. n is the number of dimensions (attributes) and p_k and q_k are the k^th^ attributes (components) of data objects p and q, respectively.

Euclidean distance in 2D

- **Example:**

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1



	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Euclidean Distance Matrix

Euclidean distance with feature importance

- **Given two vectors** $\mathbf{p}=\{p_1\ , p_2\ ,\ p_k\ ,\ \ldots,\ p_n\ \}$ $\mathbf{q}=\{q_1\ , q_2\ ,\ q_k\ ,\ \ldots,\ q_n\ \}$
- **May not want to treat all attributes the same**
- **We use weights wk to indicate the importance of each feature**
- **wk is between 0 and 1 and**

$$\sum_{k=1}^{n} w_k = 1$$

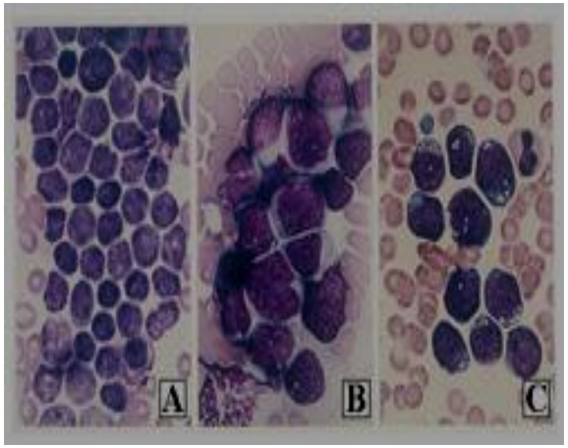
$$dist = \sqrt{\sum_{k=1}^{n} w_k (p_k - q_k)^2}$$

Gene expression profile classification

- **Diagnosis of childhood acute lymphoblastic leukemia and optimization of risk-benefit ratio of therapy**

Childhood ALL

- **6 Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse: suffer deterioration after a period of improvement.
- **The subtypes look similar**



- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- **Unavailable in most ASEAN countries**

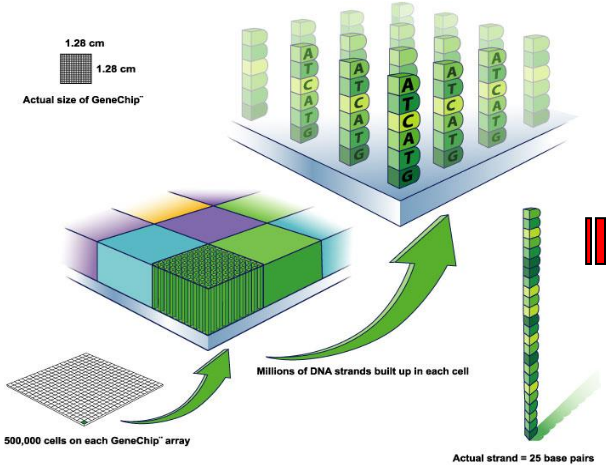
Mission

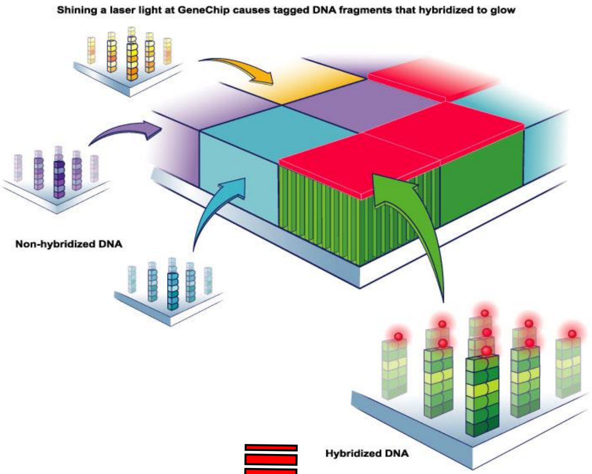
- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of *multiple specialists***

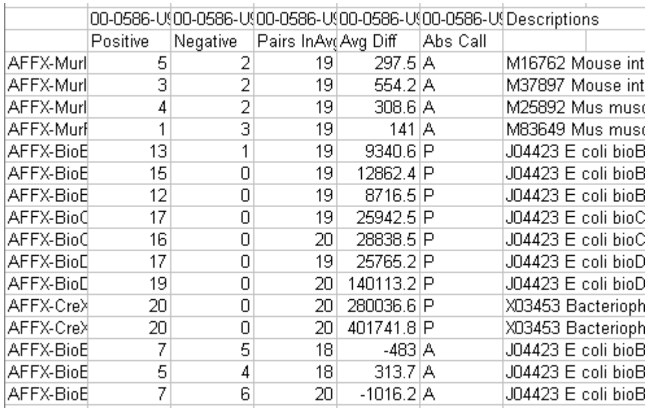
- **Generally available only in major advanced hospitals**

⇒ **Can we have a single-test easy-to-use platform instead?**

Single-test platform of microarray & machine learning

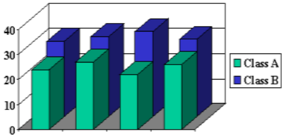

1.28 cm
1.28 cm
Actual size of GeneChip™
500,000 cells on each GeneChip™ array
Millions of DNA strands built up in each cell
Actual strand = 25 base pairs
A T C A T G
A T C A T G
A T C A T G
A T C A T G
A T C A T G


Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

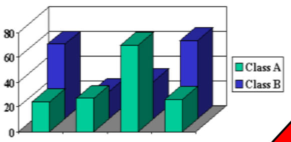


	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAvg	Avg Diff	Abs Call	
AFFX-Murl	5	2	19	297.5	A	M16762 Mouse int
AFFX-Murl	3	2	19	554.2	A	M37897 Mouse int
AFFX-Murl	4	2	19	308.6	A	M25892 Mus musc
AFFX-Murf	1	3	19	141	A	M83649 Mus musc
AFFX-BioE	13	1	19	9340.6	P	J04423 E coli bioB
AFFX-BioE	15	0	19	12862.4	P	J04423 E coli bioB
AFFX-BioE	12	0	19	8716.5	P	J04423 E coli bioB
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacterioph
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacterioph
AFFX-BioE	7	5	18	-483	A	J04423 E coli bioB
AFFX-BioE	5	4	18	313.7	A	J04423 E coli bioB
AFFX-BioE	7	6	20	-1016.2	A	J04423 E coli bioB

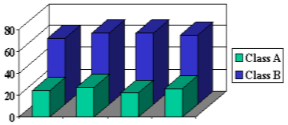
(I) Inter-class distance is too small



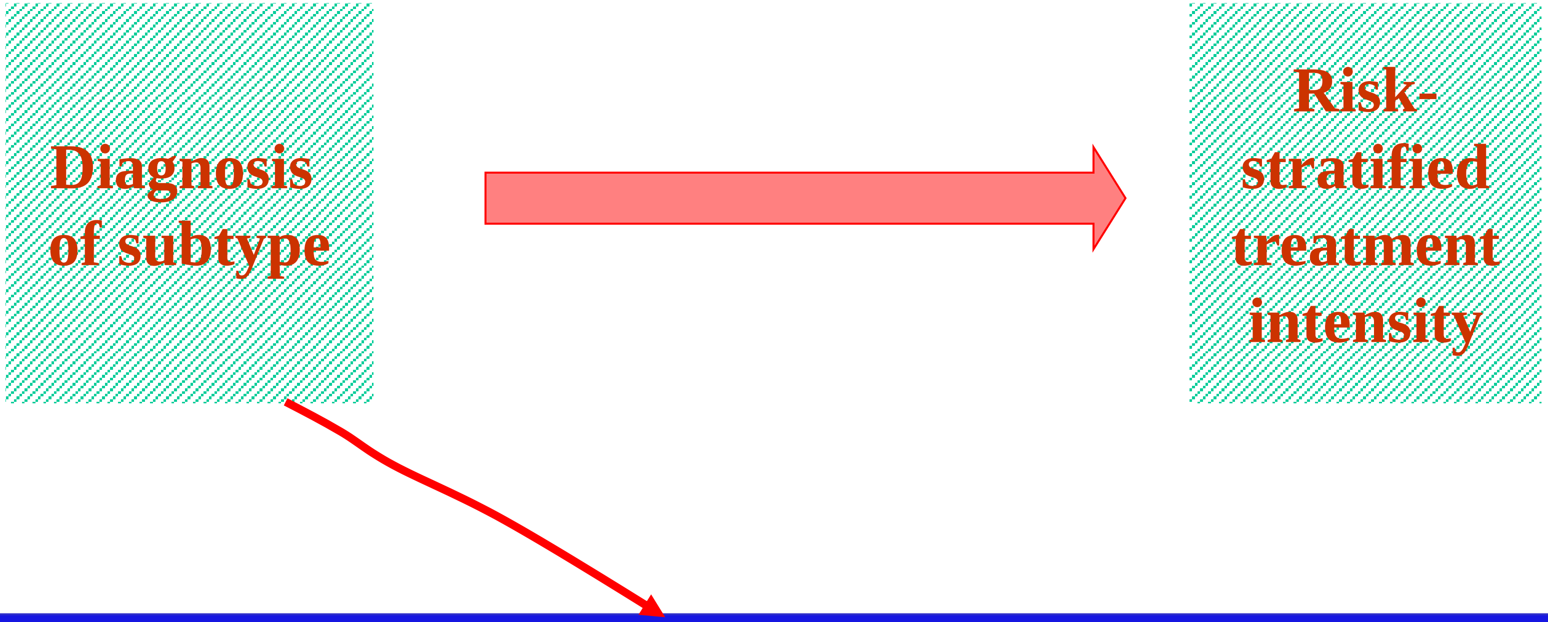
(II) Intra-class distance is too large



(III) Inter- and intra-class distances of a good signal



Overall strategy



Diagnosis of subtype

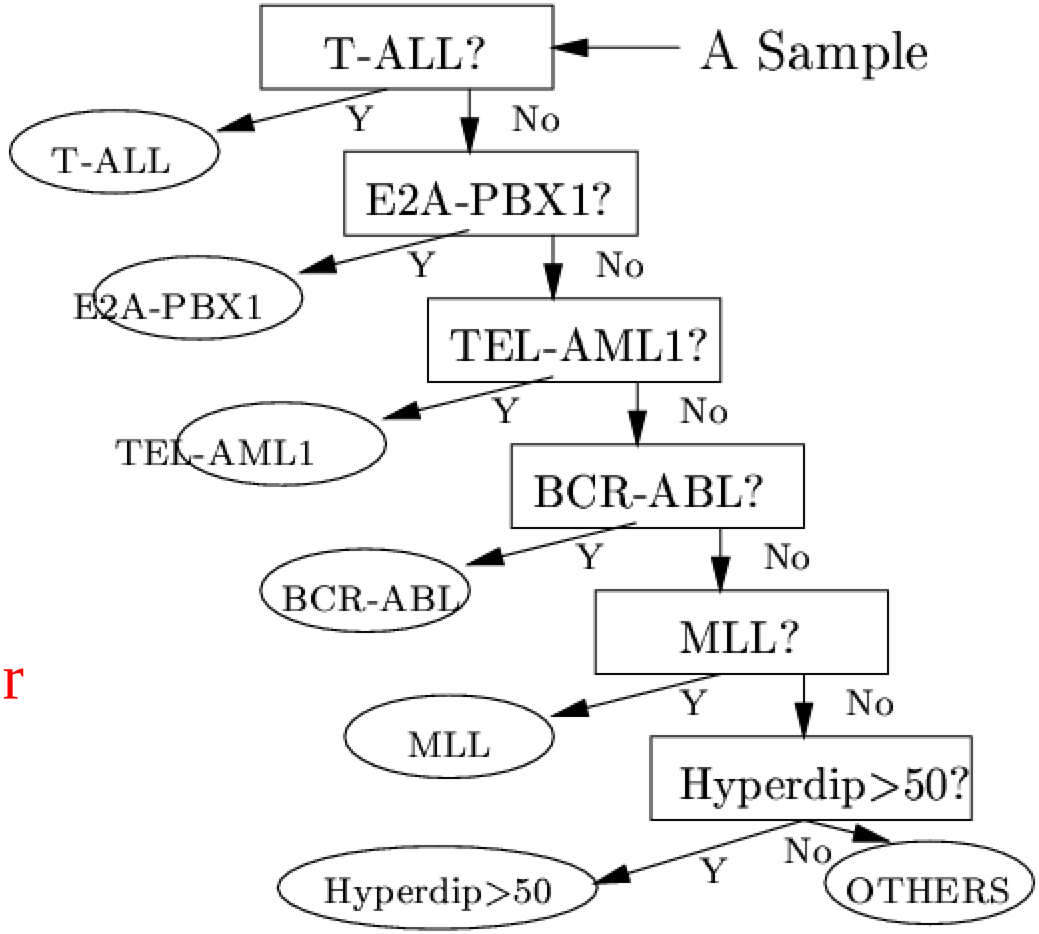
Risk-stratified treatment intensity

For each subtype, select genes to develop classification model for diagnosing that subtype

Subtype diagnosis by PCL

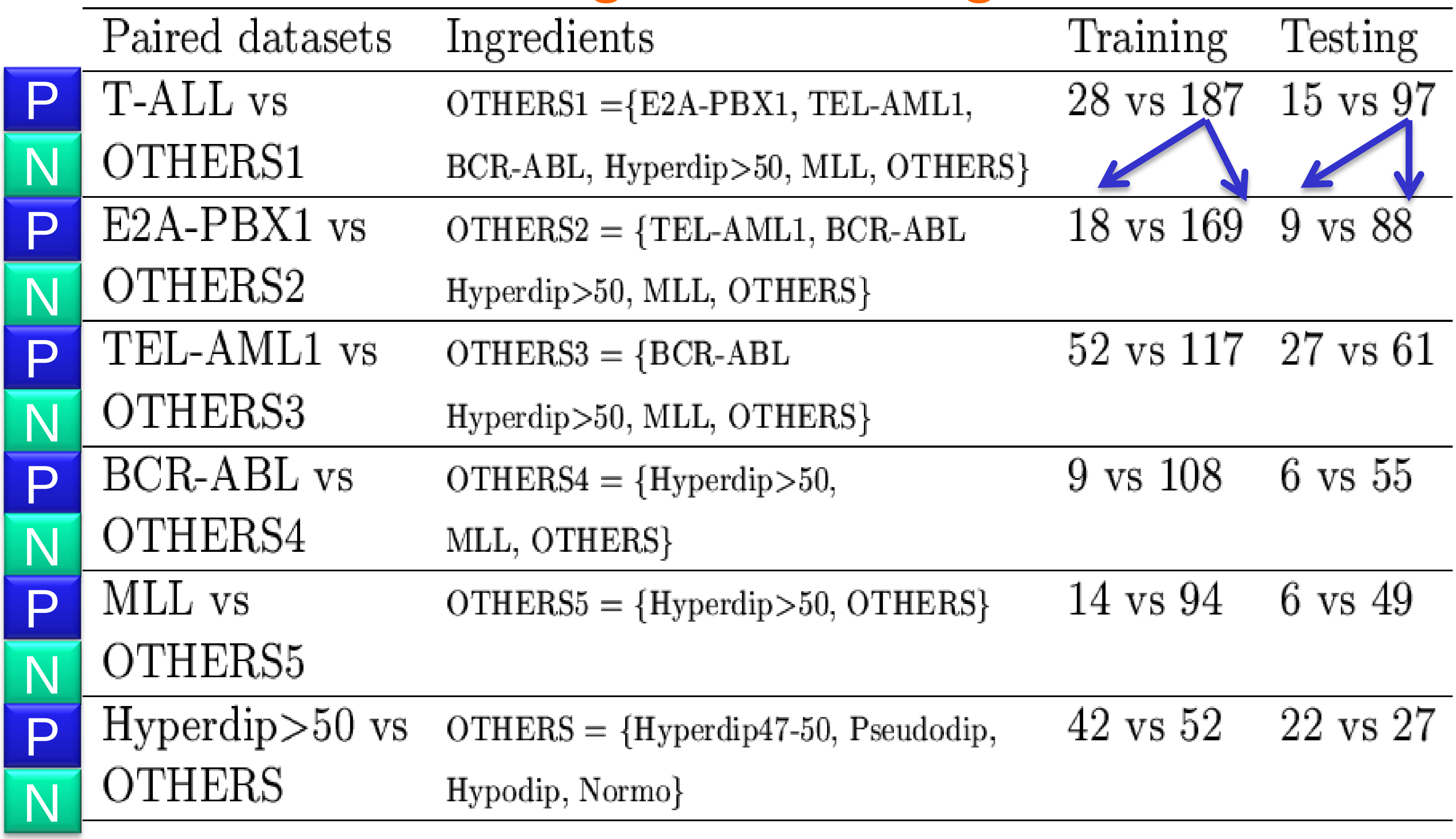
- **Gene expression data collection**
- **Classifier training by emerging pattern**
- **Apply classifier for diagnosis of future cases by PCL**

Childhood ALL subtype diagnosis workflow



A tree-structured diagnostic workflow was recommended by Prof Limsoon's doctor collaborator

Training and testing sets



	Paired datasets	Ingredients	Training	Testing
P N	T-ALL vs OTHERS1	OTHERS1 ={E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}	28 vs 187	15 vs 97
P N	E2A-PBX1 vs OTHERS2	OTHERS2 = {TEL-AML1, BCR-ABL Hyperdip>50, MLL, OTHERS}	18 vs 169	9 vs 88
P N	TEL-AML1 vs OTHERS3	OTHERS3 = {BCR-ABL Hyperdip>50, MLL, OTHERS}	52 vs 117	27 vs 61
P N	BCR-ABL vs OTHERS4	OTHERS4 = {Hyperdip>50, MLL, OTHERS}	9 vs 108	6 vs 55
P N	MLL vs OTHERS5	OTHERS5 = {Hyperdip>50, OTHERS}	14 vs 94	6 vs 49
P N	Hyperdip>50 vs OTHERS	OTHERS = {Hyperdip47-50, Pseudodip, Hypodip, Normo}	42 vs 52	22 vs 27

Training Data	Type1	Type2	Type3	Type4	Type5	Type6	Others
# Examples	28	18	52	9	14	42	52
Negatives	187	169	117	108	94	52	

Emerging patterns

- **An emerging pattern is a set of conditions**
  - usually involving several features
  - that most members of a class satisfy
  - but none or few of the other class satisfy

- **A jumping emerging pattern (JEP) is an emerging pattern that**
  - some members of a class satisfy
  - but no members of the other class satisfy

- **We only study jumping emerging patterns**

Examples of JEP

Patterns	Frequency (P)	Frequency(N)
{9, 36}	38 instances	0
{9, 23}	38	0
{4, 9}	38	0
{9, 14}	38	0
{6, 9}	38	0
{7, 21}	0	36
{7, 11}	0	35
{7, 43}	0	35
{7, 39}	0	34
{24, 29}	0	34

Easy interpretation

Reference number 9: the expression of gene 37720_at > 215
Reference number 36: the expression of gene 38028_at ≤ 12

PCL: Prediction by Collective Likelihood

- Let $EP_1^P, \ldots, EP_i^P$ be the most general EPs of D^P in descending order of support.

- Suppose the test sample T contains these most general EPs of D^P (in descending order of support):

$$EP_{i_1}^P, EP_{i_2}^P, \cdots, EP_{i_x}^P$$

T contains part of JEPs

- Use k top-ranked most general EPs of D^P and D^N. Define the score of T in the D^P class as

$$score(T, D^P) = \sum_{m=1}^{k} \frac{frequency(EP_{i_m}^P)}{frequency(EP_m^P)}$$

Pos support score: example

- Ditto for $score(T, D^N)$.

Neg support score

- If $score(T, D^P) > score(T, D^N)$, then T is class P. Otherwise it is class N.

PCL learning from training data

Top-Ranked EPs in Positive class

EP_1^P (90%)
EP_2^P (86%)
EP_3^P (85%)
EP_4^P (83%)
EP_5^P (80%)
EP_6^P (79%)
.
EP_n^P (68%)

Top-Ranked EPs in Negative class

EP_1^N (100%)
EP_2^N (95%)
EP_3^N (92%)
EP_4^N (89%)
EP_5^N (85%)
EP_6^N (80%)
.
EP_n^N (80%)

The idea of summarizing multiple top-ranked EPs is intended to avoid some rare tie cases

Test example T (k=3)

Top-Ranked EPs in Positive class

EP_1^P (90%) √
EP_2^P (86%)
EP_3^P (85%) √
EP_4^P (83%)
EP_5^P (80%) √
EP_6^P (79%)
.
EP_n^P (68%)

Top-Ranked EPs in Negative class

EP_1^N (100%) √
EP_2^N (95%)
EP_3^N (92%)
EP_4^N (89%) √
EP_5^N (85%) √
EP_6^N (80%)
.
EP_n^N (80%)

The idea of summarizing multiple top-ranked EPs is intended to avoid some rare tie cases

PCL testing (classify a test sample, k=3)

Most freq EP of pos class in the test sample

Top-k ranked EP of pos class in the test sample

$$\text{Score}^P = EP_1^{P'} / EP_1^{P} + \ldots + EP_k^{P'} / EP_k^{P} = 90/90+85/86+80/85$$

Most freq EP of pos class

Top-k ranked EP of pos class

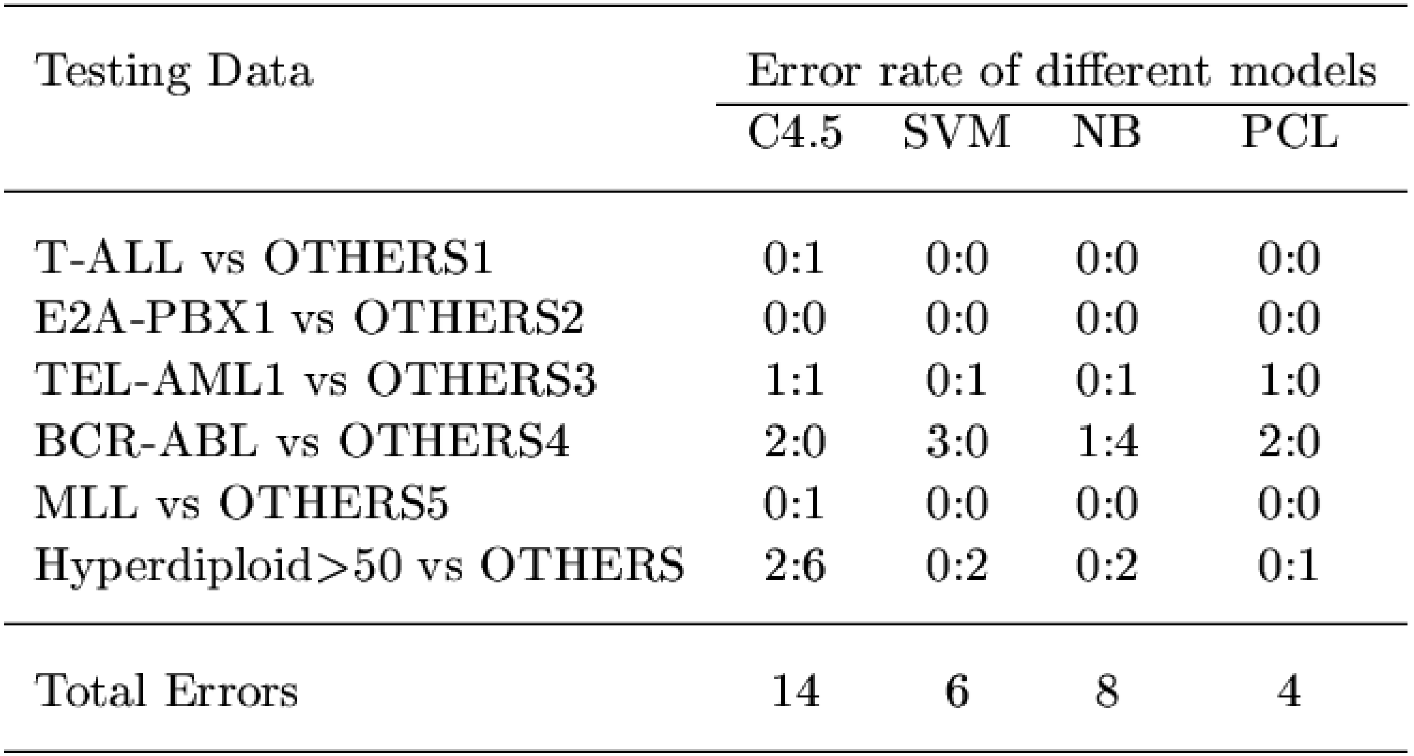
Similarly,

$$\text{Score}^N = EP_1^{N'} / EP_1^{N} + \ldots + EP_k^{N'} / EP_k^{N}$$

If Score^P > Score^N, then positive class, Otherwise negative class

If test sample contains more freq positive JEPs and less negative JEPs, then it is a positive sample; otherwise it is a negative sample.

Accuracy of PCL (vs. other classifiers)



Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS1	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS2	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS3	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS4	2:0	3:0	1:4	2:0
MLL vs OTHERS5	0:1	0:0	0:0	0:0
Hyperdiploid>50 vs OTHERS	2:6	0:2	0:2	0:1
Total Errors	14	6	8	4

The classifiers are all applied to the 20 genes selected by $\chi 2$ at each level of the tree.

x:y: # errors in positive class vs # errors in negative class

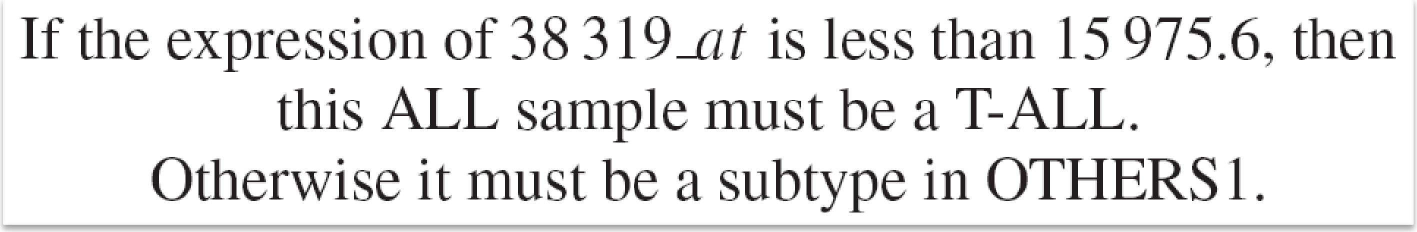
Understandability of PCL

- **E.g., for T-ALL vs. OTHERS1, one ideally discriminatory gene 38319_at was found, inducing these 2 EPs**

EP1 only occurs in P $\{gene_{\_(38\,319\_at)}@(-\infty, 15\,975.6)\}$ and

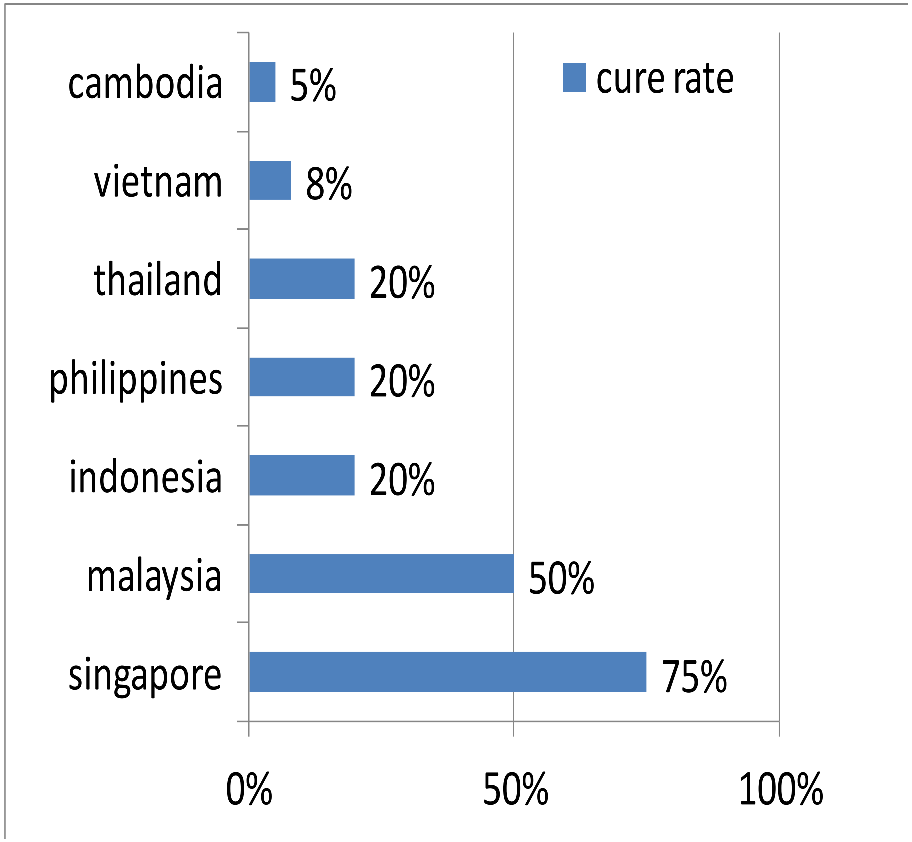
EP2 only occurs in N $\{gene_{\_(38\,319\_at)}@[15\,975.6, +\infty)\}$.

- **These give us the diagnostic rule for test example**



If the expression of 38 319_*at* is less than 15 975.6, then this ALL sample must be a T-ALL.
Otherwise it must be a subtype in OTHERS1.

Childhood ALL cure rates



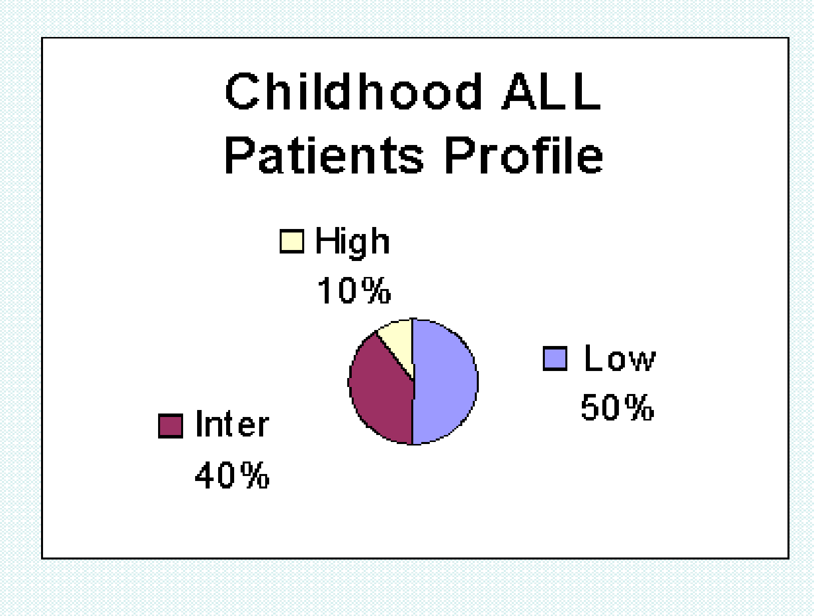
- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists**

- **Not available in less advanced ASEAN countries**

Childhood ALL treatment cost

- **Treatment for childhood ALL over 2 yrs**
  - Low intensity: US$36k
  - Intermediate intensity: US$60k
  - High intensity: US$72k
- **Treatment for relapse: US$150k**
- **Cost for side-effects: Unquantified**

Current situation (2000 new cases/yr in ASEAN)



- **Intermediate intensity conventionally applied in less advanced ASEAN countries**

Low: US$36k, Intermediate: US$**60k**, High: US$72k, relapse: US$**150k**

- **Over intensive for 50% of patients, thus more side effects (50% patients are supposed to use Low, but now we use intermediate intensity-> over)**
- **Under intensive for 10% of patients, thus more relapse**

  (should use high but use intermediate > under)

Current Cost for these 2000 cases

- **US$120m (US$60k * 2000) for intermediate intensity tx**
- **US$30m (US$150k * 2000 * 10%) for *relapse* tx (should use high)**
- **Total US$150m/yr plus un-quantified costs for dealing with side effects**

Using Prof Limsoon's platform

- Low intensity applied to 50% of patients
- Intermediate intensity to 40% of patients
- High intensity to 10% of patients

⇒ Reduced side effects

⇒ Reduced relapse

⇒ 75-80% cure rates

Total cost for new solution

- US$36m (US$36k * 2000 * 50%) for low intensity
- US$48m (US$60k * 2000 * 40%) for intermediate intensity
- US$14.4m (US$72k * 2000 * 10%) for high intensity

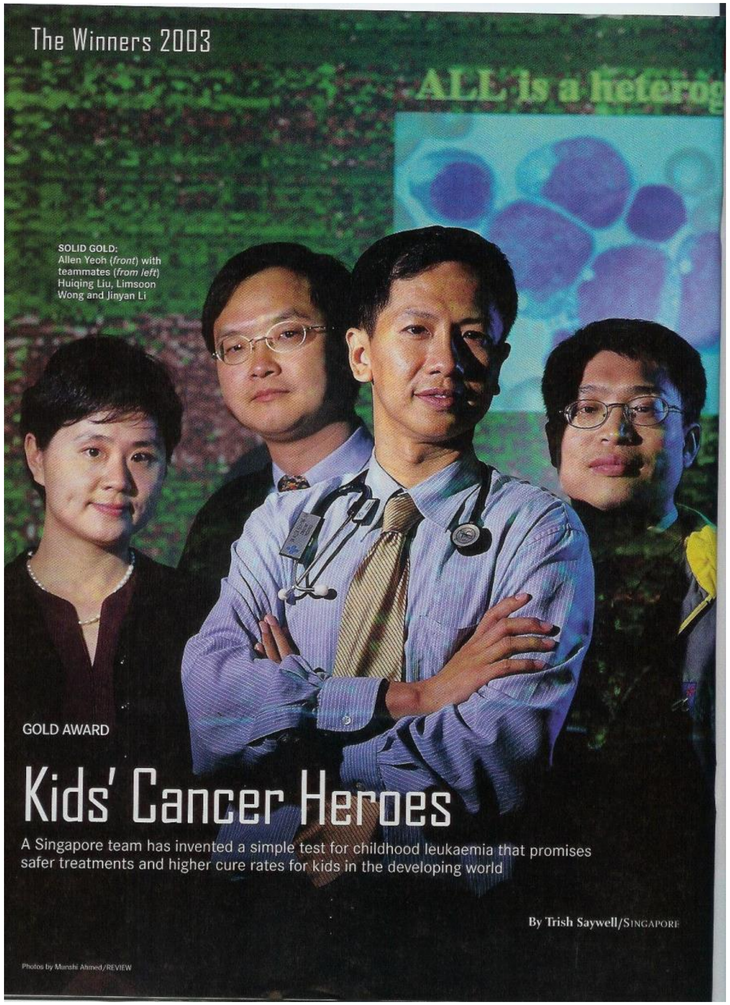
- Total US$98.4m/yr

⇒ Save US$51.6m/yr

Low: US$36k, Intermediate: US$**60k**,
High: US$72k, relapse: US$**150k**

A nice ending…

- **Asian Innovation Gold Award 2003**



The Winners 2003

ALL is a heterog

SOLID GOLD: Allen Yeoh (*front*) with teammates (*from left*) Huiqing Liu, Limsoon Wong and Jinyan Li

GOLD AWARD

Kids' Cancer Heroes

A Singapore team has invented a simple test for childhood leukaemia that promises safer treatments and higher cure rates for kids in the developing world

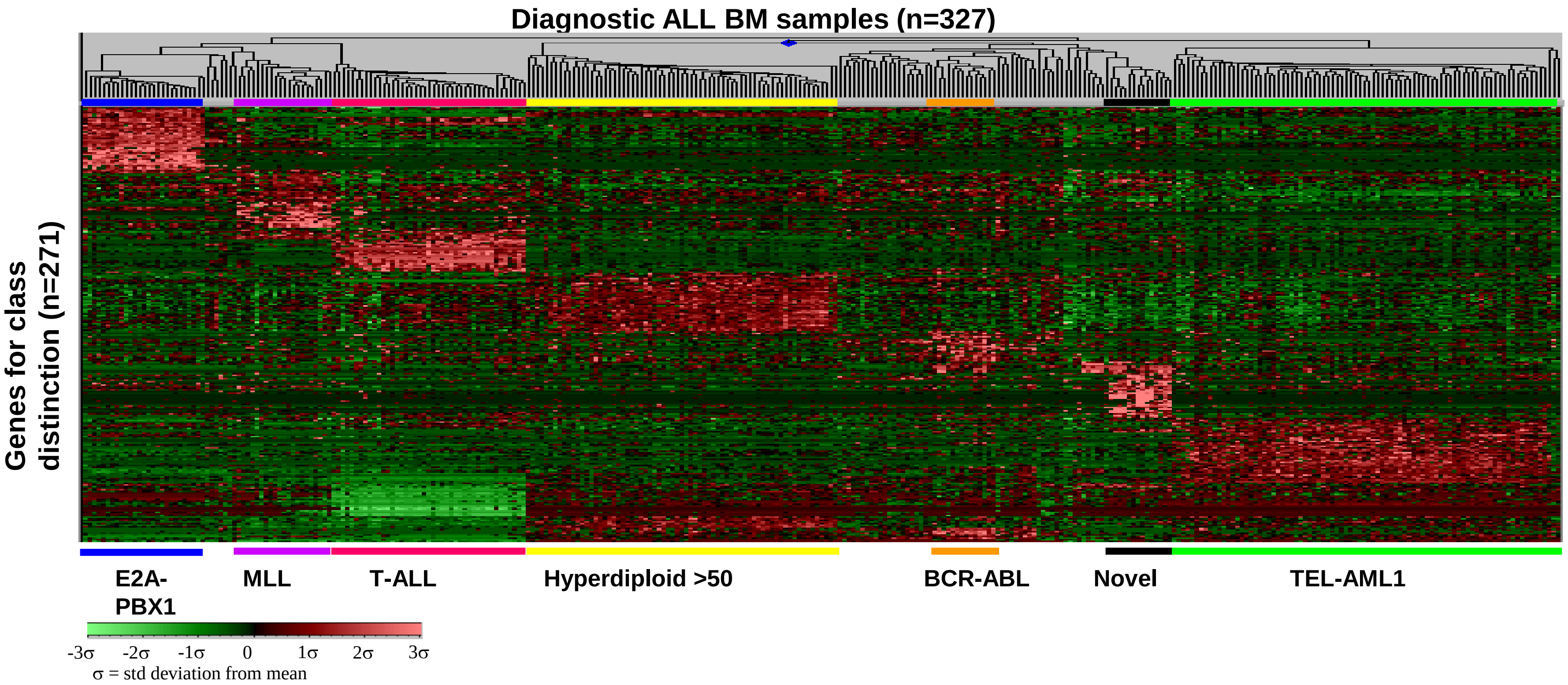
By Trish Saywell/Singapore

Photos by Munshi Ahmed/REVIEW

Gene Interaction Prediction

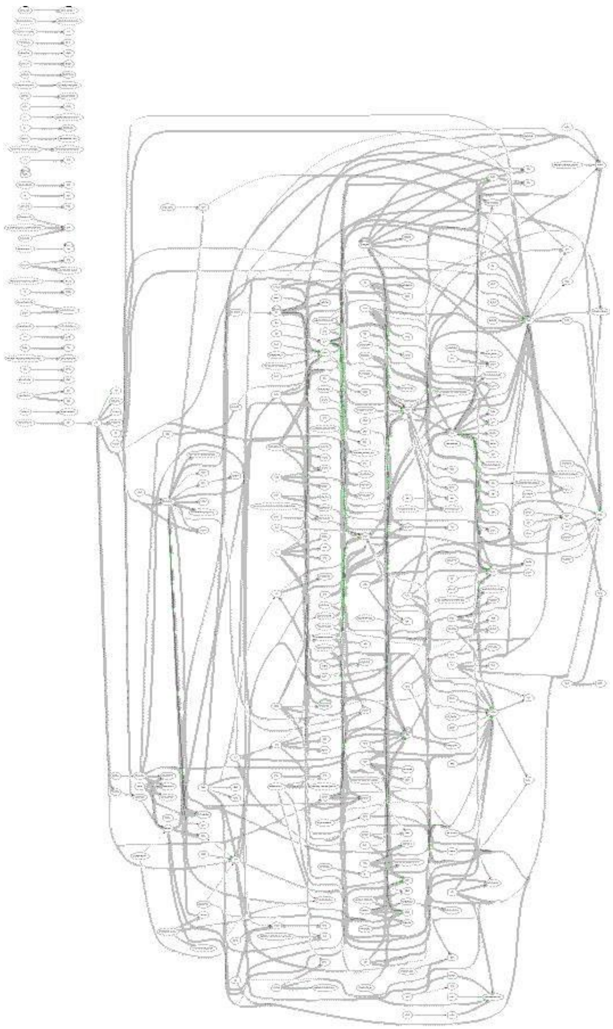
Beyond classification of gene expression profiles

- **After identifying the candidate genes by feature selection, do we know which ones are causal genes and which ones are surrogates?**



Gene regulatory circuits

- Genes are "connected" in "circuit" or network
- Expression of a gene in a network depends on expression of some other genes in the network
- Can we reconstruct the gene network from gene expression data?



Key questions

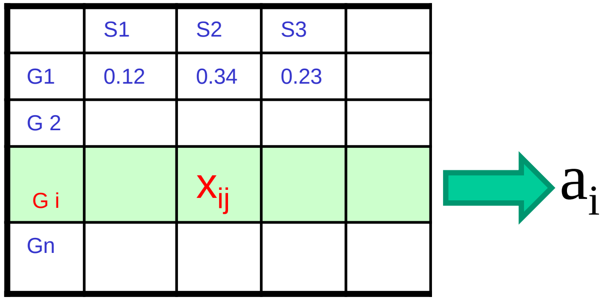
- **For each gene in the network:**
  - Which genes affect it?
  - How they affect it?

Some techniques

- **Bayesian Networks**
  - Friedman et al., *JCB* 7:601--620, 2000
- **Boolean Networks**
  - Akutsu et al., *PSB* 2000, pages 293--304
- **Differential equations**
  - Chen et al., *PSB* 1999, pages 29--40
- **Classification-based method**
  - Soinov et al., “Towards reconstruction of gene network from expression data by supervised learning”, *Genome Biology* 4:R6.1--9, 2003

A classification-based technique

Soinov et al., *Genome Biology* 4:R6.1-9, Jan 2003



	S1	S2	S3	
G1	0.12	0.34	0.23	
G 2				
G i		x_{ij}		
Gn				

a_i

- **Given a gene expression matrix X**
  - each row is a gene
  - each column is a sample
  - each element x_{ij} is expression of gene i in sample j

G i	↓	↑	↓	↓

- **Find the average value a_i of each gene i**

- **Denote s_{ij} as state of gene i in sample j,**
  - s_{ij} = up if $x_{ij} > a_i$
  - s_{ij} = down if $x_{ij} \leq a_i$

A classification-based technique

Soinov et al., *Genome Biology* 4:R6.1-9, Jan 2003

- **To see *whether* the state of gene *g* is determined by the state of *other genes i***
  - see whether ⟨s_{ij} | i ≠ g⟩ can predict s_{gj} (use other gene's same sample values to predict current gene's sample value)
  - if can predict with high accuracy, then "yes"
  - Any classifier can be used, such as C4.5, PCL, SVM, etc.

- **To see *how* the state of gene *g* is determined by the state of other genes**
  - apply C4.5 (or PCL or other "rule-based" classifiers) to predict s_{gj} from ⟨s_{ij} | i ≠ g⟩ (Rules are easy to understand)
  - and extract the decision tree or rules used

Simple Introduction of Gene Ontology

Gene Ontology (GO terms/concepts and relationships)

- **URL: http://www.geneontology.org/**

- **Download Ontology**
  - ftp://ftp.geneontology.org/pub/go/ontology-archive {Archive, including all the three parts of GO}
  - 10/31/2014 06:05PM 3,917,025 gene_ontology_edit.obo.2014-11-01.gz (consist of the following three parts; always updated one)
  - component.ontology (namespace: cellular_component)
  - function.ontology (namespace: molecular_function)
  - process.ontology (namespace: biological_process)

Associate genes with functions

- **How to get a gene/gene product's function info:**
  - 1. Download whole file (for large scale analysis)
    - http://geneontology.org/page/download-annotations

- **Saccharomyces cerevisiae**

•**Saccharomyces cerevisiae** •Stanford University	6381	94556 (48665 non-IEA)	11/1/2014	README	gene_association.sgd.gz (1 mb)

1: DB, database contributing the file (always "SGD" for this file). 2: DB_Object_ID, SGDID (SGD's unique identifier for genes and features). **3: DB_Object_Symbol**, see below 4: Qualifier (optional), one or more of 'NOT', 'contributes_to', 'colocalizes_with' as qualifier(s) for a GO annotation, when needed, multiples separated by pipe (|) **5: GO ID, unique numeric identifier for the GO term** **6: DB:Reference(|DB:Reference), the reference associated with the GO annotation** **7: Evidence, the evidence code for the GO annotation** 8: With (or) From (optional), any With or From qualifier for the GO annotation **9: Aspect, which ontology the GO term belongs (Function, Process or Component)** 10: DB_Object_Name(|Name) (optional), a name for the gene product in words, e.g. 'acid phosphatase' 11: DB_Object_Synonym(|Synonym) (optional), see below 12: DB_Object_Type, type of object annotated, e.g. gene, protein, etc. 13: taxon(|taxon), taxonomic identifier of species encoding gene product 14: Date, date GO annotation was defined in the format YYYYMMDD 15: Assigned_by, source of the annotation (always "SGD" for this file)

More detailed description of GO

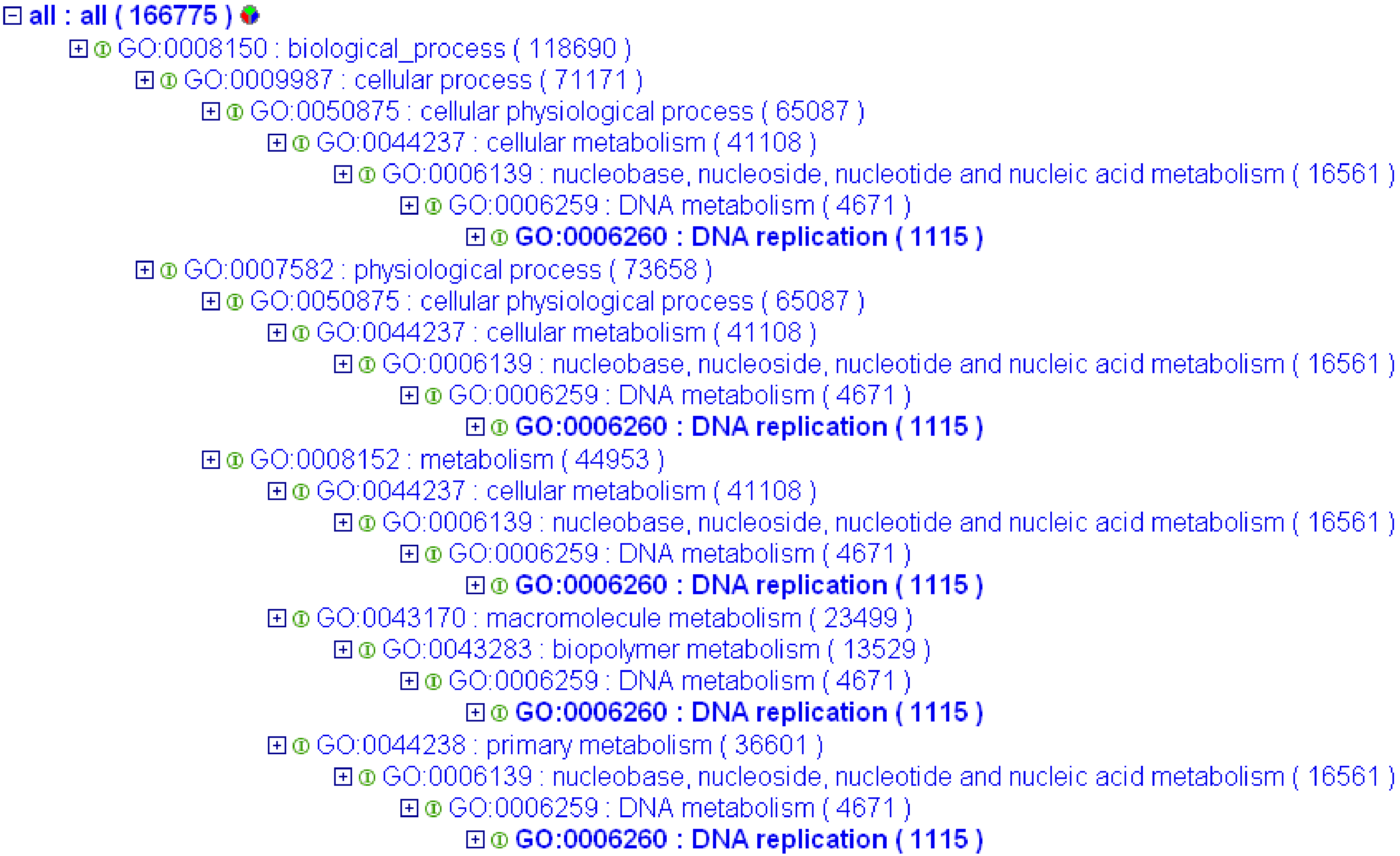
- **The Gene Ontology provides a way to capture and represent biological knowledge in a computable form**

GO slides from Jennifer Clark, Gene Ontology Consortium editorial office

How does the Gene Ontology work?

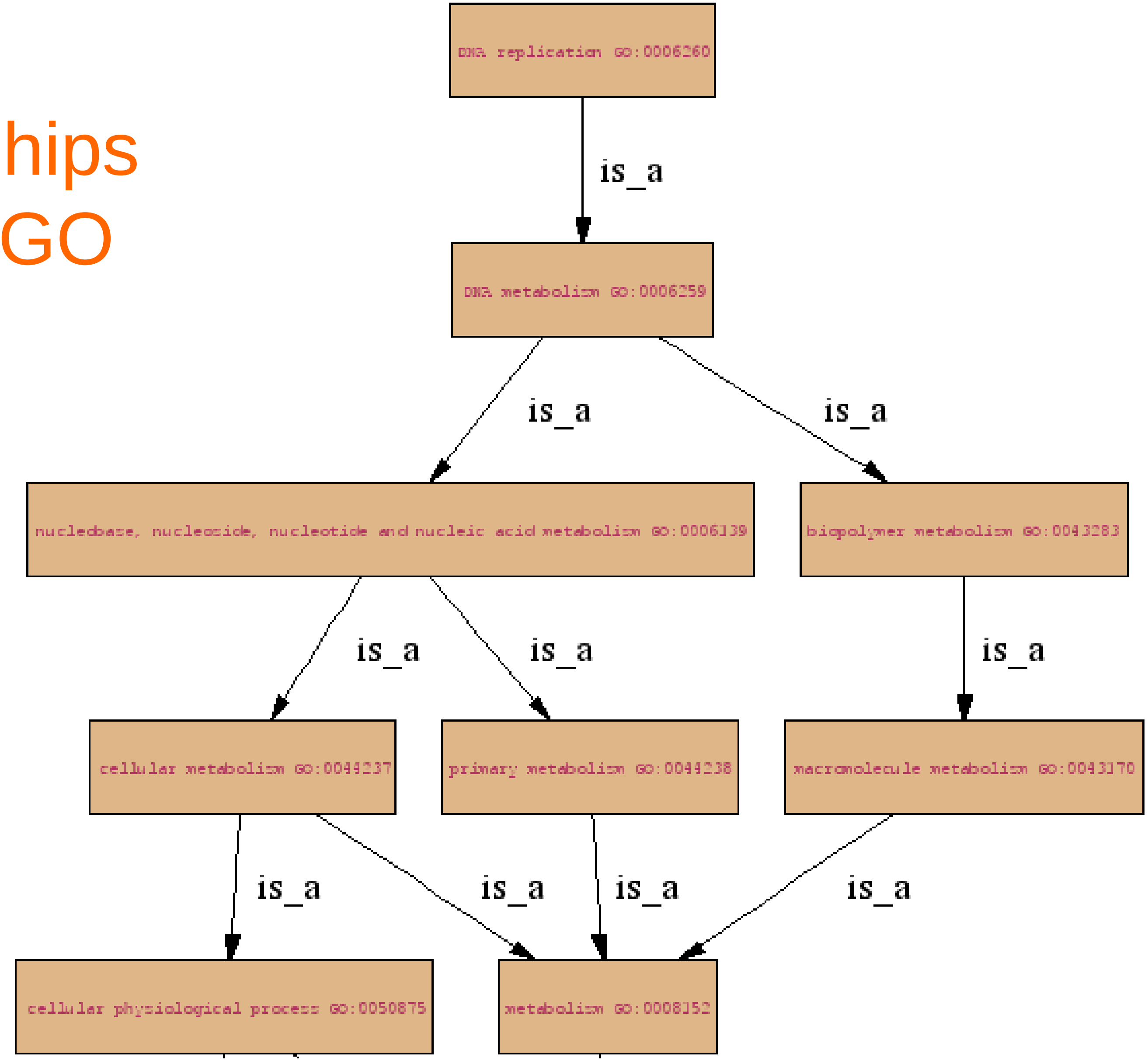
- **GO isn’t just a flat list of biological terms**
- **Terms are related within a hierarchy**

GO structure

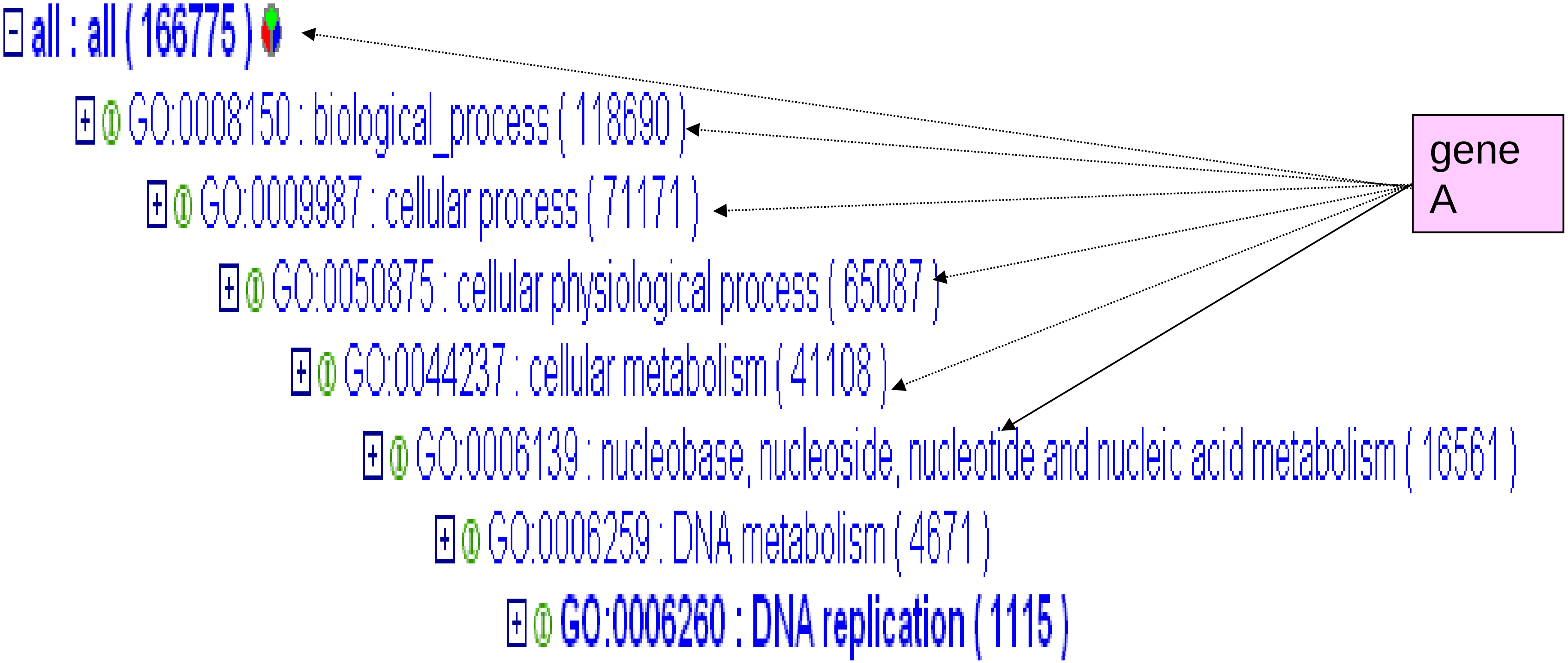


```
⊟ all : all ( 166775 )
    ⊞ GO:0008150 : biological_process ( 118690 )
        ⊞ GO:0009987 : cellular process ( 71171 )
            ⊞ GO:0050875 : cellular physiological process ( 65087 )
                ⊞ GO:0044237 : cellular metabolism ( 41108 )
                    ⊞ GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic acid metabolism ( 16561 )
                        ⊞ GO:0006259 : DNA metabolism ( 4671 )
                            ⊞ GO:0006260 : DNA replication ( 1115 )
        ⊞ GO:0007582 : physiological process ( 73658 )
            ⊞ GO:0050875 : cellular physiological process ( 65087 )
                ⊞ GO:0044237 : cellular metabolism ( 41108 )
                    ⊞ GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic acid metabolism ( 16561 )
                        ⊞ GO:0006259 : DNA metabolism ( 4671 )
                            ⊞ GO:0006260 : DNA replication ( 1115 )
            ⊞ GO:0008152 : metabolism ( 44953 )
                ⊞ GO:0044237 : cellular metabolism ( 41108 )
                    ⊞ GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic acid metabolism ( 16561 )
                        ⊞ GO:0006259 : DNA metabolism ( 4671 )
                            ⊞ GO:0006260 : DNA replication ( 1115 )
                ⊞ GO:0043170 : macromolecule metabolism ( 23499 )
                    ⊞ GO:0043283 : biopolymer metabolism ( 13529 )
                        ⊞ GO:0006259 : DNA metabolism ( 4671 )
                            ⊞ GO:0006260 : DNA replication ( 1115 )
                ⊞ GO:0044238 : primary metabolism ( 36601 )
                    ⊞ GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic acid metabolism ( 16561 )
                        ⊞ GO:0006259 : DNA metabolism ( 4671 )
                            ⊞ GO:0006260 : DNA replication ( 1115 )
```

Relationships between GO terms


DNA replication GO:0006260
is_a
DNA metabolism GO:0006259
is_a
is_a
nucleobase, nucleoside, nucleotide and nucleic acid metabolism GO:0006139
biopolymer metabolism GO:0043283
is_a
is_a
is_a
cellular metabolism GO:0044237
primary metabolism GO:0044238
macromolecule metabolism GO:0043170
is_a
is_a
is_a
is_a
cellular physiological process GO:0050875
metabolism GO:0008152

Gene function



⊟ all : all (166775)

⊞ GO:0008150 : biological_process (118690)

⊞ GO:0009987 : cellular process (71171)

⊞ GO:0050875 : cellular physiological process (65087)

⊞ GO:0044237 : cellular metabolism (41108)

⊞ GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic acid metabolism (16561)

⊞ GO:0006259 : DNA metabolism (4671)

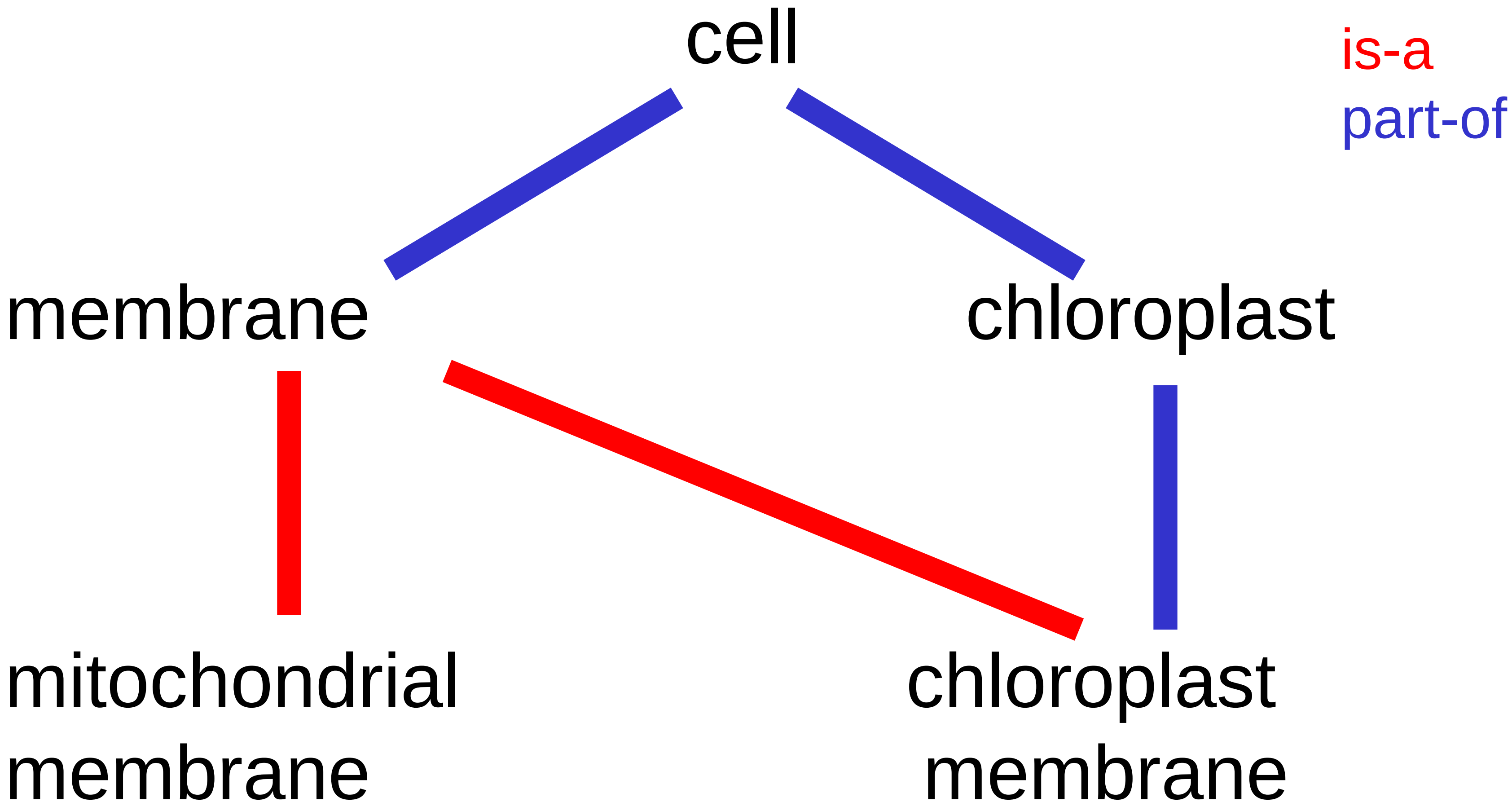
⊞ GO:0006260 : DNA replication (1115)

gene A

Ontology structure

- **Terms are linked by two relationships**
  - is-a ⓘ
  - part-of Ⓟ

Ontology structure



cell

is-a

part-of

membrane

chloroplast

mitochondrial membrane

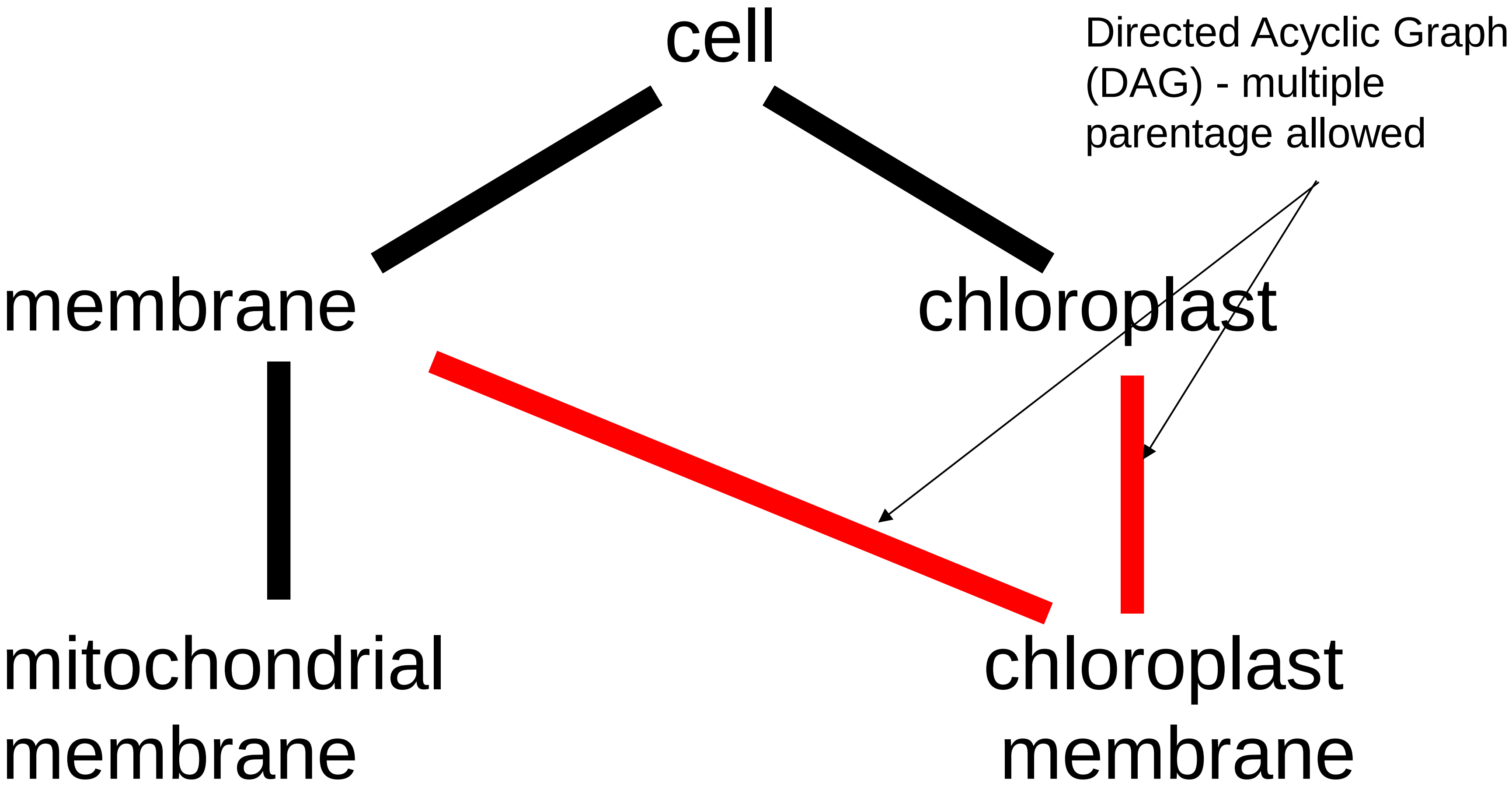
chloroplast membrane

Ontology structure

- **Ontologies are structured as a hierarchical directed acyclic graph (DAG) [NO LOOP]**

- **Terms can have more than one parent and zero, one or more children**

Ontology structure



cell

membrane

chloroplast

mitochondrial membrane

chloroplast membrane

Directed Acyclic Graph (DAG) - multiple parentage allowed

How does GO work?

What information might we want to capture about a gene product?

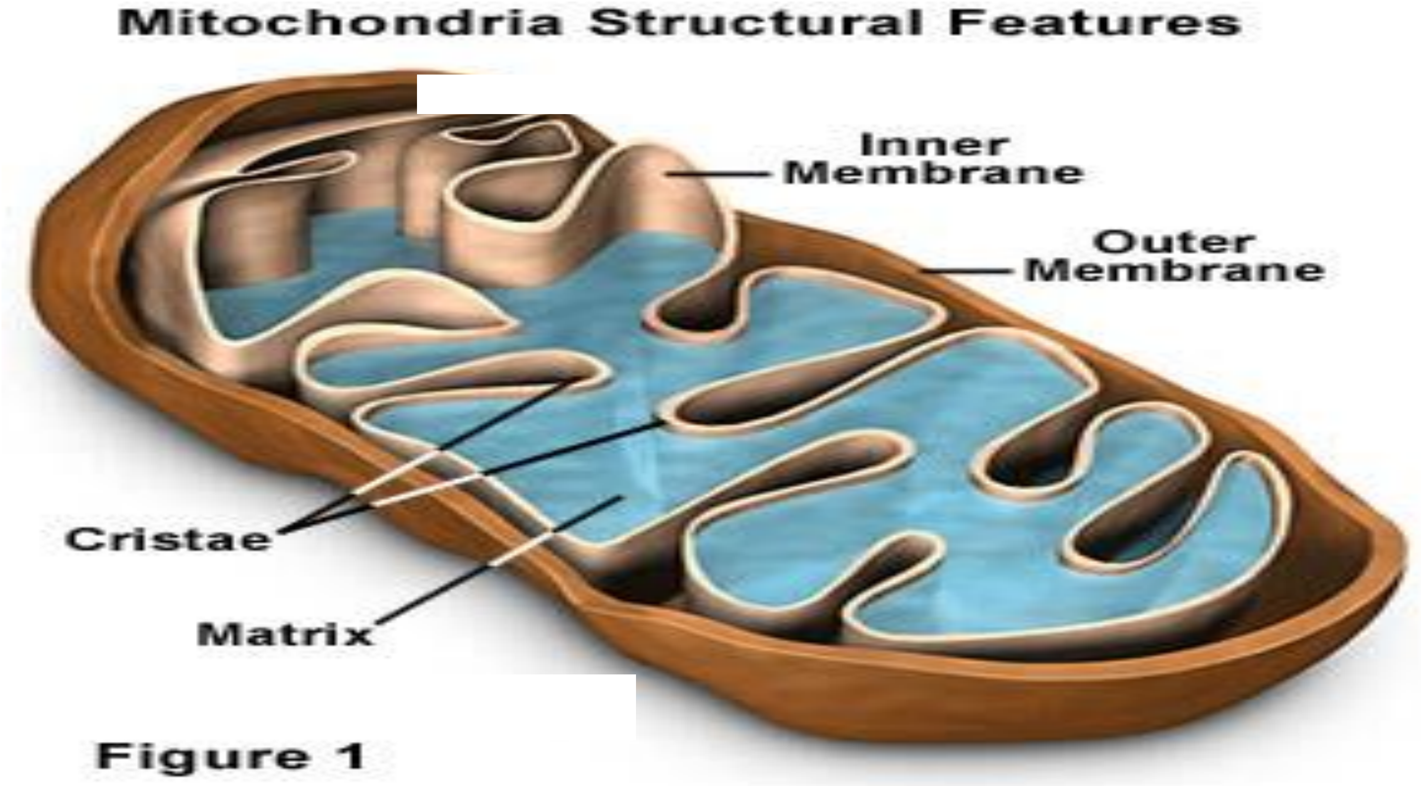
- **What does the gene product do?**
- **Where and when does it act?**
- **Why does it perform these activities?**

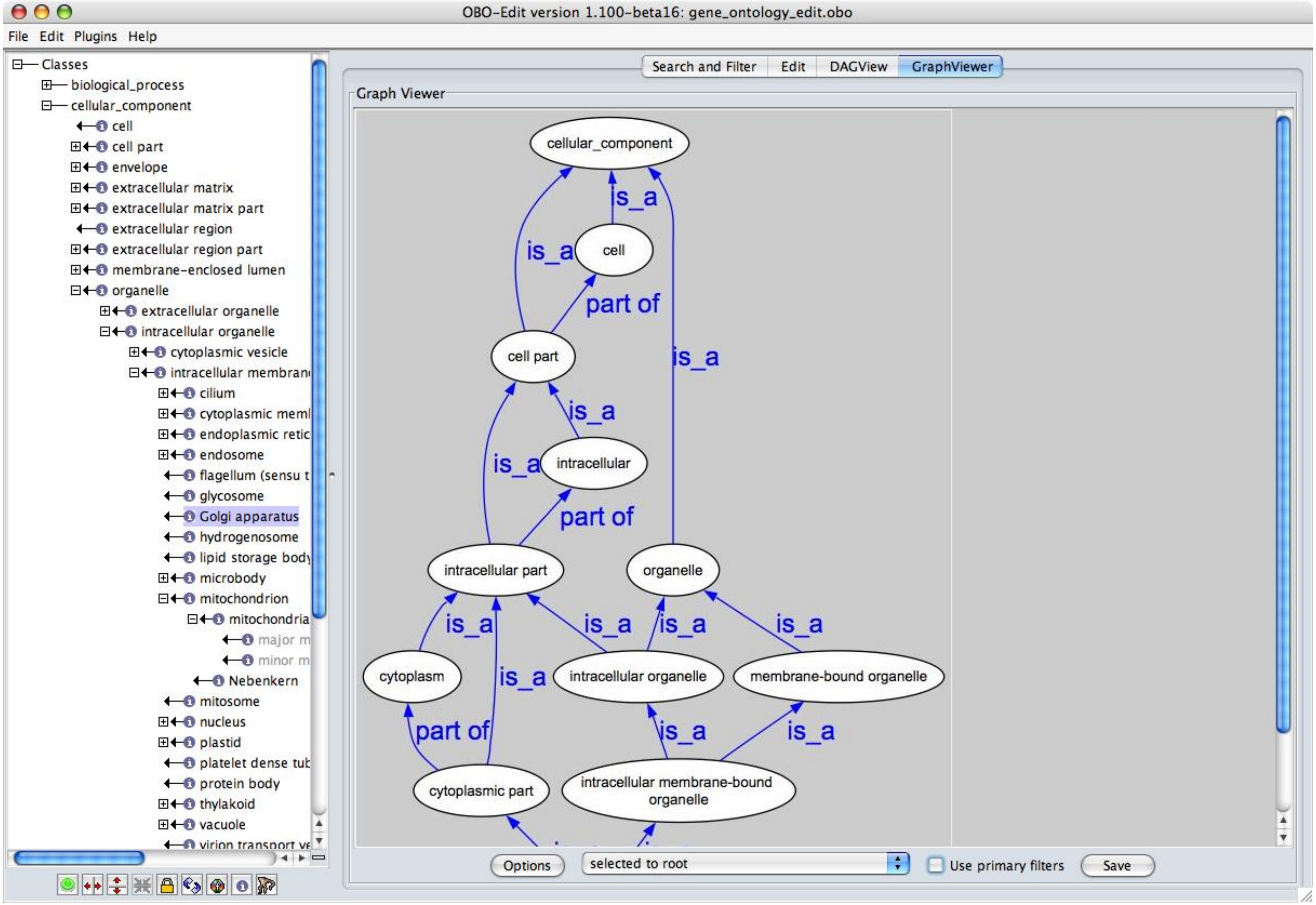
GO structure

- **GO terms divided into three parts:**
  - cellular component
  - molecular function
  - biological process
- **What each of the three parts tell us???**

Cellular Component

- **Where a gene product acts**





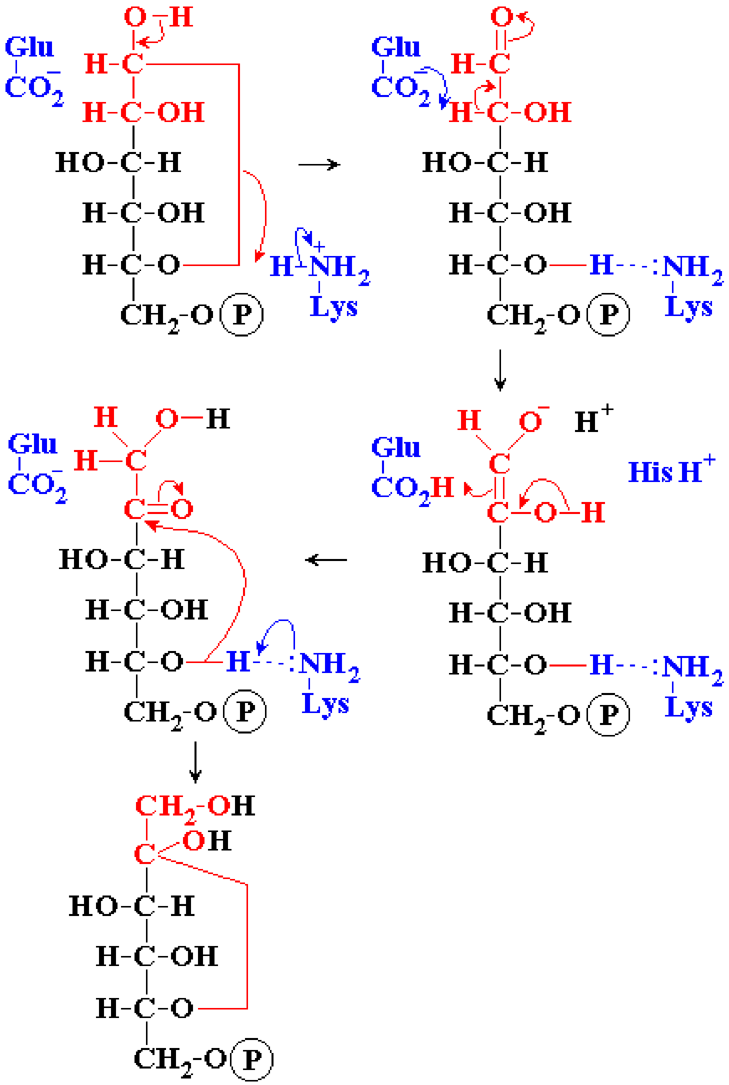
OBO-Edit version 1.100-beta16: gene_ontology_edit.obo

selected to root

☐ Use primary filters

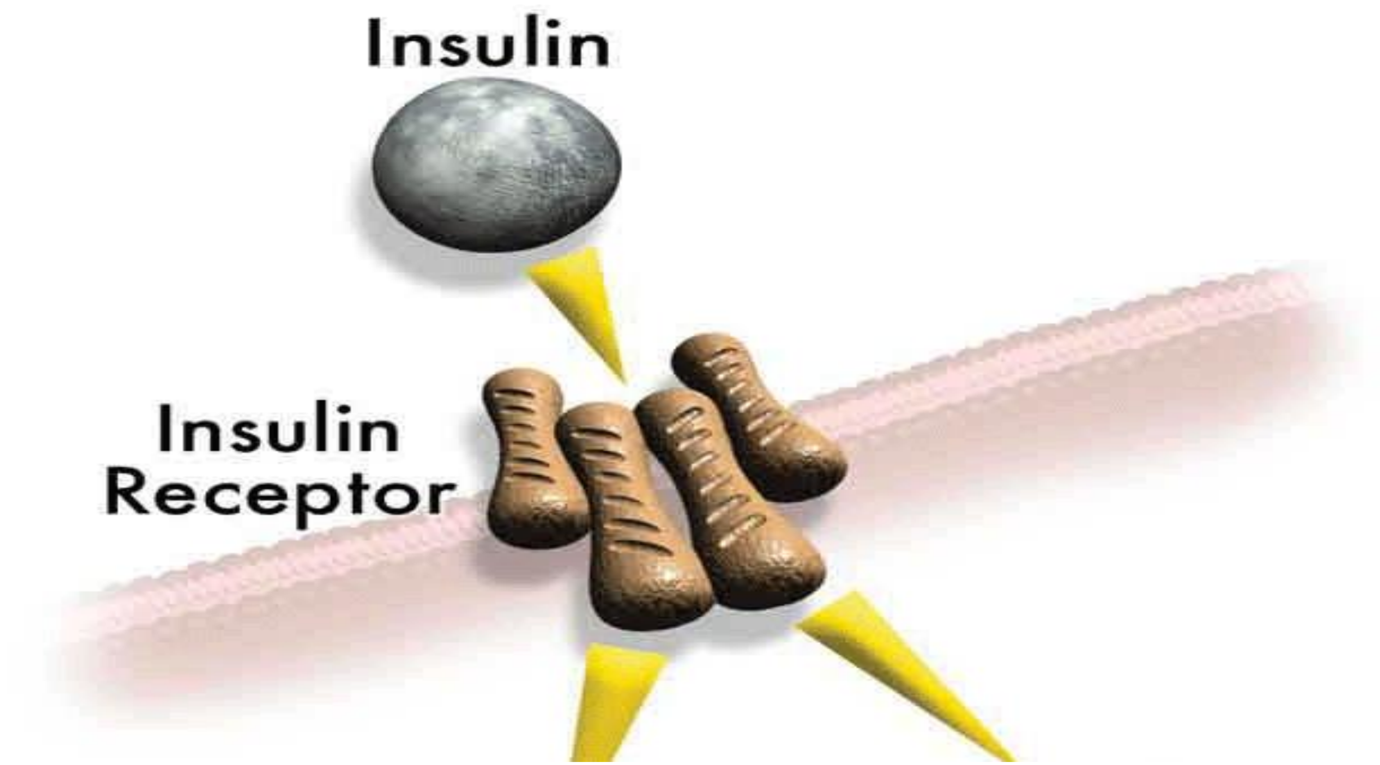
Molecular function

- **Activities or “jobs” of a gene product**



glucose-6-phosphate isomerase activity

Molecular function



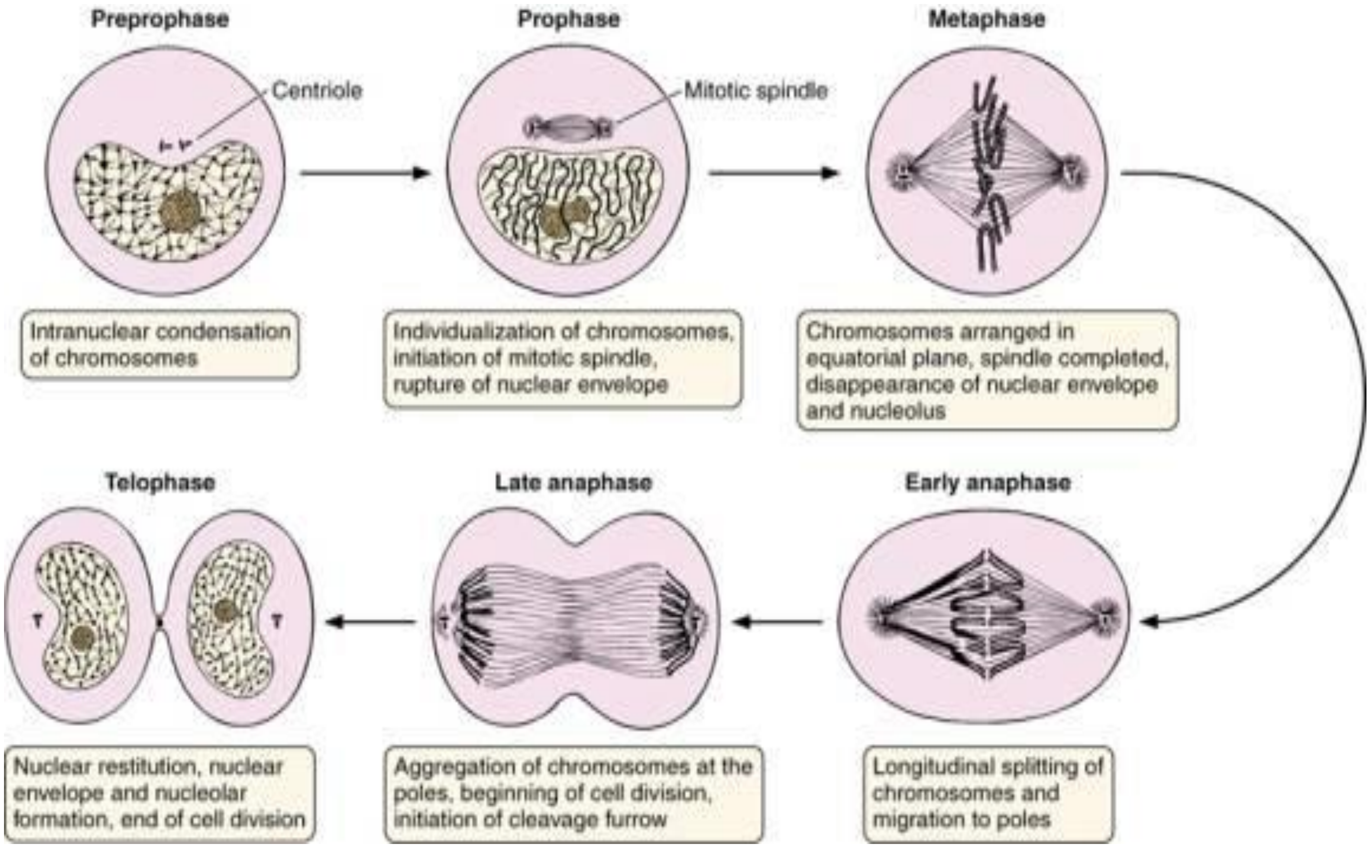
- **insulin binding**
- **insulin receptor activity**

Molecular function

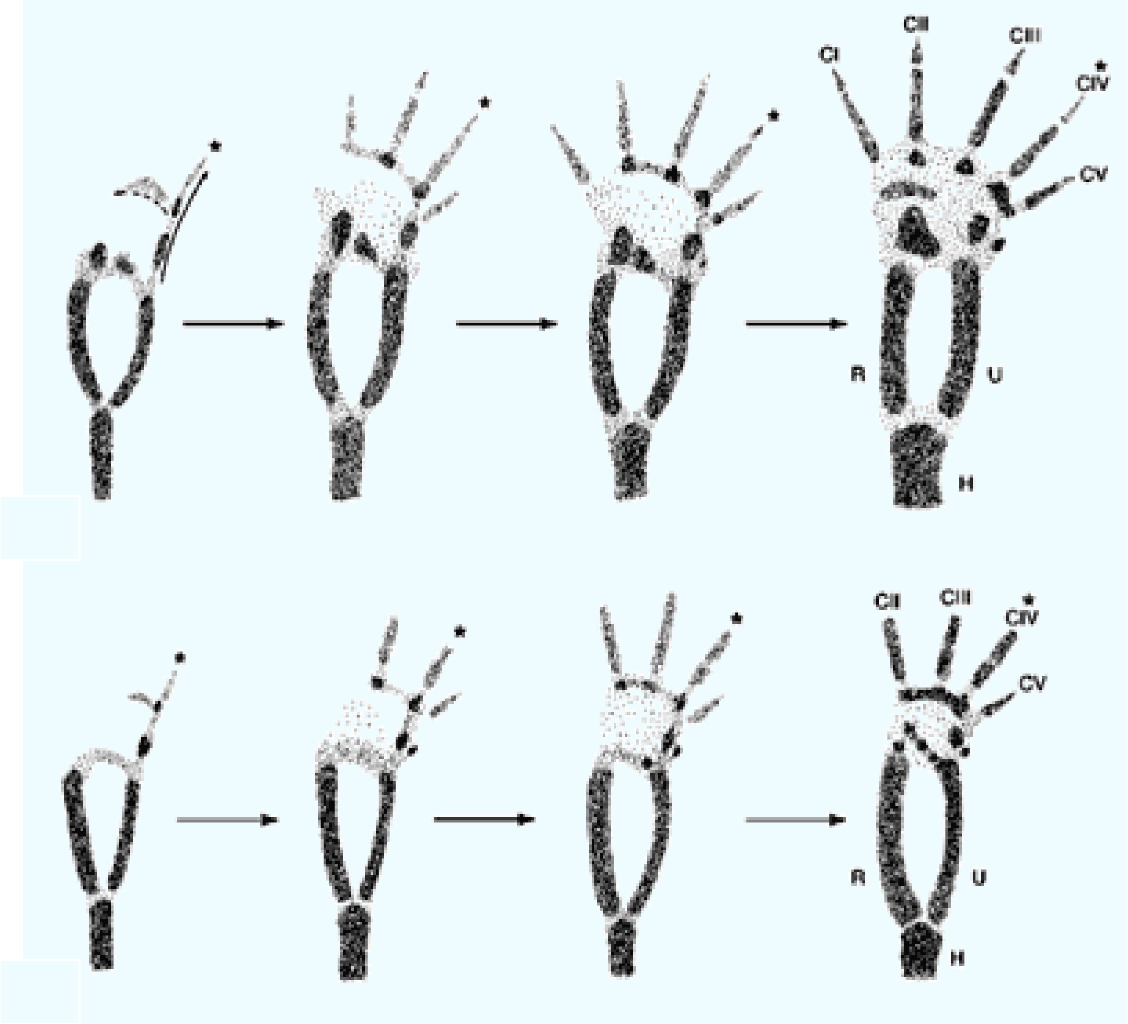
- A gene product may have several functions; a function term refers to a reaction or activity
- Sets of functions make up a biological process

Biological process

- **A commonly recognized series of events, e.g. cell division**


Preprophase
Centriole
Intranuclear condensation of chromosomes
Prophase
Mitotic spindle
Individualization of chromosomes, initiation of mitotic spindle, rupture of nuclear envelope
Metaphase
Chromosomes arranged in equatorial plane, spindle completed, disappearance of nuclear envelope and nucleolus
Telophase
Nuclear restitution, nuclear envelope and nucleolar formation, end of cell division
Late anaphase
Aggregation of chromosomes at the poles, beginning of cell division, initiation of cleavage furrow
Early anaphase
Longitudinal splitting of chromosomes and migration to poles

Biological process: limb development

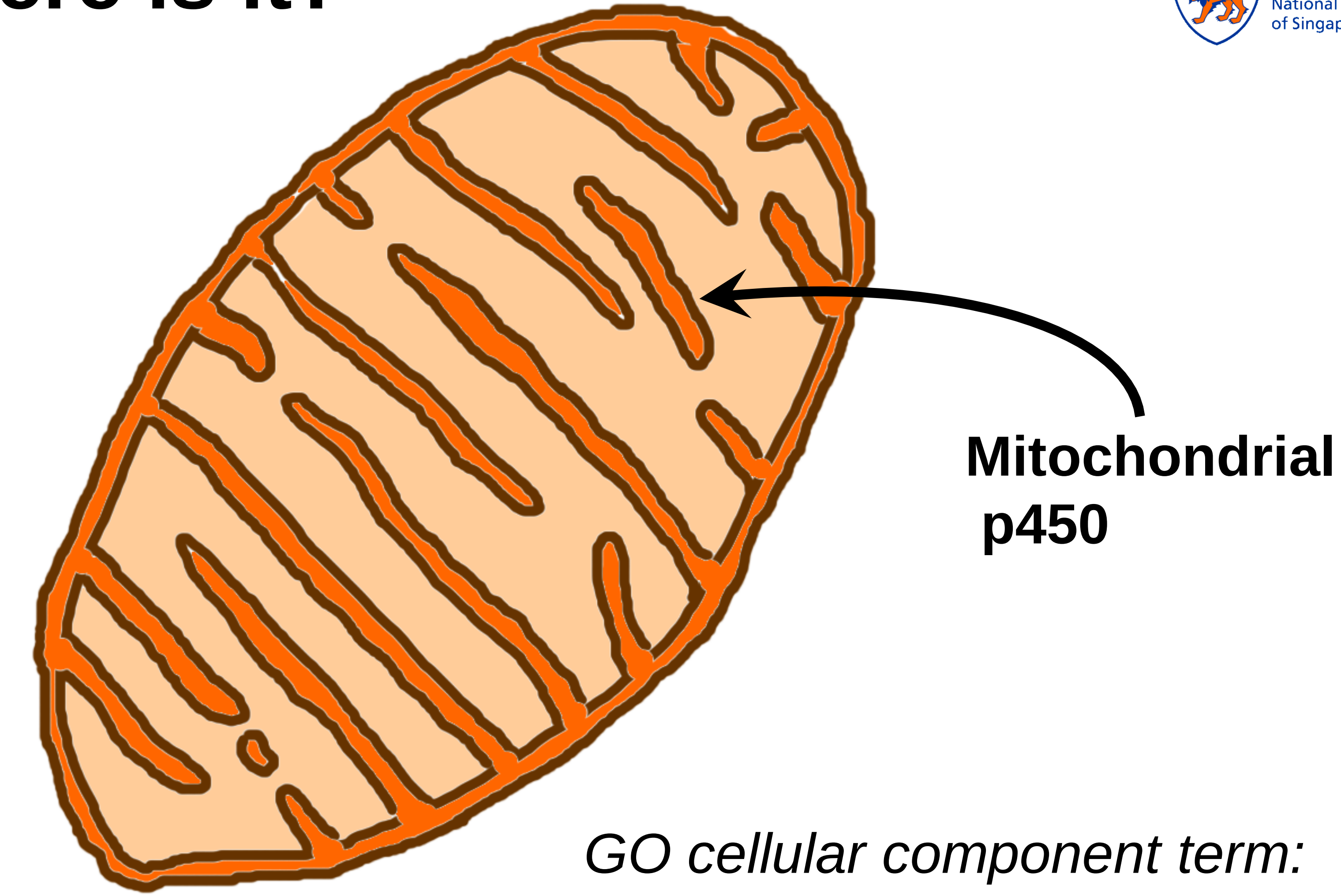


Annotation for Genes

Mitochondrial P450

This is a gene product that has already been annotated to all three gene ontologies. It is the Mitochondrial P450 gene product.

Where is it?



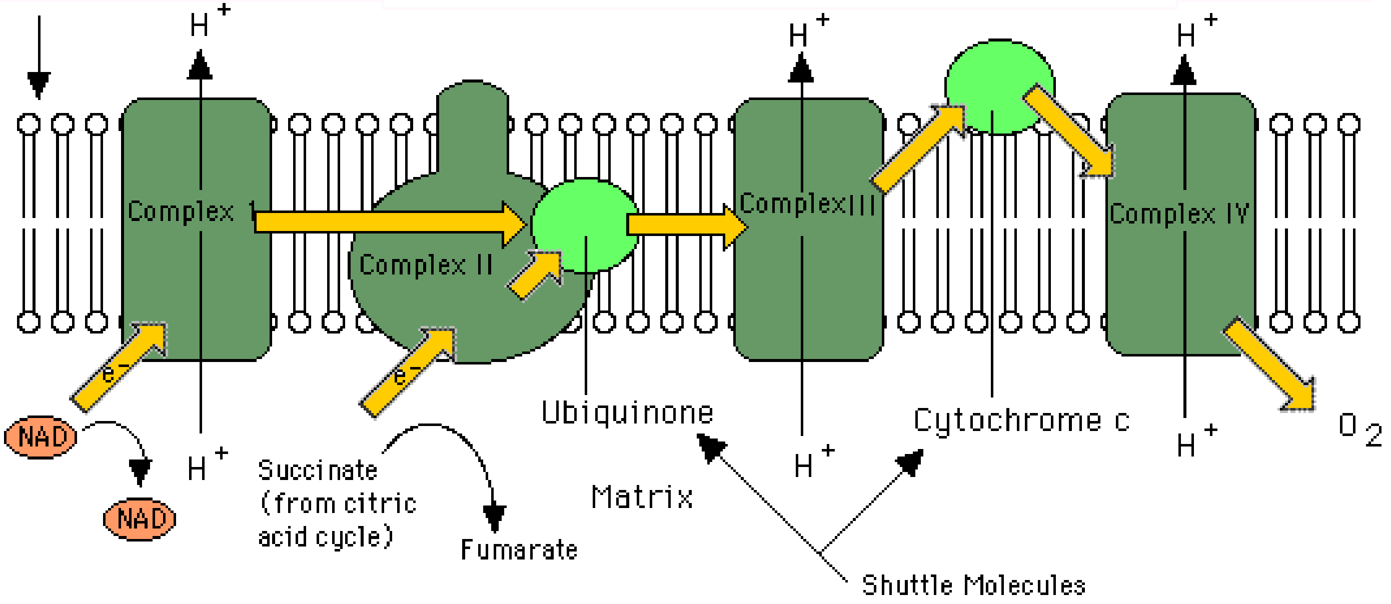
GO cellular component term:
mitochondrial inner membrane ;
GO:0005743

What does it do?

substrate + O_2 = CO_2 +H_20 product

GO molecular function term:
monooxygenase activity ; GO:0004497

Which process is this?



GO biological process term:
electron transport ; GO:0006118

http://ntri.tamuk.edu/cell/mitochondrion/krebpic.html

References on gene expression data classification

- E.-J. Yeoh et al., “Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling”, *Cancer Cell*, 1:133--143, 2002
- H. Liu, J. Li, L. Wong. Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data. *Bioinformatics*, 21(16):3377--3384, 2005.
- L.D. Miller et al., “Optimal gene expression analysis by microarrays”, *Cancer Cell* 2:353--361, 2002
- J. Li, L. Wong, “Techniques for Analysis of Gene Expression”, *The Practical Bioinformatician*, Chapter 14, pages 319—346, WSPC, 2004
- B. Bolstad et al. “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”. *Bioinformatics*, 19:185–193. 2003

Thank You

Contact: xlli@i2r.a-star.edu.sg if you have questions