Scribing: Clustering

Name: Lai Jacobi Wing Ki   SID: 1155159737

1. What is Clustering?

- clustering analysis is defined as 'finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups'

Benefit of clustering

- **Cluster items**
  - Better organization
  - Faster searching
- **Cluster people**
  - Patients: different treatment for different groups
  - Customers: different groups with different needs
    - Optimize the product based on the need of the targeting group
- **Cluster genes**
  - Identify co-expressed genes
    - Involved in the same pathway
  - Identify differentially expressed genes
    - Related to diseases
- **Cluster samples/cells**
  - Identify new disease sub-types
  - Identify new cell types
  - Discover new group
- **General Clustering**
  - Reduce the size of large data sets
  - Preserve privacy

What are needed to do clustering in gene aspect?

- Two sequences
- Dynamic programming algorithm
- A scoring matrix

2. Similarity and dissimilarity

Similarity

Dissimilarity

- Also call as distance

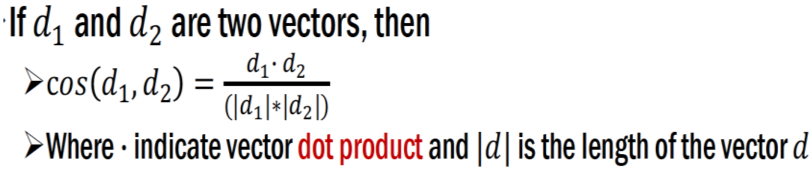
- Numerical measure of how alike two data objects are
- Higher when objects are more alike

- Numerical measure of how different two data objects are
- Lower when objects are more alike

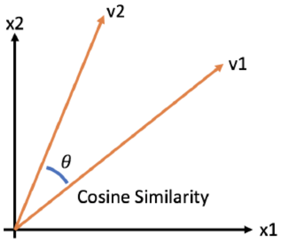
Method to find the similarity between data

- Cosine similarity
- Correlation
- Euclidean distance
- Minkowski distance

Cosine similarity


If d_1 and d_2 are two vectors, then

- $cos(d_1, d_2) = \frac{d_1 \cdot d_2}{(|d_1| * |d_2|)}$
- Where $\cdot$ indicate vector dot product and $|d|$ is the length of the vector d



Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**
d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481

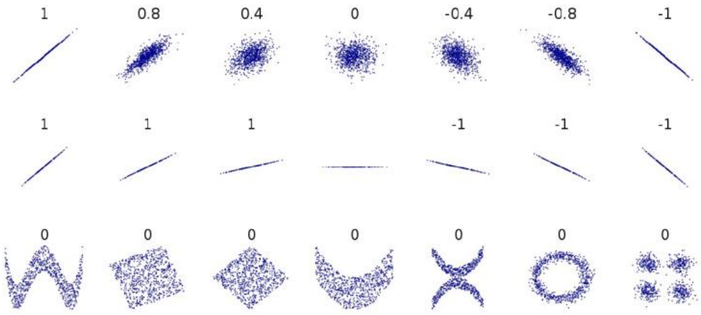
$||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

cos(d_1, d_2) = 0.3150

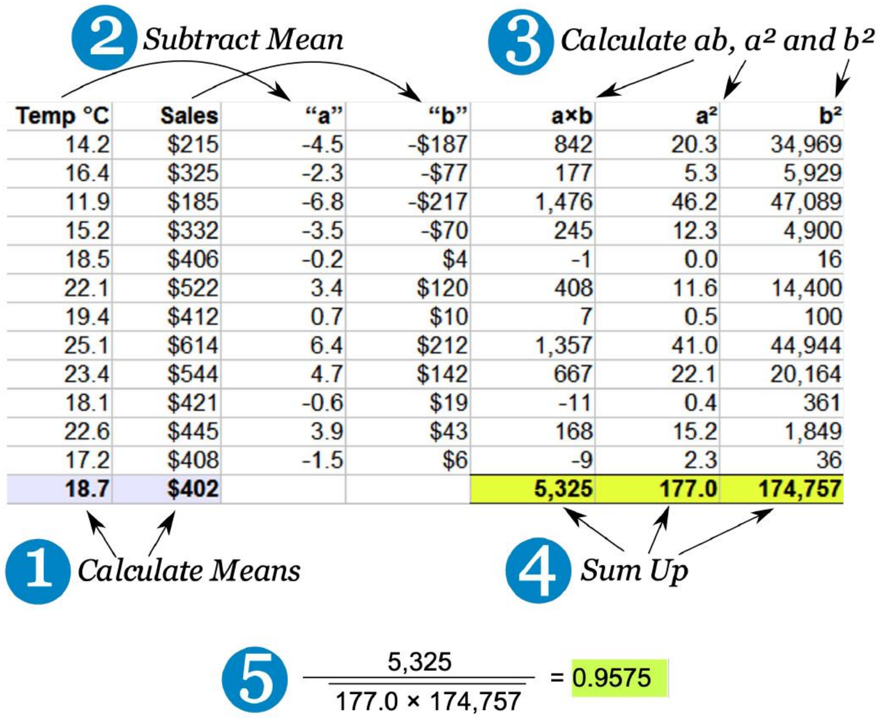
<u>Correlation</u>

- Correlation measures the linear relationship between objects

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



Example:


2 Subtract Mean
3 Calculate ab, a² and b²
1 Calculate Means
4 Sum Up
5

Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

$$\frac{5{,}325}{177.0 \times 174{,}757} = 0.9575$$

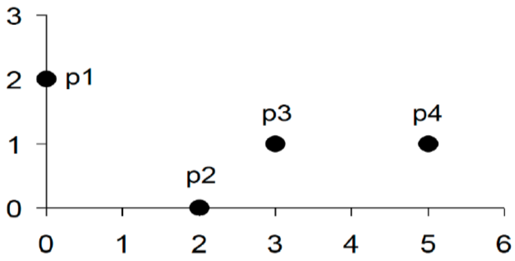
Euclidean distance

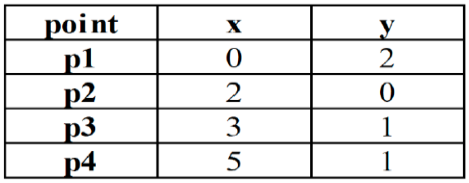
Euclidean distance

$$Ed(\boldsymbol{p},\boldsymbol{q})=\sqrt{\sum_{k=1}^{m}(p_k-q_k)^2}$$

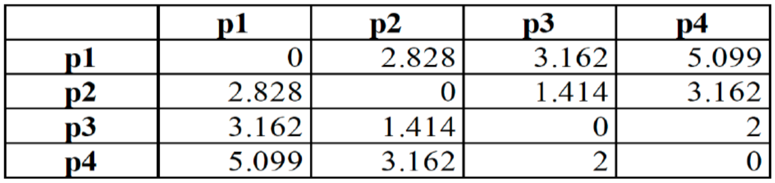
- Where m is the number of dimensions and pk and qk are, respectively, the k -th attributes of data objects $\boldsymbol{p}$ and $\boldsymbol{q}$.

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

Minkowski distance

- Minkowski Distance is a generalization of Euclidean Distance
  - r is a parameter
  - m is the number of dimensions
  - pk is the k-th attributes of data objects $\boldsymbol{p}$
  - qk is the k-th attributes of data objects $\boldsymbol{q}$

$$dist(\boldsymbol{p},\boldsymbol{q})=\left(\sum_{k=1}^{m}|p_k-q_k|^r\right)^{\frac{1}{r}}$$

$\boldsymbol{r}$ = 1 City block (Manhattan, taxicab, L1 norm) distance.

- Example: Hamming distance

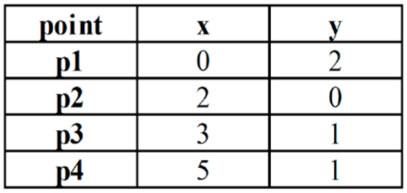
- number of bits that are different between two binary vectors

r = 2 Euclidean distance

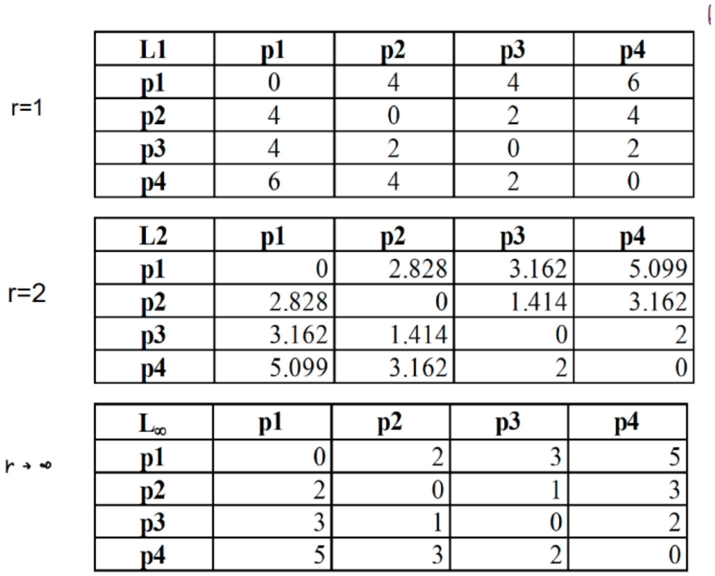
$r \to \infty$ "supremum" (Lmax norm, $L\infty$ norm) distance.

- maximum difference between any component of the vectors

Example:



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



r=1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

r=2

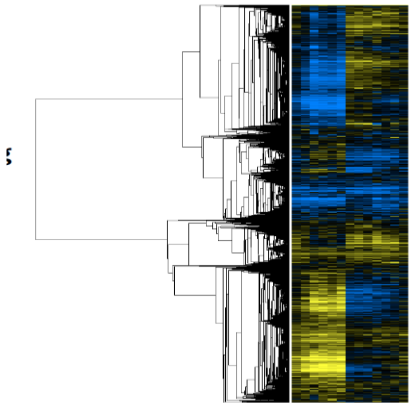
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

$r \to \infty$

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

3. Hierarchical clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
- A tree like diagram that records the sequences of merges
- They may correspond to meaningful taxonomies, like Gene clusters, phylogeny reconstruction

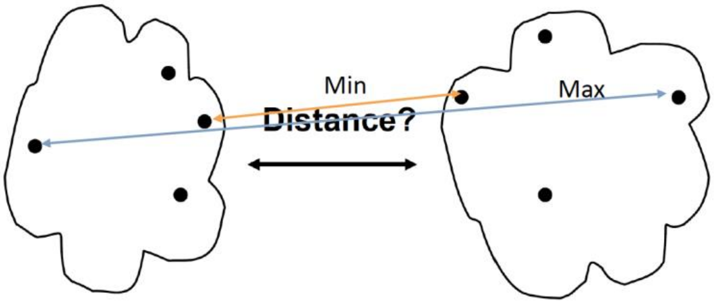


Steps of hierarchical clustering

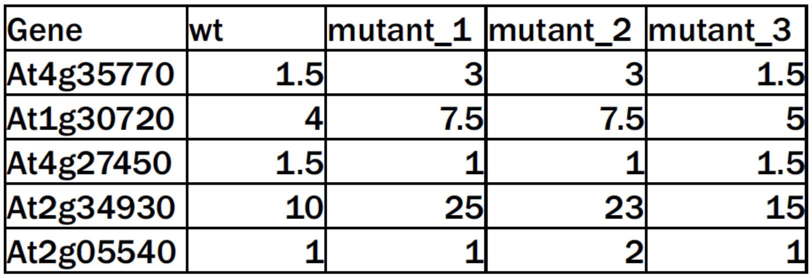
1. Compute the Similarity or Distance matrix
2. Let each data point be a cluster
3. Merge the two closest clusters
4. Update the similarity or distance matrix (first time)
5. Merge the two closest clusters
6. Update the similarity or distance matrix (second time)
7. Continue the previous two steps
8. Until only a single cluster remains

Ways to update the distance matrix after merging

- Min
- Max
- Group Average
- Distance between centroids



Example:

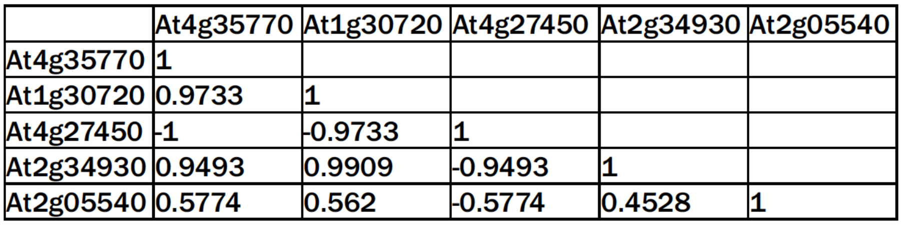


Gene	wt	mutant_1	mutant_2	mutant_3
At4g35770	1.5	3	3	1.5
At1g30720	4	7.5	7.5	5
At4g27450	1.5	1	1	1.5
At2g34930	10	25	23	15
At2g05540	1	1	2	1

Step 1: Use correlation

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

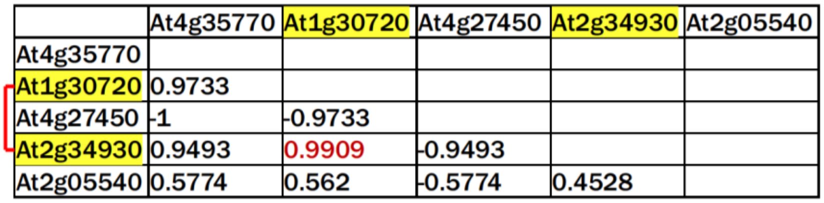
Then we will get a new graph



	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770	1				
At1g30720	0.9733	1			
At4g27450	-1	-0.9733	1		
At2g34930	0.9493	0.9909	-0.9493	1	
At2g05540	0.5774	0.562	-0.5774	0.4528	1

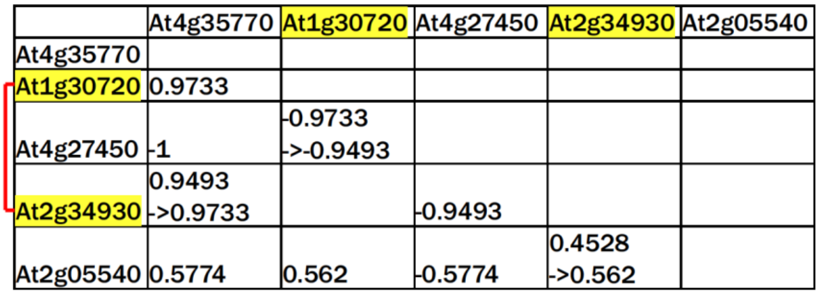
Step 2: remove the 1

Step 3: Merge the two closest matrix (At2g34930 and At1g30720)



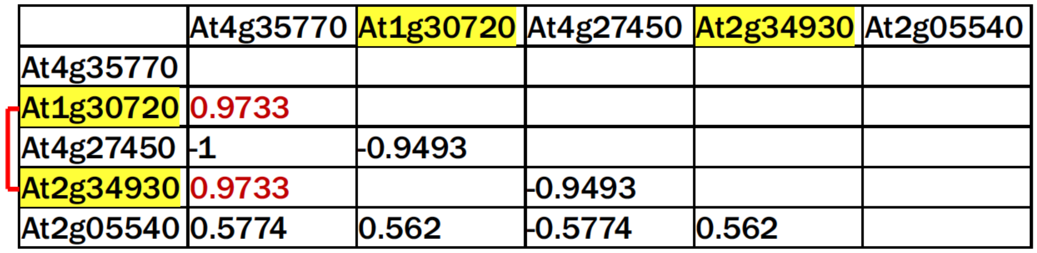
	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733			
At2g34930	0.9493	0.9909	-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528	

Step 4: Update with minimum distance (largest correlation)



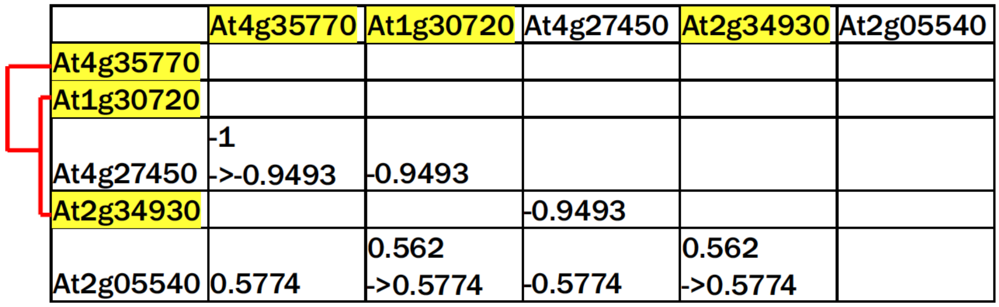
	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733 ->-0.9493			
At2g34930	0.9493 ->0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528 ->0.562	

Step 5: Merge the two closest matrix (At2g34930 , At1g30720 and At4g35770)



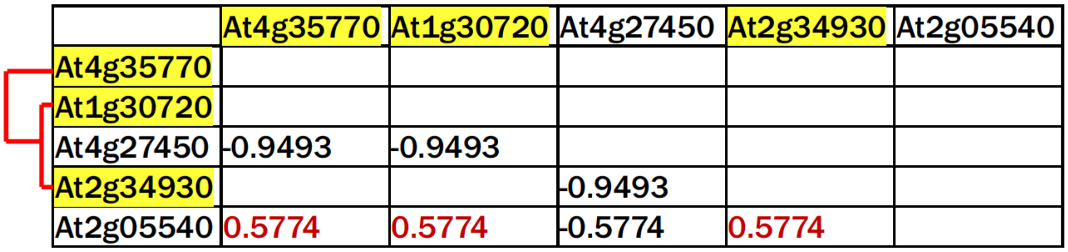
	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9493			
At2g34930	0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.562	

Step 6: Update with minimum distance (largest correlation)



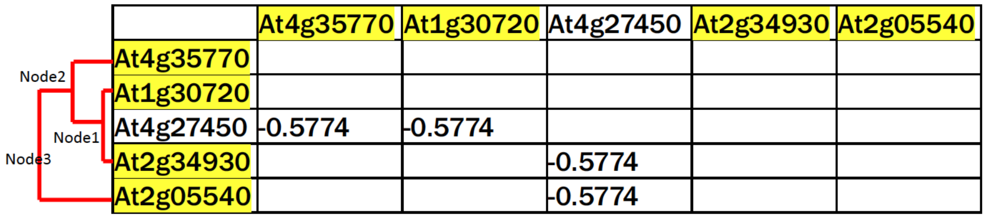
	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-1 ->-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.562 ->0.5774	-0.5774	0.562 ->0.5774	

Step 7: Merge the two closest matrix (At2g34930 , At1g30720 , At4g35770 and At2g05540)



	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.5774	-0.5774	0.5774	

Step 8: Update with minimum distance (largest correlation)


Node2
Node1
Node3

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.5774	-0.5774			
At2g34930			-0.5774		
At2g05540			-0.5774		

This is the end of forming a hierarchical clustering