Algorithms in Bioinformatics
Spring 2023
Lecture 5

Jason Ernst

University of California, Los Angeles

Announcements

- Office hours Tue 4/18 and 4/25 4-5pm moved to BSRB 350B (also OHRC) and just with Prof. Ernst
- We will not be assigning paper 2
- We have revised the due dates for HW3, HW4, Project 1, and Project 2:
  - HW3 - chapter 8 now due 4/25 (previously due 4/21)
  - HW4 - chapter 2 now due 5/1 (previously due 4/28)
  - Project 1a now due 4/27 (previously due 4/20)
  - Project 1b now due 5/4 (previously due 4/25)
  - Project 2a now due 5/9 (previously due 5/4)
  - Project 2b now due 5/11 (previously due 5/9)

You are encouraged to work ahead of the due dates!

Clustering – part 2

Lecture 5

April 18th, 2023

Some slides based on Andrew Moore and Manolis Kellis slides

Topics

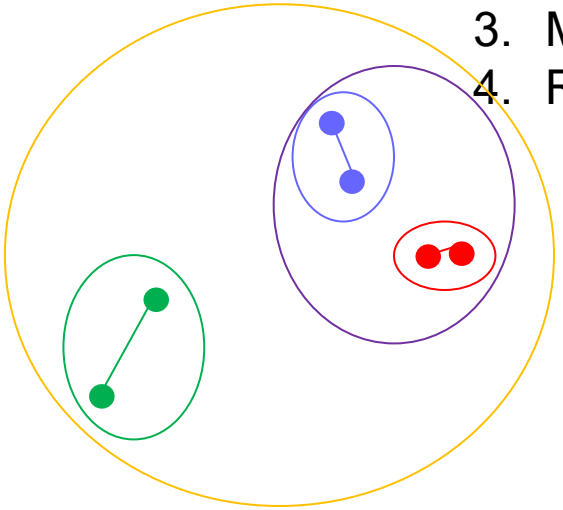
- Hierarchical clustering (review/follow-up)
- K-means clustering (review/follow-up)
- Soft-clustering/Gaussian Mixture Models/Expectation-Maximization algorithm
- Clustering short time series gene expression data

Topics

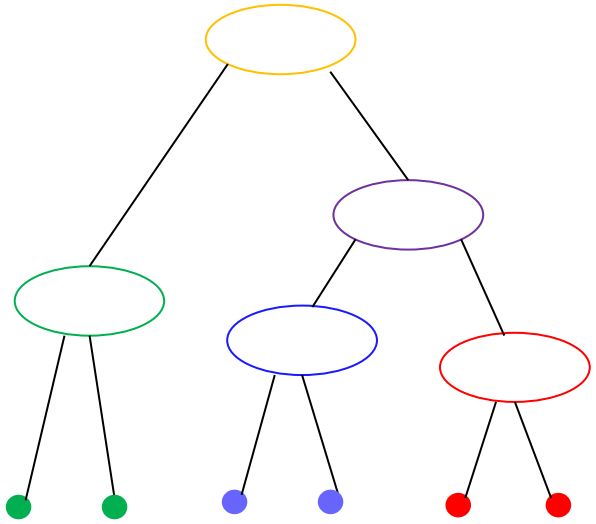
- Hierarchical clustering (review/follow-up)
- K-means clustering (review/follow-up)
- Soft-clustering/Gaussian Mixture Models/Expectation-Maximization algorithm
- Clustering short time series gene expression data

Hierarchical Clustering

1. Initially each point is its own cluster
2. Find pair of clusters with smallest distance between them or equivalently are the most similar
3. Merge into parent cluster
4. Repeat



Suppose we continue with complete linkage



Measuring Distance between Clusters

- Single Linkage Clustering

$$\mathrm{D}(\mathrm{X},\mathrm{Y}) = min_{x\in X, y\in Y} d(x,y)$$

- Complete Linkage Clustering

$$\mathrm{D}(\mathrm{X},\mathrm{Y}) = max_{x\in X, y\in Y} d(x,y)$$

- Average Linkage Clustering

$$\mathrm{D}(\mathrm{X},\mathrm{Y}) = \frac{1}{|X||Y|}\sum_{x\in X}\sum_{y\in Y} d(x,y)$$

- Centroid Linkage Clustering

$$\mathrm{D}(\mathrm{X},\mathrm{Y}) = ||c_X - c_Y||$$

where c_X and c_Y are the mean of *X* and *Y* and data assumed to be in $\mathbb{R}^d$

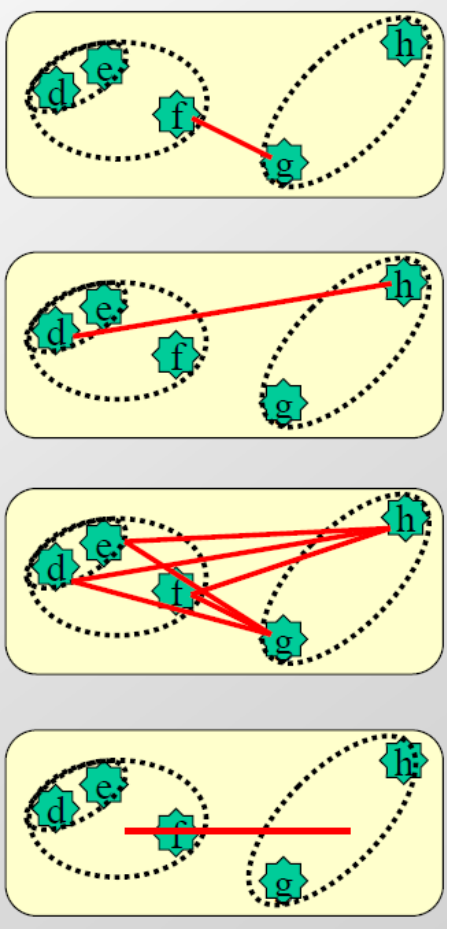


Image from Manolis Kellis

Hierarchical clustering

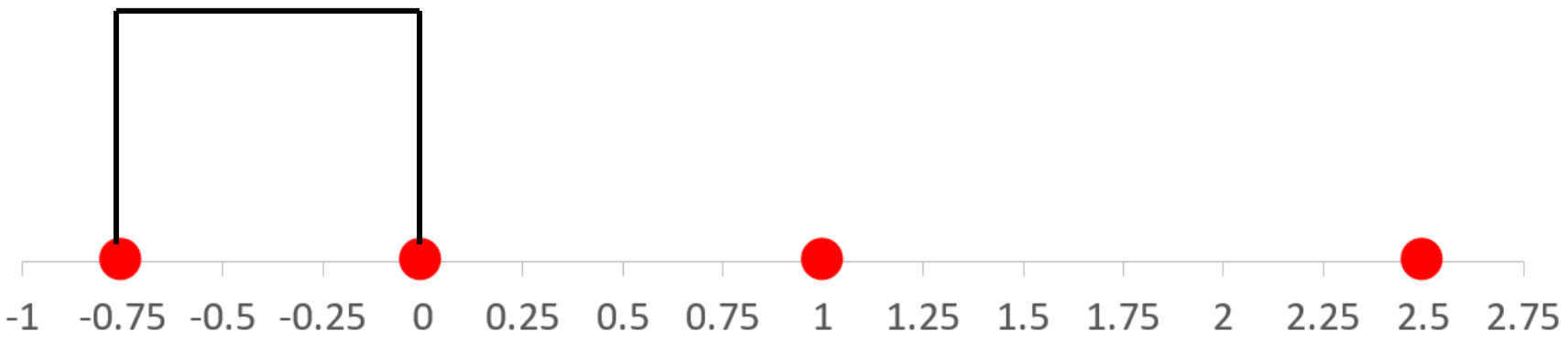
Suppose we want to perform hierarchical clustering with Euclidean distance using single linkage clustering to the four data points below.

Hierarchical clustering

Suppose we want to perform hierarchical clustering with Euclidean distance using single linkage clustering to the four data points below.
Question: What would the first merge be?

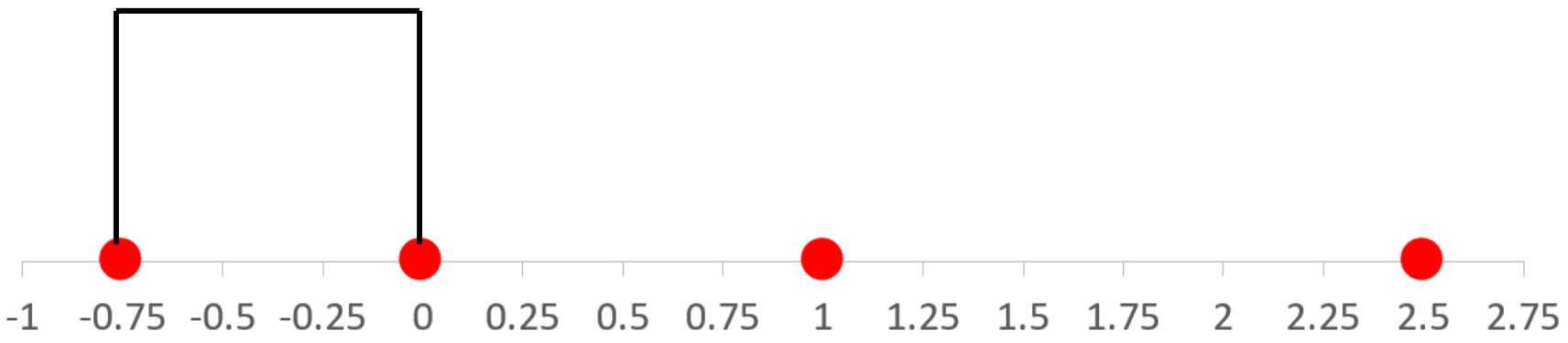
Hierarchical clustering

Suppose we want to perform hierarchical clustering with Euclidean distance using single linkage clustering to the four data points below.
Question: What would the first merge be?



Hierarchical clustering

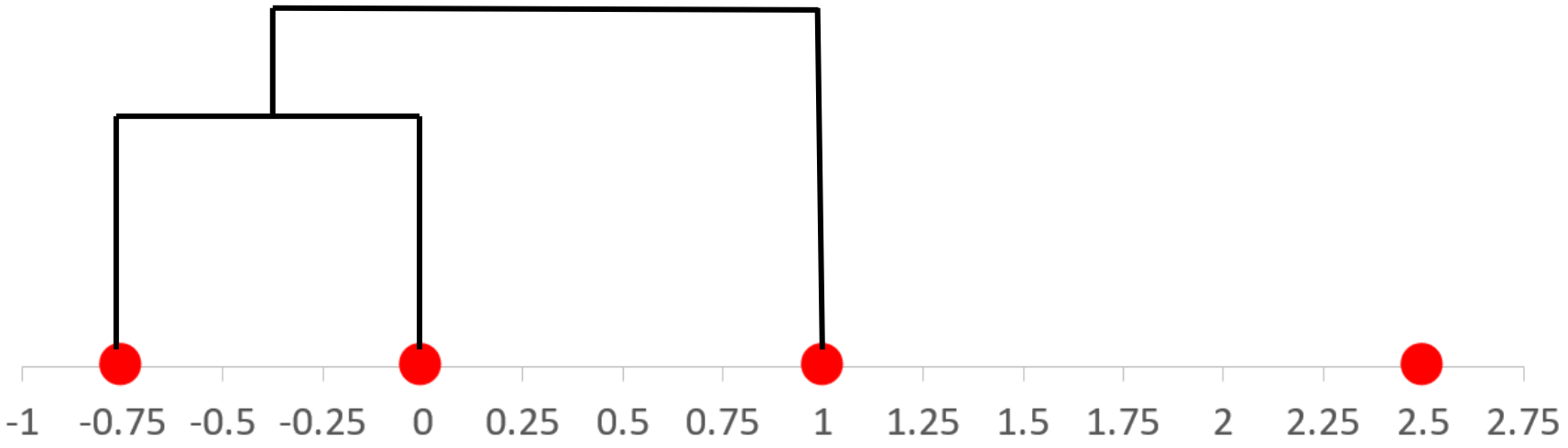
Suppose we want to perform hierarchical clustering with Euclidean distance using single linkage clustering to the four data points below.
Question: What would the second merge be?



Hierarchical clustering

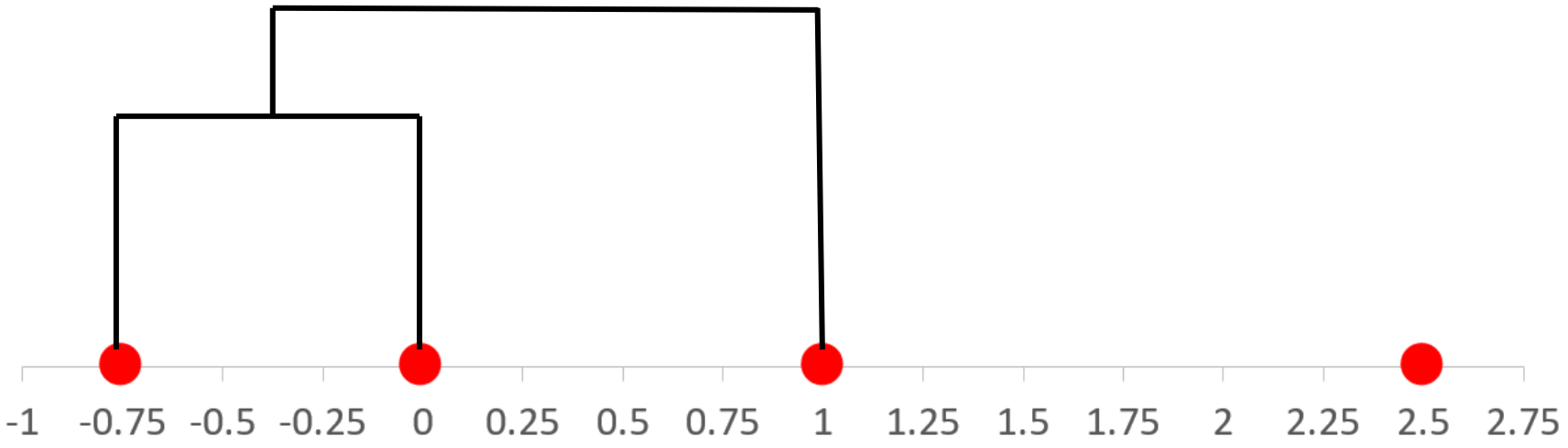
Suppose we want to perform hierarchical clustering with Euclidean distance using single linkage clustering to the four data points below.
Question: What would the second merge be?

Since (1 < 1.5)



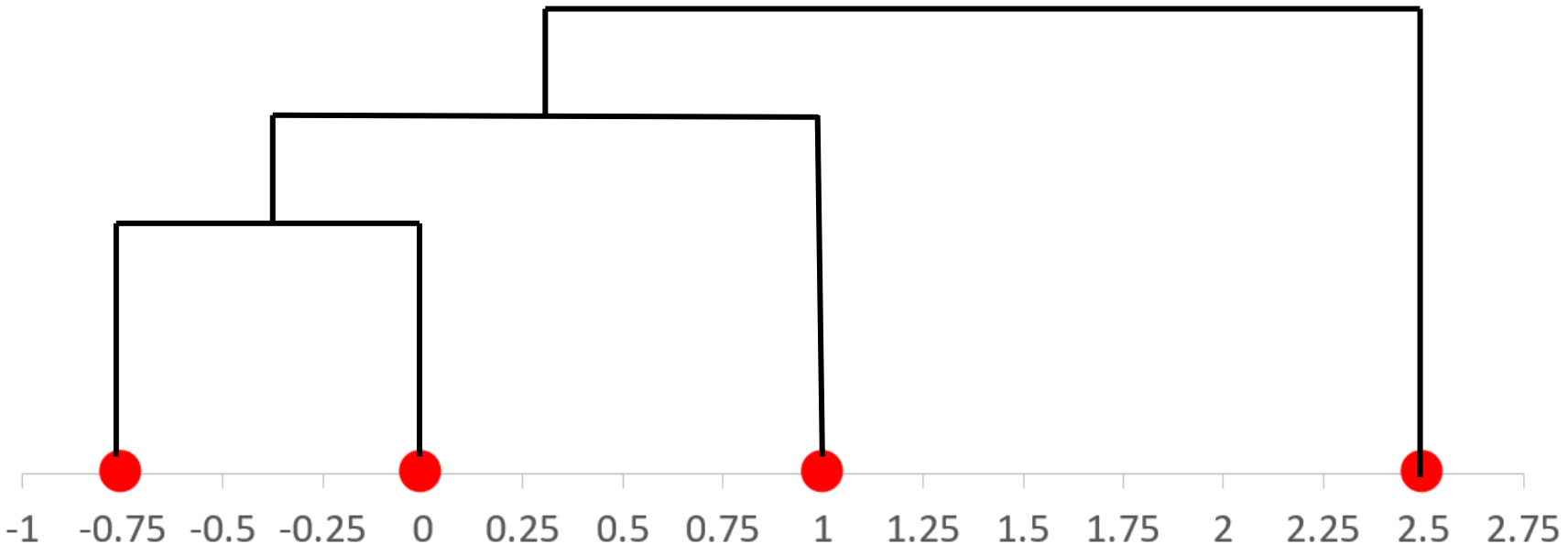
Hierarchical clustering

Suppose we want to perform hierarchical clustering with Euclidean distance using single linkage clustering to the four data points below.
Question: What would the final merge be?



Hierarchical clustering

Suppose we want to perform hierarchical clustering with Euclidean distance using single linkage clustering to the four data points below.
Question: What would the final merge be?

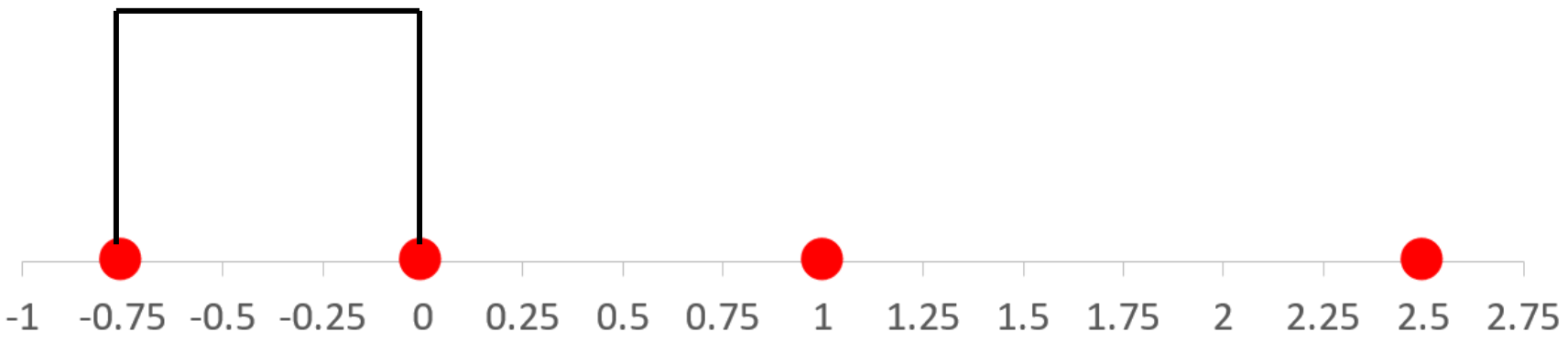


Hierarchical clustering

Now, suppose we want to perform hierarchical clustering with Euclidean distance using complete linkage clustering to the four data points below.
Question: What would the first merge be?

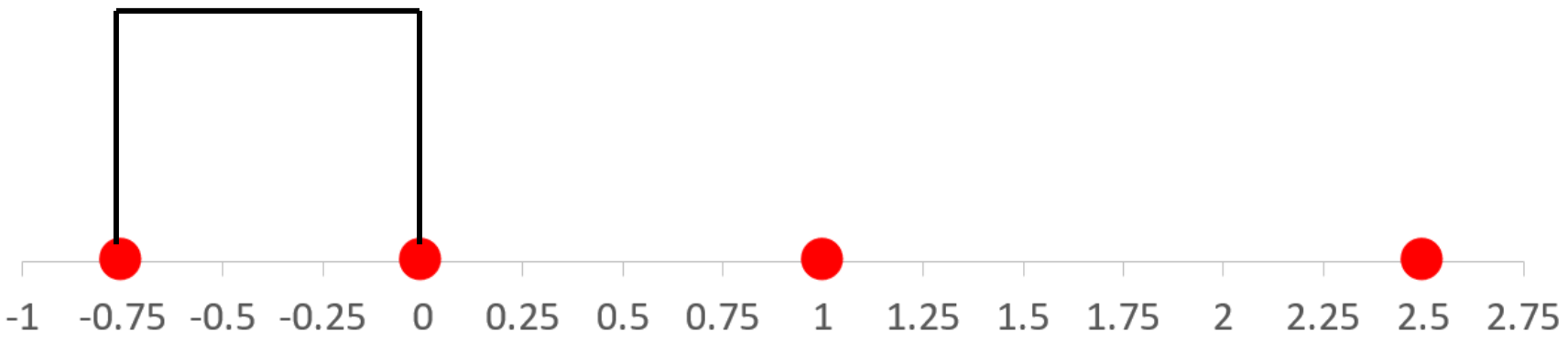
Hierarchical clustering

Now, suppose we want to perform hierarchical clustering with Euclidean distance using complete linkage clustering to the four data points below.
Question: What would the first merge be?



Hierarchical clustering

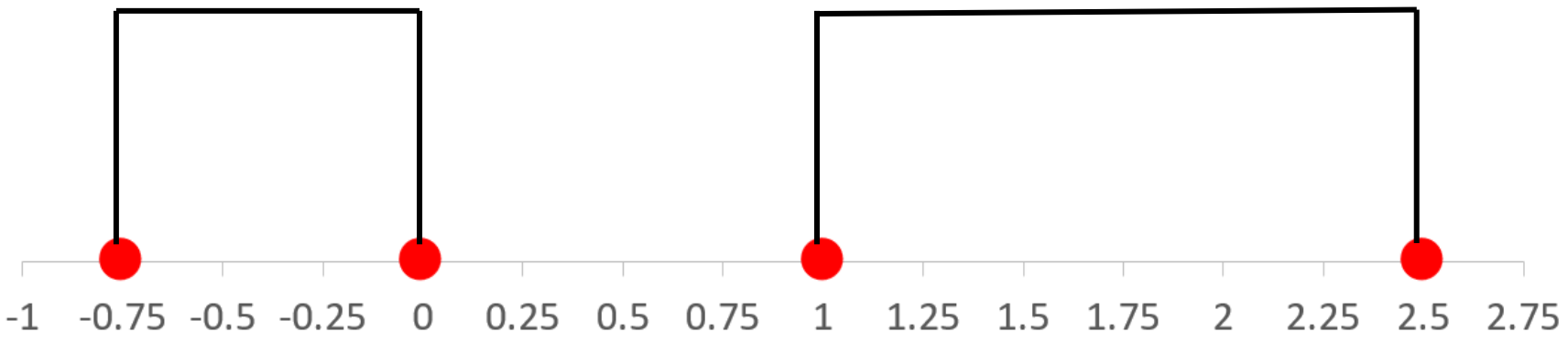
Now, suppose we want to perform hierarchical clustering with Euclidean distance using complete linkage clustering to the four data points below.
Question: What would the second merge be?



Hierarchical clustering

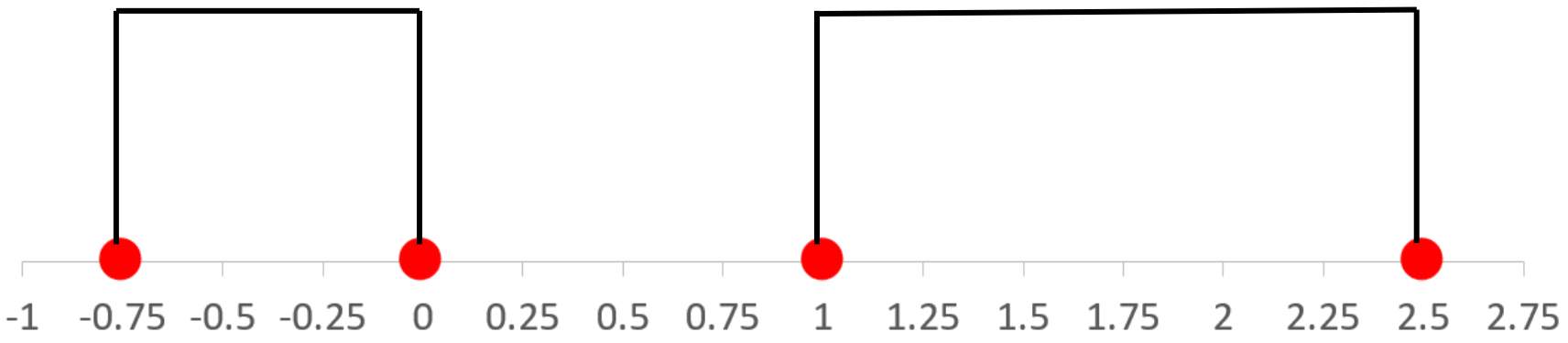
Now, suppose we want to perform hierarchical clustering with Euclidean distance using complete linkage clustering to the four data points below.
Question: What would the second merge be?

Since (1.5 < 1.75)



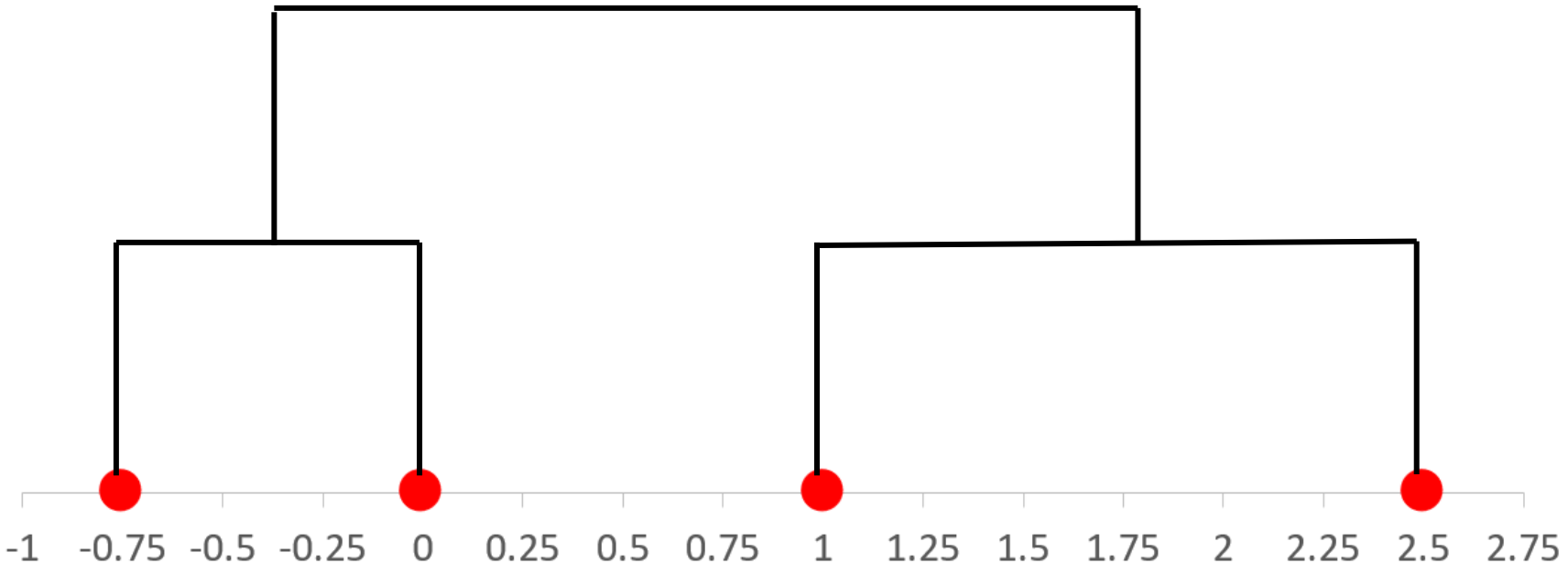
Hierarchical clustering

Now, suppose we want to perform hierarchical clustering with Euclidean distance using complete linkage clustering to the four data points below.
Question: What would the final merge be?



Hierarchical clustering

Now, suppose we want to perform hierarchical clustering with Euclidean distance using complete linkage clustering to the four data points below.
Question: What would the final merge be?



Hierarchical clustering

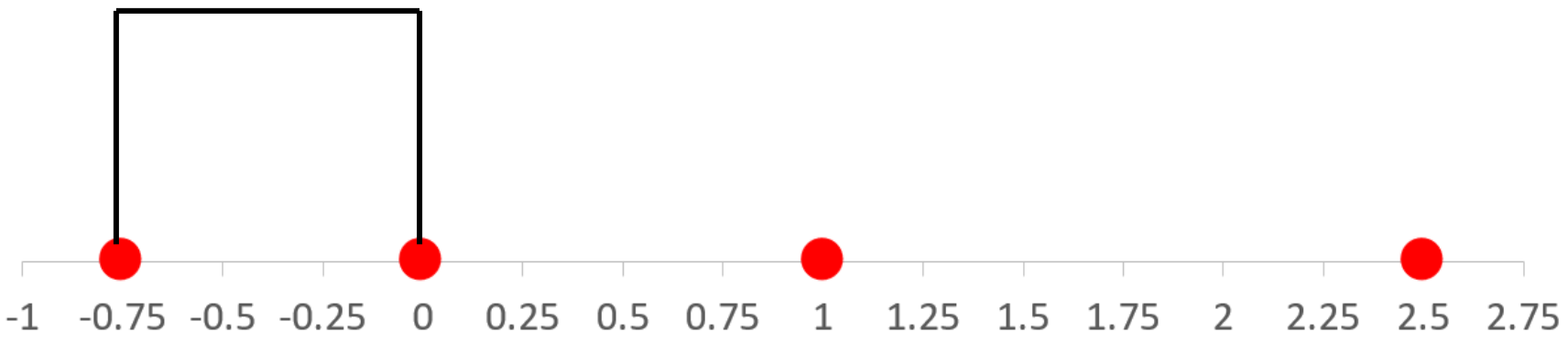
Suppose we want to perform hierarchical clustering with Euclidean distance using average linkage clustering to the four data points below.

Hierarchical clustering

Suppose we want to perform hierarchical clustering with Euclidean distance using average linkage clustering to the four data points below.
Question: What would the first merge be?

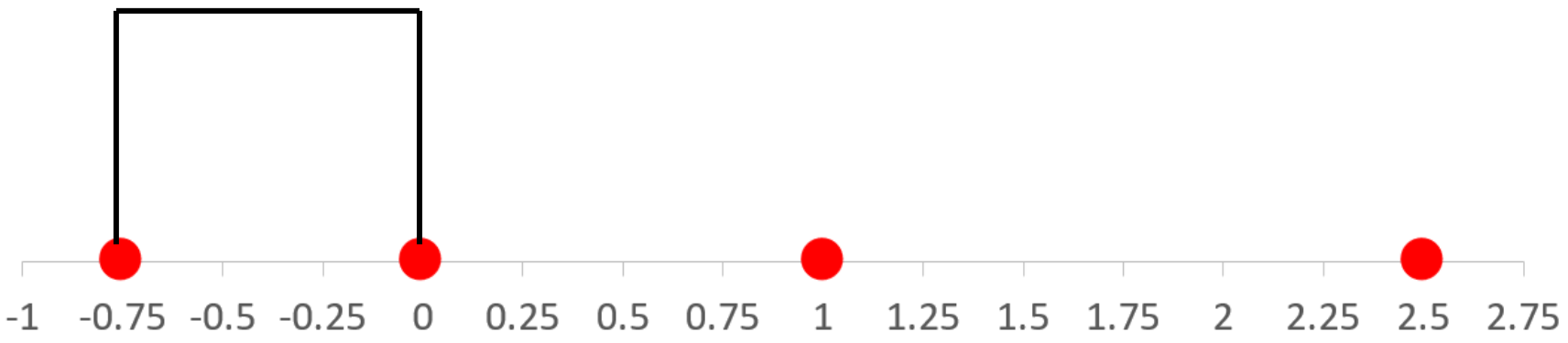
Hierarchical clustering

Suppose we want to perform hierarchical clustering with Euclidean distance using average linkage clustering to the four data points below.
Question: What would the first merge be?



Hierarchical clustering

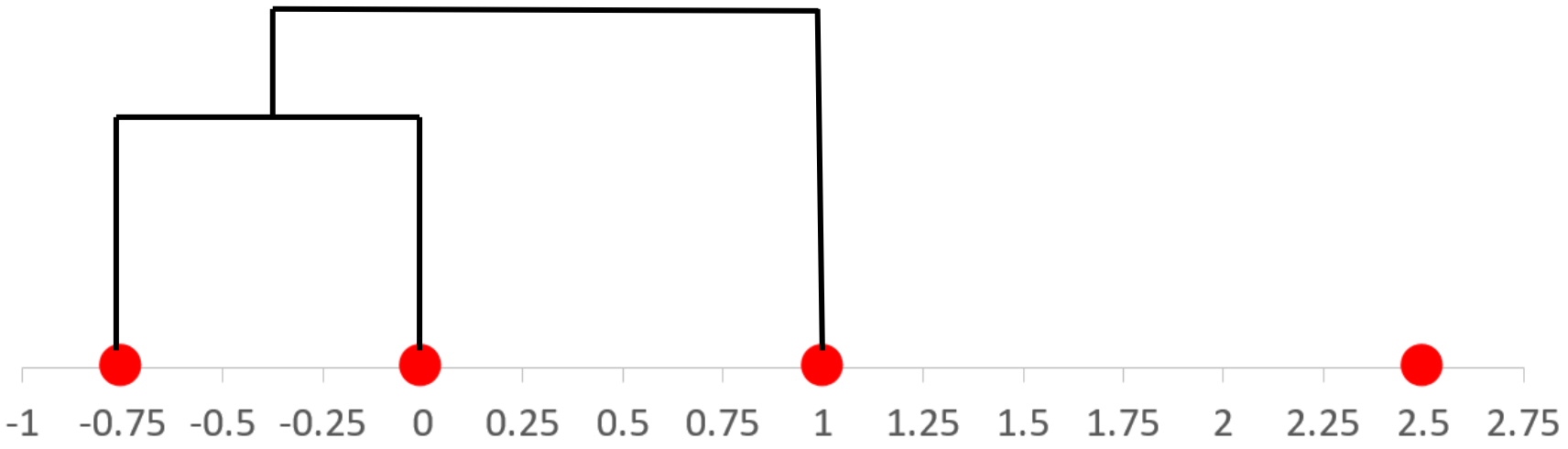
Suppose we want to perform hierarchical clustering with Euclidean distance using average linkage clustering to the four data points below.
Question: What would the second merge be?



Hierarchical clustering

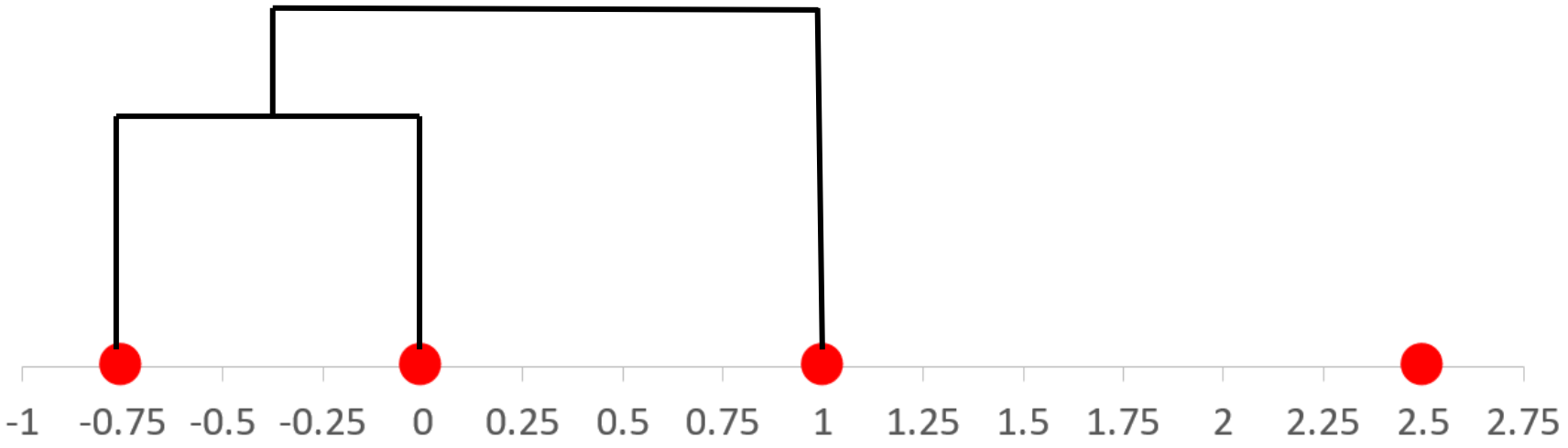
Suppose we want to perform hierarchical clustering with Euclidean distance using average linkage clustering to the four data points below.
Question: What would the second merge be?

Since ((0.75+1.75)/2=1.375< 1.5)



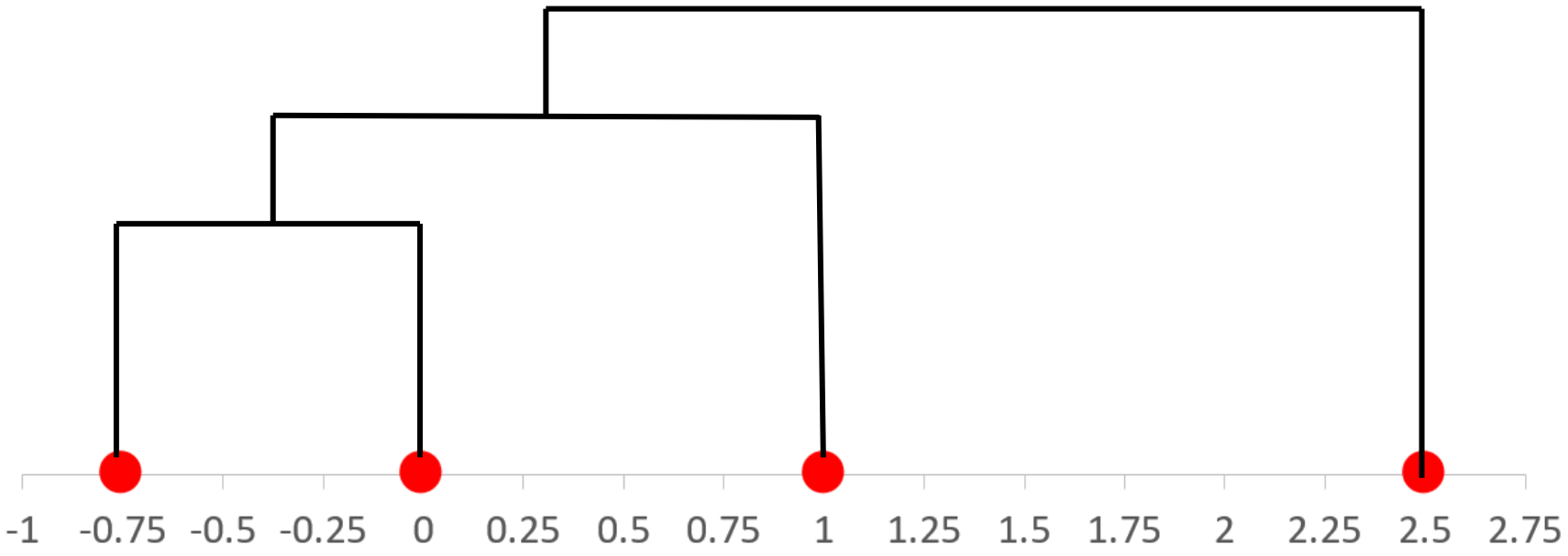
Hierarchical clustering

Suppose we want to perform hierarchical clustering with Euclidean distance using average linkage clustering to the four data points below.
Question: What would the final merge be?



Hierarchical clustering

Suppose we want to perform hierarchical clustering with Euclidean distance using average linkage clustering to the four data points below.
Question: What would the final merge be?



Topics

- Hierarchical clustering (review/follow-up)
- K-means clustering (review/follow-up)
- Soft-clustering/Gaussian Mixture Models/Expectation-Maximization algorithm
- Clustering short time series gene expression data

K-means Objective Function

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

μ_i Mean of cluster *i*

S_i Data points assigned to cluster *i*

K-means Objective Function

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

μ_i Mean of cluster *i*

S_i Data points assigned to cluster *i*

After initializing clustering centers, iterating between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

K-means algorithm example

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns…
5. …and jumps there
6. …Repeat until terminated!

(no improvement in objective function)

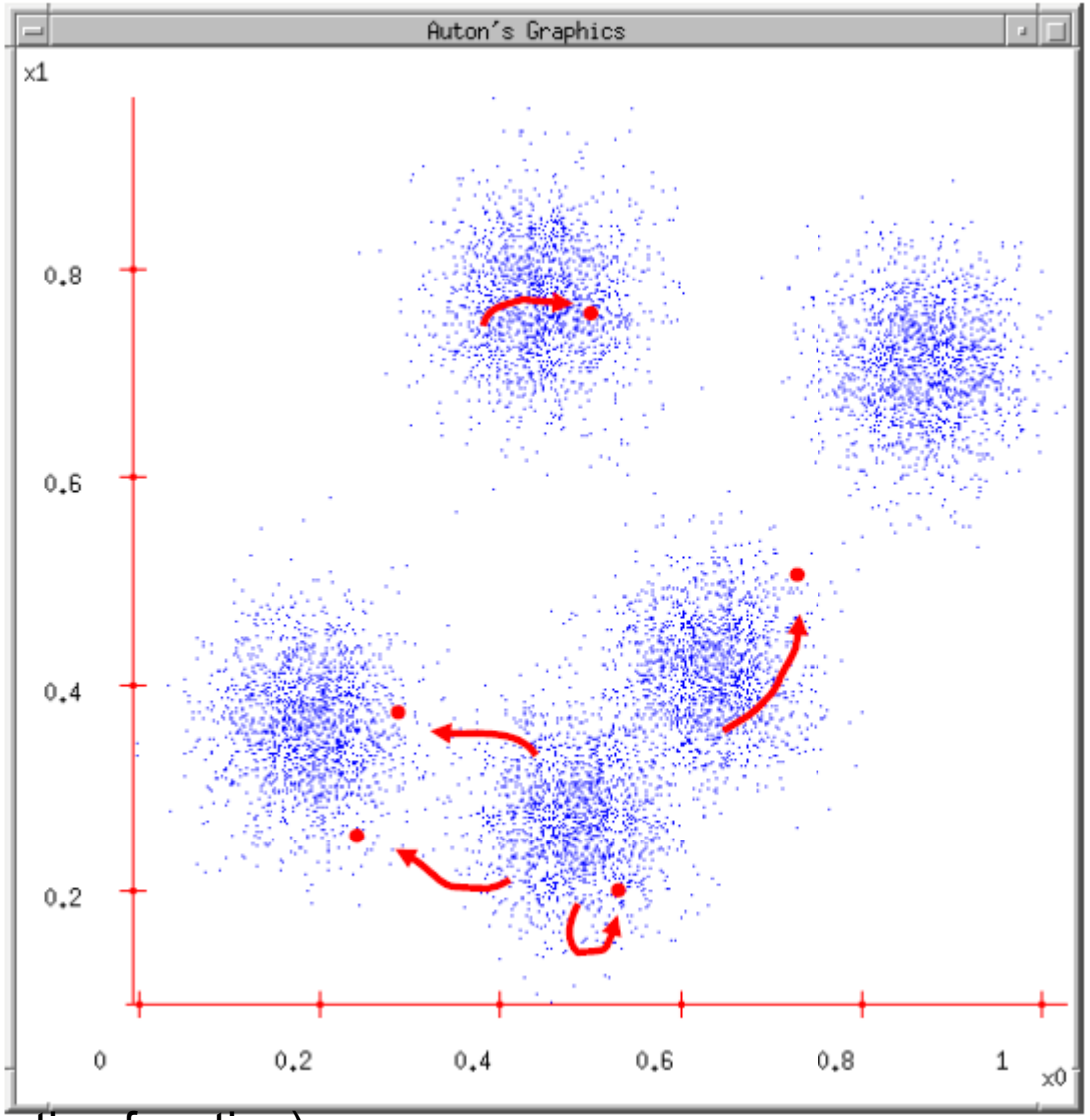


Image from Andrew Moore slides

K-means Animation

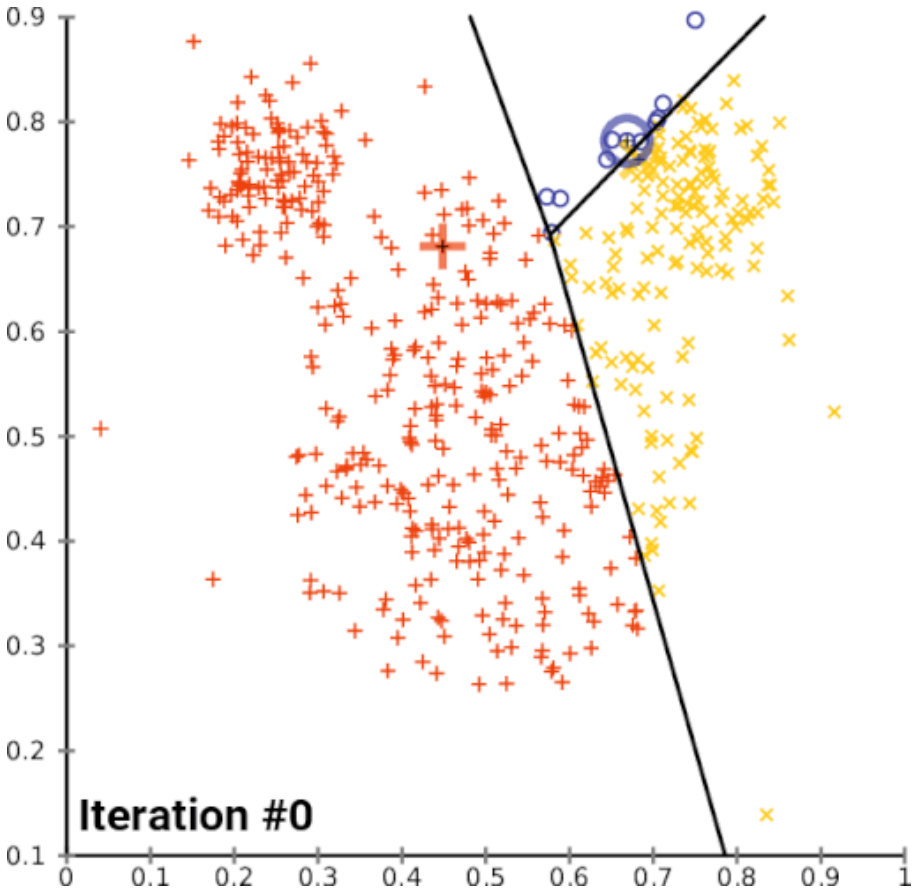


Image source : https://towardsdatascience.com/gaussian-mixture-models-vs-k-means-which-one-to-choose-62f2736025f0

K-means Objective Function

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

μ_i Mean of cluster *i*

S_i Data points assigned to cluster *i*

After initializing clustering centers, iterating between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

Claim that neither step would increase objective function

K-means Objective Function

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

μ_i Mean of cluster *i*

S_i Data points assigned to cluster *i*

After initializing clustering centers, iterating between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

Claim that neither step would increase objective function

For each point $x_j \in S_i$, either

it is closer to its current center *i* → no change

it is closer to another center *m* → objective function improves since

$$||x_j - \mu_m||^2 < ||x_j - \mu_i||^2$$

K-means Objective Function

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

μ_i Mean of cluster *i*

S_i Data points assigned to cluster *i*

After initializing clustering centers, iterating between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

Claim that neither step would increase objective function

K-means Objective Function

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

μ_i Mean of cluster *i*

S_i Data points assigned to cluster *i*

After initializing clustering centers, iterating between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

Claim that neither step would increase objective function

To find the minimum take the partial derivatives and set equal to 0

$$\frac{\partial}{\partial \mu_{i,d}} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 = \frac{\partial}{\partial \mu_{i,d}} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 = -2 \sum_{x_j \in S_i} (x_{j,d} - \mu_{i,d})$$

K-means Objective Function

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

μ_i Mean of cluster *i*

S_i Data points assigned to cluster *i*

After initializing clustering centers, iterating between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

Claim that neither step would increase objective function

To find the minimum take the partial derivatives and set equal to 0

$$\frac{\partial}{\partial \mu_{i,d}} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 = \frac{\partial}{\partial \mu_{i,d}} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 = -2 \sum_{x_j \in S_i} (x_{j,d} - \mu_{i,d})$$

the above is 0 when $\mu_{i,d} = \frac{\sum_{x_j \in S_i} (x_{j,d})}{|S_i|}$ i.e. cluster center

Topics

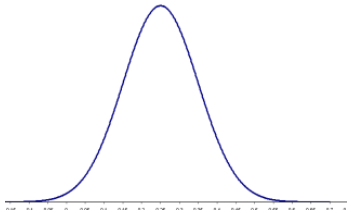
- Hierarchical clustering (review/follow-up)
- K-means clustering (review/follow-up)
- Soft-clustering/Gaussian Mixture Models/Expectation-Maximization algorithm
- Clustering short time series gene expression data

Motivating Example

- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - Genes that are repressed

Motivating Example

- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - Genes that are repressed

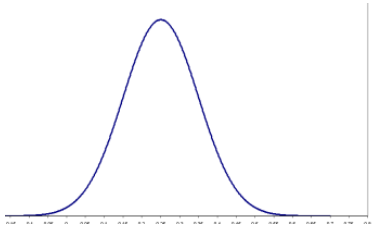


Gaussian (normal distribution)

- Further, suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.

Motivating Example

- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - Genes that are repressed

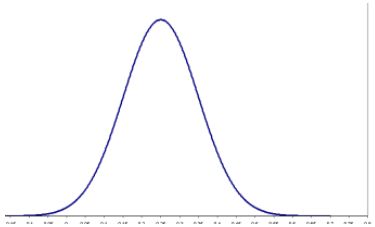


Gaussian (normal distribution)

- Further, suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.

Motivating Example

- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - Genes that are repressed



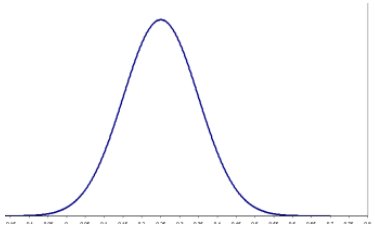
Gaussian (normal distribution)

- Further, suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.

Suppose we simulate 20 data points from each group and observe the following

Motivating Example

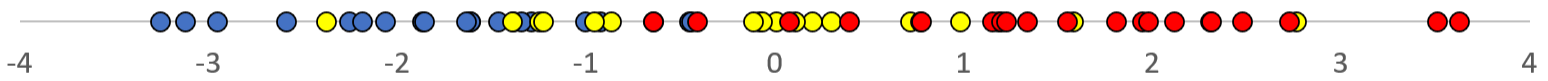
- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - Genes that are repressed



Gaussian (normal distribution)

- Further, suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.

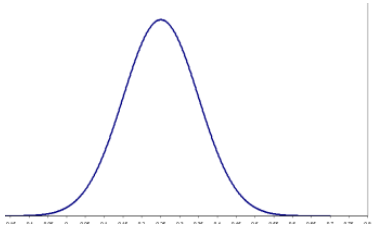
Suppose we simulate 20 data points from each group and observe the following



since we simulated the data we know the true groups.

Motivating Example

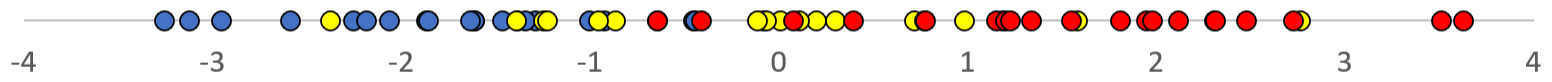
- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - Genes that are repressed



Gaussian (normal distribution)

- Further, suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.

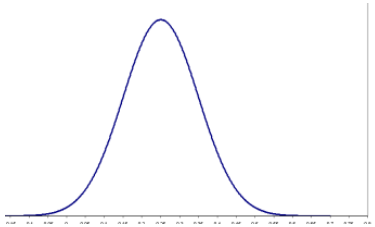
Suppose we simulate 20 data points from each group and observe the following



Question: Is there a partitioning into three intervals that will recover the true groups?

Motivating Example

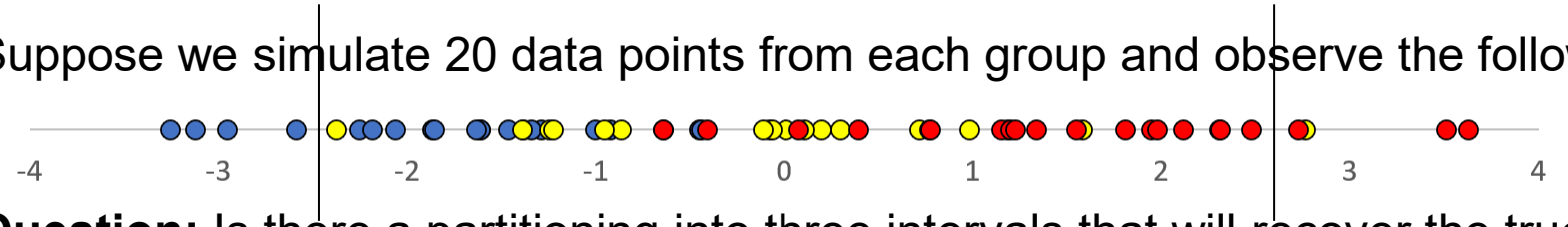
- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - Genes that are repressed



Gaussian (normal distribution)

- Further, suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.

Suppose we simulate 20 data points from each group and observe the following

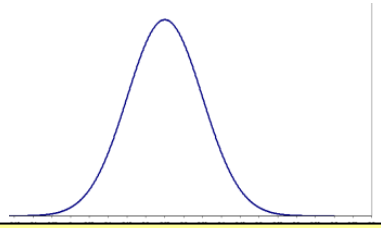


Question: Is there a partitioning into three intervals that will recover the true groups?

There is no partitioning into three intervals that will recover the true groups.

Motivating Example

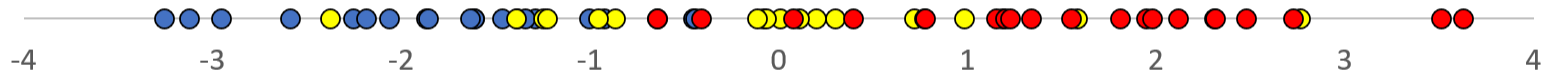
- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - [illegible]



Not valid to assume each point's true group has to correspond to its closest center mean.

- Further, suppose a priori a gene is equally likely to be in any one of the three groups.

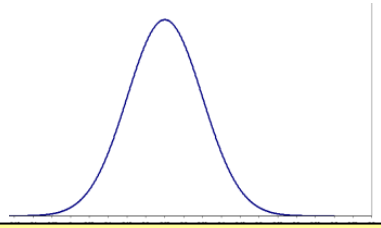
Suppose we simulate 20 data points from each group and observe the following



There is no partitioning into three intervals that will recover the true groups, which we know in this case since we simulated.

Motivating Example

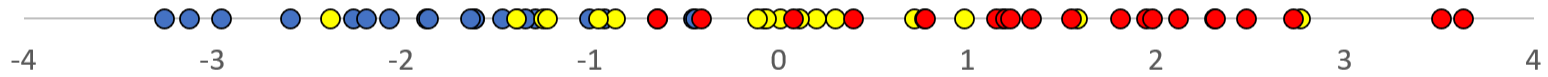
- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected
  - [illegible]



Not valid to assume each point's true group has to correspond to its closest center mean.
Question: What can we do instead?

- Further, suppose a priori a gene is equally likely to be in any one of the three groups.

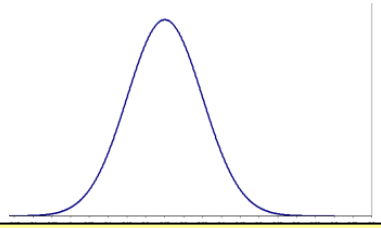
Suppose we simulate 20 data points from each group and observe the following



There is no partitioning into three intervals that will recover the true groups, which we know in this case since we simulated.

Motivating Example

- Suppose in a response to a biological stimulus we have three groups of genes:
  - Genes that are activated
  - Genes that are unaffected



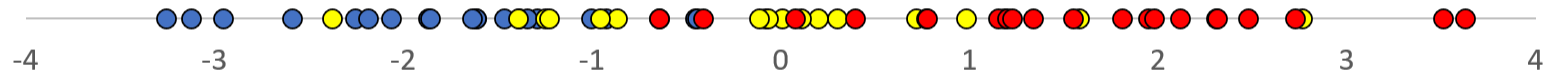
Not valid to assume each point's true group has to correspond to its closest center mean.

Question: What can we do instead?

Use probabilistic soft-assignments that reflect the uncertainty

- Further, suppose a priori a gene is equally likely to be in any one of the three groups.

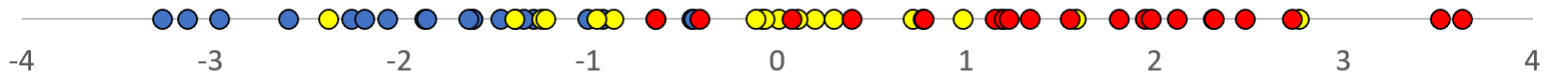
Suppose we simulate 20 data points from each group and observe the following



There is no partitioning into three intervals that will recover the true groups, which we know in this case since we simulated.

Estimating assignment probabilities

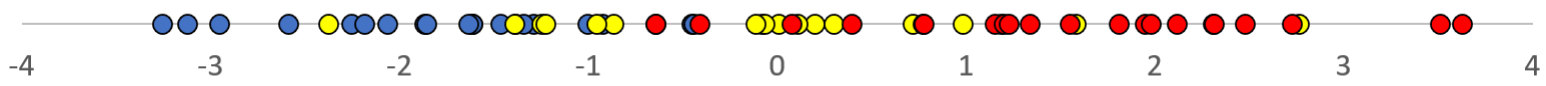
- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.
Question: How do we compute the probability it is in the repressed group?

Estimating assignment probabilities

- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.

Question: How do we compute the probability it is in the repressed group?

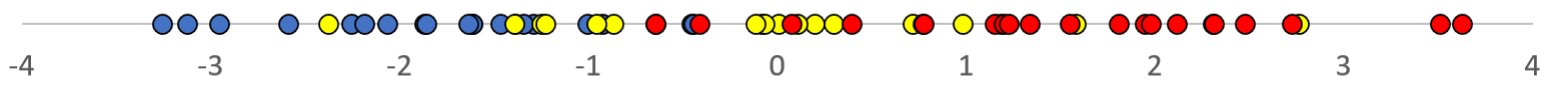
$$\frac{f(x{=}{-}1|\mu=-1.5,\sigma^2=1)}{f(x{=}{-}1|\mu=-1.5,\sigma^2=1)+f(x{=}{-}1|\mu=0,\sigma^2=1)+f(x{=}{-}1|\mu=1.5,\sigma^2=1)}$$

where $f(x|\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$ gaussian density function

since needs to be in one of three groups and apriori each equally likely.

Estimating assignment probabilities

- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



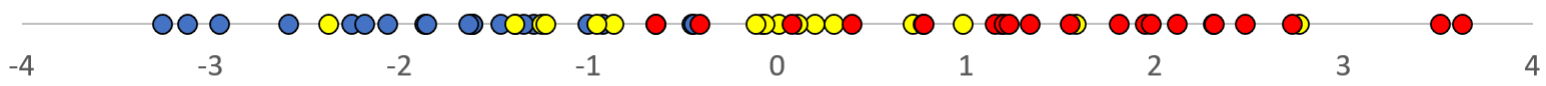
Say a gene has expression value -1.
Question: How do we compute the probability it is in the repressed group?

$f(x{=}-1|\mu = -1.5,\sigma^2 = 1)$ = 0.3521 (repressed)
$f(x{=}-1|\mu = 0, \sigma^2 = 1)$ = 0.2420 (unaffected)
$f(x{=}-1|\mu = 1.5,\sigma^2 = 1)$ = 0.0175 (activated)

where $f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$

Estimating assignment probabilities

- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.

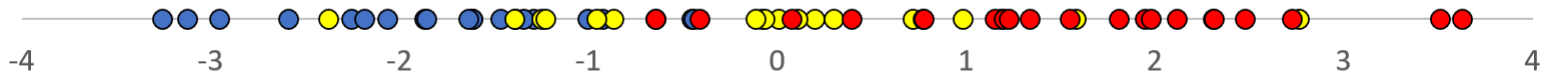
Question: How do we compute the probability it is in the repressed group?

$$\frac{f(x{=}{-}1|\mu=-1.5,\sigma^2=1)}{f(x{=}{-}1|\mu=-1.5,\sigma^2=1)+f(x{=}{-}1|\mu=0,\sigma^2=1)+f(x{=}{-}1|\mu=1.5,\sigma^2=1)}$$

$$=\frac{0.3521}{0.3521+0.2420+0.0175}=0.576$$

Estimating assignment probabilities

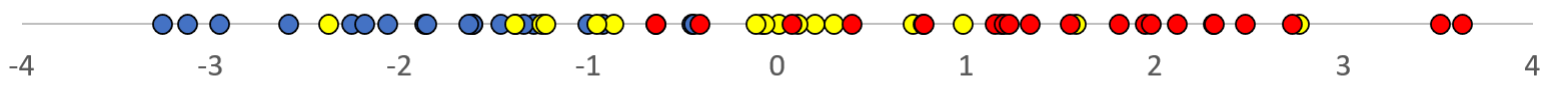
- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.
Question: How do we compute the probability it is in the unaffected group?

Estimating assignment probabilities

- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.

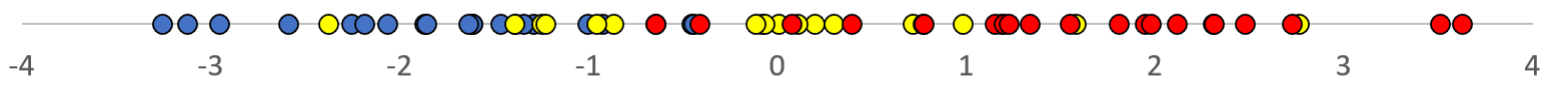
Question: How do we compute the probability it is in the unaffected group?

$$\frac{f(x{=}{-}1|\mu = 0,\sigma^2 = 1)}{f(x{=}{-}1|\mu = -1.5,\sigma^2 = 1){+}f(x{=}{-}1|\mu = 0,\sigma^2 = 1){+}f(x{=}{-}1|\mu = 1.5,\sigma^2 = 1)}$$

$$=\frac{0.2420}{0.3521+0.2420+0.0175} = 0.396$$

Estimating assignment probabilities

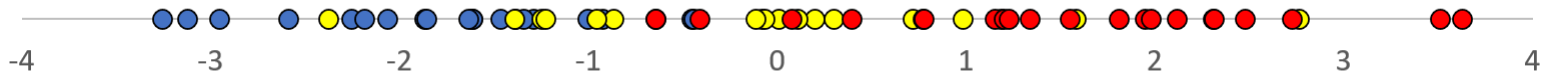
- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.
Question: How do we compute the probability it is in the activated group?

Estimating assignment probabilities

- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.

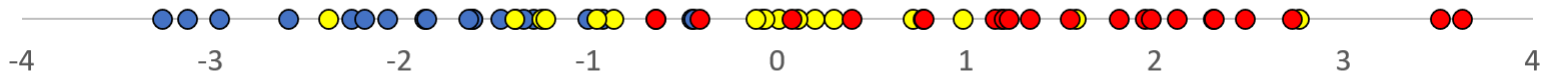
Question: How do we compute the probability it is in the activated group?

$$\frac{f(x{=}{-}1|\mu = 1.5,\sigma^2 = 1)}{f(x{=}{-}1|\mu = -1.5,\sigma^2 = 1){+}f(x{=}{-}1|\mu = 0,\sigma^2 = 1){+}f(x{=}{-}1|\mu = 1.5,\sigma^2 = 1)}$$

$$=\frac{0.0175}{0.3521+0.2420+0.0175} = 0.029$$

Estimating assignment probabilities

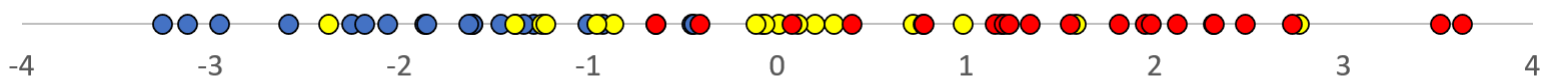
- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.
Question: How would the probability of repressed change if we assume apriori a gene had one-half probability of being in the unaffected group and one-quarter probability each for the repressed and activated?

Estimating assignment probabilities

- Suppose within each group we know the gene expression follows a gaussian (normal) distribution with variance one, and means -1.5, 0, 1.5 for repressed, unaffected, and activated groups respectively.
- Further, suppose a priori a gene is equally likely to be in any one of the three groups.



Say a gene has expression value -1.

Question: How would the probability of repressed change if we assume apriori a gene had one-half probability of being in the unaffected group and one-quarter probability each for the repressed and activated?

$$\frac{0.25 \times f(x{=}-1|\mu = -1.5,\sigma^2 = 1)}{0.25 \times f(x{=}-1|\mu = -1.5,\sigma^2 = 1){+}0.5 \times f(x{=}-1|\mu = 0,\sigma^2 = 1){+}0.25 \times f(x{=}-1|\mu = 1.5,\sigma^2 = 1)}$$

$$= \frac{0.25 \times 0.3521}{0.25 \times 0.3521 + 0.5 \times 0.2420 + 0.25 \times 0.0175} = 0.412$$

Suppose we did not know gaussian mean parameters

- Suppose we only have unlabeled points and do not know the gaussian means
- We will assume we know there are three underlying gaussian distributions with variance 1 and a priori a gene is equally likely to belong to any of the three
- **Question**: How should we estimate mean parameters of the gaussian distribution and soft-assignment probabilities? What should our objective function be?

Suppose we did not know gaussian mean parameters

- Suppose we only have unlabeled points and do not know the gaussian means
- We will assume we know there are three underlying gaussian distributions with variance 1 and a priori a gene is equally likely to belong to any of the three
- **Question**: How should we estimate mean parameters of the gaussian distribution and soft-assignment probabilities? What should our objective function be?

$$\arg\max \sum_{j=1}^{N} log \sum_{i=1}^{k} w_{ij}\, f(x_j|\mu_i, 1)$$

Here *k=3* is the number of gaussians/clusters and *N* is the number of data-points.

w_{ij} is the assignment probability of data point *j* to gaussian/cluster *i*

μ_i is the mean of gaussian/cluster *i*

$f(x_j|\mu_i, 1)$ – is gaussian distribution density value at x_j for a gaussian with mean μ_i and variance 1

Optimizing objective function

$$\arg\max \sum_{j=1}^{N} log \sum_{i=1}^{k} w_{ij}\, f(x_j|\mu_i, 1)$$

Here *k*=3 is the number of gaussians/clusters and *N* is the number of data-points.

w_{ij} is the assignment probability of data point *j* to gaussian/cluster *i*

μ_i is the mean of gaussian/cluster *i*

$f(x_j|\mu_i, 1)$ – is gaussian distribution density value at x_j for a gaussian with mean μ_i and variance 1

Question: How should we try to optimize this objective function?

Optimizing objective function

$$\arg\max \sum_{j=1}^{N} log \sum_{i=1}^{k} w_{ij}\, f(x_j|\mu_i, 1)$$

Here *k*=3 is the number of gaussians/clusters and *N* is the number of data-points.

w_{ij} is the assignment probability of data point *j* to gaussian/cluster *i*

μ_i is the mean of gaussian/cluster *I*

$f(x_j|\mu_i, 1)$ – is gaussian distribution density value at x_j for a gaussian with mean μ_i and variance 1

Question: How should we try to optimize this objective function?

Expectation-maximization (EM) algorithm. Finds local maximum.

Expectation-maximization algorithm

In this setting can be thought of a “soft” version of the k-means algorithm

In the k-means after initializing clustering centers, it iterates between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

Expectation-maximization algorithm

In this setting can be thought of a “soft” version of the k-means algorithm

In the k-means after initializing clustering centers, it iterates between two steps:

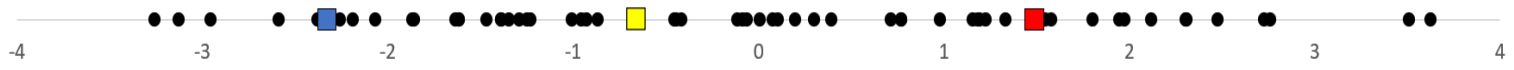
- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

In the EM algorithm, after initializing cluster centers, it iterates between two analogous steps:

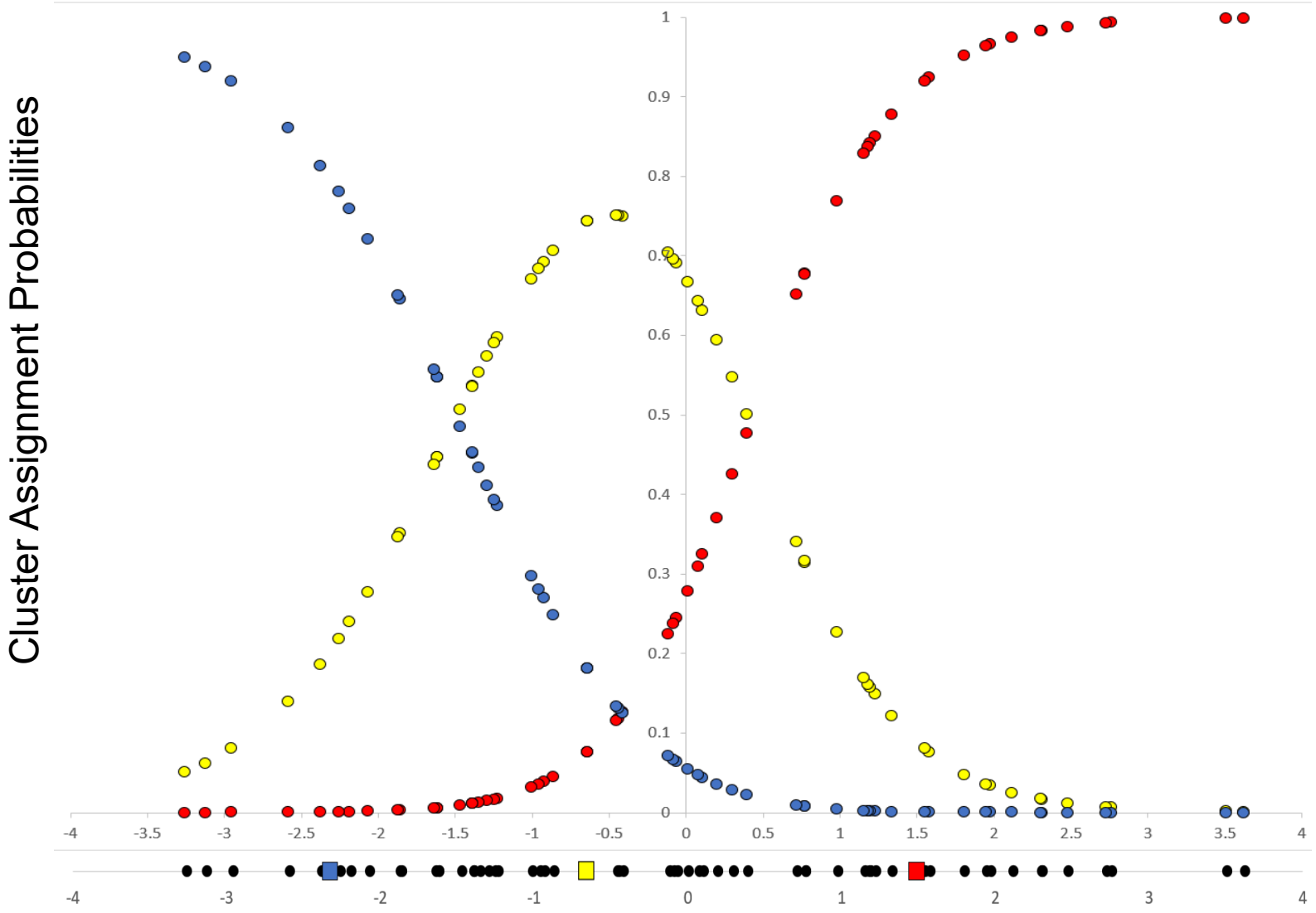
- Expectation (E)-step: Re-compute assignment probabilities of each data point to each cluster mean
- Maximization (M)-step: Recompute the weighted cluster mean based on all assignment probabilities

If we know the cluster means, straightforward to compute assignment probabilities. This we saw how to do earlier.

Random Initialization



E-step


Cluster Assignment Probabilities
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4

Expectation-maximization algorithm

In this setting can be thought of a “soft” version of the k-means algorithm

In the k-means after initializing clustering centers, it iterates between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

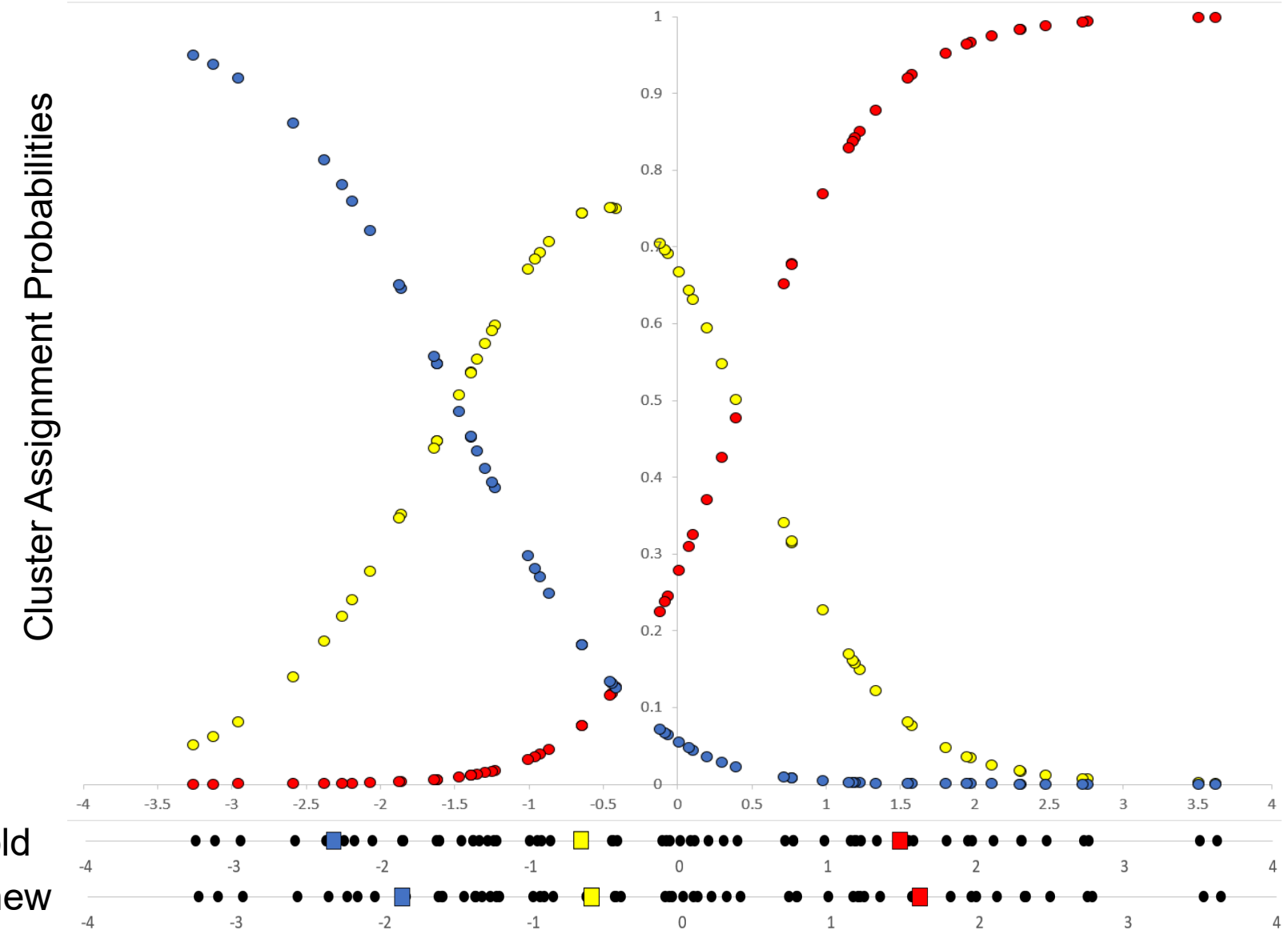
In the EM algorithm, after initializing cluster centers, it iterates between two analogous steps:

- Expectation (E)-step: Re-compute assignment probabilities of each data point to each cluster mean
- Maximization (M)-step: Recompute the weighted cluster mean based on all assignment probabilities

If we know the assignment probabilities, straightforward to recompute the weighted means

$$\mu_i = \frac{\sum_{j=1}^{N} w_{ij}\, x_j}{\sum_{j=1}^{N} w_{ij}}$$

M-step


Cluster Assignment Probabilities
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
old
new

Expectation-maximization algorithm

In this setting can be thought of a “soft” version of the k-means algorithm

In the k-means after initializing clustering centers, it iterates between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

In the EM algorithm, after initializing cluster centers, it iterates between two analogous steps:

- Expectation (E)-step: Re-compute assignment probabilities of each data point to each cluster mean
- Maximization (M)-step: Recompute the weighted cluster mean based on all assignment probabilities

If we know the assignment probabilities, straightforward to recompute the weighted means

$$\mu_i = \frac{\sum_{j=1}^{N} w_{ij}\, x_j}{\sum_{j=1}^{N} w_{ij}}$$

If re-estimating cluster priors

$$\frac{\sum_{j=1}^{N} w_{ij}}{N}$$

Expectation-maximization algorithm

In this setting can be thought of a “soft” version of the k-means algorithm

In the k-means after initializing clustering centers, it iterates between two steps:

- Re-assign data points to its closest cluster mean
- Recompute the cluster mean of the points assigned to each cluster

In the EM algorithm, after initializing cluster centers, it iterates between two analogous steps:

- Expectation (E)-step: Re-compute assignment probabilities of each data point to each cluster mean
- Maximization (M)-step: Recompute the weighted cluster mean based on all assignment probabilities

If we know the assignment probabilities, straightforward to recompute the weighted means

$$\mu_i = \frac{\sum_{j=1}^{N} w_{ij}\, x_j}{\sum_{j=1}^{N} w_{ij}}$$

If re-estimating cluster priors

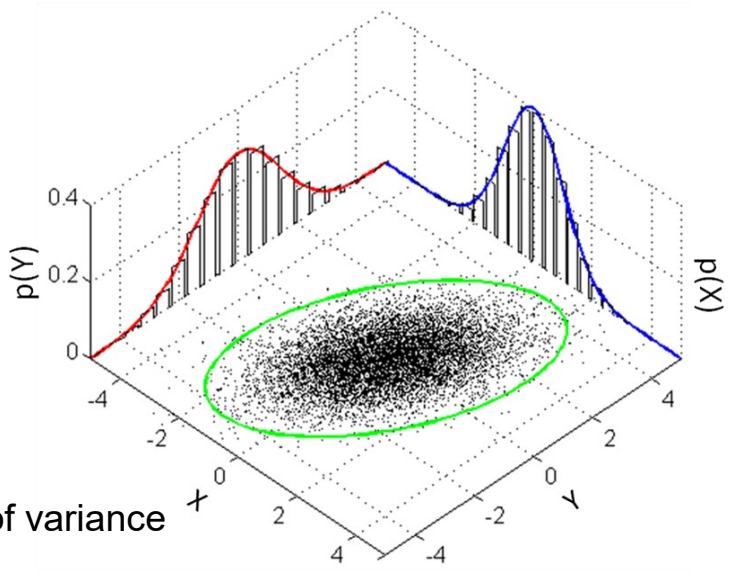
$$\frac{\sum_{j=1}^{N} w_{ij}}{N}$$

Can also re-estimate gaussian variance parameters

Approach generalizes to multiple-dimensions

Multivariate gaussian distributions

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$



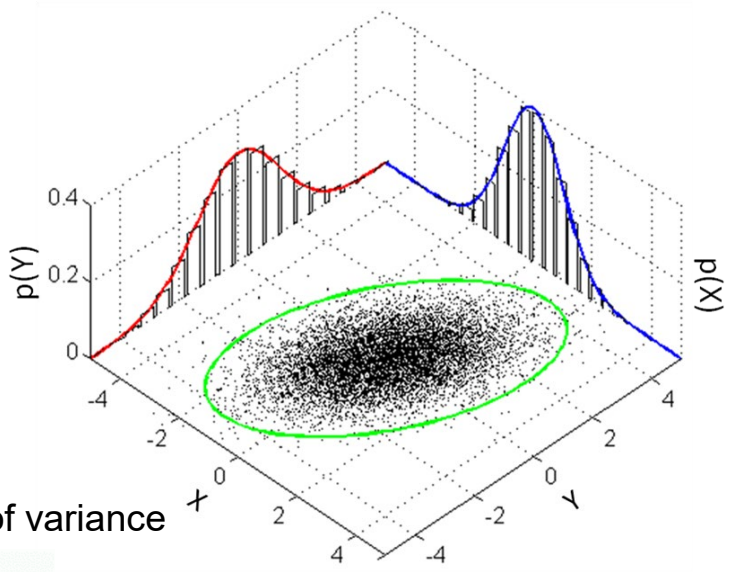
Σ - Covariance matrix – multi-dimensional generalization of variance

Image source: wikipedia

Approach generalizes to multiple-dimensions

Multivariate gaussian distributions

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$



Σ - Covariance matrix – multi-dimensional generalization of variance

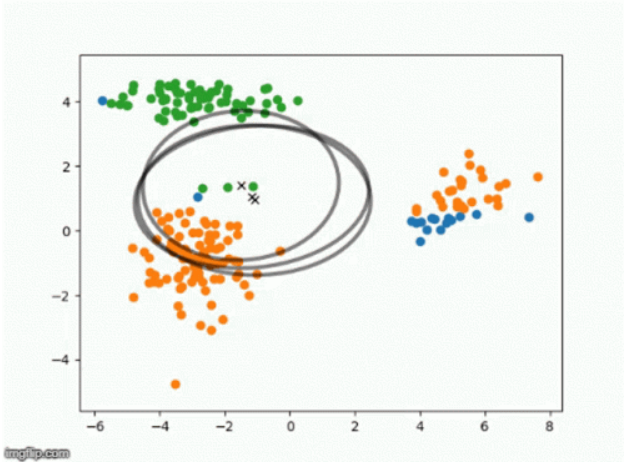


Image source: wikipedia

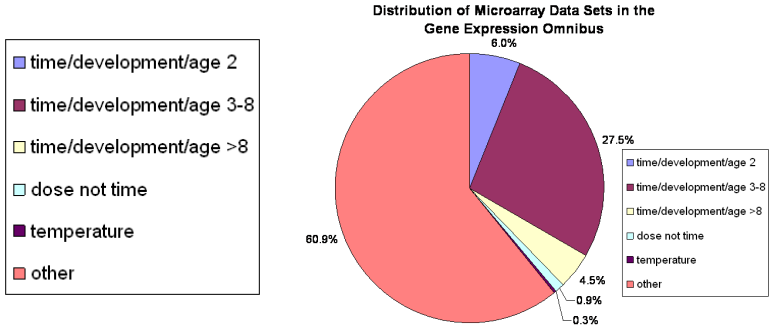
Visualization of EM https://towardsdatascience.com/gaussian-mixture-models-vs-k-means-which-one-to-choose-62f2736025f0

Topics

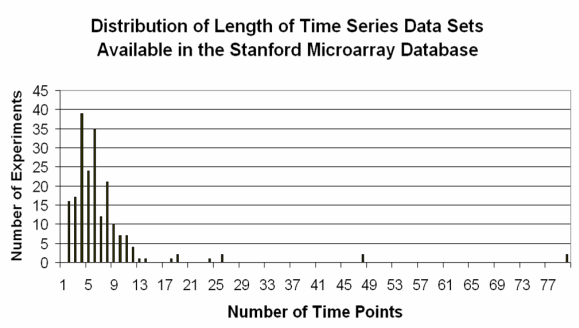
- Hierarchical clustering (review/follow-up)
- K-means clustering (review/follow-up)
- Soft-clustering/Gaussian Mixture Models/Expectation-Maximization algorithm
- Clustering short time series gene expression data

Importance of Clustering Short Time-series Data

- Biological processes occur over time (e.g. stress response, immune response, development) and frequently studied with gene expression experiments
- Most time series gene expression data sets are short (3-8 time points)



Ernst and Bar-Joseph *BMC Bioinformatics*, 2006

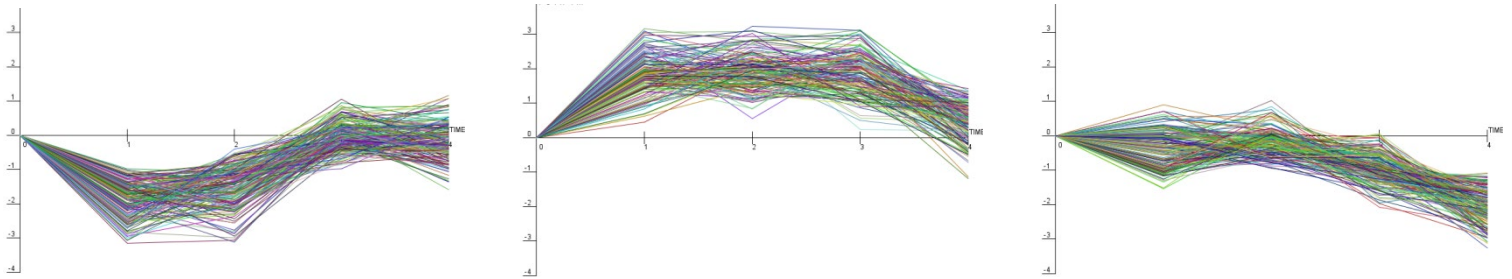


Ernst et al. *Bioinformatics*, 2005

- Genes with similar expression patterns over time are often involved in the same biological process or are co-regulated

Limitations of Standard Clustering Methods for Time Series Data

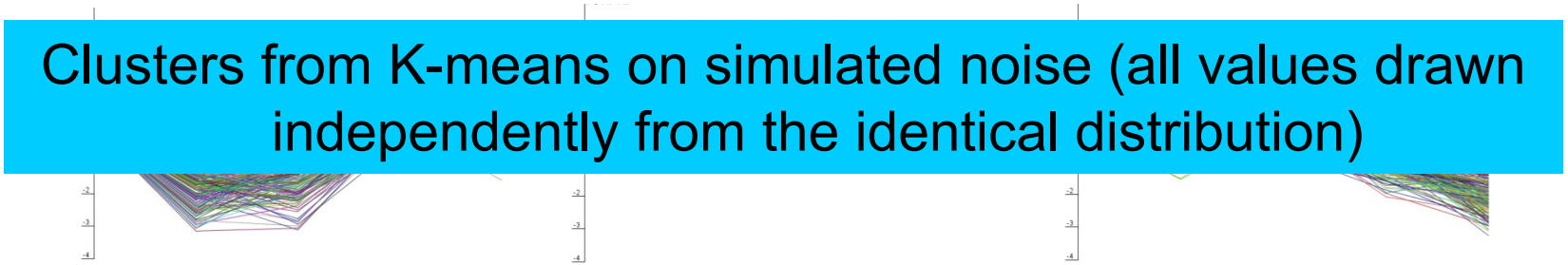
- Having few time points can pose a challenge for traditional time series models (e.g. autoregressive equations)
- Commonly used methods such as k-means and hierarchical clustering do not use the temporal ordering of experiments
- Thousands of genes and few time points ➡ many patterns by random chance
  - Standard clustering methods do not differentiate between real and random patterns



Ernst et al. ISMB/Bioinformatics, 2005

Limitations of Standard Clustering Methods for Time Series Data

- Having few time points can pose a challenge for traditional time series models (e.g. autoregressive equations)
- Commonly used methods such as k-means and hierarchical clustering do not use the temporal ordering of experiments
- Thousands of genes and few time points ➡ many patterns by random chance
  - Standard clustering methods do not differentiate between real and random patterns



Clusters from K-means on simulated noise (all values drawn independently from the identical distribution)

Ernst et al. ISMB/Bioinformatics, 2005

Method Overview

Approach: Determine temporal patterns with significantly more genes than expected compared to a random ordering of time

STEP 1: Define temporal profiles independent of data

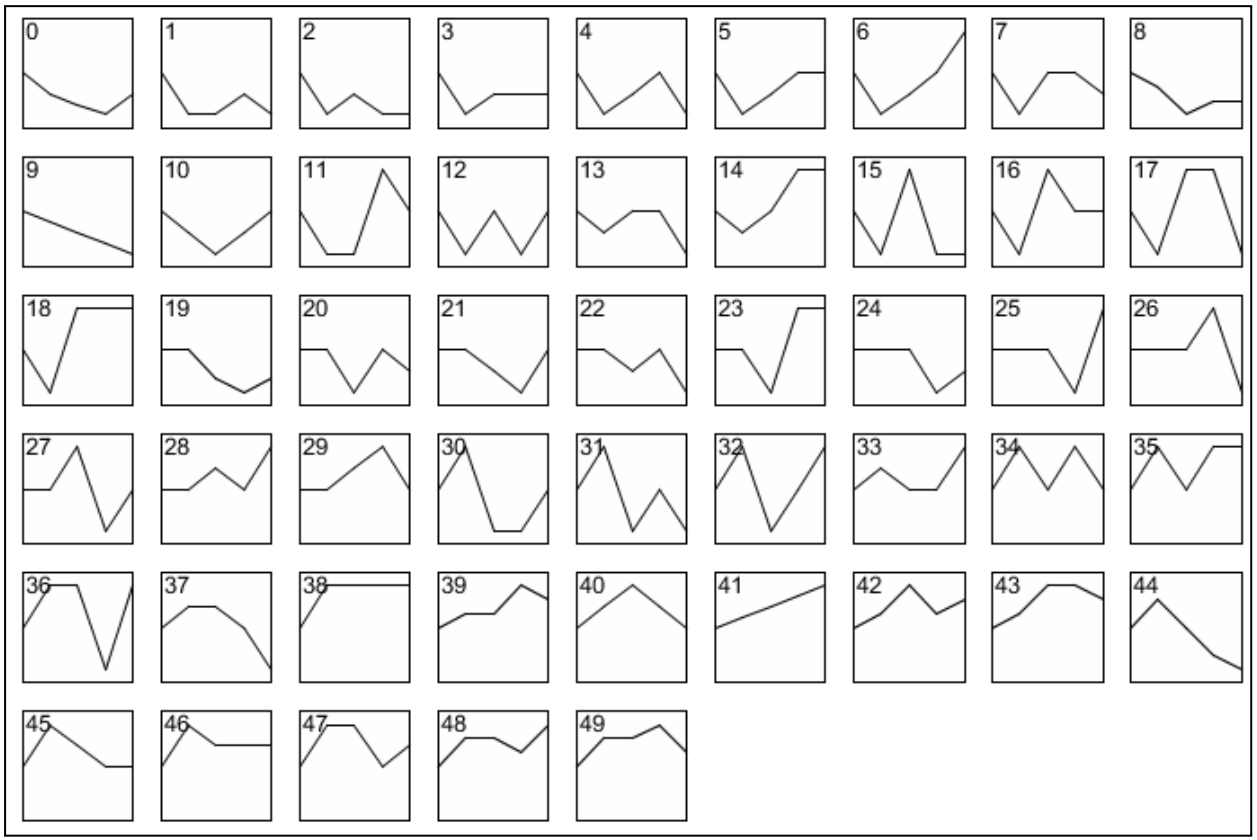
STEP 2: Assign genes to most closely matching profile

STEP 3: Compute expected number of genes per profile

STEP 4: Determine statistically significant profiles

Ernst et al. ISMB/Bioinformatics, 2005

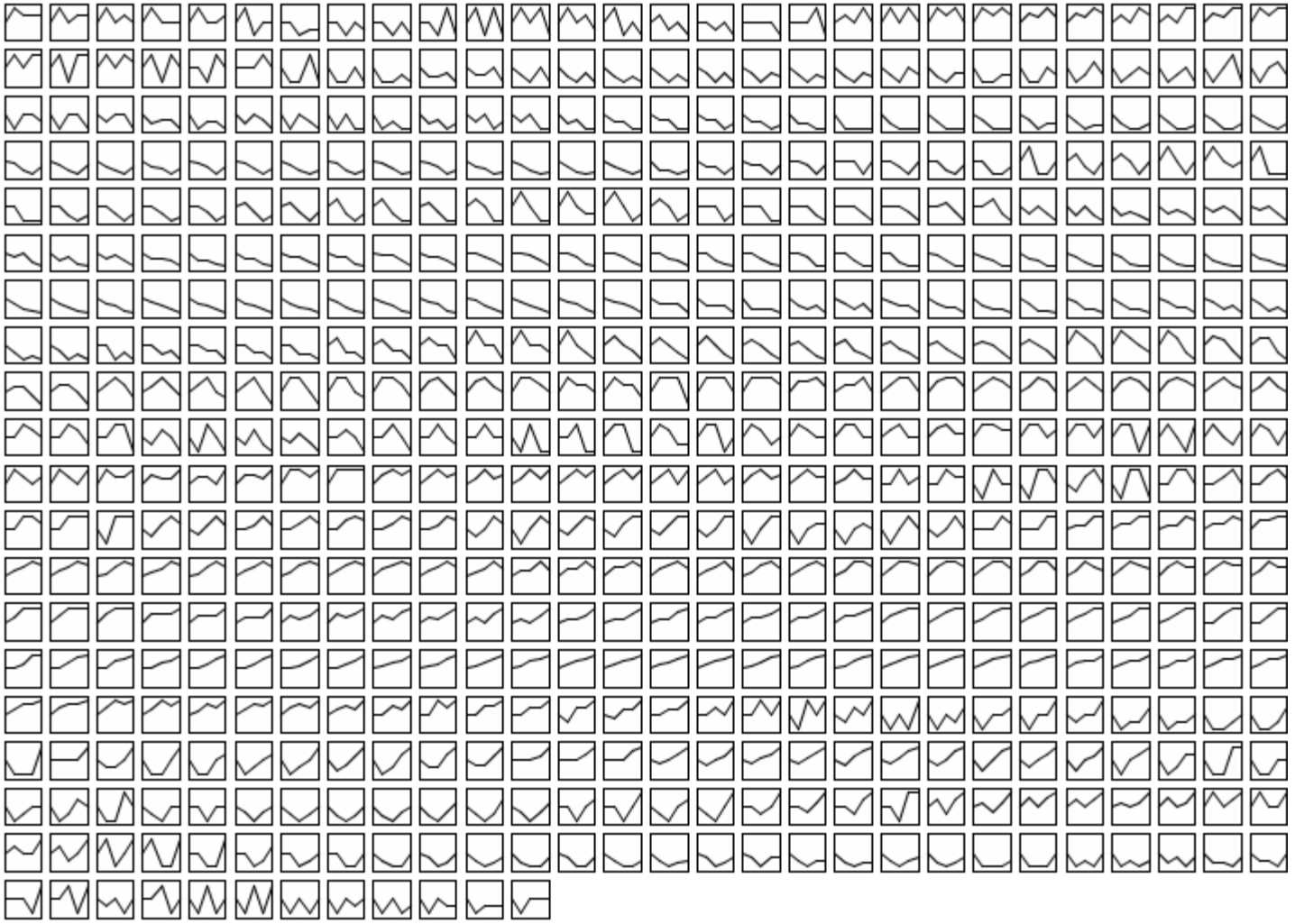
STEP 1: Define a set of distinct and representative model temporal profiles **independent** of the data.



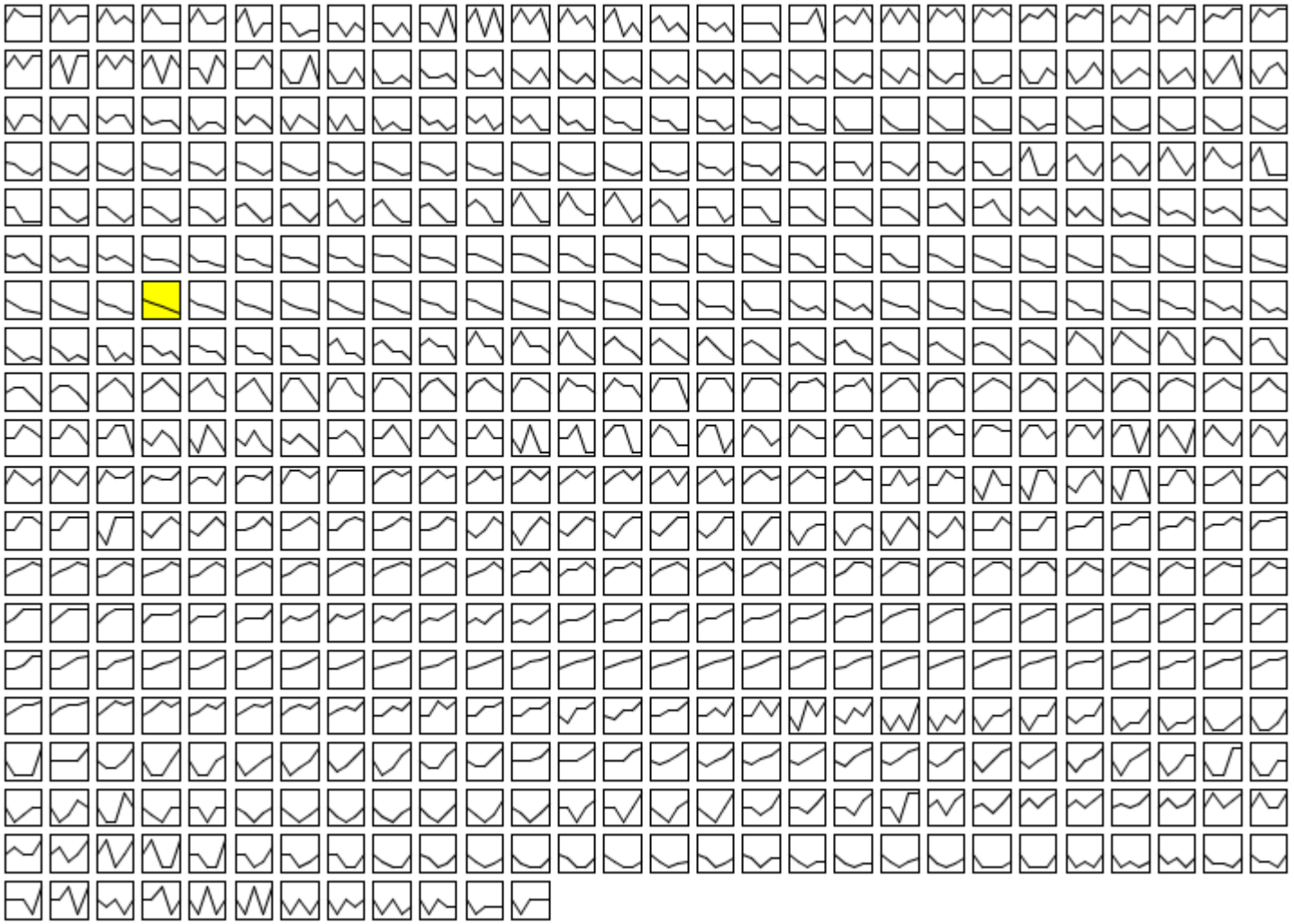
STEP 1. Define temporal profiles independent of data

Ernst et al. ISMB/Bioinformatics, 2005

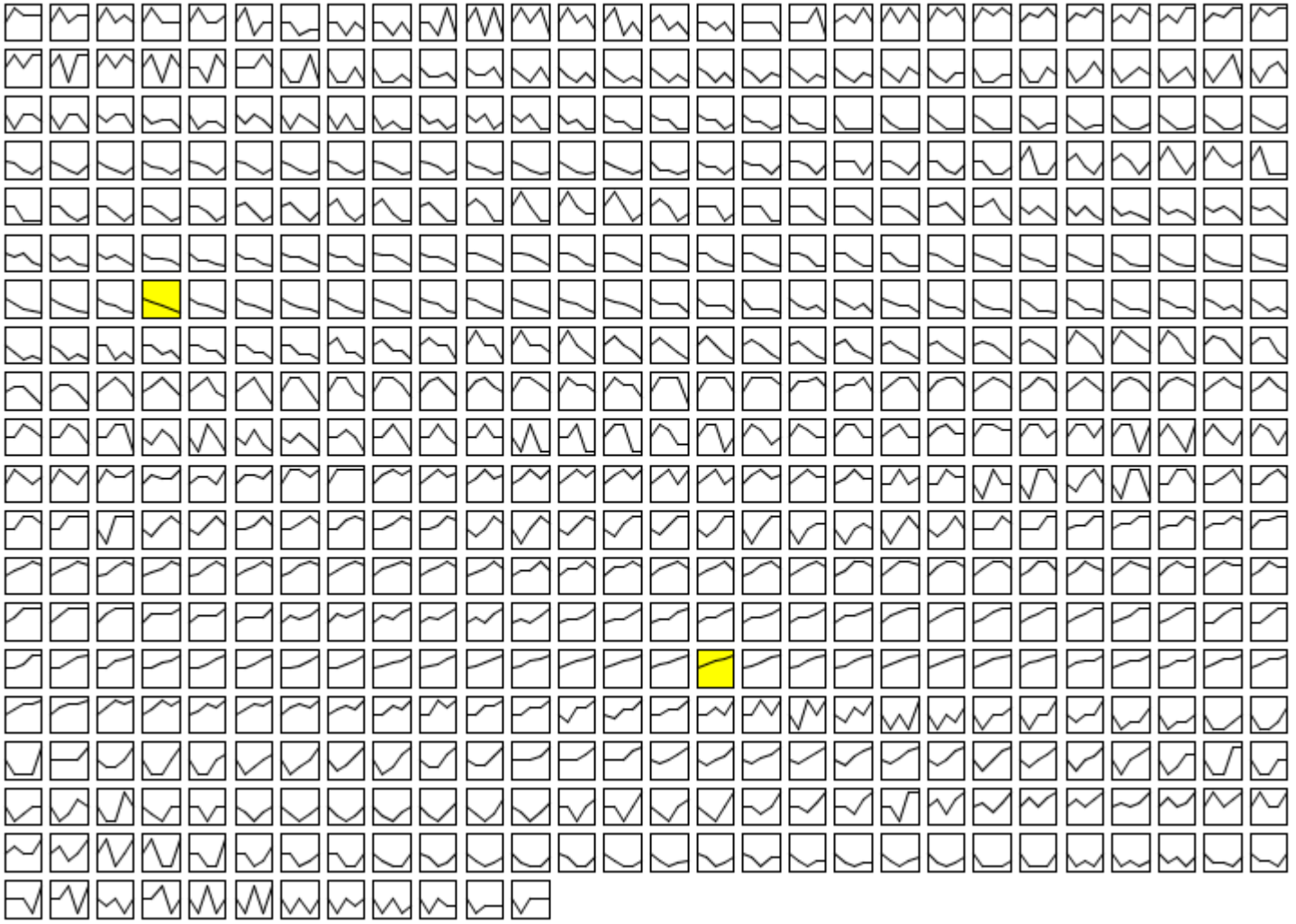
Start with all temporal profile shapes with at most *c* unit change between time points (here *c*=2)



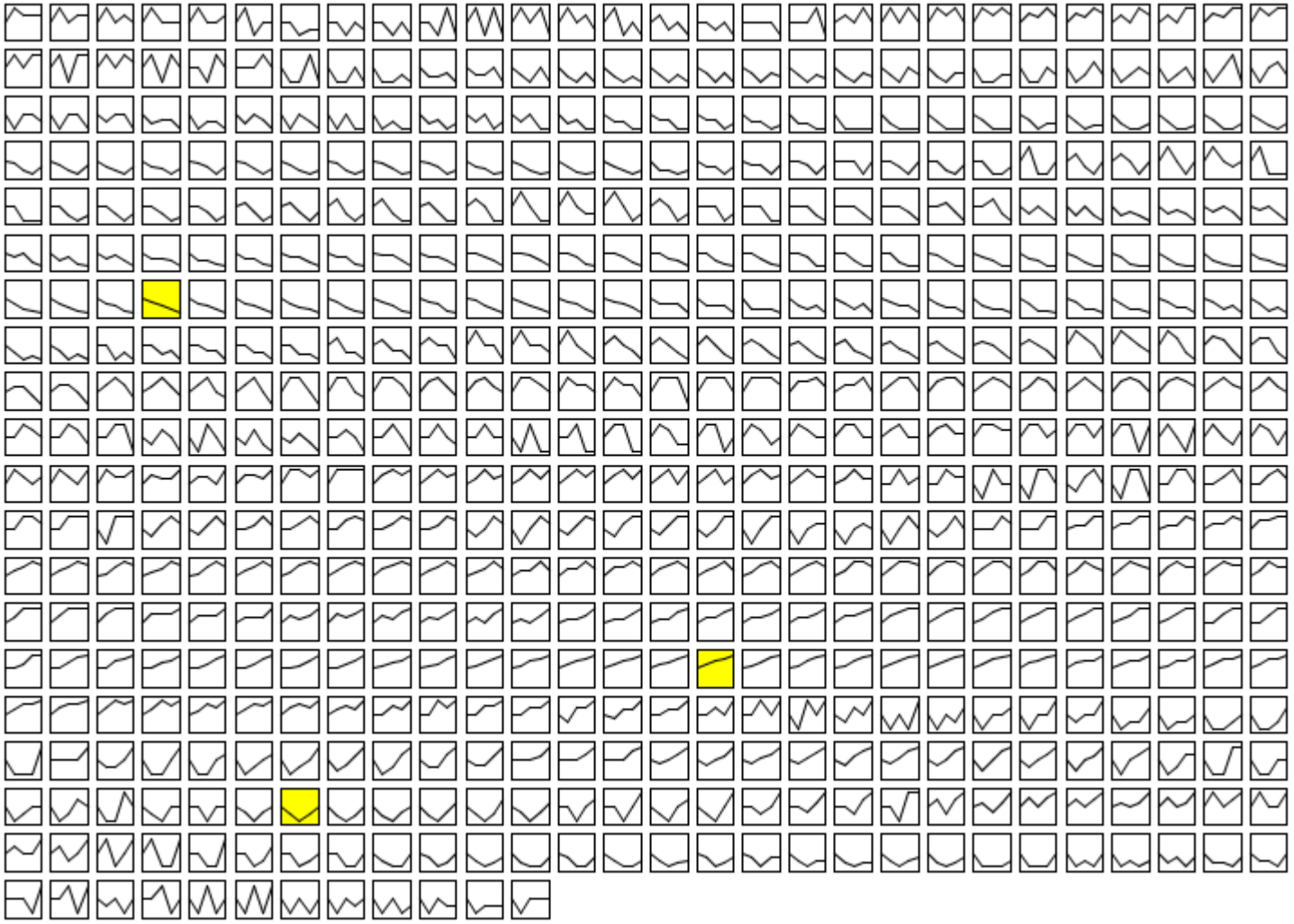
Greedily select a set of *k* profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1-\text{correlation coefficient})}$



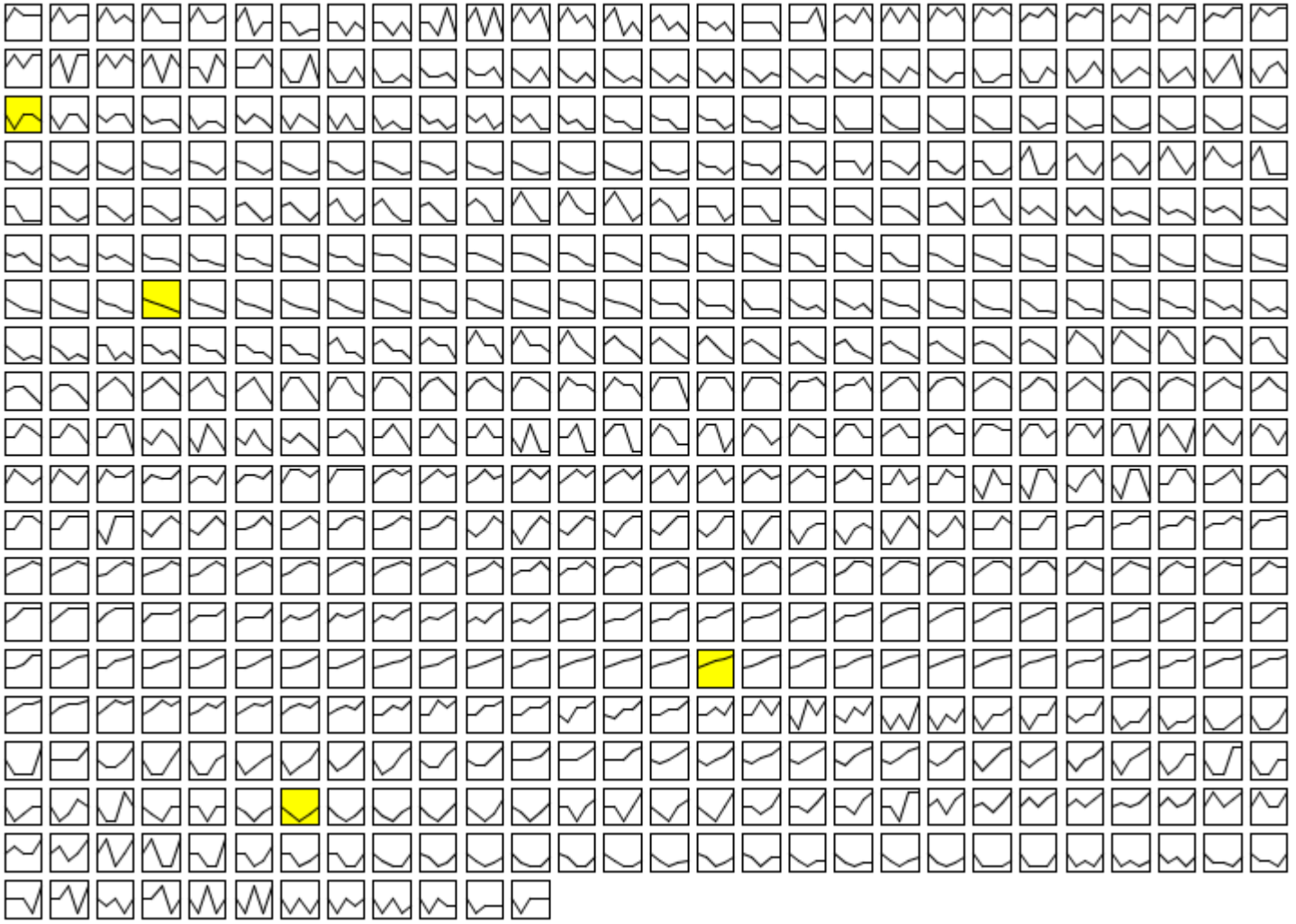
Greedily select a set of *k* profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1-\text{correlation coefficient})}$



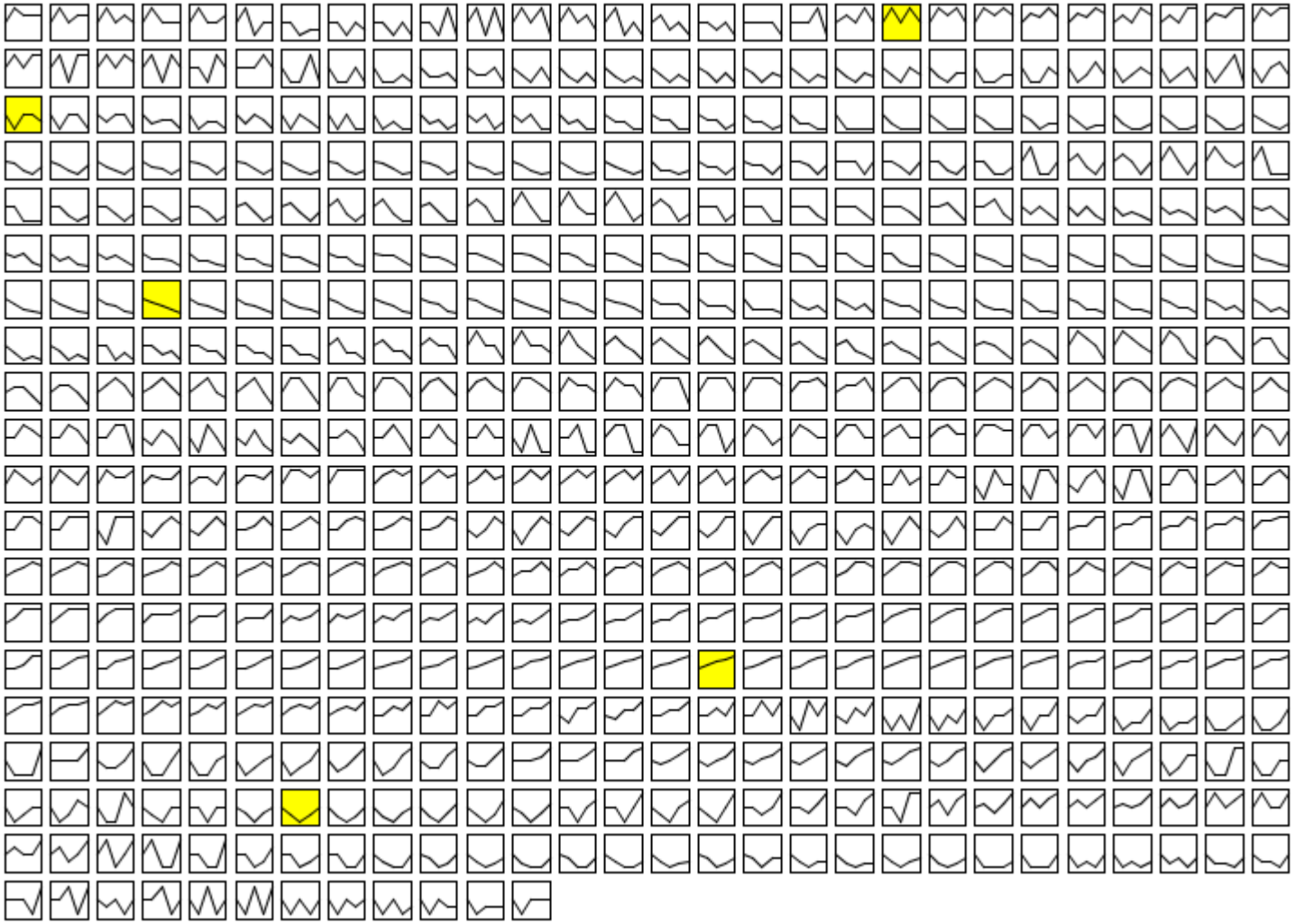
Greedily select a set of *k* profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1-\text{correlation coefficient})}$



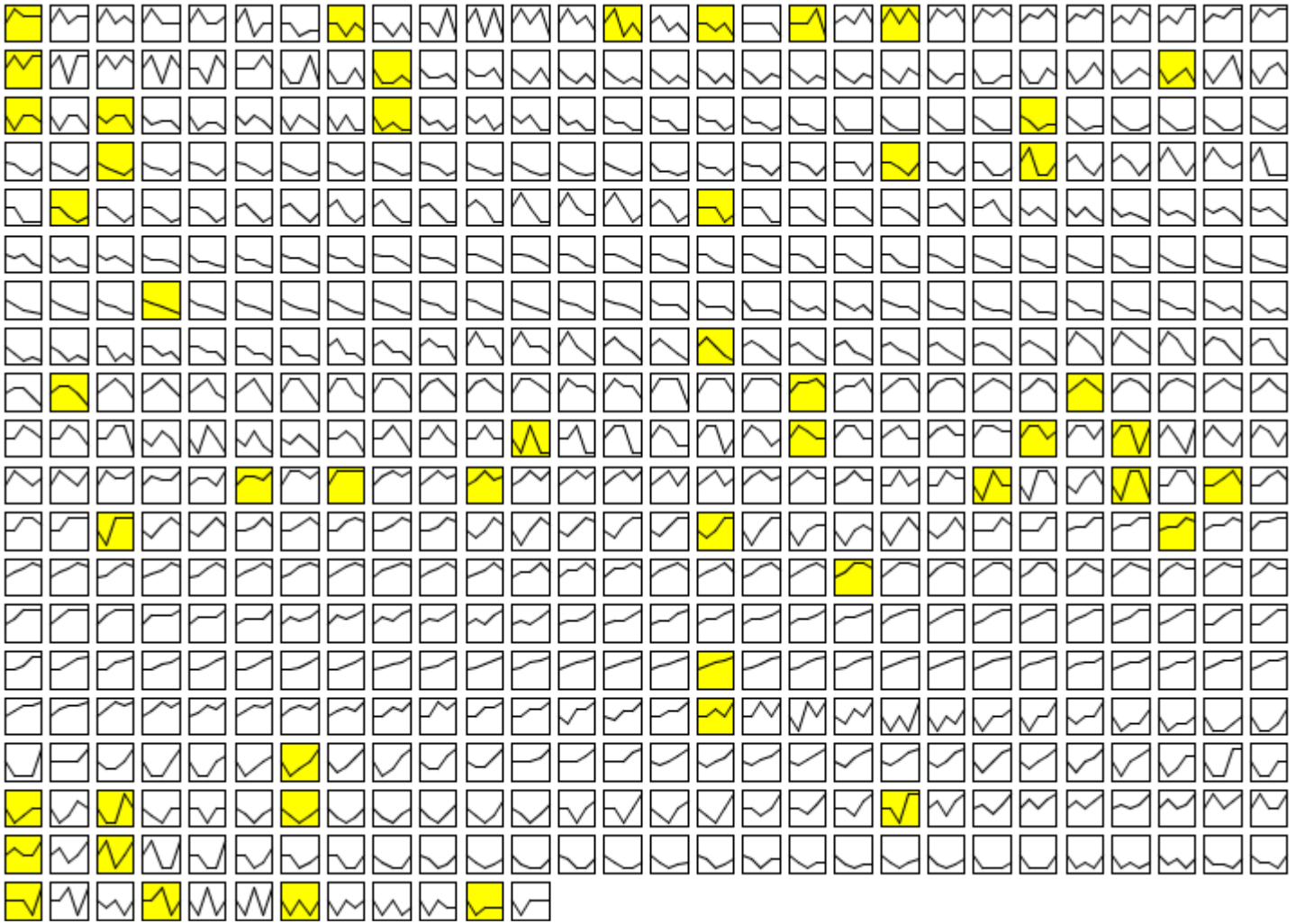
Greedily select a set of *k* profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1-\text{correlation coefficient})}$



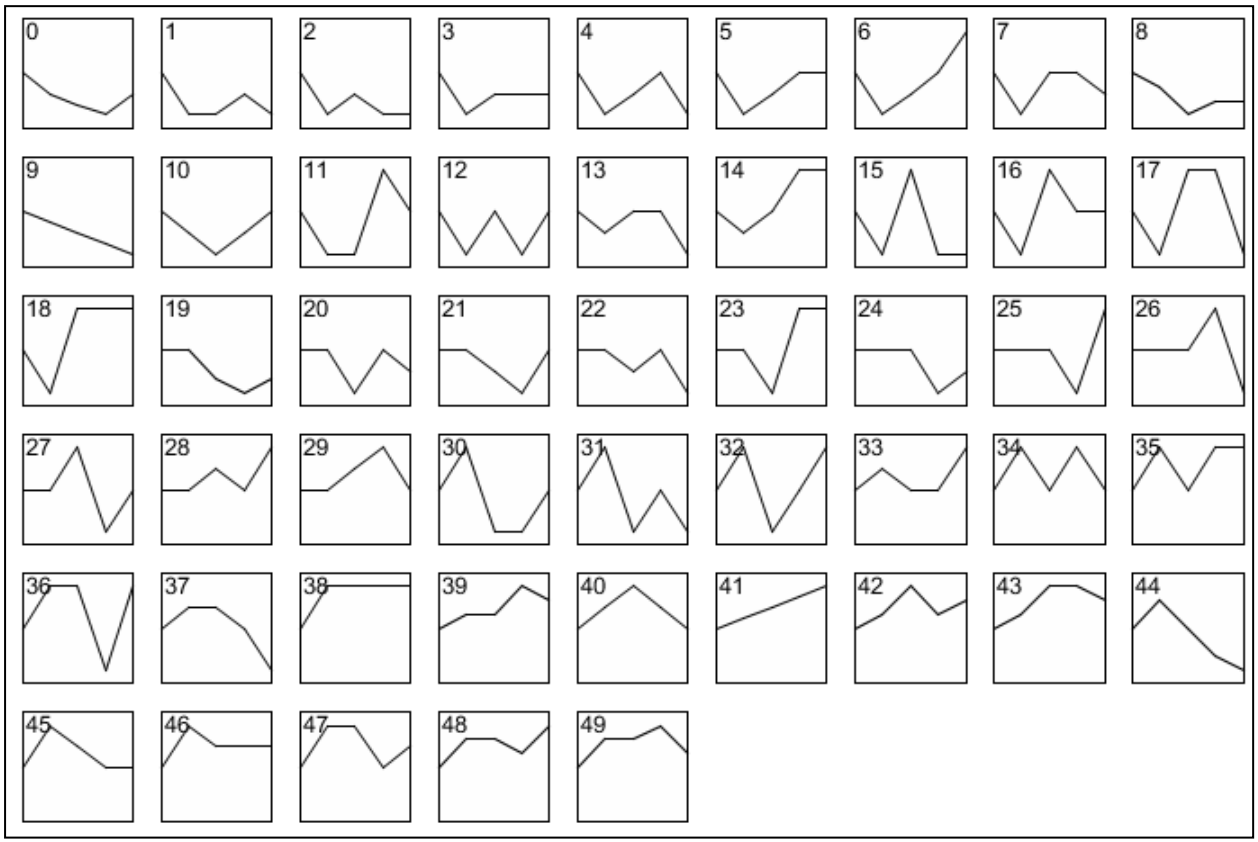
Greedily select a set of *k* profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1-\text{correlation coefficient})}$



Greedily select a set of *k* profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1 - \text{correlation coefficient})}$



STEP 2: Filter flat genes, and then assign the remaining genes to the most closely matching model profile based on the correlation coefficient.

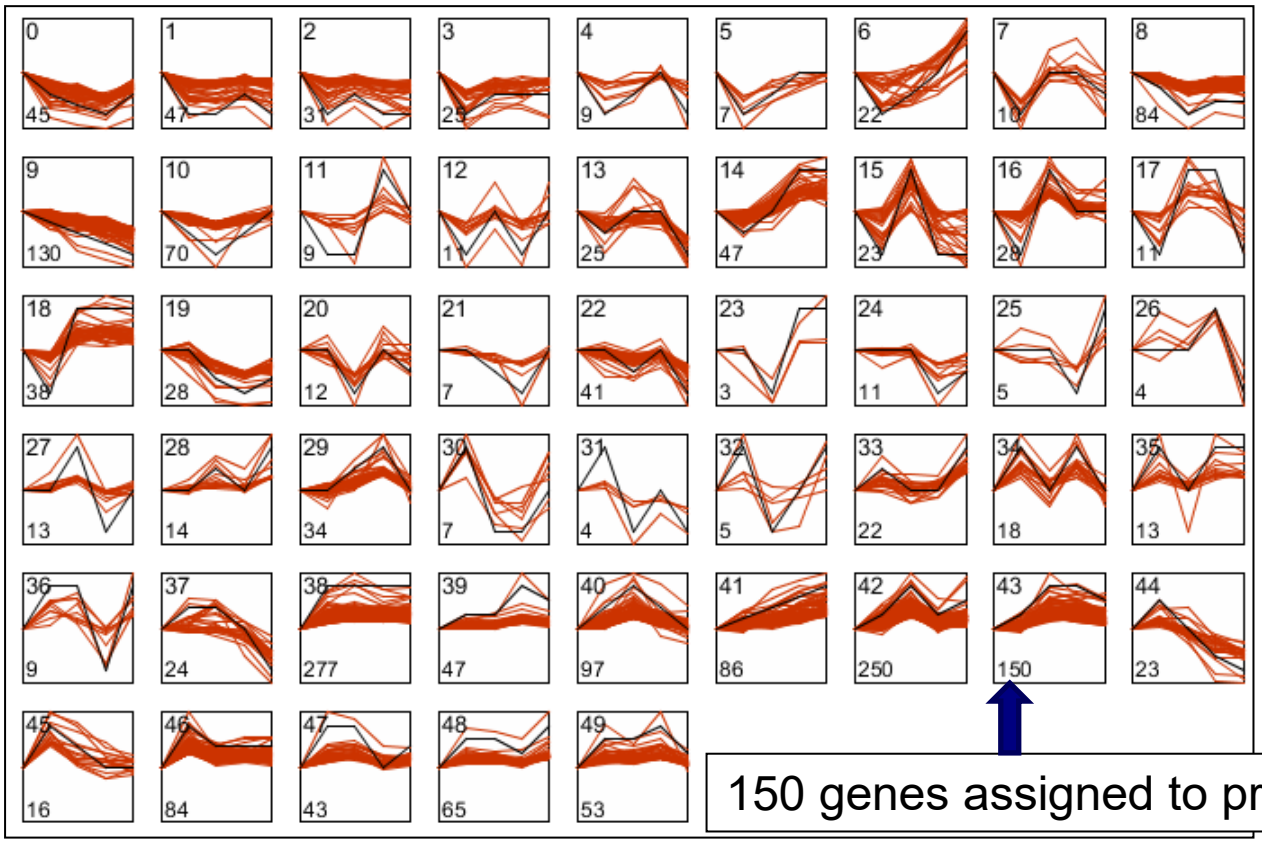


STEP 1. Define temporal profiles independent of data

STEP 2. Assign genes to most closely matching profile

STEP 2: Filter flat genes, and then assign the remaining genes to the most closely matching model profile based on the correlation coefficient.

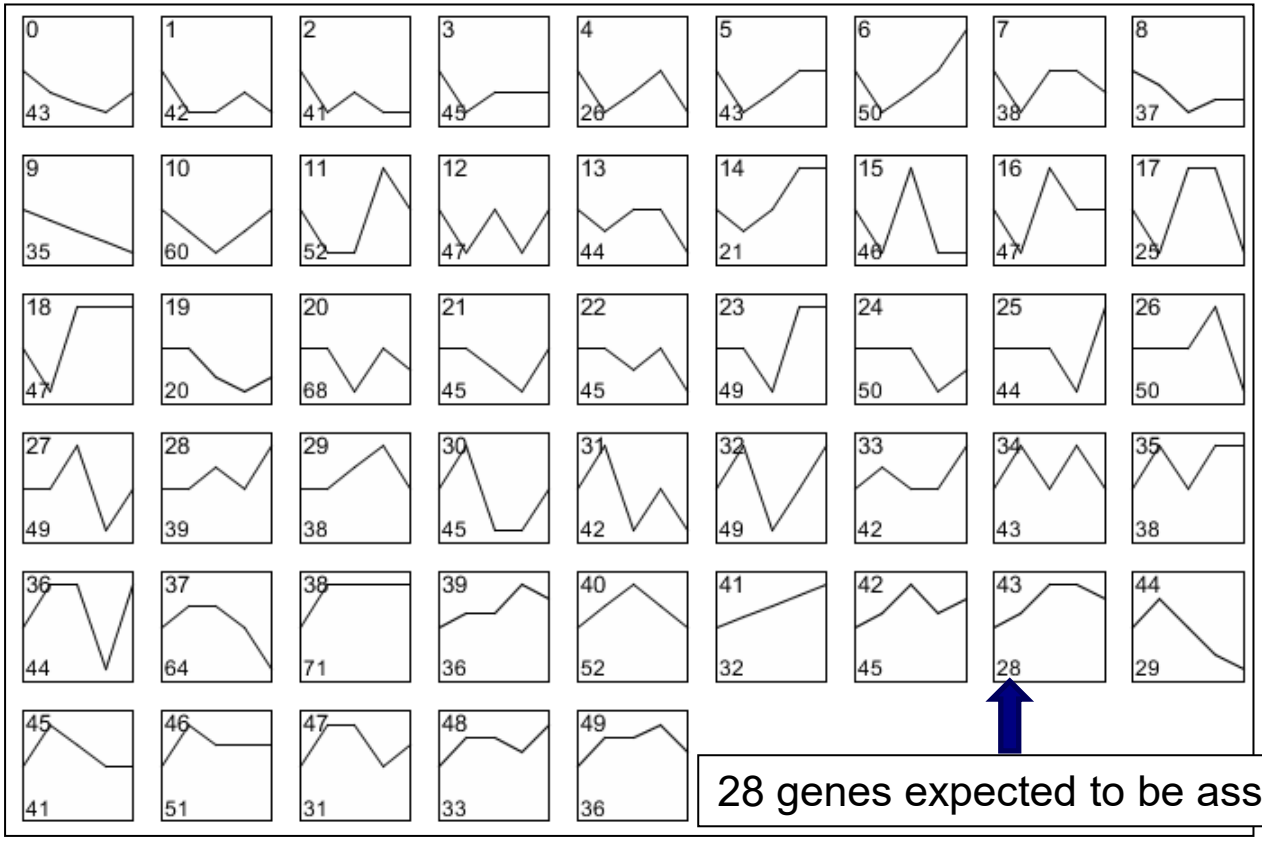
Data for immune response to a pathogen infection (Guillmen et al PNAS, 2002)



STEP 1. Define temporal profiles independent of data

STEP 2. Assign genes to most closely matching profile

STEP 3: Compute the expected number of genes assigned to a profile based on a permutation test on the time points.



28 genes expected to be assigned to profile 43

STEP 1. Define temporal profiles independent of data

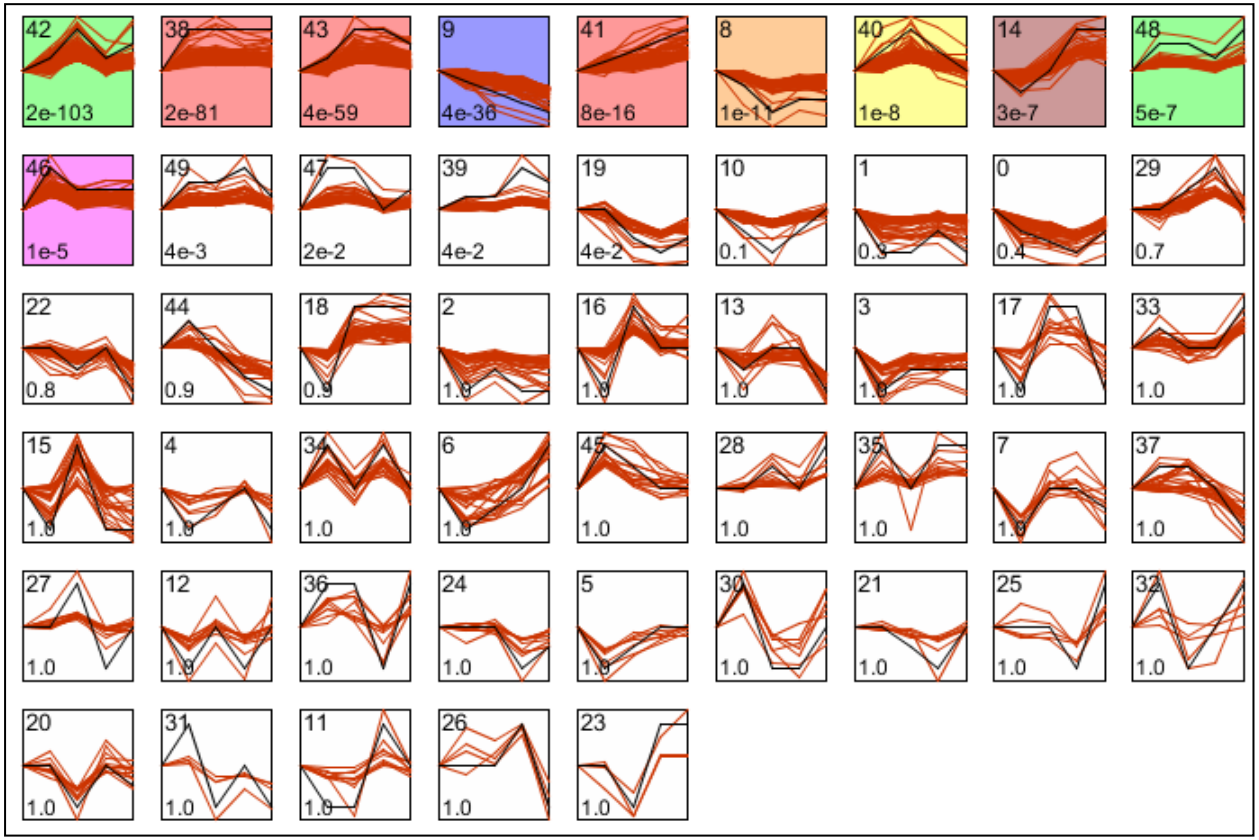
STEP 2. Assign genes to most closely matching profile

STEP 3. Compute expected number of genes per profile

Gene	0h	0.5h	3h	6h	12h
SERPINA7	-0.70	-0.54	-0.09	-0.15	-0.10
THBD	-1.19	-1.03	-0.62	-0.62	-0.59
EPHA2	0.13	0.36	0.57	-0.07	0.35
RBM5	-0.01	-0.01	-0.01	-0.23	0.36
SFRS10	0.30	0.22	0.29	0.41	-0.34

Gene	12h	3h	6h	0h	0.5h
SERPINA7	-0.10	-0.09	-0.15	-0.70	-0.54
THBD	-0.59	-0.62	-0.62	-1.19	-1.03
EPHA2	0.35	0.57	-0.07	0.13	0.36
RBM5	0.36	-0.01	-0.23	-0.01	-0.01
SFRS10	-0.34	0.29	0.41	0.30	0.22

STEP 4: Using the binomial distribution and the counts from steps 2 and 3 associate statistical significance with the number of genes assigned to each profile



Profiles ordered based on significance; Colored profiles are significant at a 0.05 bonferroni corrected level

STEP 1. Define temporal profiles independent of data

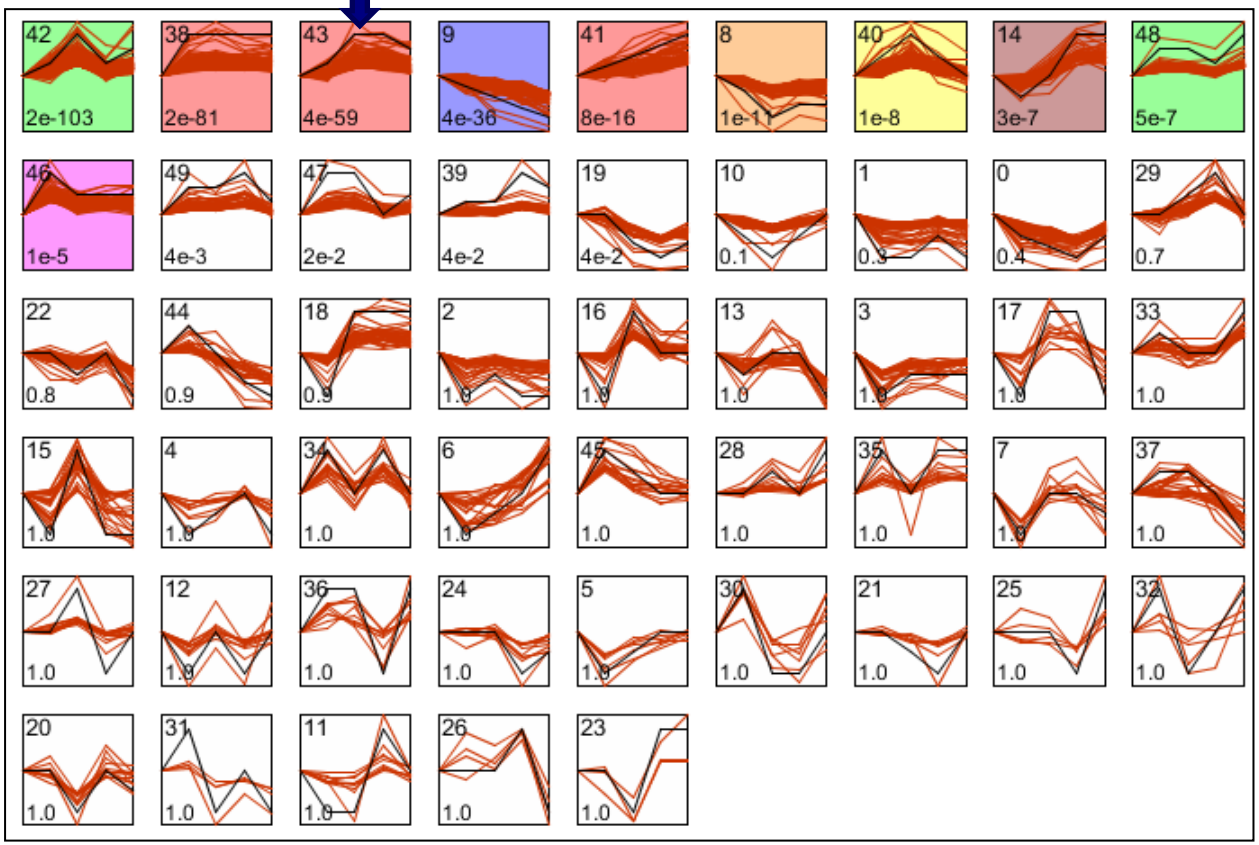
STEP 2. Assign genes to most closely matching profile

STEP 3. Compute expected number of genes per profile

STEP 4. Determine statistically significant profiles

STEP 4: Using the binomial distribution and the counts from steps 2 and 3 associate statistical significance with the number of genes assigned to each profile

Profile 43 genes enriched for negative regulation of cell death genes ($p<10^{-3}$)



Profiles ordered based on significance; Colored profiles are significant at a 0.05 bonferroni corrected level

STEP 1. Define temporal profiles independent of data

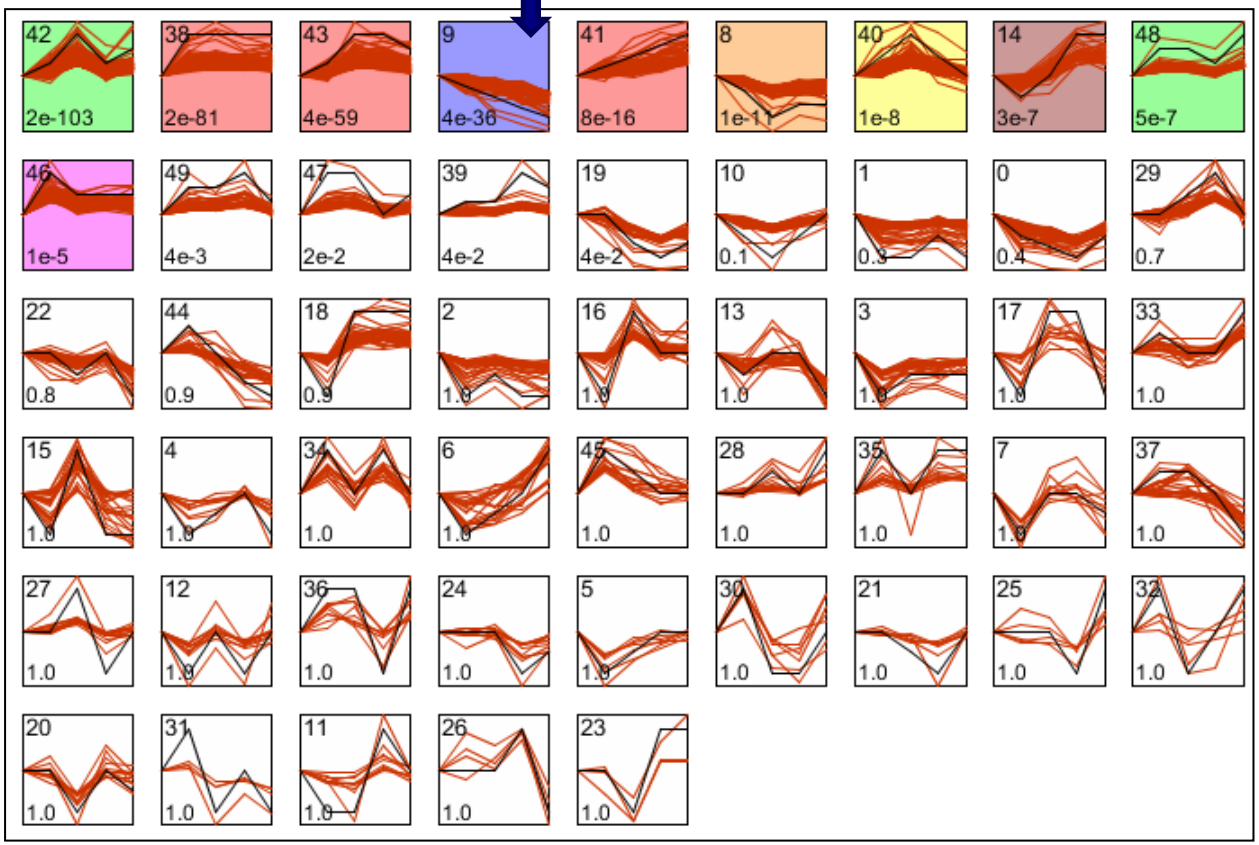
STEP 2. Assign genes to most closely matching profile

STEP 3. Compute expected number of genes per profile

STEP 4. Determine statistically significant profiles

STEP 4: Using the binomial distribution and the counts from steps 2 and 3 associate statistical significance with the number of genes assigned to each profile

Profile 9 genes enriched for DNA replication genes ($p<10^{-10}$)



Profiles ordered based on significance; Colored profiles are significant at a 0.05 bonferroni corrected level

STEP 1. Define temporal profiles independent of data

STEP 2. Assign genes to most closely matching profile

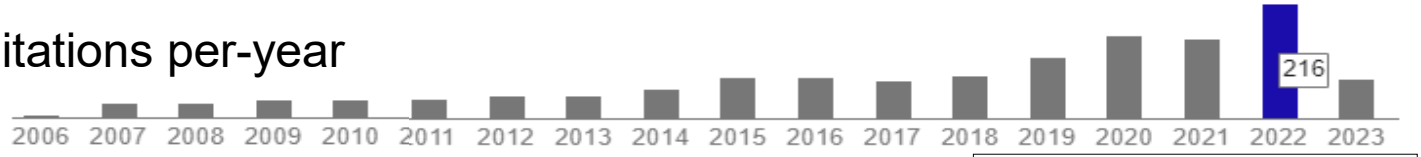
STEP 3. Compute expected number of genes per profile

STEP 4. Determine statistically significant profiles

The Short Time-series Expression Miner (STEM) software is facilitating an increasing and diverse range of biological discoveries

Software available at www.sb.cs.cmu.edu/stem

Citations per-year



blood
2008 112: 2318-2326
Prepublished online Jul 9, 2008;
doi:10.1182/blood-2008-05-156331
Transcriptional profiling of VEGF-A and VEGF-C target genes in lymphatic endothelium reveals endothelial-specific molecule-1 as a novel mediator of lymphangiogenesis

Vorinostat interferes with the signaling transduction pathway of T-cell receptor and synergizes with phosphoinositide-3 kinase inhibitors in cutaneous T-cell lymphoma
haematologica | 2010; 95(4)

OPEN ACCESS Freely available online
PLoS one
Dissecting Interferon-Induced Transcriptional Programs in Human Peripheral Blood Cells

BMC Genomics
BioMed Central
Research article
Open Access
Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines

TOXICOLOGICAL SCIENCES 117(2), 381–392 (2010)
doi:10.1093/toxsci/kfq214
Advance Access publication July 12, 2010
Time Series Analysis of Benzo[A]Pyrene-Induced Transcriptome Changes Suggests That a Network of Transcription Factors Regulates the Effects on Functional Gene Sets

research articles
Journal of proteome research
In Vitro Neurotoxicity of PBDE-99: Immediate and Concentration-Dependent Effects on Protein Expression in Cerebral Cortex Cells

Spatiotemporal compartmentalization of key physiological processes during muscle precursor differentiation
4224–4229 | PNAS | March 2, 2010 | vol. 107 | no. 9

Longitudinal shift in diabetic wound microbiota correlates with prolonged skin defense response
PNAS | August 17, 2010 | vol. 107 | no. 33 | 14799–14804

FEMS
RESEARCH LETTER
Cold-induced gene expression profiles of *Vibrio parahaemolyticus*: a time-course analysis

Available online at www.sciencedirect.com
ScienceDirect
Journal of Biotechnology 134 (2008) 162–170
Journal of BIOTECHNOLOGY
ELSEVIER
Cardiomyogenic gene expression profiling of differentiating human embryonic stem cells

Microbiology (2009), 155, 80–94
DOI 10.1099/mic.0.021790-0
Metabolite and transcriptome analysis of *Campylobacter jejuni in vitro* growth reveals a stationary-phase physiological switch

FEMS
RESEARCH LETTER
Global gene expression profiling of wild type and *lysC* knockout *Escherichia coli* W3110

Molecular Vision 2007; 13:1259-73 <http://www.molvis.org/molvis/v13/a137>
Received 19 March 2007 | Accepted 20 June 2007 | Published 23 July 2007
©2007 Molecular Vision
Analysis of the RPE transcriptome reveals dynamic changes during the development of the outer blood-retinal barrier

Online Submissions: http://www.wjgnet.com/1007-9327office
wjg@wjgnet.com
doi:10.3748/wjg.v16.i11.1385
World J Gastroenterol 2010 March 21; 16(11): 1385-1396
ISSN 1007-9327 (print)
© 2010 Baishideng. All rights reserved.
ORIGINAL ARTICLE
Time-series gene expression profiles in AGS cells stimulated with *Helicobacter pylori*

Journal of Molecular Microbiology and Biotechnology
J Mol Microbiol Biotechnol 2007;12:82–95
DOI: 10.1159/000096463
Transcriptome Analysis of Temporal Regulation of Carbon Metabolism by CcpA in *Bacillus subtilis* Reveals Additional Target Genes

IMV Molecular Vision 2008; 14:2237-2262 <http://www.molvis.org/molvis/v14/a261>
Received 15 July 2008 | Accepted 25 November 2008 | Published 8 December 2008
© 2008 Molecular Vision
Diffusible retinal secretions regulate the expression of tight junctions and other diverse functions of the retinal pigment epithelium

Journal of General Virology (2007), 88, 570–581
DOI 10.1099/vir.0.82187-0
Distinct gene subsets are induced at different time points after human respiratory syncytial virus infection of A549 cells

OPEN ACCESS Freely available online
PLoS one
Comparative Developmental Expression Profiling of Two *C. elegans* Isolates

PHYTOTHERAPY RESEARCH
Phytother. Res. **24**: 531–537 (2010)
Published online 4 August 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/ptr.2976
Reciprocal Regulation of Gene Expression by *Ephedra herba* in Mouse Brain

Biochemical and Biophysical Research Communications 370 (2008) 514–518
Gene expression patterns in glucose-stimulated podocytes ☆

BMC Genomics
BioMed Central
Database
Open Access
An expression database for roots of the model legume *Medicago truncatula* under salt stress

OPEN ACCESS Freely available online
PLoS one
Striatal Proteomic Analysis Suggests that First L-Dopa Dose Equates to Chronic Exposure

The Plant Cell, Vol. 22: 655–671, March 2010, www.plantcell.org © 2010 American Society of Plant Biologists
Apomictic and Sexual Ovules of *Boechera* Display Heterochronic Global Gene Expression Patterns

BMC Genomics
BioMed Central
Research article
Open Access
Gene expression during *Drosophila melanogaster* egg development before and after reproductive diapause

BMC Plant Biology
BioMed Central
Research article
Open Access
Comprehensive transcriptional profiling of NaCl-stressed Arabidopsis roots reveals novel classes of responsive genes

OPEN ACCESS Freely available online
PLoS one
Temporal and Regional Regulation of Gene Expression by Calcium-Stimulated Adenylyl Cyclase Activity during Fear Memory

Ernst and Bar-Joseph, BMC Bioinformatics 2006