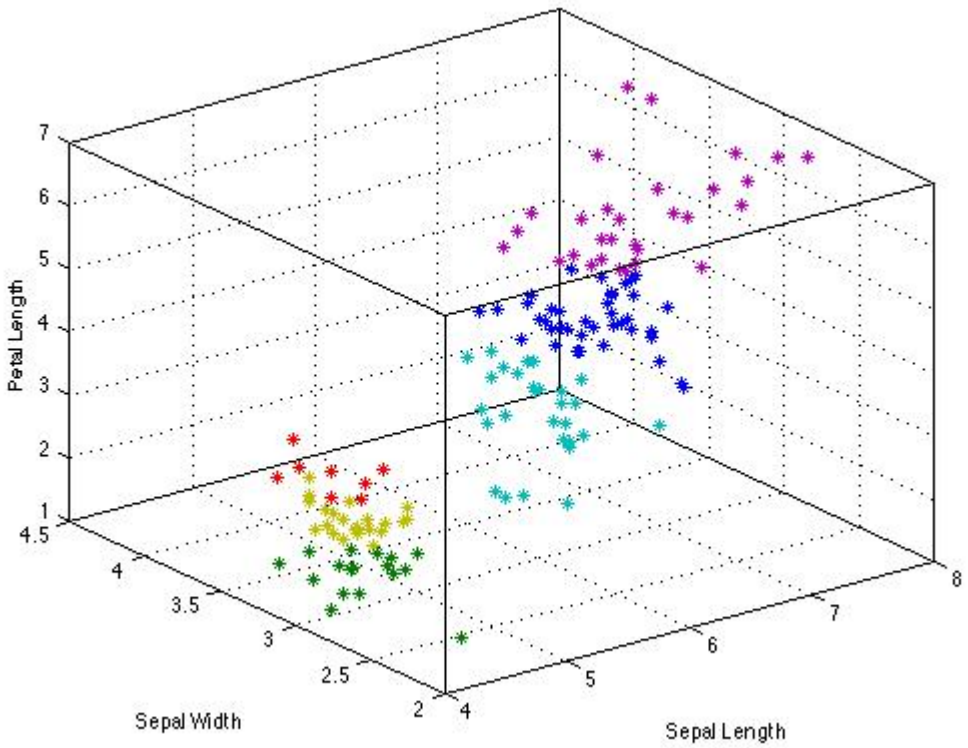
Data Clustering



Agrupamiento de Datos / Data Clustering

- **El *agrupamiento de datos* consiste en la clasificación de objetos diferentes grupos, de manera que objetos similares son agrupados en el mismo grupo.**
- **Otra definición: particionar un conjunto de datos en subconjuntos o *clusters* de tal manera que estos tengan "algo en común".**
  - **El problema: cuantificar "algo en común"**
    - Proximidad
    - Similitud
- **Es un tipo de aprendizaje *no supervisado***
- **Es un problema combinatorio difícil**

Agrupamiento de Datos / Data Clustering

1. La motivacion: - Encontrar patrones en un conjunto enorme de datos

2. La Entrada (Input):

- Un gran numero de Datos (Puntos o elementos) - N
- Una medida de La distancia entre los puntos de datos – d_{ij}

3. La Salida (Output):

- **Agrupación** (Clustering) de los Elementos en K clases de similitud. K – es un número (de Clusters) proporcionado por el usuario o puede ser automáticamente determinado
- **Homogeneidad** - Los Elementos asignados a un cluster deben ser similares entre si (es decir, sus distancias deben ser las minimas)
- **Separacion** - Los elementos de clusters diferentes estan separados lo mas posible

Agrupamiento de Datos / Data Clustering

Se pueden agrupar:

- Secuencias (DNA, RNA)
  - Ej: Agrupar por similitud/identidad global
  - Ej: Agrupar por presencia de motivos o señales
- Medidas de expresión de genes
  - Ej: Agrupar todos los genes que tienen alta expresión
- Abstracts en PubMed
  - Ej: Agrupar abstracts en base a número de palabras compartidas
- Marcadores morfológicos
  - Ej: Puntos fluorescentes en una imagen de microscopía (por ej para delinar una membrana o cualquier otra estructura celular)
- O todo a la vez
  - Vectores multidimensionales

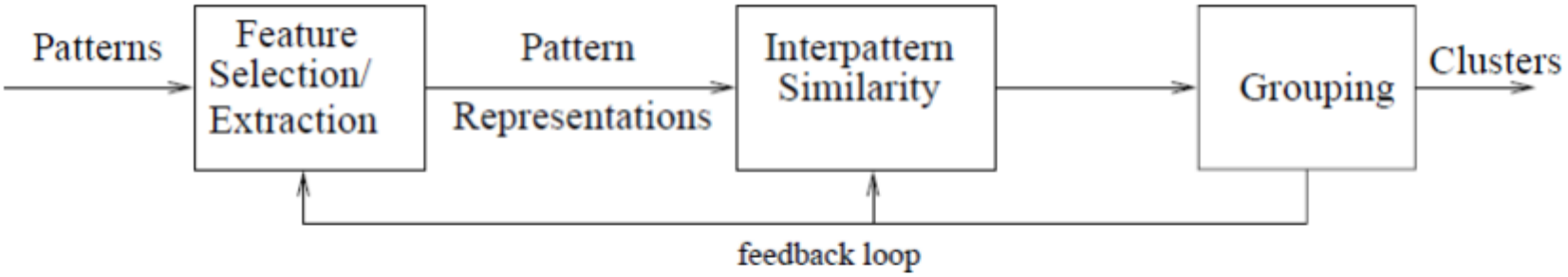
Vectores multidimensionales

Ejemplo: un vector por gen

Conteniendo información para:

- Presencia ausencia de señales (motivos)
- Niveles de expresión en distintos tejidos
- Información sobre interacciones (protein-protein)
- Etc.

Pasos en el Data Clustering



Feature selection:

- Identificar en el dataset el subset de características (features) más informativo para agrupar objetos

Pattern representations:

- La manera de representar una característica afecta directamente a las medidas de similitud

Pattern proximity:

- Hay muchas maneras de medir proximidad (distancias). En general se calculan distancias de a pares, para todos los objetos a agrupar

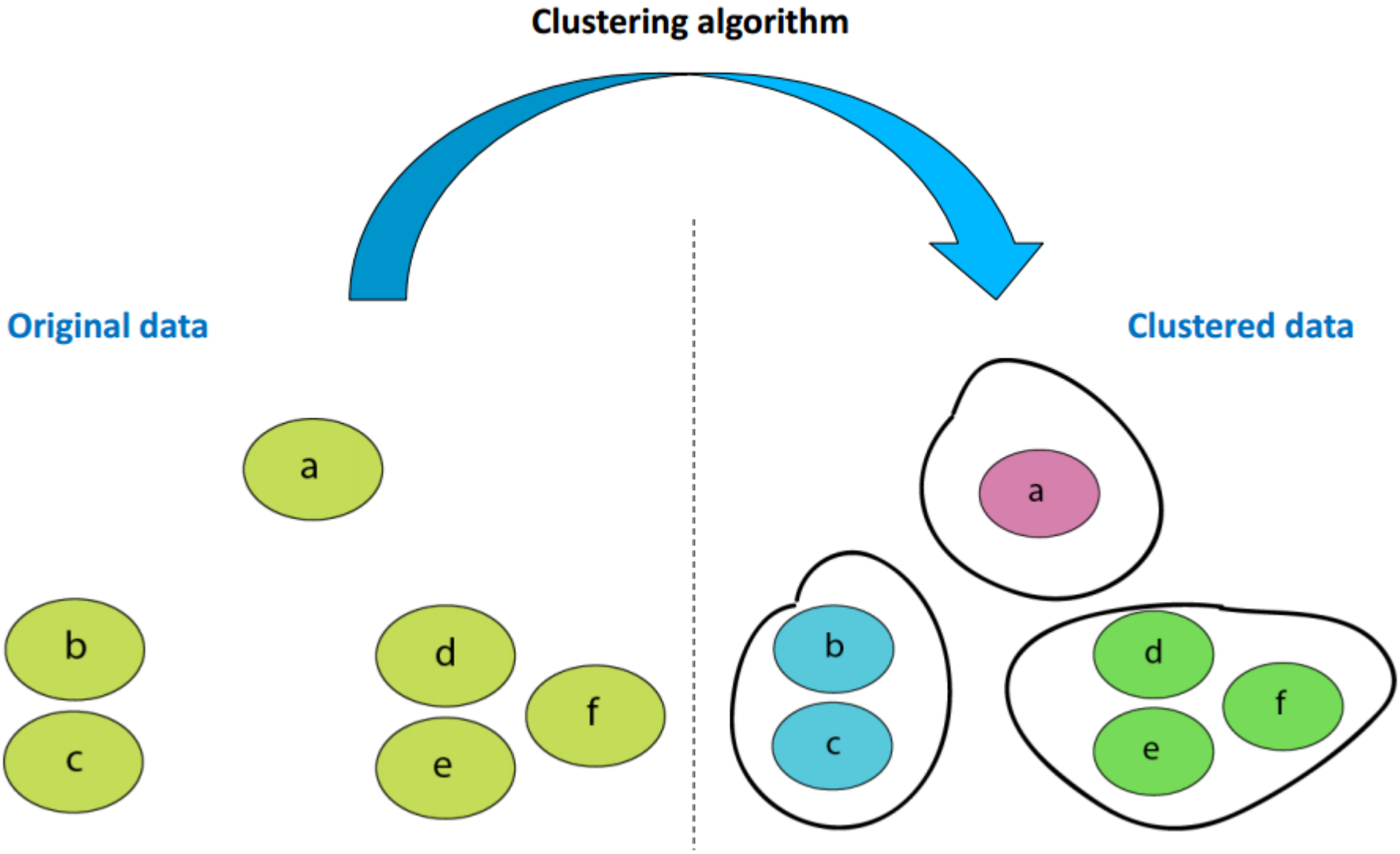
Clustering:

- Hay muchos algoritmos (estrategias) de clustering

Cluster validation analysis

- La estructura de agrupamiento es válida si no puede obtenerse simplemente por azar o no es producto de un artefacto del método

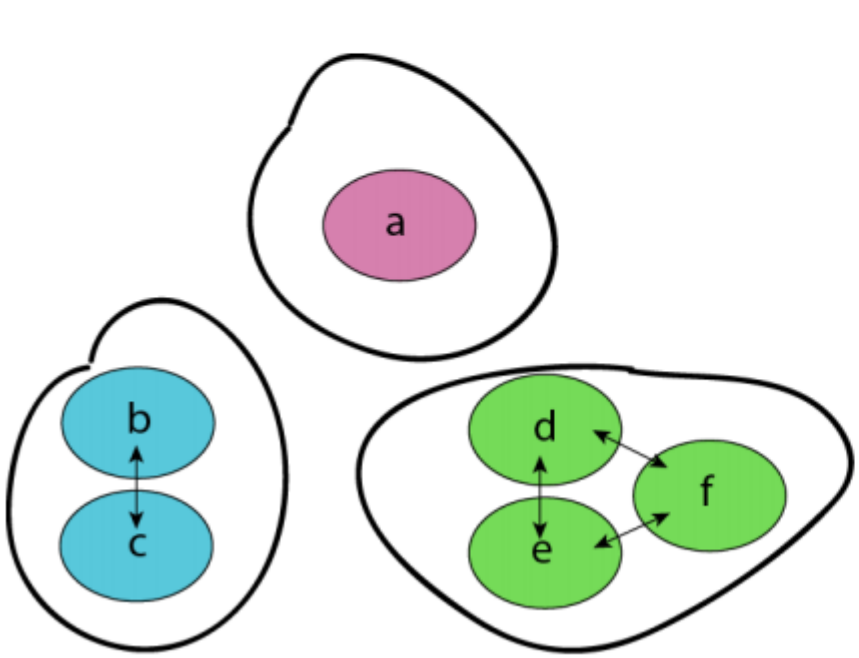
Objetivo del algoritmo de clustering

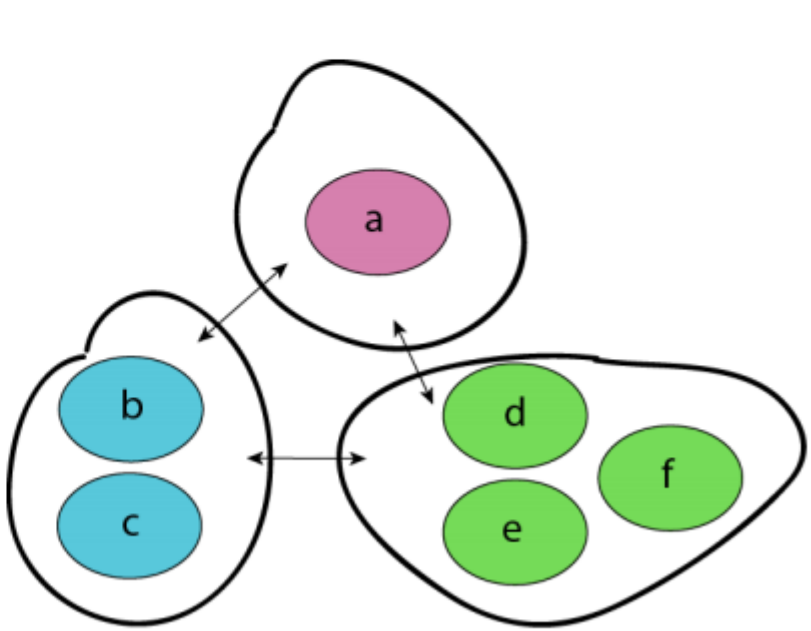


Objetivo del algoritmo de clustering

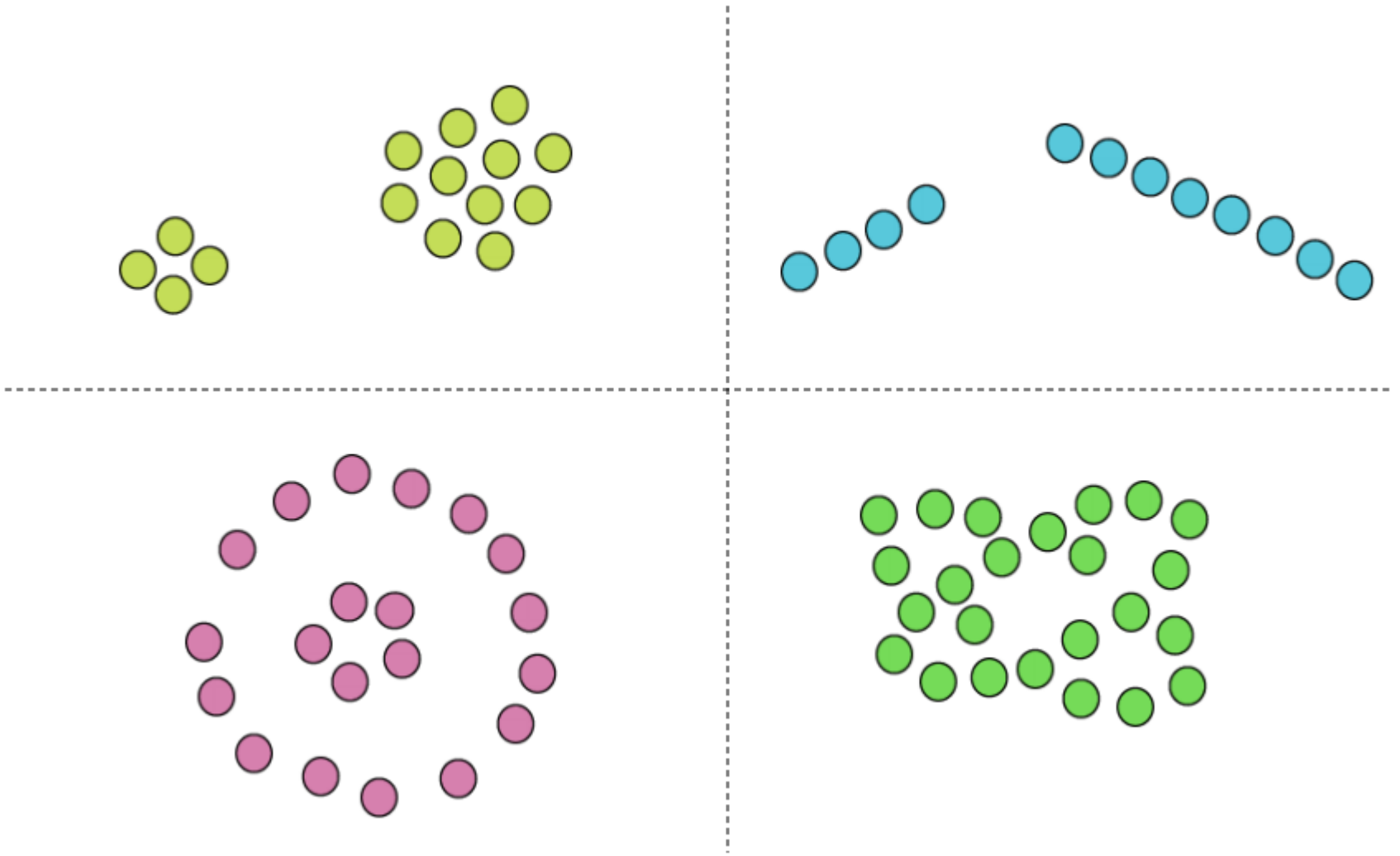
Minimizar la distancia intracluster

Maximizar la distancia entre clusters

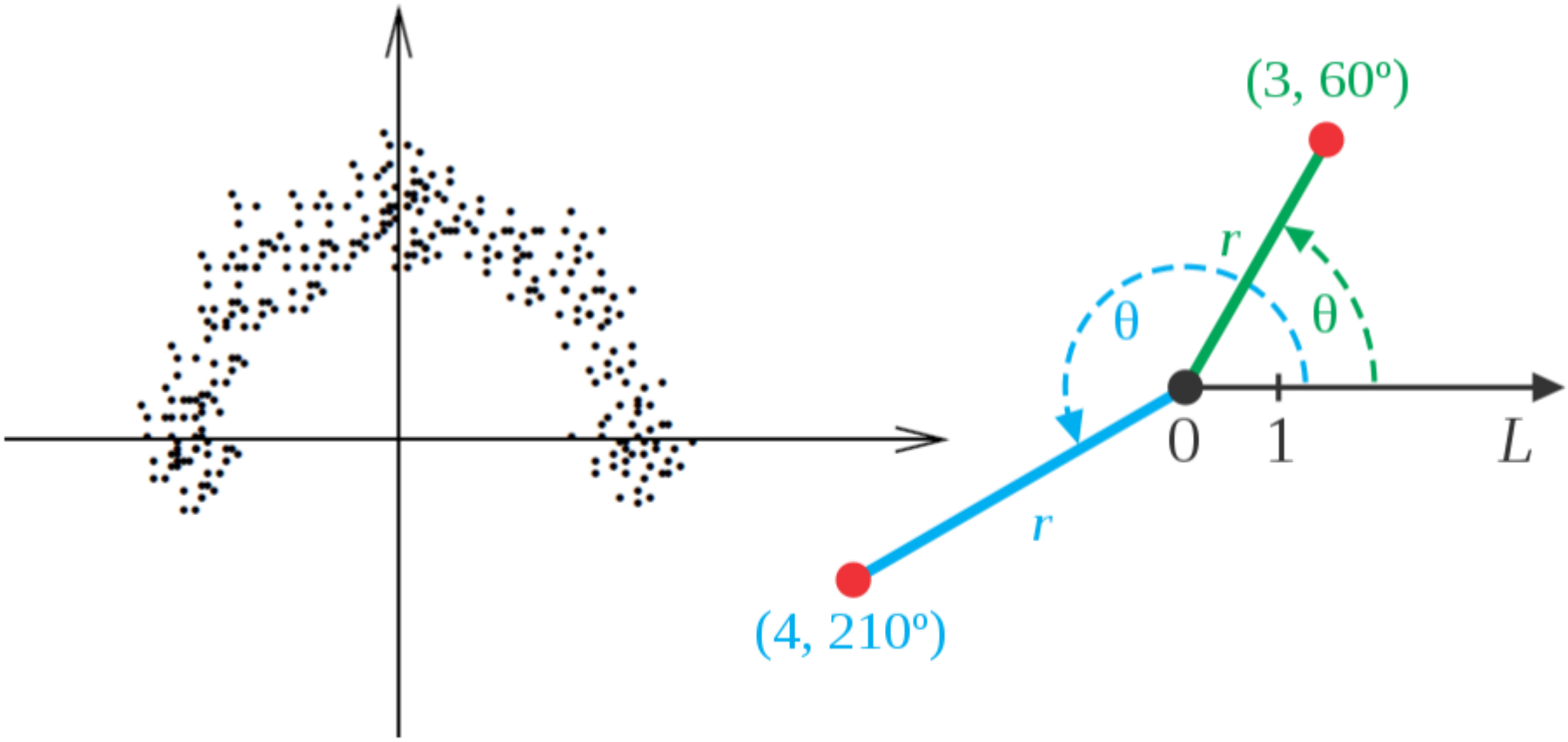




Formas de Clusters



Ejemplo de Representacion de datos

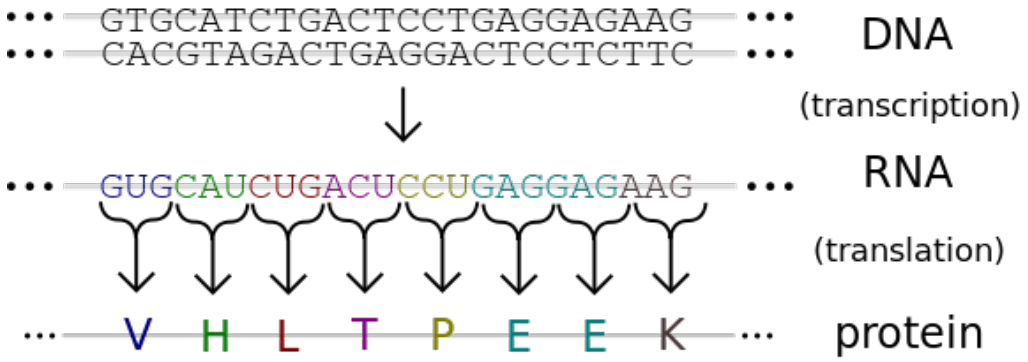


Cluster curvilineo, donde los puntos están mas o menos equidistantes del origen. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review.

Ejemplo de Representacion de datos

Expresion génica

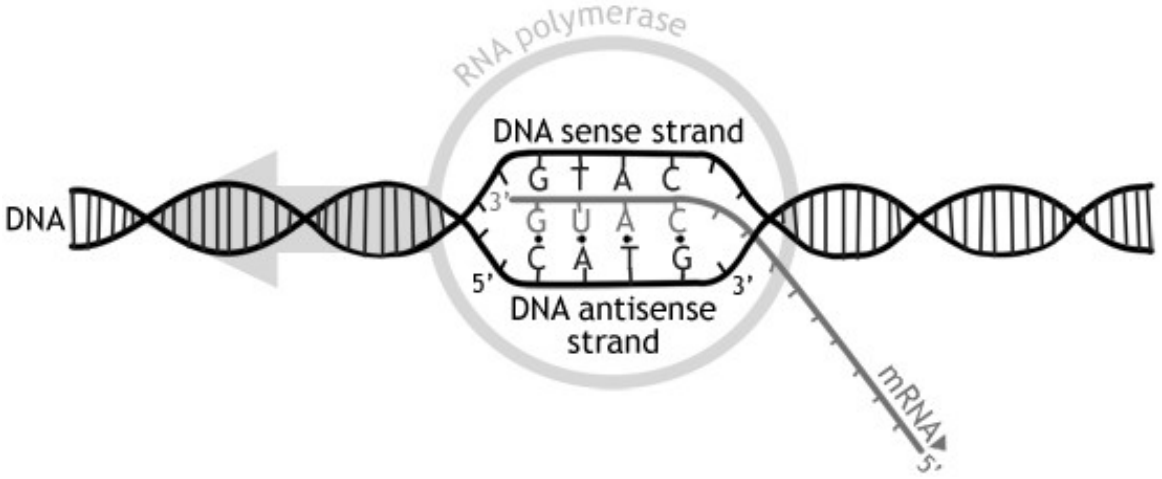
1.Cual es la funcion de un “nuevo” Gen ?
2. Simplemente comparando secuencias de nuevos genes con las secuencias de Genes conocidos no necesariamente ayudará a saber la funcion de un determinado Gen. De hecho, para el 40 % de los Genes secuenciados esta afirmación se cumple.
3. Los Genes que cumplen funciones similares o complementarias a las de Genes conocidos, seran transcritos o se expresaran conjuntamente con los genes conocidos



Ejemplo de Representacion de datos

Expresion génica

La transcripción del ADN es el primer proceso de la expresion Génica, mediante el cual se transfiere la información contenida en la secuencia del ADN hacia la secuencia de la proteina utilizando diversos ARN como intermediarios. Durante la transcripción genética, las secuencias de ADN son copiadas a ARN mediante una enzima llamada ARN polimerasa la cual sintetiza un ARN mensajero que mantiene la información de la secuencia del ADN.

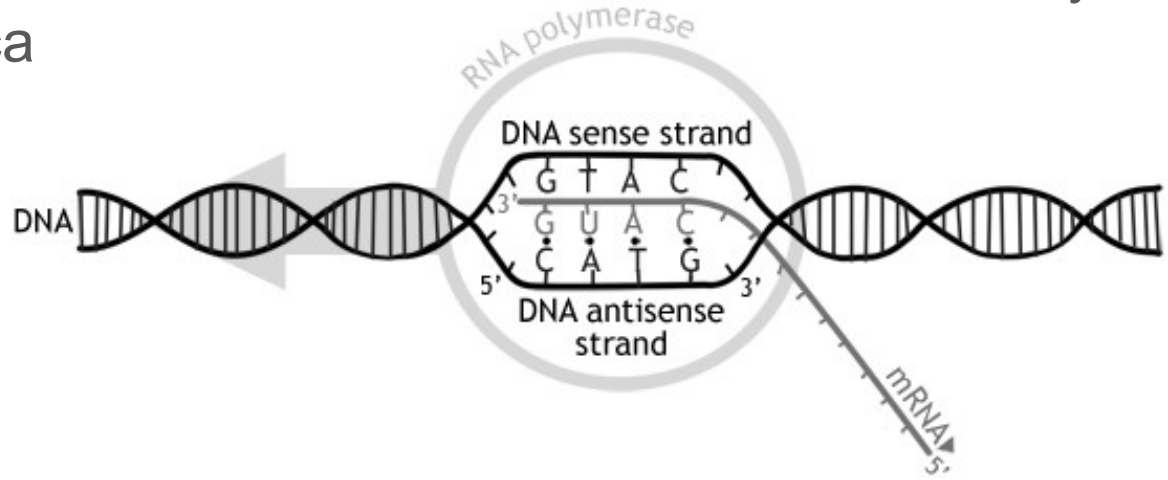


Ejemplo de Representacion de datos

Expresion génica

Los genes que tienen funcion antagónica podrían ser expresados en un punto en el tiempo posterior o anterior al momento de la expresion de Genes Conocidos.
El nivel de expresion se estima midiendo la cantidad de ARNm (ARN -mensajero) generado para ese Gen en particular.

- Un Gen es activo, cuando se manifiesta el proceso de transcripcion
- Una mayor cantidad de ARNm indicará entondes mayor actividad Genica



Ejemplo de Representacion de datos

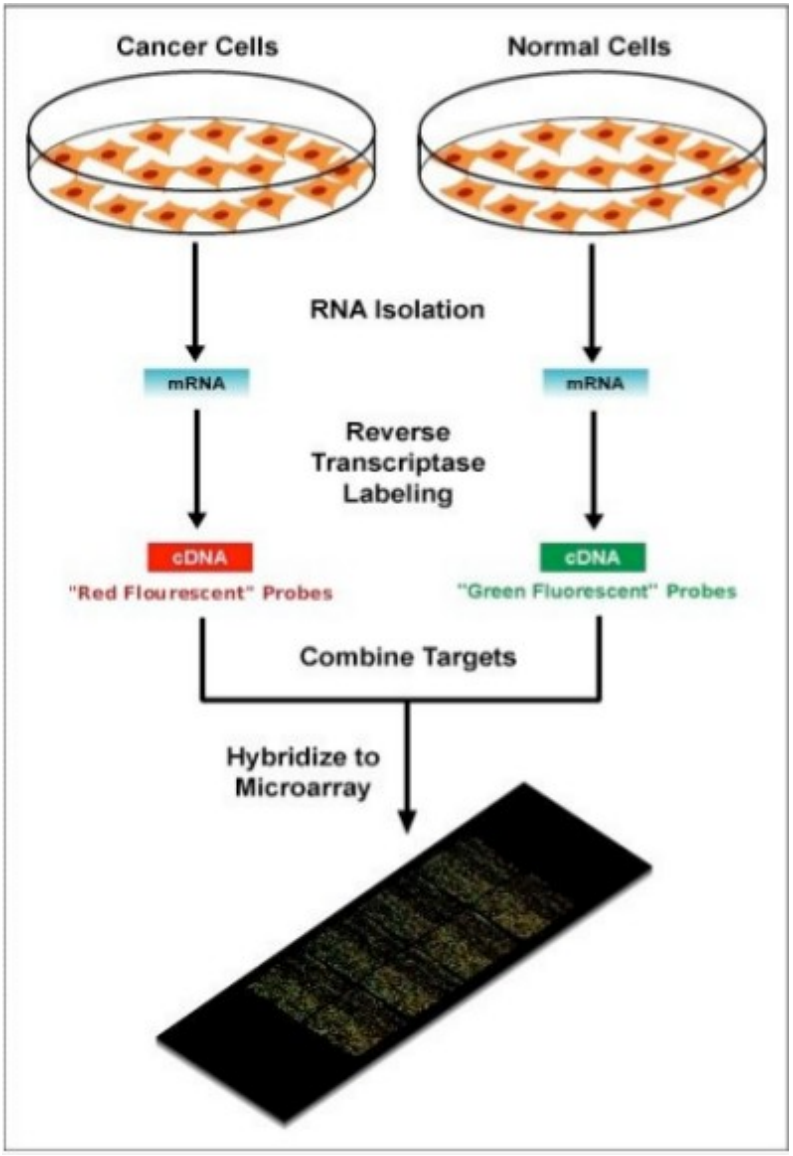
Micro Arrays

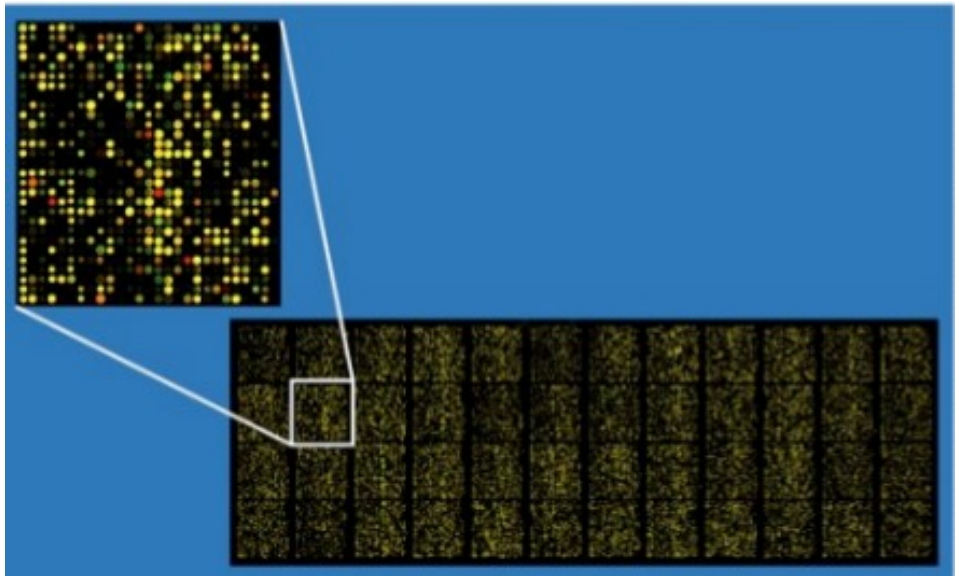
Los experimentos con Micro Arrays permiten estudiar los niveles de expresion génica de diferentes genes en lapso de tiempo.

1. Se sintetiza ADN Complementario (cDNA))a partir de ARN – mensajero (mRNA), ya que cDNA es más estable. - utilizando una enzima denominada transcriptasa inversa.
2. Se mezcla cDNA con un tinte fluorescente, para que pueda ser detectado
3. Se vierte el cDNA sobre el Micro Array que contiene miles de sondas de alta densidad que se acoplan con cadenas complementarias en la muestra y las inmobilizan en la superficie.
4. Se lee el Micro Array con Laser o camaras de alta resolucion.
5. La iluminación revela cuales son los genes transcritos o que se han expresado.

“El ADN complementario es una hebra de ADN de doble cadena una de las cuales constituye una secuencia totalmente complementaria del ARN mensajero a partir del cual se ha sintetizado. Se suele utilizar para la clonación de genes propios de células eucariotas en células procariotas, debido a que, dada la naturaleza de su síntesis, carece de intrones.” (https://es.wikipedia.org/wiki/ADN_complementario)

Clustering en Biologia Computacional





- Green: expressed only in control
- Red: expressed only in an experimental cell
- Yellow: equally expressed in both samples
- Black: NOT expressed in either control or sample

Clustering en Biologia Computacional

6. La Informacion recopilada usualmente se transfiere a una matriz de intensidad relativamente normalizada.
7. La matriz de intensidad (o Matriz de Expresion) le permite a los Biologos inferir correlaciones entre genes, - (incluso si estos son muy diferentes entre si)- y entender
como las funciones de los genes podrían estar relacionadas entre si.

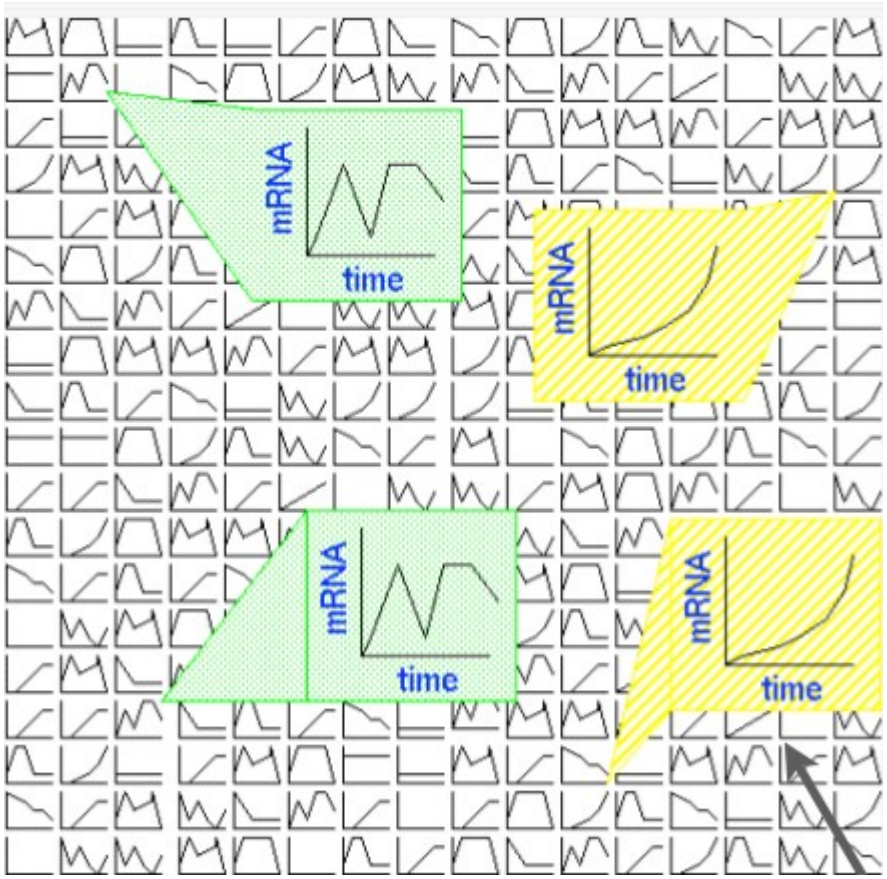
Gene	Time 1	Time 2	Time 3
1	10	8	10
2	10	0	9
3	4	8.5	3
4	9.5	0.5	8.5
5	4.5	8.5	3
6	10.5	9	12
7	5	8.5	11
8	2.7	8.7	2
9	9.7	2	9
10	10.2	1	9.2

$$D(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

Distancia euclidiana en d dimensiones

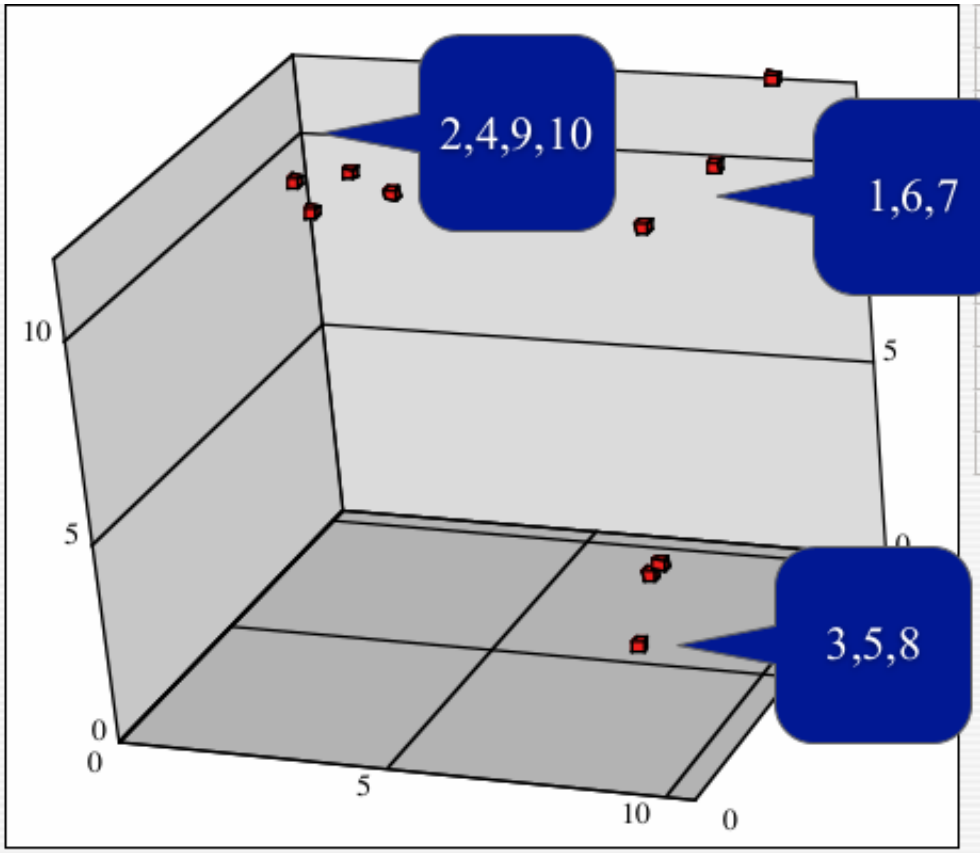
Clustering en Biologia Computacional

6. Se puede monitorear un Micro array en el tiempo para observar los cambios en la expresion génica al pasar el tiempo.
7. Tambien se pueden monitorear 2 muestras que se encuentren bajo las mismas condiciones para observar la expresion diferencial.



Cada cuadro representa La expresion Génica de un Gen en La unidad de tiempo

Clustering en Biologia Computacional



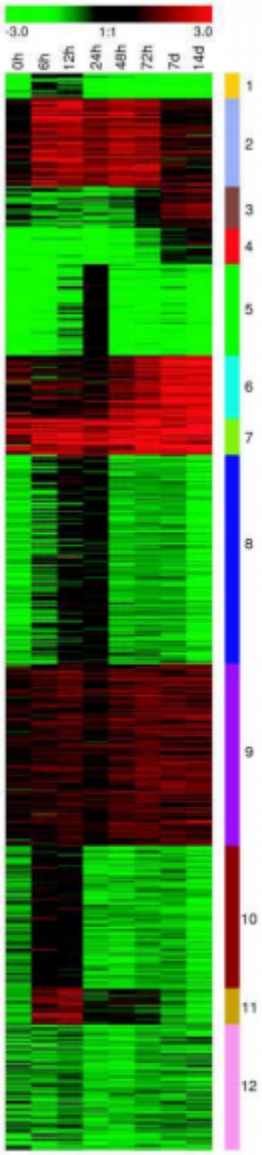
	1	2	3	4	5	6	7	8	9	10
1		8.1	9.2	7.7	8.9	2.3	5.1	10.9	6.1	7.0
2	8.1		12.0	0.9	11.8	9.5	10.1	13.3	2.0	1.0
3	9.2	12.0		11.2	0.5	11.1	8.1	1.7	10.5	11.5
4	7.7	0.9	11.2		10.9	9.2	9.5	12.5	1.6	1.1
5	8.9	11.8	0.5	10.9		10.8	8.0	2.1	10.3	11.3
6	2.3	9.5	11.1	9.2	10.8		5.6	12.7	7.7	8.5
7	5.1	10.1	8.1	9.5	8.0	5.6		9.3	8.3	9.3
8	10.9	13.3	1.7	12.5	2.1	12.7	9.3		12.0	12.9
9	6.1	2.0	10.5	1.6	10.3	7.7	8.3	12.0		1.1
10	7.0	1.0	11.5	1.1	11.3	8.5	9.3	12.9	1.1	

PAIRWISE DISTANCES

	1	6	7	2	4	9	10	3	5	8
1	0.0	2.3	5.1	8.1	7.7	6.1	7.0	9.2	8.9	10.9
6	2.3	0.0	5.6	9.5	9.2	7.7	8.5	11.1	10.8	12.7
7	5.1	5.6	0.0	10.1	9.5	8.3	9.3	8.1	8.0	9.3
2	8.1	9.5	10.1	0.0	0.9	2.0	1.0	12.0	11.8	13.3
4	7.7	9.2	9.5	0.9	0.0	1.6	1.1	11.2	10.9	12.5
9	6.1	7.7	8.3	2.0	1.6	0.0	1.1	10.5	10.3	12.0
10	7.0	8.5	9.3	1.0	1.1	1.1	0.0	11.5	11.3	12.9
3	9.2	11.1	8.1	12.0	11.2	10.5	11.5	0.0	0.5	1.7
6	8.9	10.8	8.0	11.8	10.9	10.3	11.3	0.5	0.0	2.1
8	10.9	12.7	9.3	13.3	12.5	12.0	12.9	1.7	2.1	0.0

REARRANGED DISTANCES

Clustering en Biologia Computacional



Expressed sequence tag (EST) - Es una sub-sequencia de una secuencia de ADN complementario (cDNA).

Ej. de Clustering de ESTs.

Expresión de genes a lo largo de un experimento.

No importa tanto si los genes se expresan mucho o poco (ej agrupar por nivel de expresión no tiene sentido)

Importa el comportamiento de cada gen a lo largo de un tratamiento experimental.

Correlation distance. Mide la dependencia entre datos.

CD = 0 si los datos son independientes.
CD = 1 si tienen dependencia.

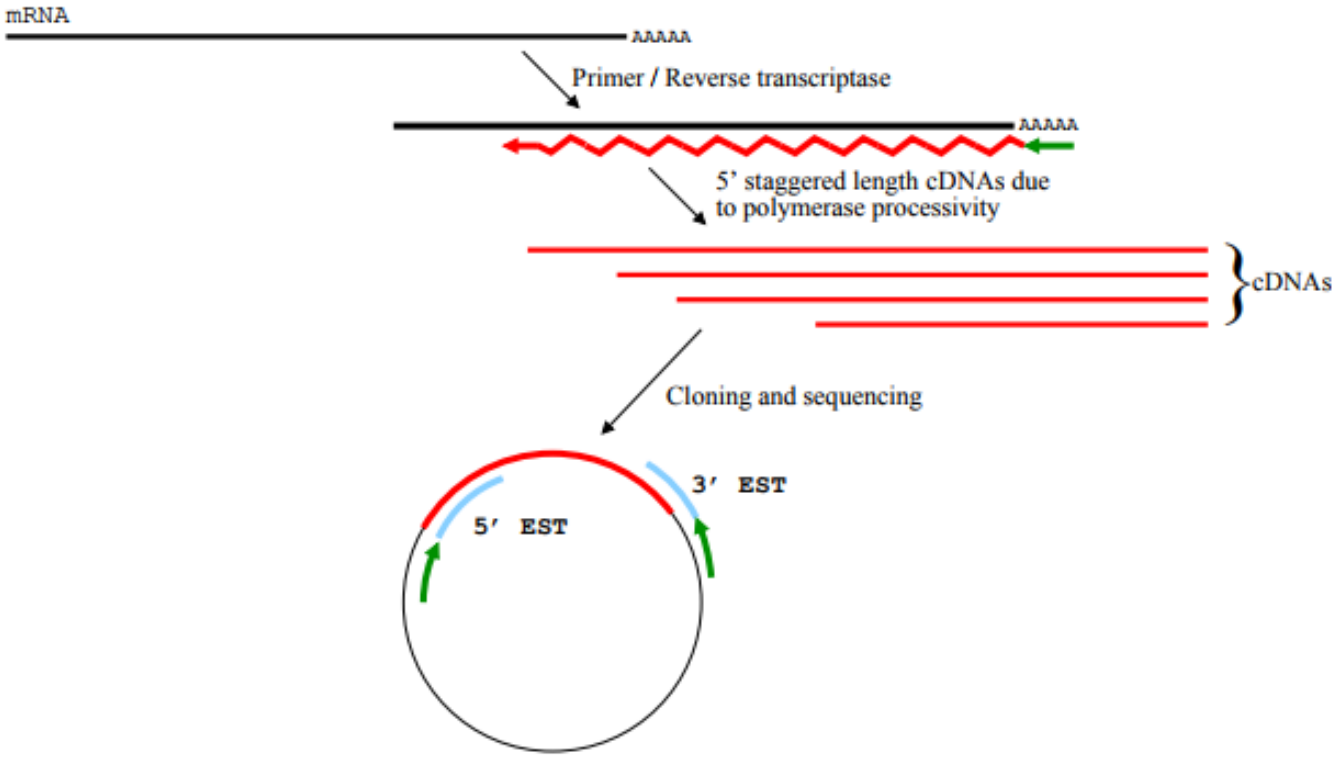
Clustering of ESTs found to be differentially expressed during fat cell differentiation. Shown is k-means clustering of 780 ESTs found to be more than twofold upregulated or downregulated at a minimum of four time points during fat cell differentiation. ESTs were grouped into 12 clusters with distinct expression profiles.
Hackl *et al. Genome Biology* 2005 **6**:R108 doi:10.1186/gb-2005-6-13-r108

Clustering en Biologia Computacional

Expressed sequence tags (ESTs)

ESTs represent partial sequences of cDNA clones (average ~ 360 bp).

Single-pass reads from the 5' and/or 3' ends of cDNA clones.



Clustering en Biologia Computacional

Interest for ESTs

ESTs represent the most extensive available survey of the transcribed portion of genomes.

ESTs are indispensable for gene structure prediction, gene discovery and genomic mapping.

Characterization of splice variants and alternative polyadenilation.

In silico differential display and gene expression studies (specific tissue expression, normal/disease states).

SNP data mining.

High-volume and high-throughput data production at low cost.

Clustering en Biologia Computacional

Low data quality of ESTs

High error rates ($\sim 1/100$) because of the sequence reading single-pass.

Sequence compression and frame-shift errors due to the sequence reading single-pass.

A single EST represents only a partial gene sequence.

Not a defined gene/protein product.

Not curated in a highly annotated form.

High redundancy in the data $\Rightarrow$ huge number of sequences to analyze.

Clustering en Biologia Computacional

Improving ESTs: Clustering, Assembling and Gene indices

The value of ESTs is greatly enhanced by clustering and assembling.

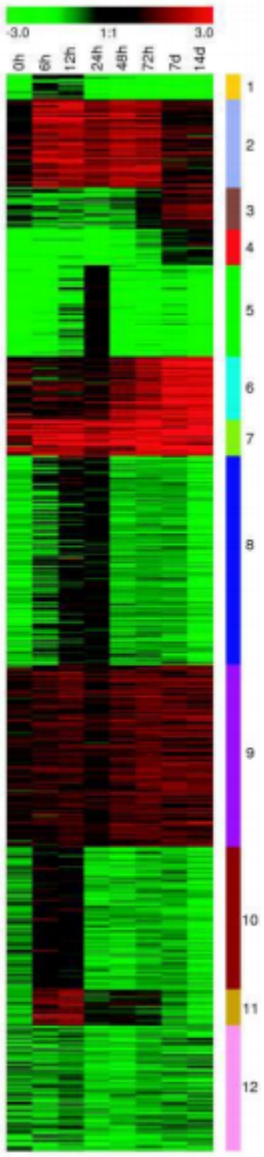
- solving redundancy can help to correct errors;
- longer and better annotated sequences;
- easier association to mRNAs and proteins;
- detection of splice variants;
- fewer sequences to analyze.

Gene indices: All expressed sequences (as ESTs) concerning a single gene are grouped in a single index class, and each index class contains the information for only one gene.

Gene Expression Omnibus - https://www.ncbi.nlm.nih.gov/geo/

GEO DataSets - https://www.ncbi.nlm.nih.gov/gds/

Clustering en Biologia Computacional



Expresión de genes a lo largo de un experimento.

No importa tanto si los genes se expresan mucho o poco (ej agrupar por nivel de expresión no tiene sentido)

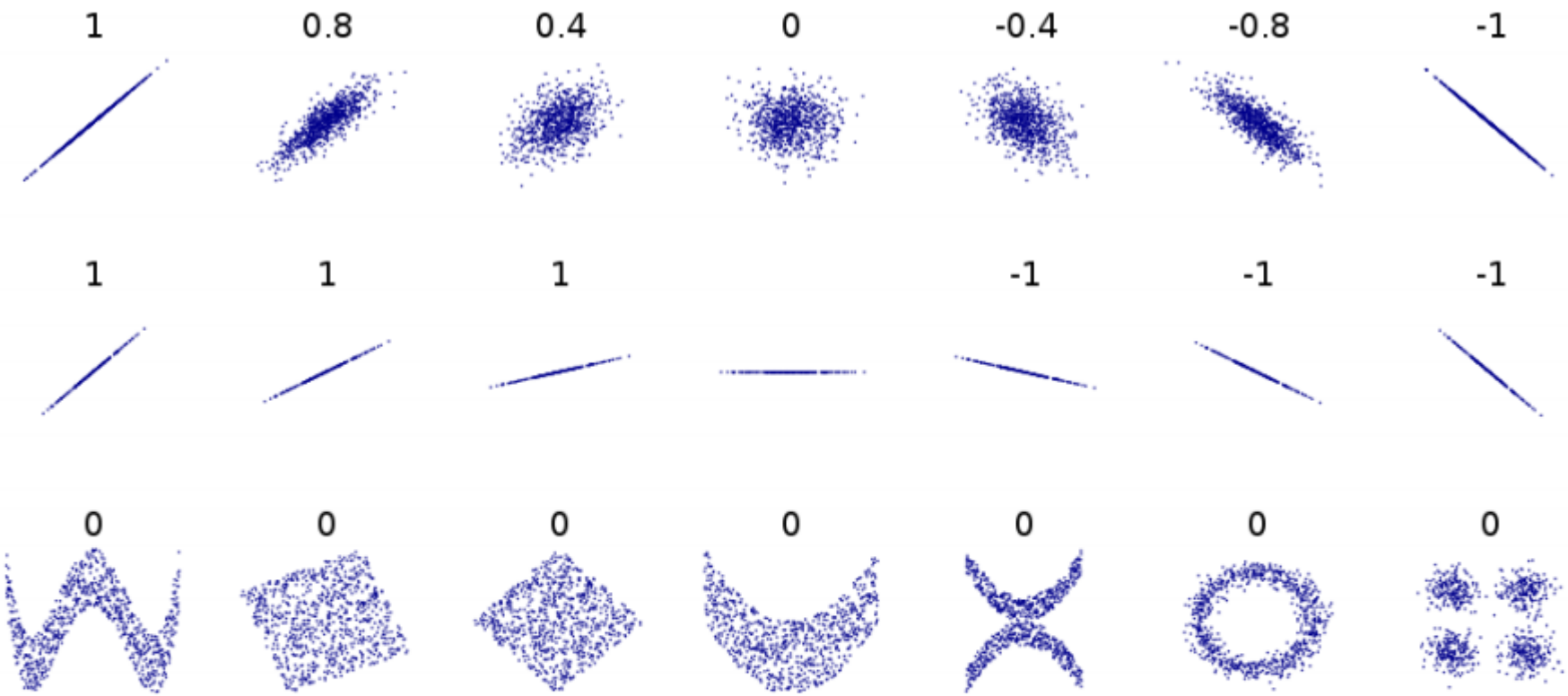
Importa el comportamiento de cada gen a lo largo de un tratamiento experimental.

Correlation distance. Mide la dependencia entre datos.

CD = 0 si los datos son independientes.
CD = 1 si tienen dependencia.

Clustering of ESTs found to be differentially expressed during fat cell differentiation. Shown is k-means clustering of 780 ESTs found to be more than twofold upregulated or downregulated at a minimum of four time points during fat cell differentiation. ESTs were grouped into 12 clusters with distinct expression profiles.
Hackl *et al. Genome Biology* 2005 **6**:R108 doi:10.1186/gb-2005-6-13-r108

Distancia de Correlación de Pearson

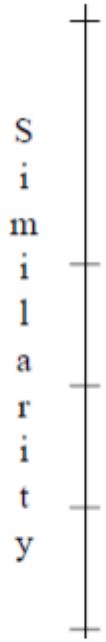


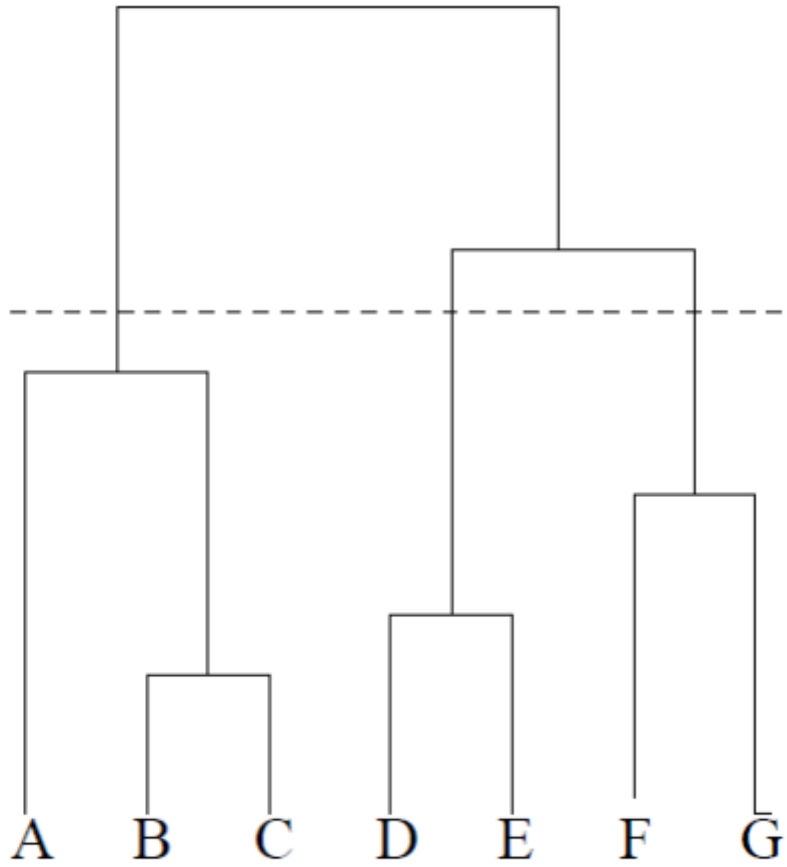
Several sets of (*x*, *y*) points, with the Pearson correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.
http://en.wikipedia.org/wiki/Correlation

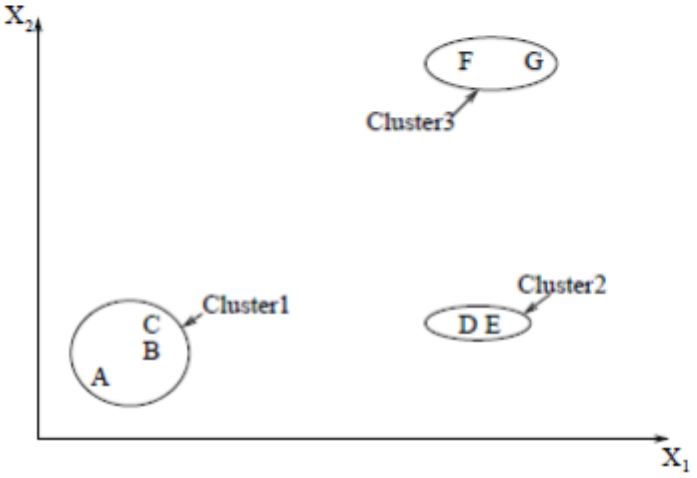
Clustering Jerárquico

Todos los algoritmos jerárquicos producen como resultado un dendograma

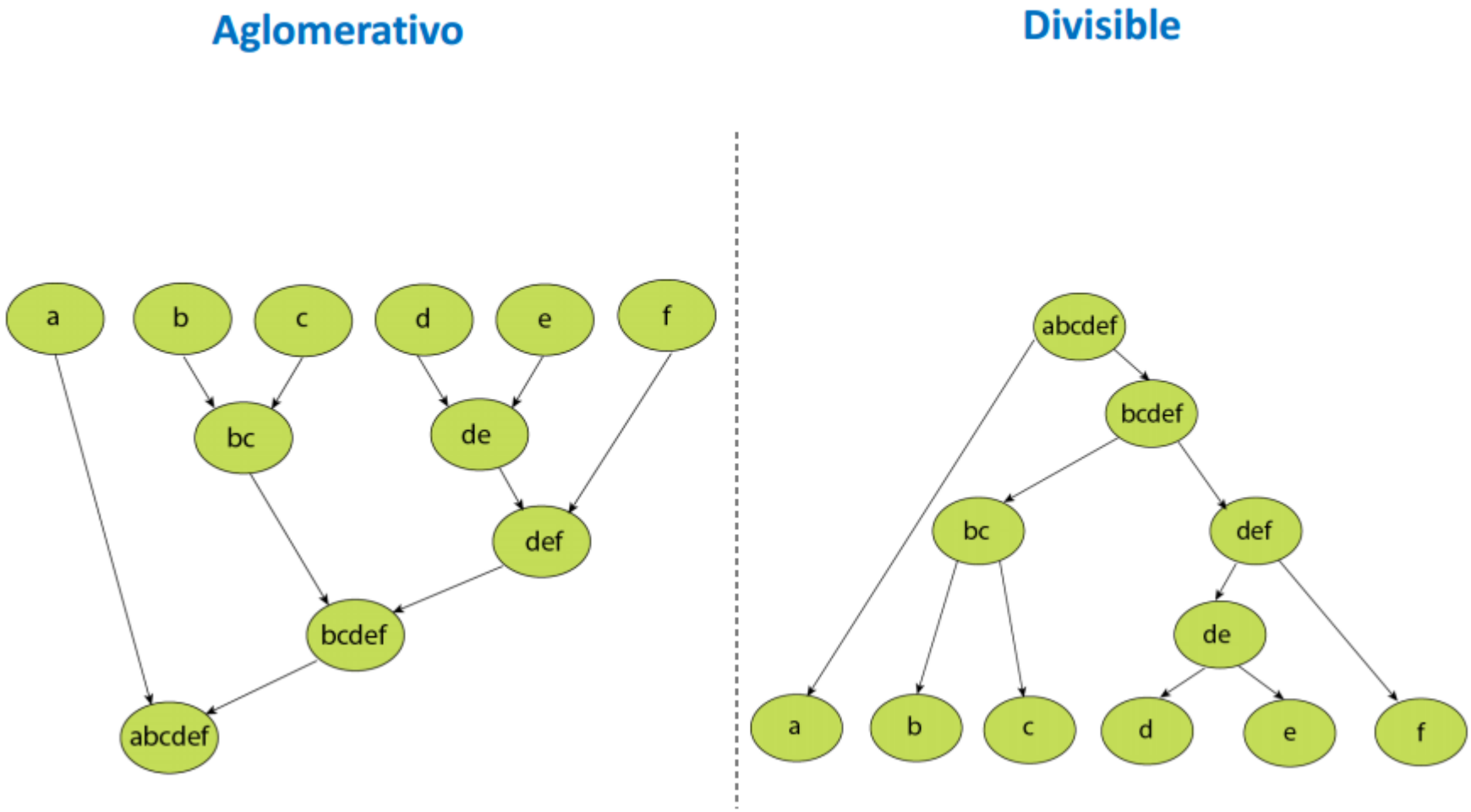
A partir del dendograma se pueden obtener varias particiones (estructuras de clusters) de los datos.





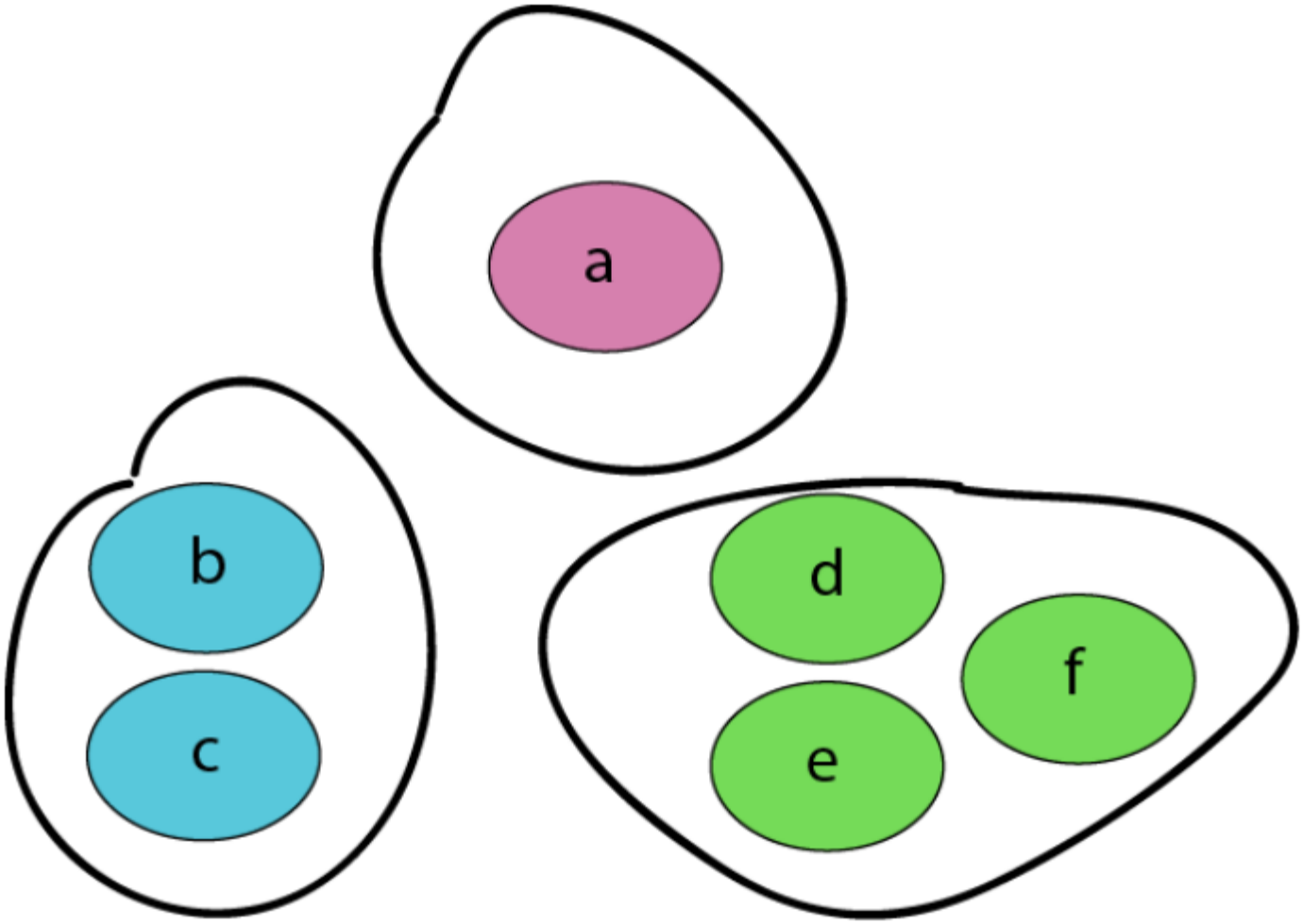


Estrategias de Clustering: Clustering Jerárquico



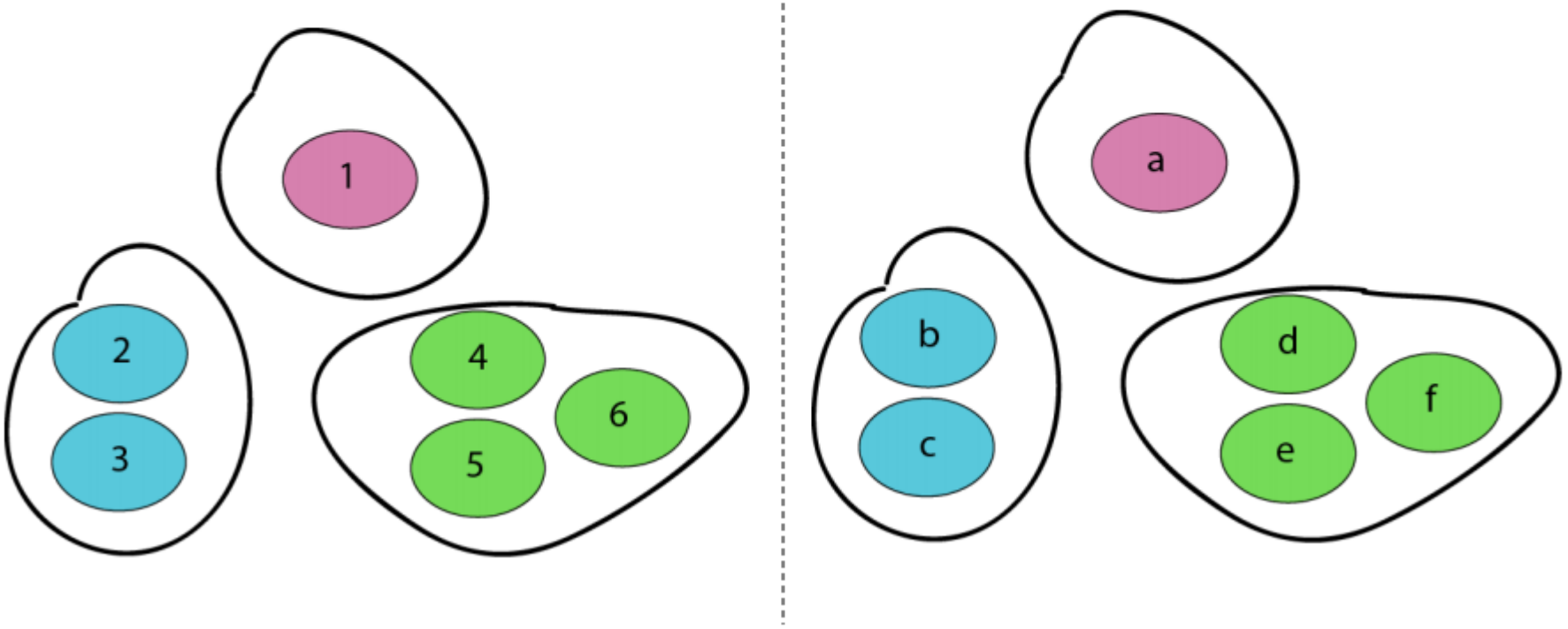
Estrategias de Clustering: Clustering Particional

A diferencia de los algoritmos jerárquicos, se obtiene ***una única partición de los datos*** (una única estructura de clusters)



Propiedades de los Clusters

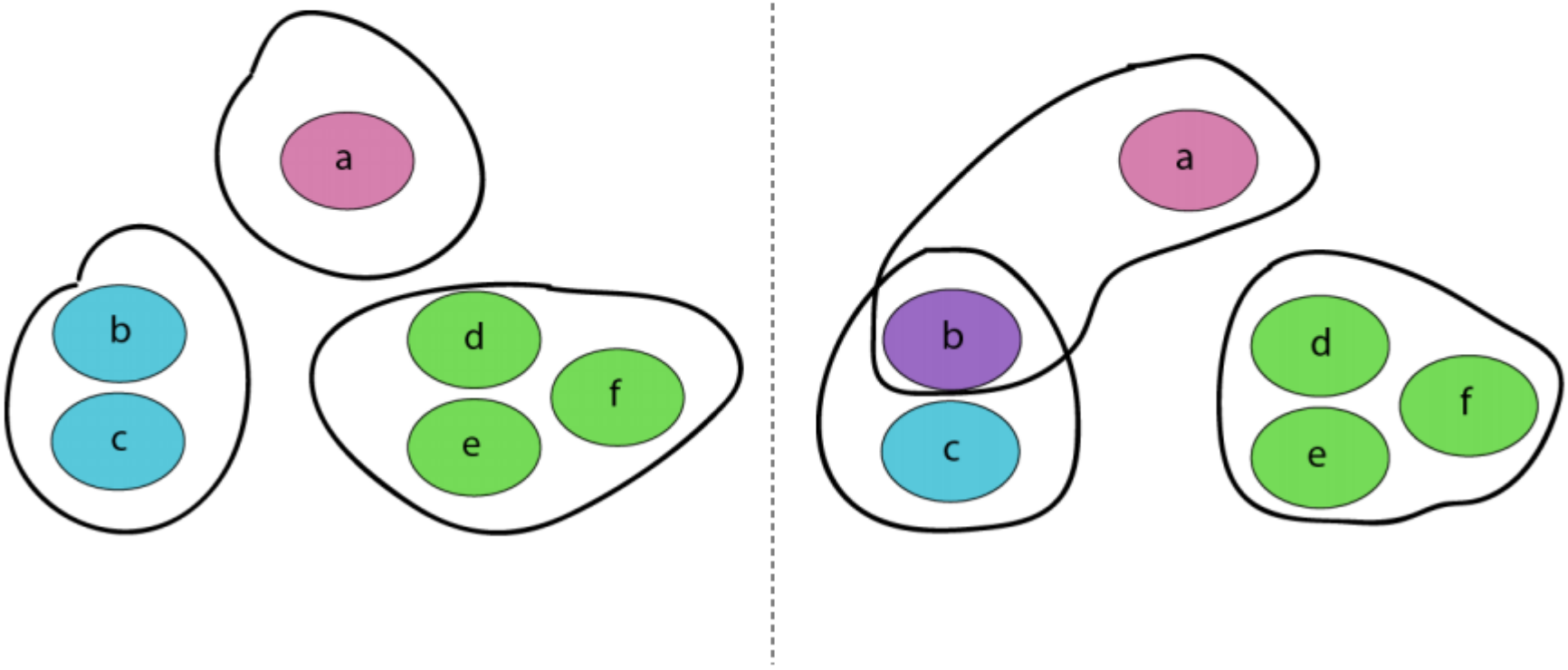
Numéricos vs. Categóricos



Propiedades de los Clusters

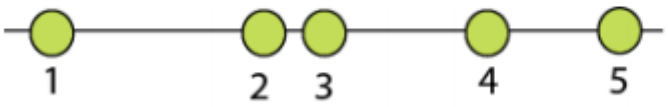
Disjuntos vs. No disjuntos

(hard) (fuzzy)



Clustering Jerárquico

Dado un conjunto de N (5) elementos a ser agrupado y una matriz de distancia (o similitud) de N x N:



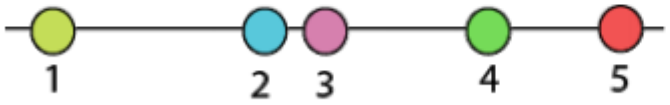
d	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

Clustering Jerárquico

Comenzar por asignar cada item a un cluster.

Tenemos 5 clusters

En este paso, las distancias entre los clusters son las mismas que entre los elementos de cada cluster

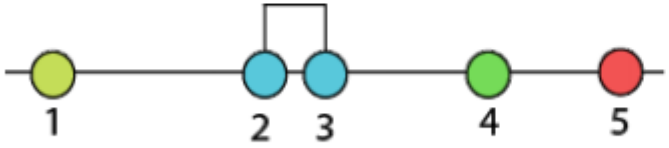


d	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

Clustering Jerárquico

Encontrar el par más cercano de clusters y unirlo en un único cluster.

Tenemos 4 clusters



d	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

Clustering Jerárquico

Calcular las distancias entre el nuevo cluster y los viejos clusters

En single-linkage la distancia que se usa es la ***mínima*** entre distintos elementos de un cluster

Los elementos se agrupan ***siempre*** encontrando la ***mínima*** distancia en la matriz

d	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

d	1	2-3	4	5
1	0	**5**	10	13
2-3	**5**	0	**4**	**7**
4	10	**4**	0	**3**
5	13	**7**	**3**	0

Clustering Jerárquico

En el algoritmo complete-linkage la distancia que se usa en la nueva matriz es la ***máxima*** entre distintos elementos de un cluster

Los elementos se agrupan ***siempre*** encontrando la ***mínima*** distancia en la matriz

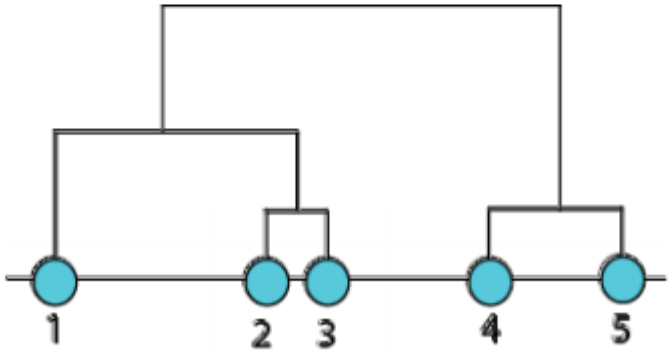
Y en average-linkage?

d	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

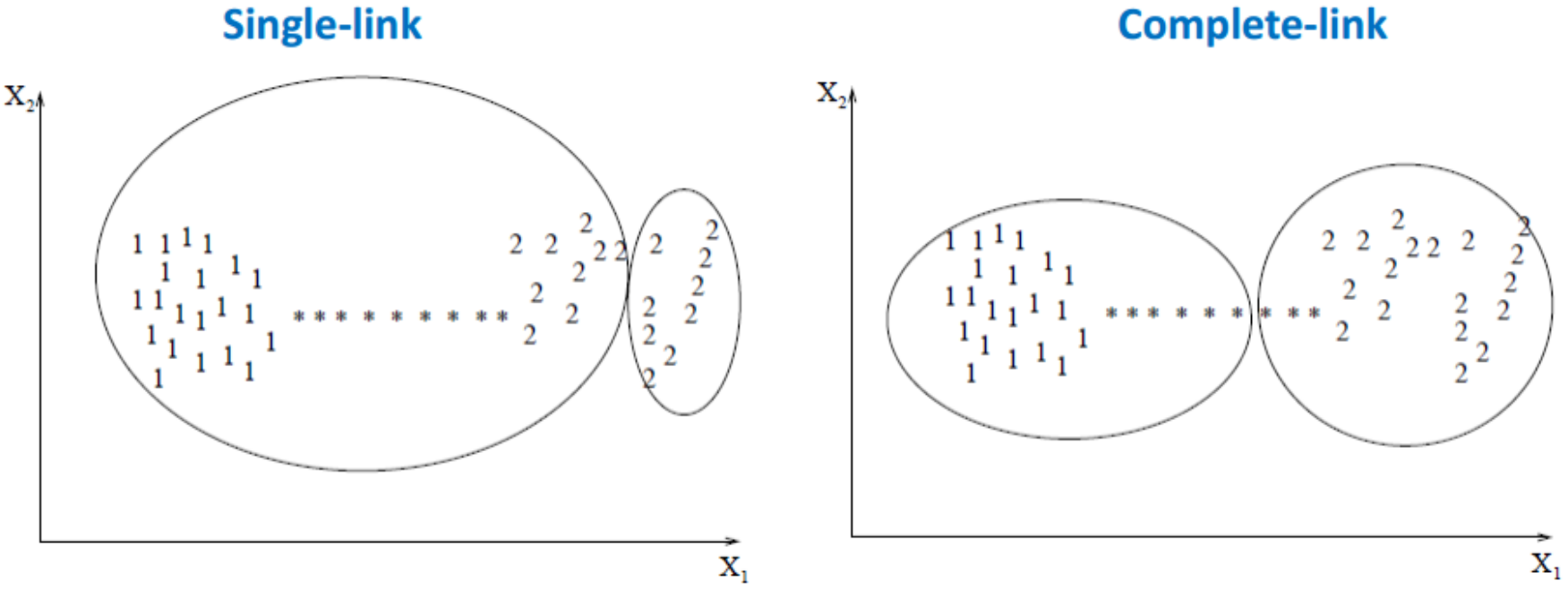
d	1	2-3	4	5
1	0	**6**	10	13
2-3	**6**	0	**5**	**8**
4	10	**5**	0	**3**
5	13	**8**	**3**	0

Single Linkage

Repetir los pasos 2 y 3 hasta que todos los elementos se encuentren en el mismo cluster de tamaño N (5)



Single Linkage vs. Complete Linkage



Ejemplo: dataset compuesto por elementos pertenecientes a dos clases, conectadas por una cadena de datos ruidosos. Tomado de Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review.

Clustering Jerárquico: Otras variantes

Hierarchical clustering techniques applied to phylogenetic reconstruction:

UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

Usado para reconstruir filogenias

Usa la media aritmética (average-link)

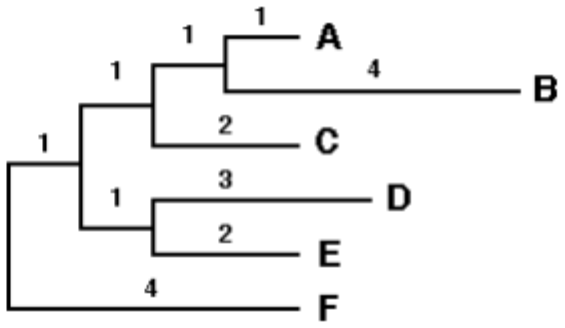
Distancias ultramétricas

Neighbor-joining

Usado para reconstruir filogenias

Usa la media aritmética

Las distancias son aditivas



Clustering Jerárquico: Otras variantes

El Algoritmo UPGMA construye un arbol con raiz (dendograma) que refleja la estructura de una matriz de similitud. En cada paso, los clusters más cercanos se combinan para formar un cluster de nivel mas alto. La distancia entre 2 clusters **A** y **B** cada uno de magnitud (Cardinalidad) | **A** | y | **B** | se toma como el promedio de todas las distancias d(x,y) entre pares de objetos x en **A** e y en **B** es decir, la distancia promedio entre elementos de cada cluster:

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y)$$

Distancia entre 2 clusters **A** y **B**

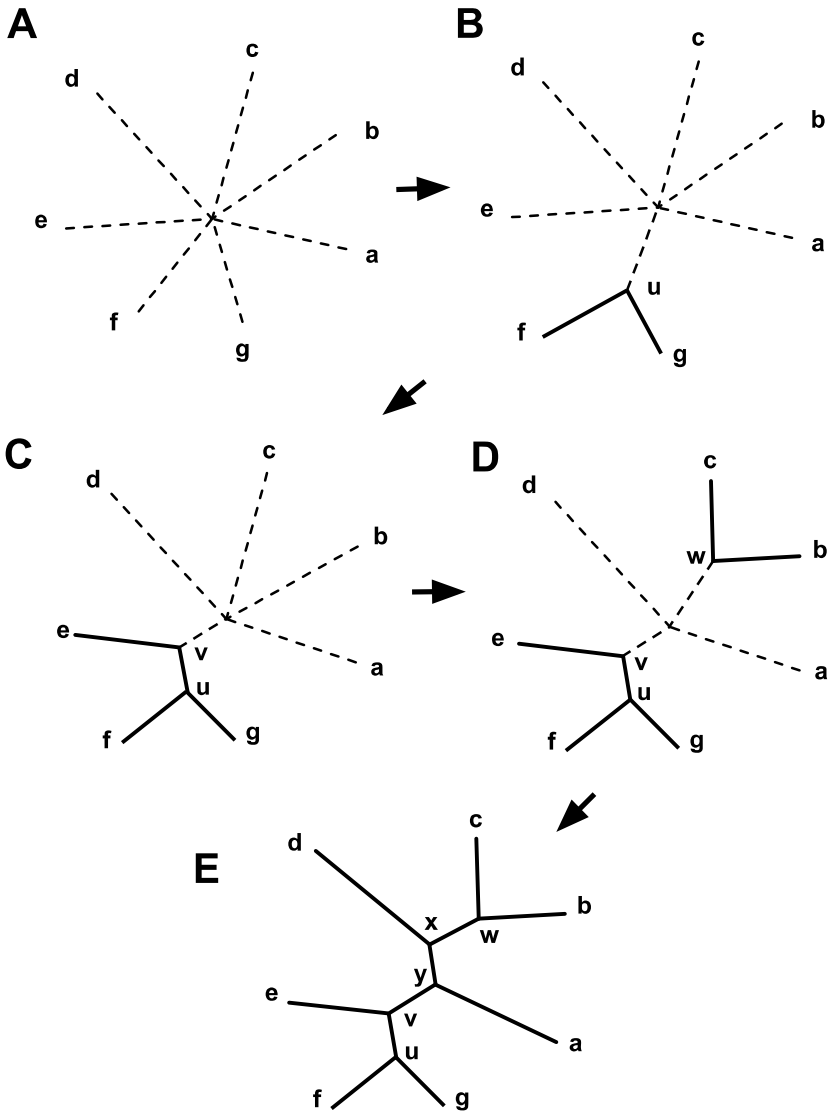
En cada paso, la distancia entre los clusters unidos $d_{(A \cup B)}$ y un nuevo cluster **X**, esta dado por el promedio proporcional de las distancias $d_{(A,X)}$ y $d_{(B,X)}$

$$d_{(\mathcal{A} \cup \mathcal{B}),X} = \frac{|\mathcal{A}| \cdot d_{\mathcal{A},X} + |\mathcal{B}| \cdot d_{\mathcal{B},X}}{|\mathcal{A}| + |\mathcal{B}|}$$

Clustering Jerárquico: Otras variantes

Neighbor Joining

1)Basados en la matriz de distancias, se calcula la matriz Q definida mas abajo.
2)Encontrar los distintos puntos i, j donde i≠j para los cuales Q(i,j) tiene el valor mas bajo. Estos puntos se unen a un nuevo nodo creado para este fin, el cual se conecta al nodo central. (En el ejemplo, f y g se unen en el nodo u)
3)Calcular la distancia de cada punto de este par al nuevo nodo.
4)Calcular la distancia de cada punto fuera de este par, al nuevo nodo.
5)Comenzar el algoritmo otra vez, reemplazando el par de vecinos unidos con el nuevo nodo, usando las distancias calculadas en el el paso previo.



Clustering Jerárquico: Otras variantes

Neighbor Joining

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

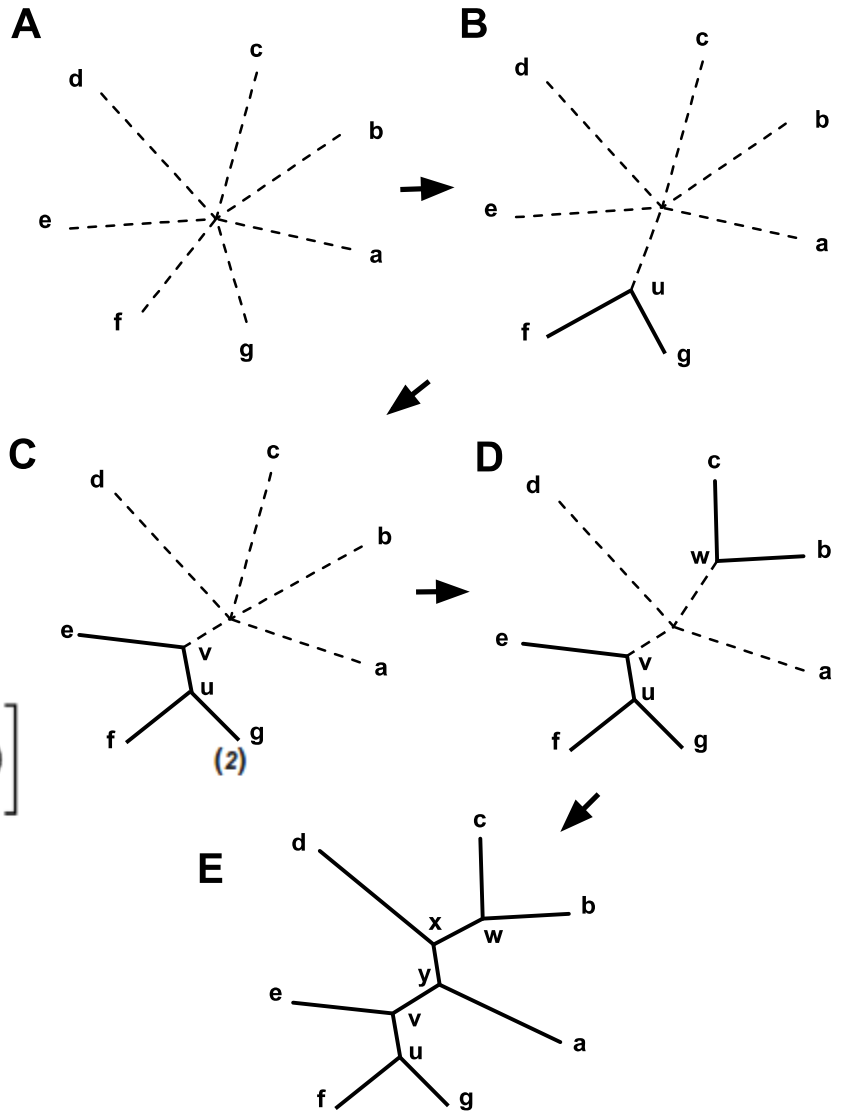
where $d(i,j)$ is the distance between taxa i and j.

Distancias de los elementos del par unido f, g al nuevo nodo u:

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

and:

$$\delta(g,u) = d(f,g) - \delta(f,u)$$



Clustering Jerárquico: Otras variantes

Neighbor Joining

Distancias de los elementos del par unido f, g al nuevo nodo u:

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

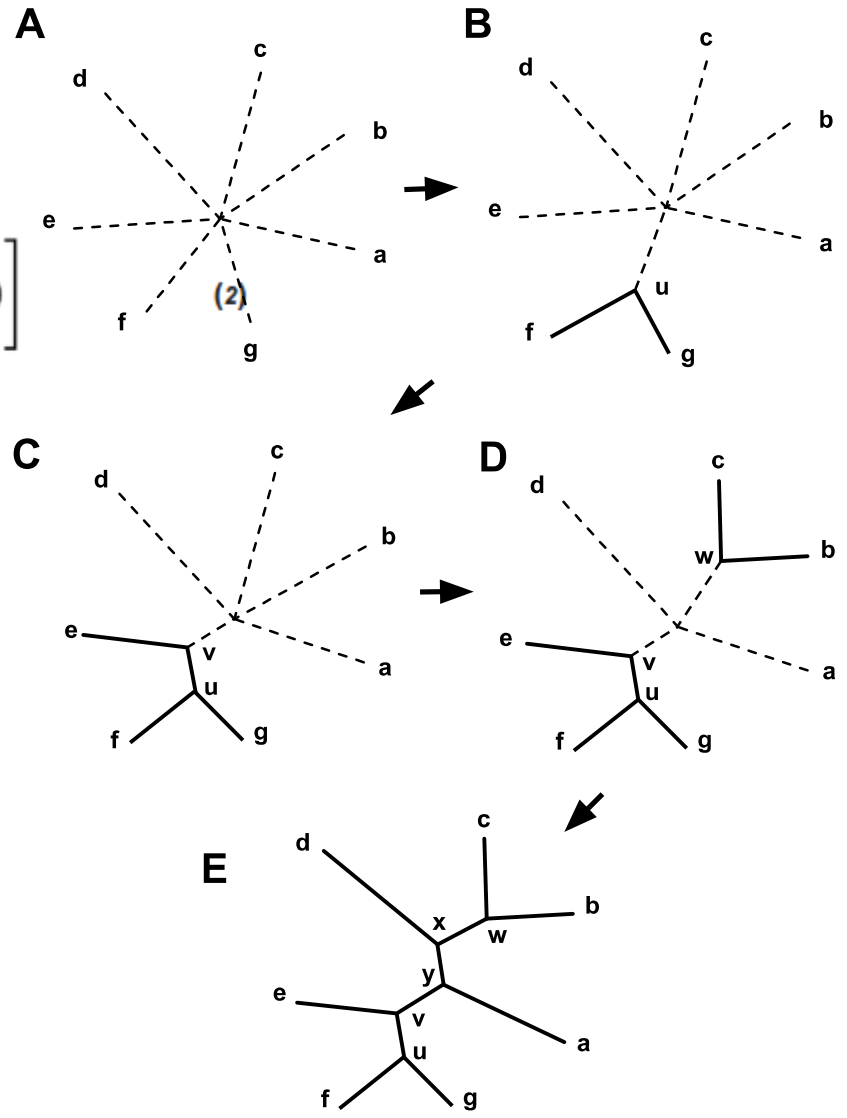
and:

$$\delta(g,u) = d(f,g) - \delta(f,u)$$

Para cada elemento que no se ha considerado en el paso previo se calcula la distancia al nuevo nodo asi:

$$d(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)]$$

Donde u es el nuevo nodo, k es el nodo al que queremos calcular la distancia



Clustering Jerárquico: Otras variantes

Neighbor Joining

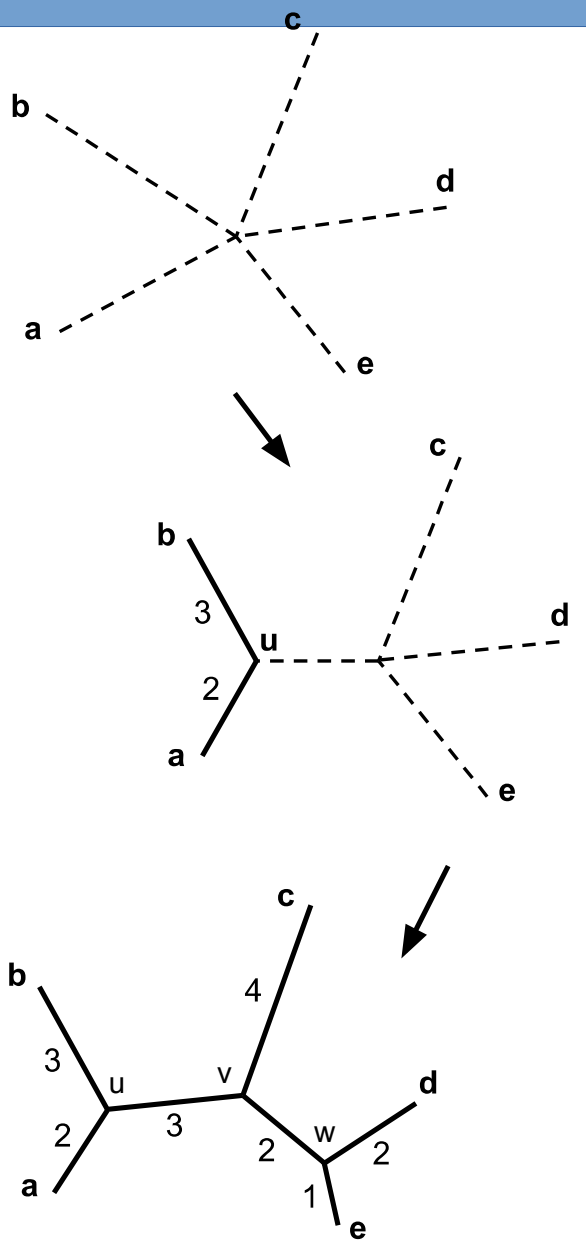
Let us assume that we have five taxa (a, b, c, d, e) and the following distance matrix D:

	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

First step [edit]

- **First joining**

We calculate the Q_1 values by equation (1). For example:



Clustering Jerárquico: Otras variantes

Neighbor Joining

	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

$$Q_1(a,b) = (n-2)d(a,b) - \sum_{k=1}^{5} d(a,k) - \sum_{k=1}^{5} d(b,k)$$

$$= (5-2) \times 5 - (5+9+9+8) - (5+10+10+9) = 15 - 31 - 34 = -50$$

We obtain the following values for the Q_1 matrix (the diagonal elements of the matrix are not used and are omitted here):

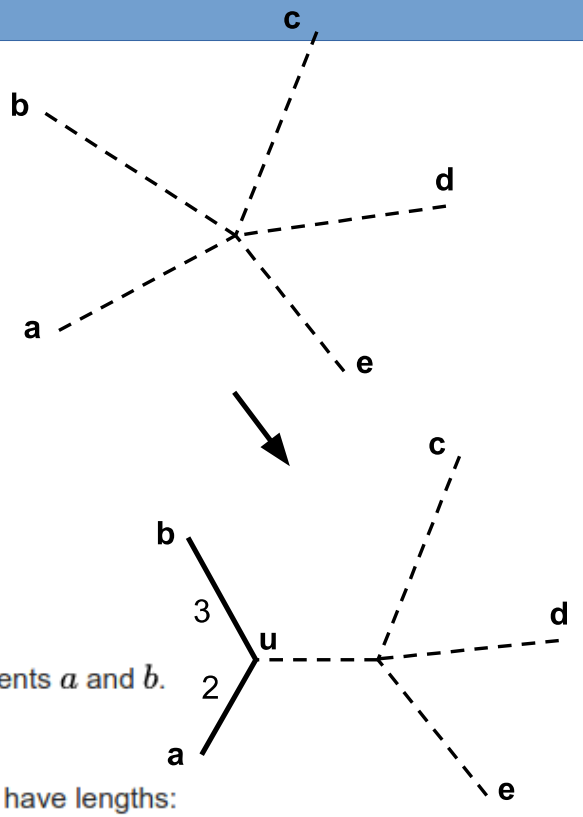
	a	b	c	d	e
a		−50	−38	−34	−34
b	−50		−38	−34	−34
c	−38	−38		−40	−40
d	−34	−34	−40		−48
e	−34	−34	−40	−48	

In the example above, $Q_1(a,b) = -50$. This is the smallest value of Q_1, so we join elements a and b.

Clustering Jerárquico: Otras variantes

Neighbor Joining

	a	b	c	d	e
a		−50	−38	−34	−34
b	−50		−38	−34	−34
c	−38	−38		−40	−40
d	−34	−34	−40		−48
e	−34	−34	−40	−48	



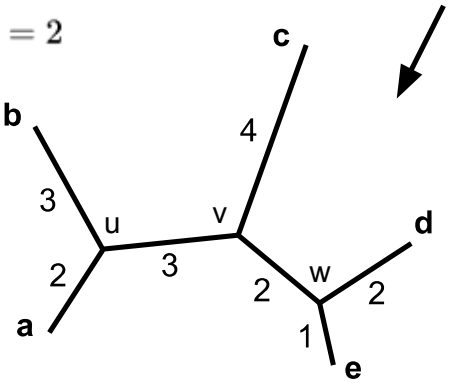
In the example above, $Q_1(a,b) = -50$. This is the smallest value of Q_1, so we join elements a and b.

- **First branch length estimation**

Let u denote the new node. By equation (2), above, the branches joining a and b to u then have lengths:

$$\delta(a,u) = \frac{1}{2}d(a,b) + \frac{1}{2(5-2)}\left[\sum_{k=1}^{5} d(a,k) - \sum_{k=1}^{5} d(b,k)\right] = \frac{5}{2} + \frac{31-34}{6} = 2$$

$$\delta(b,u) = d(a,b) - \delta(a,u) = 5 - 2 = 3$$



Clustering Jerárquico: Otras variantes

Neighbor Joining

- **First distance matrix update**

We then proceed to update the initial distance matrix D into a new distance matrix D_1 (see below), reduced in size by one row and one column because of the joining of a with b into their neighbor u. Using equation (3) above, we compute the distance from u to each of the other nodes besides a and b. In this case, we obtain:

$$d(u,c) = \frac{1}{2}[d(a,c) + d(b,c) - d(a,b)] = \frac{9 + 10 - 5}{2} = 7$$
$$d(u,d) = \frac{1}{2}[d(a,d) + d(b,d) - d(a,b)] = \frac{9 + 10 - 5}{2} = 7$$
$$d(u,e) = \frac{1}{2}[d(a,e) + d(b,e) - d(a,b)] = \frac{8 + 9 - 5}{2} = 6$$

The resulting distance matrix D_1 is:

	u	c	d	e
u	0	**7**	**7**	**6**
c	**7**	0	*8*	*7*
d	**7**	*8*	0	*3*
e	**6**	*7*	*3*	0

Bold values in D_1 correspond to the newly calculated distances, whereas italicized values are not affected by the matrix update as they correspond to distances between elements not involved in the first joining of taxa.

Clustering Jerárquico: Otras variantes

Neighbor Joining

Second step [edit]

- **Second joining**

The corresponding Q_2 matrix is:

	u	c	d	e
u		−28	−24	−24
c	−28		−24	−24
d	−24	−24		−28
e	−24	−24	−28	

We may choose either to join u and c, or to join d and e; both pairs have the minimal Q_2 value of -28, and either choice leads to the same result. For concreteness, let us join u and c and call the new node v.

- **Second branch length estimation**

The lengths of the branches joining u and c to v can be calculated:

$$\delta(u,v) = \frac{1}{2}d(u,c) + \frac{1}{2(4-2)}\left[\sum_{k=1}^{4} d(u,k) - \sum_{k=1}^{4} d(c,k)\right] \quad = \frac{7}{2} + \frac{20-22}{4} = 3$$

$$\delta(v,c) = d(u,c) - \delta(u,v) \quad = 7 - 3 = 4$$

Clustering Jerárquico: Otras variantes

Neighbor Joining

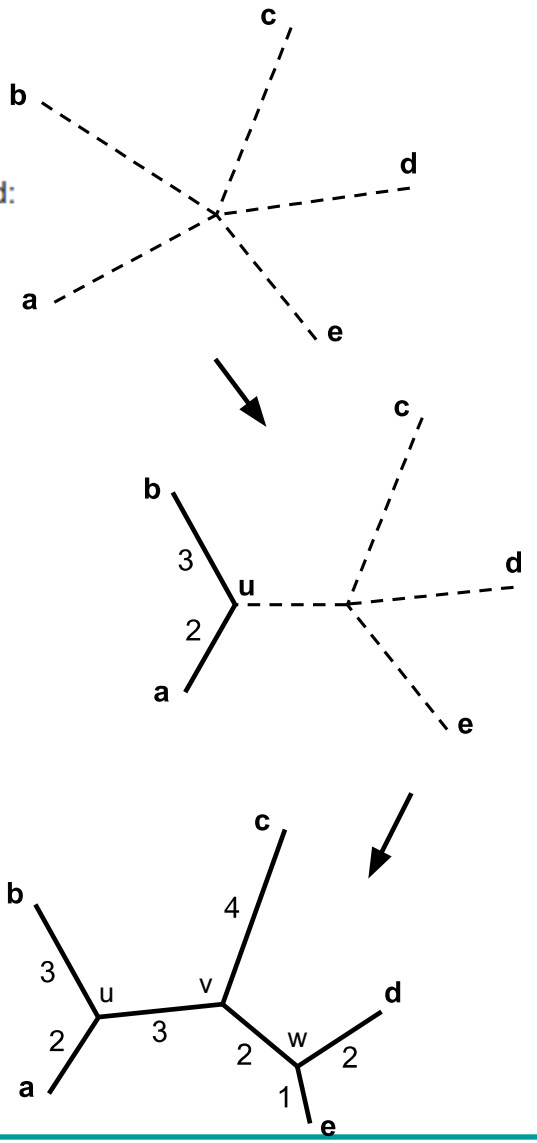
- **Second distance matrix update**

The updated distance matrix D_2 for the remaining 3 nodes, v, d, and e, is now computed:

$$d(v,d) = \frac{1}{2}[d(u,d) + d(c,d) - d(u,c)] = \frac{7+8-7}{2} = 4$$

$$d(v,e) = \frac{1}{2}[d(u,e) + d(c,e) - d(u,c)] = \frac{6+7-7}{2} = 3$$

	v	d	e
v	0	**4**	**3**
d	**4**	0	*3*
e	**3**	*3*	0



Clustering Jerárquico: Otras variantes

Neighbor Joining

Final step [edit]

The tree topology is fully resolved at this point. However, for clarity, we can calculate the Q_3 matrix. For example:

$$Q_3(v,e) = (3-2)d(v,e) - \sum_{k=1}^{3} d(v,k) - \sum_{k=1}^{3} d(e,k) = 3 - 7 - 6 = -10$$

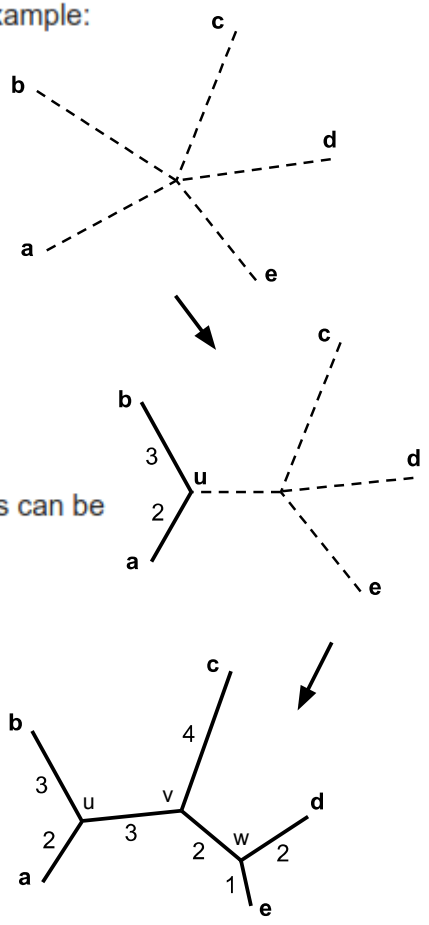
	v	d	e
v		−10	−10
d	−10		−10
e	−10	−10	

For concreteness, let us join v and d and call the last node w. The lengths of the three remaining branches can be calculated:

$$\delta(v,w) = \frac{1}{2}d(v,d) + \frac{1}{2(3-2)}\left[\sum_{k=1}^{3} d(v,k) - \sum_{k=1}^{3} d(d,k)\right] \quad = \frac{4}{2} + \frac{7-7}{2} = 2$$

$$\delta(w,d) = d(v,d) - \delta(v,w) = 4 - 2 = 2$$

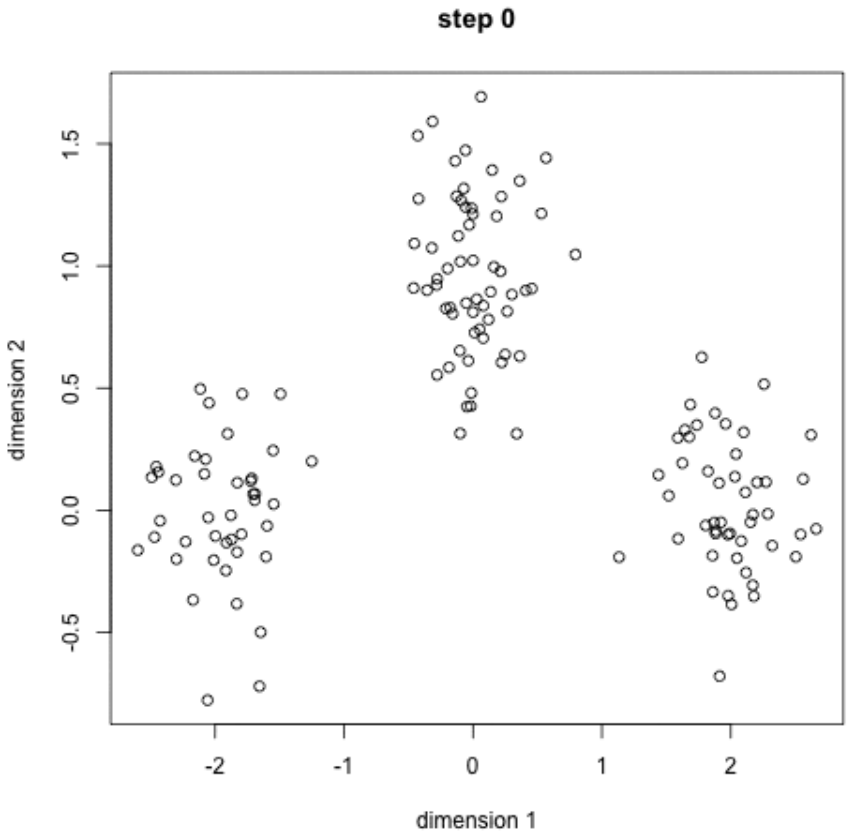
$$\delta(w,e) = d(v,e) - \delta(v,w) = 3 - 2 = 1$$



Clustering No Jerárquico

1. Los Métodos No Jerarquicos están diseñados para clasificar individuos en K Clusters, donde K se define a priori.

2. La idea central de estos métodos es elegir una particion inicial de individuos y después intercambiar los miembros de estos clusters para obtener una mejor particion.
3. Los algoritmos difieren entre sí en como llegar a obtener la mejor particion y los metodos que se utilizan para ello.



Clustering No jerarquico: K-means

Es muy rapido!

Particional

Usa Distancia euclídea

Necesita el valor de *k* (Nro de clusters)

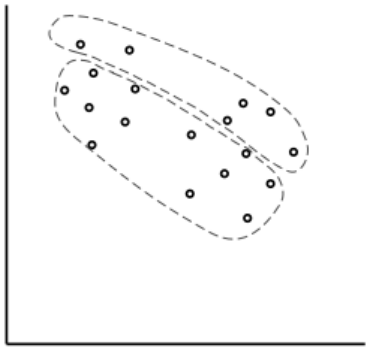
Util para búsqueda de prototipos

Sensible a outliers

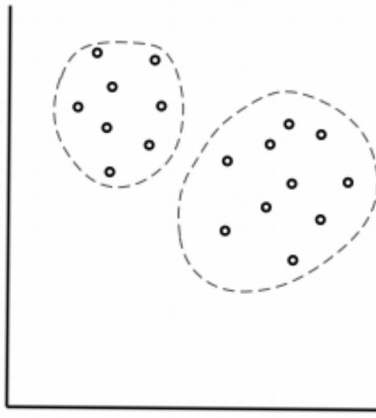
Clustering No jerarquico: K-means

1. Se puede considerar a las n filas de una matriz de expresion de m columnas
(n x m) como un conjunto de n puntos en un espacio m – dimensional.

2. en este método, el problema consiste en encontrar un conjunto de k puntos o centroides en un espacio m – dimensional que minimiza el error de distorsion cuadrática.



K=2



Clustering No jerarquico: K-means

1. Dado un data point ***v*** .
2. y Dado un conjunto de **k** centroides ***X*** = { x_1, …., x_k}
3. se define la distancia de ***v*** al conjunto de los centroides ***X*** , como la distancia de ***v*** al punto (centroide) mas cercano en ***X***

Es decir:

$$d(v, X) = \min_{1 < i < k} d(v, x_i)$$

Donde $d(v, x_i)$ es la distancia Euclidiana en m dimensiones

$$D(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

Distancia euclidiana para d dimensiones

Clustering No jerarquico: K-means

1. El error de distorsion cuadrática para un set de puntos $\boldsymbol{v} = \{ \boldsymbol{v}_1, \ldots \boldsymbol{v}_n \}$.

y un conjunto de **k** centroides $\boldsymbol{X} = \{ x_1, \ldots, x_k \}$

Esta definido como el promedio del cuadrado de la distancia de cada punto del set de datos a su centroide mas cercano

$$d(\boldsymbol{v}, \boldsymbol{X}) = \frac{\sum_{i=1}^{n} d(\boldsymbol{v}_i, \boldsymbol{X})^2}{n}$$

Entrada: $\boldsymbol{v}$

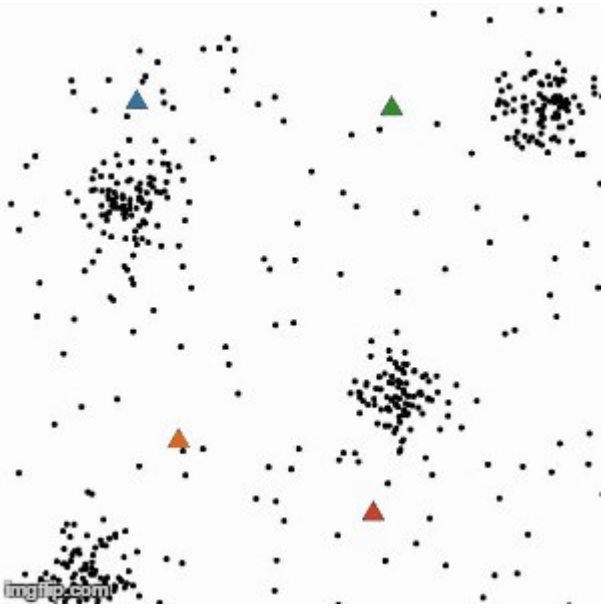
Salida : $\boldsymbol{X}$

Condicion: minimizar $d(\boldsymbol{v}, \boldsymbol{X})$ en el conjunto de todos los $\boldsymbol{X}$ posibles

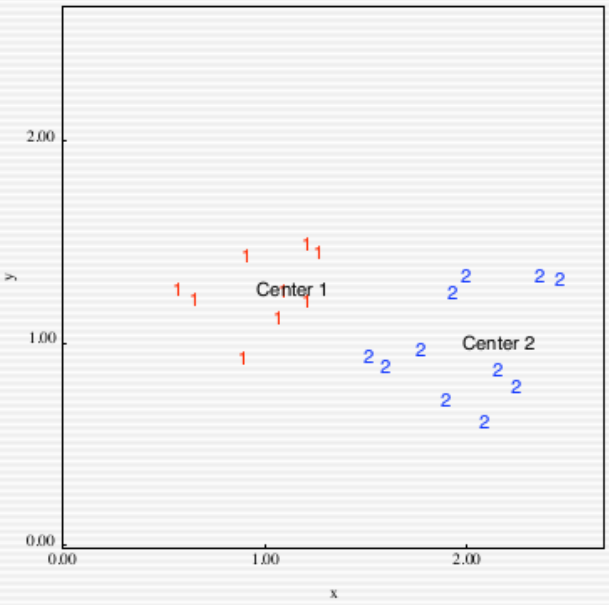
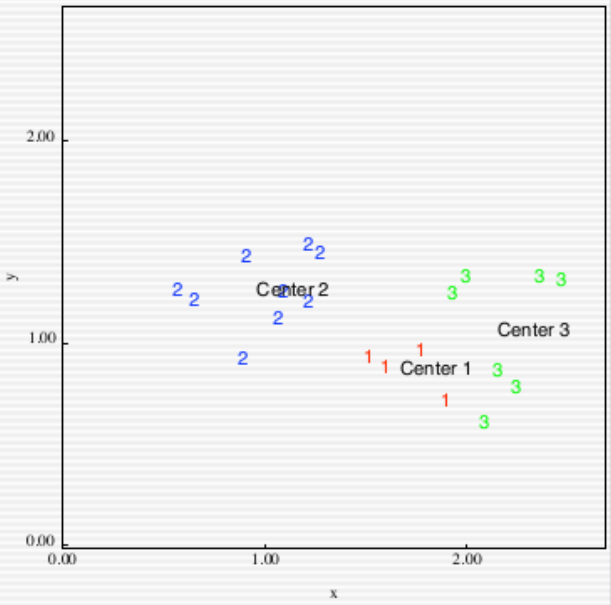
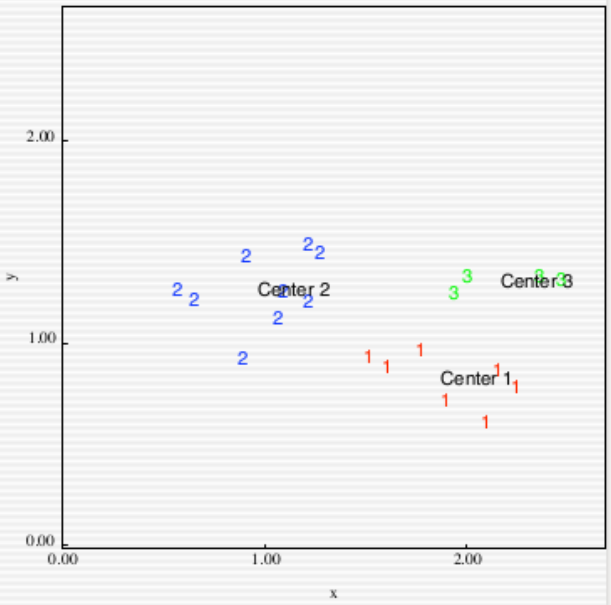
Clustering No jerarquico: K-means

Este problema no es un problema que se pueda resolver con algoritmos eficientes (Polinomiales), cuando el numero K es mayor o igual a 2, por lo que se utilizan variantes que introducen algunas heuristicas, como por ejemplo: **El algoritmo de Lloyd:**

1. Asignar abitrariamente los K centros de clusters.
2. Mientras los centros de clusters van cambiando:
A. Calcular la distancia de cada punto al centro del cluster actual X_i, para $1 < i < k$ y asignar el punto al cluster más cercano.
B. Despues de asignar todos los puntos a sus clusters, calcular nuevos centros para cada cluster tomando el centroide de todos los puntos en ese cluster.
3. Entregar los centros de clusters y las asignaciones.



Clustering No jerarquico: K-means

K=2	K=3	K=3 (different starting points)
		

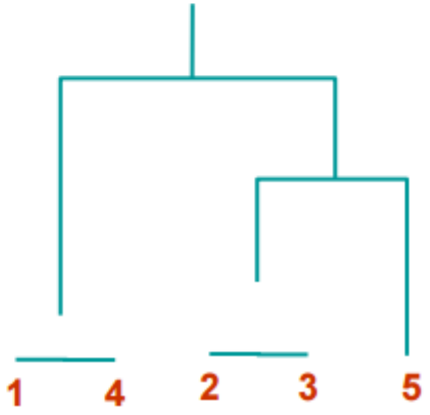
Filogenia:

Filogenia
Reconstrucción filogenética
Inferencia de filogenias

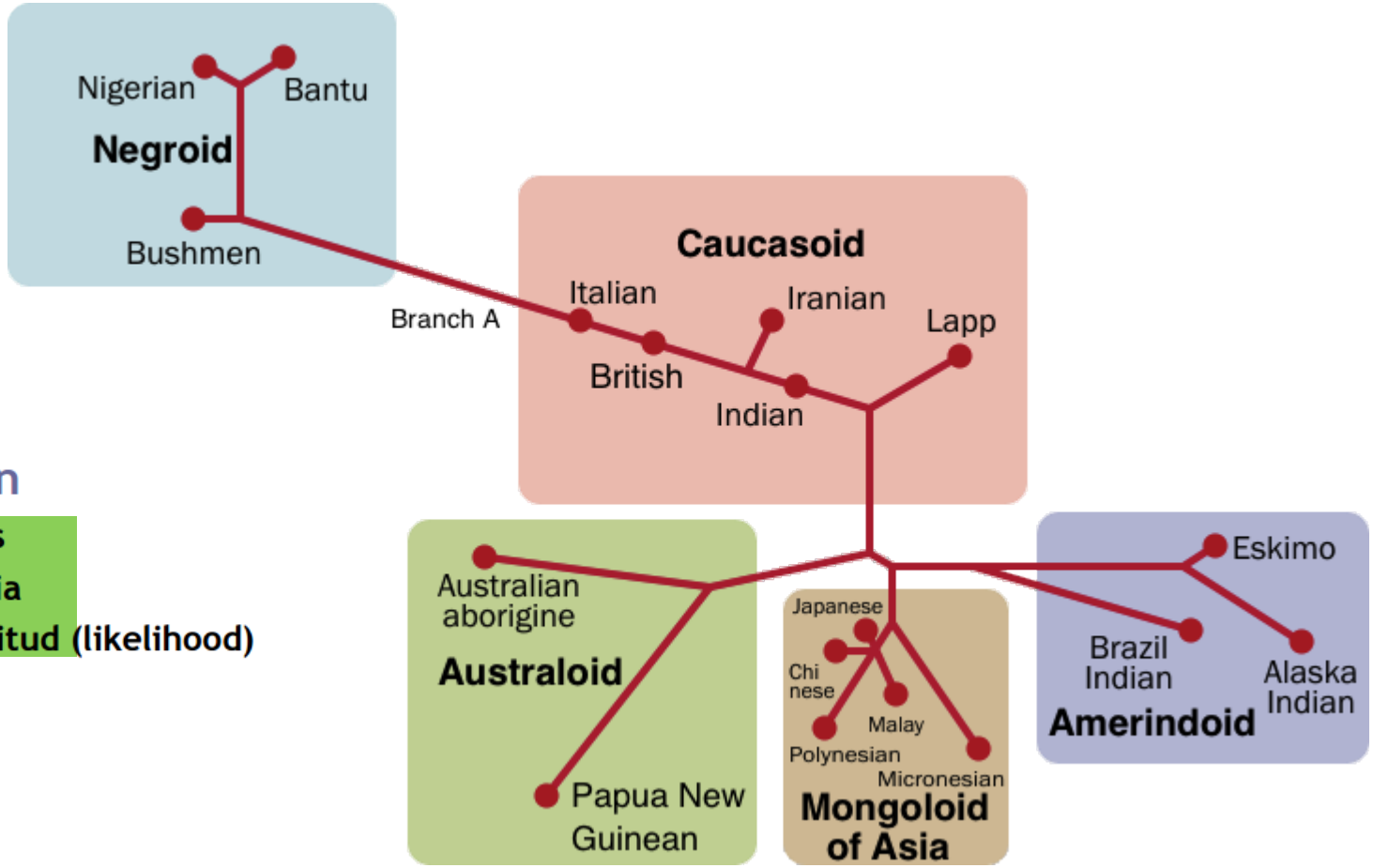
Filogenia:

- Una filogenia es un árbol que describe la secuencia de eventos que llevó a producir los caracteres que observamos en la actualidad
- Es una hipótesis!
- Los eventos pasados son desconocidos. Se *infieren*

- Un árbol es un grafo
  - Nodos y ejes
- En particular:
  - Los nodos exteriores (hojas del árbol) son los eventos observados (especies actuales)
  - Los nodos internos son los eventos (ancestros) postulados
  - La longitud de los ejes (ramas) representa el tiempo de evolucion entre nodos



Filogenia: Metodos de reconstrucción



- Basados en
  - Distancias
  - Parsimonia
  - Verosimilitud (likelihood)

Filogenia: Metodos de reconstrucción

- Cómo inferir la filogenia?
  - Definir los caracteres a seguir
  - Construir una matriz de distancias
  - Seleccionar un algoritmo para reconstruir la filogenia a partir de los datos de distancias

- Caracteres y estados
  - Los caracteres deben evolucionar en forma independiente
  - Los estados observados comparten un origen común

los caracteres se describen generalmente en términos de sus estados, por ejemplo: "cabello presente" vs. "cabello ausente", donde "cabello" es el personaje y "presente" y "ausente" son sus estados

Para secuencias de ADN un caracter corresponde a una posición en la secuencia y los estados posibles, son los nucleótidos A, T, C, G.

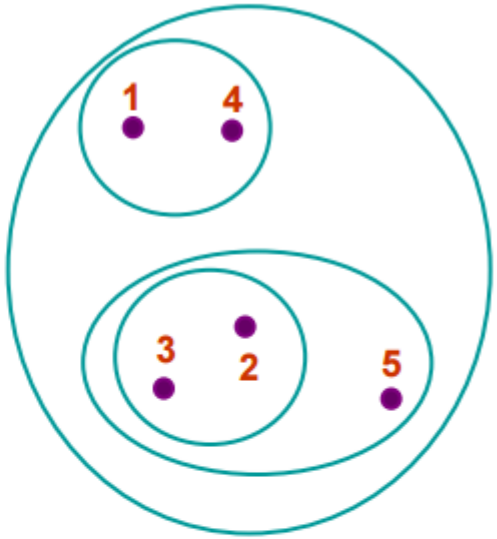
Filogenia: Tipos de caracteres

MORFOLÓGICOS Medidas Corporales Medidas Parciales Presencia de estructuras	**MOLECULARES** Hibridación DNA-DNA RFLP Secuencias (DNA ó Proteínas)
CONTINUOS Medidas Corporales Medidas Parciales Hibridación de DNA-DNA	**DISCRETOS** Presencia de estructuras RFLP Secuencias (DNA ó Proteínas)

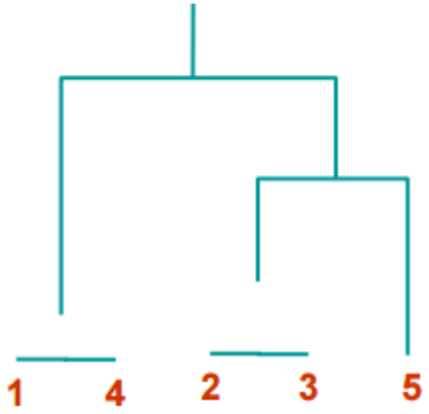
RLFP- Restriction Fragment Length Polymorphism- se refiere a secuencias específicas de nucleótidos en el ADN que son reconocidas y cortadas por las enzimas de restricción

Filogenia: Metodos de reconstrucción – Basados en distancias

- Los pares de secuencias más cercanos (neighbors) comparten un ancestro común y están unidos a él por ramas
- El objetivo del método es encontrar un árbol que acomode a todos los *vecinos* correctamente
- El largo de las ramas tiene que concordar con los datos de distancia



- Usan métodos de *clustering* para agrupar *vecinos*



Filogenia: Distintos tipos de distancias

- **Aditivas**
  - La suma de las longitudes de las ramas de dos especies con su nodo ancestral es igual a la distancia calculada entre las especies

	Sp. 1	Sp. 2	Sp. 3	Sp. 4
Sp. 1	-			
Sp. 2	3	-		
Sp. 3	7	6	-	
Sp. 4	8	7	3	-

- **Ultramétricas**
  - Cada ancestro común está equidistante de sus descendientes
  - Util para visualizar similitud en contextos no evolutivos

Filogenia: Metodos de reconstrucción Maxima Parsimonia

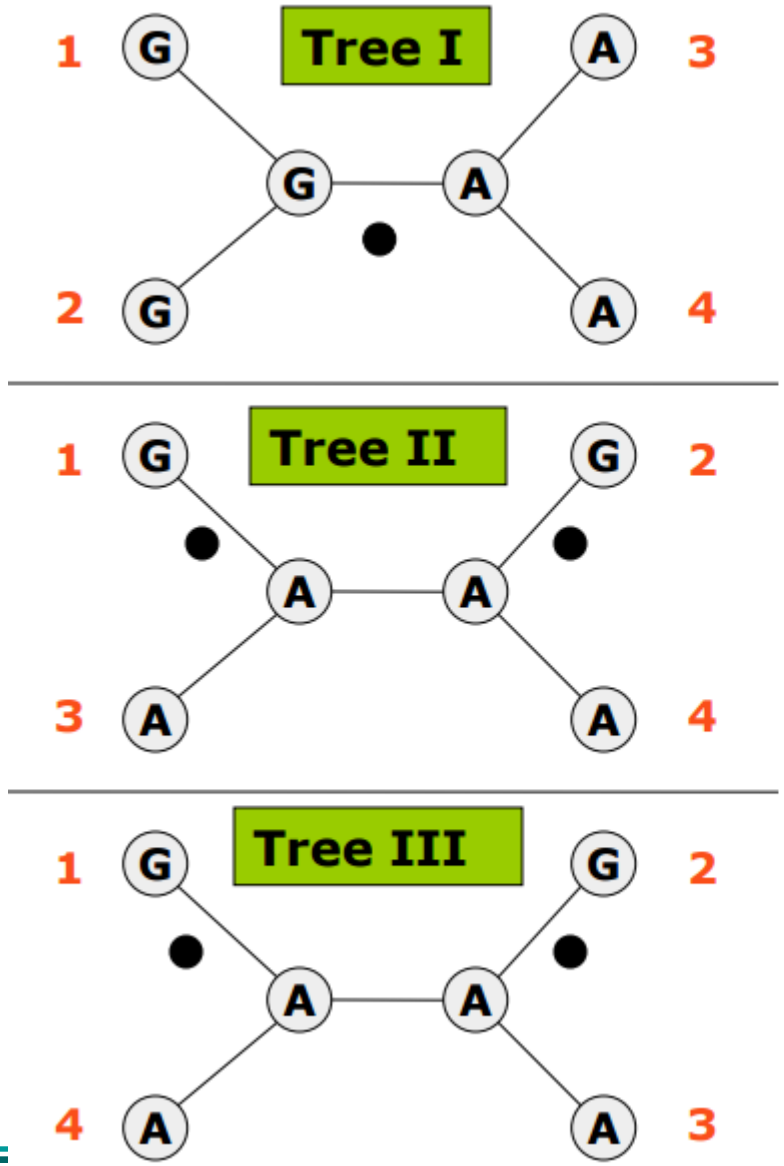
- Predicen el árbol (o árboles) que minimizan el número de cambios (o pasos) que es necesario hacer para generar la variación observada entre las secuencias
- También conocido como *método de evolución mínima*

	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

Especies

- Ejemplo
  - Para ser informativo un sitio debe tener dos estados presentes en al menos dos especies
  - Sitios no informativos: 1, 2, 3, 4, 6 y 8
  - Sitios informativos: 5, 7 y 9
  - Sólo se analizan los sitios informativos

Filogenia: Metodos de reconstrucción Maxima Parsimonia



- Hay 3 árboles posibles (sin raíz) para describir la evolución de 4 especies
- Menor número de cambios para explicar la evolución: árbol 1 (1 cambio)
- El mismo análisis se repite para cada uno de los sitios informativos
- El resultado es el árbol que provee el menor número de pasos para acomodar los datos en los sitios informativos (el más parsimonioso)

Pasar de A -> G genera un cambio

Filogenia: Metodos de reconstrucción Maxima Parsimonia

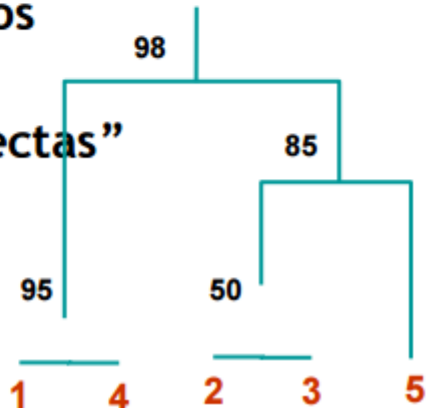
- Asume que la velocidad de evolución es similar en todas las ramas
  - La inferencia obviamente falla cuando esto no se cumple
  - Ejemplo: cambio de G a A en forma independiente en dos especies
    - Especie 1: G > A
    - Especie 2: G > C > T > G > C > A
- Se pueden asignar puntajes a los árboles
  - En lugar de contar cambios se pueden asignar distintos valores a los cambios (por ejemplo usando una matriz)
- A diferencia de los métodos de distancia, el método permite obtener la secuencia postulada de cualquier ancestro

Filogenia: Metodos de reconstrucción Maximum Likelihood

- Maximum likelihood
  - Similar al método de máxima parsimonia: usa todas las columnas del alineamiento, considera todos los árboles posibles
  - Usa probabilidades

Filogenia: Metodos de Testeo

- **Bootstrap test**
  - Bootstrap resampling technique (Efron 1982)
  - Dado un número de secuencias M de longitud N (un alineamiento), y un árbol calculado por un método cualquiera, se genera un nuevo set de secuencias M' en el cual N' bases/residuos elegidos al azar son reemplazados, también al azar.
  - En base a este nuevo set M' se recalcula el árbol utiilizando el mismo método y se comparan las topologías del árbol.
  - Esto se repite varias veces (100, 1000 repeticiones) y se calcula, para cada rama un valor de bootstrap
  - Bootstrap value: % de veces que la rama aparece en los distintos árboles
  - Bootstrap values >= 95% corresponden a ramas "correctas"



Filogenia: Metodos de Testeo

- Jacknife
  - Muy similar al test de bootstrapping
  - Se generan nuevos data sets por muestreo parcial del original
  - Usualmente se muestrea el 50% de los datos originales
  - Se rehacen los árboles y se verifica la topología
  - Se hacen varios re-muestreos (100-1000 veces)
  - Se construye un árbol consenso con valores de confidencia para cada rama