

TÉCNICAS DE COMPUTACIÓN EN BIOINFORMÁTICA

Profesor: Mario Inostroza Ponta
Ayudante: Jorge Párraga-Álava

- ❑ Datos generados por toda la comunidad científica deben ser de alguna manera ordenados y controlados: **NCBI**, **PubMed**, etc.
- ❑ Se requieren herramientas para extraer información útil de datos producidos por técnicas biológicas de alta productividad.
- ❑ Principales técnicas de computación aplicadas en la bioinformática incluyen:
 - ✓ Alineamiento de secuencias
 - ✓ Análisis de datos de expresión génica
 - ✓ Predicción de estructura de proteínas
 - ✓ Interacciones proteína-proteína, etc.

Técnicas para alineamiento de secuencias

¿En qué consiste?

En comparar dos o más secuencias de ADN, o ARN, o estructuras primarias protéicas para determinar:

- ✓ Si evolucionaron desde un ancestro común
- ✓ Si tienen funciones compartidas
- ✓ Y en el caso de proteínas, si tienen formas similares.

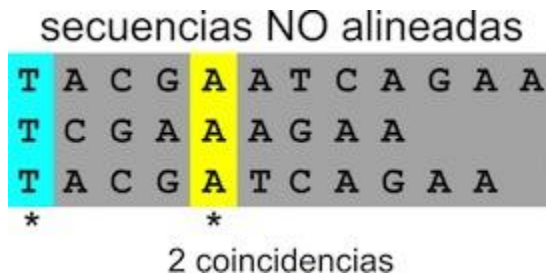
¿Cómo se representan?

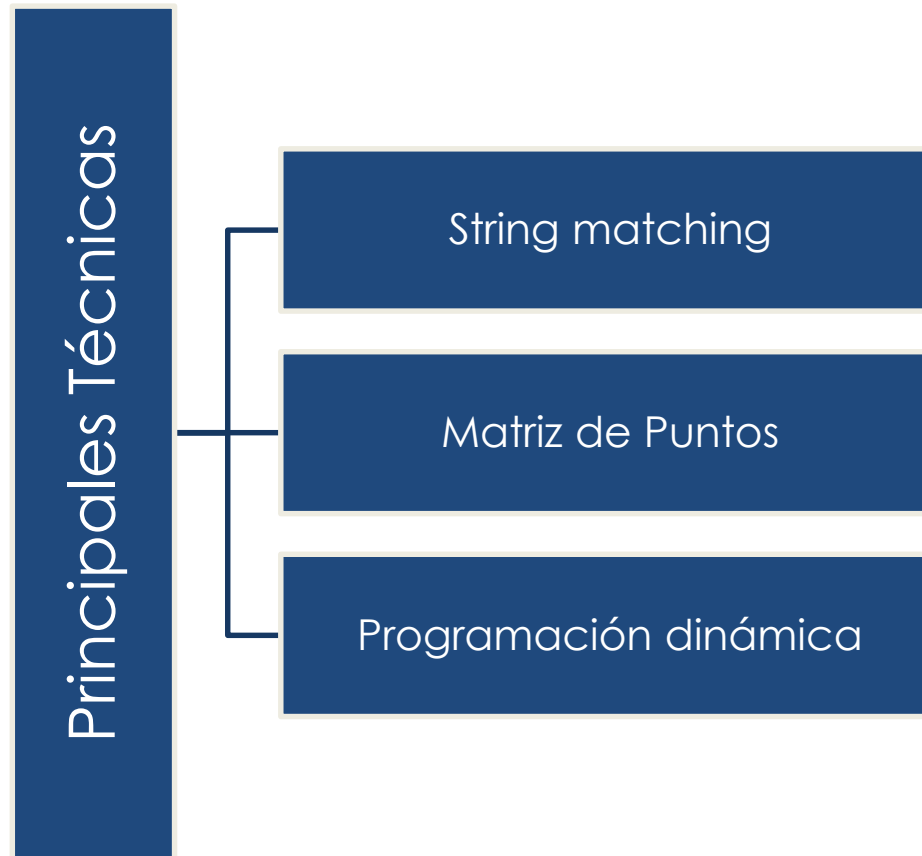
Se escriben con letras (aminoácidos, nucleótidos, etc.) en forma de filas en las que, si es necesario, se insertan espacios (*gaps*) para que las zonas con idéntica o similar estructura se alineen.

```
ATCTGAGGAAA_T
AT__G_GGAAGT
```

¿Qué se espera de las técnicas computacionales?

El algoritmo debe ser capaz de encontrar la(s) secuencia(s) de un base de datos de referencia que tienen mayor parecido a la secuencia de consulta.

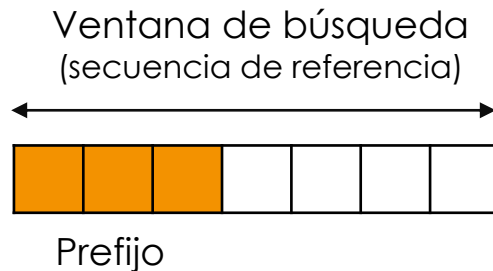




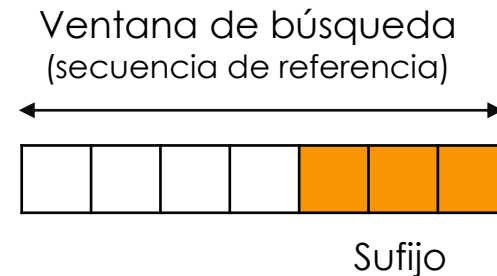
1.- STRING MATCHING

- ❑ Encontrar todas las ocurrencias de un patrón de *string* dado en un *string* de referencia.

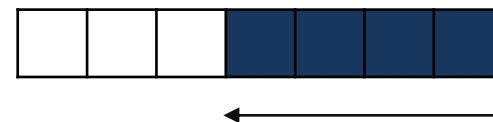
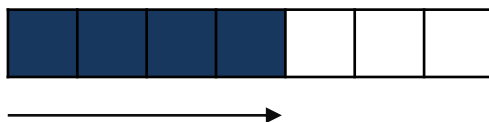
Basado en prefijo



Basado en sufijo



Matching: mover prefijo de largo m
a lo largo de la venta de búsqueda



1.- STRING MATCHING (Ejemplo)

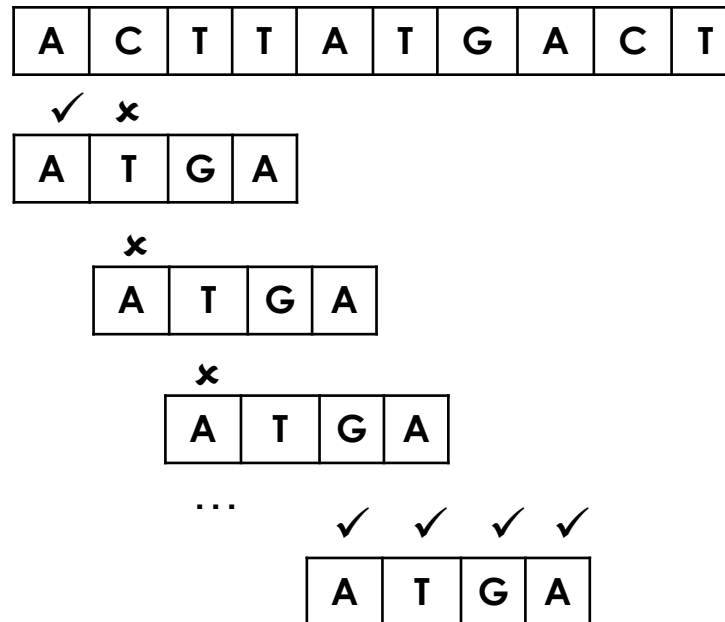
Sec. de refer.

A	C	T	T	A	T	G	A	C	T
---	---	---	---	---	---	---	---	---	---

Sec. de consulta

A	T	G	A
---	---	---	---

Matching: mover prefijo de largo m a lo largo de la secuencia de referencia.



1.- **STRING MATCHING**

- ☐ Cuando las secuencias son muy cortas o muy similares es un problema trivial.
- ☐ En la realidad, se necesitan alinear secuencias largas, muy variables y extremadamente numerosas que no pueden ser alineadas fácilmente.
- ☐ Por ello es adecuado el uso de técnica computacionales más idóneas las cuales dividen el alineamiento dos categorías:
 - ☐ **Alineamiento global**
 - ☐ **Alineamiento local**

1.- STRING MATCHING

- ❑ **Alineamiento global:** las secuencias se alinean a lo largo de toda su longitud, intentando alinear secuencias completas.

```
seq1  EARDF-NQYYSSIKRSGSIQ
      . : ..... : .
seq2  LPKLFIDQYYSSIKRTMG-H
```

- ❑ **Alineamiento local:** sólo se alinean las partes más parecidas de la secuencia.

```
seq1  NQYYSSIKRS
      . : ..... :
seq2  DQYYSSIKRT
```

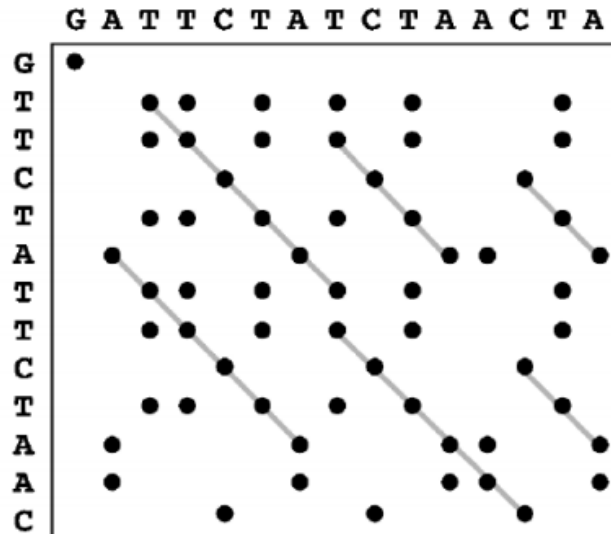
Más información en:

<https://ab.inf.uni-tuebingen.de/teaching/ws09/bioinformatics-i/06-stringmatch.pdf/view>

2.- MATRIZ DE PUNTOS

- ❑ Método gráfico para comparar dos secuencias utilizando una **matriz bidimensional**, donde las filas corresponden a la secuencia de consulta y las columnas a la secuencia de referencia.
- ❑ La comparación es realizada verificando la similitud entre cada letra de una secuencia contra los de la otra.
 - ✓ Si se encuentra un matching, un punto se coloca en coordenada de la gráfica, caso contrario coordenada queda vacía.
 - ✓ Cuando hay regiones substancialmente similares, los puntos forman líneas diagonales continuas, las cuales revelan el alineamiento de las secuencias.

2.- MATRIZ DE PUNTOS



Herramienta online

<http://www.cbs.dtu.dk/services/MatrixPlot/>

Limitantes

- ❑ El usuario debe construir el alineamiento completo visualmente al ir uniendo las diagonales.
- ❑ En la mayoría de los casos hay demasiados puntos en la gráfica, lo que dificulta la identificación del verdadero alineamiento.

3.- PROGRAMACIÓN DINÁMICA

- ❑ El problema se divide en sub-problemas, de manera que la solución a los sub-problemas simplifica la solución del problema global.
- ❑ Permite determinar el alineamiento óptimo de secuencias al verificar las **coincidencias para todos los posibles pares de caracteres** entre las dos secuencias.
- ❑ Al igual que la técnica anterior usa una **matriz bidimensional** pero reemplaza **puntos** por una **matriz de puntajes** para contar las coincidencias y diferencias entre las secuencias.



Fases

3.- PROGRAMACIÓN DINÁMICA (Ejemplo)

Alineamiento Global Algoritmo Needleman-Wunsch (NW)

Secuencia 1	A	T	C	C	G
Secuencia 2	A	G	T	C	G

Esquema de puntajes y penalidades

$S_{i,j}=1$

- Si hay matching entre la letra en la posición i de la secuencia 1 y la letra en la posición j de la secuencia 2.

$S_{i,j}=0$

- En cualquier otro caso

$w=0$ (no usaremos penalidad por gaps)

3.- PROGRAMACIÓN DINÁMICA (Ejemplo) (Algoritmo NW)

Paso 1: Inicialización

	A	T	C	C	G
A	0	0	0	0	0
G	0				
T	0				
C	0				
G	0				

Paso 2: Matriz de puntaje

Se inicia en la esquina superior izquierda y se encuentra el máximo puntaje $M_{i,j}$ para cada posición i, j .

$$M_{i,j} = \text{Max} [M_{i-1,j-1} + S_{i,j}, M_{i,j-1} + w, M_{i-1,j} + w]$$

3.- PROGRAMACIÓN DINÁMICA (Ejemplo) (Algoritmo NW)

M₁₁

$$M_{1,1} = \text{Max}[M_{0,0} + 1, M_{1,0} + 0, M_{0,1} + 0]$$

$$M_{1,1} = \text{Max}[1, 0, 0]$$

$$M_{1,1} = 1$$

Como $w = 0$, el resto de la fila 1 y columna 1 puede llenarse con 1's

	A	T	C	C	G
	0	0	0	0	0
A	0	1	1	1	1
G	0	1			
T	0	1			
C	0	1			
G	0	1			

$$M_{i,j} = \text{Max} [M_{i-1,j-1} + S_{i,j}, M_{i,j-1} + w, M_{i-1,j} + w]$$

3.- PROGRAMACIÓN DINÁMICA (Ejemplo) (Algoritmo NW)

M₂₂

$$M_{2,2} = \text{Max}[M_{1,1} + 0, M_{2,1} + 0, M_{1,2} + 0]$$

$$M_{2,2} = \text{Max}[1, 1, 1]$$

$$M_{2,2} = 1$$

M₃₂

$$M_{3,2} = \text{Max}[M_{2,2} + 1, M_{3,1} + 0, M_{2,2} + 0]$$

$$M_{3,2} = \text{Max}[2, 1, 1]$$

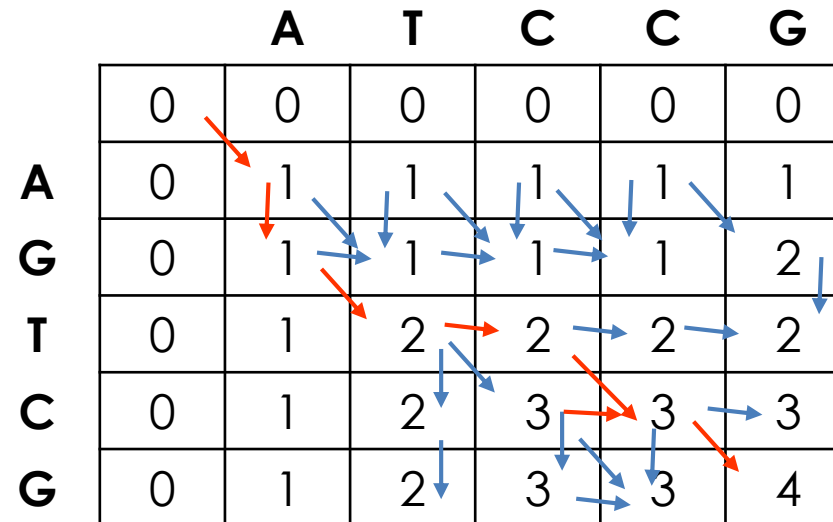
$$M_{3,2} = 2$$

		A	T	C	C	G
		0	0	0	0	0
A		0	1	1	1	1
G		0	1	1		
T		0	1	2		
C		0	1			
G		0	1			

$$M_{i,j} = \text{Max} [M_{i-1,j-1} + S_{i,j}, M_{i,j-1} + w, M_{i-1,j} + w]$$

3.- PROGRAMACIÓN DINÁMICA (Ejemplo) (Algoritmo NW)

	A	T	C	C	G
A	0	0	0	0	0
G	0	1	1	1	1
T	0	1	2	2	2
C	0	1	2	3	3
G	0	1	2	3	4



Paso 3: Rastreo del alineamiento

De donde proviene:

- Vecino de la izquierda (gap en secuencia 2)
- Vecino de la diagonal (matching/mismatching)
- Vecino de arriba (gap en secuencia 1)

Secuencia 1

A_TCCG

Secuencia 2

AGT_CG

RESÚMEN

Técnicas computacionales para alineamiento de secuencias:

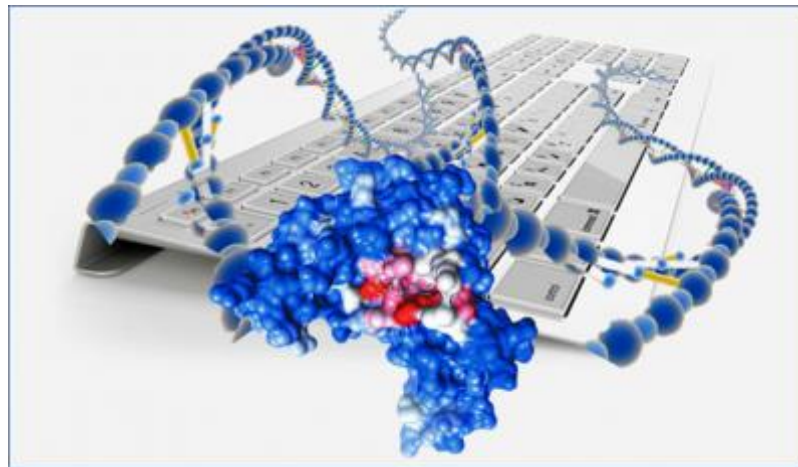
- **String matching**
- **Matriz de puntos**
- **Programación dinámica**

HABILIDADES REQUERIDAS

- ❑ Manipulación de caracteres, vectores, matrices, uso de métodos y funciones para realizar procesos de matching y sub-dividir procesos en forma adecuada.

Técnicas de Análisis de Datos Génicos (Minería de Datos)

- ❑ La minería de datos corresponde a un conjunto de métodos para la extracción de conocimiento desde enormes repositorios de datos.
- ❑ Entre las técnicas mas utilizadas en bioinformática se destacan:
 - El reconocimiento de patrones (string matching)
 - Selección de características
 - Agrupamiento

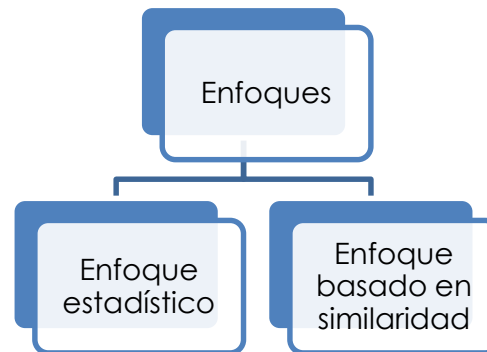


Selección de características

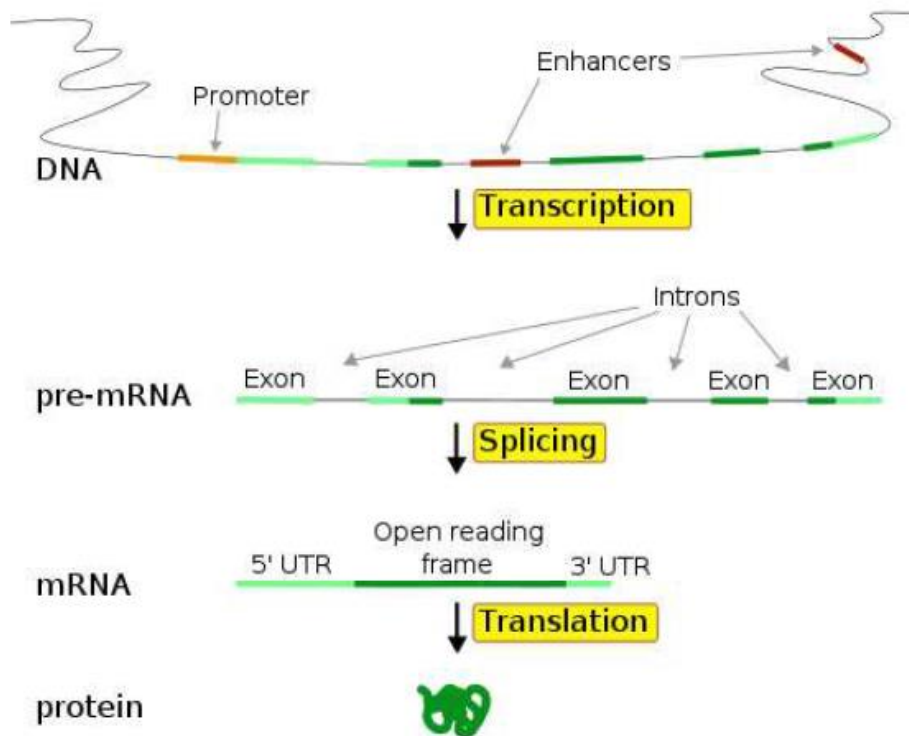
Busca *reducir* las entradas o encontrar las más significativas para un posterior procesamiento y análisis.

1.- PREDICCIÓN DE GENES

Permite la localización de patrones de comportamiento genéticos, ubicación de secuencias que corresponden a regiones regulatorias, etc.



1.- PREDICCIÓN DE GENES (Enfoque Estadístico)



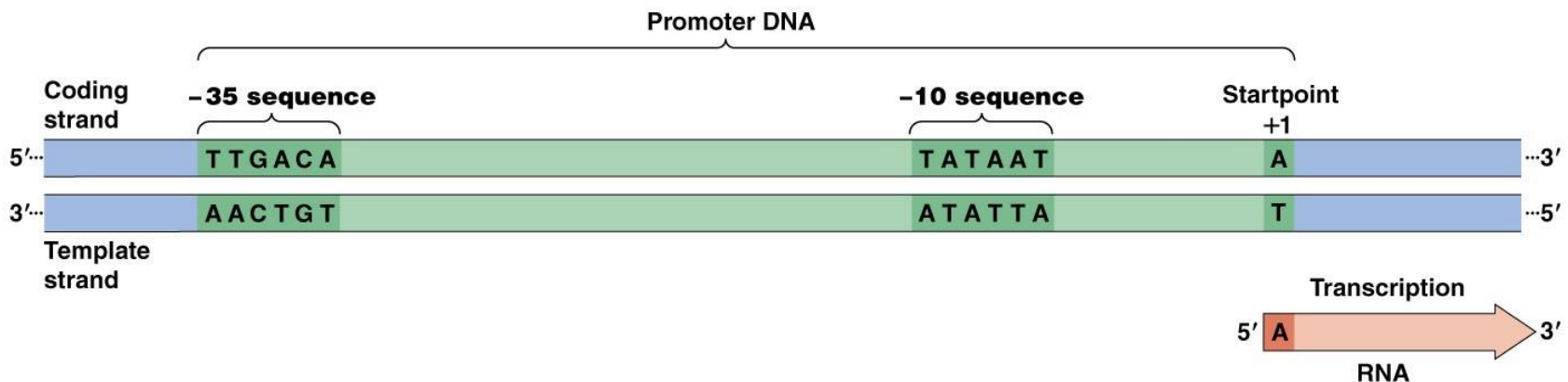
- ❑ Buscar patrones que aparecen frecuentemente en genes y no en otros lugares.
- ❑ Se basan en detectar leves variaciones estadísticas entre regiones codificantes (exons) y no codificantes.
- ❑ Técnicas habituales
 - Hidden Markov Models
 - Likelihood ratio

1.- PREDICCIÓN DE GENES (Enfoque por Similitud)

- ❑ Se basan en un enfoque **más biológico** al utilizar genes previamente secuenciados y las proteínas que codifican.

❑ Predicción de regiones promotoras

En los organismos procariontes, la secuencia específica de ADN que precede a los genes, y señala el comienzo de la transcripción del ADN a ARN se llama **promotor**.



1.- PREDICCIÓN DE GENES (Enfoque por Similitud)

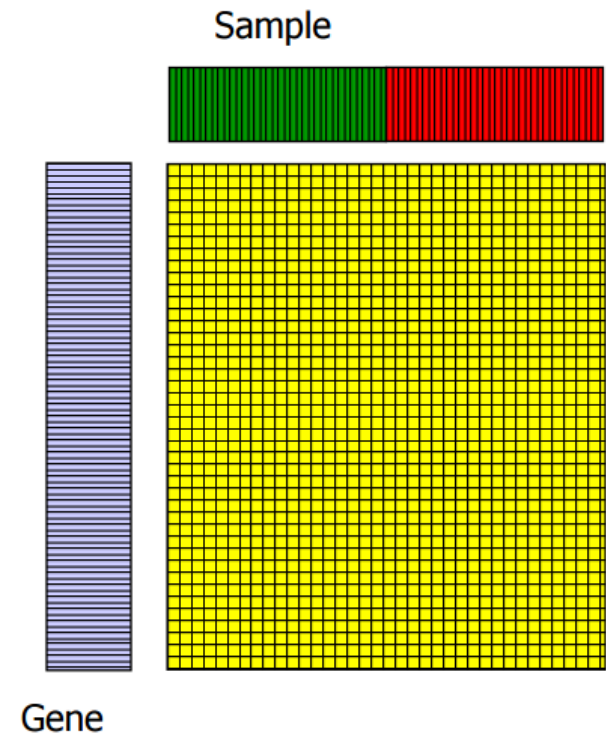
Predicción de regiones promotoras Técnica matriz de pesos posicionales (PWM).

secuencia 1	C C A A T G	} Perfiles de regiones promotoras.
secuencia 2	C C A T T G	
secuencia 3	C C A T T G	
secuencia 4	C C A T T G	

- ❑ Cuenta el número de nucleótidos en cada columna y establece un porcentaje de ocurrencia para cada una de las posiciones en la secuencia.
- ❑ Establece un **umbral** que servirá para establecer el grado de similitud que tiene una secuencia de consulta con los perfiles de regiones promotoras ya descubiertas.

2.- SELECCIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

- ☐ Selección de conjunto de genes que presentan cambios significativos en su expresión génica entre dos condiciones (habitualmente).
- ☐ Se compara la expresión de los genes de un grupo A con los de un grupo B.
- ☐ Entre las técnicas usadas se destacan:
 - **fold-change**
 - t-student moderada
 - ANOVA
 - SAM



2.- SELECCIÓN DE GENES DIFERENCIALMENTE EXPRESADOS (Ejemplo)

- ❑ **Fold-change**, mide el cambio de proporción de expresión génica a lo largo de las muestras.

$$FC = \log_2 \left| \frac{avg(N)}{avg(C)} \right|$$

- ❑ Consideremos las muestras N, y C correspondientes a tejido normal y con cáncer.

	Normal	Normal	Normal	Cancer	Cancer
	1	2	3	1	2
gen A	4	1	4	6	12
gen B	12	5	6	1	3
gen C	3	6	7	11	10
gen F	2	11	4	5	7

2.- SELECCIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

$$FC = \log_2 \left| \frac{avg(N)}{avg(C)} \right|$$

	Normal 1	Normal 2	Normal 3	Cancer 1	Cancer 2
gen A	11	6	3	1	2
gen B	3	5	6	4	3
gen C	3	6	7	11	10
gen F	2	11	4	5	7

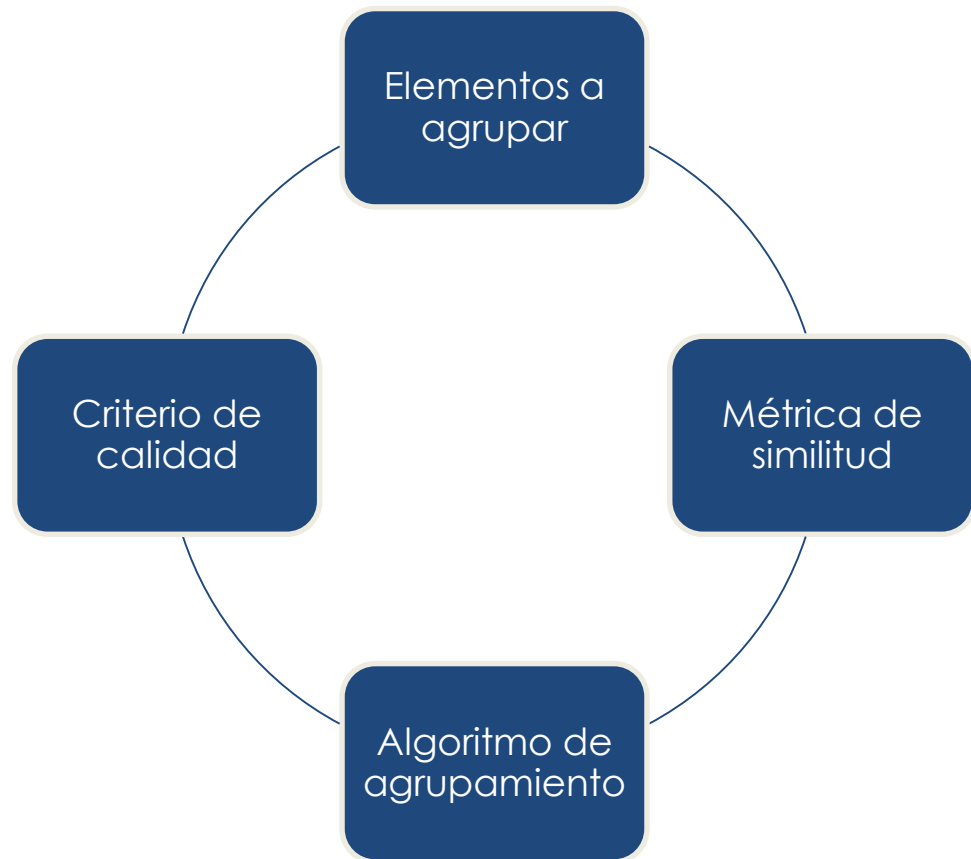
	Avg(N)	Avg(C)	FC
gen A	6.66	1.5	2.15
gen B	4.66	3.5	0.42

Umbral de Fold-change

- $FC < 2$ no es diferencialmente expresado
- $2 < FC < 4$ es diferencialmente expresado
- $FC > 4$ está altamente diferencialmente expresado

Agrupamiento

- ❑ Es un procedimiento de agrupación de una serie de objetos de acuerdo con una medida de distancia.
- ❑ La idea es que los elementos dentro de un grupo sean lo mas parecidos entre ellos.
- ❑ Los grupos sean lo mas diferentes entre ellos.



Agrupamiento (Ejemplo)

- Agrupar genes en base a su niveles de expresión génica.

	15min	30min	60min	90min	3hr	6hr	9hr	24hr
AT4G16610	0.2559	-0.5427	-1.3302	-0.4752	0.0759	-0.5427	2.0108	0.5484
AT4G29010	-0.5087	-0.1735	-1.8016	-0.5087	0.4249	0.4568	0.5127	1.5981
AT2G34420	-0.1999	-0.6569	-0.6569	-1.0885	-0.1745	-0.2507	1.1710	1.8566
AT2G43080	0.0915	0.1892	-0.7387	-0.0793	-0.3602	-1.6912	1.1173	1.4714
AT3G12120	0.5163	-0.3426	-1.5392	0.2074	0.1688	-1.0085	0.2364	1.7612

- Métrica de distancia

$$D(g_1, g_2) = \sqrt{(g_1s_1 - g_2s_1)^2 + (g_1s_2 - g_2s_2)^2 + \dots + (g_1s_8 - g_2s_8)^2}$$

Agrupamiento (Ejemplo)

❑ Matriz de distancia

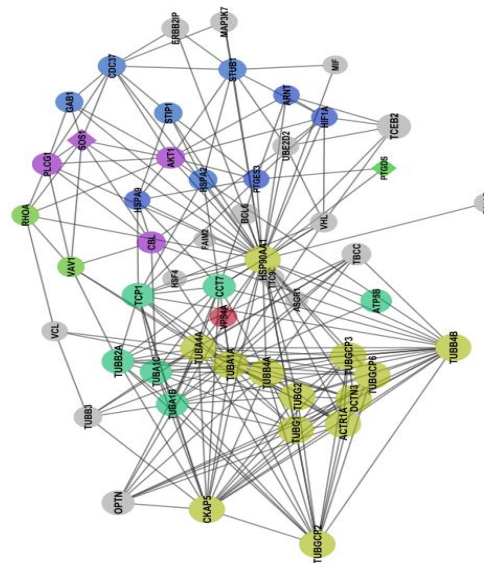
	AT4G16610	AT4G29010	AT2G34420	AT2G43080	AT3G12120
AT4G16610	0	2.3262	1.9012	2.0564	2.3372
AT4G29010	2.3262	0	1.8264	2.7237	1.9942
AT2G34420	1.9012	1.8264	0	2.0216	2.1552
AT2G43080	2.0564	2.7237	2.0216	0	1.6710
AT3G12120	2.3372	1.9942	2.1552	1.6710	0

Agrupamiento (Ejemplo)

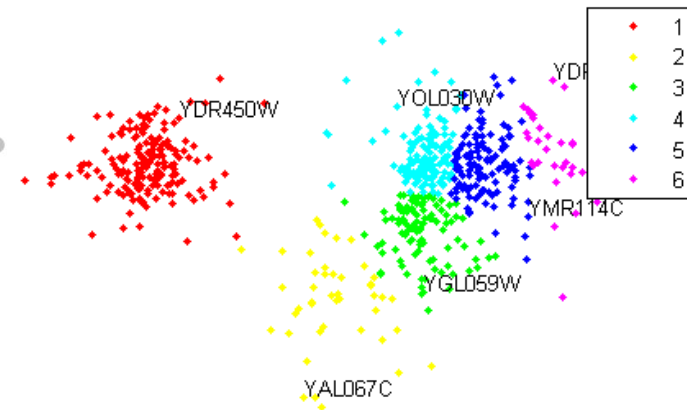
❑ Algoritmo de agrupamiento



Jerárquico



Densidad

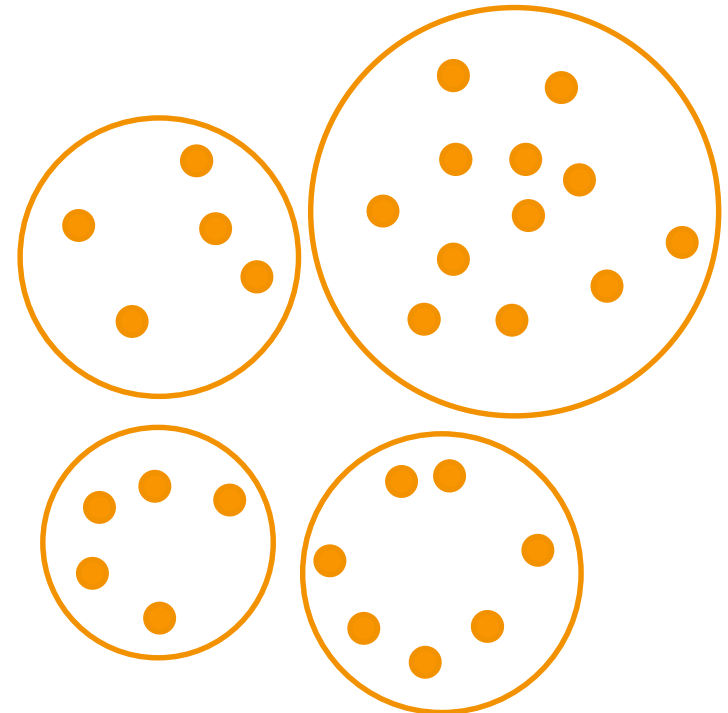
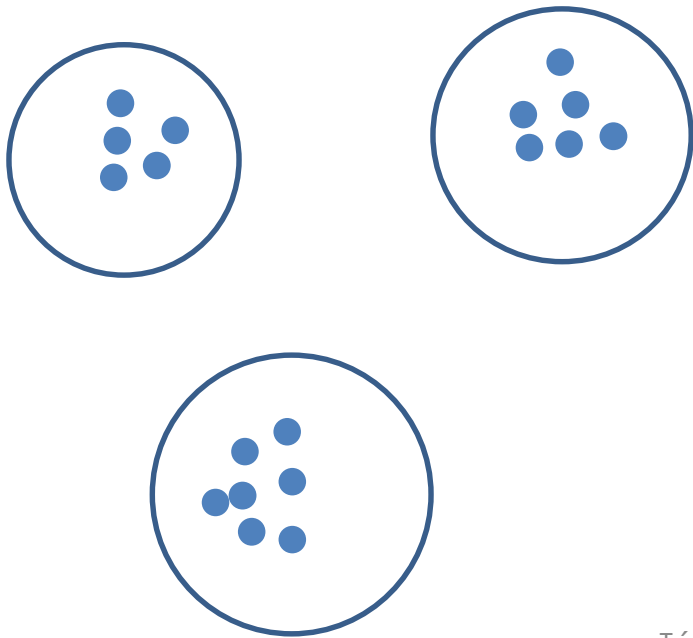


Particional

Agrupamiento (Ejemplo)

❑ Criterio de calidad

- ✓ Homogeneidad
- ✓ Separación



RESÚMEN

Técnicas computacionales para análisis de datos génicos:

- **Minería de datos**
 - Selección de características
 - Agrupamiento
- **Métodos estadísticos**

HABILIDADES REQUERIDAS

- ❑ Manipulación de vectores, matrices, uso de métodos y técnicas estadísticas y minería de datos con un énfasis biológico.

Técnicas de Optimización Combinatorial

- ❑ Gran variedad de problemas que pueden formularse como problemas de optimización. Por ejemplo:
 - Minimizar el número de gaps en un alineamiento
 - Hallar el número de particiones idóneas del agrupamiento de genes
 - Maximizar el número de coincidencias de pares de nucleótidos
- ❑ Estos problemas y otros son muy complejos por lo que no es posible resolverlos de **forma exacta** en tiempo razonable.
- ❑ Alternativa, recurrir a **algoritmos aproximados que entregan soluciones de alta calidad en un tiempo razonable.**
- ❑ Estos algoritmos incluyen las técnicas **heurísticas** y **meta-heurísticas**.

Heurísticas

- Procedimiento con un alto grado de confianza de encontrar buenas soluciones con costo computacional razonable.



Meta-heurísticas

- Estrategias para diseñar y/o mejorar las técnicas heurísticas orientados a obtener un alto rendimiento.
- Usar cuando las heurísticas no son efectivas o no existe un algoritmo específico para un determinado problema.
- Constan de mecanismos para escapar de soluciones óptimas locales en su intento por encontrar la solución óptima global.

Algoritmos genéticos



Simulated Annealing



Algoritmos de
enjambre



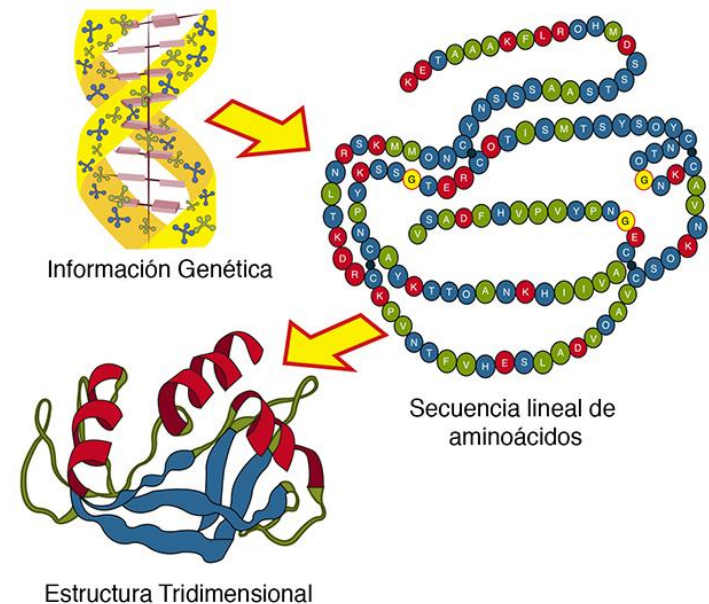
Búsqueda Tabú



Búsqueda Dispersa



Algoritmos
meméticos

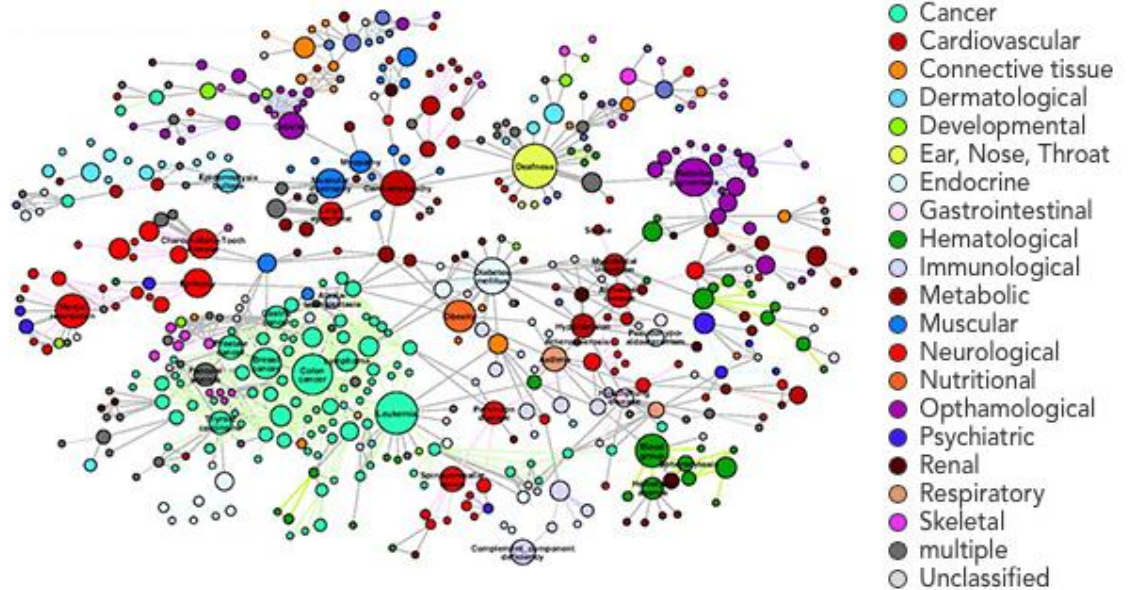


- **Agrupamiento de Genes**
- Plegado de Proteínas
- Inferencia Filogenética

Teoría de Grafos

- ❑ Un **grafo** es una representación de las interacciones que tienen lugar entre las entidades de un sistema.
- ❑ Los nodos representan las entidades (genes, proteínas, metabolitos, etc).
- ❑ Las aristas entre distintos nodos indican que las correspondientes entidades están relacionadas entre sí de alguna forma.

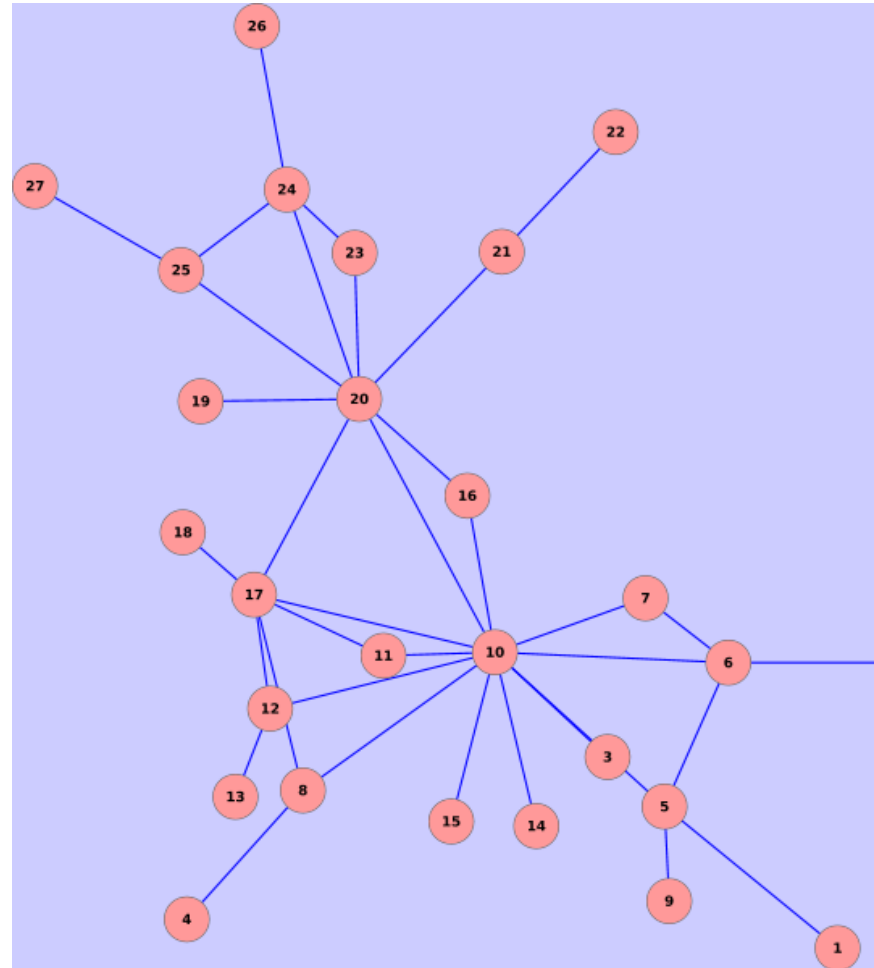
Human Disease Network



- ❑ Una red o grafo G no dirigido es un par de conjuntos $G=(V,E)$.

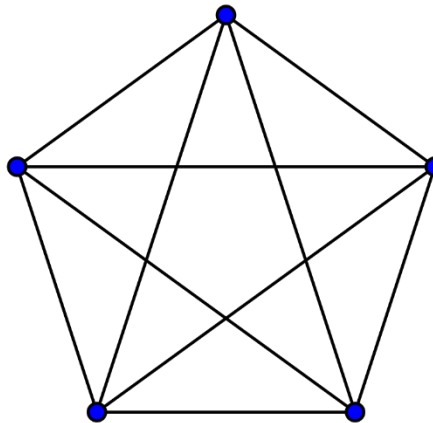
- $V = \{v_1, v_2, \dots, v_n\}$ es el conjunto de vértices o nodos.

- $E = \{ (v_i, v_j), (v_k, v_l), \dots \}$ es un conjunto de aristas entre elementos de V .

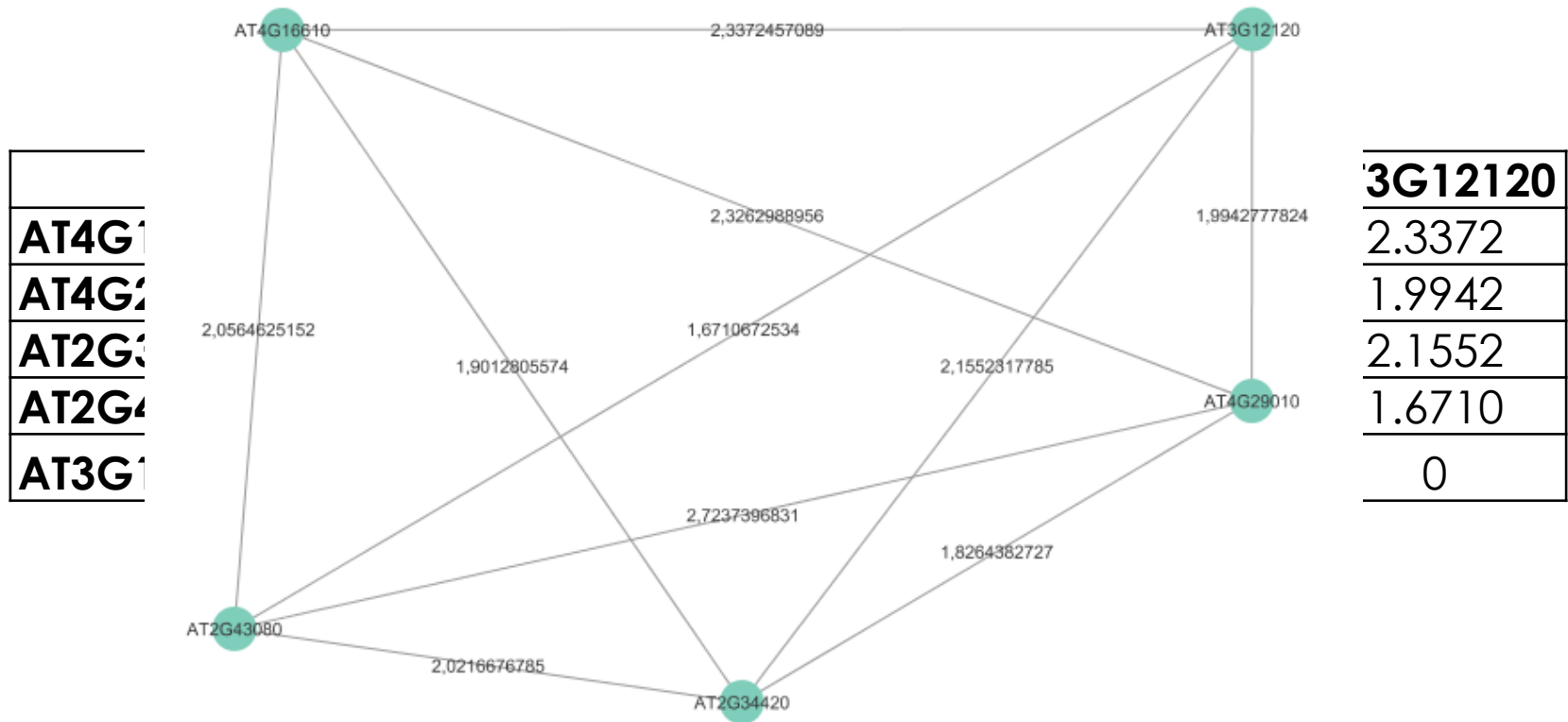


Problema del *CLIQUE* y su relación en análisis de redes biológicas

- ❑ Un clique en un grafo no dirigido G es un conjunto de vértices V tal que para todo par de vértices de V , existe una arista que las conecta.
- ❑ El Problema del clique, que consiste en **dado un grafo, decidir si existe en él un clique con un tamaño particular, habitualmente el máximo posible.**

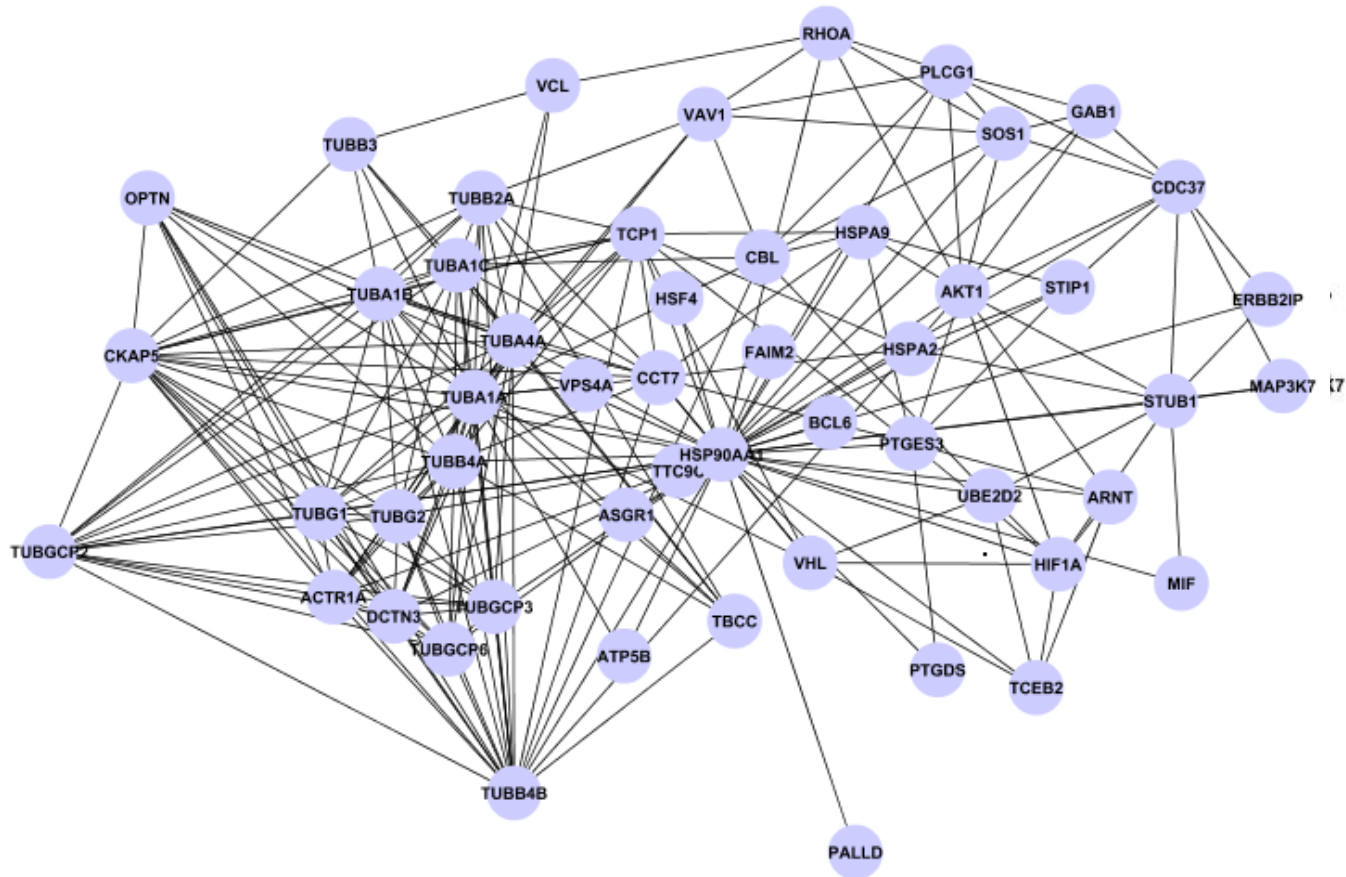


Problema del **CLIQUE** y su relación en análisis de redes biológicas



Grafo (CLIQUE) de interacción entre genes de Arabidopsis Thaliana

Grafos y redes biológicas



¿Preguntas?