Seminario sobre Biocomputación
Granada, 10 diciembre 2004

Líneas de Trabajo en Bioinformática UMA

C. Cotta

Dpto. Lenguajes y Ciencias de la Computación
Universidad de Málaga

ccottap@lcc.uma.es
http://www.lcc.uma.es/~ccottap

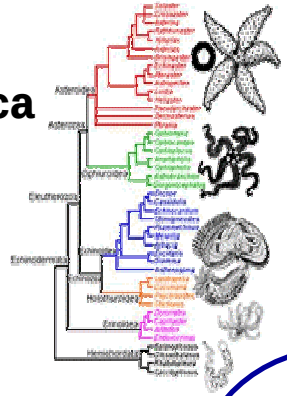
Universidad de Málaga

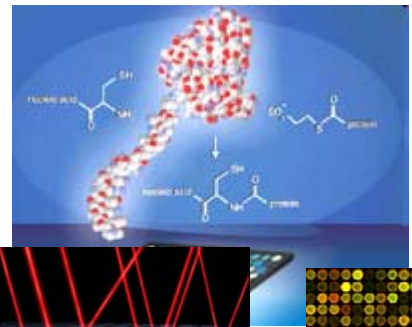
A Vista de Pájaro

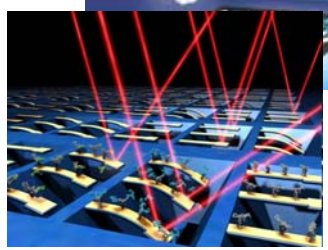
Inferencia Filogenética

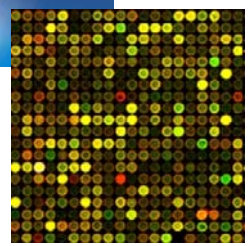
Análisis de Datos de Expresión Genética

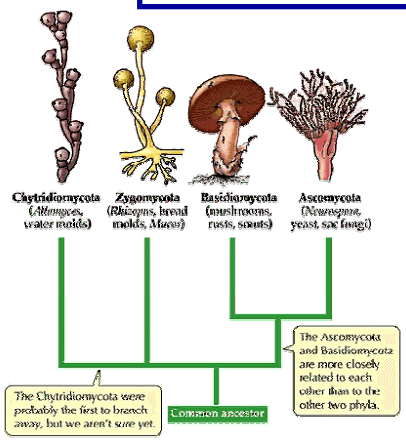
Redes Genéticas y Metabólicas

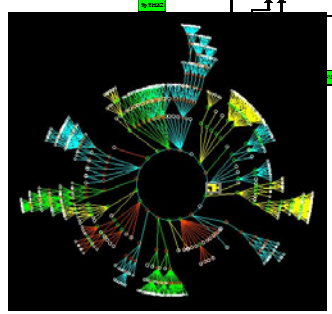








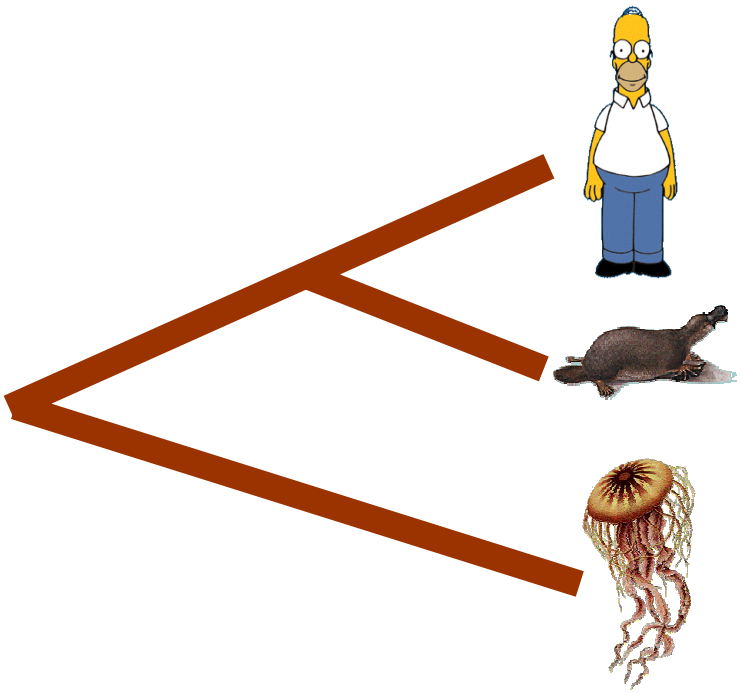


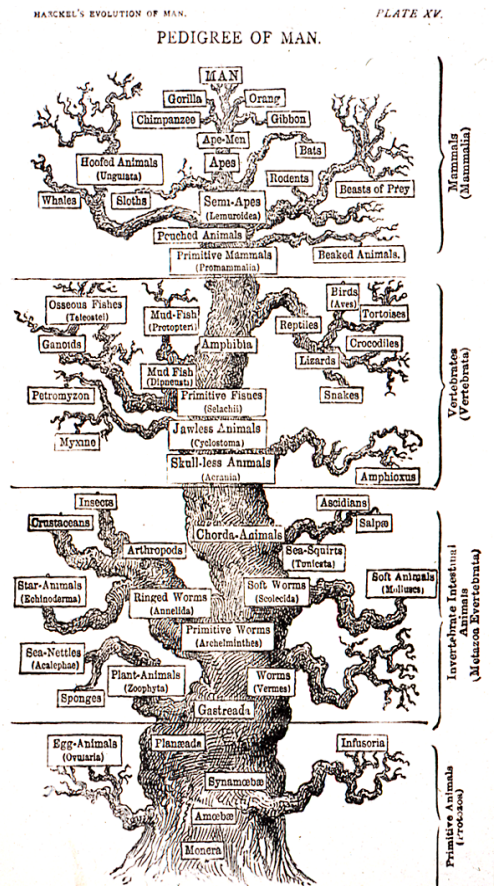


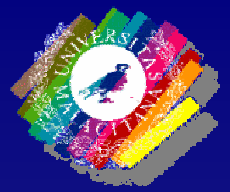
Protein Structure Prediction
DNA Sequencing and Alignment
...

Inferencia Filogenética

- Las Teorías de la Evolución nos indican que todas las especies tienen un antecesor común.







Universidad de Málaga

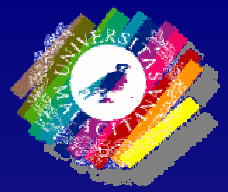
Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Inferencia Filogenética

- Tradicionalmente, el análisis se basaba en características morfológicas.
- Hoy en día, se usa información biomolecular.
- El análisis filogenético tiene implicaciones no sólo en taxonomía, sino también en:
  - Alineamiento de secuencias
  - Estudios epidemiológicos
  - Medicina forense
  - Predicción de la estructura de proteínas
  - ...



Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Enfoques para la Inferencia

- Métodos basados en secuencias
  - Máxima verosimilitud (ML)
  - Parsimonia
- Métodos basados en distancia
- Puede demostrarse la NP-dificultad de la inferencia bajo estos enfoques.
- Necesidad de recurrir a (meta)heurísticas.
- Consideramos algoritmos evolutivos y métodos basados en distancia.

Métodos Basados en Distancia

- Se obtiene una matriz de distancias $D=\{d_{ij}\}_{1\leq i,j\leq n}$ entre taxa.
- Se construye un árbol con pesos asociados a cada rama. Este árbol T induce una matriz de distancias inferidas D^T.
- Minimizar distancia entre D^T y D (si D es aditiva, sólo hay un T para el que $D^T=D$).
- Hipótesis consideradas:

$$\begin{cases} d_{ij}^T \geqslant d_{ij} & \text{(homoplasia)} \\ d_{ij}^T \leqslant \max\left(d_{ik}^T, d_{jk}^T\right) & \text{(ultrametricidad)} \end{cases}$$

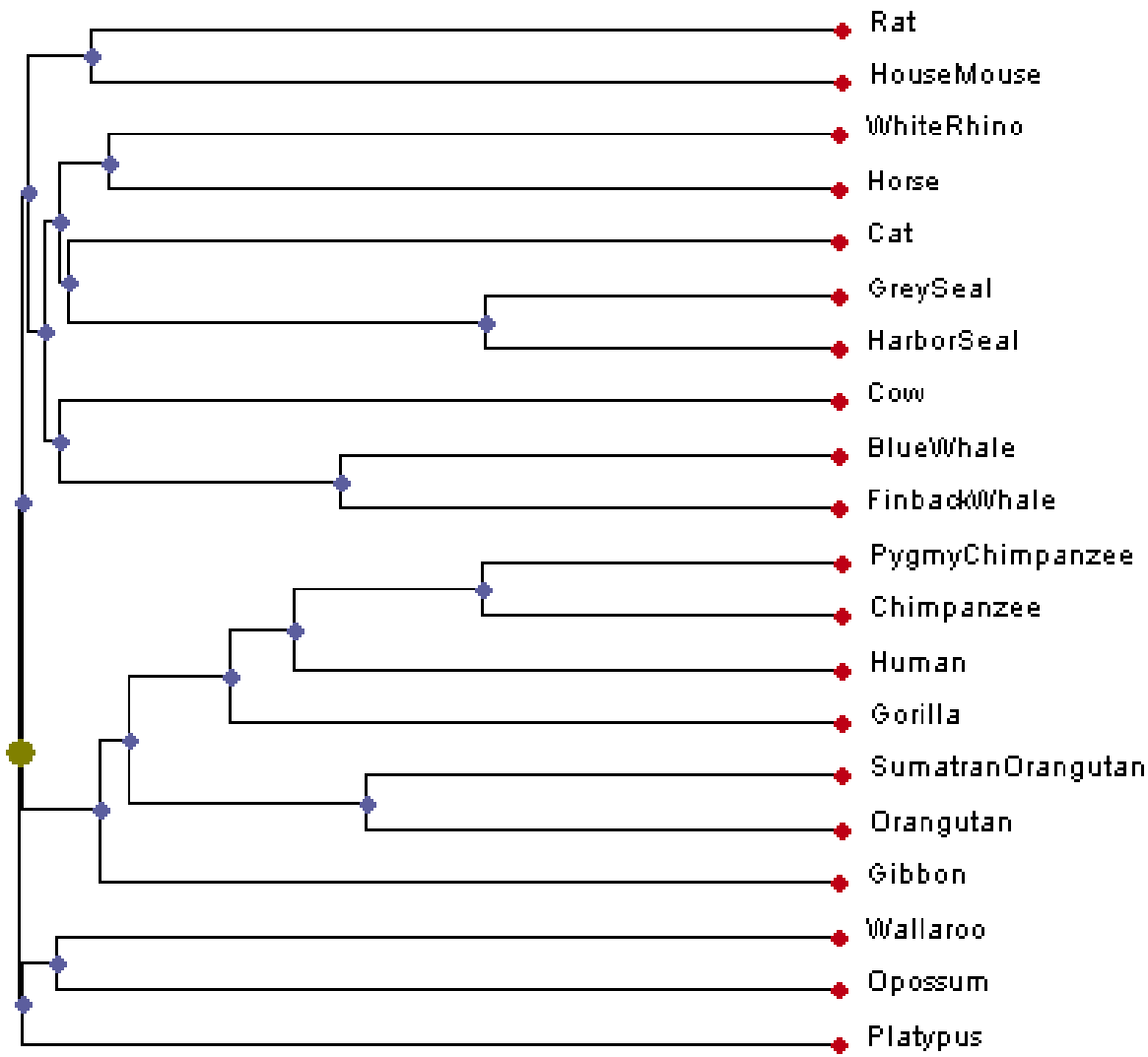
Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

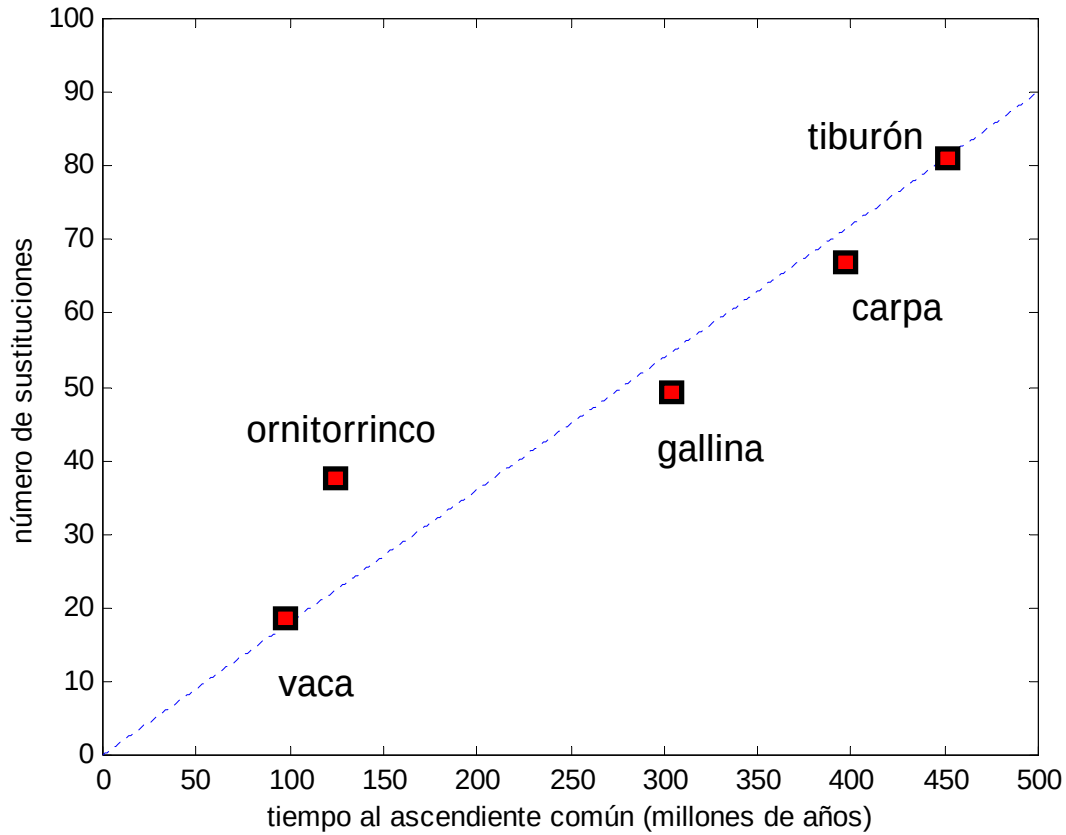
Redes Genéticas y Metabólicas

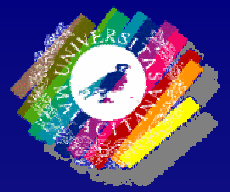
Árboles Ultramétricos



El Reloj Molecular

- Ritmo de sustituciones en la hemoglobina animal [ZuckerkandlPauling62]





Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Árboles Ultramétricos

- La hipótesis del reloj molecular está desacreditada hoy en día, pero...

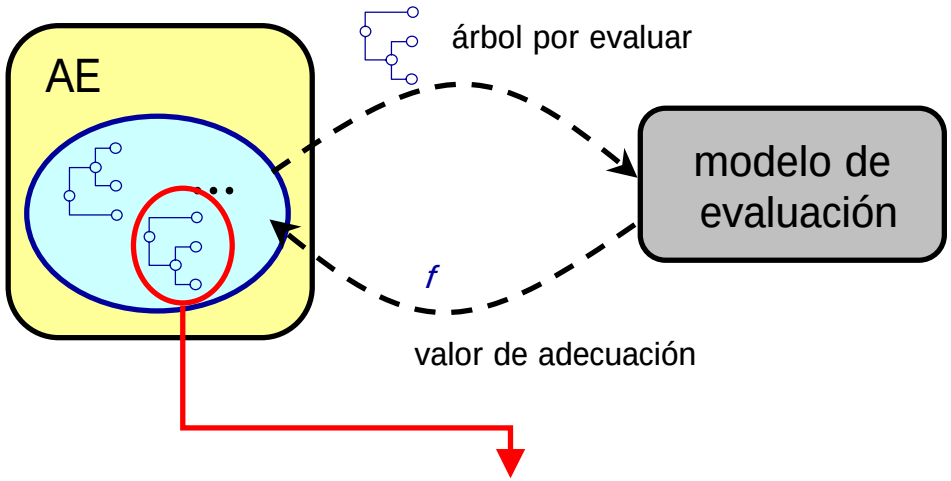
  ... el modelo ultramétrico proporciona una muy buena aproximación al aditivo, y...

  ... los pesos de las ramas pueden calcularse en tiempo $O(n^2)$ [WCT99].

- Si d_{ij} se calcula empleando distancias de Hamming, entonces se puede simular parsimonia.

Enfoques Evolutivos

- Aproximación directa.

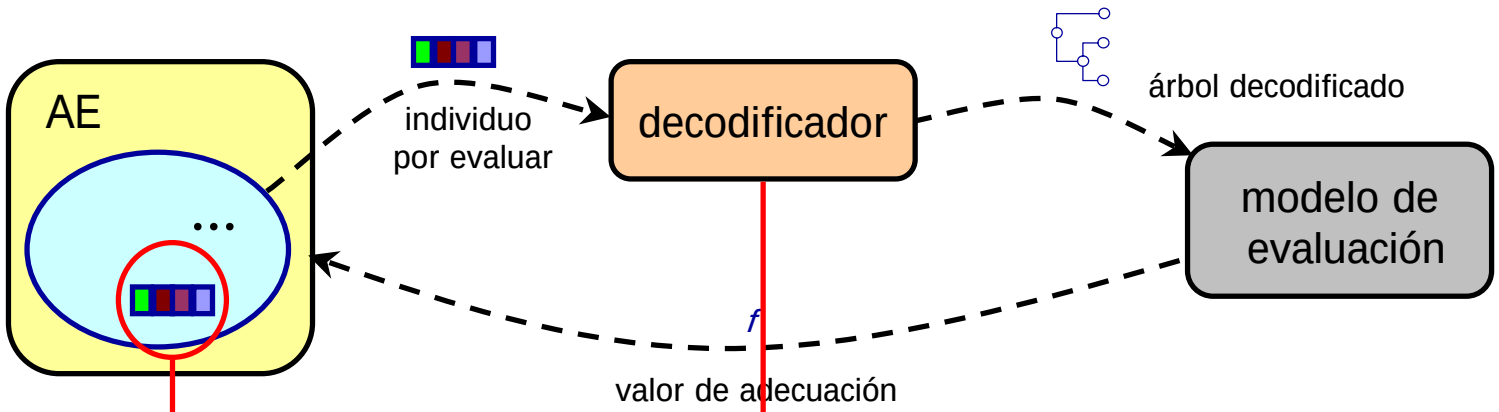


Cada individuo de la población es un árbol filogenético.

Deben definirse operadores reproductivos que manipulen árboles, y produzcan nuevas soluciones válidas

Enfoques Evolutivos

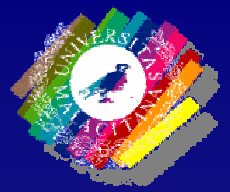
- Aproximación indirecta.



Cada individuo de la población es una *semilla* para la construcción de un árbol filogenético.

Se emplea un *decodificador* para construir el árbol que cada individuo contiene.

Se definen operadores que manipulan estas *semillas.*



Universidad de Málaga

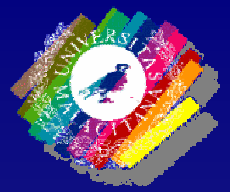
Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Comparativa Experimental

- Instancias filogenia mamíferos.
- Mammals.20 [COH97] $\Rightarrow$ mostrar orden de especiación entre primates, ferungulados, y roedores.
- Mammals.34 [RGPCS00] $\Rightarrow$ mostrar posición filogenética de los roedores.
- Matrices de distancia obtenidas a partir de ADN mitocondrial.



Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

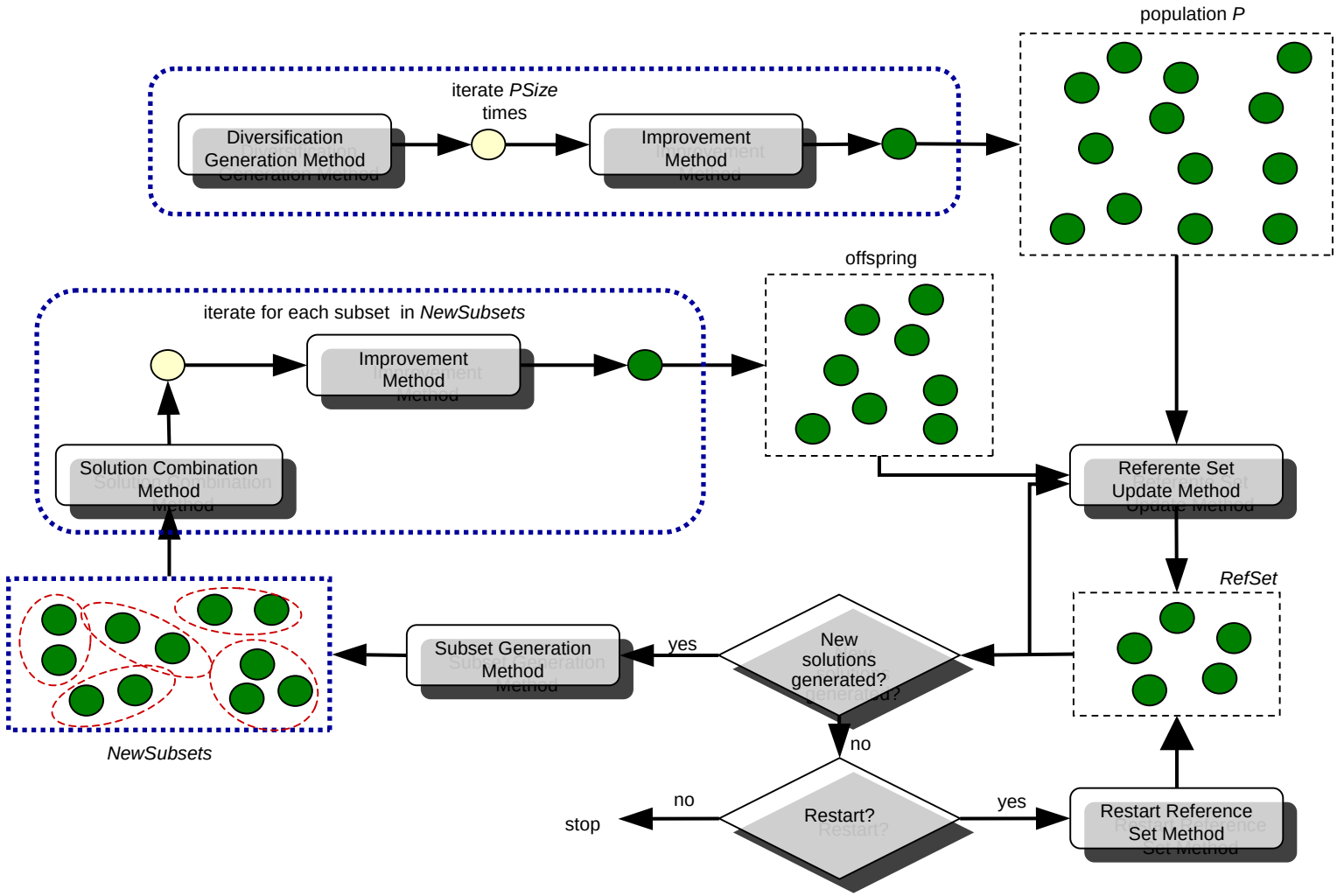
Redes Genéticas y Metabólicas

Comparativa Experimental

- El tamaño de estas instancias permite su resolución mediante *Branch & Bound*.
- El enfoque directo encuentra la solución óptima un número moderado de veces.
- El enfoque ordinal no consigue encontrar la solución óptima del BnB, usando múltiples operadores diferentes.
- El enfoque permutacional encuentra consistentemente la solución óptima...

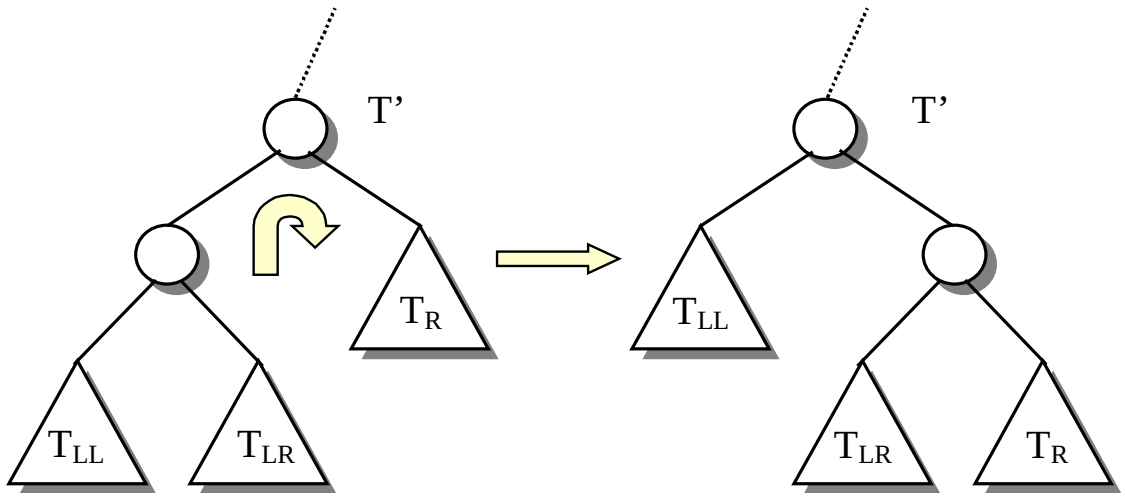
  ... pero escala peor.

Scatter Search y MAs



Mejora Local

- Rotación interna:



movimiento aceptable si

$$\max_{x,y \in \mathcal{L}(T_{LL}) \cup \mathcal{L}(T_{LR})} \{d_{x,y}\} > \max_{x,y \in \mathcal{L}(T_{LR}) \cup \mathcal{L}(T_R)} \{d_{x,y}\}$$

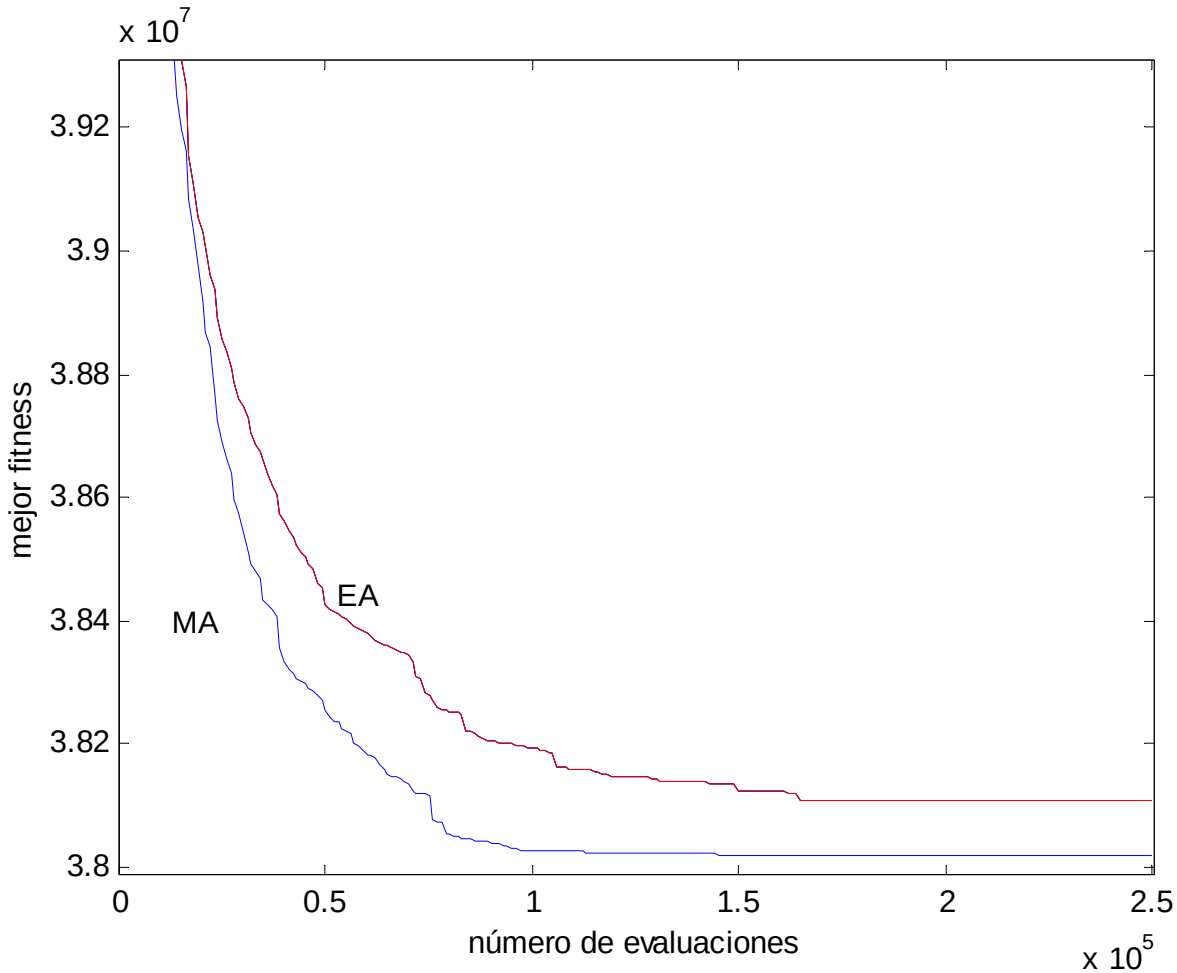
Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

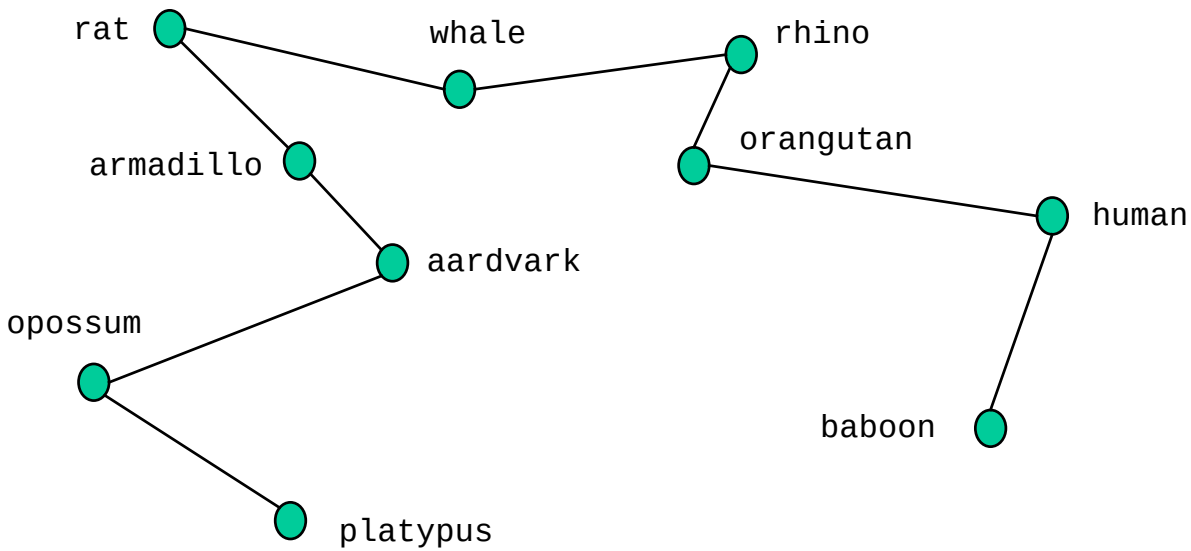
Comparativa Experimental



- Resultados sobre una instancia de 84 especies

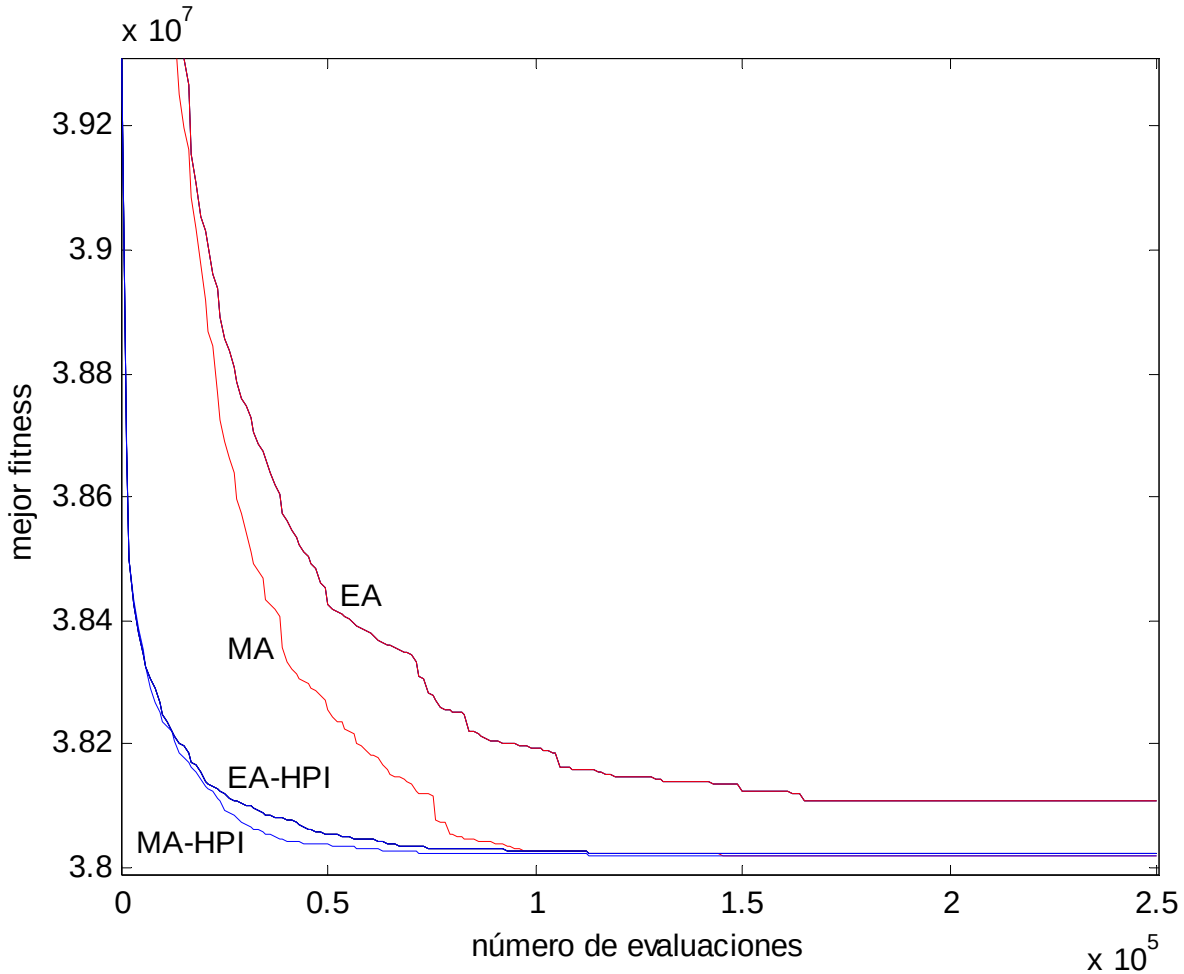
Inicialización Heurística

- Construir una población inicial con soluciones de calidad.

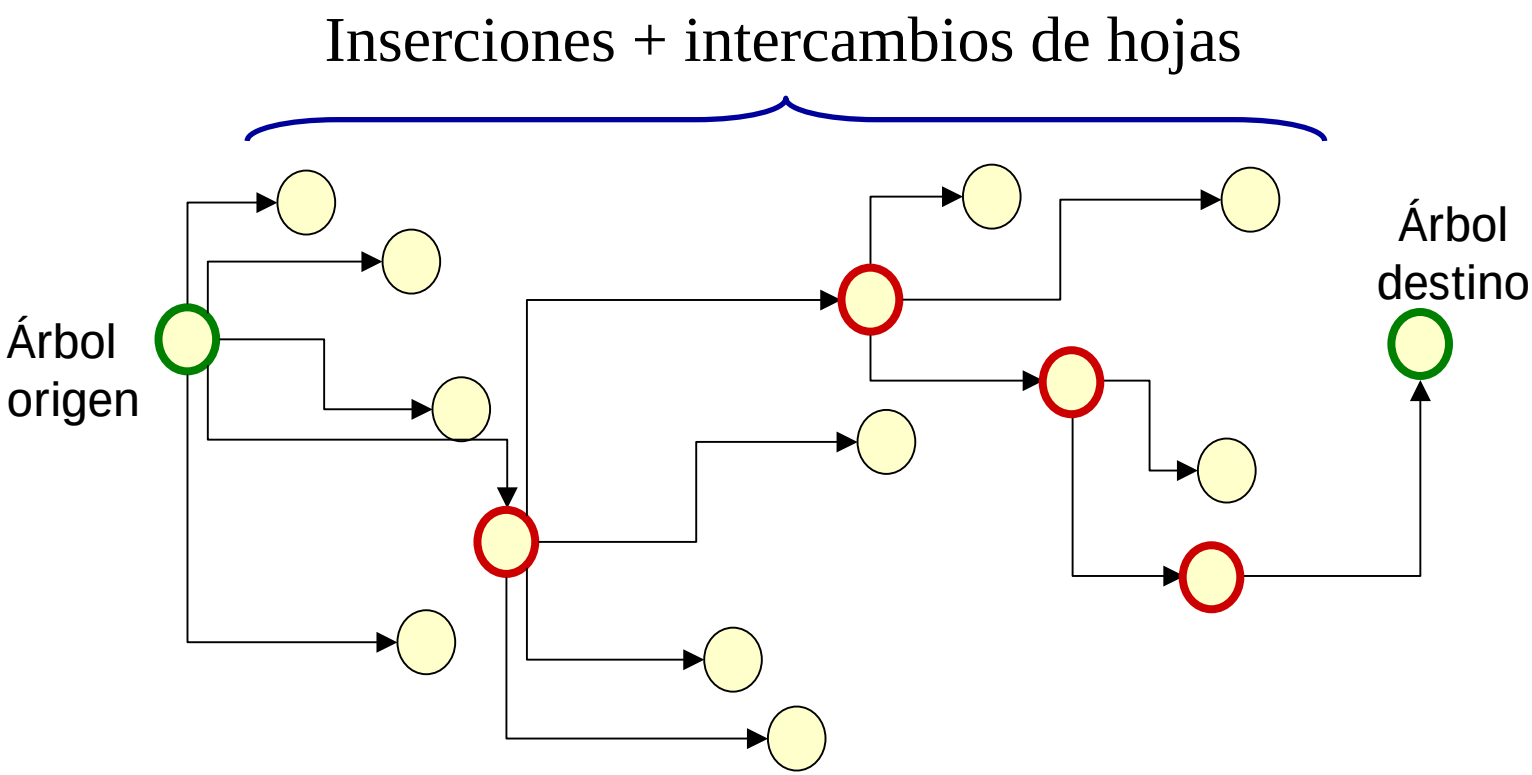


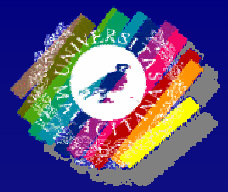
- Forzar que el orden de las hojas de los árboles corresponda al camino Hamiltoniano.

Inicialización Heurística



Path Relinking





Universidad de Málaga

Algunos Resultados

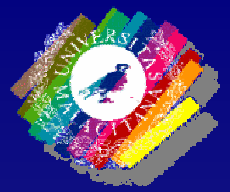
Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

	M420	M1097	M877	M971	M808
single-link	2.93600	1.26385	53.82255	6.47470	66.51205
complete-link	2.35685	1.01205	10.15965	4.82025	11.58550
average-link	2.50110	1.04855	10.89800	5.15390	12.45225
neighbor-joining	3.44775	1.24985	18.21510	5.46825	38.20275
Fitch-Margoliash	2.63675	1.09675	29.09415	5.91770	31.86065

	Scatter Search (static update)			
Problem	best	mean ± std. dev.	worst	median
M420	2.3491	2.3545 ± 0.0028042	2.3565	2.3556
M1097	1.0114	1.0114 ± 0.0000300	1.0115	1.0114
M877	10.1451	10.1528 ± 0.0035985	10.1554	10.1541
M971	4.8049	4.8106 ± 0.0035018	4.8164	4.8100
M808	11.5306	11.5396 ± 0.0078184	11.5505	11.5362



Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Bibliografía

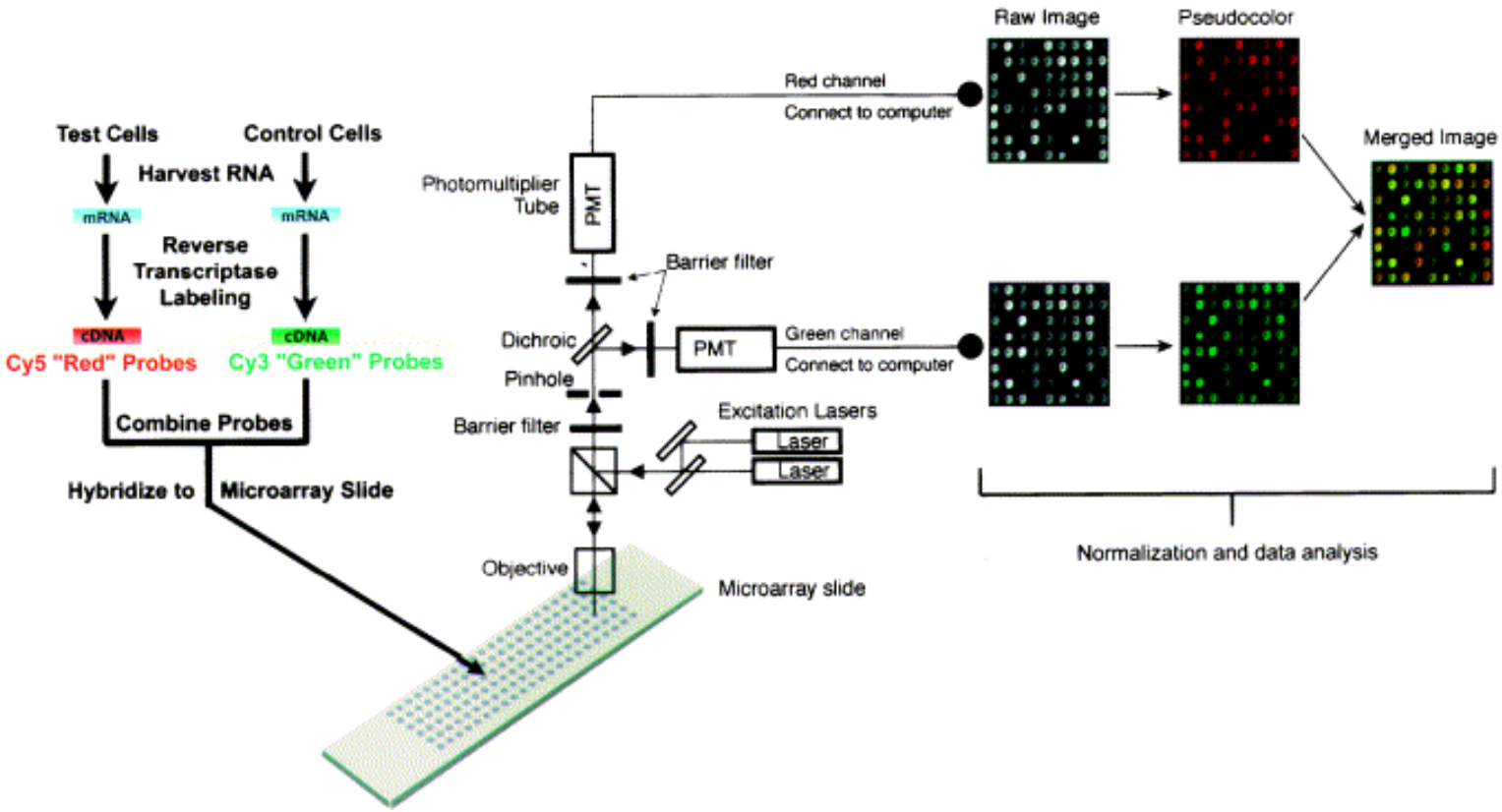
[CottaMoscato02] C. Cotta, P. Moscato, **"Inferring Phylogenetic Trees Using Evolutionary Algorithms"**, *Parallel Problem Solving From Nature VII*, JJ. Merelo *et al.* (eds.), LNCS 2439:720-729, Springer Verlag, 2002

[MBC02] P. Moscato, L. Buriol, C. Cotta, **"On the Analysis of Data Derived from Mitochondrial DNA: Kolmogorov and a Traveling Salesman give their Opinion"**, *Advances in Nature Inspired Computation: the PPSN VII Workshops*, D. Corne et al. (eds.), pp. 37-38, PEDAL, University of Reading, 2002

[CottaMoscato03] C. Cotta, P. Moscato, **"A Memetic-Aided Approach to Hierarchical Clustering from Distance Matrices: Application to Gene Expression Clustering and Phylogeny"**, *Biosystems* 72(1-2):75-97, 2003

[Cotta04] C. Cotta, **"Scatter Search with Path Relinking for Phylogenetic Inference"**, *European Journal of Operational Research*, 2004 (en prensa)

Micromatrices de ADN



¡Se monitoriza el estado simultáneo de miles de genes!

Universidad de Málaga

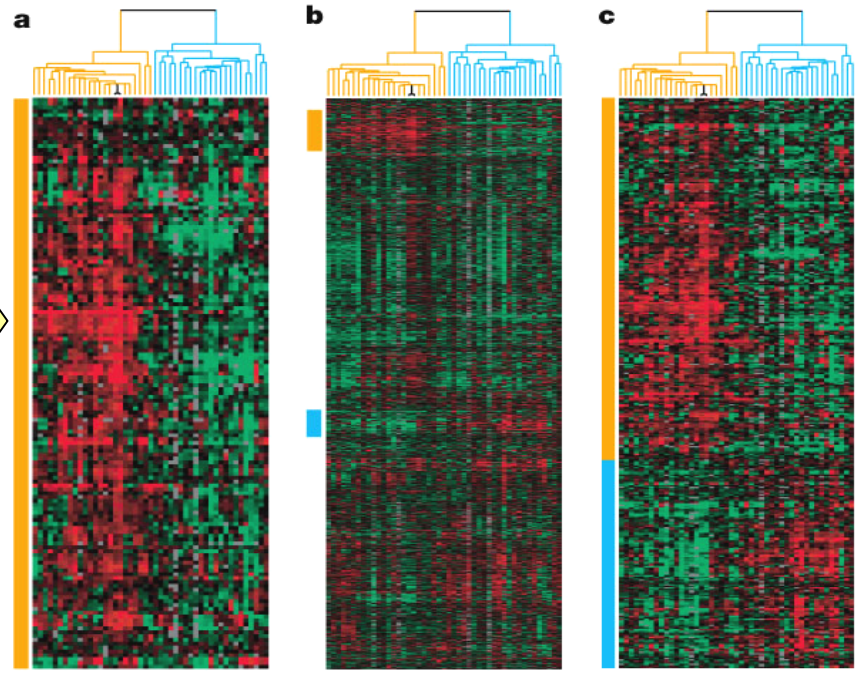
Agrupamiento de Genes

Inferencia Filogenética

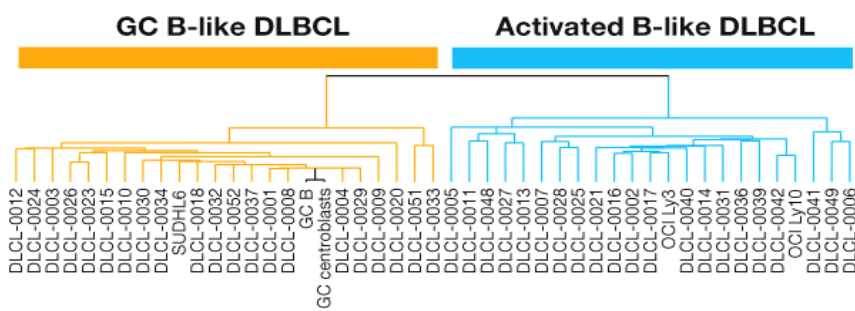
Análisis de Datos de Expresión Genética

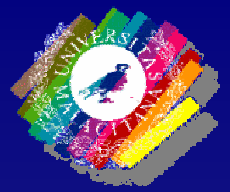
Redes Genéticas y Metabólicas

Datos de expresión génica para dos tipos de leucemia



El agrupamiento de genes con patrones de expresión similar es fundamental.





Universidad de Málaga

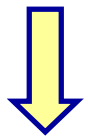
Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Procesado de la Información

- Agrupamiento jerárquico: análogo a los problemas de filogenia.



Se puede usar el arsenal de técnicas descritas anteriormente.

- Problema relacionado: presentación de la información.

¿Cómo se le puede hacer la vida más fácil al usuario?

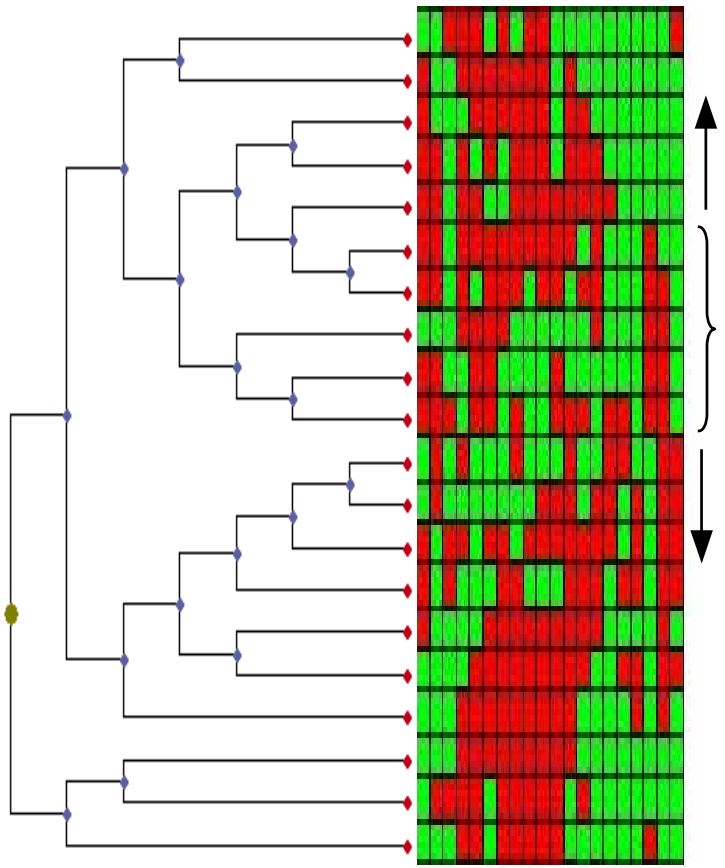
Universidad de Málaga

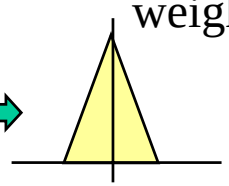
El Problema de la Ordenación de Genes

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas





$$\Psi(G) = \sum_{i=1}^{N} \Phi(g_i, G)$$

Homogeneidad del perfil génico en torno a g_i

$$\Phi(g_i, G) = \sum_{j=i-s}^{j=i+s} \Delta(g_i, g_j)$$

Universidad de Málaga

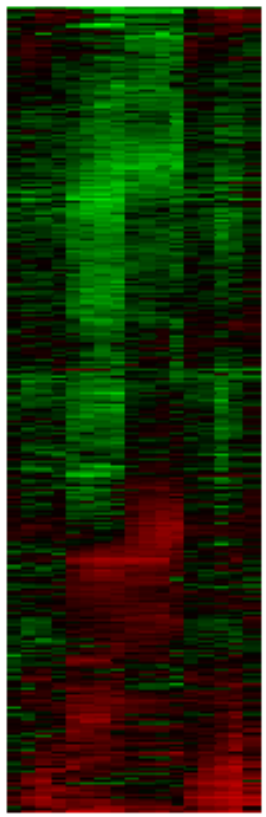
MAs make my day!

Inferencia Filogenética

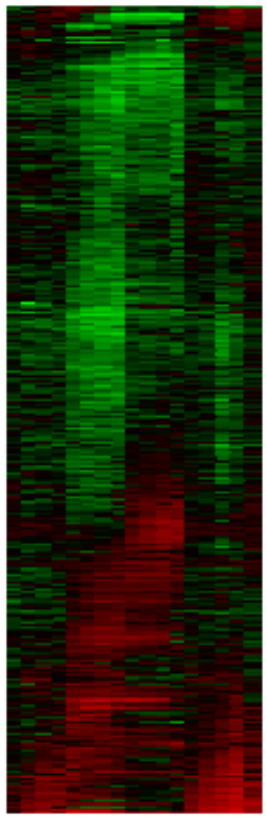
Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

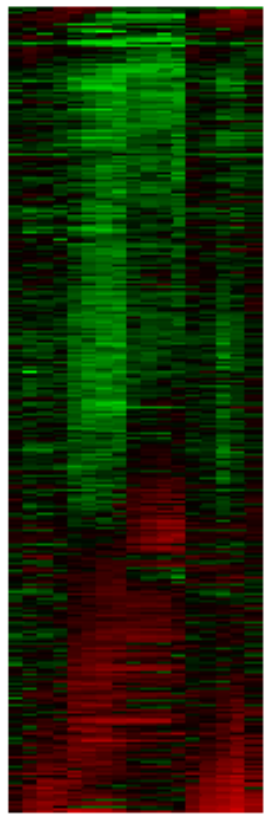
Discriminación visual evidente



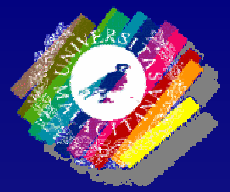
s = 5%



s = 10%



s = 20%



Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Selección de Genes

- *k*-Feature Selection (*k*-FS)
  - *m* ejemplos $e_i, \cdots, e_m$
  - $e_i = (x_i, t_i) \in \Sigma^n \times \mathcal{C}$

    n atributos etiquetas de clase
  - Encontrar $S \subseteq \{1, \cdots, n\}$ tal que
    - $|S| = k$
    - $f : \Sigma^k \to \mathcal{C}$ existe, con $f\left(x_i^{S_1}, \cdots, x_i^{S_k}\right) = t_i$
- El problema es *NP-hard*. So what?
- El problema es *W[2]-hard*. Gotcha!

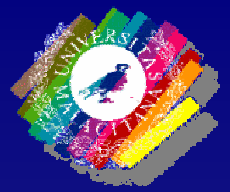
Selección Robusta

- (α,β)-*k*-Feature Selection:
  - Encontrar $S \subseteq \{1, \cdots, n\}$ tal que
    - $|S| = k$
    - $t_i \neq t_j \Rightarrow \exists S_{ij} \subseteq S$ tal que
      - $|S_{ij}| \geqslant \alpha$
      - $\forall r \in S_{ij}\ x_i^r \neq x_j^r$

      } Discriminación inter-clase
    - $t_i = t_j \Rightarrow \exists S'_{ij} \subseteq S$ tal que
      - $|S'_{ij}| \geqslant \beta$
      - $\forall r \in S'_{ij}\ x_i^r = x_j^r$

      } Consistencia intra-clase

El problema es también *NP-hard* y *W[2]-hard*.



Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Resolución Multi-Nivel

- Uso de un algoritmo evolutivo para discretización de los datos.
- Kernelización segura a través de una reducción a Red-Blue Dominating Set.
- Aplicación de un criterio de selección voraz, y vuelta a kernelizar.
- Se consiguen clasificadores robustos.

Universidad de Málaga

Inferencia Filogenética

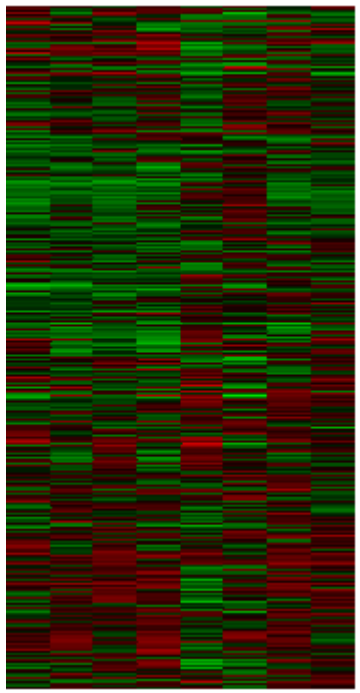
Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

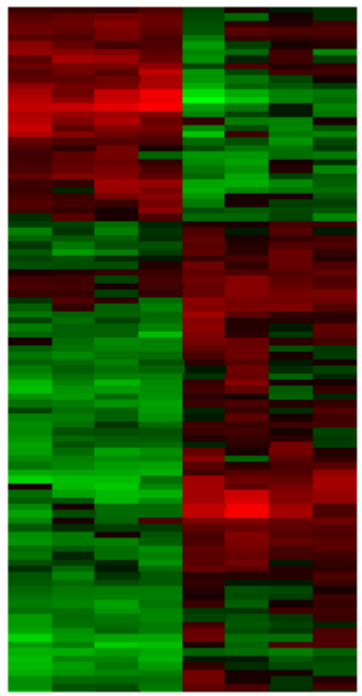
Algunos Resultados

Datos Leucemia

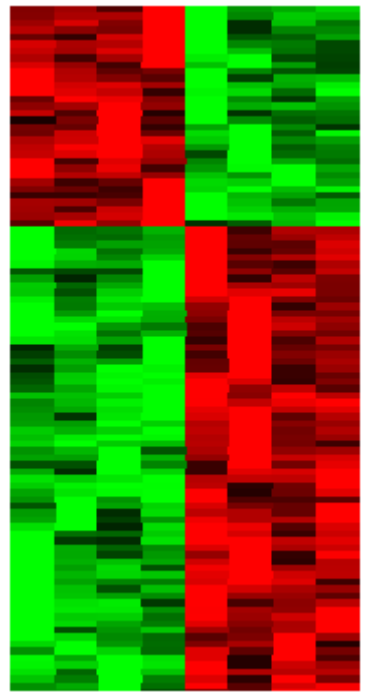
$\alpha = \beta = 100$



Datos originales (2984 genes)



Genes seleccionados (100 genes)



Re-normalización de los datos de expresión

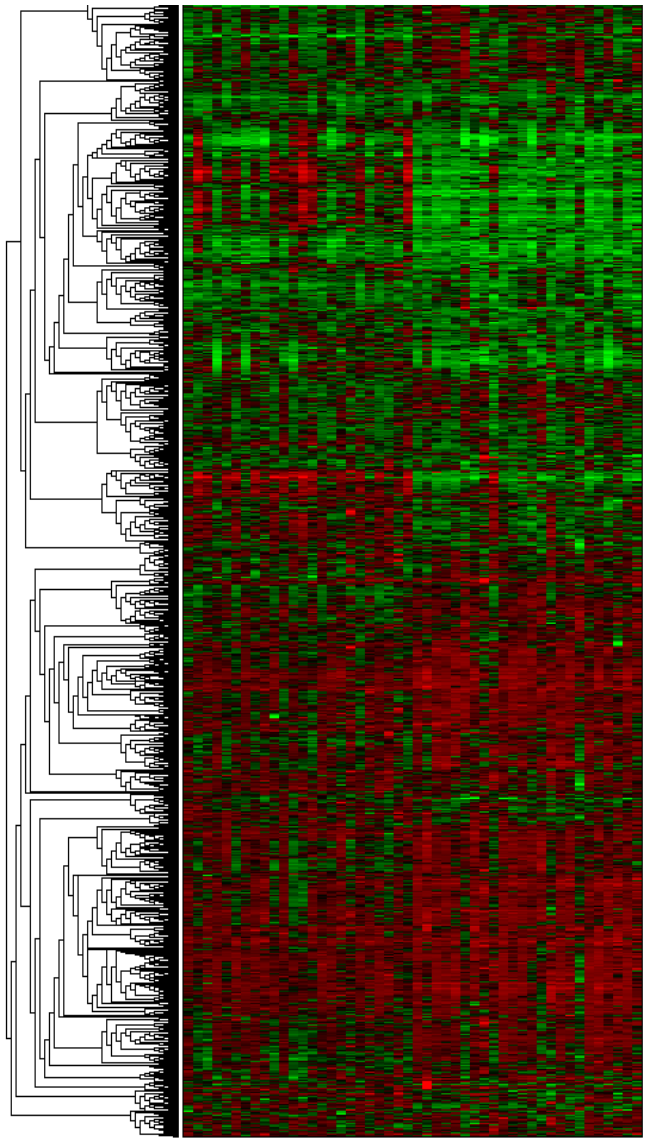
Universidad de Málaga

Aplicación a la Enfermedad de Alzheimer

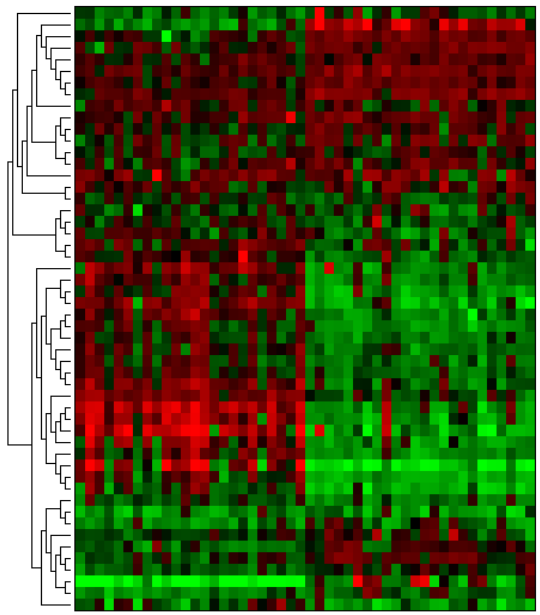
Inferencia Filogenética

Análisis de Datos de Expresión Genética

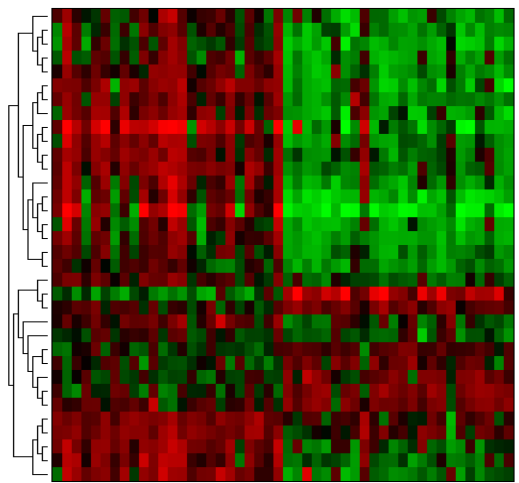
Redes Genéticas y Metabólicas



Conjunto completo de datos (2100 genes)



Modelo Programación entera (52 genes)



Brown *et al.* (2002) (34 genes)

Universidad de Málaga

La Unión Hace la Fuerza

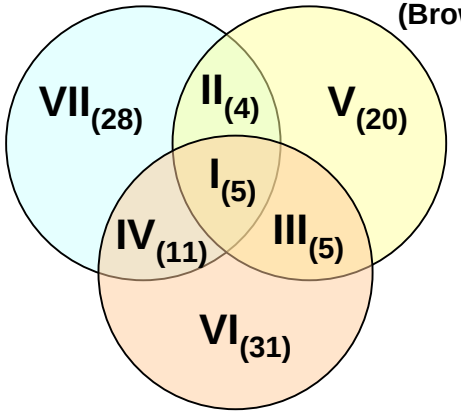
Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

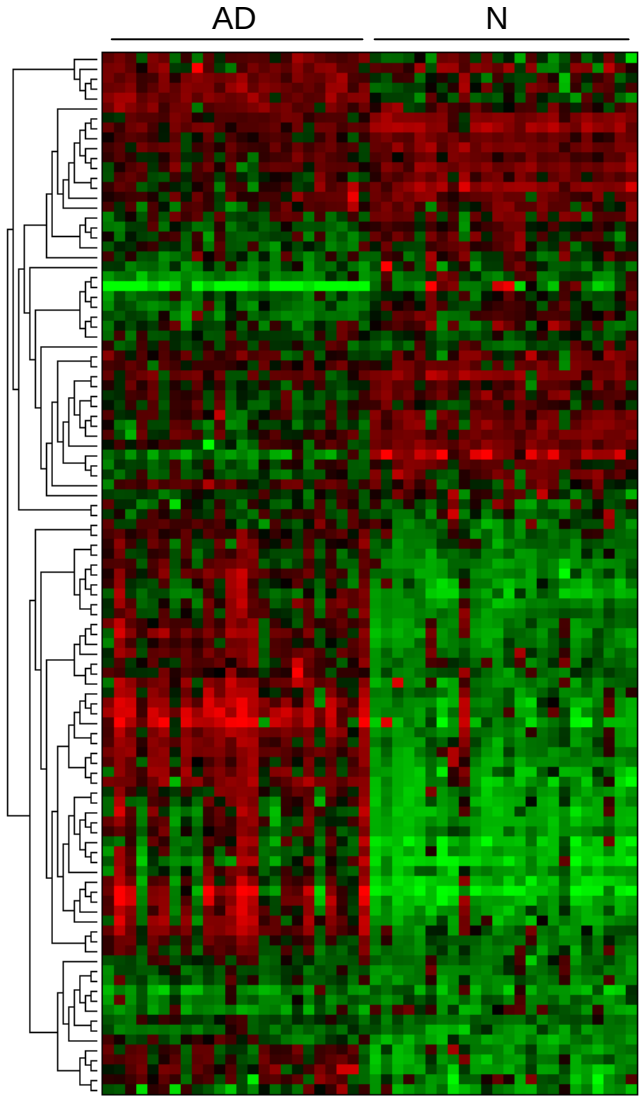
Búsqueda evolutiva

Single Value Decomposition (Brown *et al*. 2002)

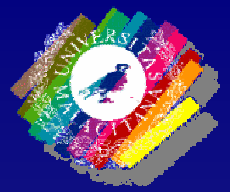


Modelo Programación entera

I – DNCI1, KIAA0069, LOC51628, NR1I3, TAF2F
II – **IDI1**, **MAPK10**, **WASF1**, RAP2A
III – ICAP-1A, FOXJ3, KIAA0992, LOC51235, **YWHAH**
IV – HAX1, LOC54460, **PRDX5** + 8 ESTs (Clone IDs 377827, 395436, 669471, 858450, 884653, 1032362, 1161775, 1500241)
V – BICD1, CCS, **COX7B**, DRAP1, DSCR1L1, IDH3A, LIMS1, NFATC3, **PRKCB1**, PSCA, PSCD2, PTPRN2, RAB2, **RARS**, SALL2, SEPW1, **SMS**, TIF1, XPO1, ZNF142
VI – ADD1, ATP6F, CANPX, CYBA, EIF2C1, EIF4B, FLJ11132, FLJ11200, GLG1, GNG10, INSL4, KIAA0154, KIAA0608, **MAPK14**, MCF2, NFIX, THBS1, TPD52, USP16 + 12 ESTs (Clone IDs 246116, 308788, 768324, 824479, 867751, 868188, 1034472, 1291971, 1292501, 1292893, 1493181, 1505783)
VII – APOC4, ATP5G3, FLJ11220, FLJ12895, FLJ20323, GAPDH, IL11RA, KIAA0308, LOC153561, NFKBIB, PPP2R1A, S100A11, SIAH1, SLC2A5, SLC9A6, SMAP, SRI, SRP46, Z39IG + 9 ESTs (Clone IDs 48906, 147192, 462944, 469379, 796548, 813813, 126858, 1493137, 1505240)



Unión de los tres enfoques (105 genes)



Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Bibliografía

[CMG+03] C. Cotta, A. Mendes, V. Garcia, P. França, P. Moscato, **Applying Memetic Algorithms to the Analysis of Microarray Data**, *Applications of Evolutionary Computing*, G. Raidl *et al.* (eds.), LNCS 2611:22-32, Springer Verlag, 2003

[CottaMoscato03] C. Cotta, P. Moscato, **"The *k*-FEATURE SET Problem is W[2]-Complete"**, *Journal of Computer and System Sciences* 67(4):686-690, 2003

[CottaMoscato03] C. Cotta, P. Moscato, **"A Memetic-Aided Approach to Hierarchical Clustering from Distance Matrices: Application to Gene Expression Clustering and Phylogeny"**, *Biosystems* 72(1-2):75-97, 2003

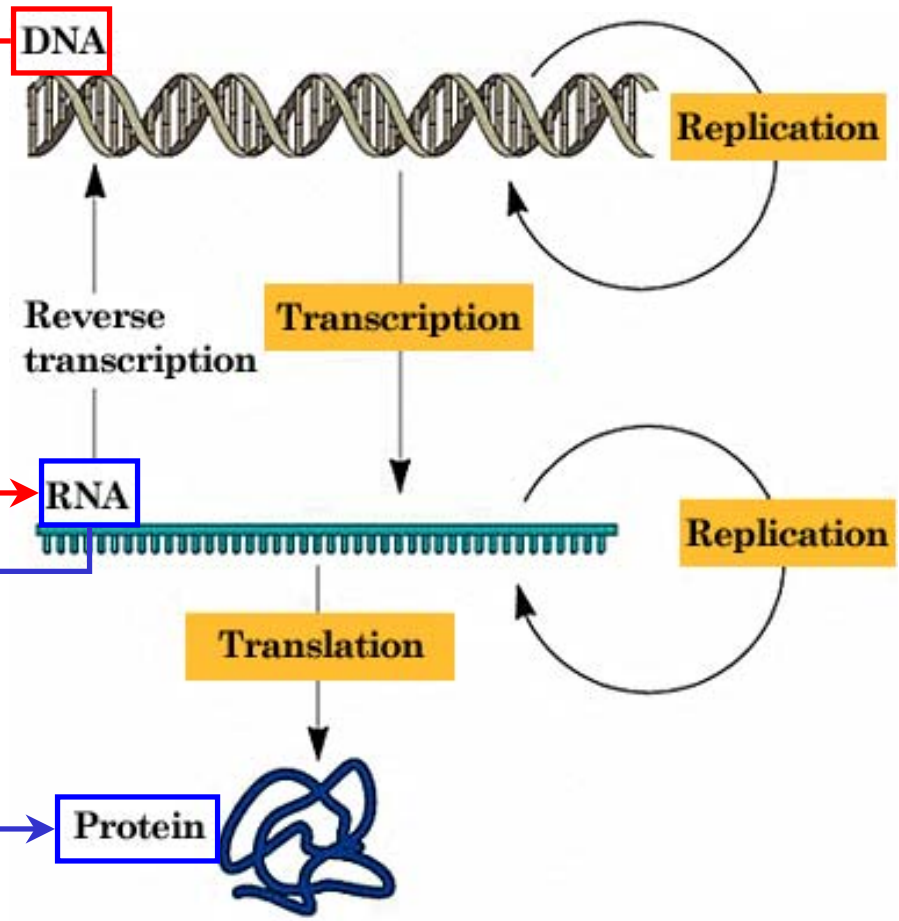
[CSM04] C. Cotta, C. Sloper, P. Moscato, **"Evolutionary Search of Thresholds for Robust Feature Selection: Application to the Analysis of Microarray Data"**, *Applications of Evolutionary Computing*, G. Raidl et al. (eds.), LNCS 3005:21-30, Springer-Verlag, 2004

Redes Genéticas

El Dogma Central de la Biología

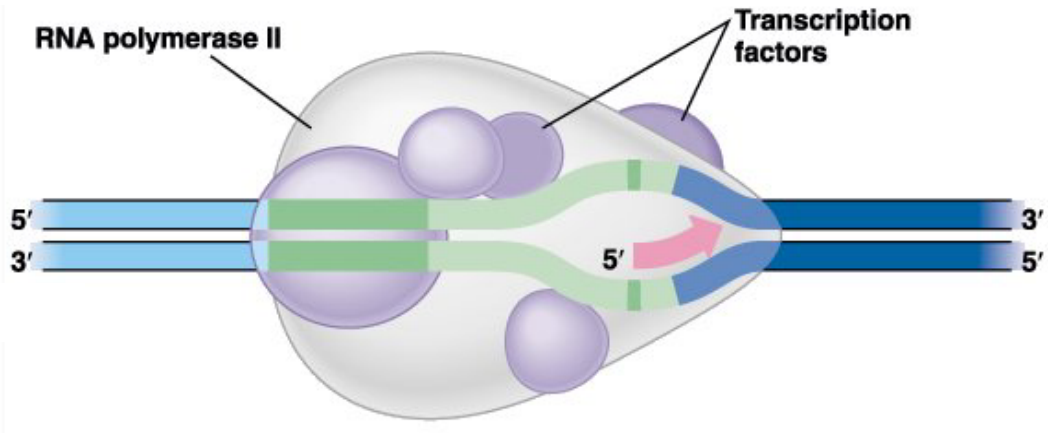
El ADN se transcribe en ARN mensajero

El ARNm se traduce en proteínas



El Proceso de Transcripción

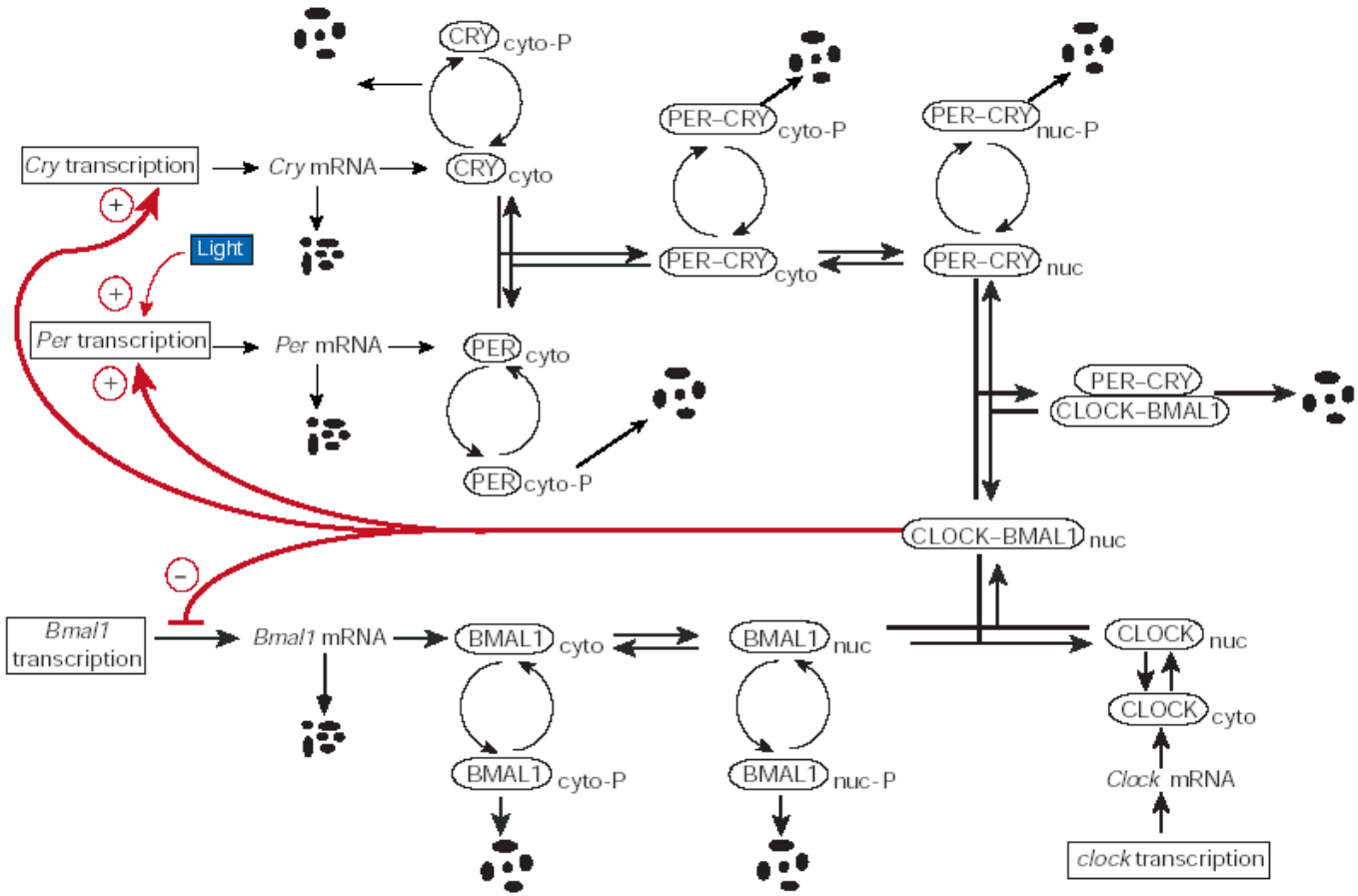
- El proceso de transcripción está controlado por encimas y factores de transcripción, e.g., proteínas



- La expresión de un gen depende por lo tanto del nivel de expresión de otros genes.

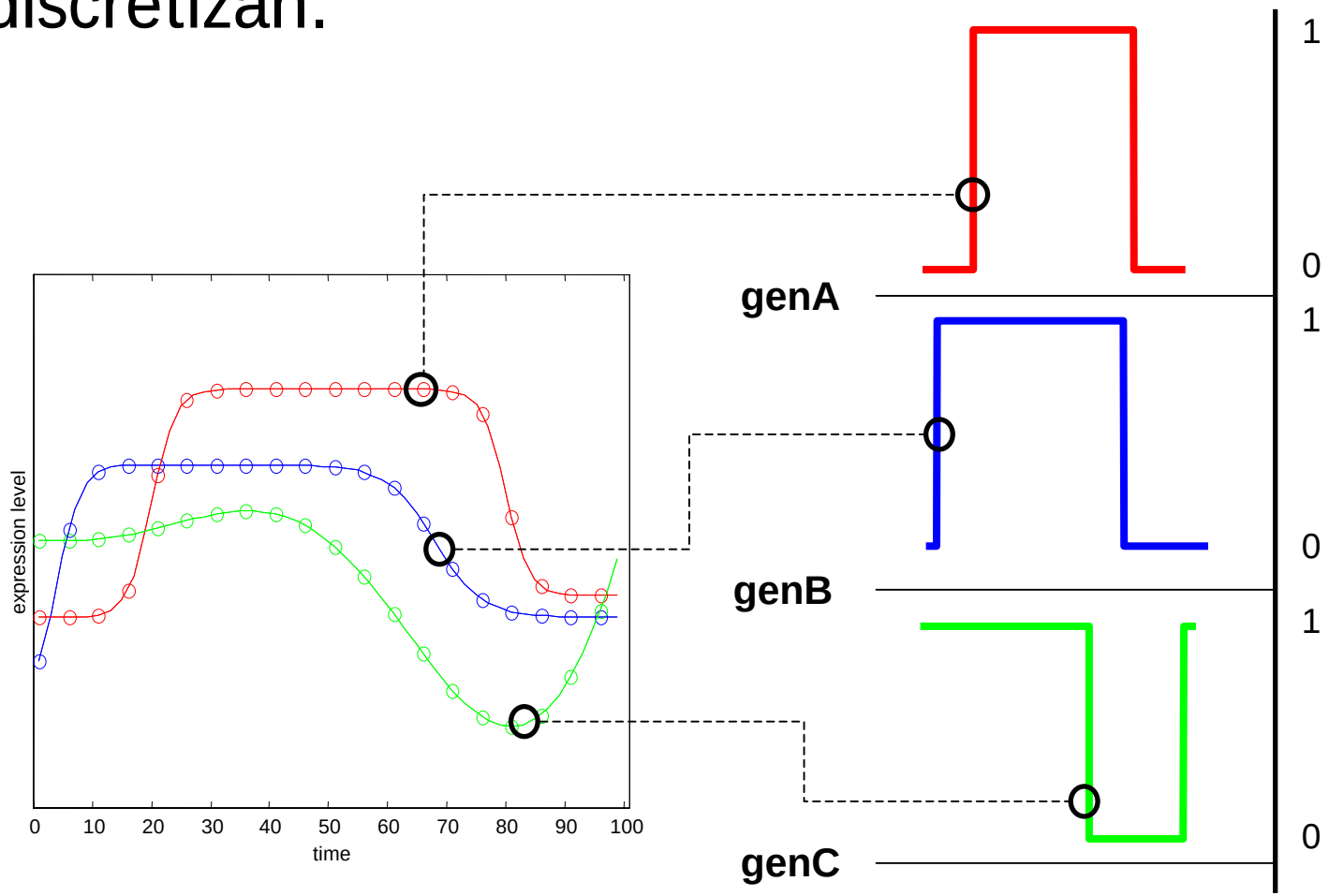
Redes Genéticas

- Esto se modela como una red genética.



Redes Lógicas

- Tiempo y niveles de expresión se discretizan.



Redes Lógicas Temporales

- **Datos de entrada**: una matriz $\Lambda = \langle \lambda_1, \cdots, \lambda_m \rangle$, $\lambda_i \in \mathrm{IB}^n$.

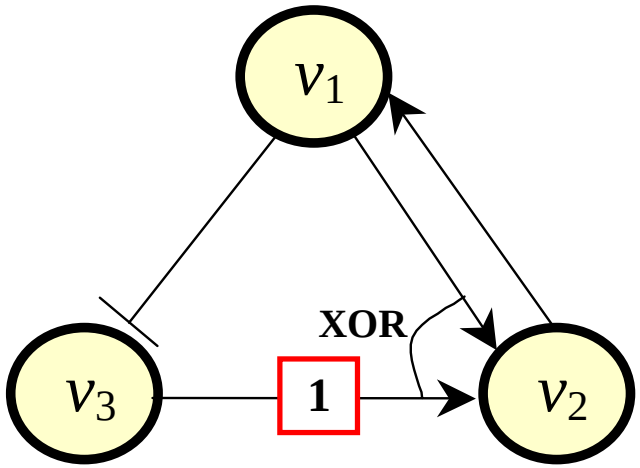
- **Red Lógica**: un triplete (V, E, Φ), donde $V = \{g_1, \cdots, g_n\}$, $E \subseteq V \times V$, y $\Phi = \{\phi_1, \cdots, \phi_n\}$ es un conjunto de funciones lógicas:

$$g_i(t+1) = \phi_i[g_{i1}(t), \cdots, g_{ik}(t)] \quad \forall g_{ij} : (g_{ij}, g_i) \in E$$

- **Redes Lógicas Temporales**: un triplete (V, E, Φ), con $E \subseteq V \times V \times \mathrm{IN}$:

$$g_i(t+1) = \phi_i[g_{i1}(t-d_{i1}), \cdots, g_{ik}(t-d_{ik})] \quad \forall g_{ij} : (g_{ij}, g_i, d_{ij}) \in E$$

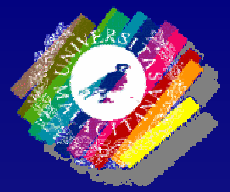
Redes Lógicas Temporales



INPUT			OUTPUT		
v_1	v_2	$v_3^{(-1)}$	v_1	v_2	v_3
0	0	0	0	0	1
0	0	1	0	1	1
0	1	0	1	0	1
0	1	1	1	1	1
1	0	0	0	1	0
1	0	1	0	0	0
1	1	0	1	1	0
1	1	1	1	0	0

$$v_1(t+1) = v_2(t)$$
$$v_2(t+1) = v_1(t) \oplus v_3(t-1)$$
$$v_3(t+1) = \neg v_1(t)$$

A partir del estado de los genes en los instantes t_{-1} y t_0, pueden computarse todos los estados posteriores.



Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Modelado de Redes Lógicas

- Enfoques basados en entropía: ID3, C4.5, ...
- Fuerza bruta: REVEAL, BOOL-1, ...

$$O(mnC_k^n) = O(mn^{k+1})$$

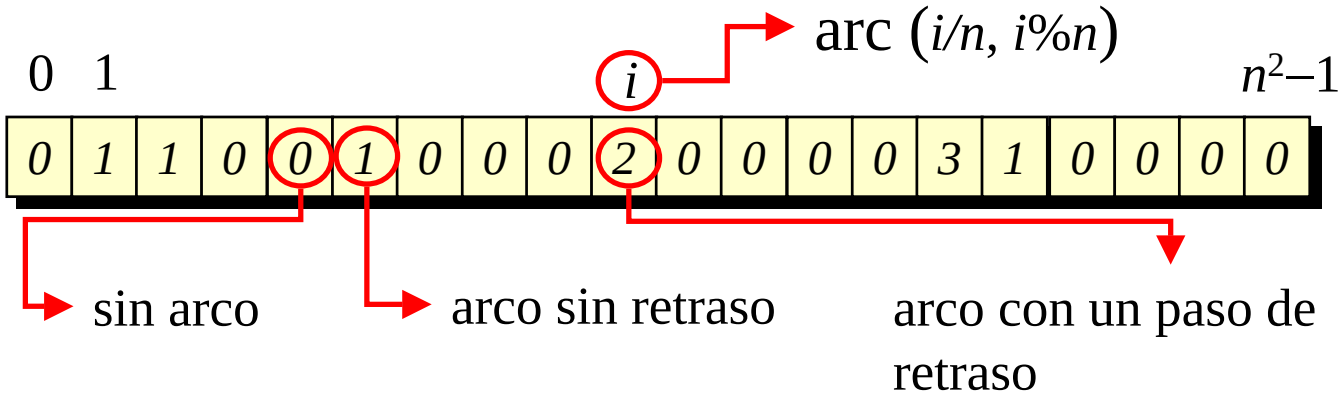
¡No es tratable en parámetro fijo!

¿Es posible mejorar esta complejidad?

¡No, imposible! El problema es W[2]-completo.

Representación

- Genotipo: un vector n^2-dimensional.



- Obtención del Fenotipo:

$$\phi_i(b_{i1},\cdots,b_{ik}) = b \Leftrightarrow$$

$$\Leftrightarrow \sum_{\lambda_j \mid g_i,g_{i1},\cdots,g_{ik} = b,b_{i1},\cdots,b_{ik}} 1 > \sum_{\lambda_j \mid g_i,g_{i1},\cdots,g_{ik} = 1-b,b_{i1},\cdots,b_{ik}} 1$$

Evaluación

- Precisión: fracción de estados predecidos correctamente por la red.

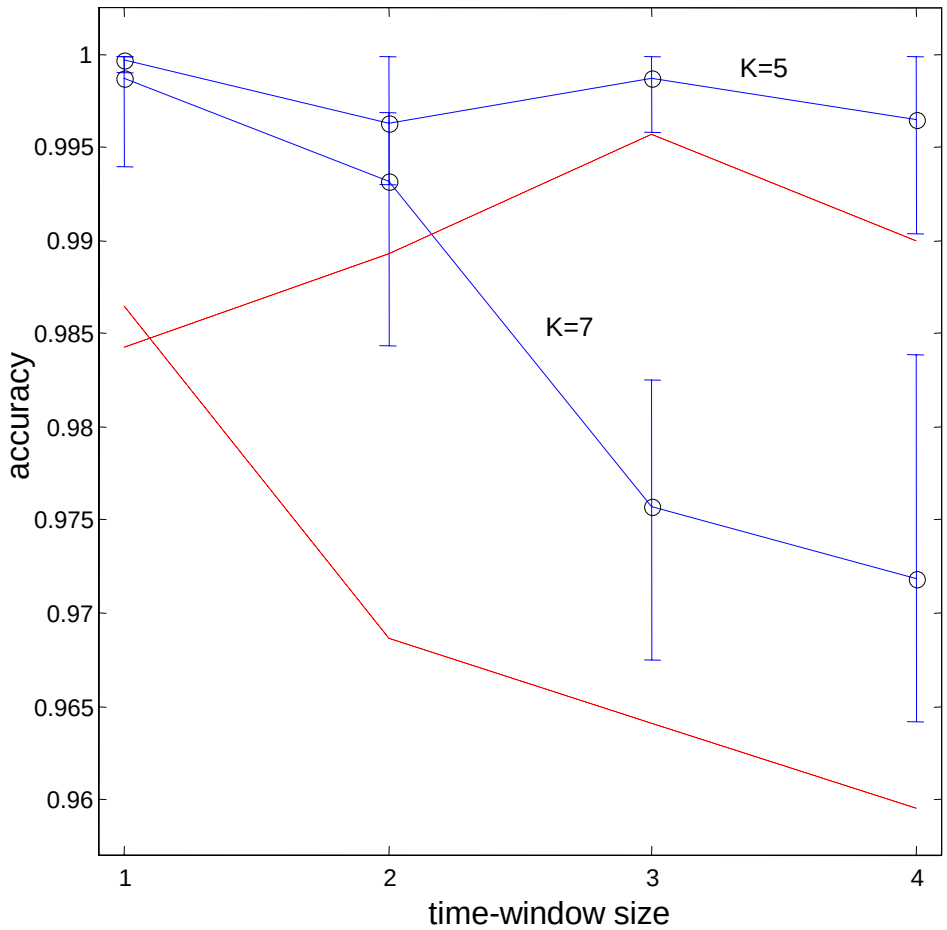
$$accuracy_{TBN}(\Lambda) = 1 - \frac{1}{|V|}\sum_{g_i \in V} \varepsilon^{\Lambda}_{TBN}(g_i)$$

$$\varepsilon^{\Lambda}_{TBN}(g_i) = 1 - \frac{1}{|\Lambda| - D}\sum_{\lambda_j \in \Lambda} \delta\left[\lambda_j|_{g_i}, \phi_i\left(\lambda_j|_{g_{i1},\cdots,g_{ik}}\right)\right]$$

máximo factor de retraso en la red

Algunos Resultados

Precisión (n=16)



Universidad de Málaga

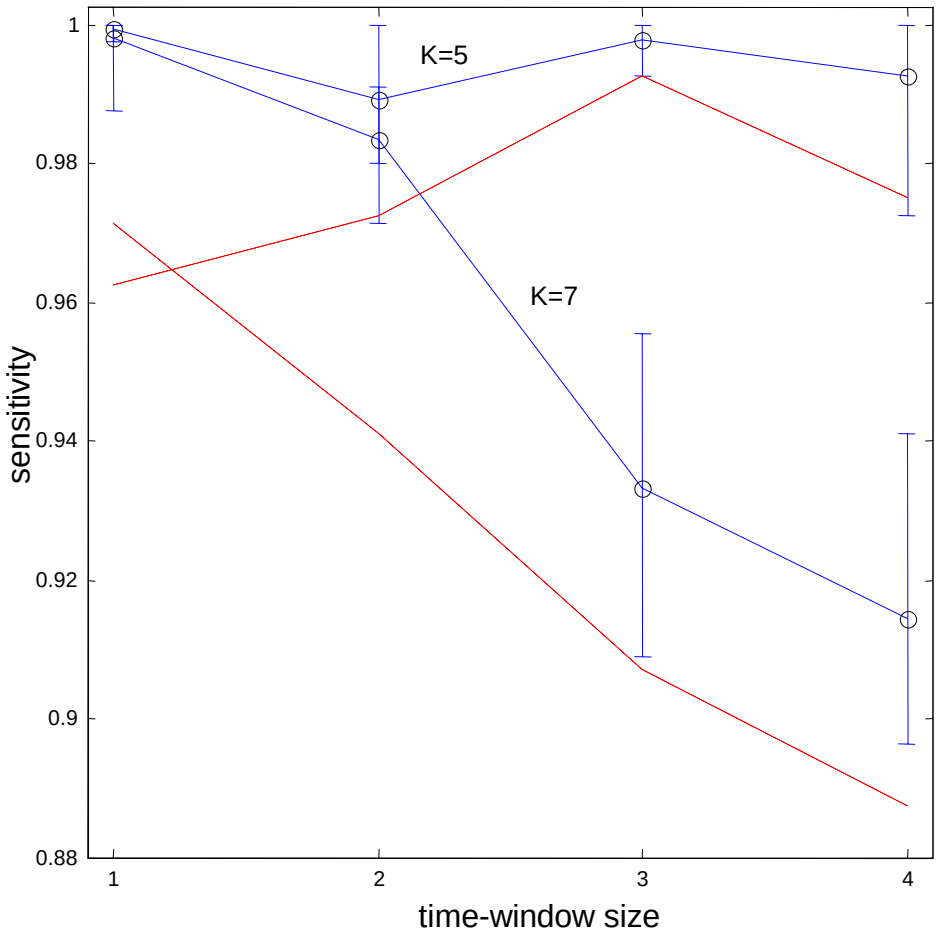
Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Algunos Resultados

Sensibilidad (n=16)



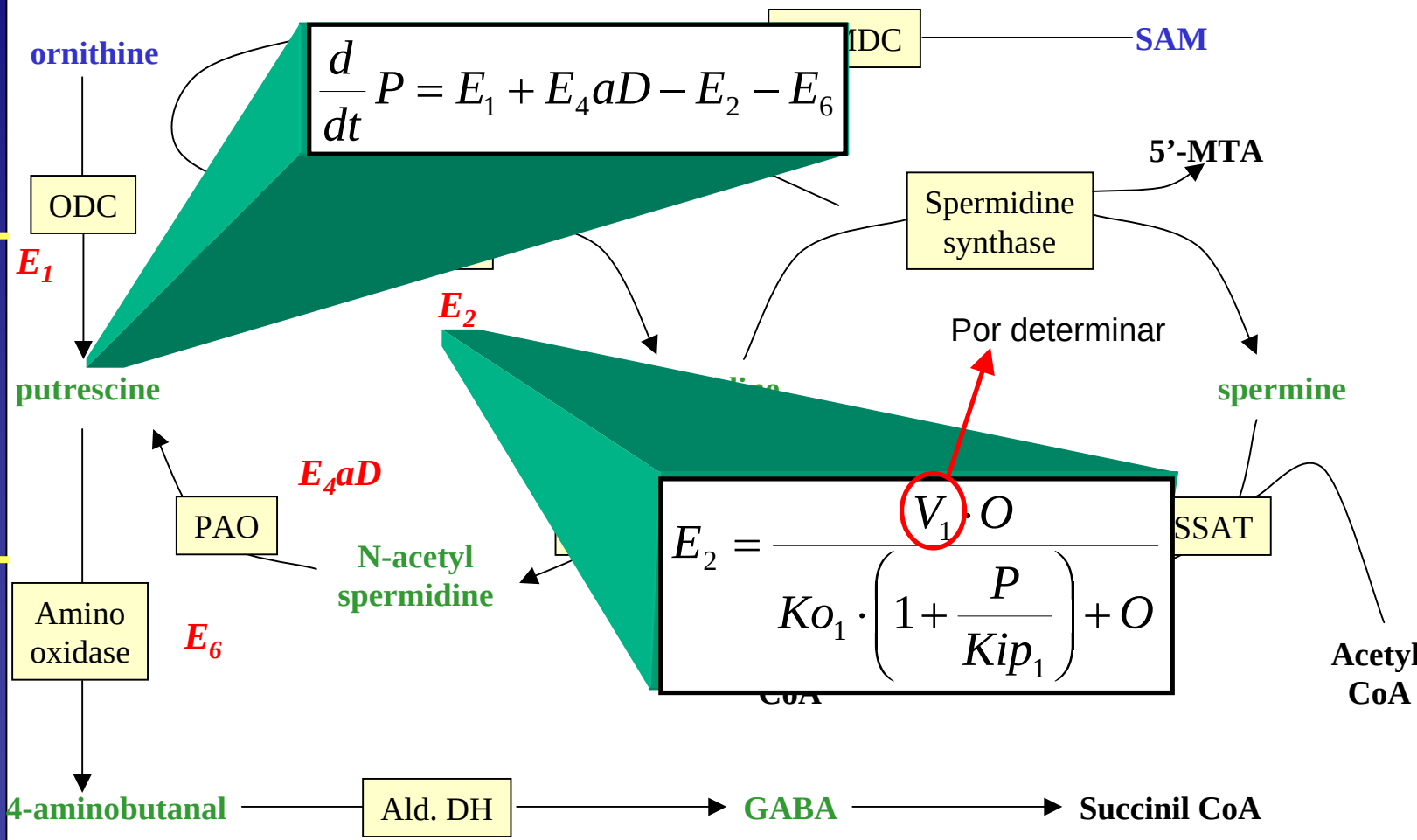
Universidad de Málaga

Redes Metabólicas

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas



Universidad de Málaga

Inferencia Filogenética

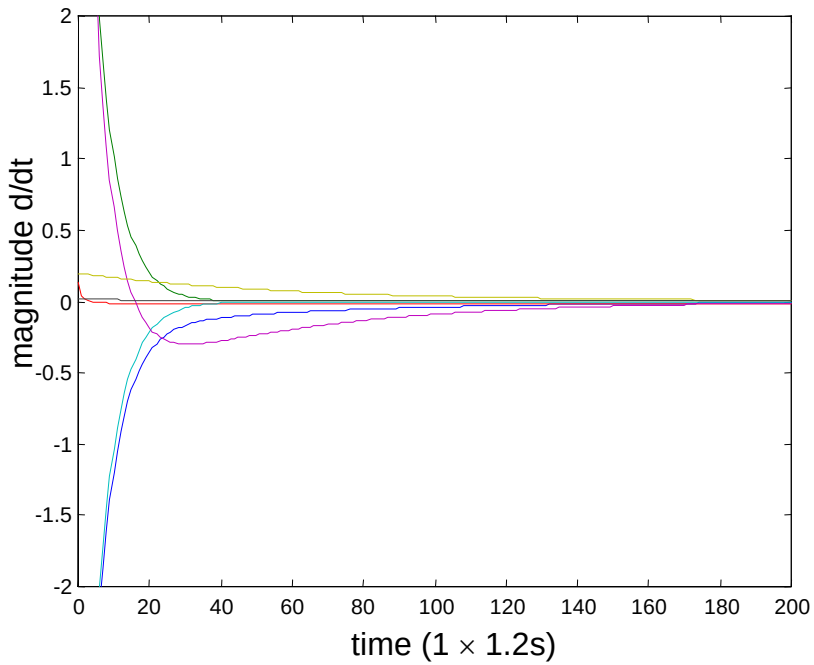
Análisis de Datos de Expresión Genética

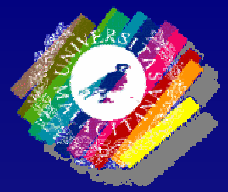
Redes Genéticas y Metabólicas

Redes Metabólicas

- 8 variables continuas por optimizar
- Objetivo: alcanzar estado estacionario en tiempo mínimo
- Simulación de ecuaciones diferenciales
- Restricciones: no-negatividad, límites máximos

El algoritmo evolutivo encuentra una solución factible, y minimiza el tiempo para alcanzar el estado estacionario.





Universidad de Málaga

Inferencia Filogenética

Análisis de Datos de Expresión Genética

Redes Genéticas y Metabólicas

Bibliografía

[Cotta03] C. Cotta, **"On the Evolutionary Inference of Temporal Boolean Networks"**, *Computational Methods in Neural Modeling*, J. Mira, J.R. Álvarez (eds.), LNCS 2686: 494-501, Springer Verlag, 2003

[CottaMoscato03] C. Cotta, P. Moscato, **"The *k*-FEATURE SET Problem is W[2]-Complete"**, *Journal of Computer and*

[CottaTroya04] C. Cotta, J.M. Troya, **"Reverse Engineering of Temporal Boolean Networks from Noisy Data using Evolutionary Algorithms"**, *Neurocomputing*, 2004 (por aparecer)