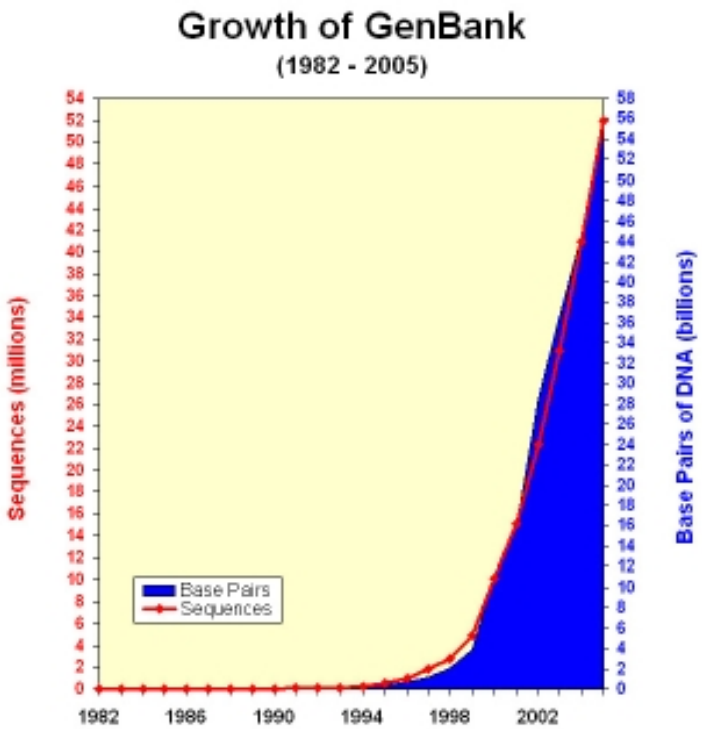
Comparació de grafs en biologia

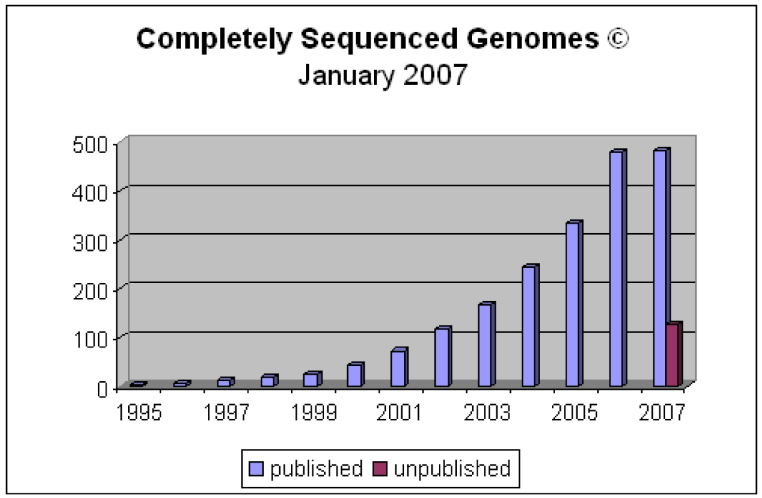
Francesc Rosselló

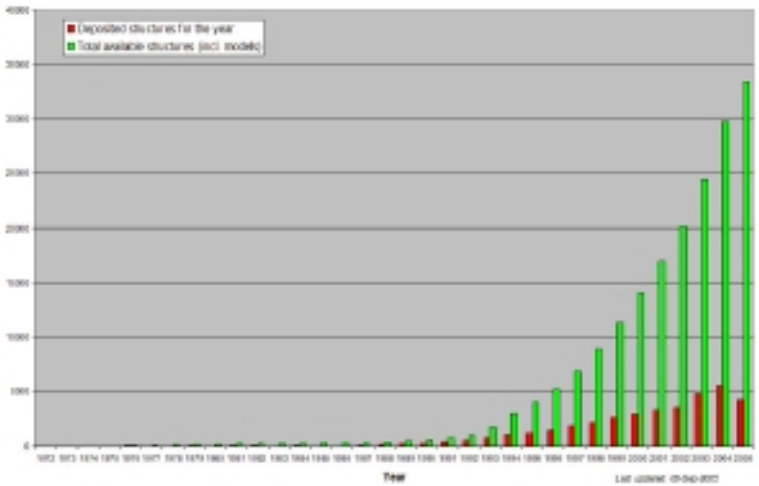
Palma, 8 d'octubre de 2007

Biologia computacional

Els avenços recents en biotecnologia han portat a un creixement exponencial en l'obtenció d'informació genòmica i biomolecular







Biologia computacional

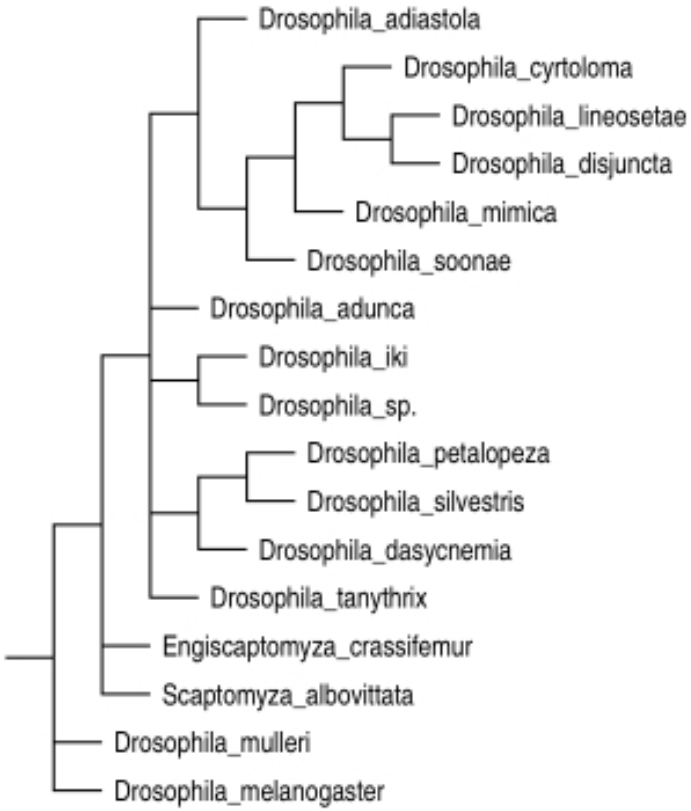
Els avenços recents en biotecnologia han portat a un creixement exponencial en l'obtenció d'informació genòmica i biomolecular

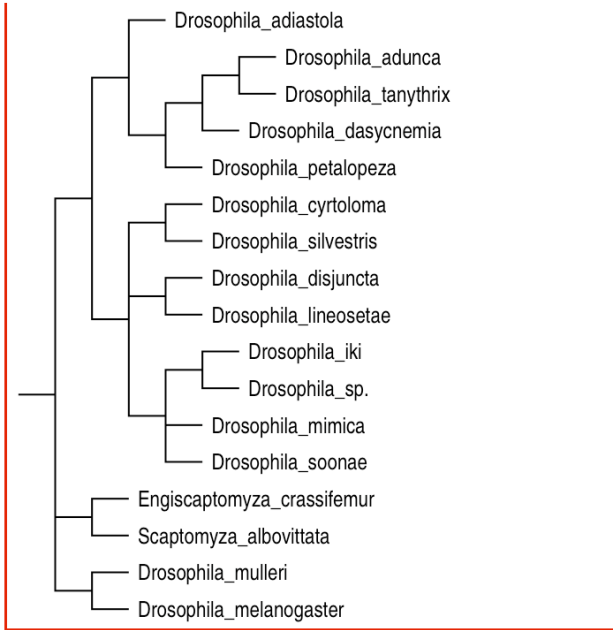
Els biòlegs necessiten eines computacionals per analitzar aquestes dades

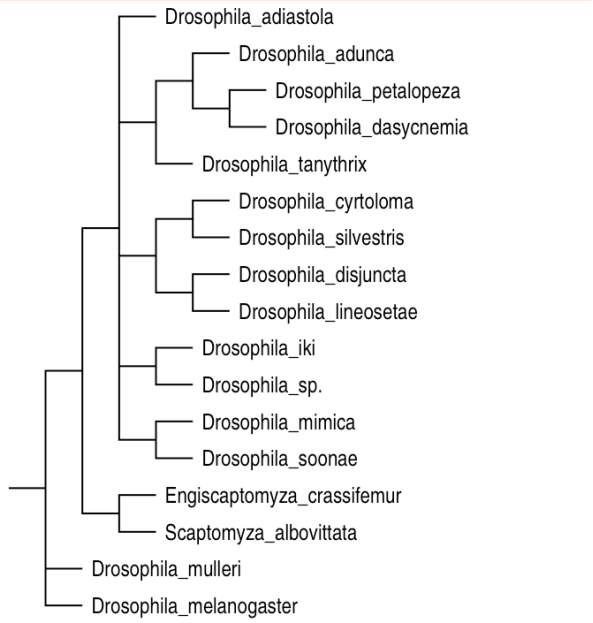
La biologia computacional (i el seu braç armat, la bioinformàtica) les forneixen

Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better (Joel Cohen, 2004)

A quin se sembla més el de l'esquerra?

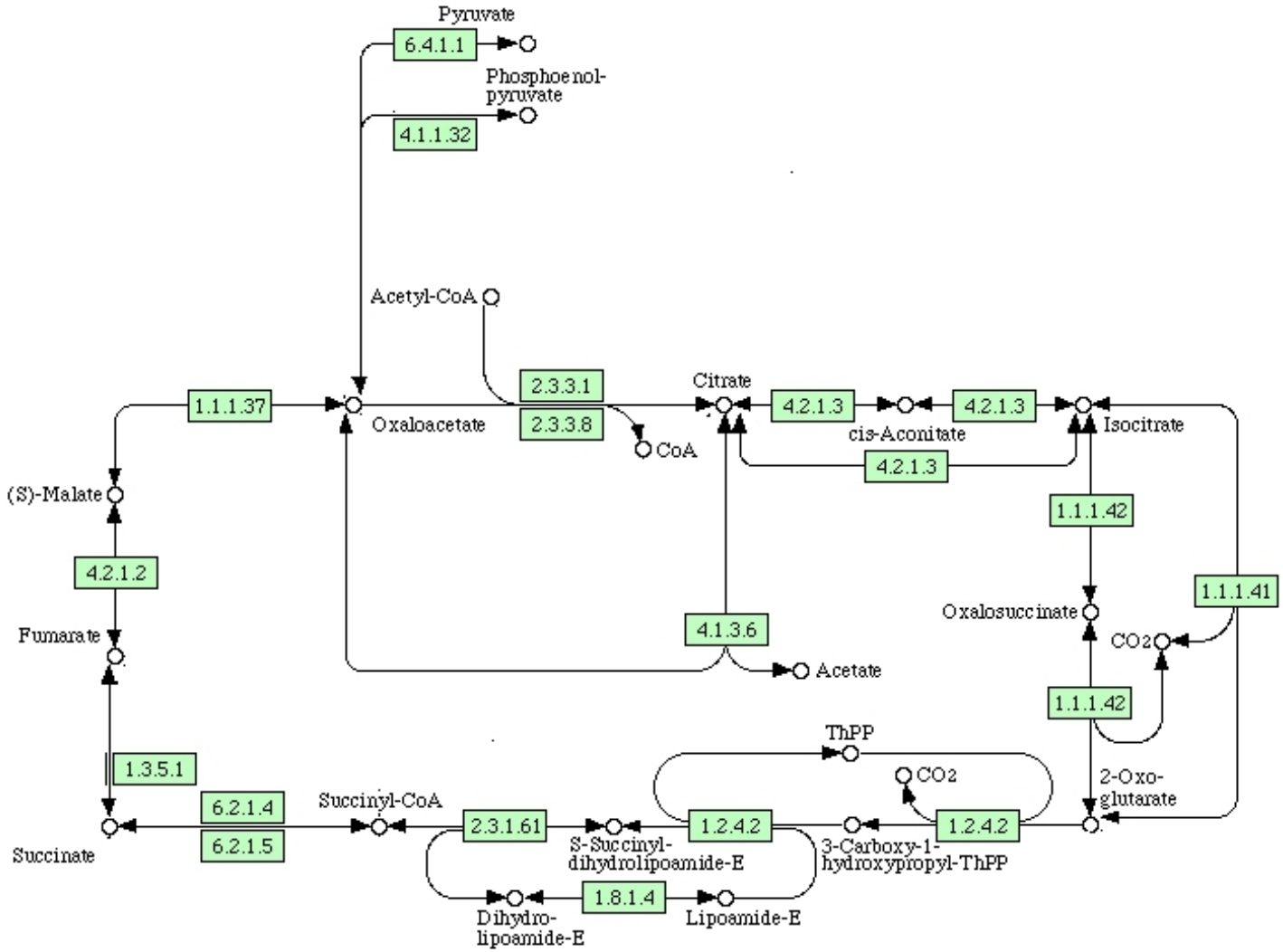






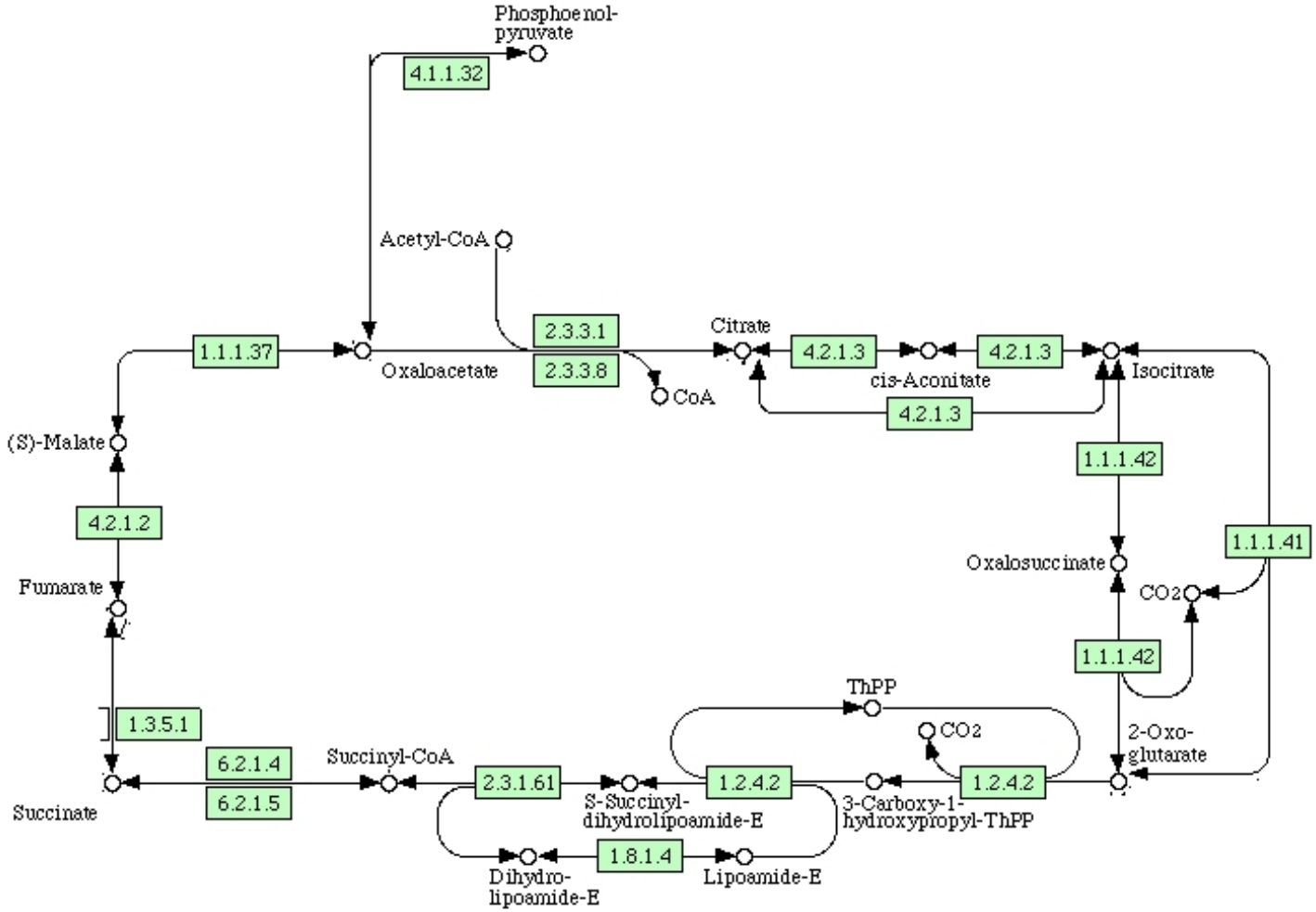
Què tenen en comú?

Cicle de Krebs (Homo sapiens)



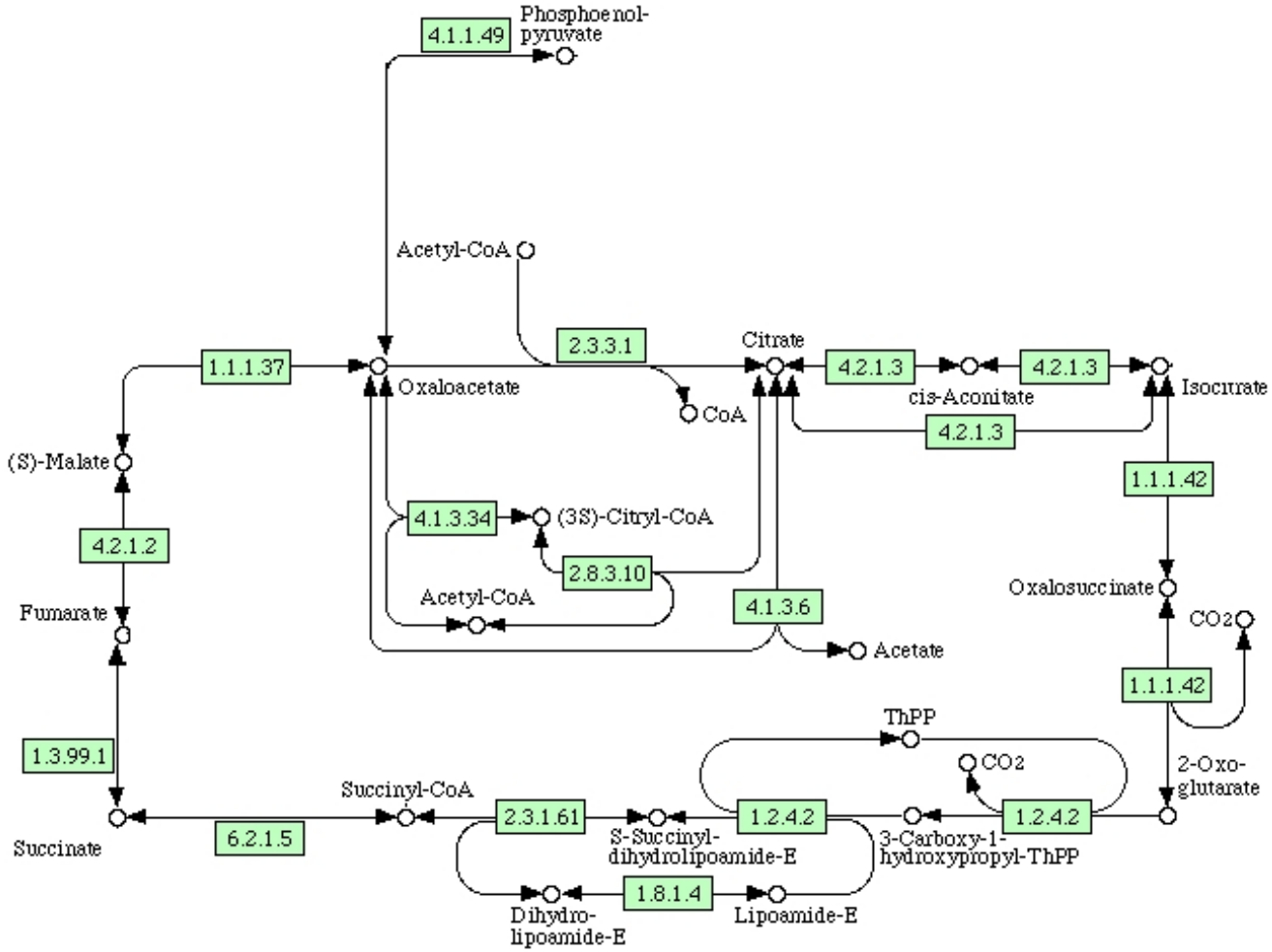
Què tenen en comú?

Cicle de Krebs (Drosophila melanogaster)

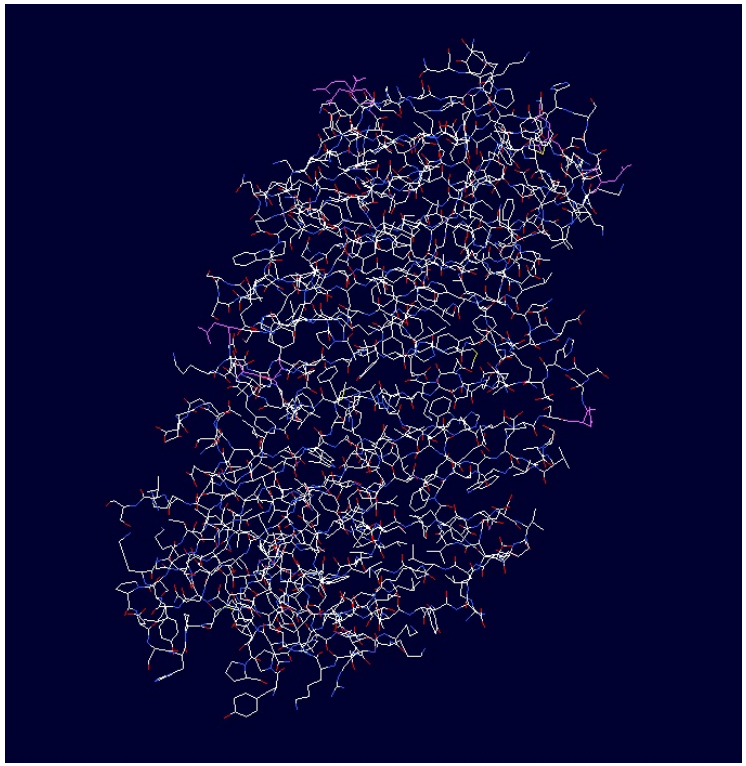


Què tenen en comú?

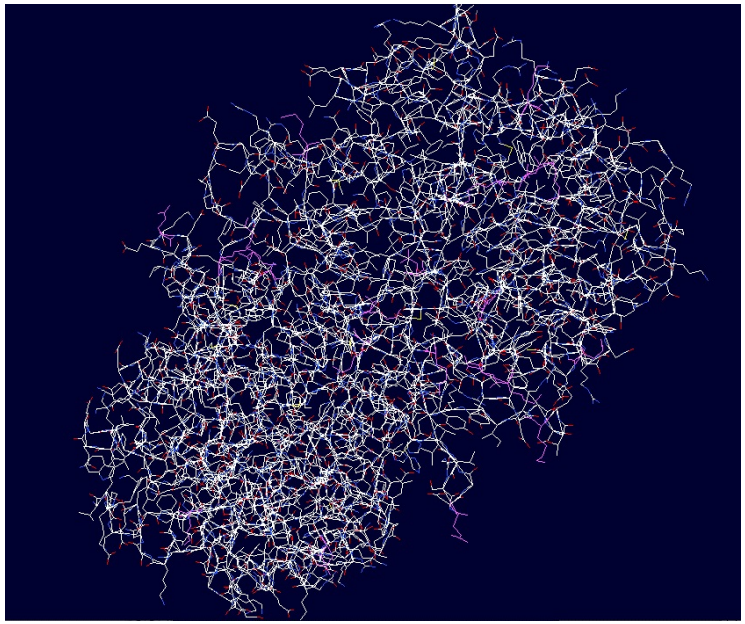
Cicle de Krebs (Escherichia Coli K-12 MG1655)



Se semblen?



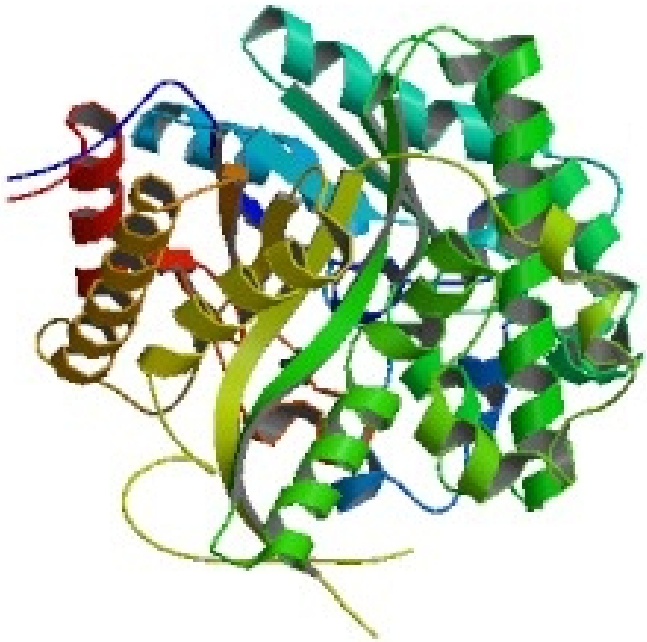
1ces



2ces

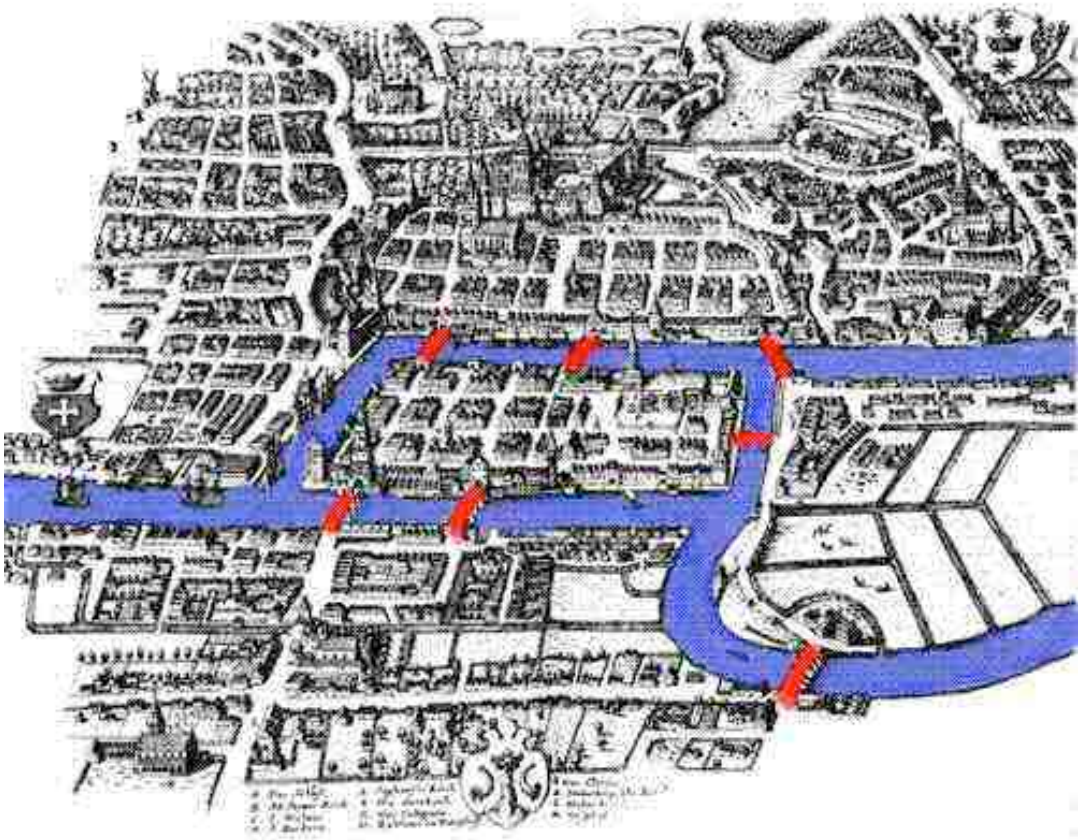
Se semblen?

1ces



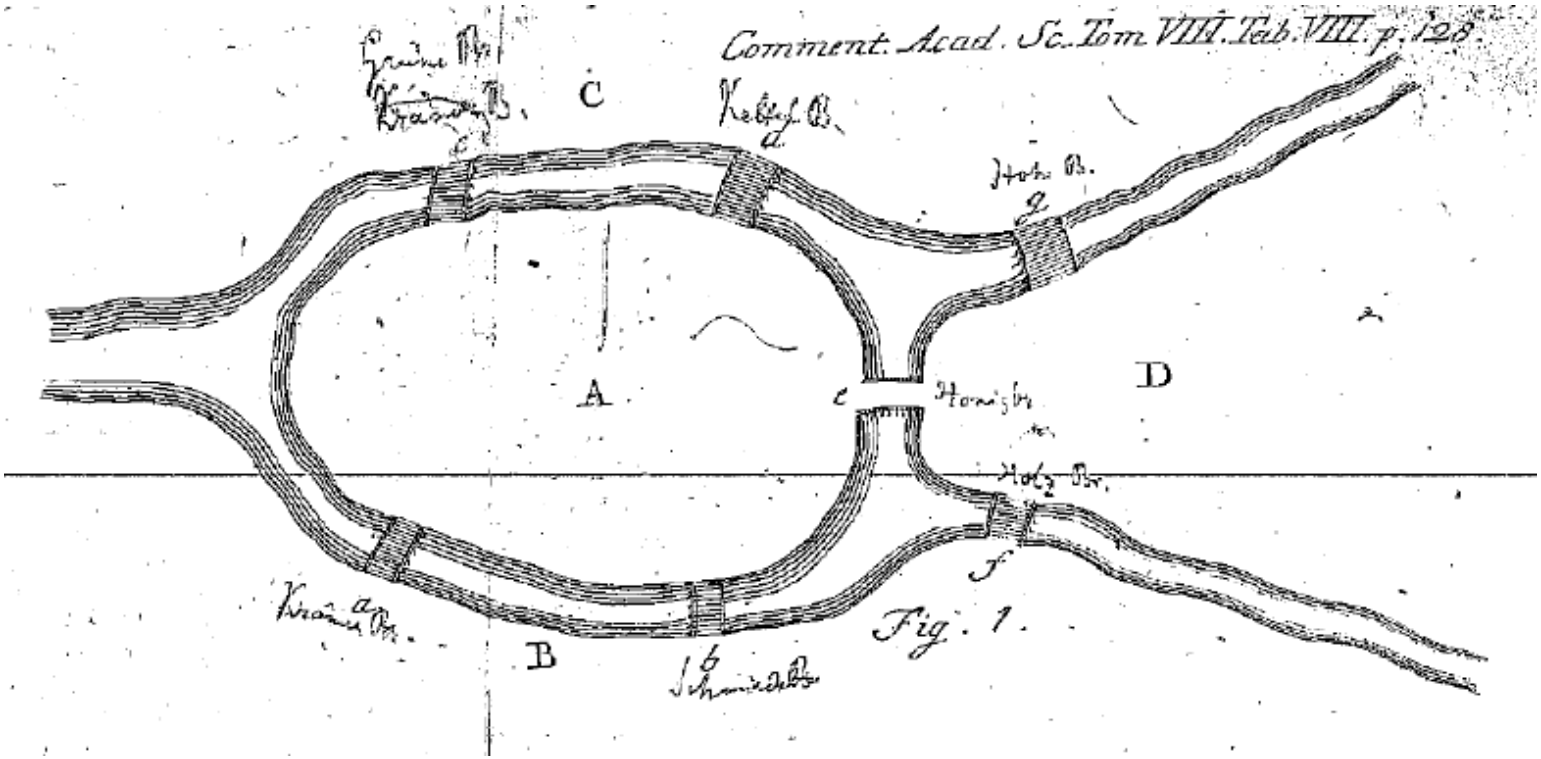
2ces

Tot començà amb Euler (1736)



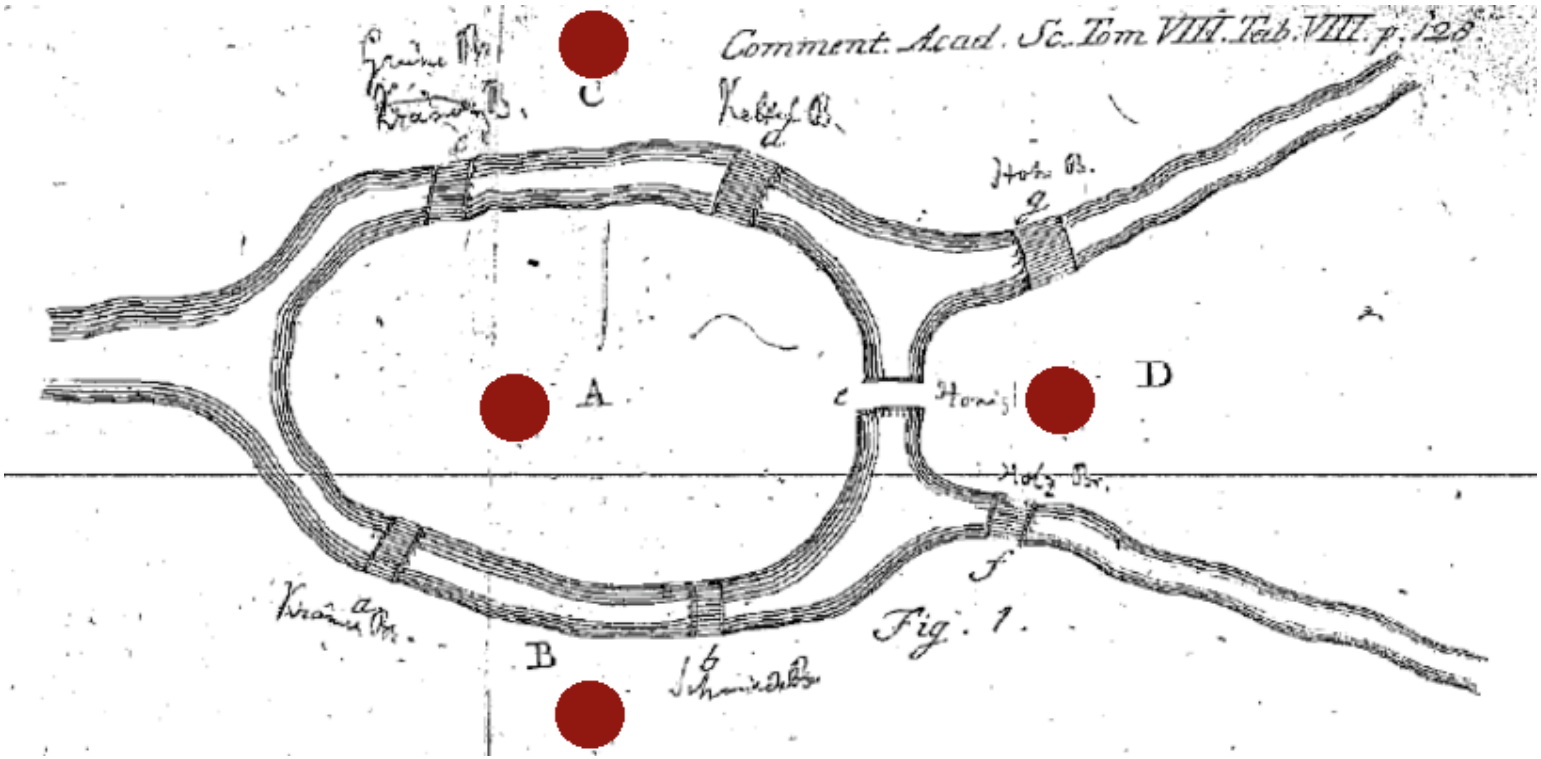
La ciutat de Königsberg

Tot començà amb Euler (1736)



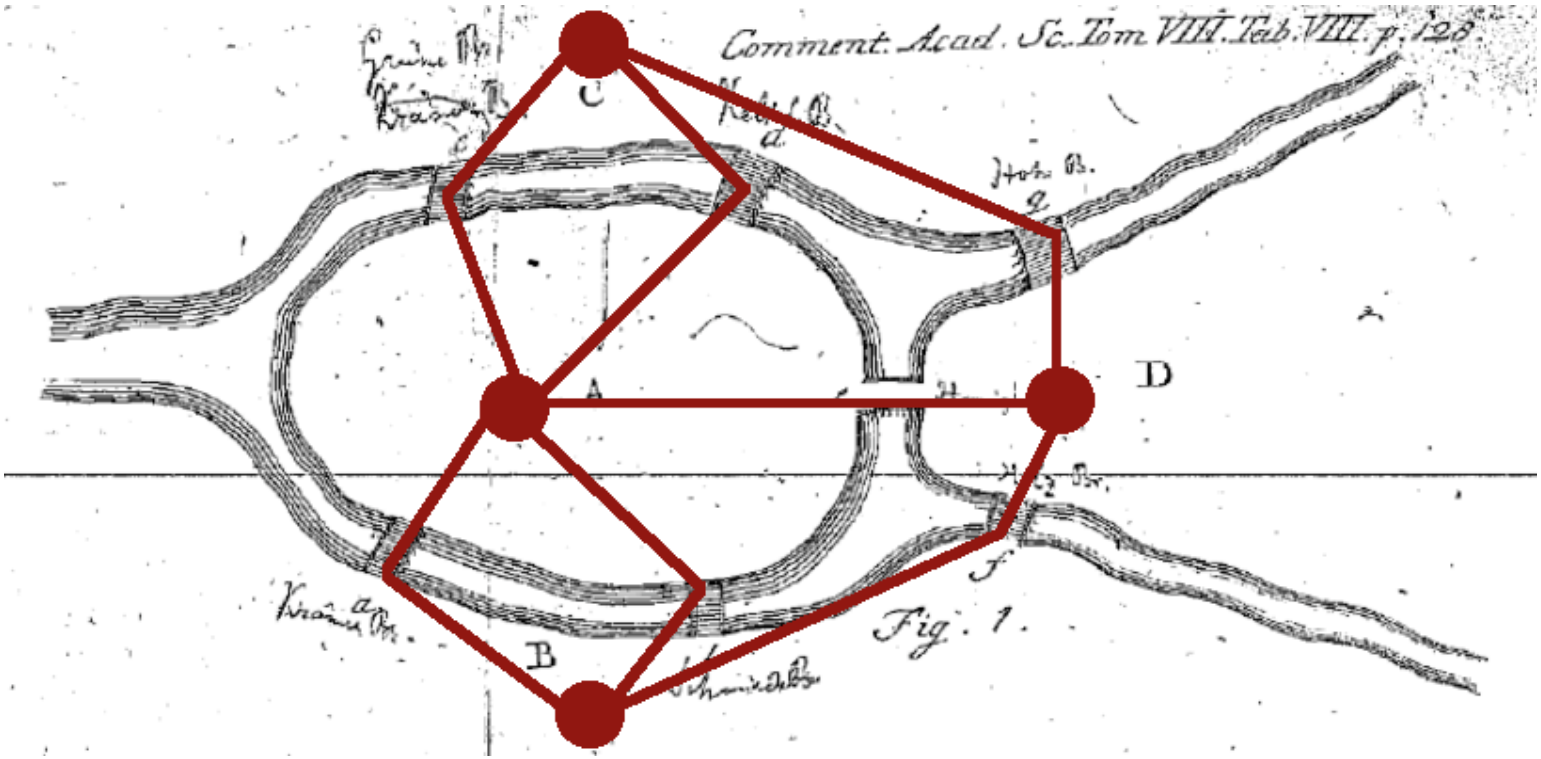
La ciutat de Königsberg, segons Euler

Tot començà amb Euler (1736)



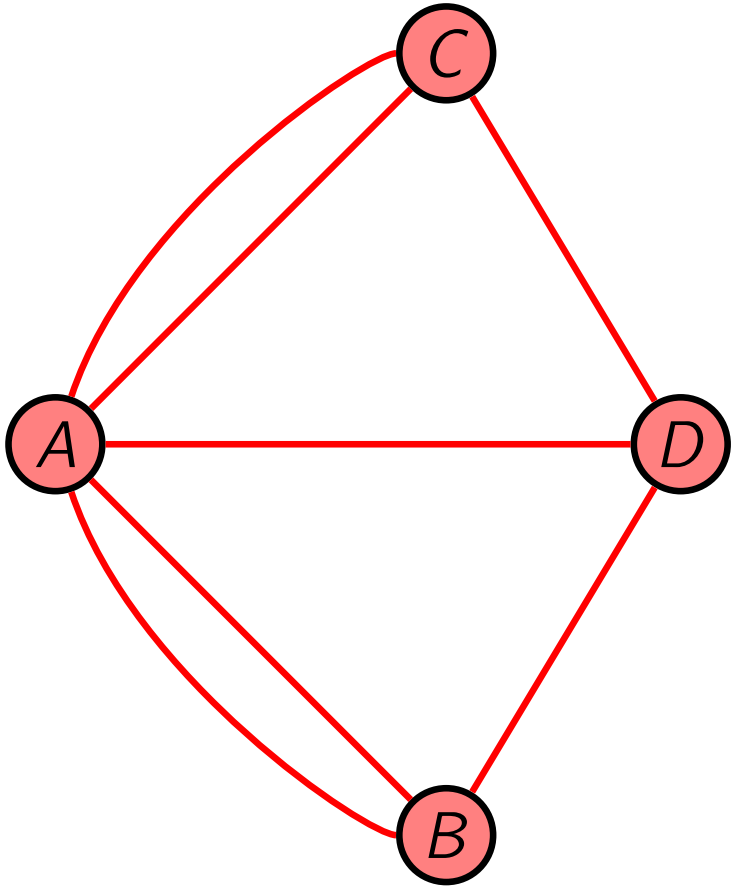
La ciutat de Königsberg, segons Euler

Tot començà amb Euler (1736)



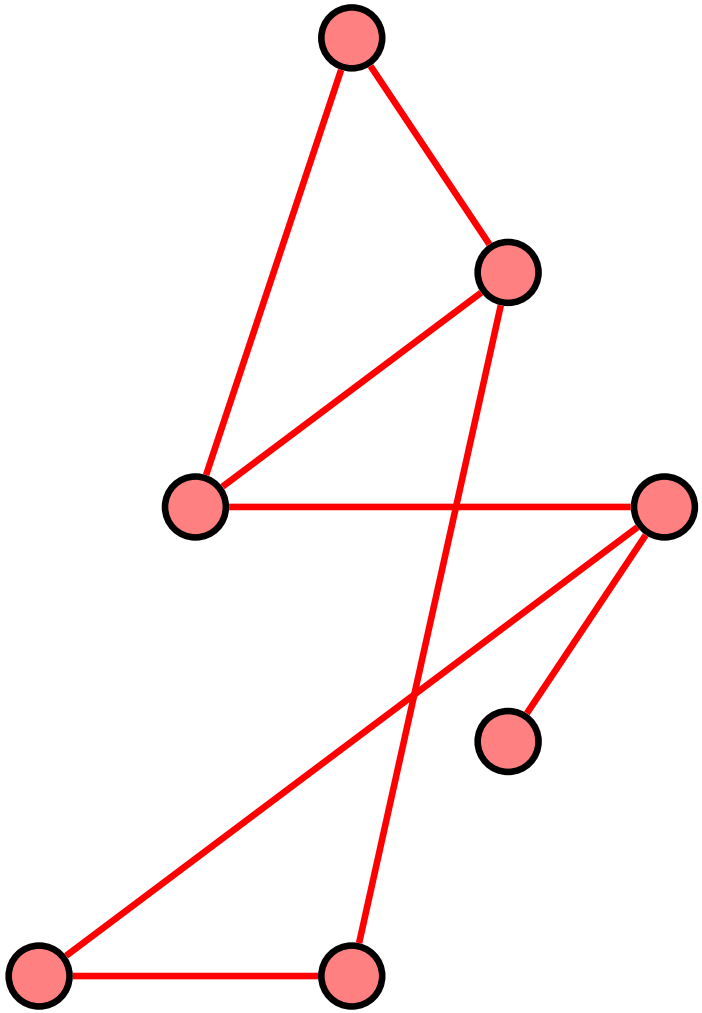
La ciutat de Königsberg, segons Euler

Tot començà amb Euler (1736)



La ciutat de Königsberg, segons Euler, era un (multi)graf

Grafs

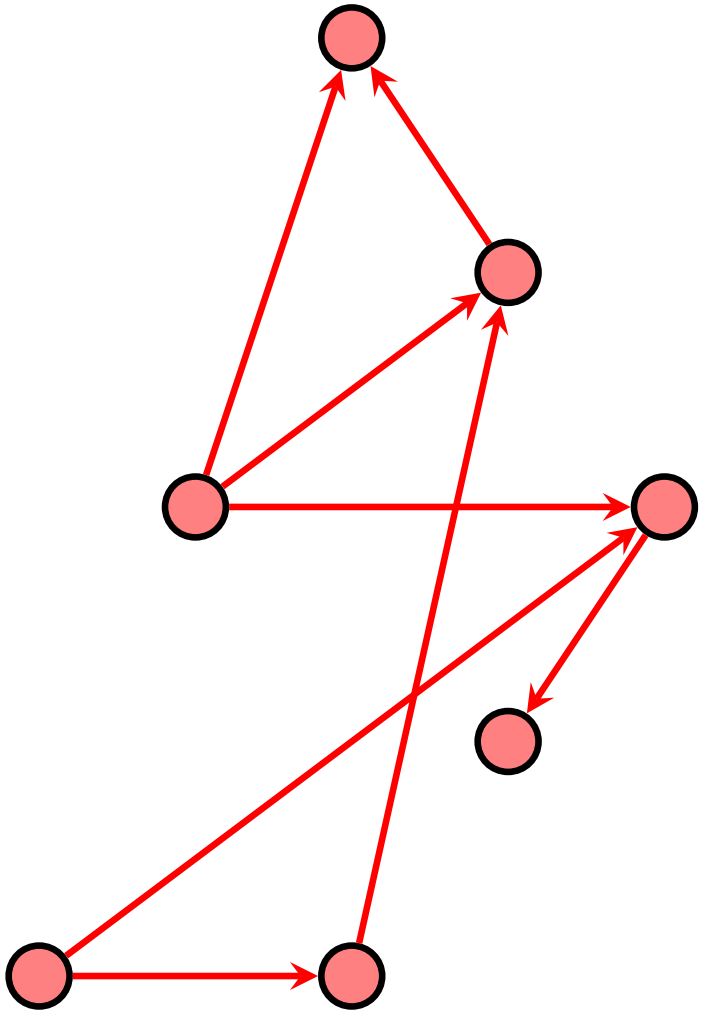


Un graf és:

- Un conjunt finit de punts (nodes)
- Un conjunt de parells no ordenats de nodes (arestes)

El grau d'un node és el nombre d'arestes incidents

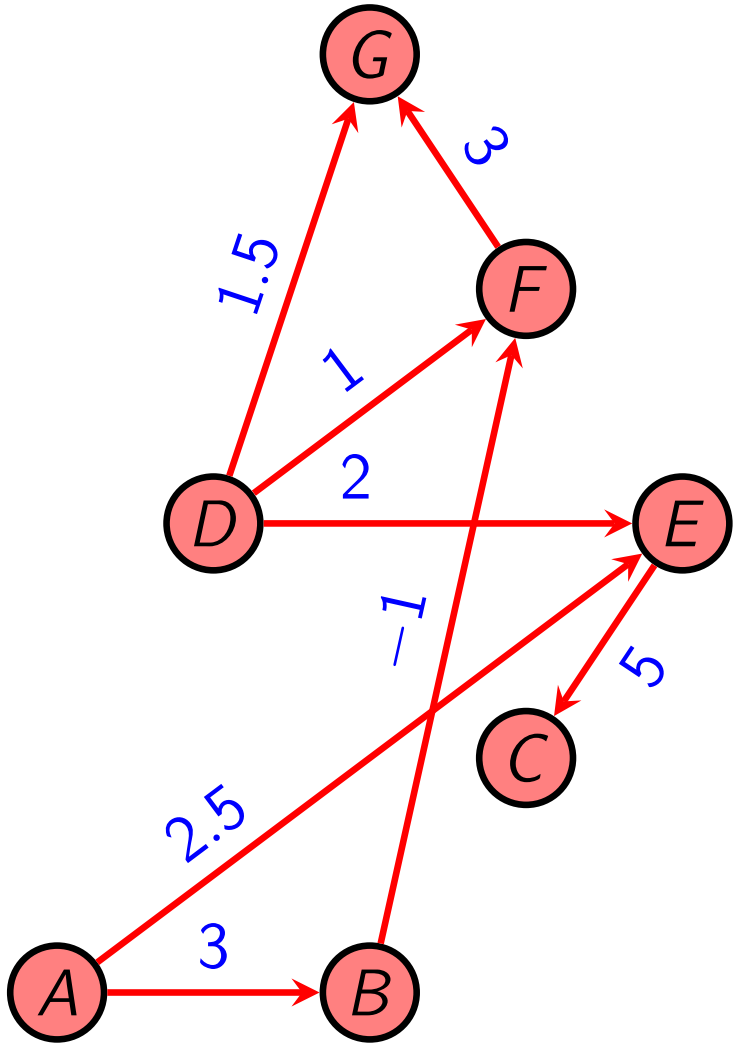
Grafs



Un graf dirigit en lloc d'arestes té arcs (parells ordenats de nodes)

Els nodes aquí tenen grau de sortida i grau d'arribada

Grafs

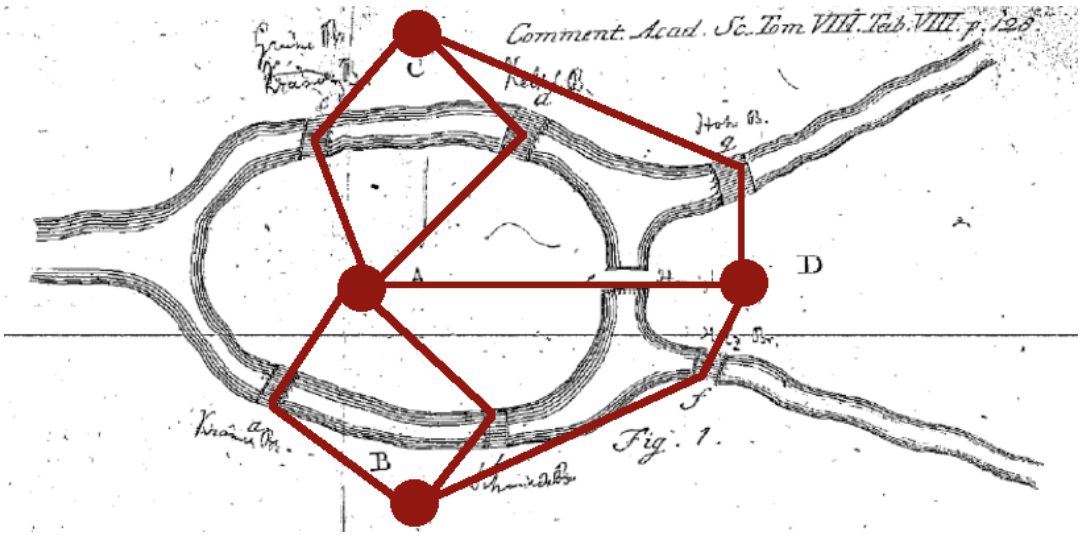


En un graf amb pesos (dirigit o no), les arestes tenen associats nombres reals

En un graf etiquetat alguns nodes i arestes tenen noms (donats per una aplicació en un conjunt d'etiquetes)

Grafs

Els grafs són una eina valuosa en ciències experimentals per ordenar i classificar informació (sobre connexions, relacions, etc.) en un nivell intermedi de detall



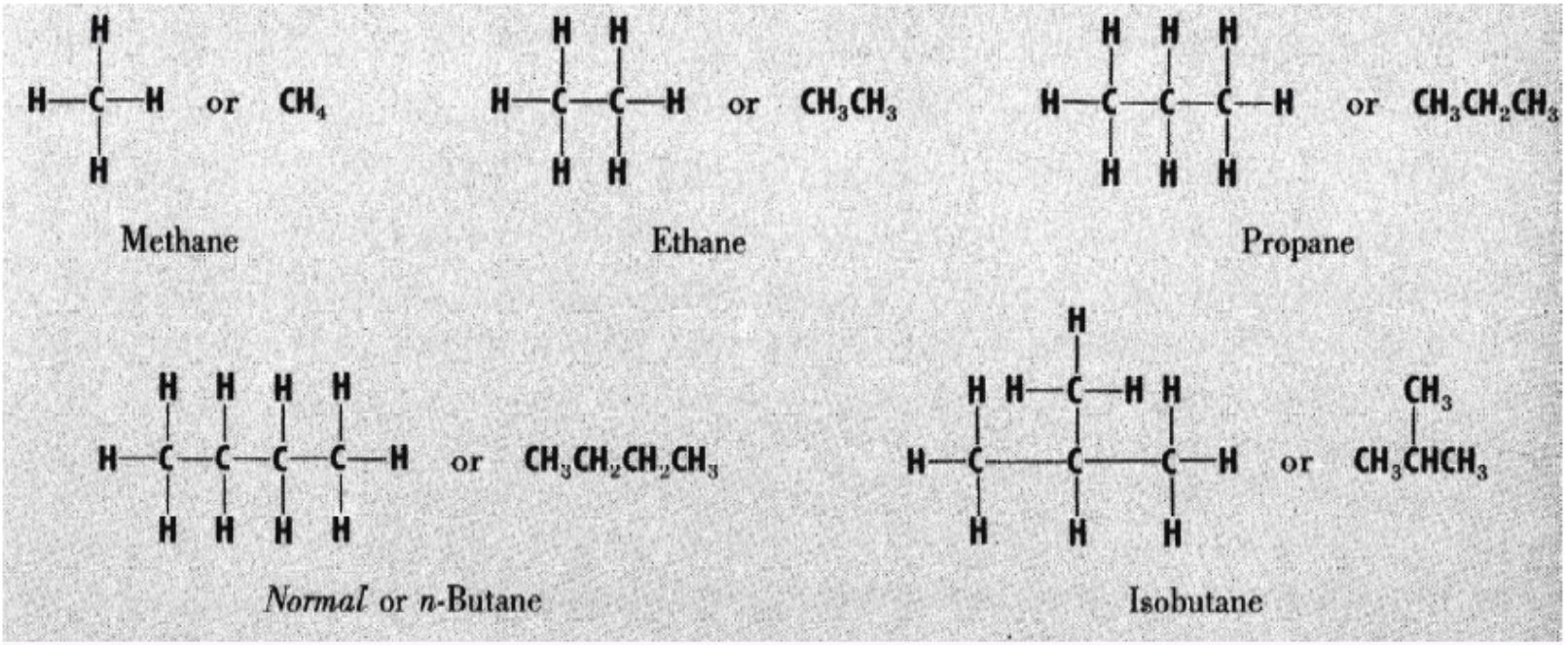
Grafs químics (Cayley, 1874)

A. Cayley introduí l'any 1874 els kenogrames, origen dels grafs químics:

- Nodes: àtoms
- Arestes: enllaços químics

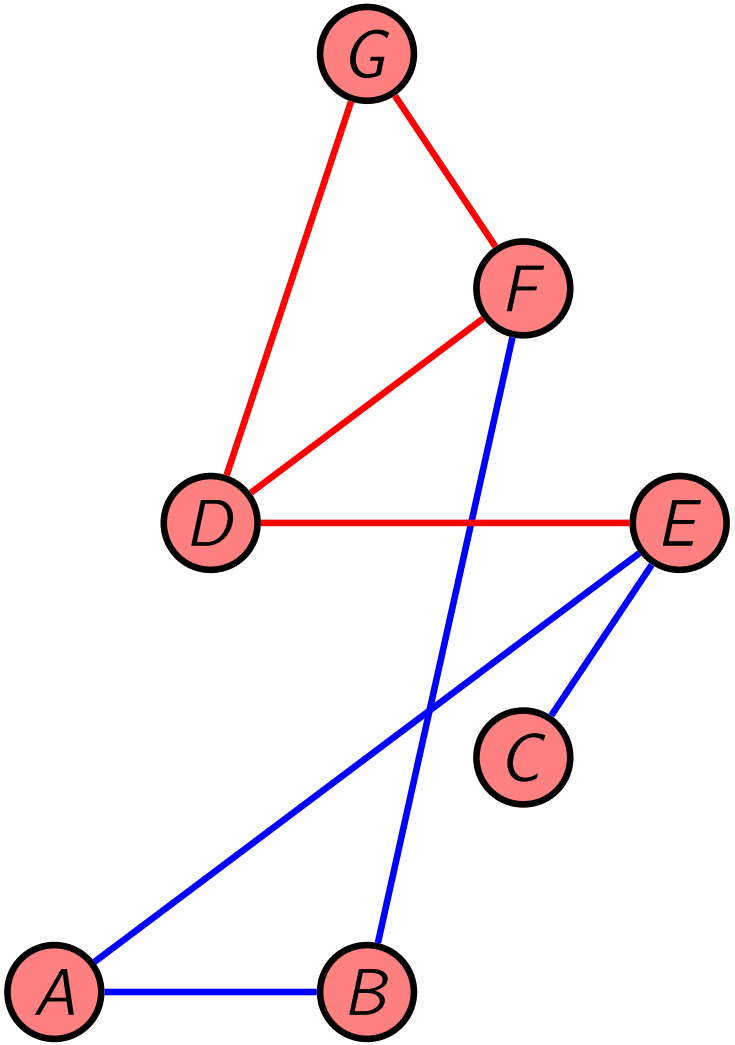
Li permeté enumerar els isòmeres (molècules diferents amb la mateixa composició química) d'hidrocarburs saturats del tipus C_nH_{2n+2}, per a $n = 1, \ldots, 13$

Grafs químics (Cayley, 1874)



C_nH_{2n+2}, $n = 1, 2, 3, 4$, segons un llibre de química orgànica

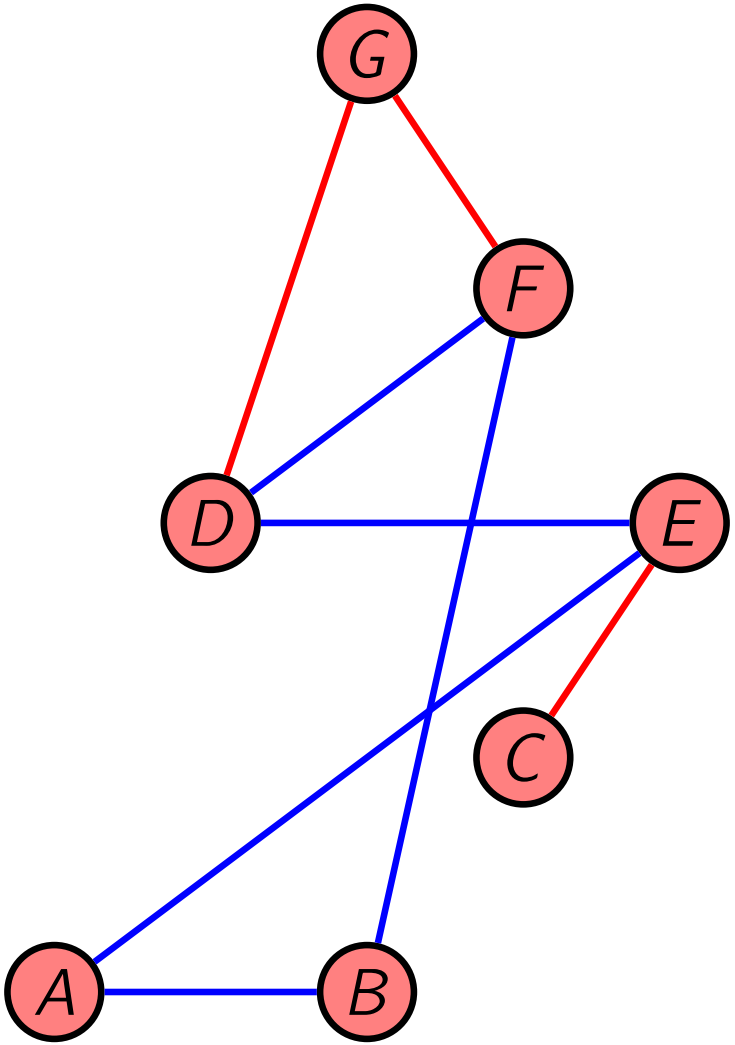
Arbres



C, E, A, B, F és un camí

Un camí és una seqüència de nodes $(v_0, v_1, \ldots, v_k)$ tal que cada $\{v_i, v_{i+1}\}$ és una aresta del graf

Arbres

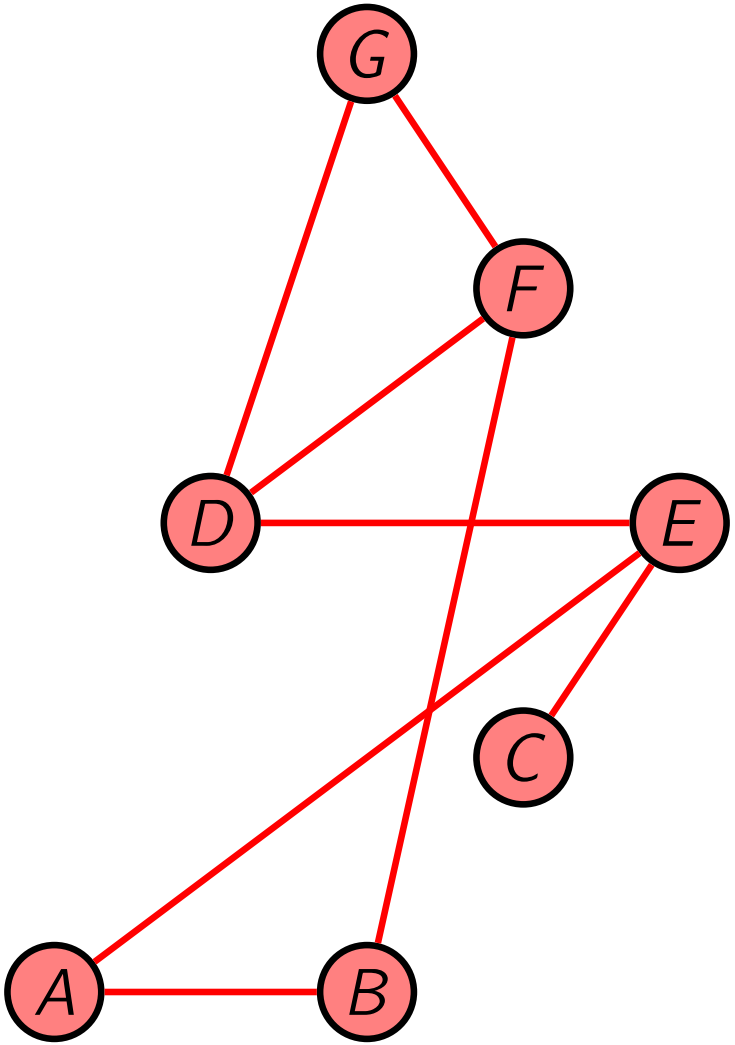


Un camí és una seqüència de nodes $(v_0, v_1, \ldots, v_k)$ tal que cada $\{v_i, v_{i+1}\}$ és una aresta del graf

Un cicle és un camí amb $v_0 = v_k$

A, B, F, D, E, A és un cicle

Arbres



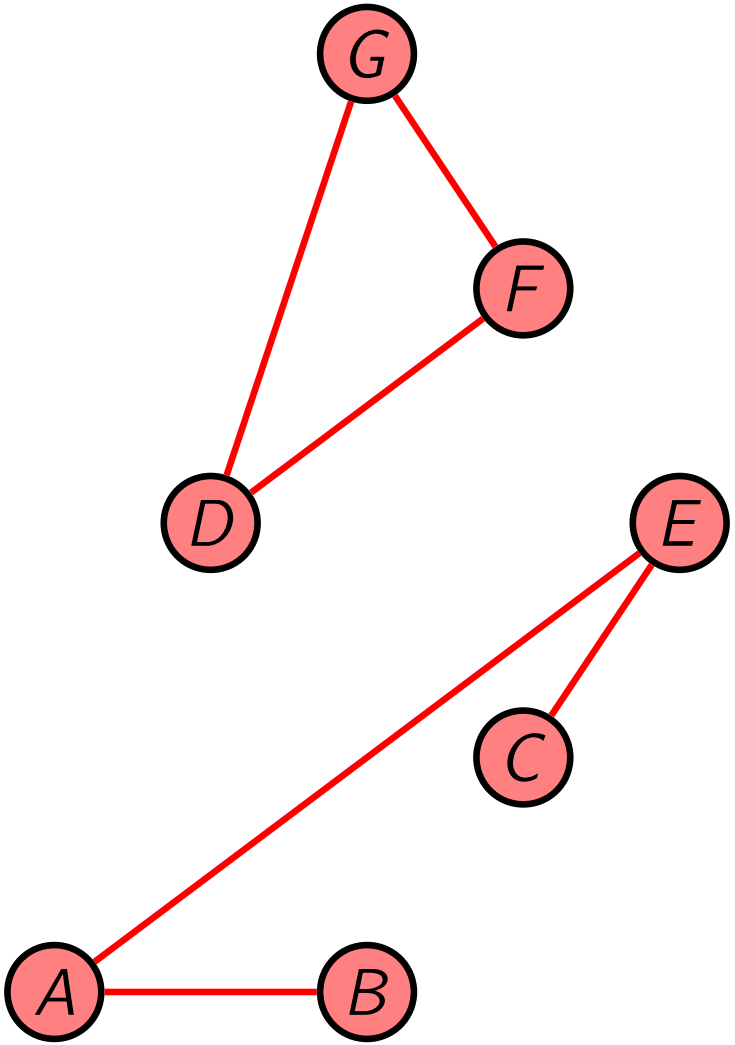
Aquest graf es connex

Un camí és una seqüència de nodes $(v_0, v_1, \ldots, v_k)$ tal que cada $\{v_i, v_{i+1}\}$ és una aresta del graf

Un cicle és un camí amb $v_0 = v_k$

Un graf és connex quan per a cada parella de nodes hi ha un camí que va de l'un a l'altre

Arbres



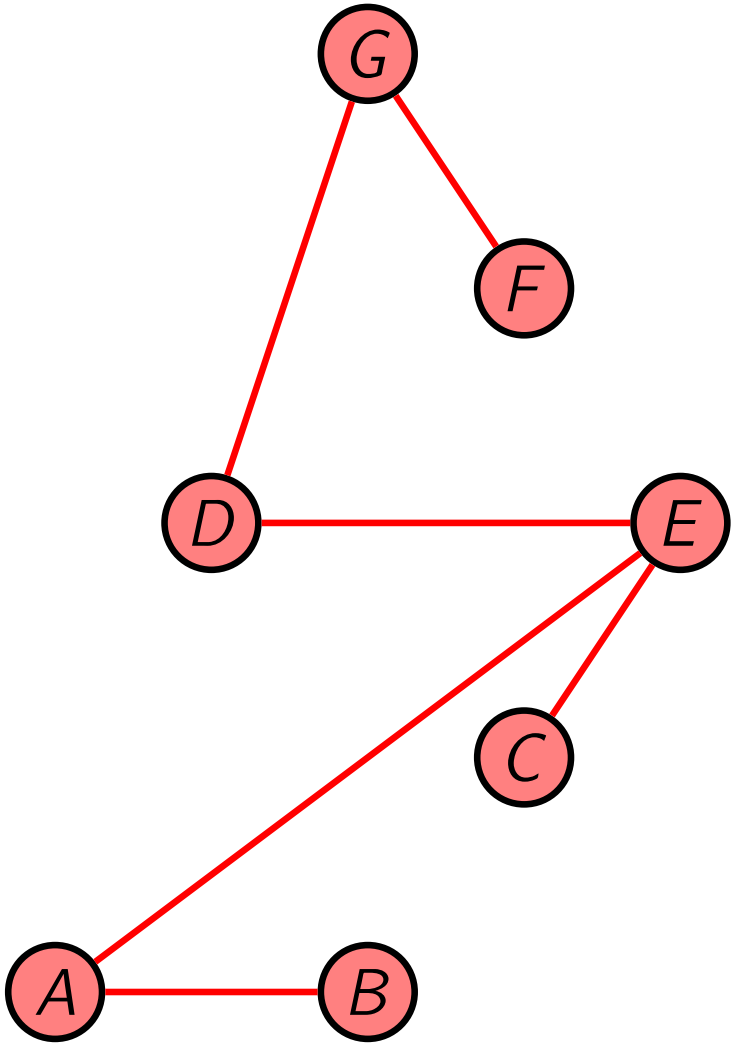
Aquest graf no és connex

Un camí és una seqüència de nodes $(v_0, v_1, \ldots, v_k)$ tal que cada $\{v_i, v_{i+1}\}$ és una aresta del graf

Un cicle és un camí amb $v_0 = v_k$

Un graf és connex quan per a cada parella de nodes hi ha un camí que va de l'un a l'altre

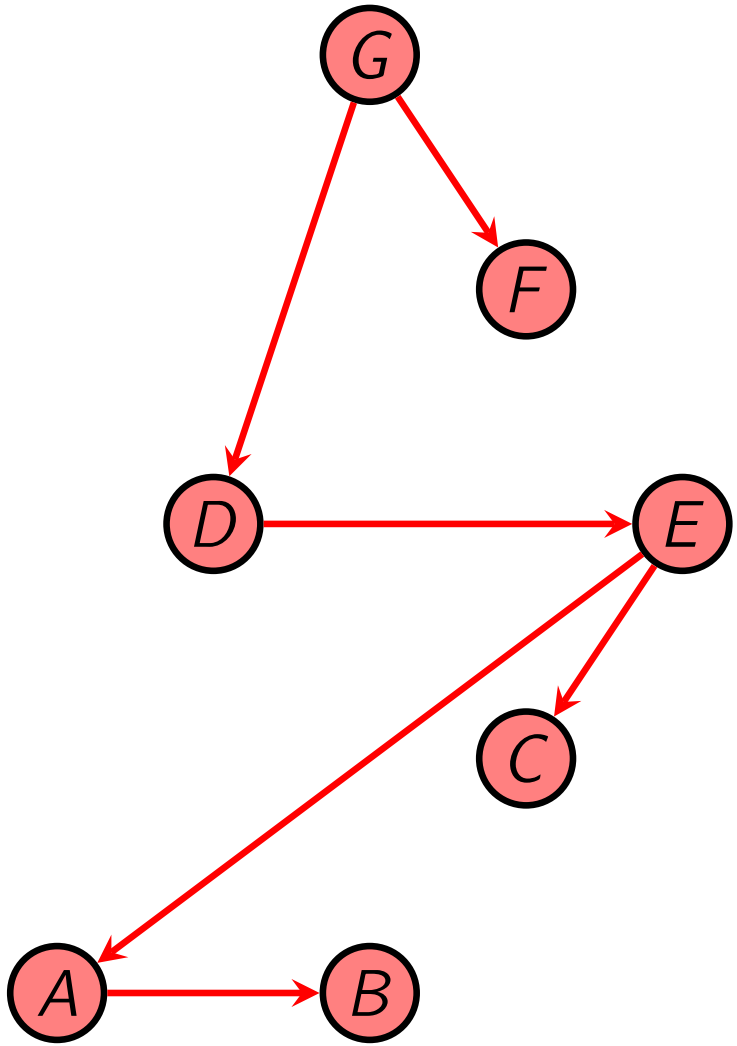
Arbres



Un arbre és graf connex i sense cicles

Els nodes de grau 1 són les fulles, i els altres, nodes interiors

Arbres



Un arbre arrelat és un arbre amb un node distingit, l'arrel

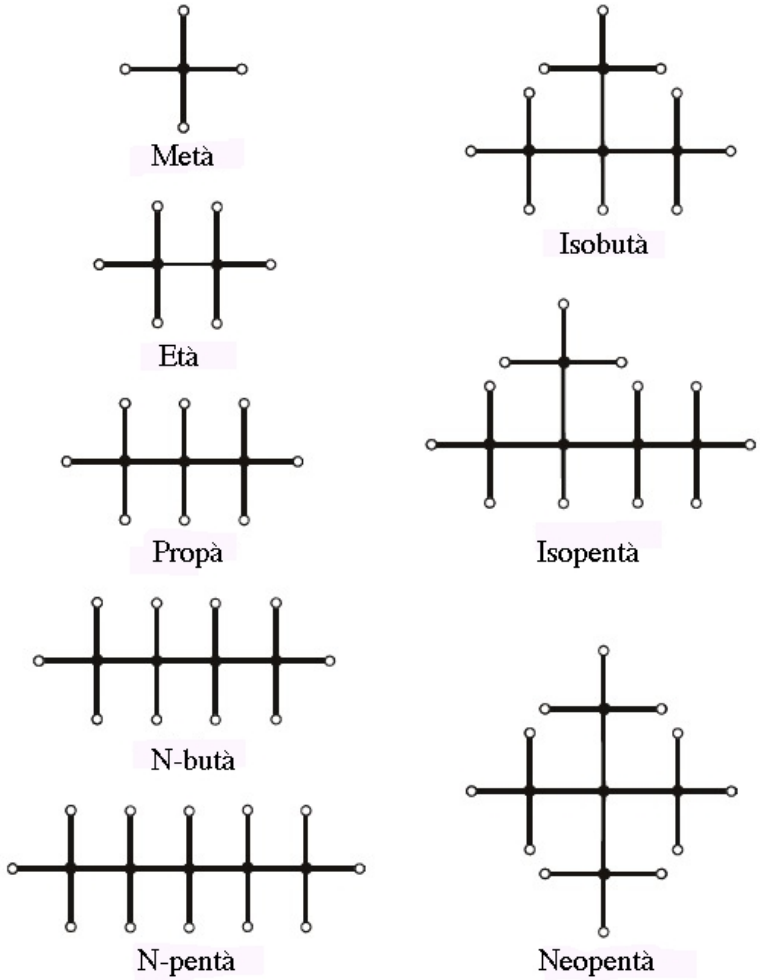
El consideram dirigit, amb direcció dels arcs allunyant-se de l'arrel

L'arrel té grau d'entrada 0, tots els altres nodes tenen grau d'entrada 1

G és l'arrel

Cayley i els arbres

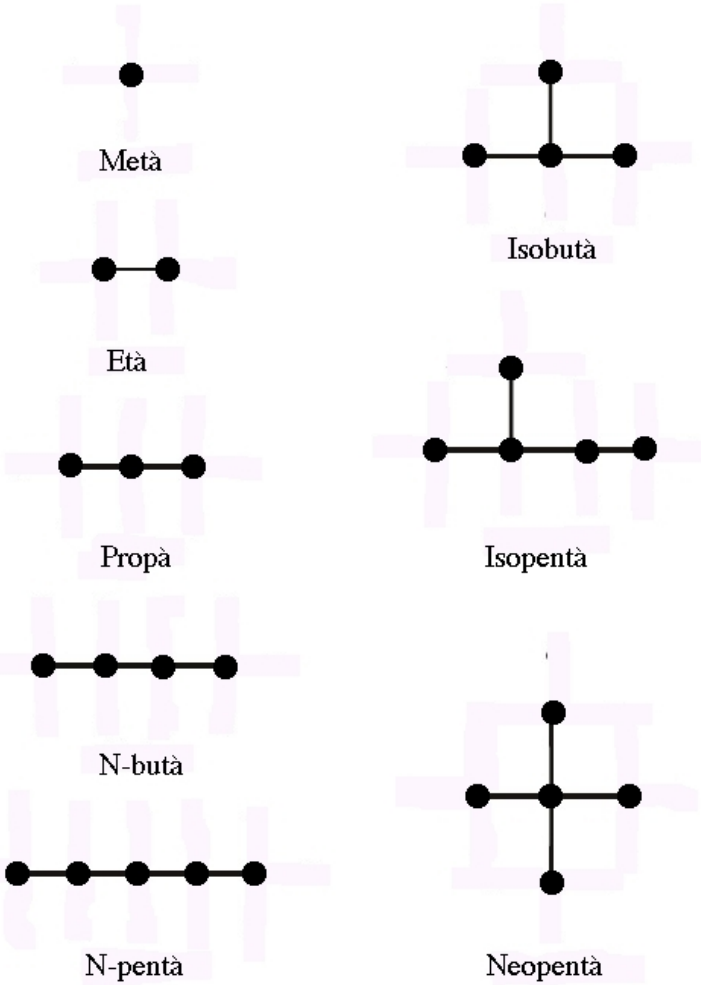
Els kenogrames de Cayley són arbres amb nodes interiors de grau 4 i fulles H



Cayley i els arbres

Els kenogrames de Cayley són arbres amb nodes interiors de grau 4 i fulles H

Equivalents a arbres amb tots els nodes de grau $\leqslant 4$ (només les C)



Grafs en biologia

Algunes aplicacions:

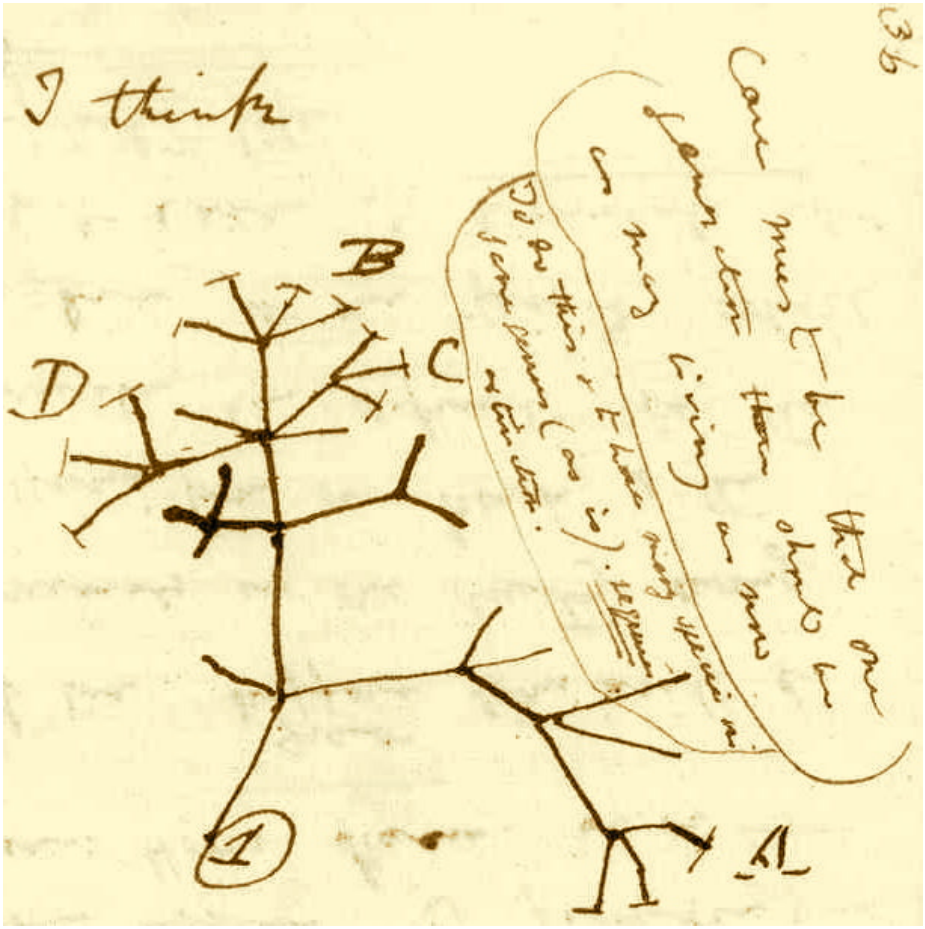
- Filogenètica
- Estructures tridimensionals de molècules
- Xarxes de reaccions i interaccions biomoleculars

Arbres i xarxes filogenètics

Des de temps de Darwin (1837), els biòlegs empren grafs per representar la descendència evolutiva

Nodes: Espècies

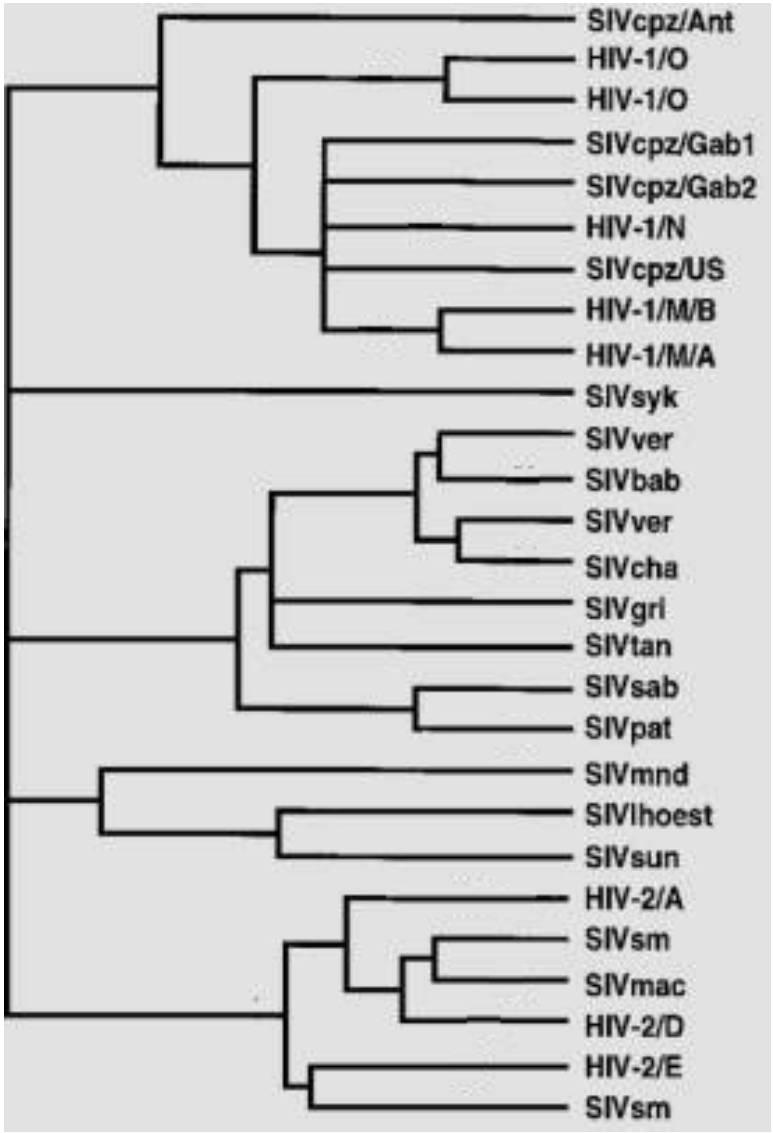
Arcs: Descendència directa



Arbres i xarxes filogenètics

Si només es tenen en compte mutacions, cada espècie té només un ascendent directe i el resultat és un arbre filogenètic:

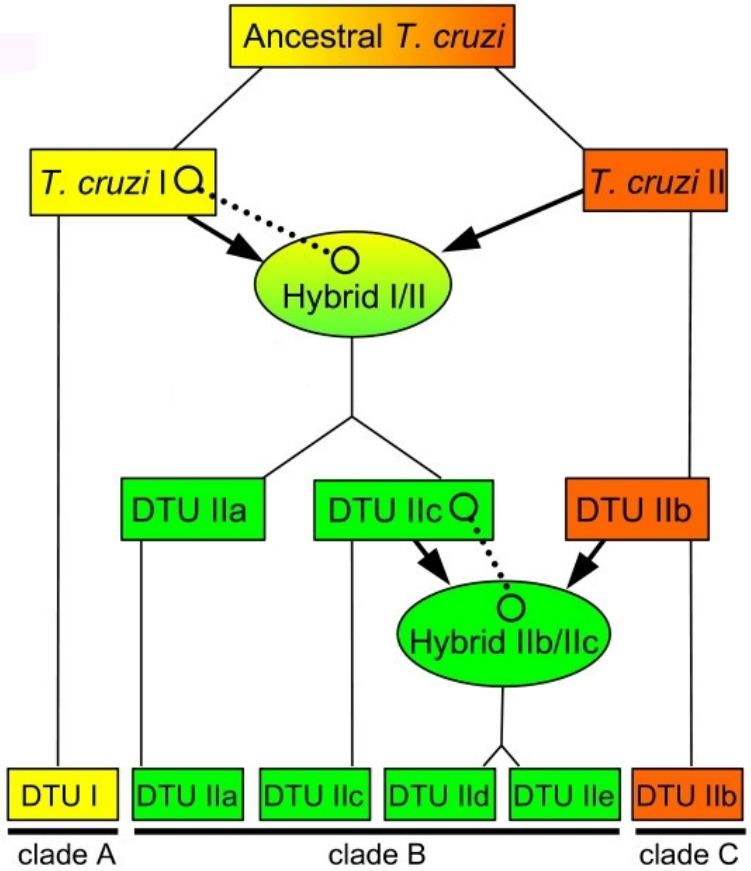
Arbre arrelat amb les fulles etiquetades de manera injectiva i sense nodes interiors de grau de sortida 1



Arbres i xarxes filogenètics

Si es permeten hibriditzacions, les espècies poden tenir més d'un ascendent directe, i el resultat és una xarxa filogenètica:

Graf dirigit acíclic arrelat (amb un únic node amb grau d'entrada 0, l'arrel) amb les fulles etiquetades de manera injectiva



Estructures tridimensionals de RNA

L'àcid ribonucleic (RNA) és un àcid nucleic que juga molts papers en processos cel·lulars

És un polímer, format per una cadena de nucleòtids, amb bases enganxades. N'hi ha 4: Adenina, Guanina, Timina i Uracil

Per tant, una molècula d'ARN es pot pensar com una paraula sobre A,C,G,U

Estructures tridimensionals de RNA

Les molècules d'ARN es pleguen dins les cèl·lules en formes bastant complicades, relacionades amb la funció

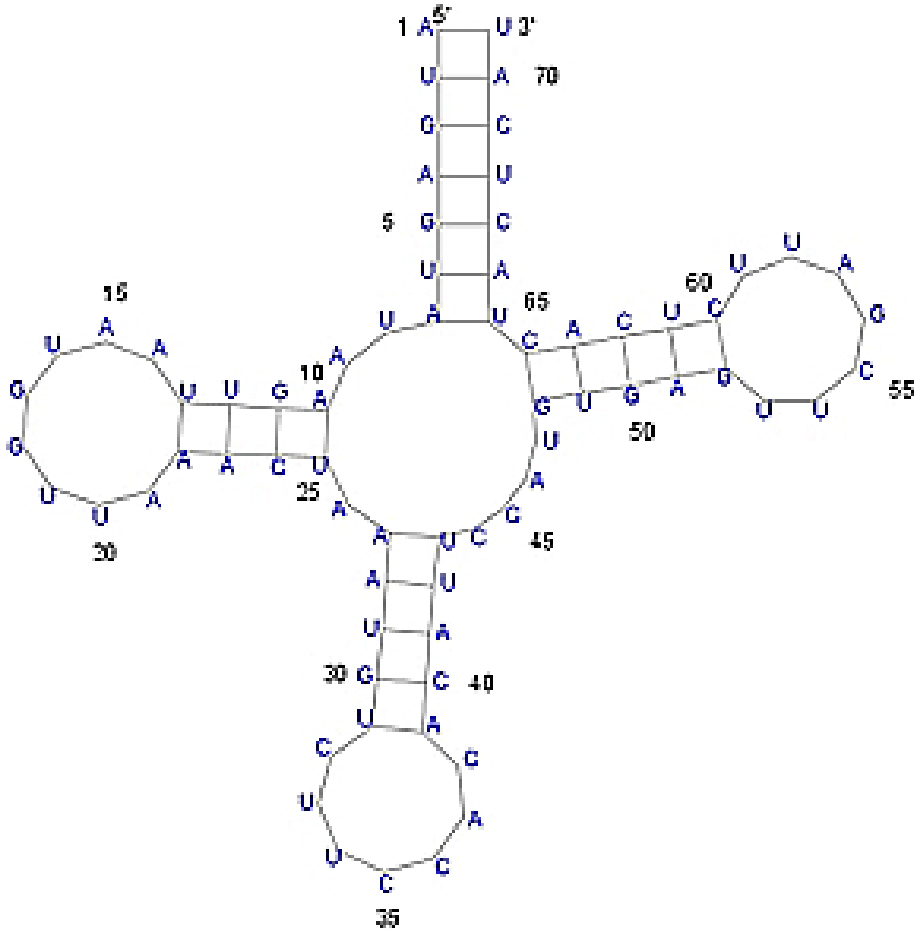
L'estructura s'aguanta per enllaços d'hidrogen entre bases.

Sovint és suficient conèixer el graf de contactes:

Nodes: Bases
Arestes: Seqüència dins la molècula (esquelet) i enllaços d'hidrogen (contactes)

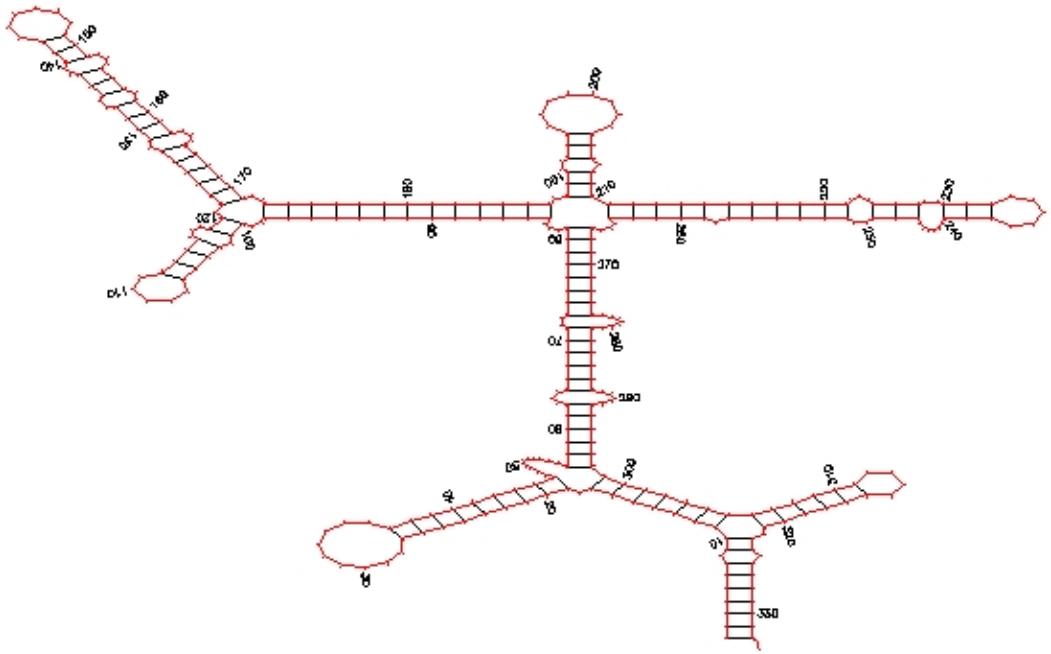
Estructures tridimensionals de RNA

Estructures secundàries: cada base té com a molt un contacte, i els contactes no es creuen



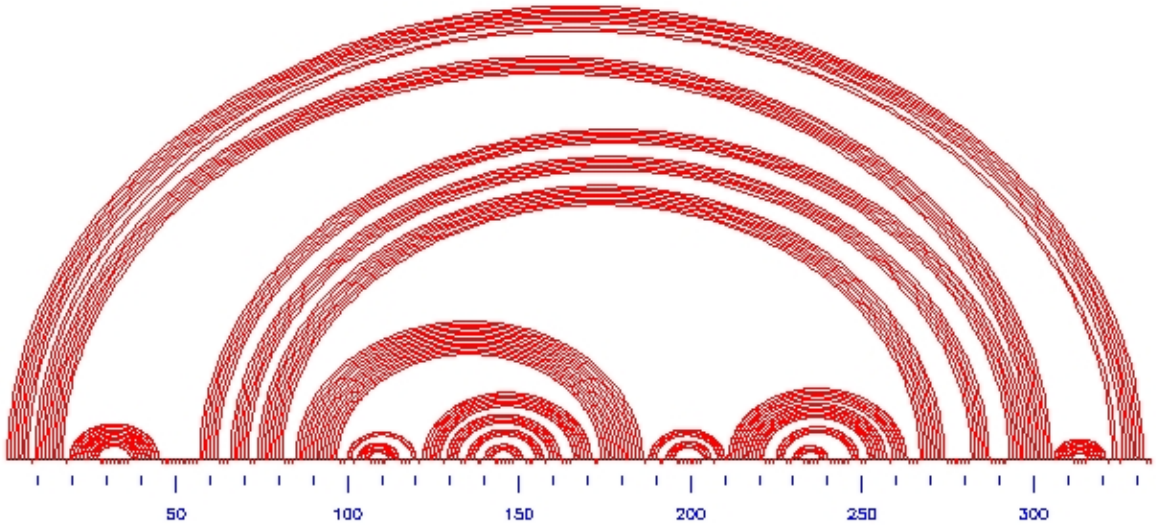
Estructures tridimensionals de RNA

Les estructures secundàries es representen com a grafs de cúpules:



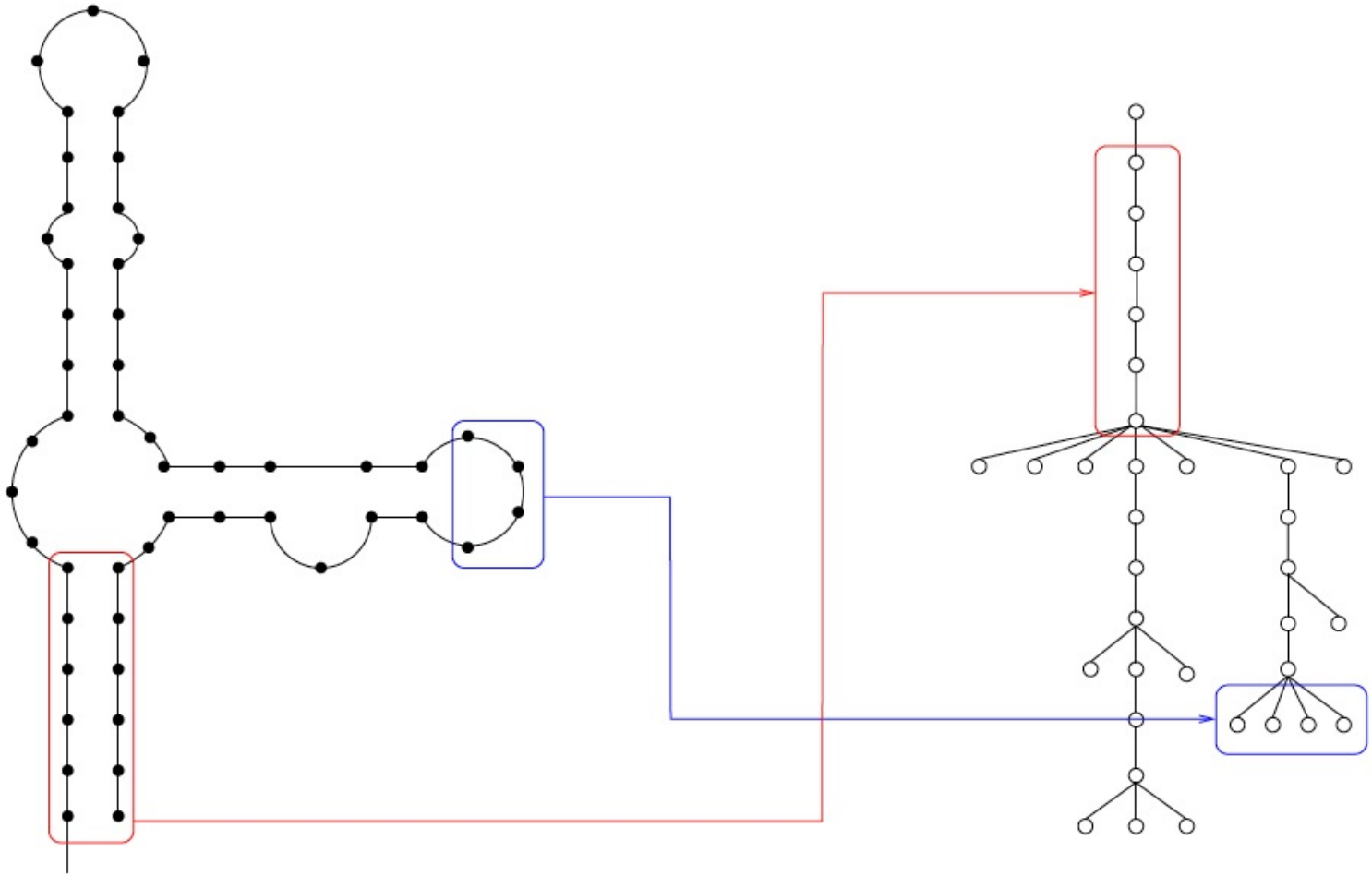
Estructures tridimensionals de RNA

Les estructures secundàries es representen com a grafs de cúpules:



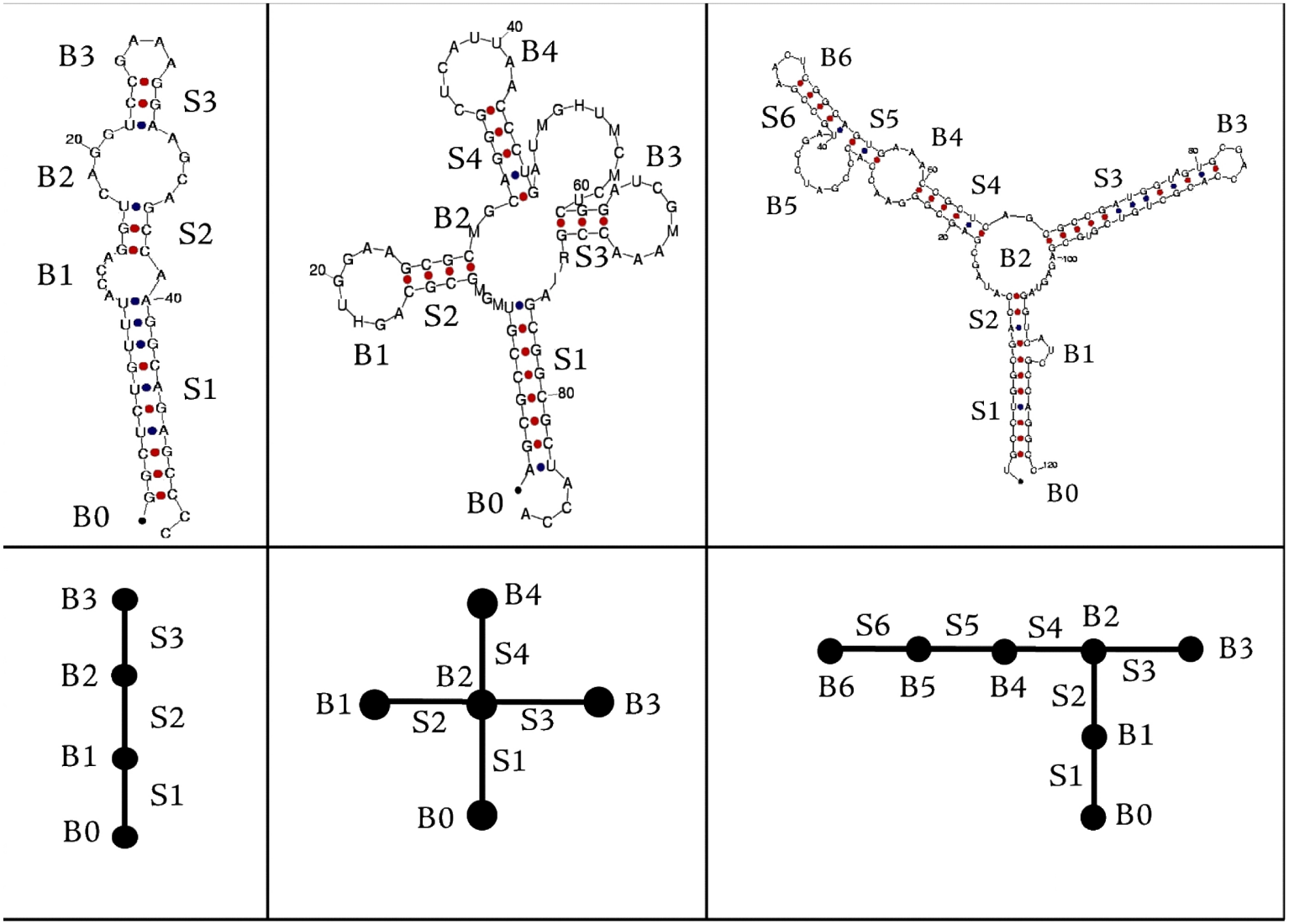
Estructures tridimensionals de RNA

Les estructures secundàries es poden representar de manera injectiva com a arbres lineals (amb fills ordenats d'esquerra a dreta)



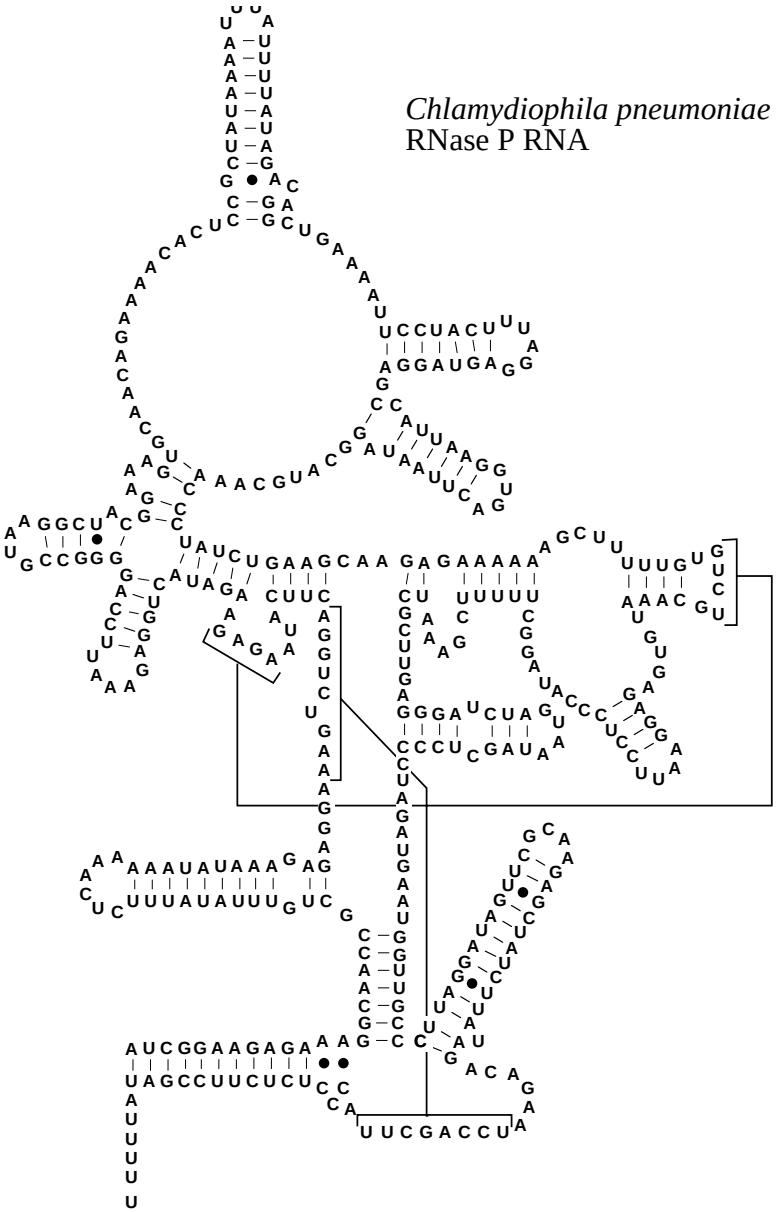
Estructures tridimensionals de RNA

Hi ha altres representacions més grolleres de les estructures secundàries com a arbres



Estructures tridimensionals de RNA

Estructures secundàries amb pseudo-nusos: cada base té com a molt un contacte, però els contactes es poden creuar



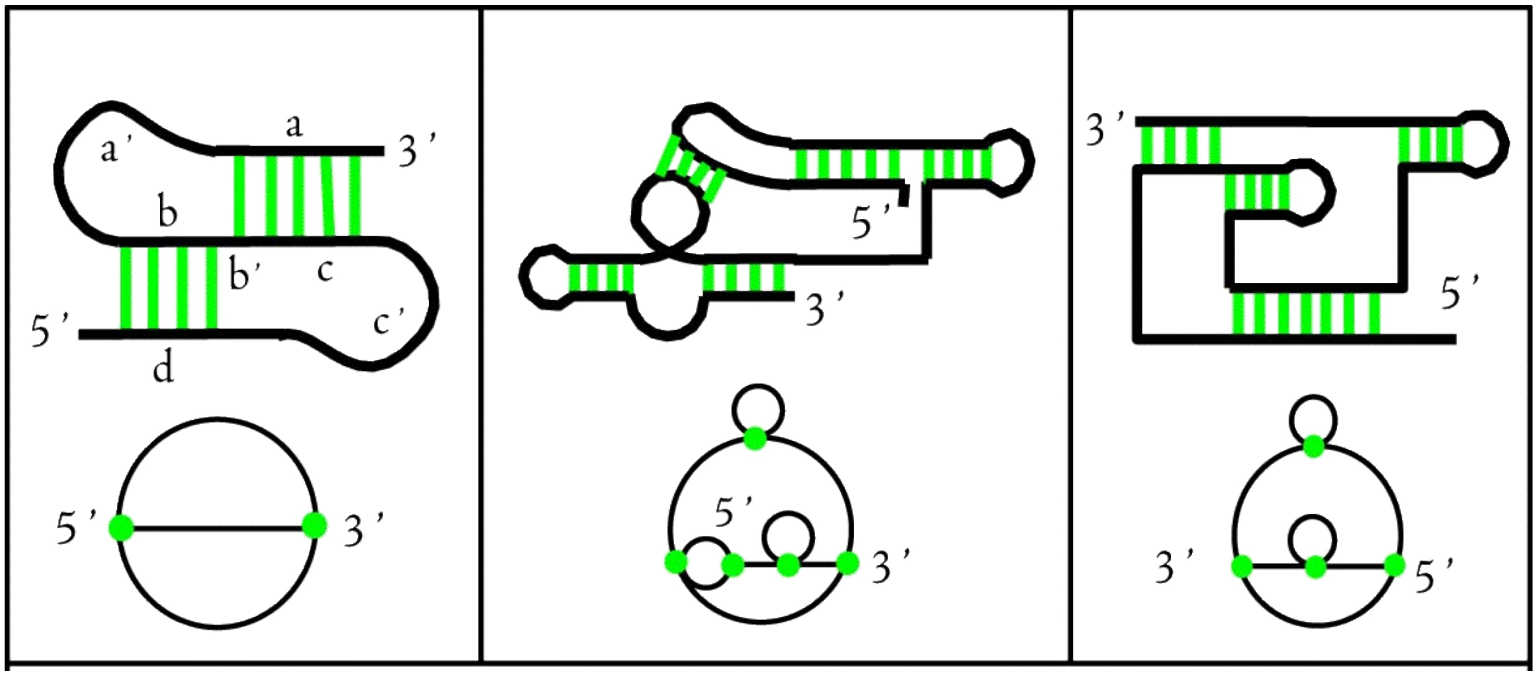
Chlamydiophila pneumoniae
RNase P RNA

Estructures tridimensionals de RNA

Les estructures amb pseudo-nusos no es poden representar com a arbres, sinó:

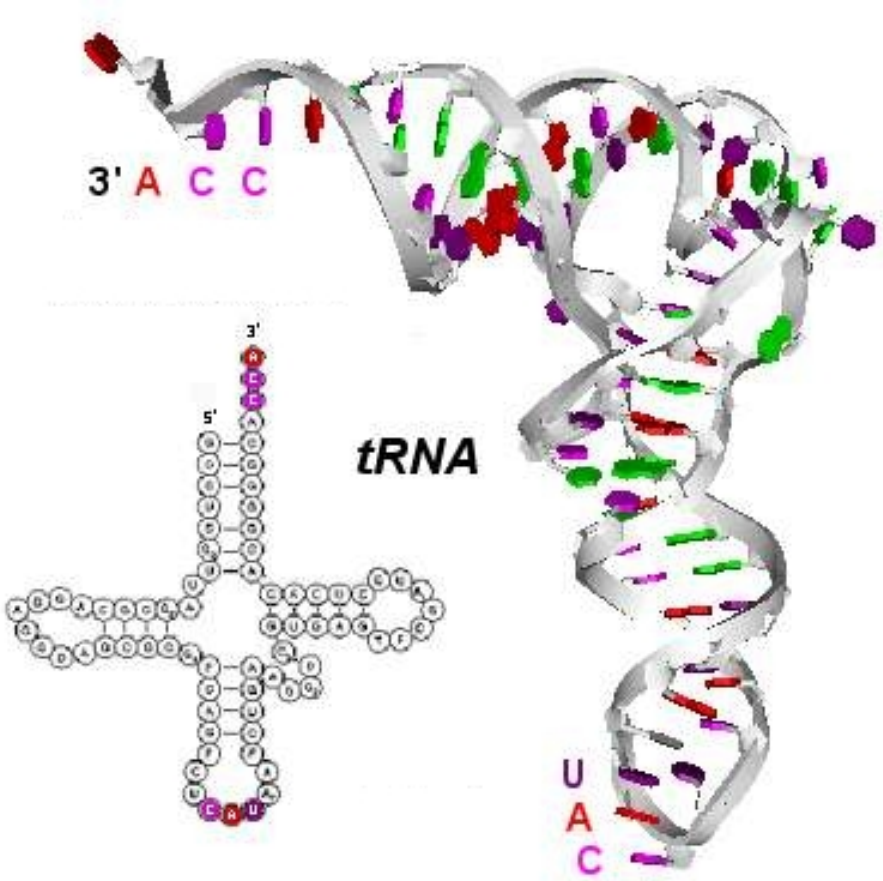
Nodes: tijes (seqüències de contactes consecutius)

Arcs: recorregut de l'esquelet



Estructures tridimensionals de RNA

L'estructura detallada és molt més complicada



Estructures tridimensionals de proteïnes

Les proteïnes són part essencial dels organismes i participen en tots els processos cel·lulars

Són polímers, formats per cadenes d'aminoàcids. Hi ha 20 aminoàcids

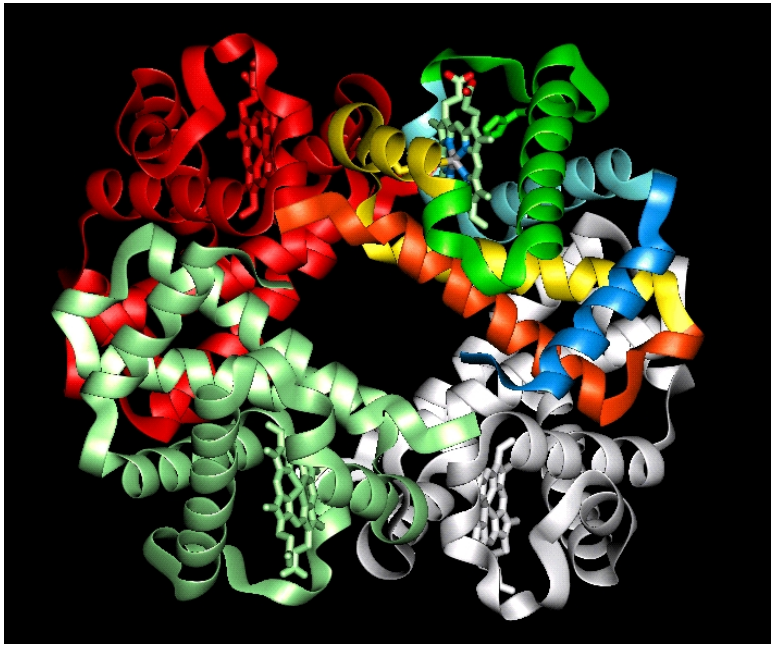
Per tant, una proteïna es pot pensar com una paraula sobre un alfabet de 20 lletres

Estructures tridimensionals de proteïnes

Les proteïnes es cargolen dins les cèl·lules formant estructures 3D molt complicades.

En general, aquesta estructura és determinada per la seqüència d'aminoàcids

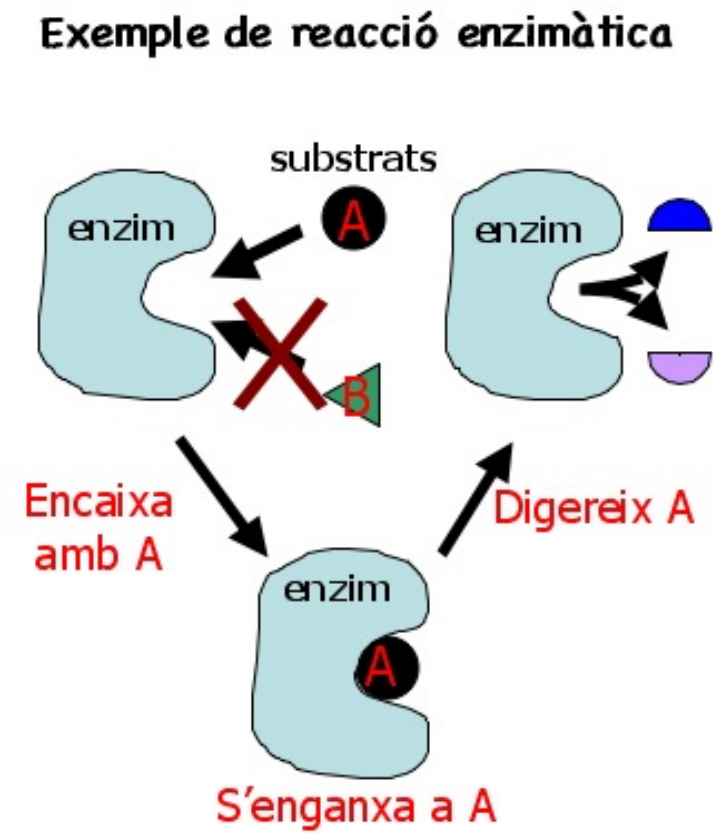
L'estructura determina la funció

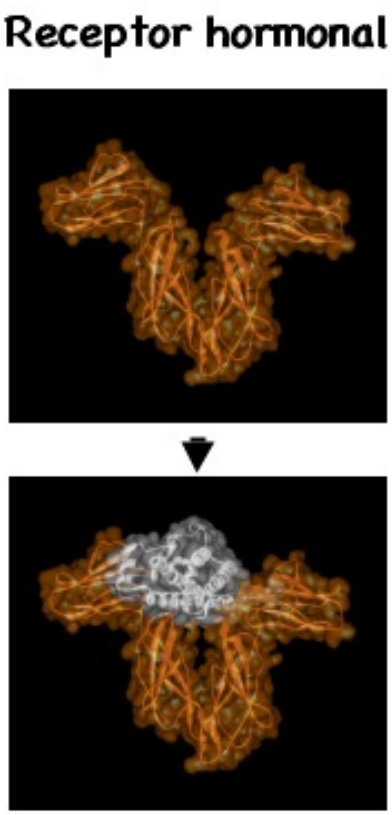


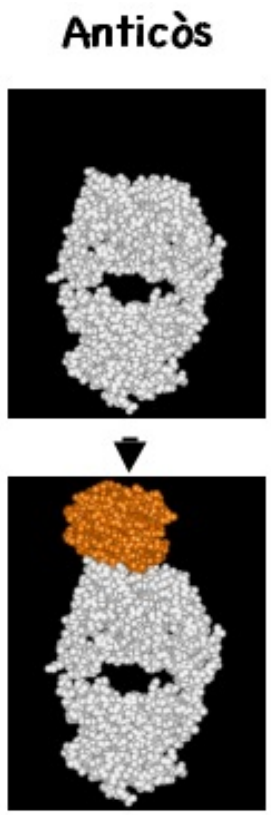
Hemoglobina

Estructures tridimensionals de proteïnes

L'estructura 3D determina la funció de la proteïna







Estructures tridimensionals de proteïnes

Jerarquia de l'estructura de les proteïnes:

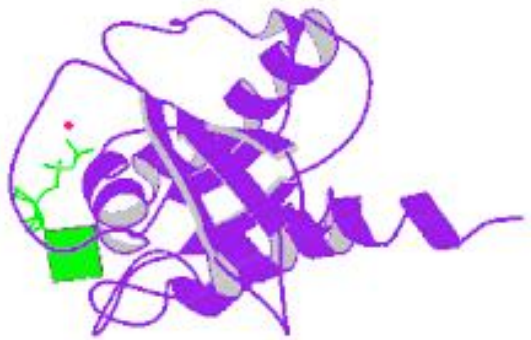
- Estructura primària: la seqüència d'aminoàcids
- Estructura secundària: estructures regulars comuns (α-hèlices, β-fulles,...)
- Estructura terciària: estructura 3D formada per les estructures secundàries d'una proteïna recargolades
- Estructura quaternària: estructura formada per més d'una proteïna recargolades

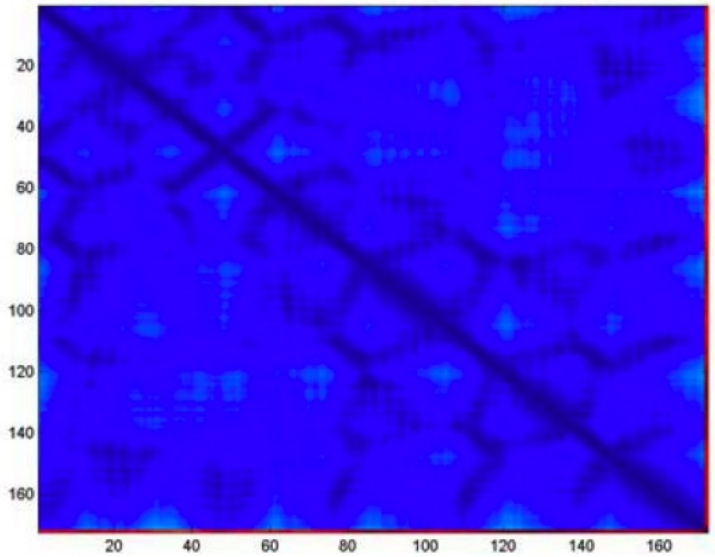
Estructures tridimensionals de proteïnes

L'estructura 3D es representa per mitjà d'un graf complet:

Nodes: Aminoàcids

Arestes: Totes les possibles, amb pesos les distàncies entre aminoàcids (matriu de distàncies)





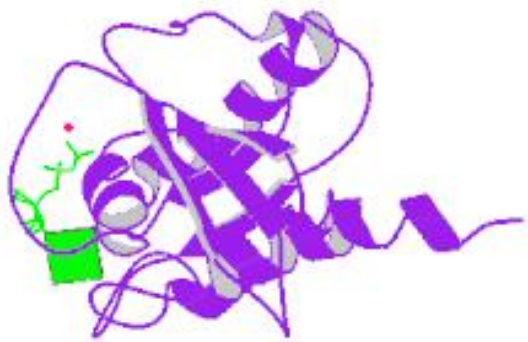
Proteïna 1aa9 i la seva matriu de distàncies

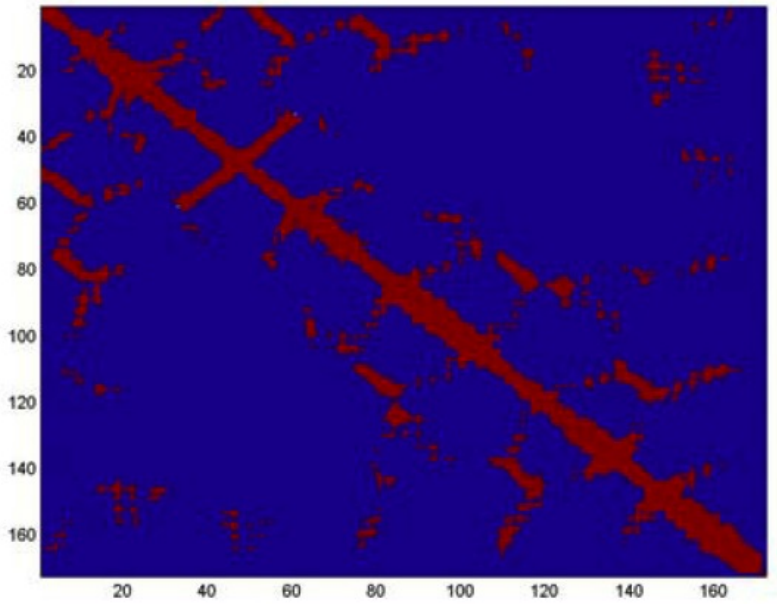
Estructures tridimensionals de proteïnes

L'estructura 3D es representa per mitjà d'un graf de contactes:

Nodes: Aminoàcids

Arestes: Els dos aminoàcids estan a distància $\leqslant \varepsilon$ fixat (9 Å)





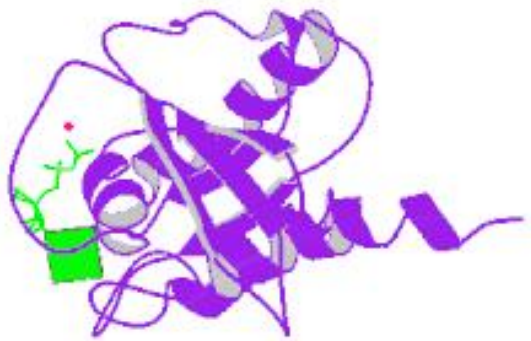
Proteïna 1aa9 i el seu graf de contactes (matriu d'adjacència)

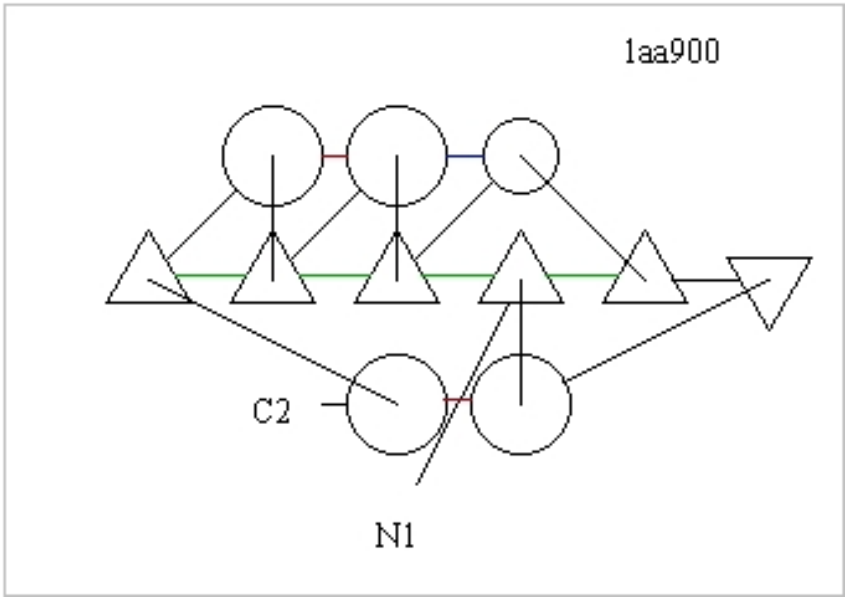
Estructures tridimensionals de proteïnes

L'estructura 3D es representa per mitjà d'un graf TOPS:

Nodes: Les estructures secundàries

Arestes: Relacions entre estructures secundàries

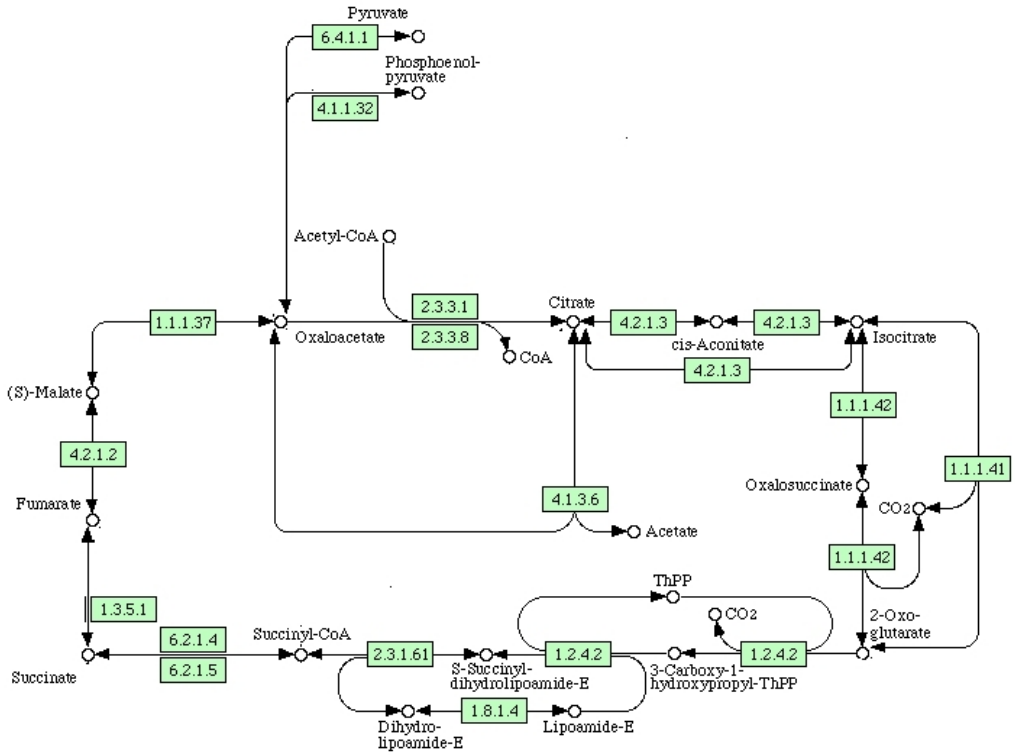




Proteïna 1aa9 i el seu graf TOPS

Rutes metabòliques

Conjunt de processos i reaccions bioquímiques catalitzades per enzims que produeixen compostos (metabolits) que són emprats o emmagatzemats per la cèl·lula

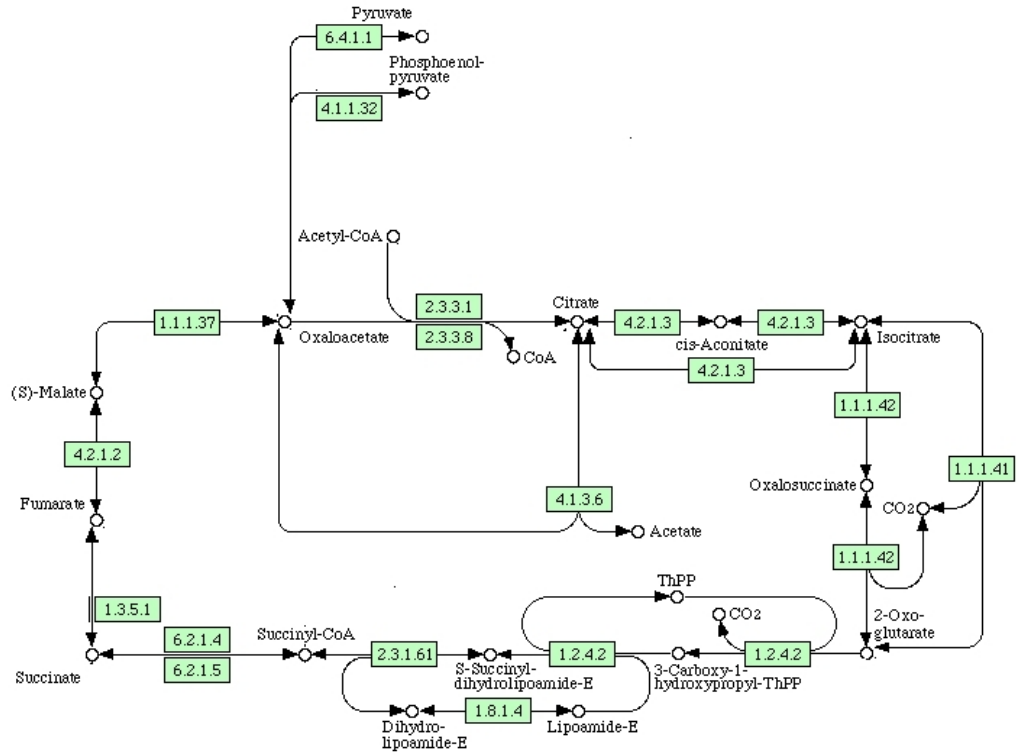


Rutes metabòliques

Les representam com a grafs bipartits:

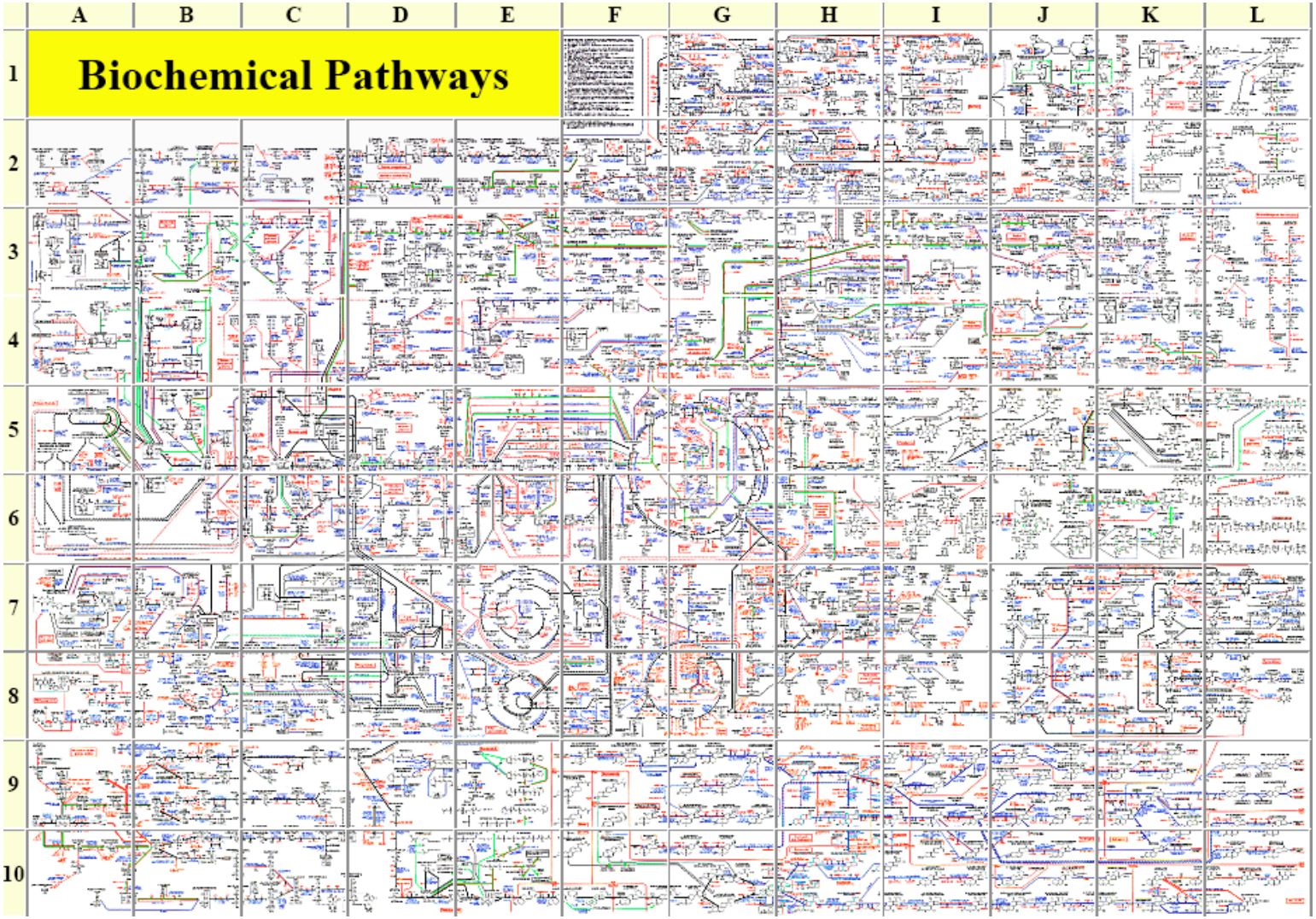
Nodes: Compostos i enzims

Arcs: Ser substrat o producte de la reacció catalitzada per l'enzim



Rutes metabòliques

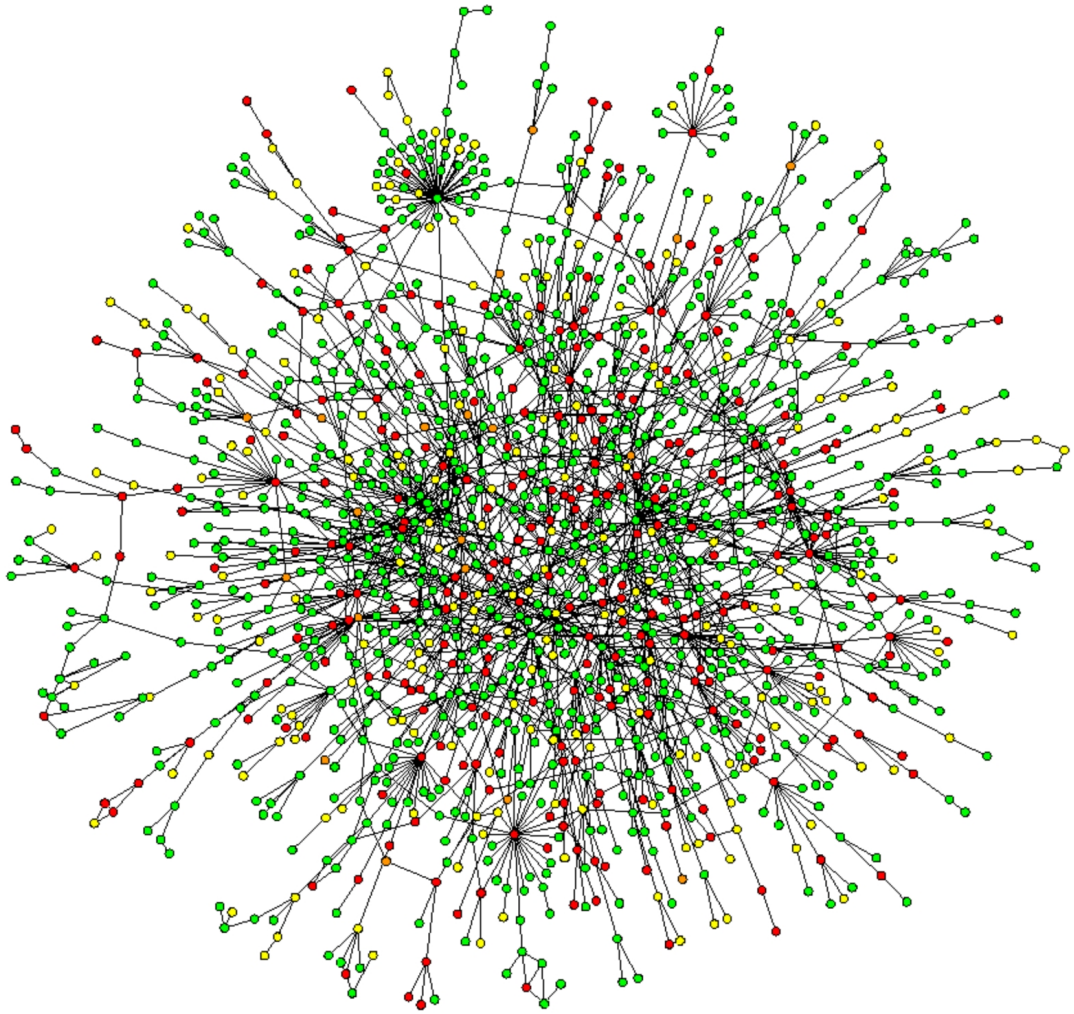
Les rutes metabòliques es combinen: xarxa metabòlica



http://www.expasy.ch/cgi-bin/show_thumbnails.pl

Xarxes de proteïnes

Conjunt d'interaccions entre les proteïnes d'una cèl·lula



Xarxa de proteïnes del llevat

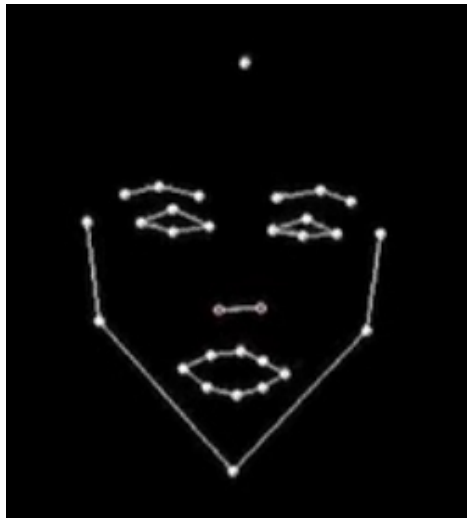
Comparació de grafs en biologia

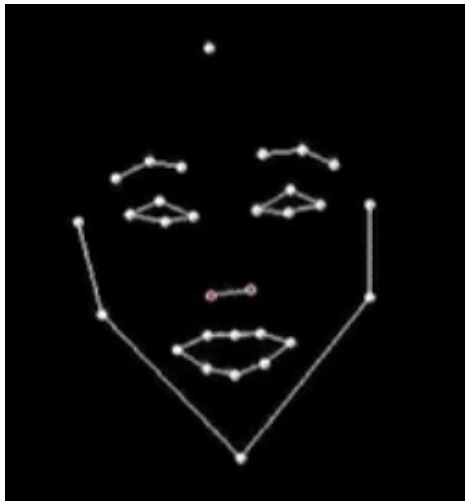
Un problema important en biologia computacional és definir mètodes per comparar aquests grafs:

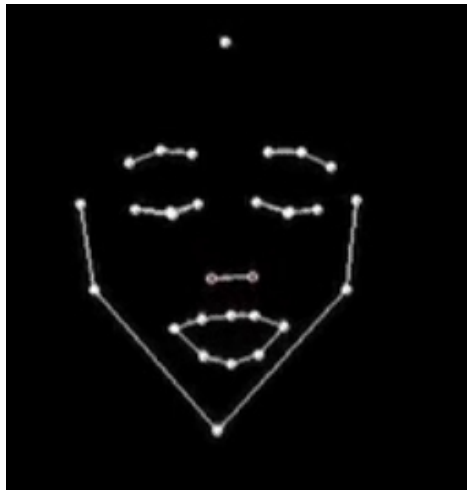
- Per poder avaluar algoritmes de predicció i reconstrucció
- Per poder predir la funció d'una proteïna (per comparació amb una proteïna coneguda)
- Per trobar estructures conservades en la mateixa xarxa biomolecular per a organismes diferents
- Per calcular filogènies
- Per xafarderia

Comparació de grafs

La comparació de grafs és un problema en moltes altres disciplines







Comparació de grafs en biologia

El mètode de comparació ha de ser bo d'entendre, ràpid (i fàcil) de calcular, i ha de tenir significat biològic

Cada ús de grafs requereix de mètodes de comparació específics

Quatre maneres bàsiques de comparar grafs:

- Definir una distància (mètrica)
- Donar una correspondència
- Donar subgraf màxim en comú
- Donar supergraf minimal en comú

Distàncies

Objectiu: definir una distància sobre un espai de grafs d'aquests de tal manera que més llunyà signifiqui més diferent

Serveix per avaluar numèricament la similitud dels grafs

No només es tracta de $d(G, G') = 0 \Leftrightarrow G = G'$

La desigualtat triangular serveix per accelerar algoritmes de cerca basats en una distància

Les distàncies permeten calcular agrupaments (filogènies)

Exemple: distàncies per a arbres filogenètics

S'han proposat més d'una dotzena de distàncies per a arbres filogenètics, cadascuna amb les seves coses bones, i encara cap de definitiva

Dues de les més populars (senzilles, es calculen ràpid):

- Robinson-Foulds
- Nodal

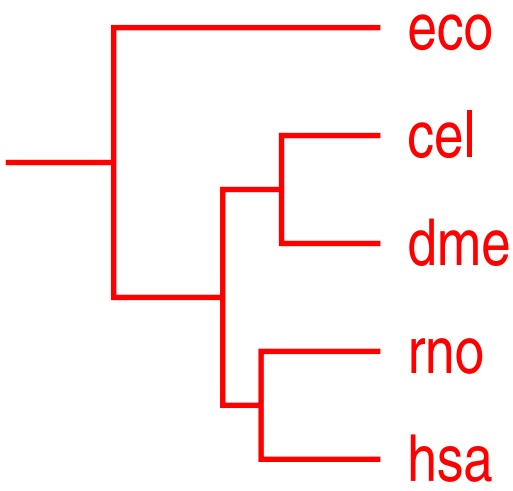
Distància de Robinson-Foulds

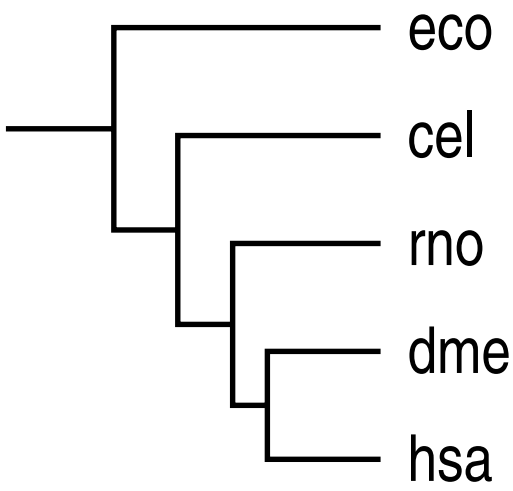
A cada node interior v d'un arbre T li assignam el conjunt $C_T(v)$ de les fulles descendents (hi ha un camí dirigit del node a la fulla)

$$C(T) = \{C_T(v) \mid v \text{ node interior}\}$$

$d_{RF} = |C(T)\Delta C(T')|$ és una distància

Distància de Robinson-Foulds





$$C(T) = \{\{h,r\},\{c,d\},\{c,r,d,h\},\{e,c,d,r,h\}\}$$

$$C(T') = \{\{d,h\},\{r,d,h\},\{c,d,r,h\},\{e,c,d,r,h\}\}$$

$$C(T)\Delta C(T') = \{\{h,r\},\{c,d\},\{d,h\},\{r,d,h\}\}$$

$$d_{RF}(T.T') = 4$$

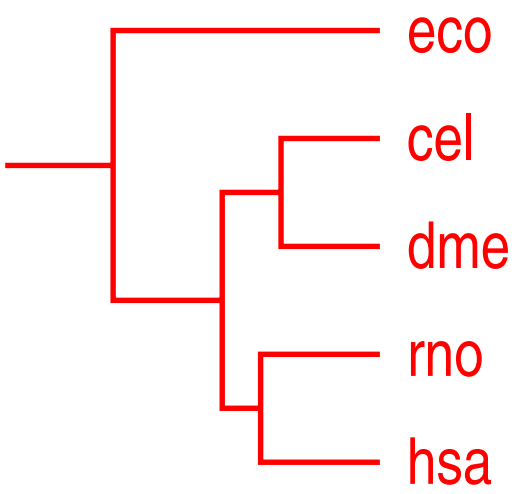
Distància nodal

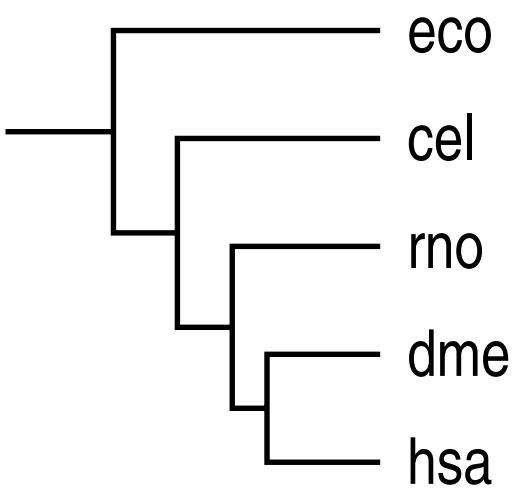
A cada parella de fulles (i, j) amb $i < j$ li assignam el nombre de nodes intermedis $d_{i,j}$ del camí (no dirigit) més curt d'una a l'altra

$$D(T) = (d_{1,2}, d_{1,3}, \ldots, d_{1,n}, d_{2,3}, \ldots)$$

$d_N(T, T') = \sum_{i,j} |d_{i,j} - d'_{i,j}|$ és una distància

Distància nodal





	cd	ce	ch	cr	de	dh	dr	eh	er	hr
T	1	3	3	3	3	3	3	3	3	1
T'	3	2	3	2	4	1	2	4	3	2

$$d_N(T, T') = 10$$

Exemple: distàncies per a RNA

Les distàncies per a arbres filogenètics no són adients per a estructures secundàries de RNA (són arbres, però de significat i estructura diferents)

S'han proposat més d'una vintena de distàncies per a estructures secundàries de RNA de la mateixa longitud, cadascuna amb les seves coses bones, i encara cap de definitiva

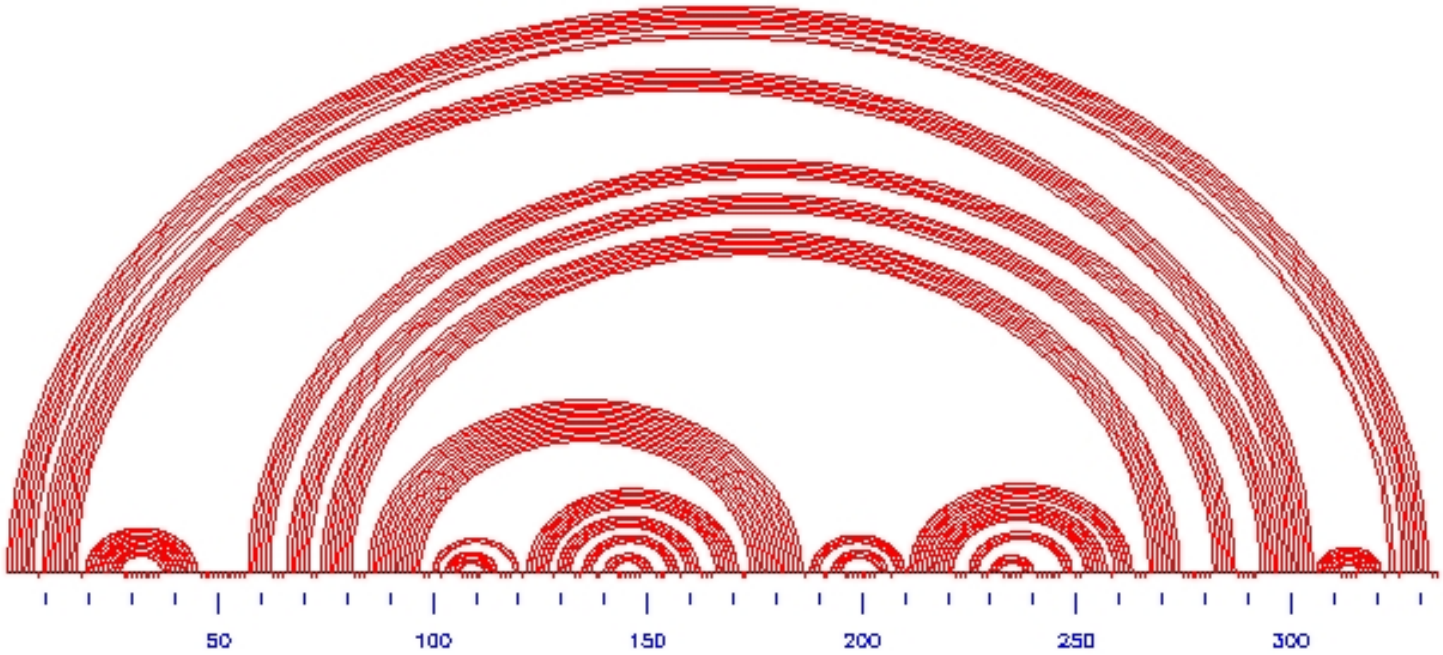
Dues de les més populars (senzilles, es calculen ràpid):

- Muntanyes
- Involucions

Cap distància acceptada per a estructures secundàries de longituds diferents

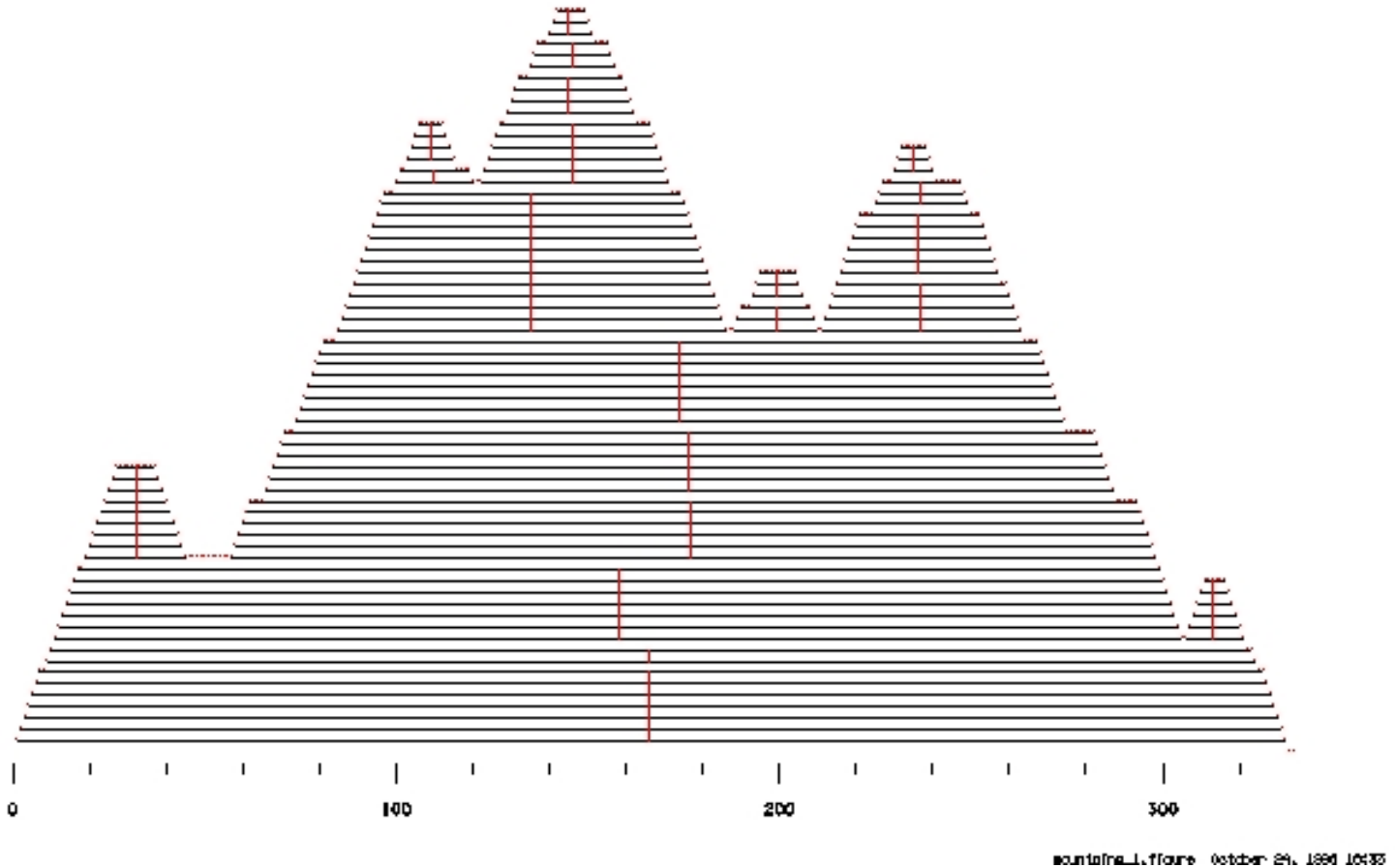
Distància de muntanyes

Donada una estructura secundària Γ de n bases, definim $f_\Gamma(k)$ com el nombre de contactes $\{i, j\}$ tals que $i < k < j$



Distància de muntanyes

Donada una estructura secundària Γ de n bases, definim $f_\Gamma(k)$ com el nombre de contactes $\{i, j\}$ tals que $i < k < j$



Distància de muntanyes

Donada una estructura secundària Γ de n bases, definim $f_\Gamma(k)$ com el nombre de contactes $\{i, j\}$ tals que $i < k < j$

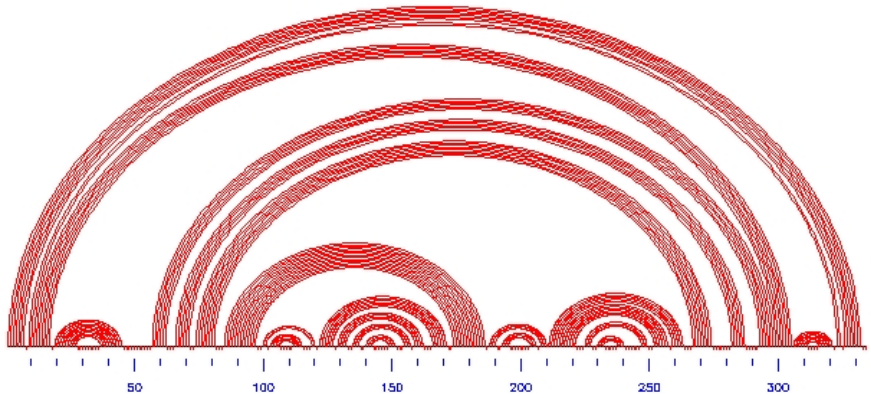
$$d_{mount}(\Gamma, \Gamma') = \sum_{k=1}^{n} |f_\Gamma(k) - f_{\Gamma'}(k)|$$

Distància d'involucions

Donada una estructura secundària Γ de bases $\{1, \ldots, n\}$, definim

$$\pi(\Gamma) = \prod_{\{i,j\} \text{ contacte}} (i,j) \qquad \text{(producte de transposicions)}$$

$d_{inv}(\Gamma, \Gamma')$= mínim nombre de transposicions en què descompon el producte $\pi(\Gamma_1) \cdot \pi(\Gamma_2)$



Correspondències

Objectiu: trobar una aplicació d'un bocí d'un graf en un bocí de l'altre graf que mostri 'on se semblen'

Se sol assignar a cada aplicació d'aquestes un pes de similitud, i se cerca una aplicació de pes màxim (alineament òptim)

Definir un pes significatiu és difícil

Si el pes és additiu (suma de pesos de parells (node, imatge)), trobar una correspondència global injectiva òptima és un problema resolt i ben conegut (algoritme hongarès)

Però per a situacions més complicades, trobar una correspondència òptima pot ser molt difícil

Exemple: Alineament d'arbres filogenètics

Siguin T i T' dos arbres filogenètics amb les mateixes etiquetes de fulles, suposem T més petit que T'

$$p(v, v') = \frac{|C(v) \cap C(v')|}{|C(v) \cup C(v')|}$$

Donada aplicació $f : T \longrightarrow T'$,

$$p(f) = \sum_{v \text{ node de } T} p(v, f(v))$$

Correspondència òptima: $f : T \longrightarrow T'$ injectiva amb $p(f)$ màxim.

Exemple: Alineament de xarxes de proteïnes

Siguin N i N' dues xarxes d'interaccions de proteïnes amb n i m nodes ($n \leqslant m$)
Sigui s una funció de similitud de proteïnes

- $p_0(i,j) = s(p_i, p_j)$
- $p_{k+1}(i,j) =$ suma de les similituds p_k dels nodes adjacents a p_i i p_j (normalitzada)
- $p = \lim_{k\to\infty} p_k$

Donada aplicació $f : N \longrightarrow N'$,

$$p(f) = \sum_{v \text{ node de } N} p(v, f(v))$$

Correspondència òptima: $f : N \longrightarrow N'$ injectiva amb $p(f)$ màxim.

Conclusió

- Els grafs s'empren en molts camps de la biologia
- La comparació d'aquests grafs és un problema important
- No està resolt, segurament no estarà resolt del tot mai, i està obert a les aportacions de tots els matemàtics i informàtics


LOONEY TUNES
"That's all Folks!"
A WARNER BROS. CARTOON