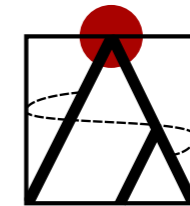


A fajfa és génfák közös rekonstrukciója



Szöllősi Gergely

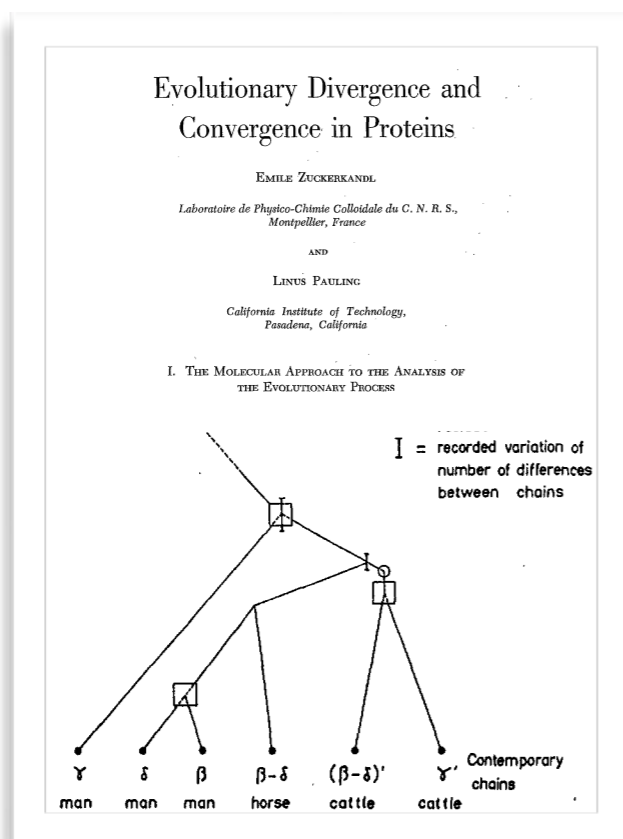
ELTE-MTA „Lendület” Biofizika Kutatócsoport
UMR 5558 CNRS LBBE Lyon, Franciaország



AGENCE NATIONALE DE LA RECHERCHE
ANR

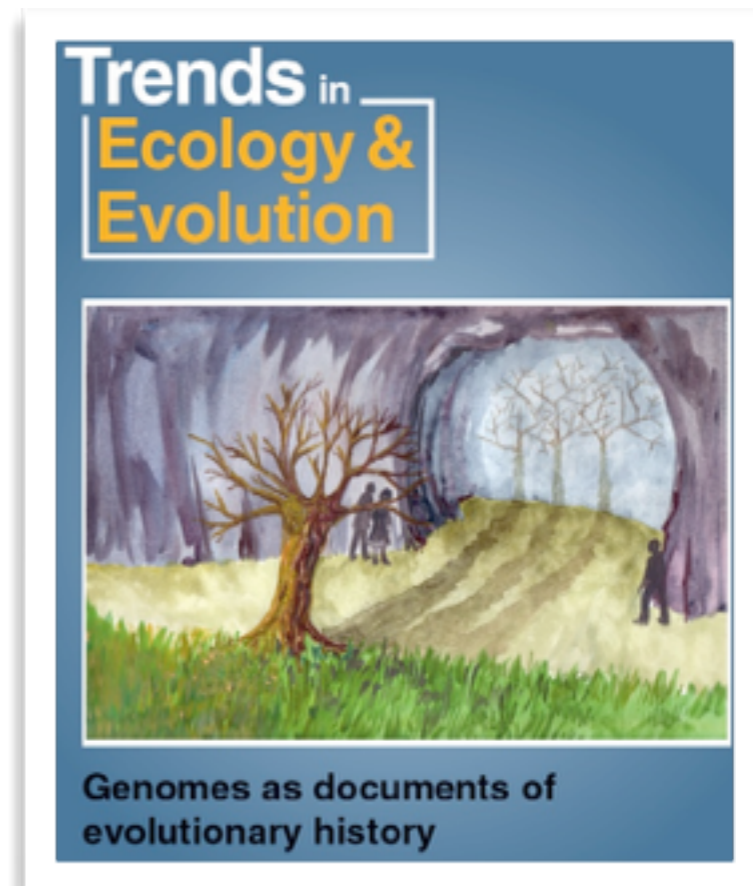
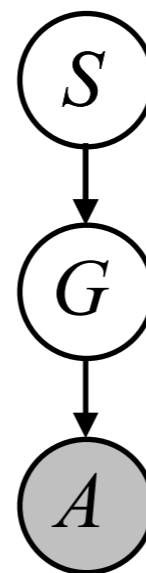


ssolo@elte.hu



Zukerkandl & Pauling 1965

DTL



Daubin & Boussau 2011

A genom az élőlények genetikai tervrajza

Minden élőlény fejlődési tervét és működési programját egy hosszú, DNS-molekulában íródott genetikai szöveg, az élőlény ún. *genomja* tartalmazza.

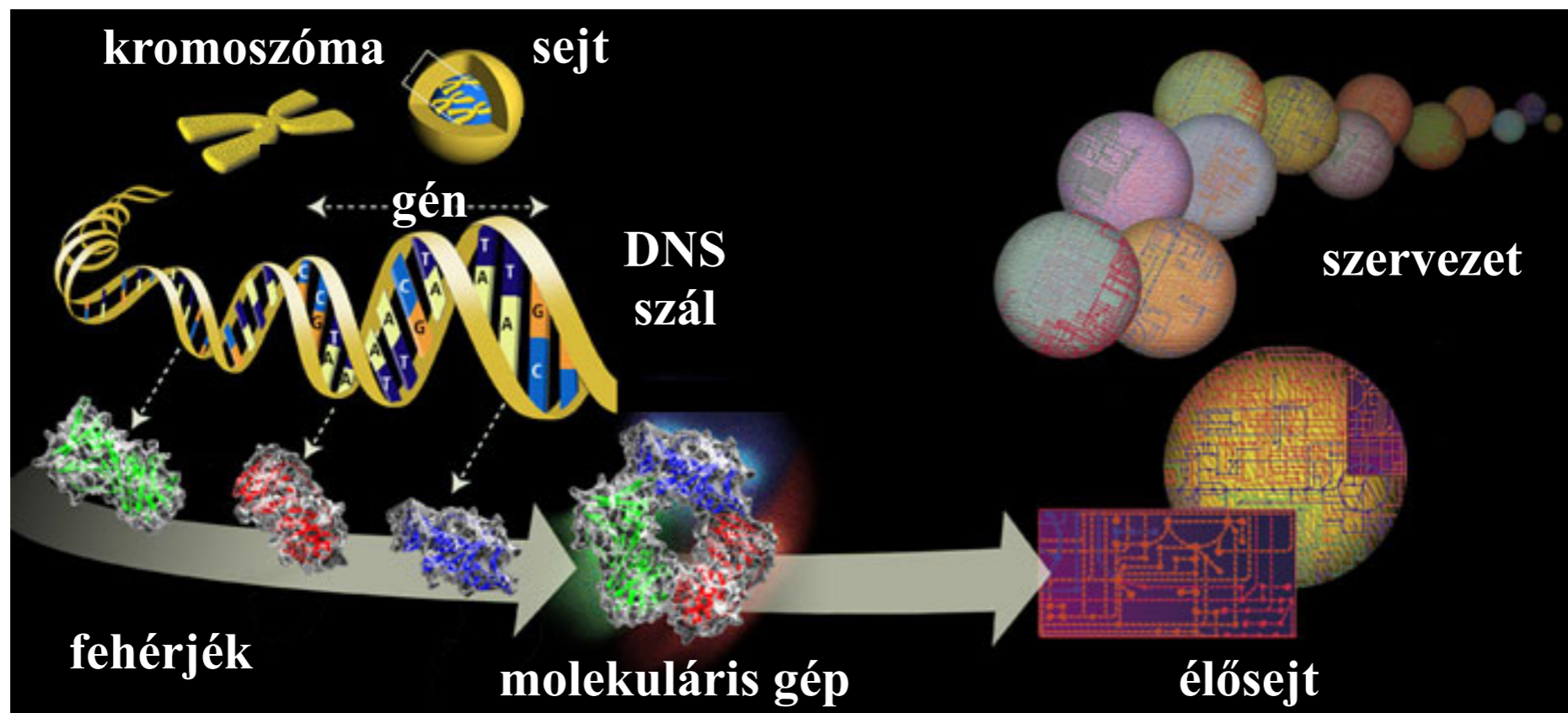
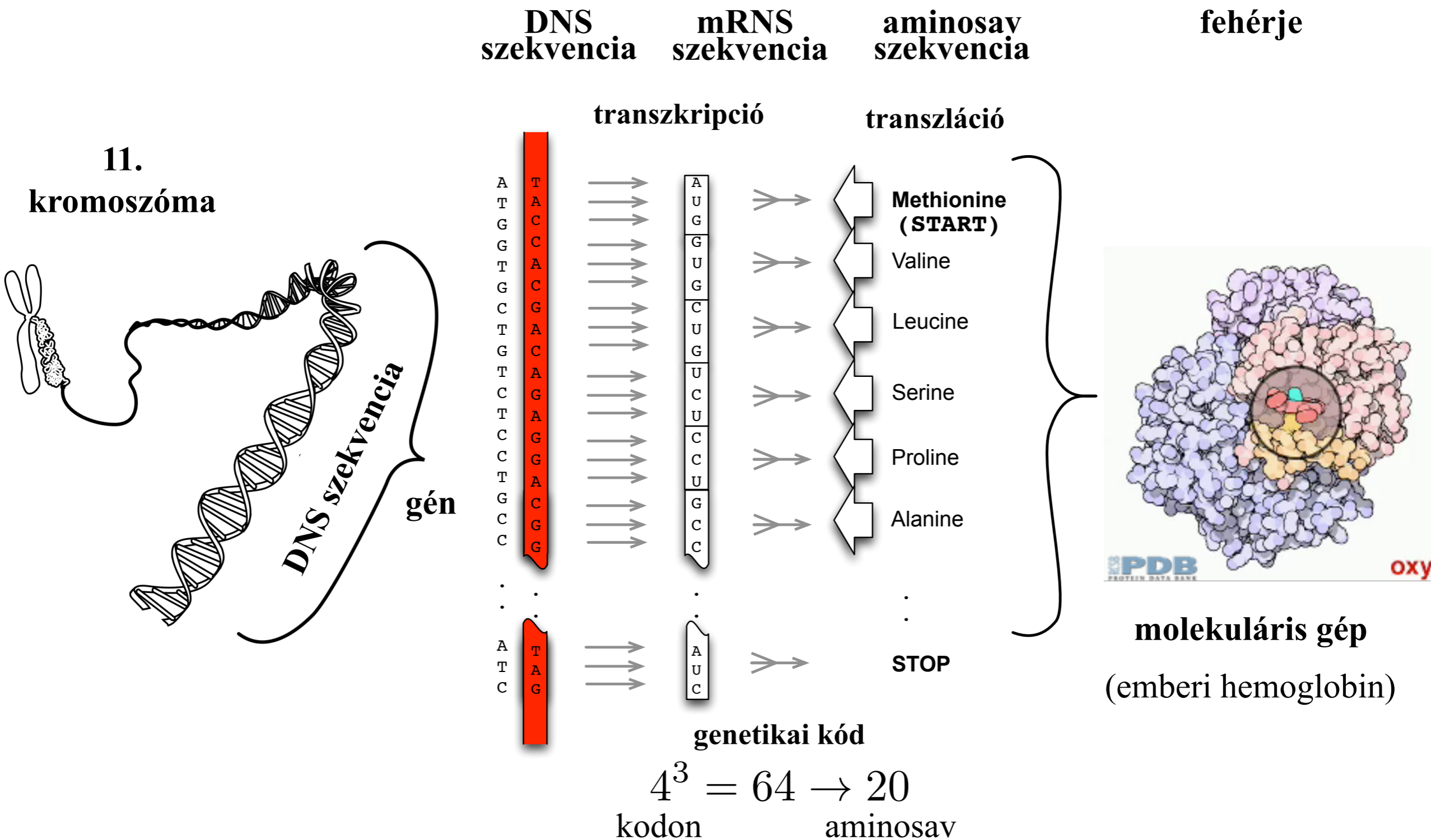


Image courtesy of U.S. Department of Energy Genome Programs
and wikimedia commons



A genom DNS-molekulákban íródott genetikai szöveg

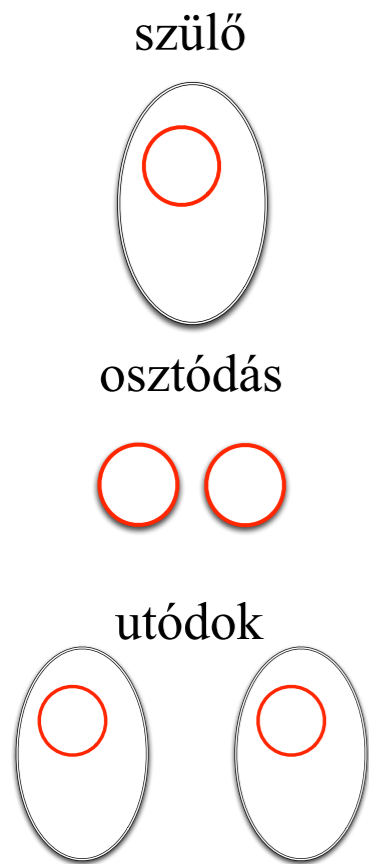
A genom genetikai szövegét alkotó egyes szavak a gének, amik fehérjéket kódolnak.



Az öröklődés során a genom szövege lemásolódik

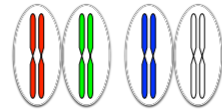
A szaporodás részleteitől függetlenül két rokon DNS szekvencia lokálisan mindig egy múltbeli DNS másolási eseményre (replikációra) vezethető vissza.

**aszexuális
szaporodás**



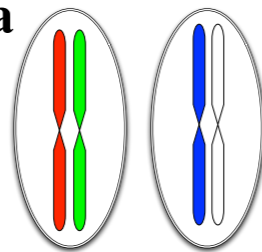
**szexuális
szaporodás**

nagyszülők

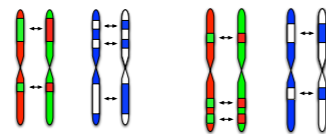


szülők

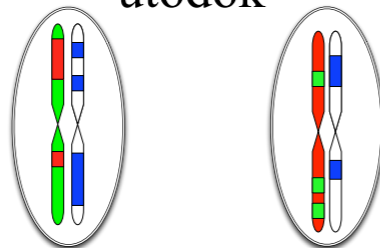
apa anya



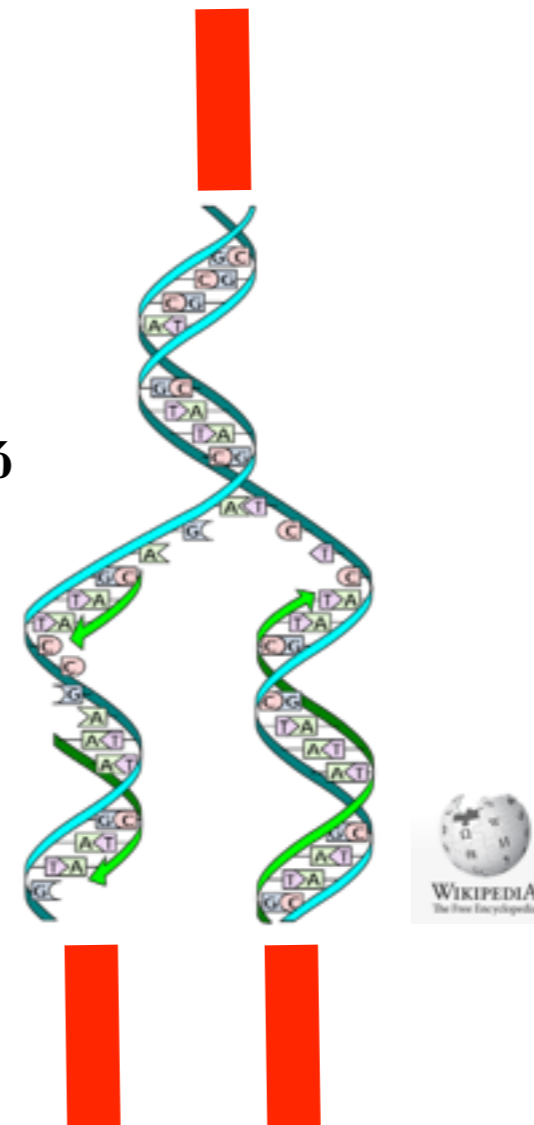
rekombináció



utódok

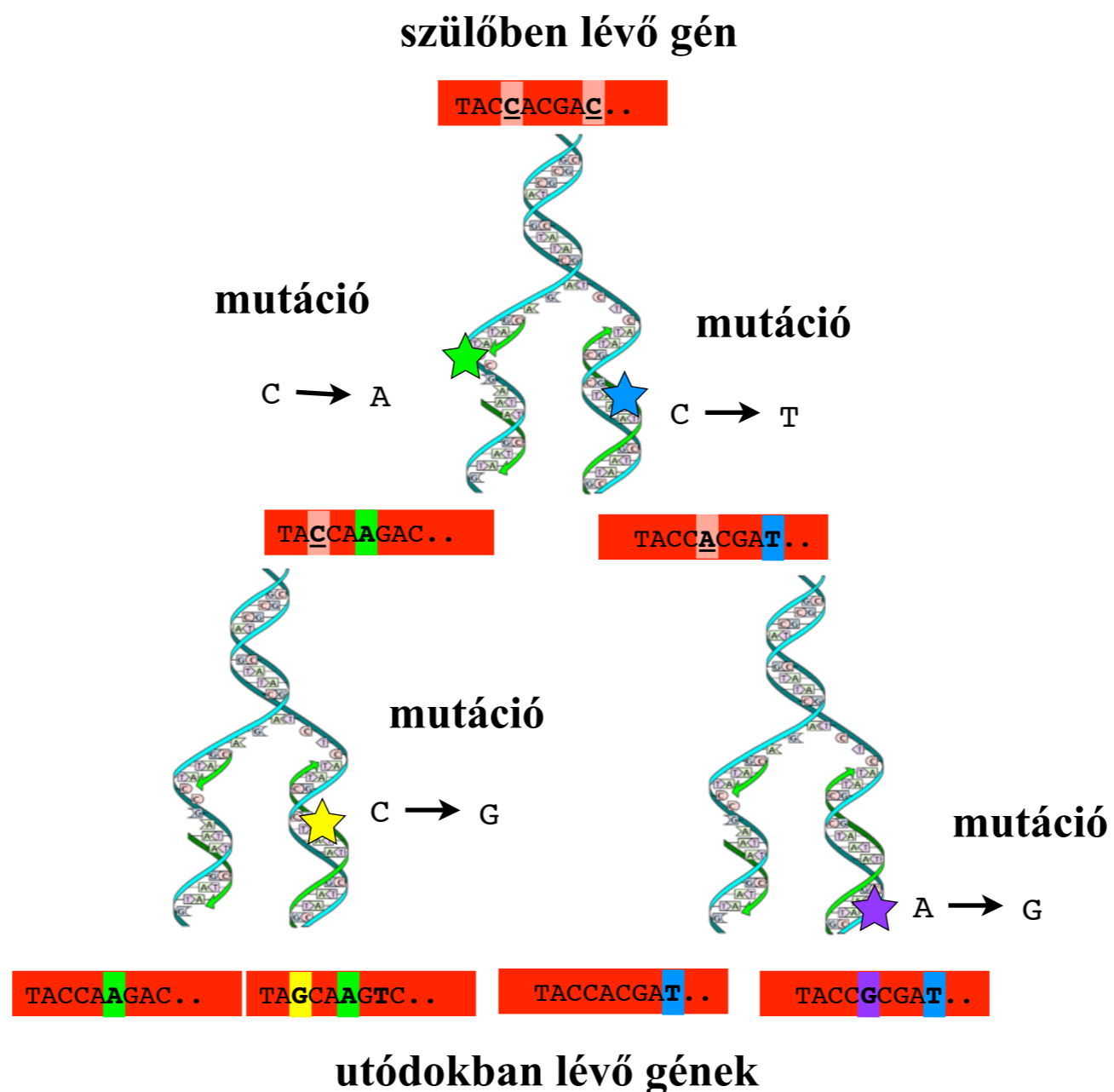


**DNS
replikáció**



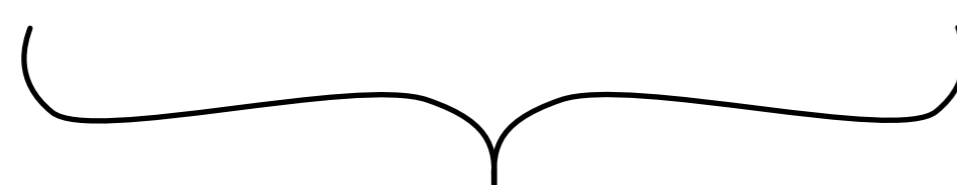
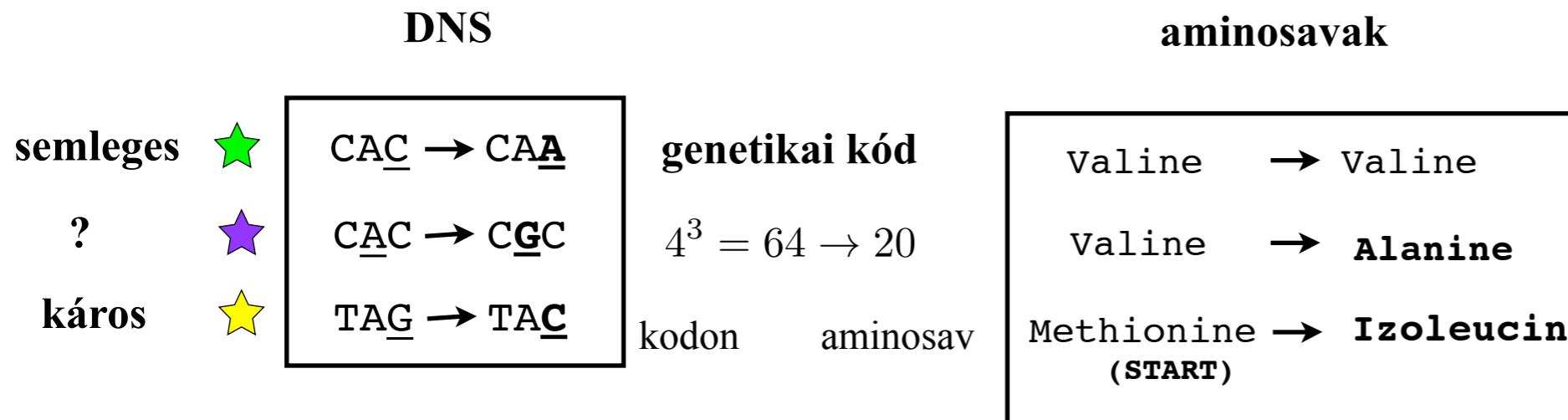
A másolás során bekövetkező hibák öröklődnek

A DNS replikáció során a génekbe kerülő hibák (mutációk) sorsa semleges változások esetén a véletlenül, egyébként pedig azon múlik, hogy az élőlény számára hasznosak-e vagy sem.



A másolás során bekövetkező hibák öröklődnek

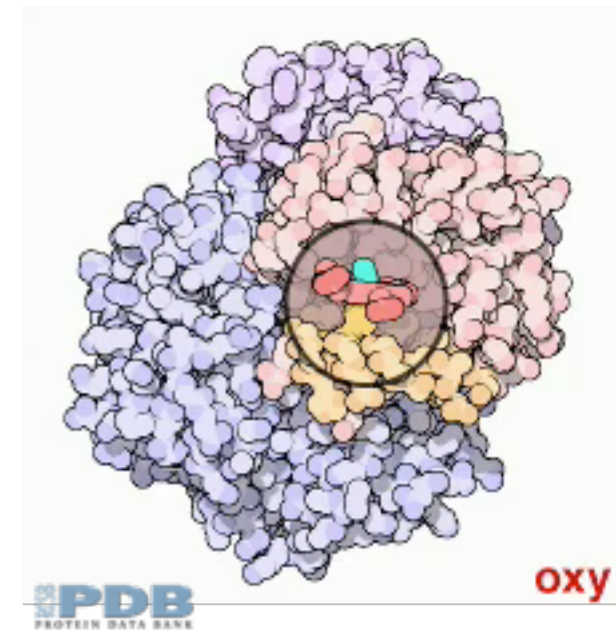
A DNS replikáció során a génekbe kerülő hibák (mutációk) sorsa semleges változások esetén a véletlenül, egyébként pedig azon múlik, hogy az élőlény számára hasznosak-e vagy sem.



empirikus *szubsztitúciós modellek*

TN92, TN93, F81, HKY85, GTR, TKF91, TKF92, WAG, BLOSUM, PAM, JTT92, LG08, REV, MTREV, GY94, MG95, NY98, M0, M1, . . . M13, CAT (és CAT újra), MKv, Dayhoff, JC69, K2P, K3P, ECM, DEC, BM, OU, EB, CATBP, GG98, TS98, G01, UCLN, UCG, RLC, ACLN, CIR, és WN

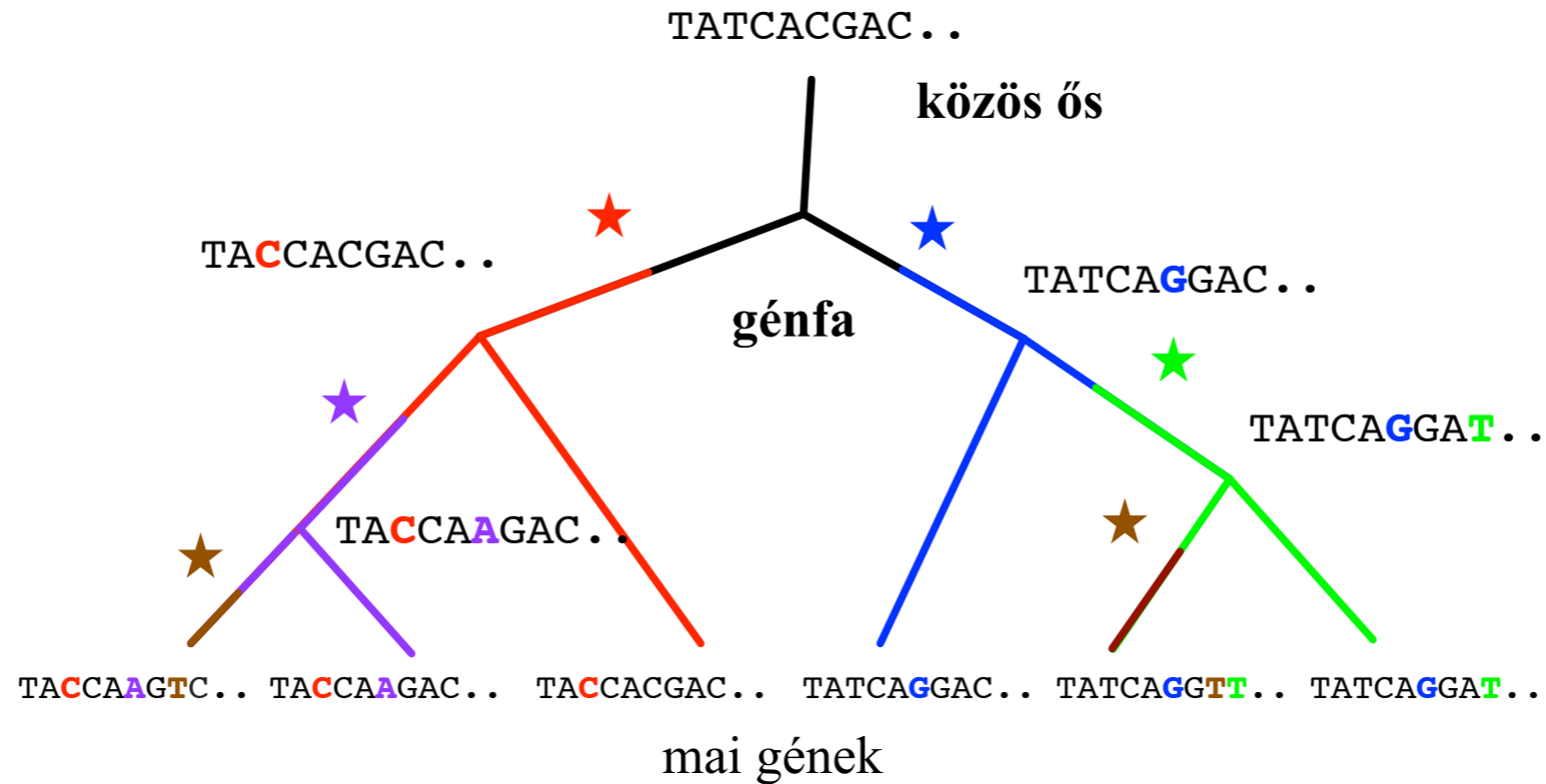
molekuláris gép



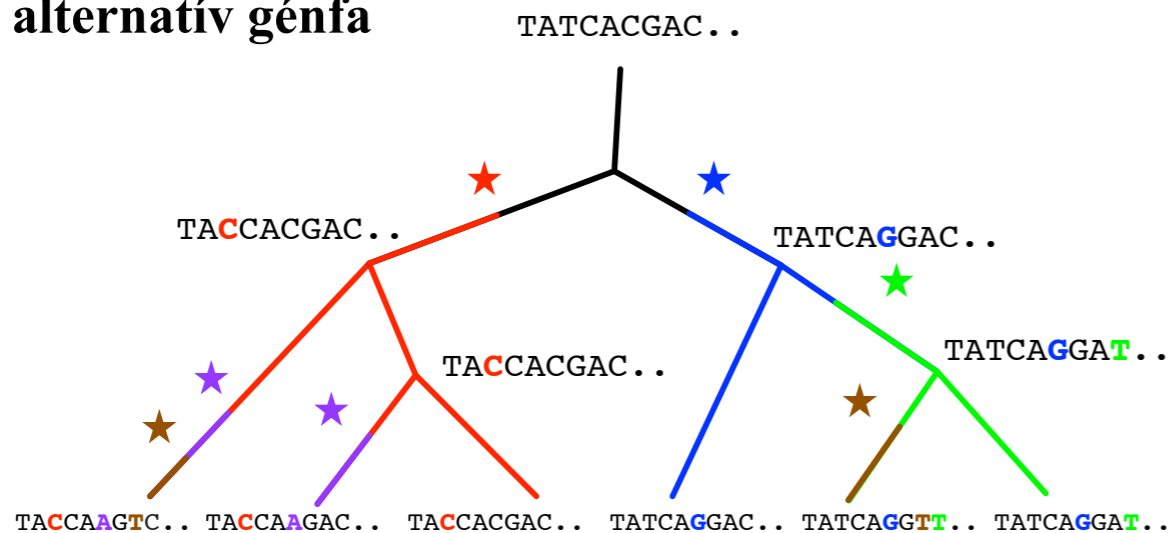
(emberi hemoglobin)

A rokon gének története rekonstruálható

Rokon gének alapján rekonstruálható a gének családfája, a *génfa*. A fa elágazásai ősi génreplikációk, gyökere a gének legutoljára létezett közös őse.

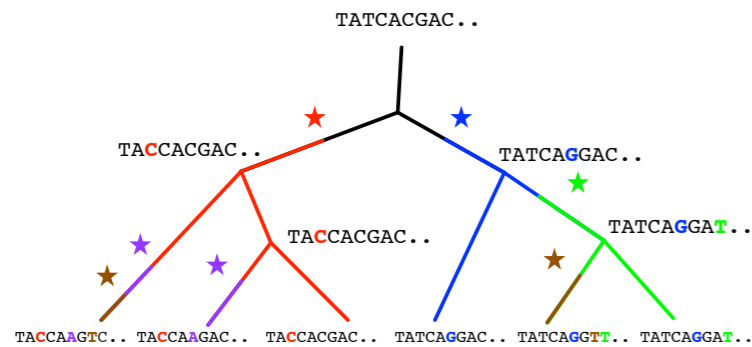
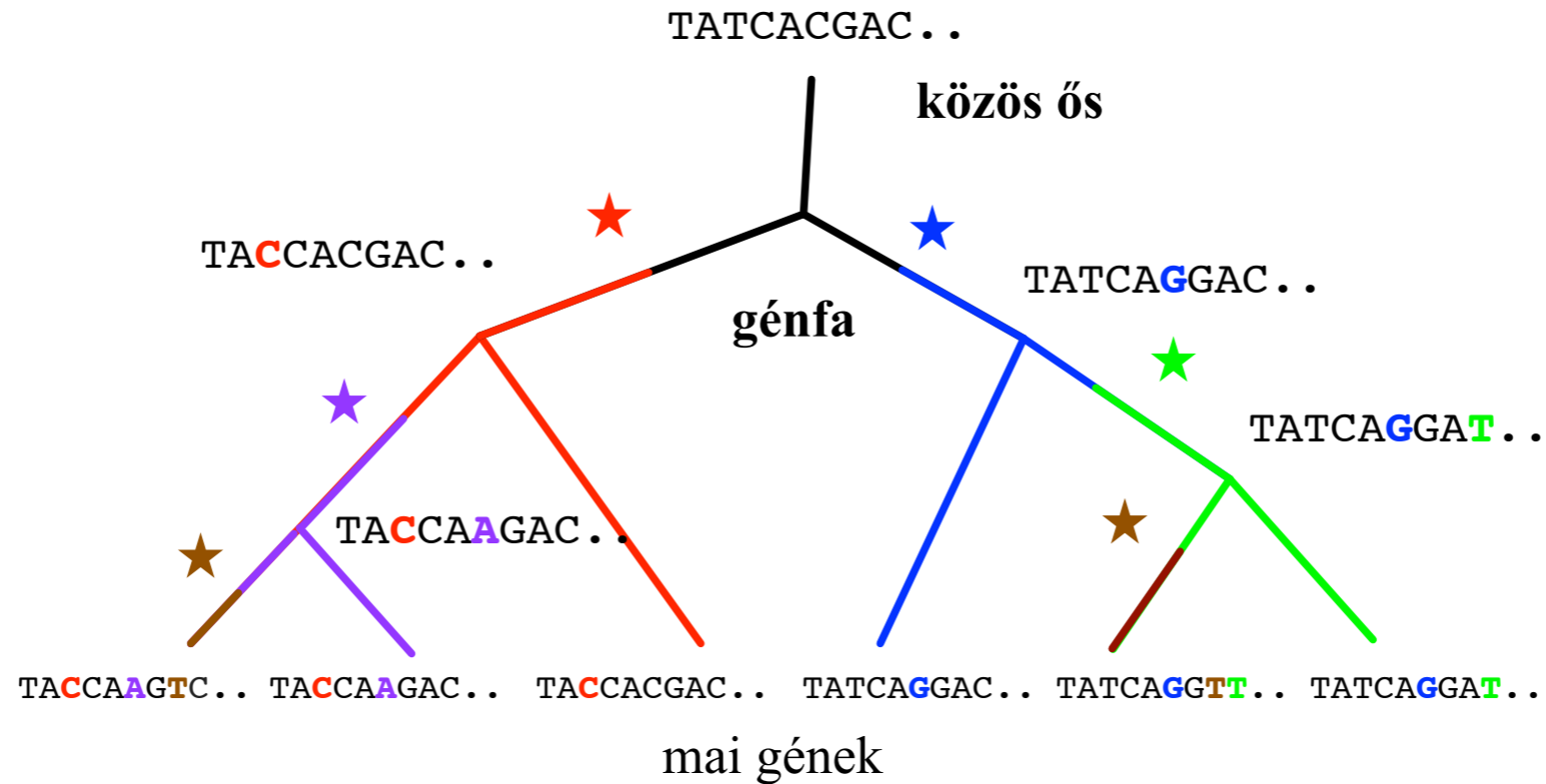


alternatív génfa

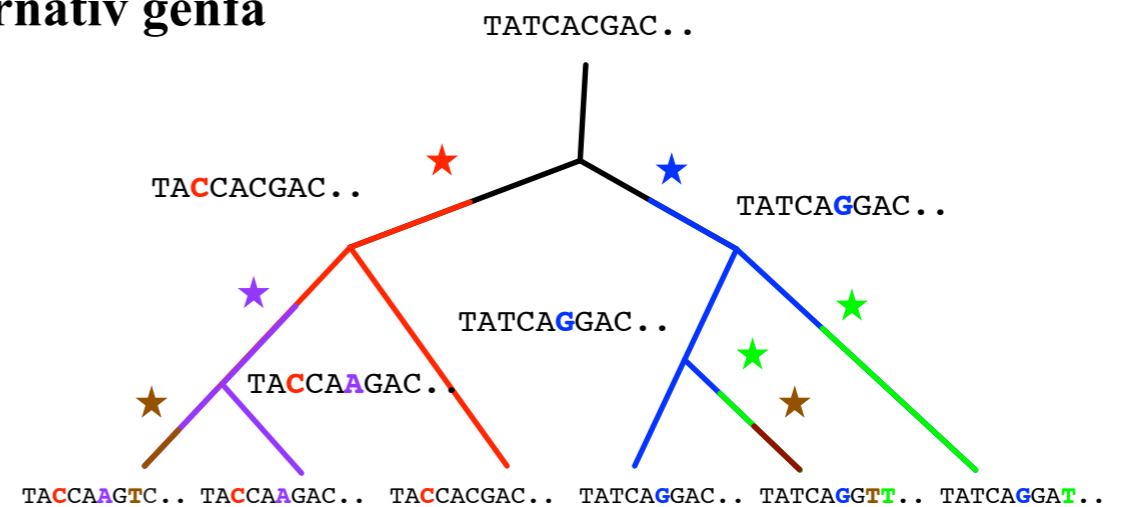


A rokon gének története rekonstruálható

Rokon gének alapján rekonstruálható a gének családfája, a *génfa*. A fa elágazásai ősi génreplikációk, gyökere a gének legutoljára létezett közös őse.



alternatív génfa



A rokon gének története rekonstruálható

Rokon gének alapján rekonstruálható a gének családfája, a *génfa*. A fa elágazásai ősi génreplikációk, gyökere a gének legutoljára létezett közös őse.

A megfigyelés a ma megtalálható génszekvenciák:

TATCAAGTC..
 TATCAAGAC..
 TATCACGAC..
 TACCAGGAC..
 TACCAGGTT..
 TACCAGGAT..

A rekonstrukció a maximum likelihood génfa:

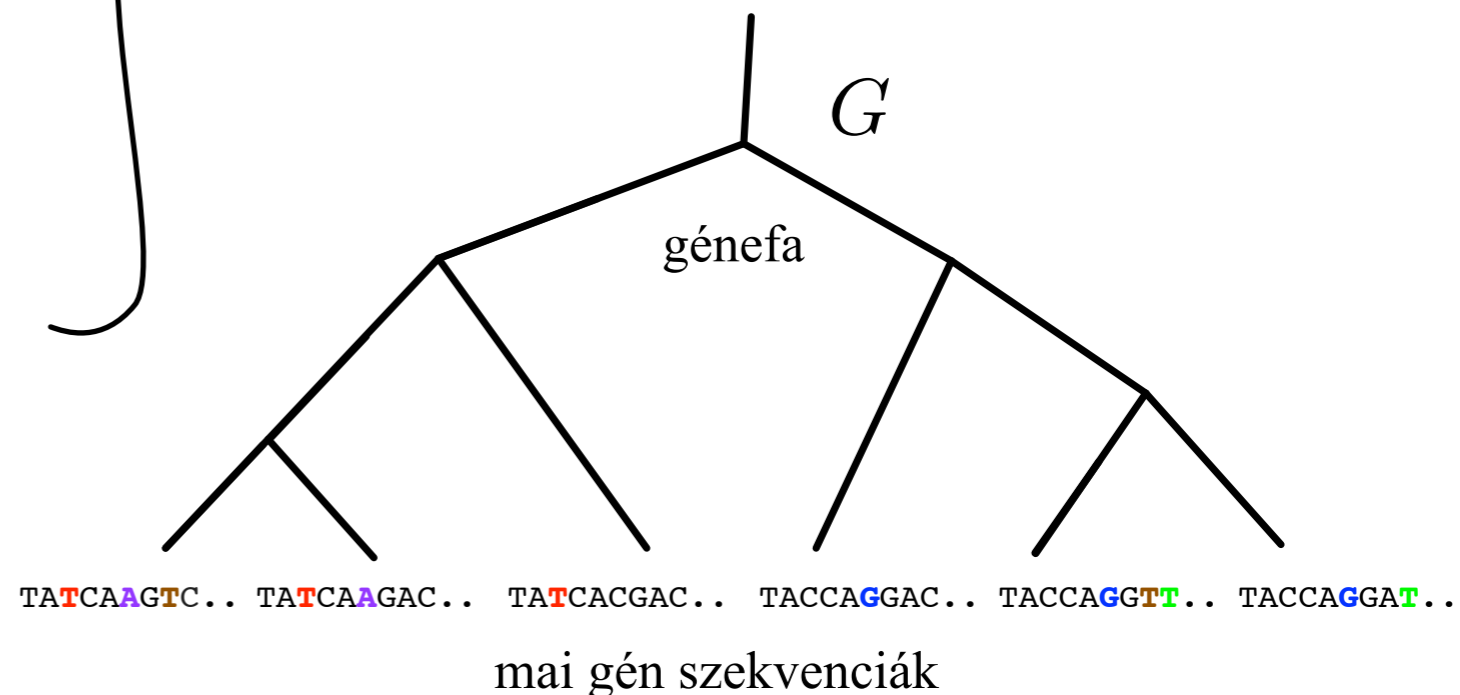
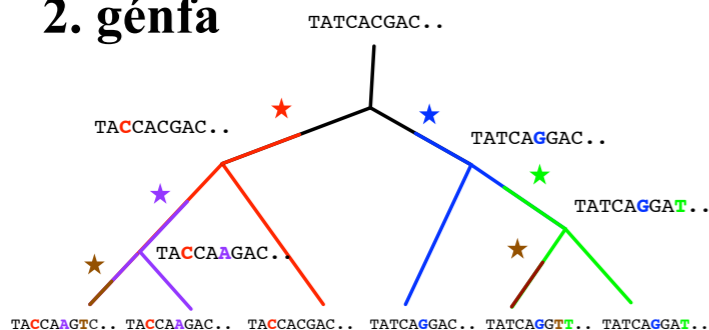
$$p(\text{szekvenciák} | G)$$

A modell egy sztochasztikus szekvenciaevolúció-modell:

1. szubsztitúciós modell

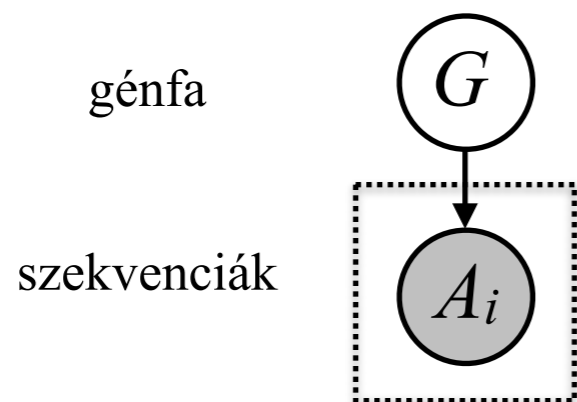
TN92, TN93, F81, HKY85, GTR, TKF91, TKF92, WAG, BLOSUM, PAM, JTT92, LG08, REV, MTREV, GY94, MG95, NY98, M0, M1, . . . M13, CAT (és CAT újra), MKv, Dayhoff, JC69, K2P, K3P, ECM, DEC, BM, OU, EB, CATBP, GG98, TS98, G01, UCLN, UCG, RLC, ACLN, CIR, és WN

2. génfa

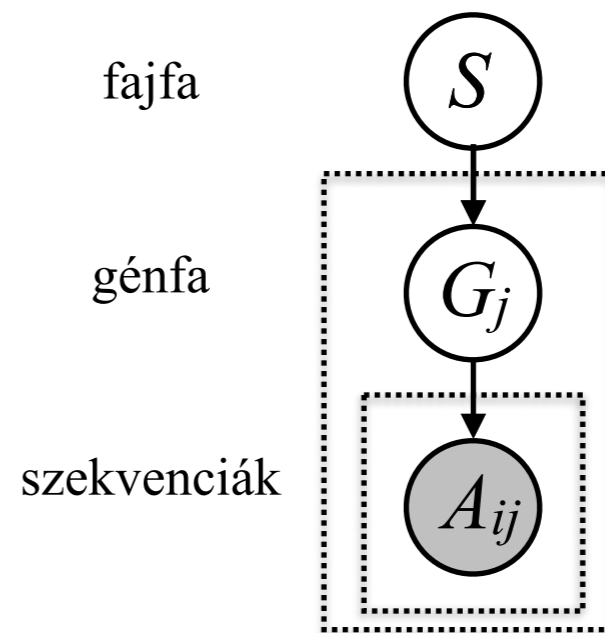


A molekuláris filogenetika szekvenciák alapján rekonstruál génfákat.

**rekonstrukció
egyedi gének alapján**



közös rekonstrukció



A rokon gének története rekonstruálható

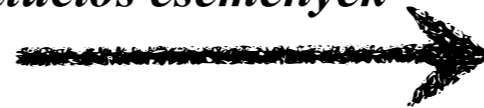
Ahhoz, hogy a *mai szekvenciák* valószínűségét kiszámoljuk, összegeznünk kell az adott *génfa mentén* történő összes lehetséges *szubsztitúciós történet* felett.

A megfigyelt szekvenciák A valószínűsége adott G -re:

$$p(A|G)$$

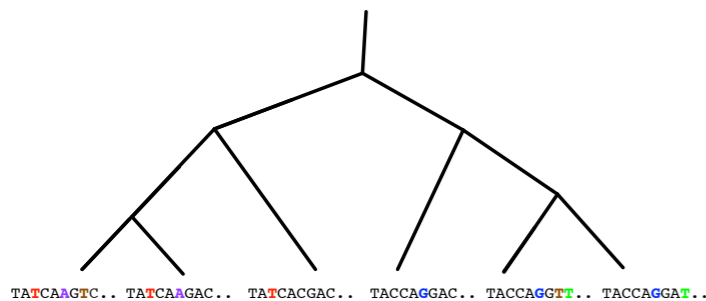
Felsenstein 1981

szubsztitúciós események



G

génfa

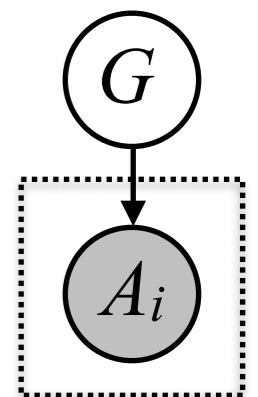
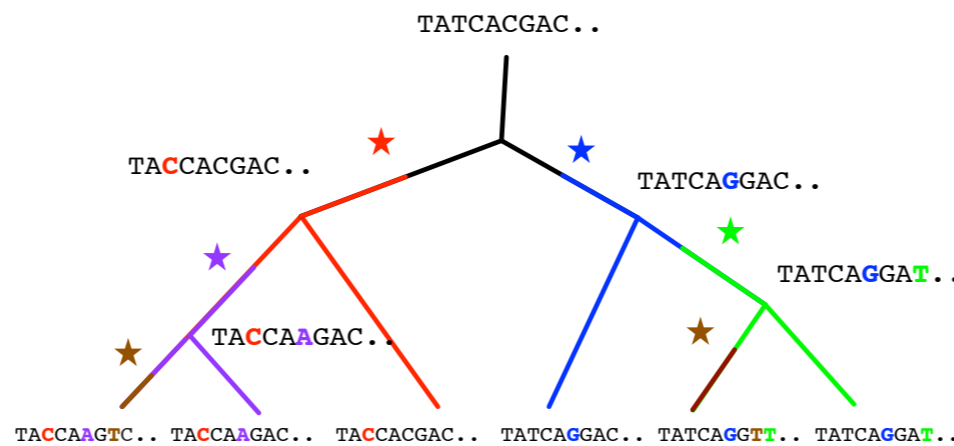


A

szekvenciák:

TATCAAGTC..
 TATCAAGAC..
 TATCACGAC..
 TACCAAGAC..
 TACCAAGTT..
 TACCAAGAT..

$i = 1 \dots L$

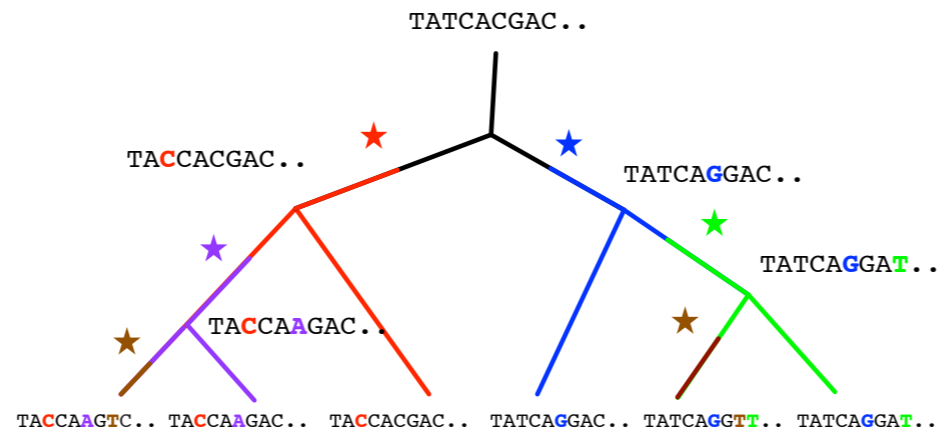


Az egyes géntörténetek elmosódottak

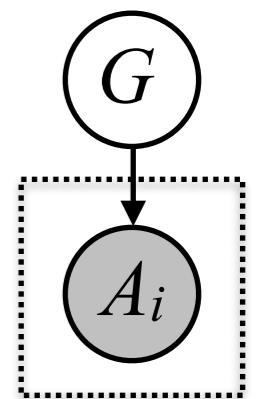
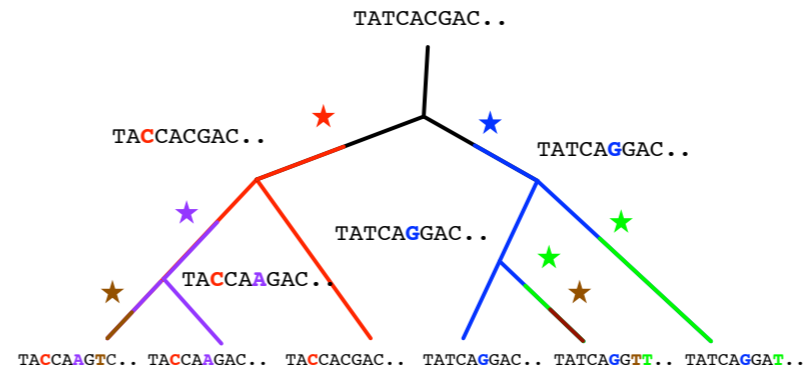
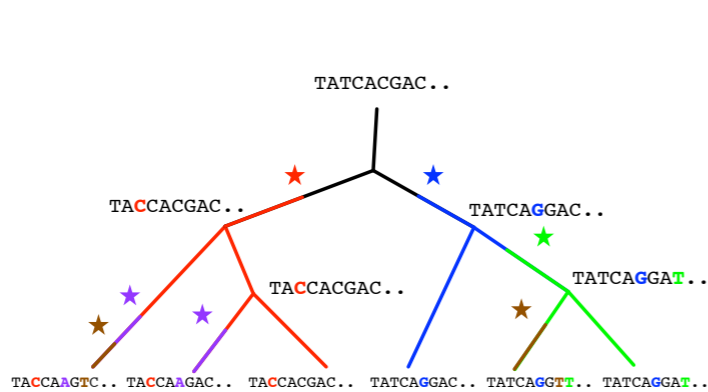
A rokon gének szekvenciája alapján jósolható génfák elmosódottak, a rekonstrukció során szinte mindig statisztikailag nem, vagy csak alig megkülönböztethető lehetőségek közül kell választanunk.

A maximum likelihood génfa

$$p(\text{szekvenciák} | G)$$

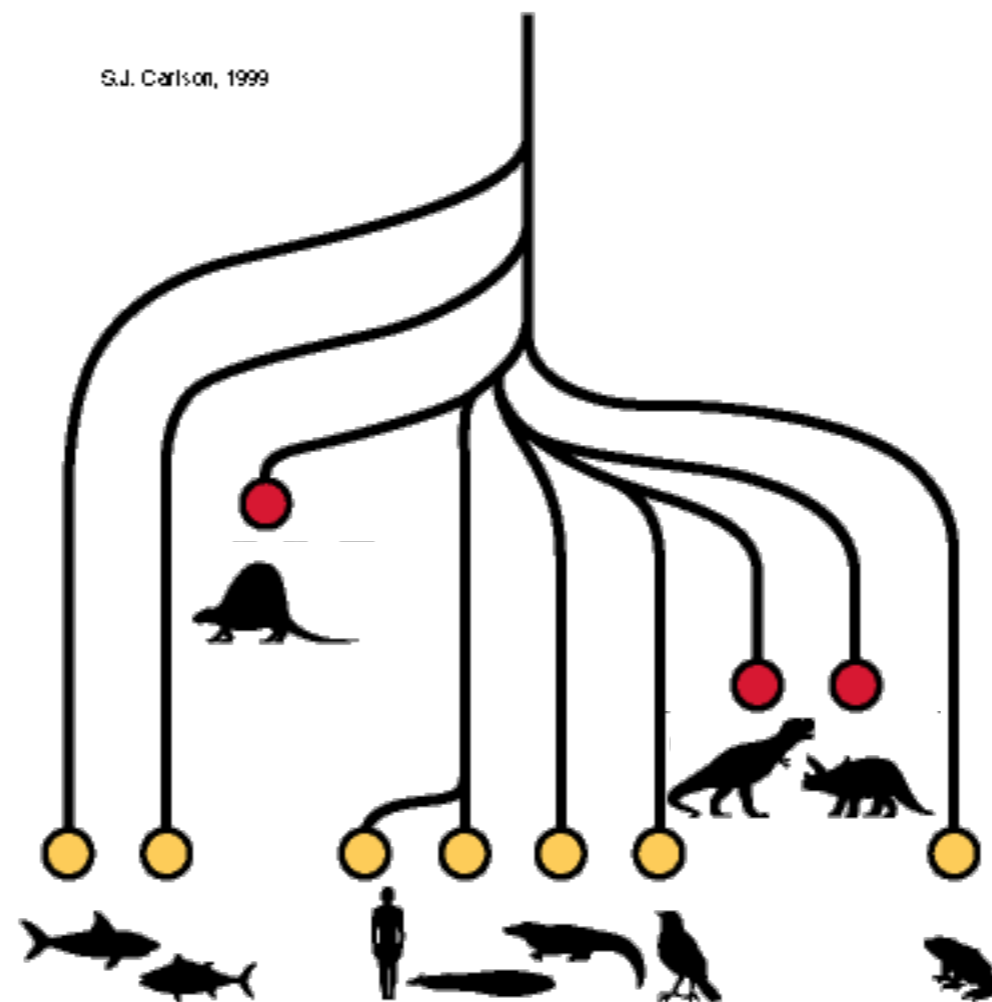


környezetében gyakran sok **statisztikailag hasonló génfa van:**



Ami igazán érdekes, az a fajfa..

A történet, amit rekonstruálni akarunk, nem az egyes DNS replikációs események története, hanem új fajok keletkezéséé és régieknek kihalásáé.

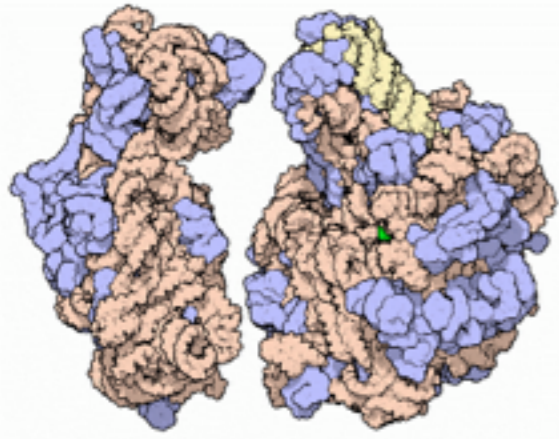


ma élő gerincesek

Kell egy különleges génfa..

Egyedi gének óvatos kiválasztásával információt kaphatunk a fajfáról, így derült fény az élet három nagy doménjére.

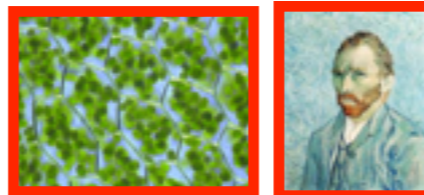
Minden faj genomjában megtalálható gének:



egyetlen gén 16S rRNS

Carl Woese, 1977

Eukarióták



Animals

Plants

Protozoa

Euryarchaeota

Crenarchaeota

Archeák



Baktériumok



Firmicutes

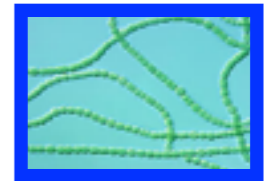
Chlamydiae

Planctomycetes

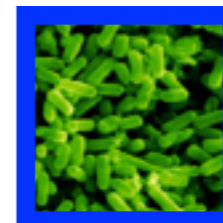
Actinobacteria

Fusobacteria

Cyanobacteria

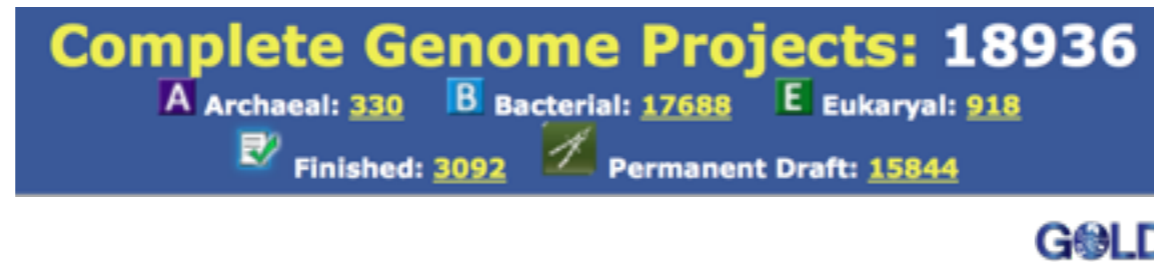


Proteobacteria

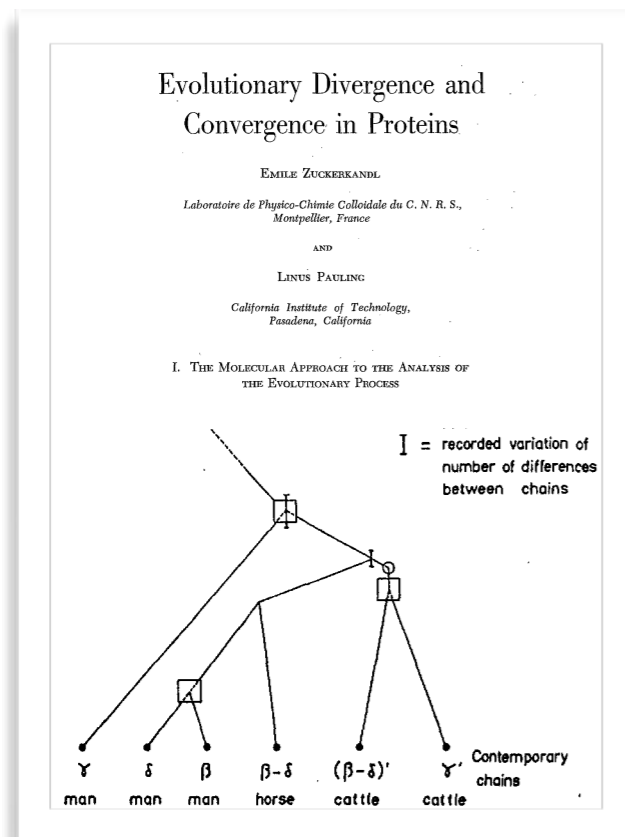


A molekuláris evolúciókutatás aranykora?

Egy organizmus teljes genetikai szövegének a meghatározásának ára az exponenciálisnál gyorsabban csökkent.

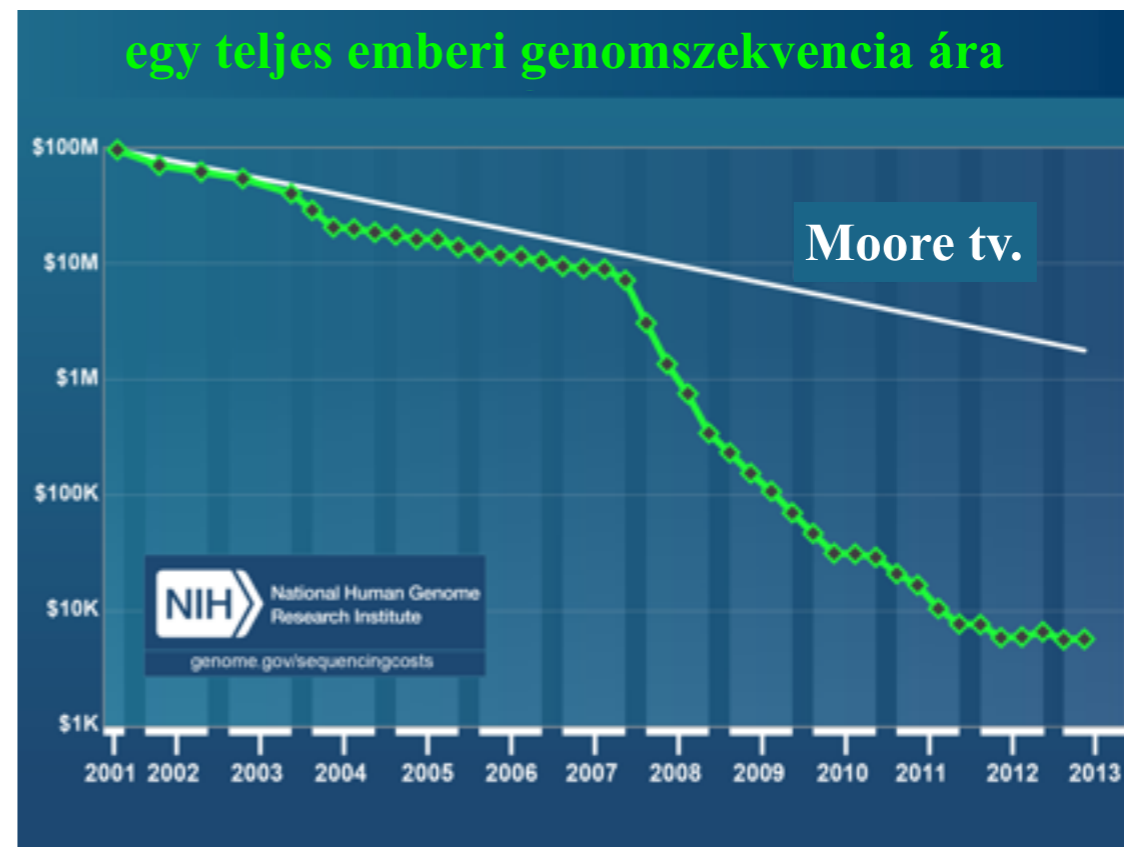


Az első génfa



Zukerkandl & Pauling 1965

Hemoglobin oldalláncok, 148 aminosav



hemoglobin

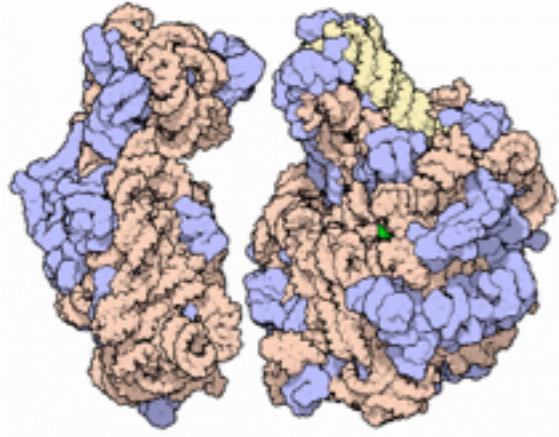
Az emberi genom 6×10^9 bázispár, kb. 2 méter.

E.coli baktérium 4.6×10^6 bázispár, kb. 1.5 milliméter.

Az 1% történelme

Egyedi gének óvatos kiválasztásával információt kaphatunk a fajfáról, így derült fény az élet három nagy doménjére.

Minden faj genomjában megtalálható gének:



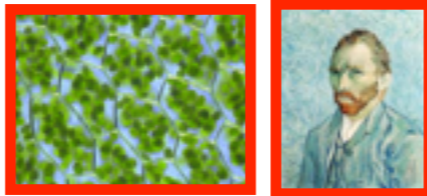
egyetlen gén 16S rRNS

Carl Woese, 1977

maximum pár tucat gén

Ciccarelli, 2006

Eukarióták



Animals

Plants

Protozoa

Euryarchaeota

Crenarchaeota

Archeák



Baktériumok



Firmicutes

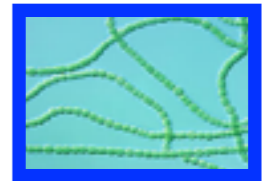
Chlamydiae

Planctomycetes

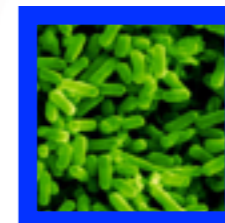
Actinobacteria

Fusobacteria

Cyanobacteria



Proteobacteria



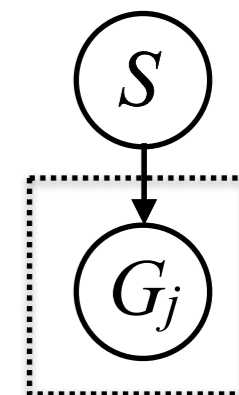
(a gének 1%-ának története)

.. de a génfák a fajfa mentén keletkeznek

Az egyedi gén családok 99%-nak a története komplikált, de közös fajfa mentén zajlanak. Az egyedi történeteket összesítve pontosabb fajfát és pontosabb génfákat rekonstruálhatunk.



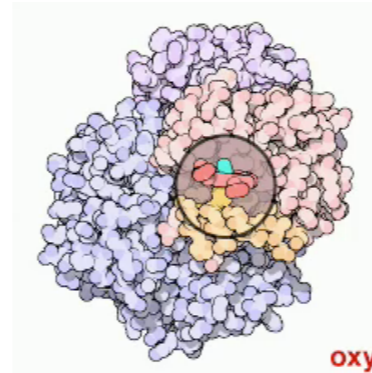
Daubin & Boussau 2011



A gének 99%-ának története komplikált

Minden gén története egyedi evolúciós események sorozata, rokon gének családjában gyakran változik a gének példányszáma és módosul a gén funkciója.

A felnőtt emberi hemoglobin



$2\alpha + 2\beta$ láncból áll.

molekuláris gép

Ember



felnőtt

$+2\beta$ (97%)

$+2\delta$ (3%)

$2\alpha +$

Tehén



felnőtt

$+2\{\beta\delta\}$

Ló



felnőtt és magzat

$+2\{\beta\delta\}$



magzat

$+2\gamma$

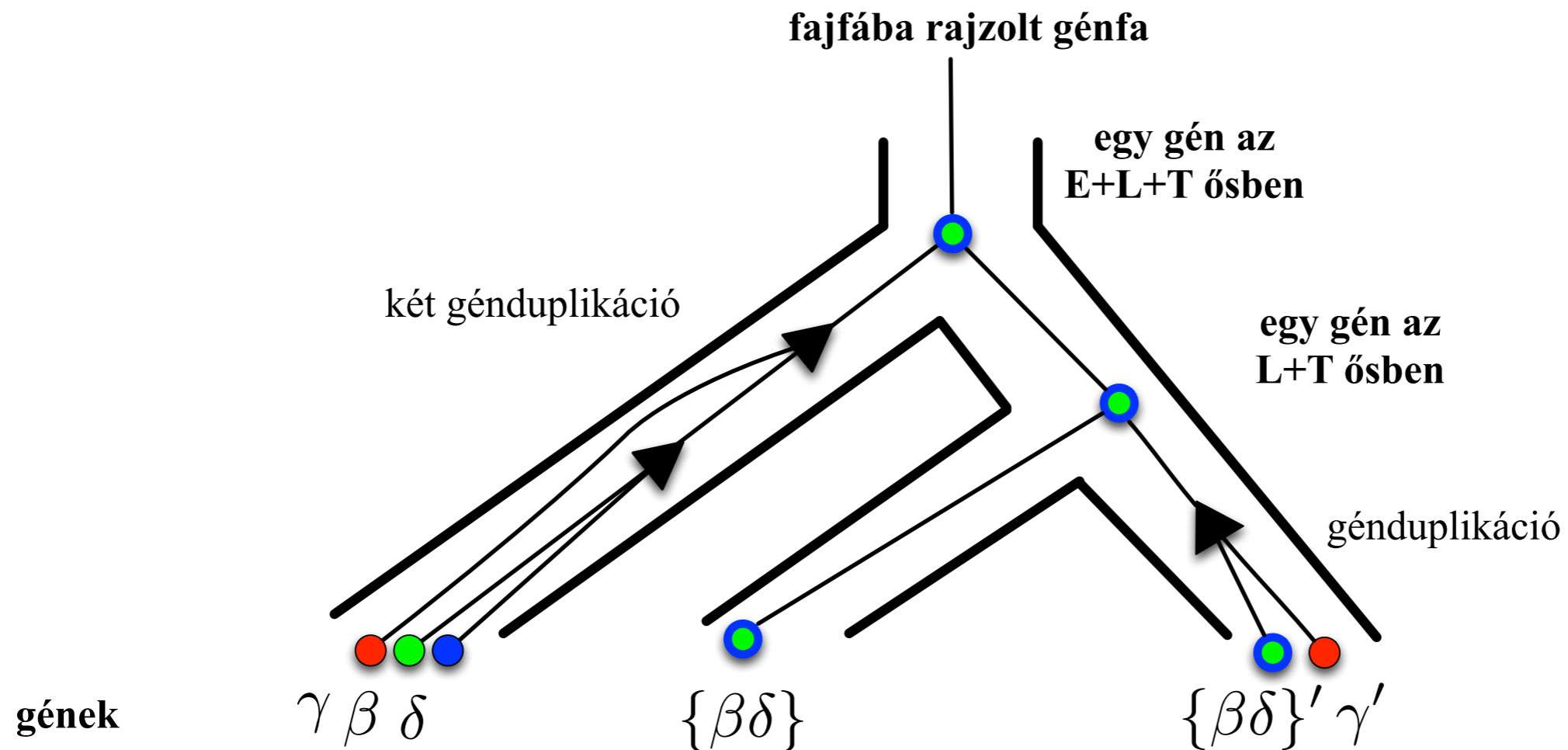


magzat

$+2\gamma$

A gének 99%-ának története komplikált

Minden gén története egyedi evolúciós események sorozata, rokon gének családjában gyakran változik a gének példányszáma és módosul a gén funkciója.



fajok



Ember



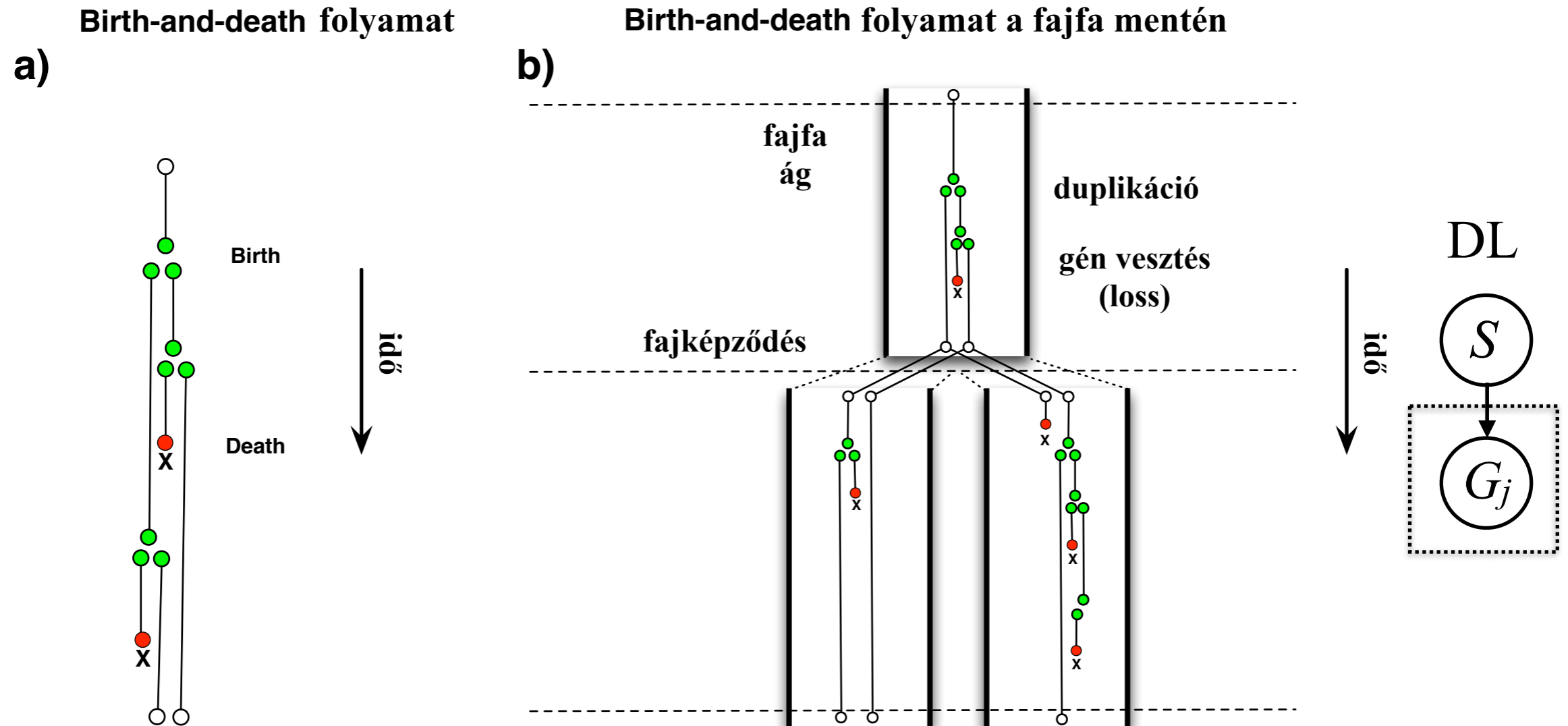
Ló



Tehén

.. de a génfák a fajfa mentén keletkeznek

Ahhoz, hogy kiszámoljuk a *génfa* valószínűségét, *összegeznünk* kell az *összes lehetséges berajzolásán a génfának a fajfába*.



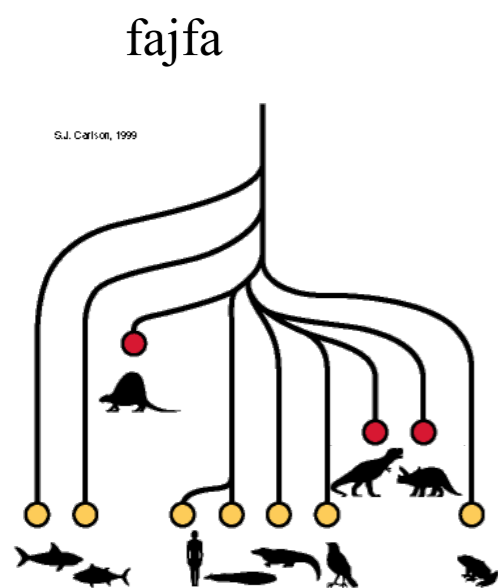
.. de a génfák a fajfa mentén keletkeznek

Ahhoz, hogy kiszámoljuk a *génfa* valószínűségét, *összegeznünk* kell az *összes lehetséges berajzolásán a génfának a fajfába*.

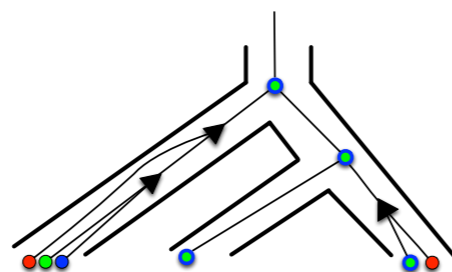
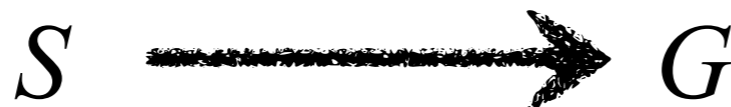
A megfigyelt génfa G valószínűsége adott S -re:

$$p(G|S)$$

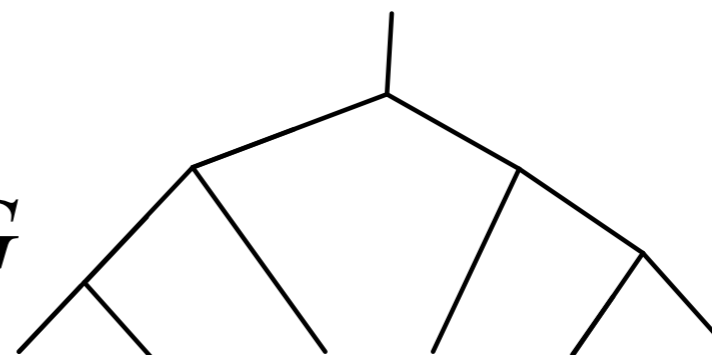
Arvestad et al. (2003) stb.



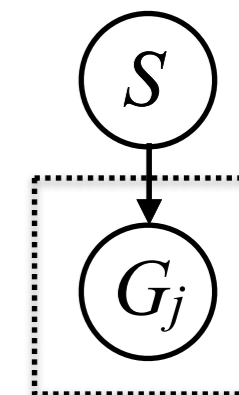
gének születése és halála
pl. duplikáció és gén vesztés



génfa

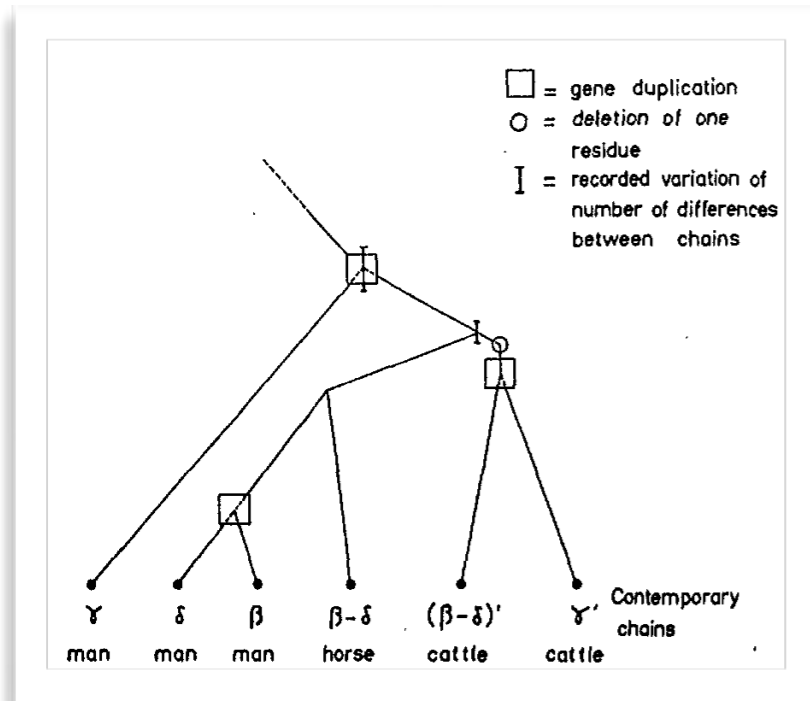


DL



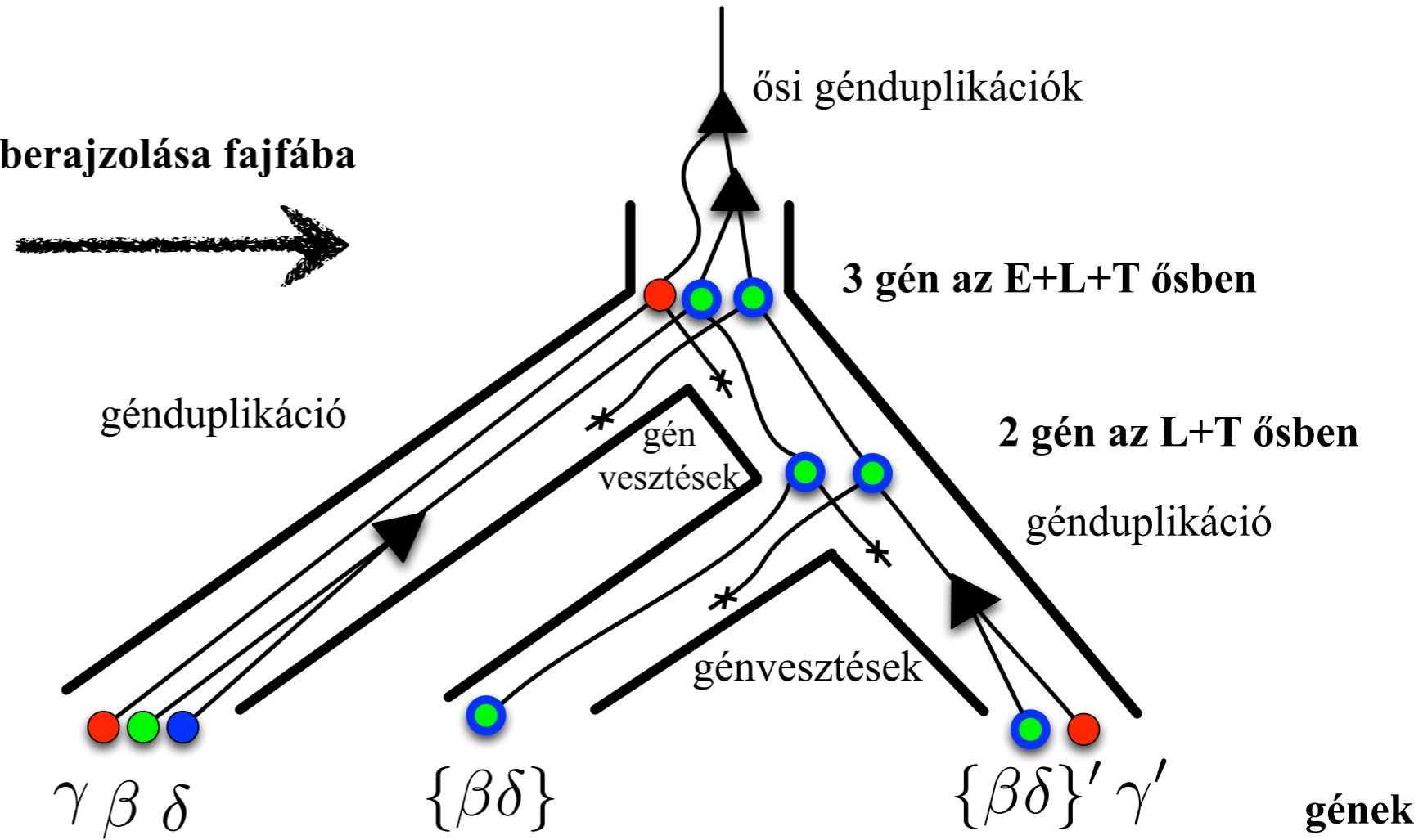
Az egyes géntörténetek elmosódottak

A rokon gének szekvenciája alapján önállóan jósolható génfák elmosódottak, gyakran hibásak, a génfákban lévő hibák erősen befolyásolhatják a rekonstruált evolúciós történetet.



Zukerkandl & Pauling 1965

génfa berajzolása fajfába



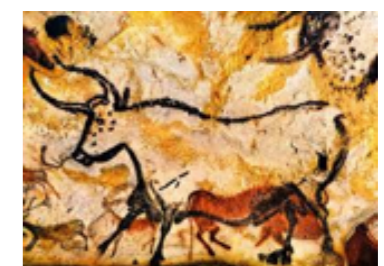
gének



Ember



Ló



Tehén

fajok

A fajfa és génfák közösen rekonstruálhatóak

A rokon gének szekvenciája alapján önállóan jósolható génfák elmosódottak, gyakran hibásak, a génfákban lévő hibák erősen befolyásolhatják a rekonstruált evolúciós történetet.

Hierarchikus generatív modell:

$$p(G|S)$$

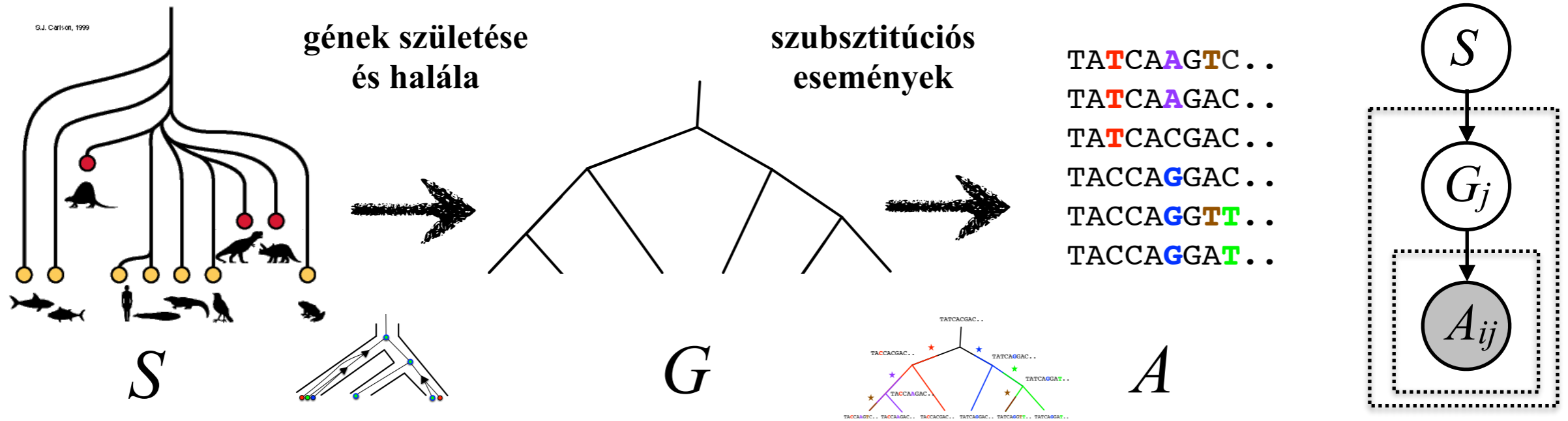
$$p(A|G)$$

faj fa

génfák

szekvenciák

DL



A fajfa és génfák közösen rekonstruálhatóak

A rokon gének szekvenciája alapján önállóan jósolható génfák elmosódottak, gyakran hibásak, a génfákban lévő hibák erősen befolyásolhatják a rekonstruált evolúciós történetet.

párhuzamos számítási eljárás

a *maximum a posteriori* fajfa keresésére

szerver:

egy közös fajfa optimalizálása

$$\mathcal{L}(\{G_j\}, S, \text{rates} | \{A_{ij}\}) :$$

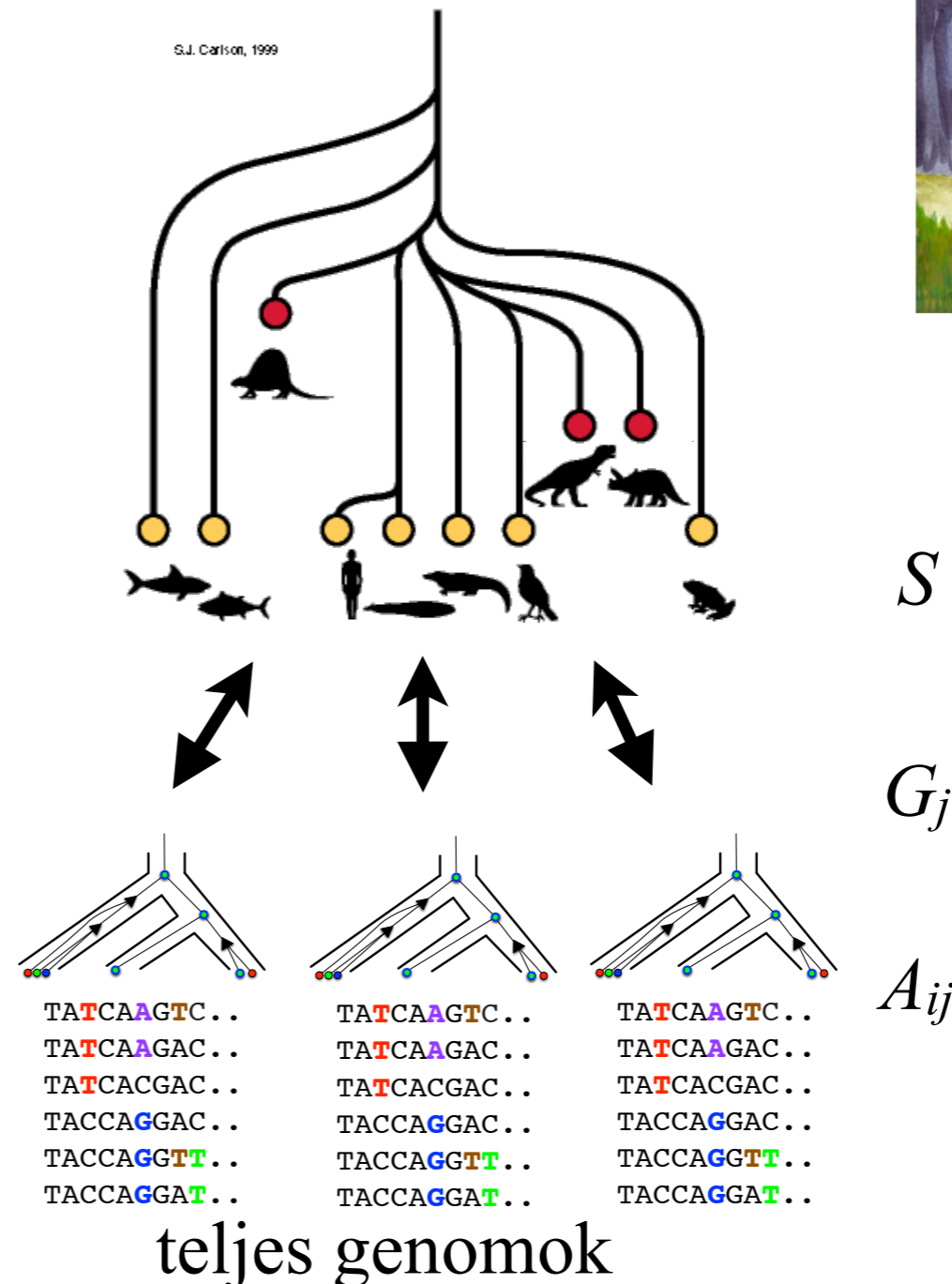
$$\prod_j$$

kliensek:

génfák optimalizálása

szekvencia- és génfa-valószínűségek szorzata

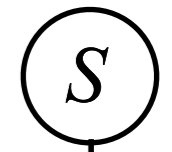
$$\prod_i p(A_{ij} | G_j) \times p(G_j | S, \text{rates})$$



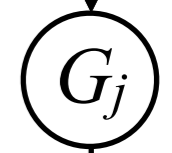
MAP

DL

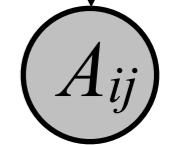
S



G_j



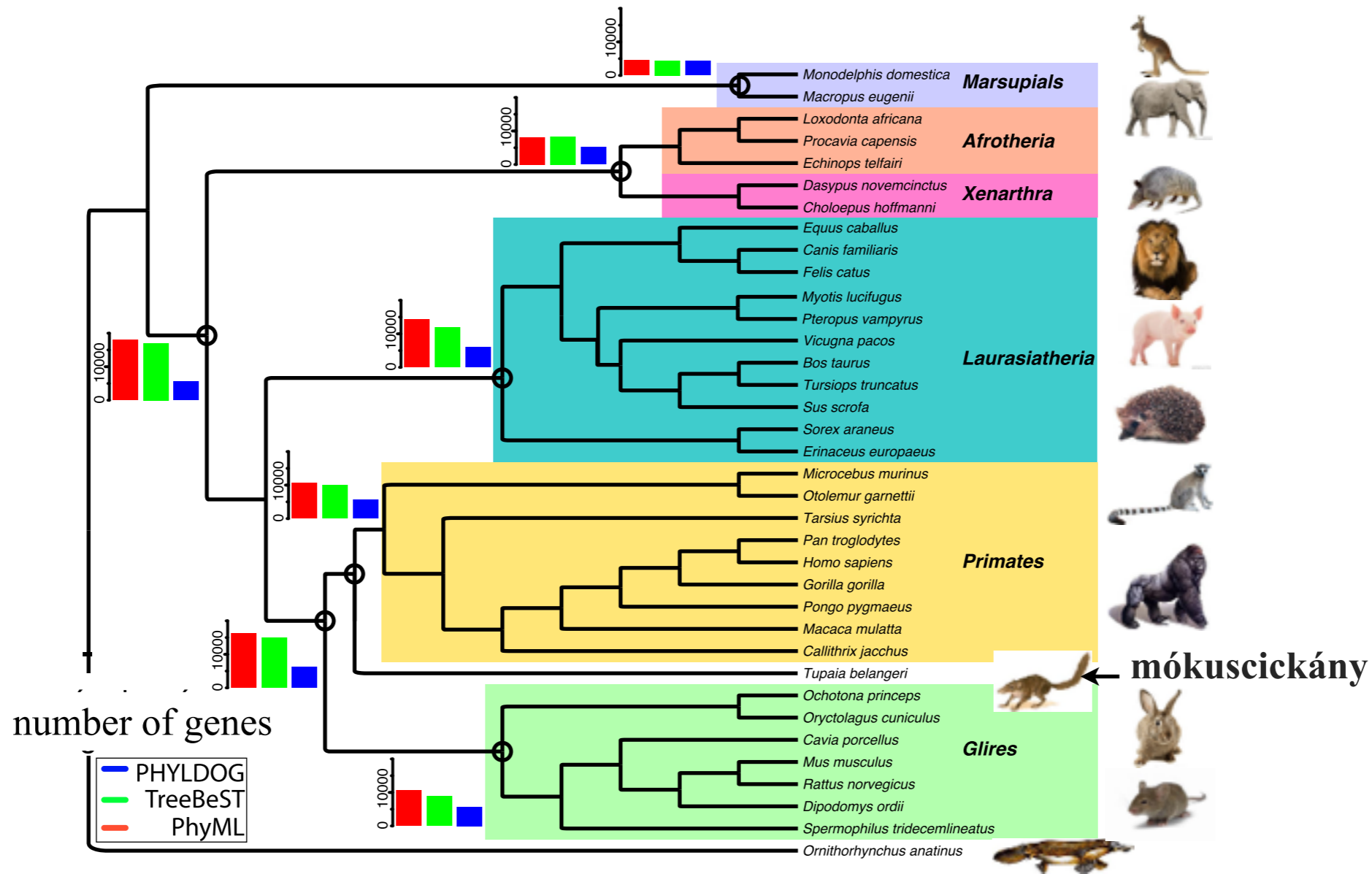
A_{ij}



Az emlősök faja

6966 gén családot használva 36 emlős genomból, **rekonstruáltuk közösen** 160 millió év evolúcióját lefedő fajtát és az a mentén futó génfákat.

emlősök faja



PHYLD OG

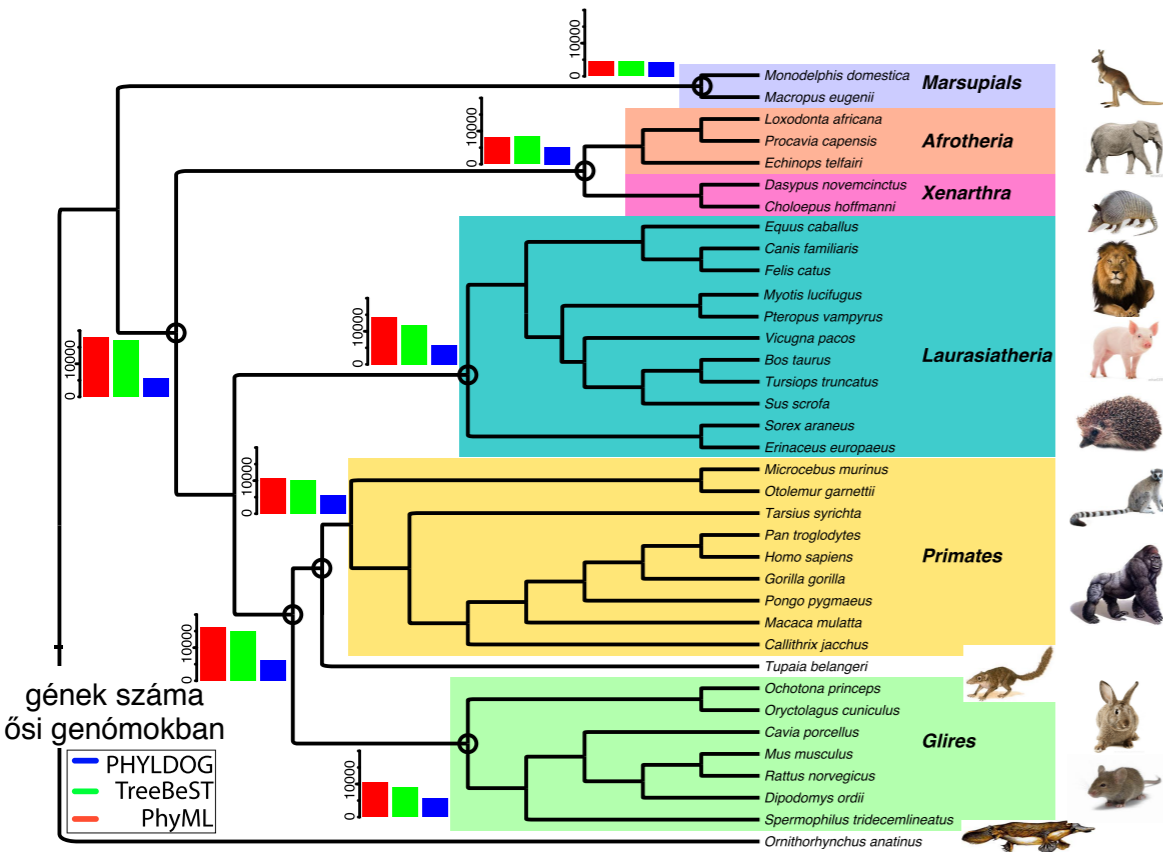


Bastien Boussau

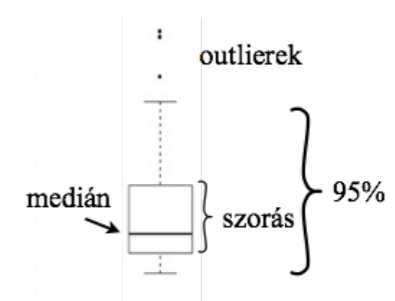
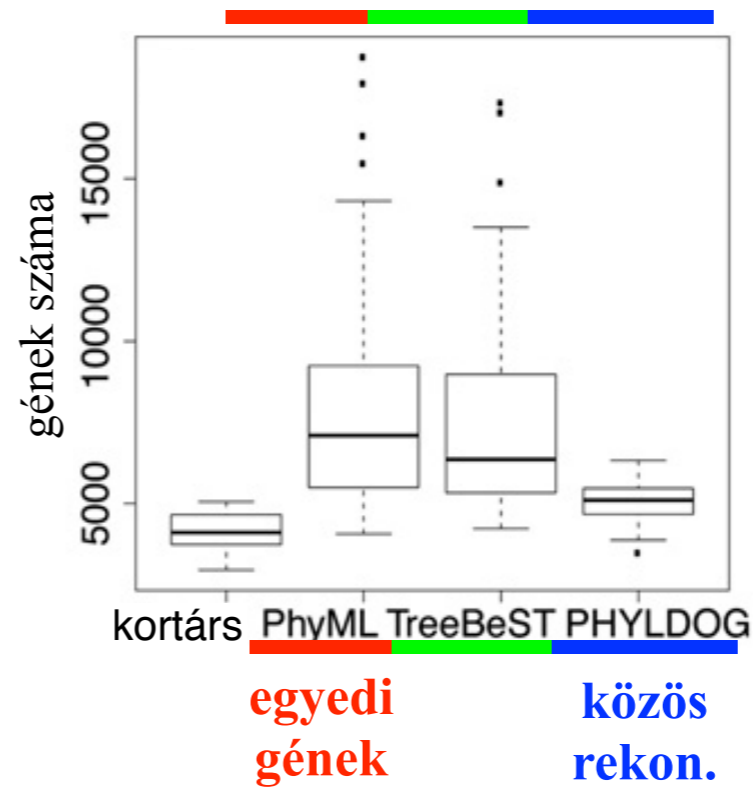
Az emlősök fajfája

6966 gén családot használva 36 emlős genomból, **rekonstruáltuk közösen** 160 millió év evolúcióját lefedő fajfát és az a mentén futó génfákat.

emlősök fajfája



A közös optimalizáció valószínűbb ősi genomméreteket eredményez.



MAP

DL

S

G_j

A_{ij}



Bastien
Boussau

Az emlősök fajfája

6966 gén családot használva 36 emlős genomból, **rekonstruáltuk közösen** 160 millió év evolúcióját lefedő fajfát és az a mentén futó génfákat.

Rekonstruálható az ősi génsorrend is..

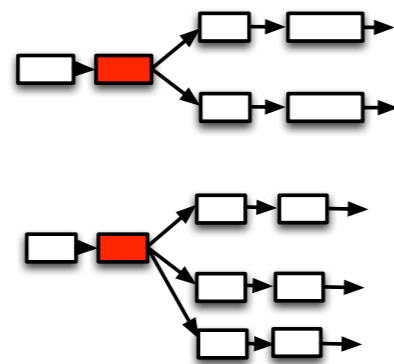
helyes génfa

két szomszéd

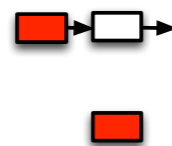


génfahibák

három vagy több szomszéd



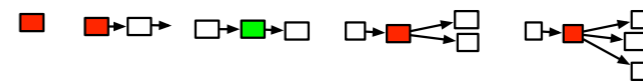
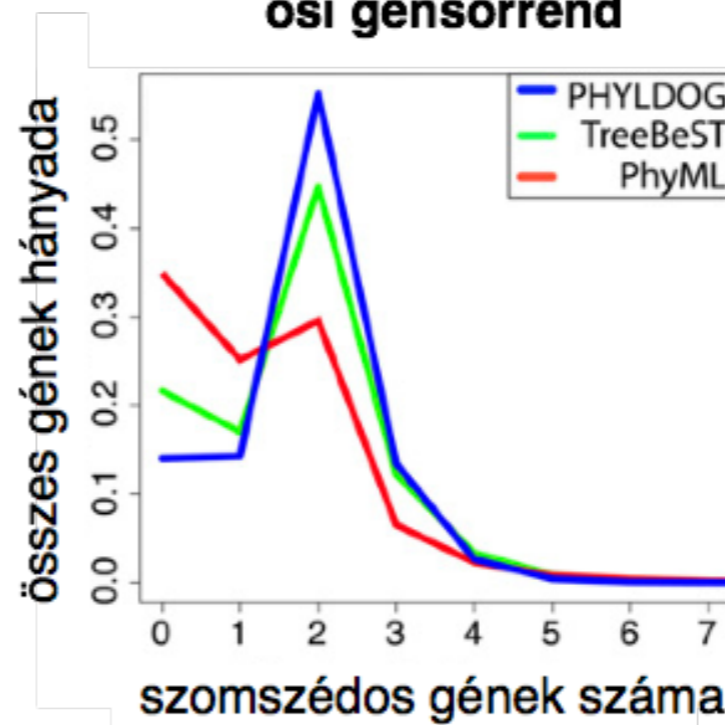
egy vagy nulla szomszéd



Eric
Tannier

A közös optimalizáció valószínűbb ősi génsorrendet eredményez.

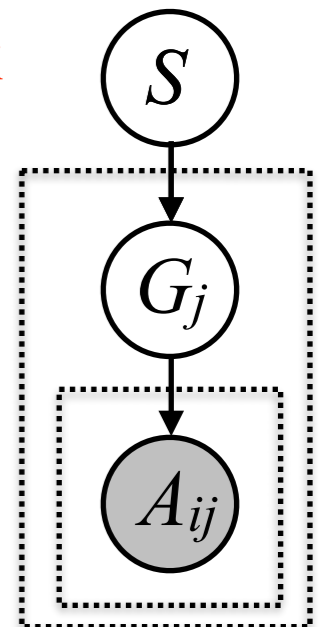
**emlős génfák alapján
rekonstruált
ősi génsorrend**



**közös
rekon.**

**egyedi
gén**

**MAP
DL**

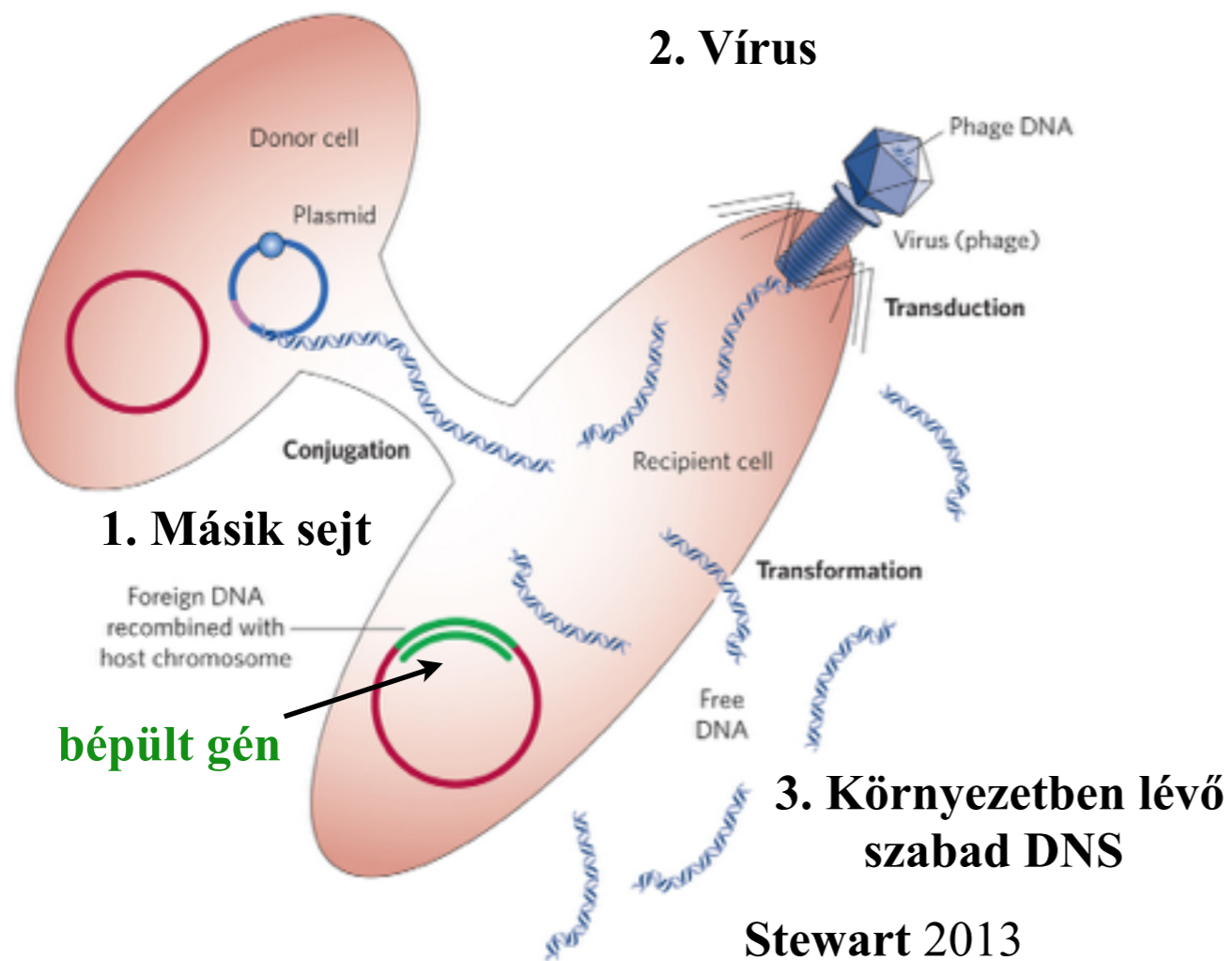


Boussau, Szöllősi, Duret, Gouy, Tannier & Daubin *Genome Res.* (2013)
Genome-scale coestimation of species and gene trees
 Bérard, Gallien, Boussau, Szöllősi, Daubin, Tannier *Bioinformatics* (2013)
Evolution of gene neighborhoods within reconciled phylogenies

Horizontális géntranszfer

A génfák nem mindig követik a fajfát. Egysejtű organizmusoknál gyakori jelenség, hogy kívülről a sejtbe került gének a genomba beépülnek és tovább öröklődnek.

Transzfermechanizmusok

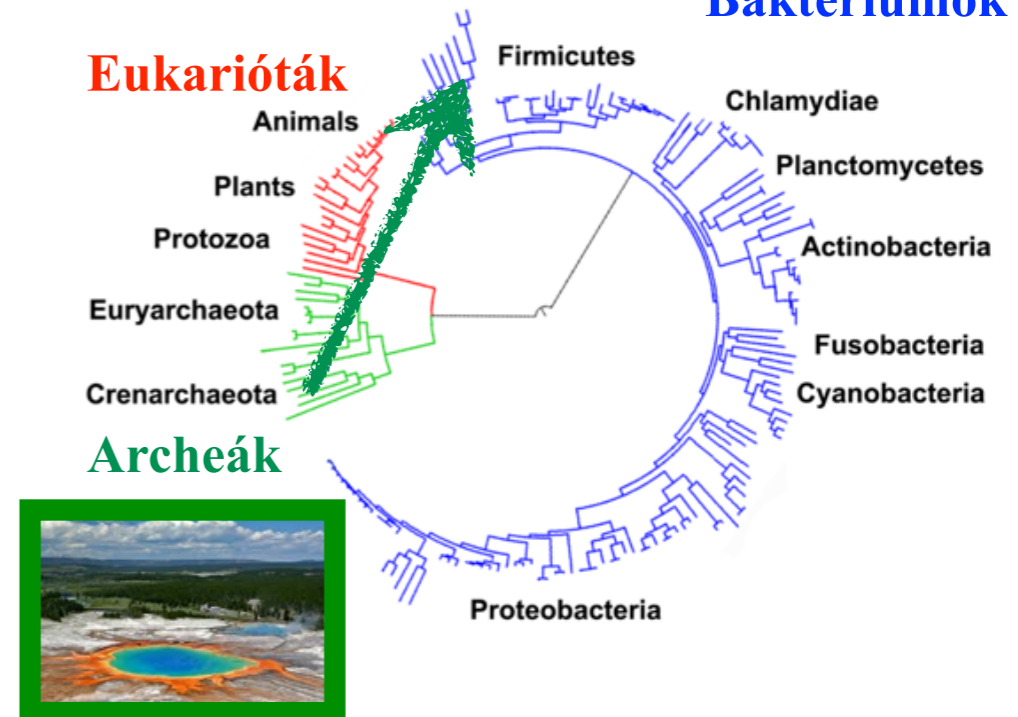


Transzfer példák

antibiotikum rezisztencia
 magas hőmérsékleten működő enzimek



Baktériumok



Horizontális géntranszfer állatokban

Egysejtűek között gyakori a géntranszfer, de többsejtű élőlényeknél, köztük az állatok közt is ismertek példák.

Karotinok

növények és planktonok

baktériumok és archeák

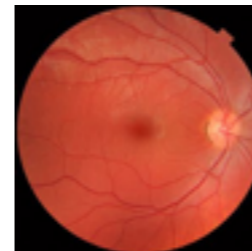
termelik



megeszik

flamingók színe

emberi szem sárgafoltja



de!



borsón élő levéltetvek egy faja is termeli!

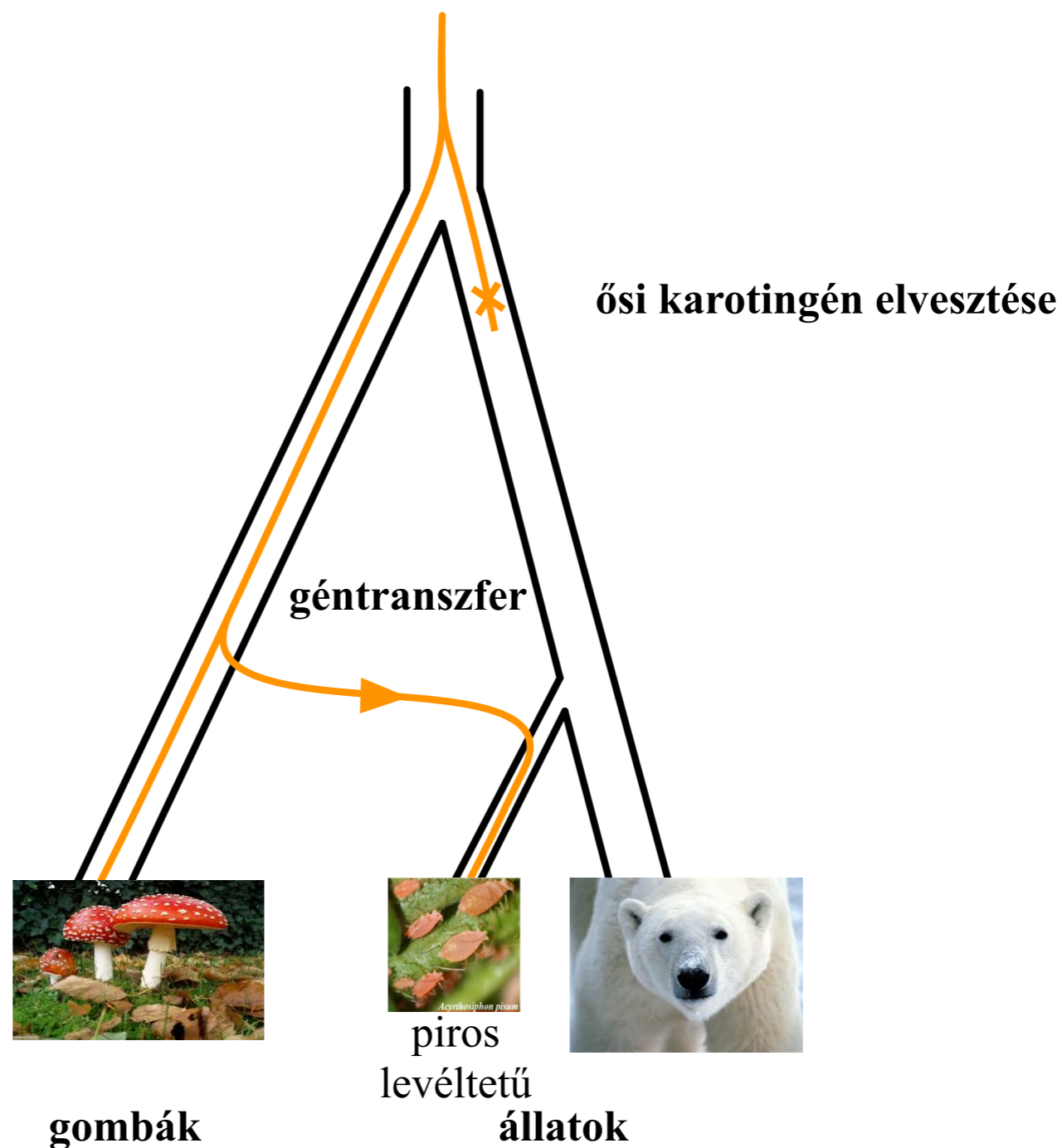
Horizontális géntranszfer állatokban

Egysejtűek között gyakori a géntranszfer, de többsejtű élőlényeknél, köztük az állatok közt is ismertek példák.

borsón élő levéltetű



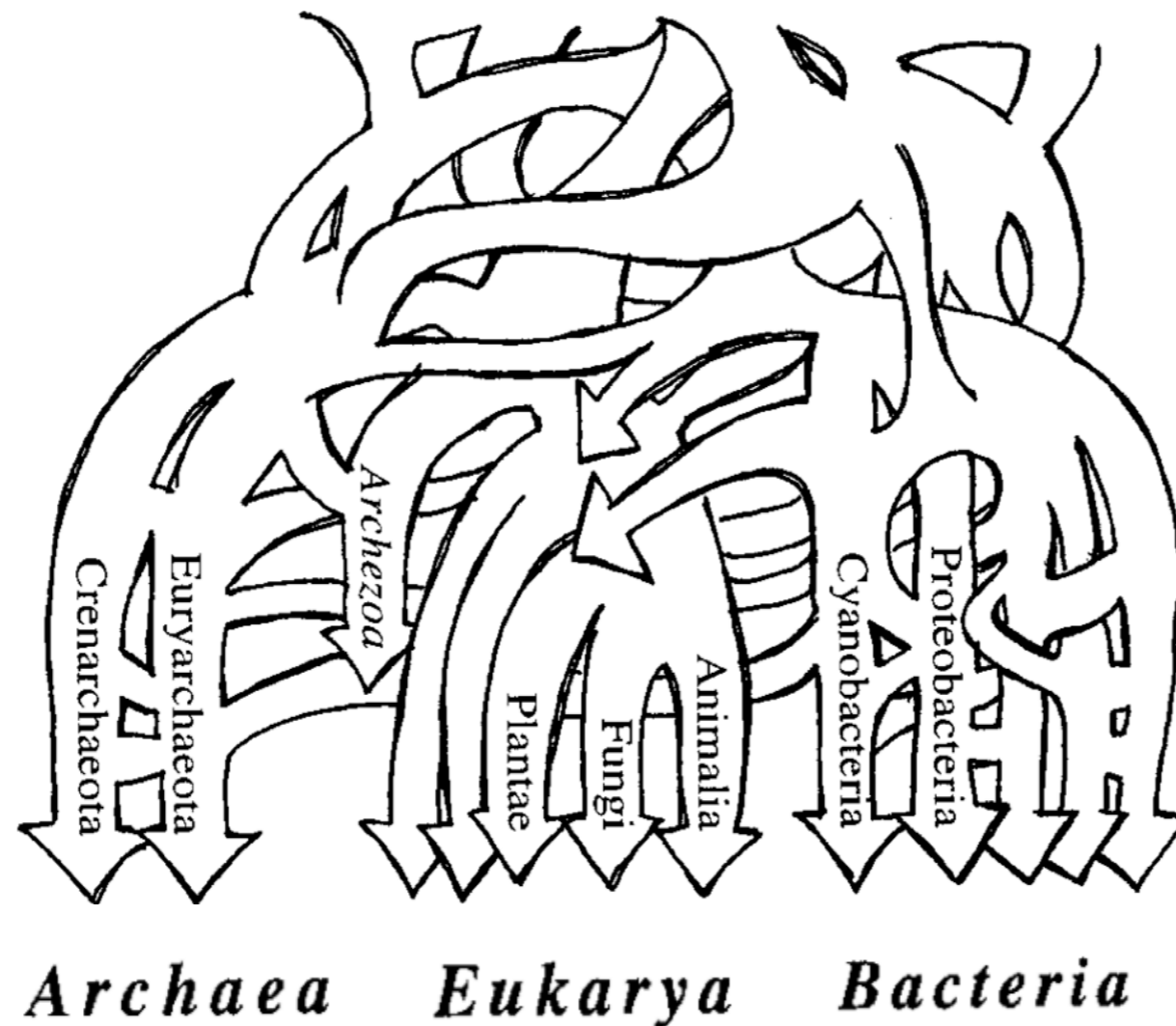
Moran & Jarvik 2010



A horizontális géntranszfer mint zaj

A géntranszfer ellentmondásos géntörténeteket produkál, a karotingének családjában a levéltetű-gén közeli rokona a gombáénak. A transzfer gyakoriságának fényében felmerült, hogy túl sok a zaj a fajfa rekonstrukciójához.

Az élet eredete

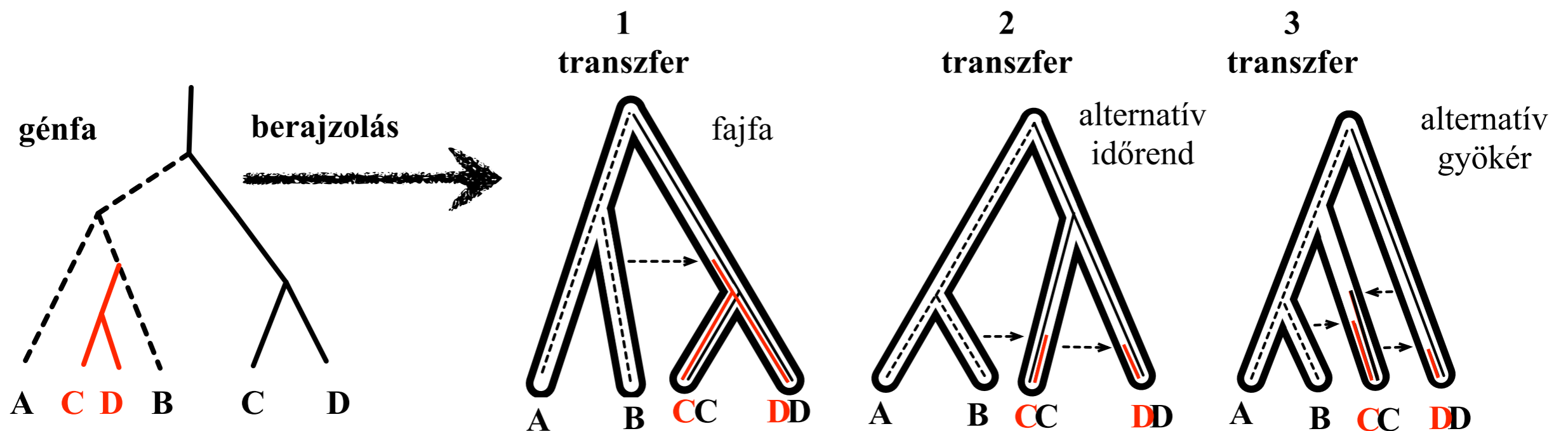


Doolittle 1999

A horizontális géntranszfer mint információ

A génfák topológiájában kódolt transzferek “*molekuláris fossziliák*”, amelyek a fajfa időrendjét rögzítik.

Ez az információ egy transzfereket tartalmazó $p(G_i|S)$ modell segítségével tárható fel.

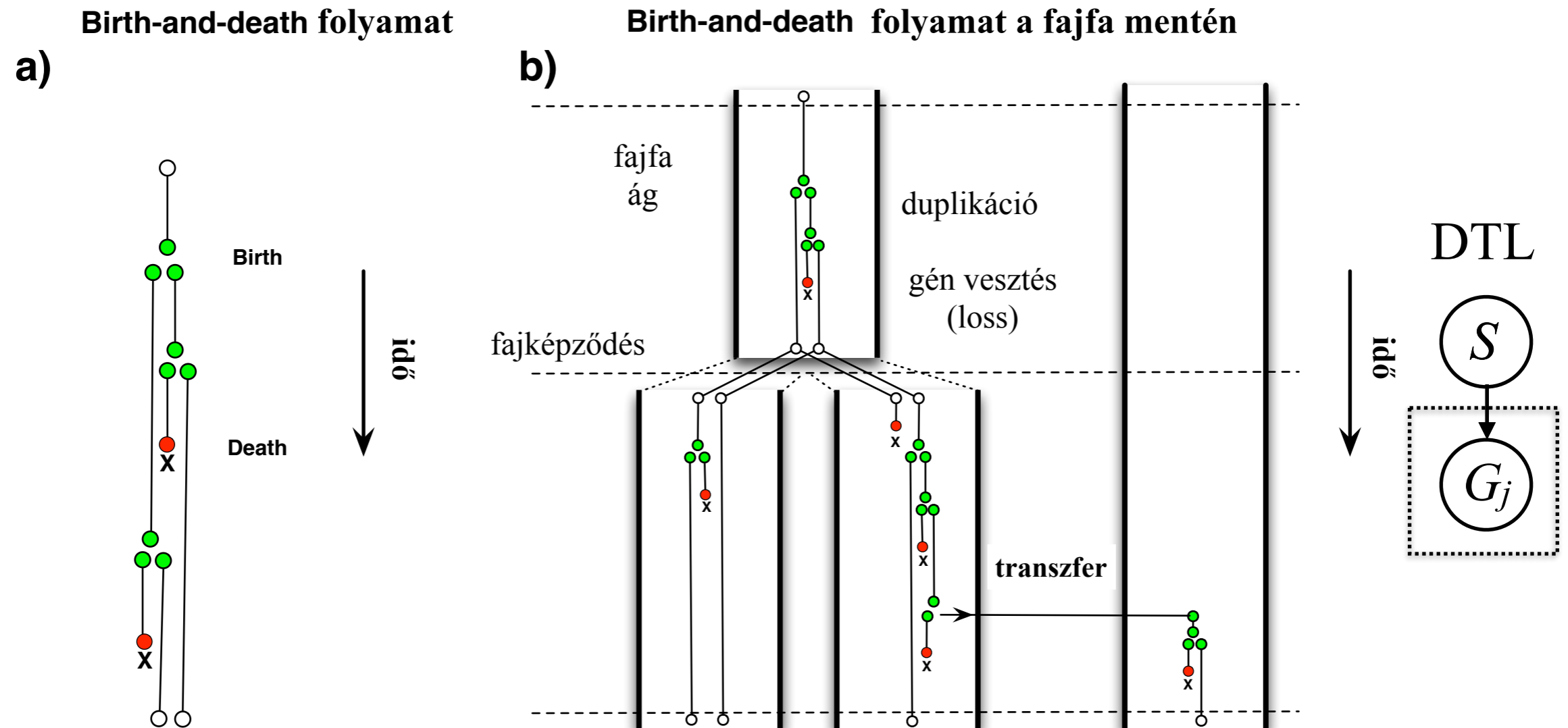


Szöllősi, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
Lateral Gene Transfer from the Dead

Szöllősi, Boussau, Abby, Tannier & Daubin *PNAS* (2012)
Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations

.. de a génfák a fajfa mentén keletkeznek

Ahhoz, hogy kiszámoljuk a *génfa* valószínűségét, *összegeznünk* kell az *összes lehetséges berajzolásán a génfának a fajfába*.



Tofigh PhD thesis 2009

Szöllősi, Boussau, Abby, Tannier & Daubin *PNAS* (2012)
Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations

A fotoszintetizáló baktériumok fájája

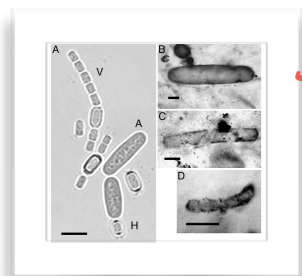
8332 gén családot használva 36 cianobaktérium genomból, rekonstruáltuk 3 milliárd év evolúcióját lefedő fajtát és a fajelágazások időbeli sorrendjét.

párhuzamos számítási eljárás

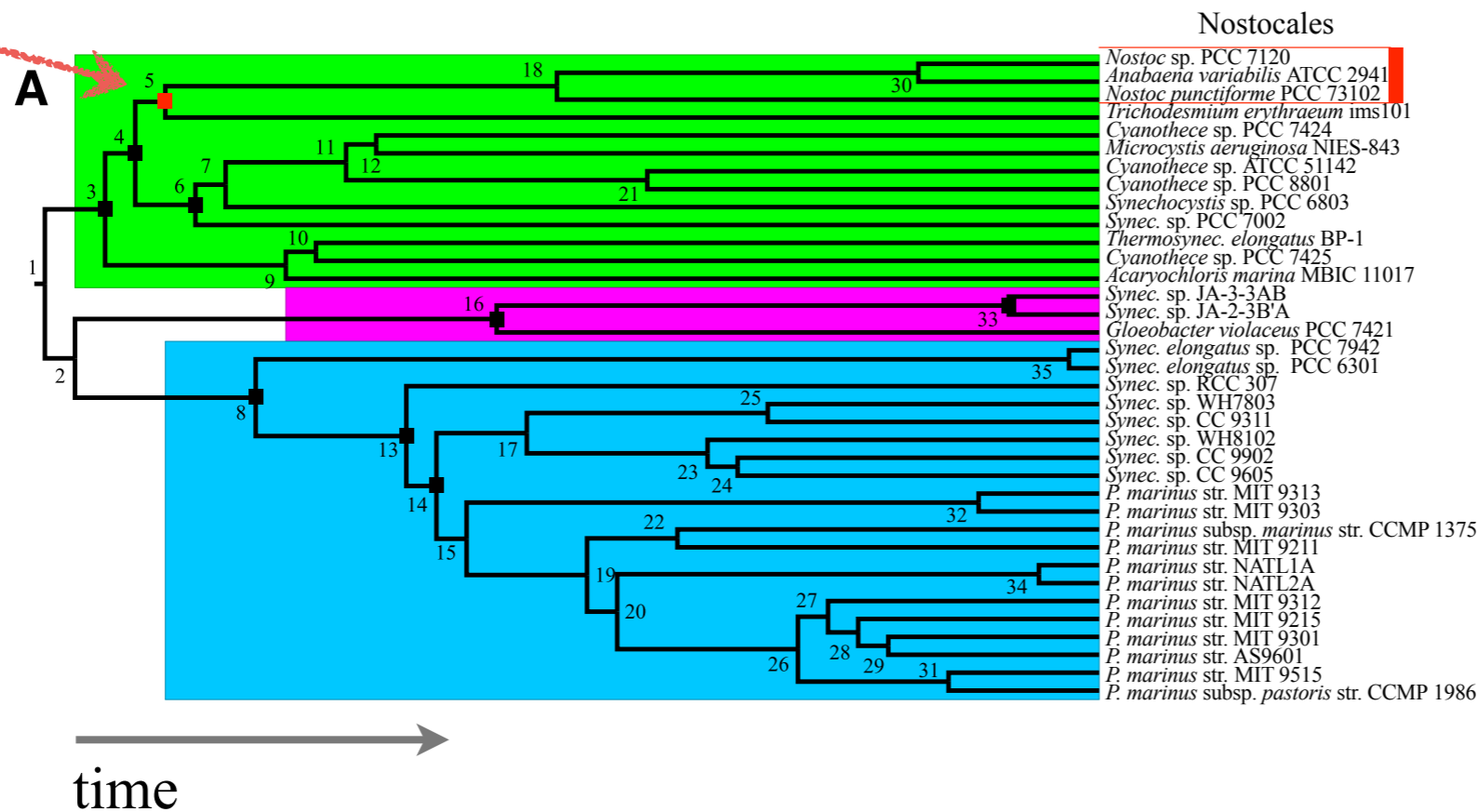
a *maximum likelihood* fajfa keresésére

$$\mathcal{L}(S, \text{rates} | \{G_j\}) \propto \prod_j p(G_j | S, \text{rates})$$

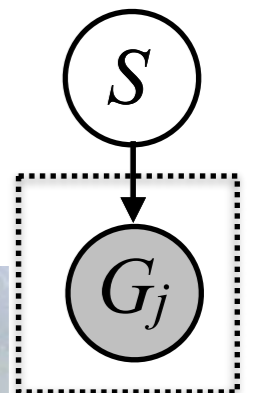
$2.450 - 2.320 \times 10^9$ év



mikrofossziliák
Tomitani *et al.* 2006



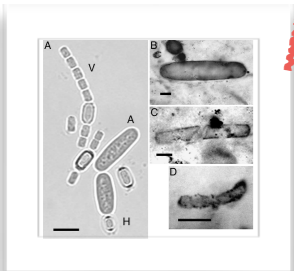
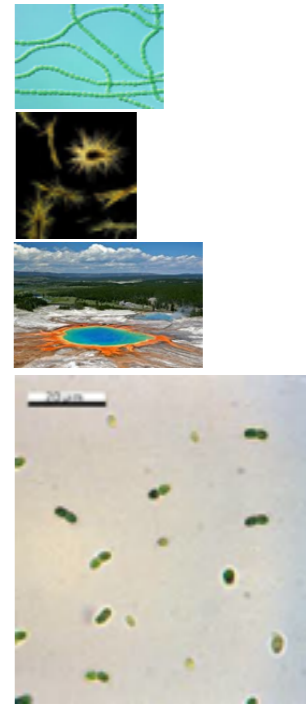
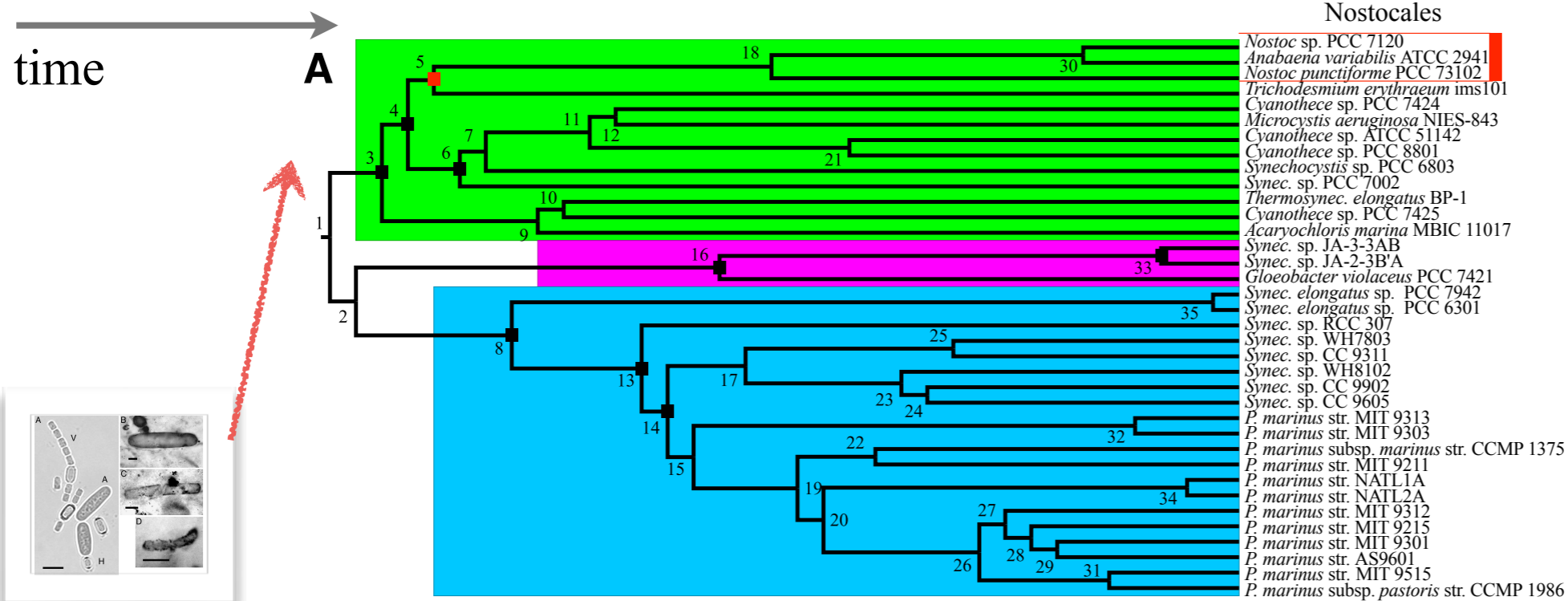
ML
DTL



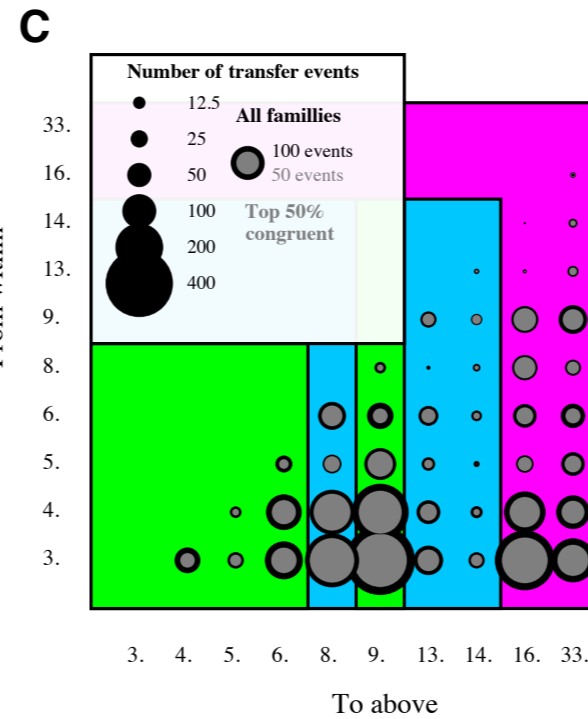
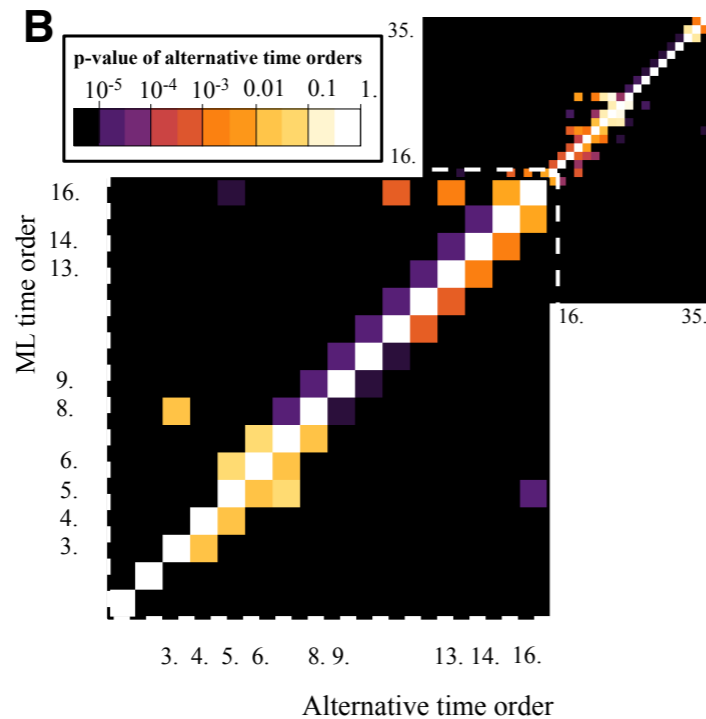
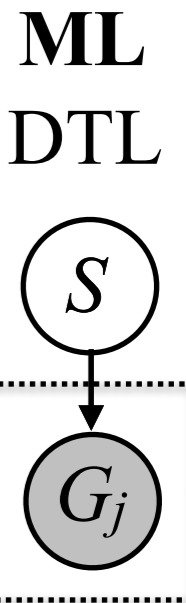
Szöllősi, Boussau, Abby, Tannier & Daubin *PNAS* (2012)
Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations

A fotoszintetizáló baktériumok fájája

8332 gén családot használva 36 cianobaktérium genomból, rekonstruáltuk 3 milliárd év evolúcióját lefedő fáját és a fajelágazások időbeli sorrendjét.

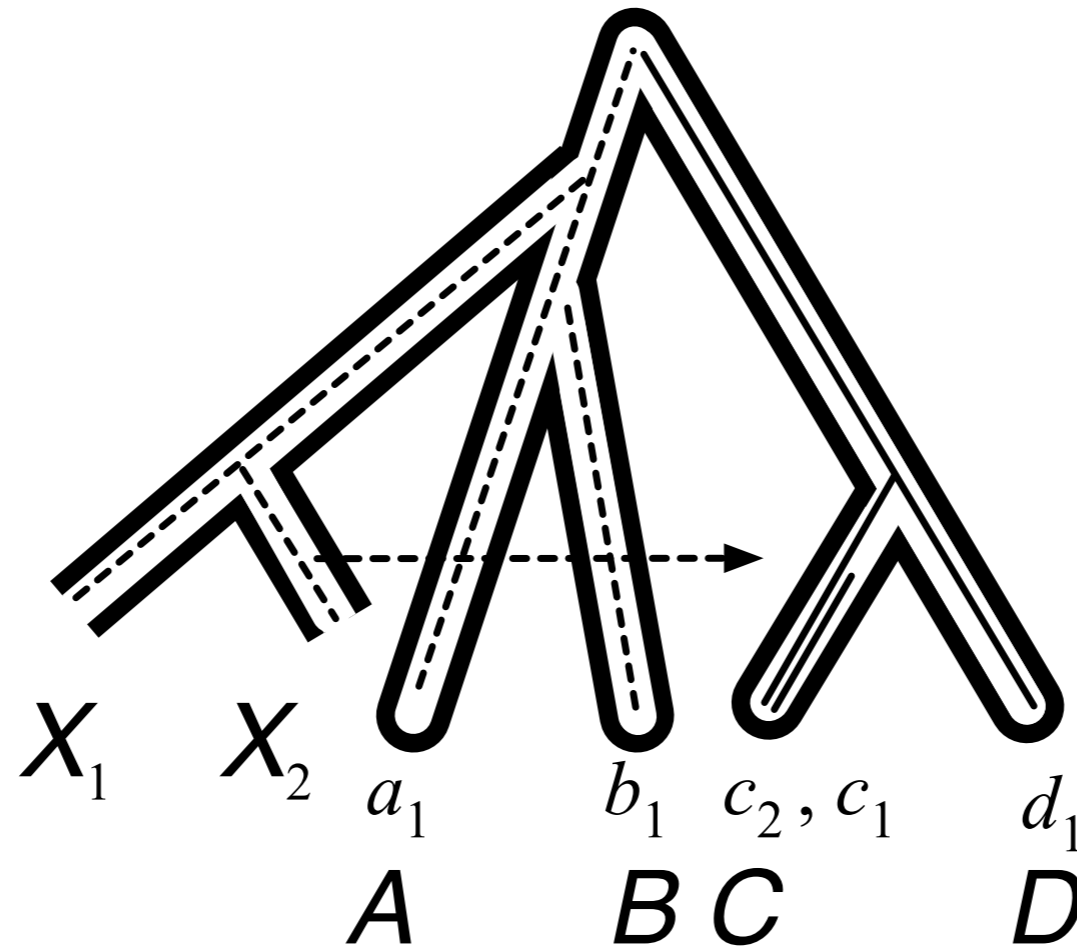


microfossils
Tomitani *et al.* 2006



Géntranszfer a halottaktól.

.. elhanyagoltuk az a lehetőséget, hogy az a faj, ahonnan a transzfer indult, kihalt vagy nem szerepel a mintákban.

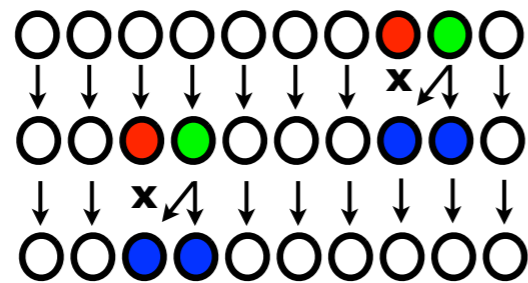


Nicolas
Lartillot

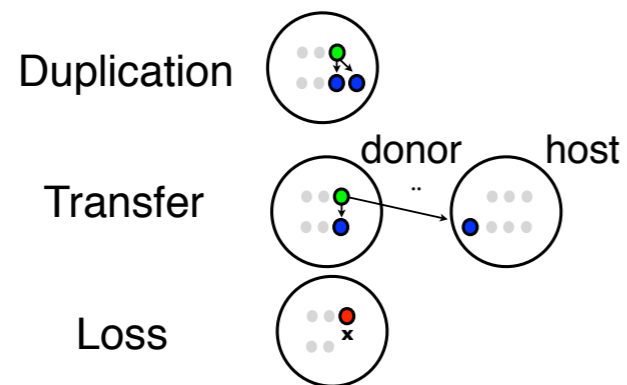
Géntranszfer a halottaktól.

Ahhoz, hogy a kihalt és nem megfigyelt fajokat figyelembe tudjuk venni, modelleznünk kell a fajok keletkezését és kihalását.

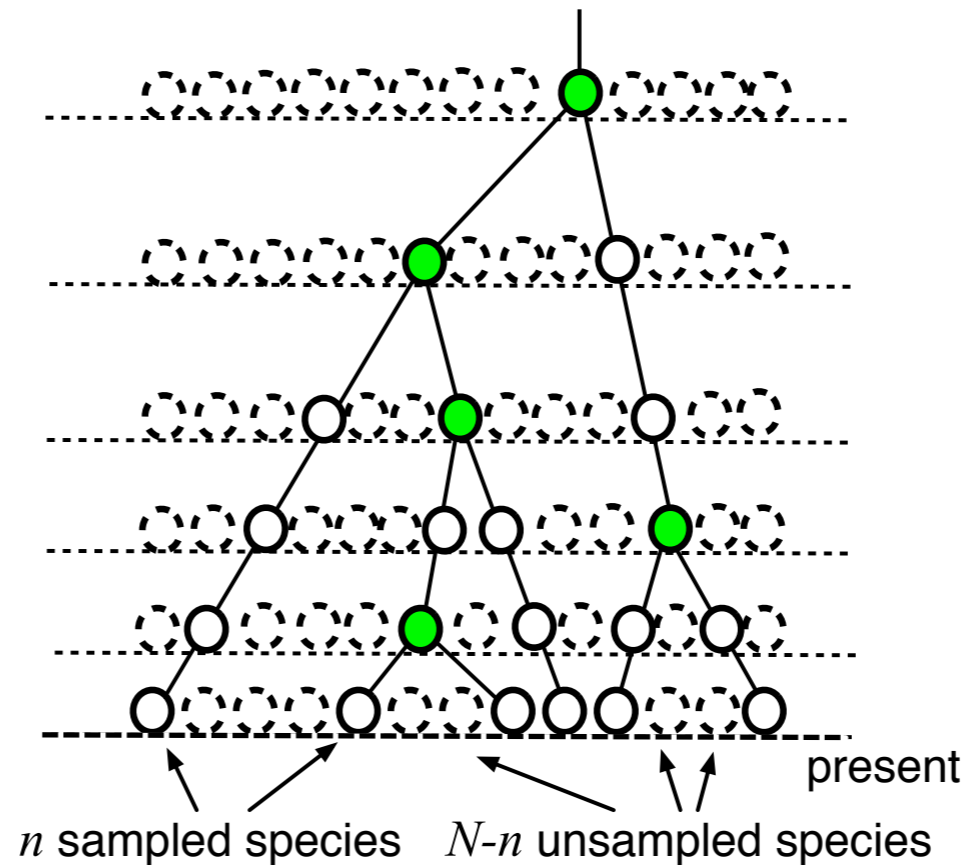
a) speciation dynamics



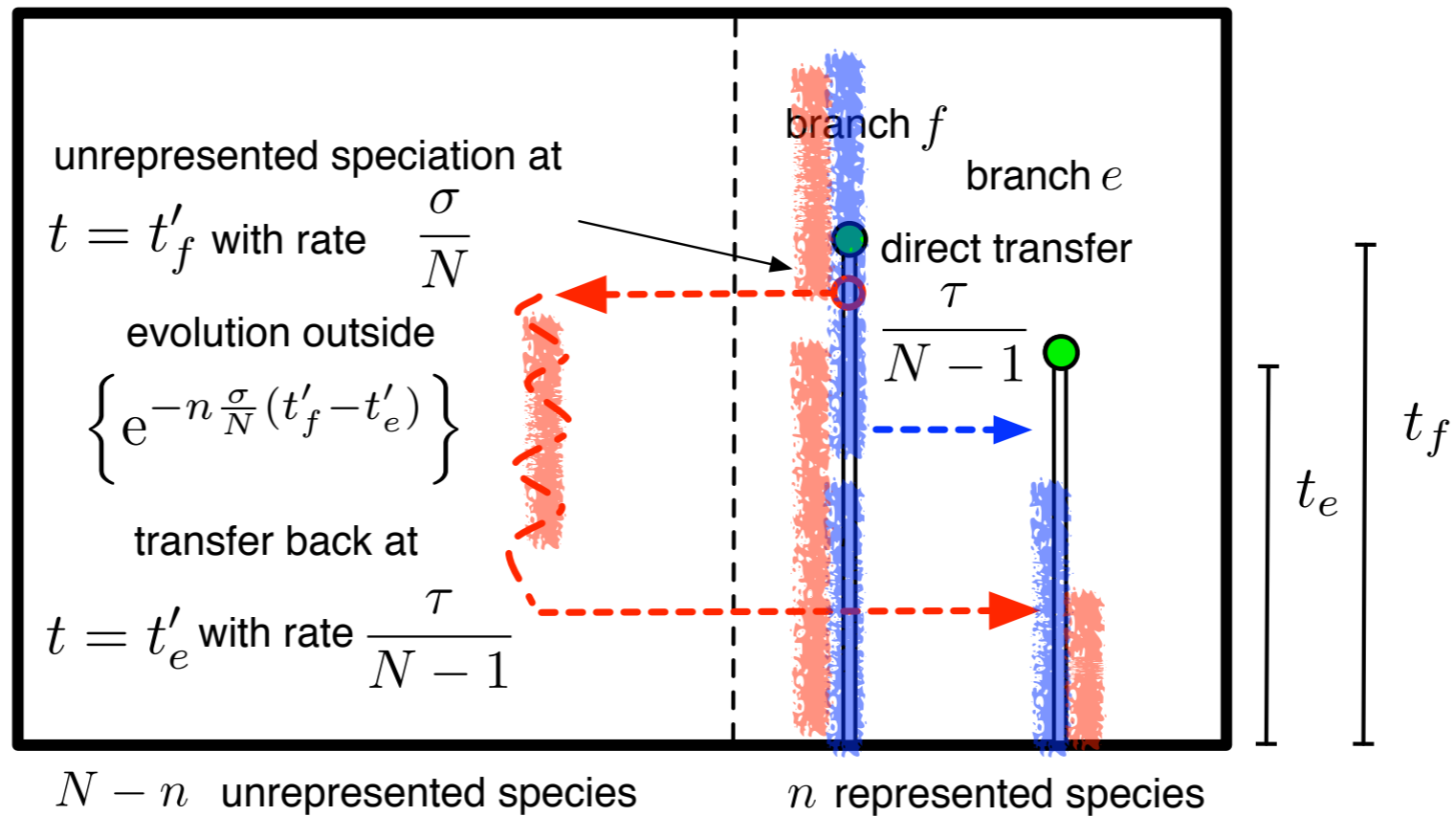
b) gene birth and death



c) represented species phylogeny



..(majdnem) minden transzfer a halottaktól jön..



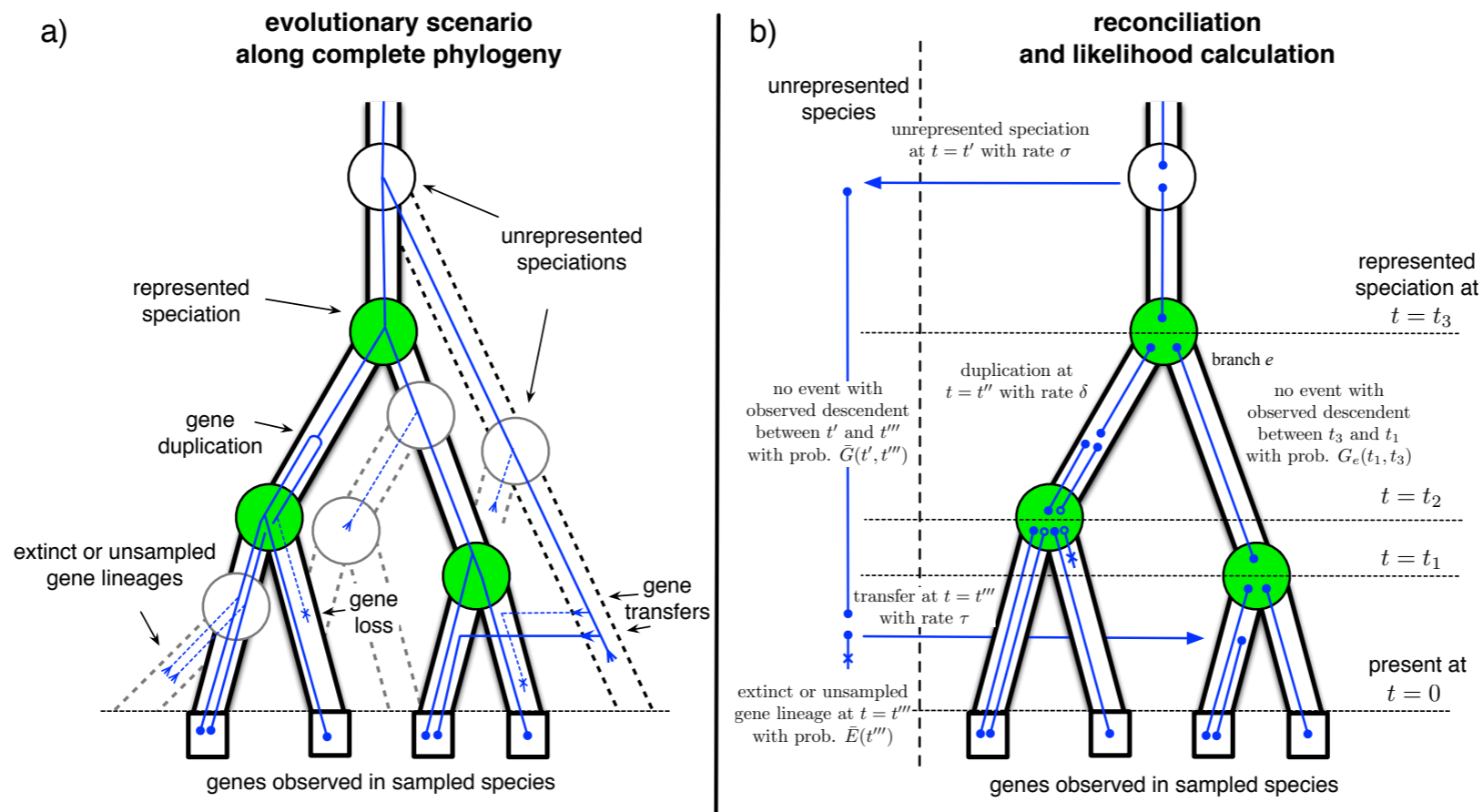
$$T_{\text{direct}} \approx \int_0^{\frac{N}{n\sigma}} \frac{\tau}{N} dt'_e = \frac{1}{N} \frac{\tau}{2n} \left[\frac{2N}{\sigma} \right],$$

$$T_{\text{indirect}} \approx \int_0^{\frac{N}{n\sigma}} \int_{t'_e}^{\frac{N}{n\sigma}} \tau \left\{ e^{-n \frac{\sigma}{N} (t'_f - t'_e)} \right\} \frac{\sigma}{N} dt'_f dt'_e = \frac{1}{en} \frac{\tau}{2n} \left[\frac{2N}{\sigma} \right],$$

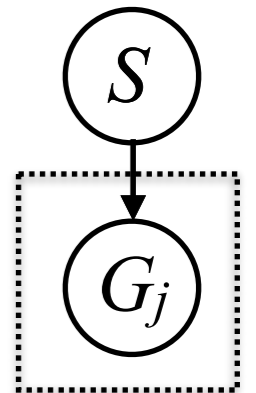
..(majdnem) minden transzfer a halottaktól jön..

473 közel univerzális géncsaládot vizsgálva 36 cyanobaktériumban, a transzferek 28%-a mutat topológiai explicit bizonyítékot a kihalt fajból való transzferre, de (csak) 6%-uk jött a cyanobaktériumokon kívülről.

egy mean-field-szerű eljárással ki tudjuk számolni $p(G_i|S)$ -t ha $n \ll N$



DTL+ex



Szöllősi, Tannier, Lartillot & Daubin *Systematic Biology* (2013)

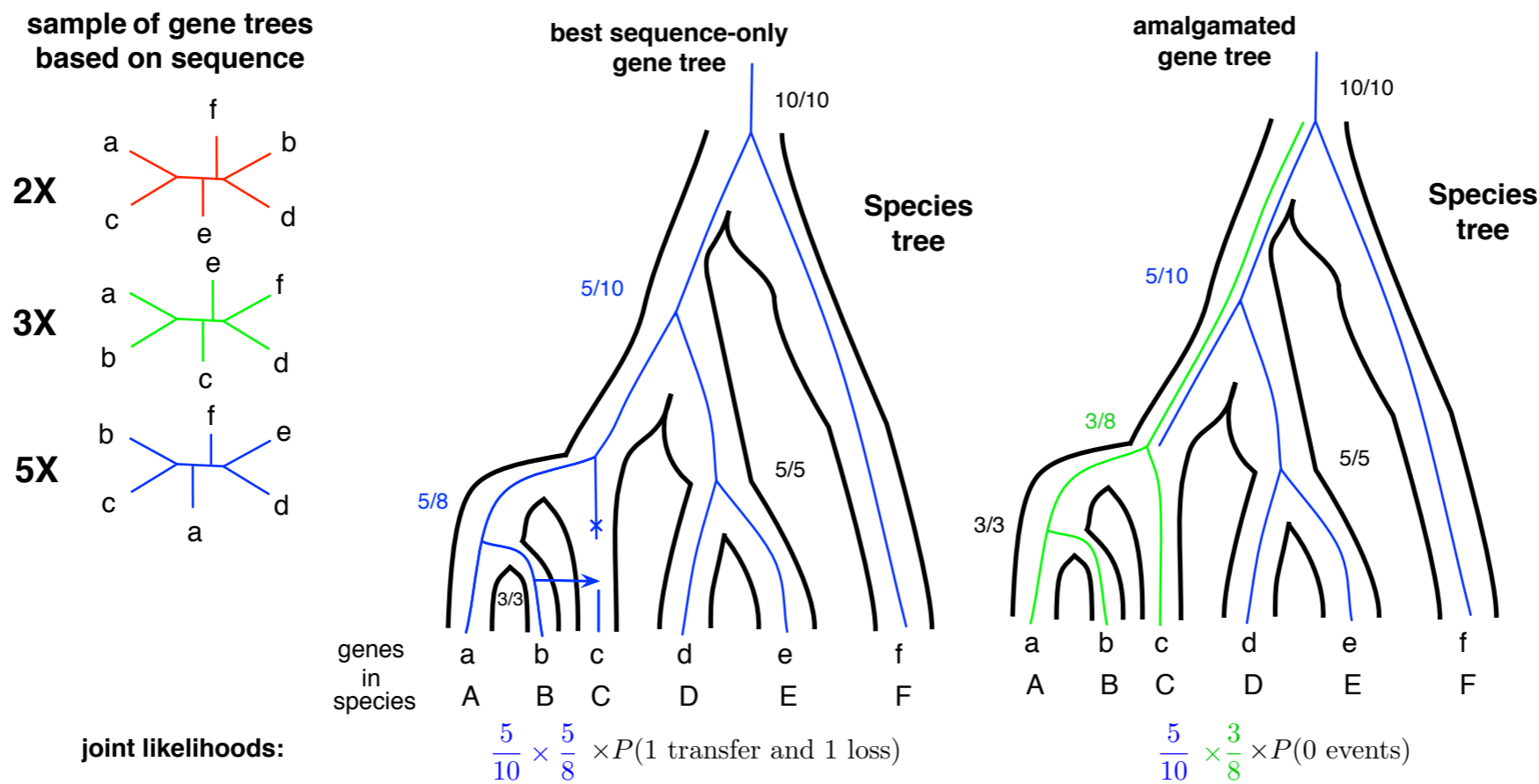
Lateral Gene Transfer from the Dead

Szöllősi, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)

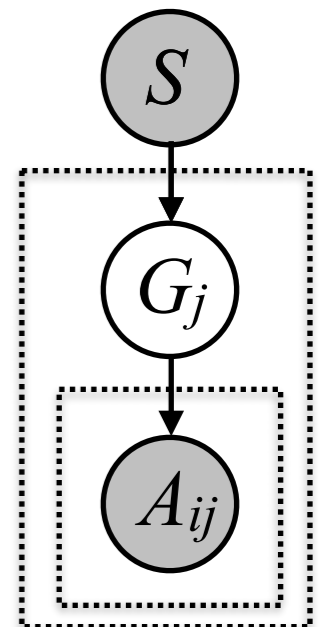
Efficient exploration of the space of reconciled gene trees

A lehetséges génfák terének hatékony bejárása

Egy feltételes függetlenséget feltételező közelítéssel kis méretű minta alapján (pl. 10^4) nagyon sok (tipikusan 10^{12} , de akár 10^{40}) génfára jól közelíthető a szekvencia valószínűsége. *Ez lehetővé teszi a közös likelihood szerint optimális génfa hatékony keresését és a génfák feletti összegzésének közelítését.*



DTL+ex



ALE nyílt forrású implementáció:

<http://github.com/ssolo/ALE>

Szöllősi, Tannier, Lartillot & Daubin *Systematic Biology* (2013)

Lateral Gene Transfer from the Dead

Szöllősi, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)

Efficient exploration of the space of reconciled gene trees

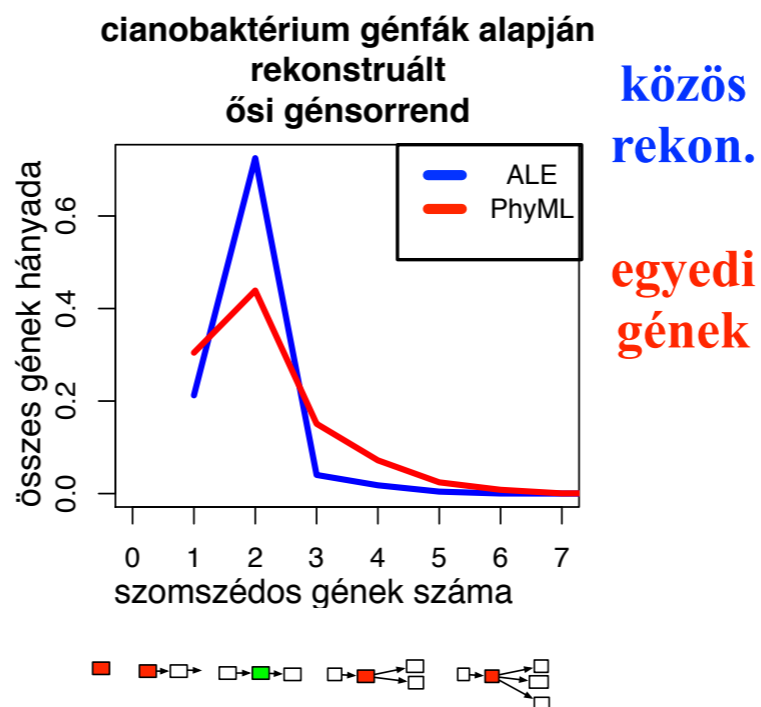
Szöllősi, Tannier, Daubin & Boussau *Systematic Biology* (2014)

The inference of gene trees with species trees (to appear)

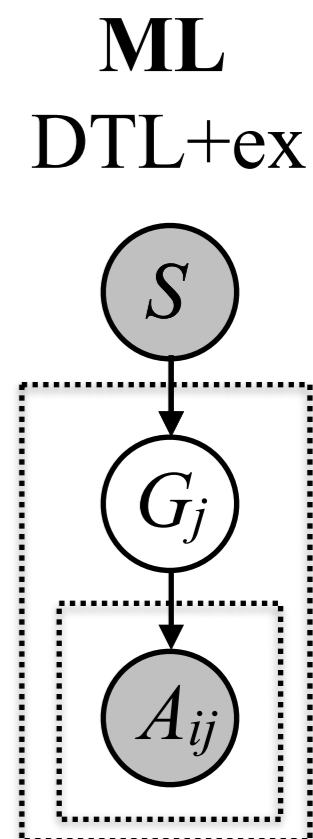
Génfák rekonstrukciója ismert fajfa esetén

1099 géncsaládot használva 36 cyanobaktériumból rekonstruáltunk génfákat a fajfát ismertnek feltételezve.

Pontosabb génfák..



..azt mutatják, hogy a vélt transzferek 60-80%-a a génfák elmosódottságának eredménye. De így is marad 1000-1500 transzfer a fotoszintetizáló baktériumoknál ..



Szöllősi, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
Lateral Gene Transfer from the Dead

Szöllősi, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)
Efficient exploration of the space of reconciled gene trees

A lehetséges génfák terének hatékony bejárása

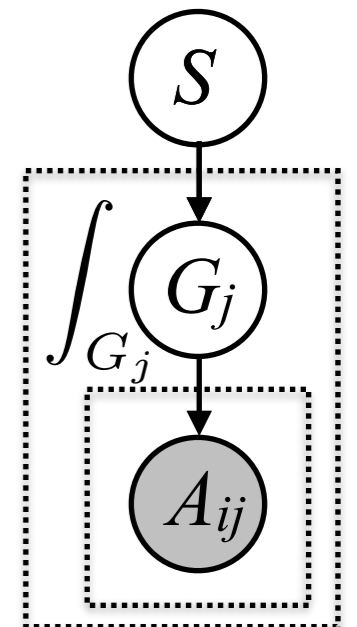
Az ALE módszerrel közelíteni tudjuk az integrált az ún. Felsenstein egyenletben:

$$P(A|S, \text{ráták}) = \int_G p(A|G)p(G|S, \text{ráták})$$

Felsenstein 1988

Ez lehetővé teszi a hatékony MCMC bejárását $\mathcal{L}(S, \text{rates}|A)$ -nek!

DTL+ex



ALE nyílt forrású implementáció:

<http://github.com/ssolo/ALE>

Szöllősi, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)
Efficient exploration of the space of reconciled gene trees

Szöllősi, Tannier, Daubin & Boussau *Systematic Biology* (2014)
The inference of gene trees with species trees (to appear)

Az élet fájának rekonstrukciója teljes genomokból

Ez folyamatban lévő munka.. posztdokot keresünk a lyoni **LBBE-n!**

párhuzamos számítási eljárás

szerver:

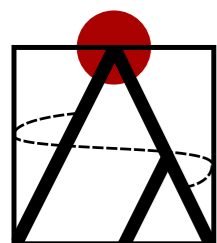
$$\prod_j$$

S és ráták optimalizálása

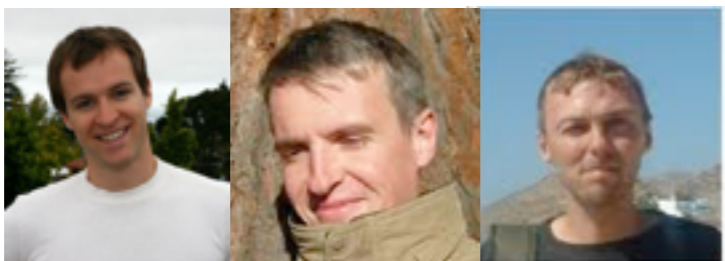
kliensek:

integrál G_j felett ALE segítségével

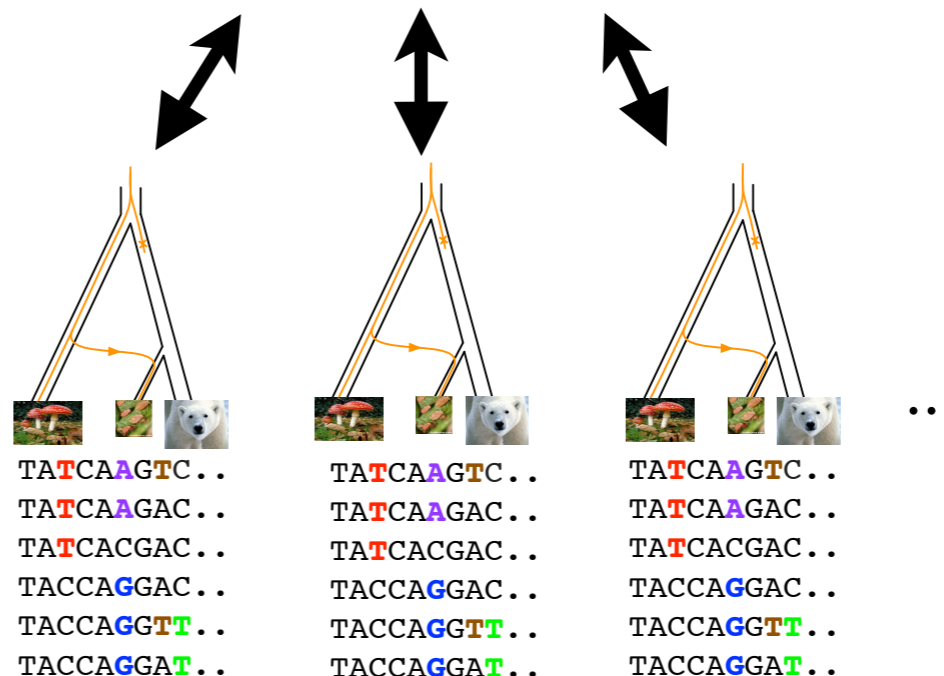
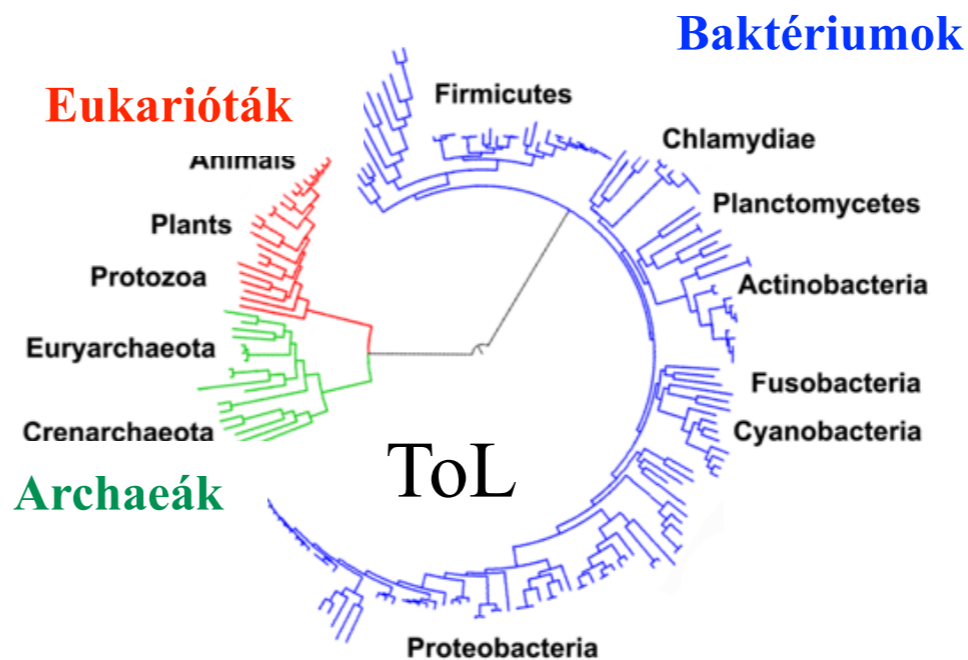
$$\prod_i p(A_i|G_j) \times p(G_j|S)$$



AGENCE NATIONALE DE LA RECHERCHE
ANR

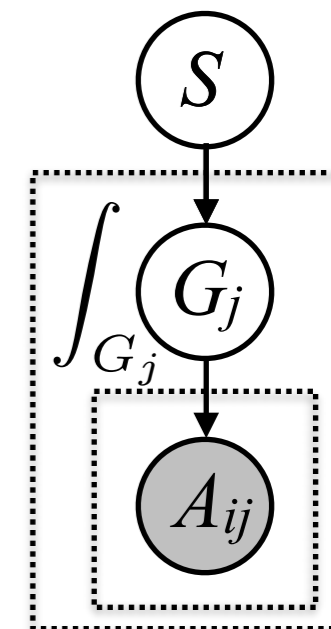


Bastien Boussau Vincent Daubin Eric Tannier



complete genomes

DTL+ex

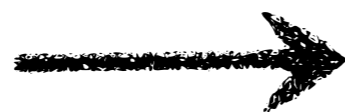


Ősi élőlények genetikai tervrajzának rekonstrukciója

A közös rekonstrukciója az eddiginél pontosabb képet ad az ősi gén tartalmáról, de minket igazából az ősi élőlények és környezetük tulajdonságai érdekelnek..

Az ELTE-n pedig doktoranduszt!

Ősi gének listája

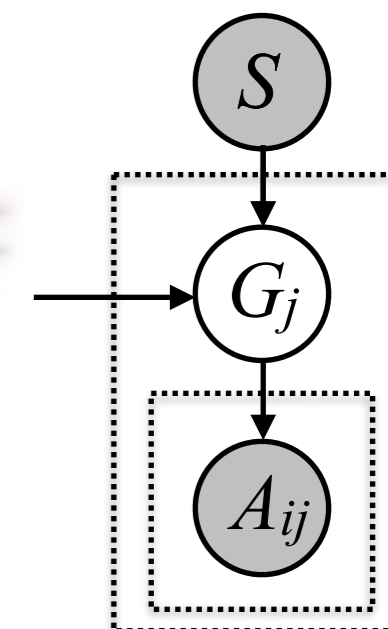


Ősi tulajdonságok,
környezet, rendszerek

metabolikus hálózatok,
molekuláris gépek,
köölcsönhatások..



DTL+ex



Derényi Imre

ELTE-MTA „Lendület” Biofizika Kutatócsoport

