Comparació de mètodes de reconstrucció filogenètica basats en quartets

Felipe Cano, Marc Felipe i Marta Casanellas*
Universitat Politècnica de Catalunya

En aquest treball estudiem i comparem dos mètodes de reconstrucció filogenètica basats en quartets. A la introducció, repassem què és un arbre filogenètic i algunes tècniques que ens ajuden a inferir-lo a partir del codi genètic d'espècies actuals. Explorem els mètodes basats en quartets i, més concretament, els programes de reconstrucció filogenètica `SplitSup` i `Erik+2`, que són els objectes d'estudi d'aquest treball. A continuació procedim a avaluar la robustesa dels mètodes utilitzant simulacions sota diferents ajustos dels paràmetres en el model evolutiu. Finalment, utilitzem unes seqüències d'ADN reals d'unes espècies de llevat per a veure si els dos mètodes són capaços d'inferir-ne l'arbre filogenètic.

I. INTRODUCCIÓ: FONAMENT TEÒRIC

La filogènia consisteix en l'estudi de la relació evolutiva entre les espècies. Sovint aquesta relació s'expressa en forma d'arbre, on es veu un ancestre comú, que correspon a l'arrel de l'arbre, des d'on surten branques, de més o menys llargària, que es van ramificant i eventualment acaben en les fulles de l'arbre, que representen les espècies actuals. La longitud de cada aresta de l'arbre és una magnitud que representa l'abundància de les mutacions que ha sofert el codi genètic durant l'especiació.

Per a trobar l'arbre filogenètic que ha donat lloc a un grapat d'espècies donades, cal primer identificar les mutacions al codi genètic. És a dir, cal trobar un fragment del codi genètic de cada espècie de manera que siguin iguals entre si, excepte perquè uns quants nucleòtids han mutat. Un cop han estat identificats aquests fragments, es posen un sota l'altre de manera que coincideixin en la major part. Aquest procés s'anomena trobar un alineament.

Podem modelar l'evolució de les espècies mitjançant cadenes de Markov, els coeficients de la qual són paràmetres desconeguts (veure [1]). Les úniques dades conegudes en aquest model són les probabilitats conjuntes $p_{N_1,N_2,\ldots,N_k}$ d'observar, en la mateixa posició, el nucleòtid N_1 a la primera espècie, el nucleòtid N_2 a la segona espècie, etc. que podem estimar segons la freqüència relativa.

La quantitat de paràmetres desconeguts és molt elevada en aquest model, i per aquest motiu es busquen el que s'anomenen invariants filogenètics, que consisteixen en relacions polinòmiques que han de complir les probabilitats d'observació independentment dels paràmetres si provenen de cert arbre filogenètic. En altres paraules, per a cada arbre filogenètic existeixen unes relacions que aquestes probabilitats han de complir, i que no compleixen si provenen d'un altre arbre filogenètic. Conèixer els invariants filogenètics per a cada topologia és doncs crucial per a identificar/descartar un cert candidat com a l'arbre filogenètic correcte.

Tot i haver modelat el procés evolutiu utilitzant un ancestre comú, sovint els arbres filogenètics es mostren sense arrel. Això és degut al fet que només amb les seqüències actuals d'ADN no es pot determinar el sentit de les mutacions i, per tant, no es pot identificar l'arrel de l'arbre. És per aquest motiu que treballarem amb arbres sense arrel.

Nosaltres avaluem dos mètodes de reconstrucció filogenètica basats en quartets. Aquests són els mètodes que reconstrueixen l'arbre filogenètic a partir de la posició relativa dels subconjunts de quatre fulles dins l'arbre: donat que un arbre de quatre fulles se'l pot dotar de tres topologies d'arbre etiquetat diferents (veure Figura 1), saber quina és la correcta dóna informació sobre l'arbre. Repetint aquest procés per a cada subconjunt de quatre fulles és possible reconstruir completament l'arbre filogenètic.

Per a esbrinar quina és la topologia correcta, ambdós programes es basen en el següent teorema (veure [2]):

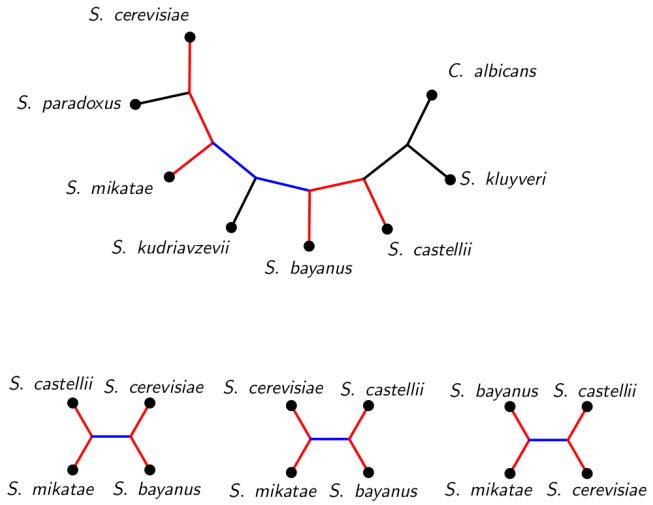

Figura 1: Un arbre filogenètic relacionant 8 espècies de llevat. Es mostra també les tres topologies possibles en l'arbre que té per fulles *S. castellii*, *S. mikatae*, *S. bayanus* i *S. cerevisiae*. La topologia induïda per l'arbre filogenètic és la del mig:
(*S. cerevisiae*, *S. mikatae* | *S. bayanus*, *S. castellii*).

*Com a tutora d'aquest treball.

Si la topologia (12|34) és la induïda per un arbre filogenètic, llavors la matriu de *flattening* construïda de la següent manera:

$$M_{(12|34)} = \begin{array}{c} \begin{matrix} **AA & **AC & **AG & \cdots & **TT \end{matrix} \\ \begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & \cdots & p_{AATT} \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & \cdots & p_{ACTT} \\ p_{AGAA} & p_{AGAC} & p_{AGAG} & \cdots & p_{AGTT} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{TTAA} & p_{TTAC} & p_{TTAG} & \cdots & p_{TTTT} \end{pmatrix} \end{array} \begin{matrix} AA** \\ AC** \\ AG** \\ \vdots \\ TT** \end{matrix}$$

té, en general, rang 4, mentre que les matrius de *flattening* corresponents a les altres topologies:

$$M_{(13|24)} = \begin{array}{c} \begin{matrix} *A*A & *A*C & *A*G & \cdots & *T*T \end{matrix} \\ \begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & \cdots & p_{ATAT} \\ p_{AACA} & p_{AACC} & p_{AACG} & \cdots & p_{ATCT} \\ p_{AAGA} & p_{AAGC} & p_{AAGG} & \cdots & p_{ATGT} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{TATA} & p_{TATC} & p_{TATG} & \cdots & p_{TTTT} \end{pmatrix} \end{array} \begin{matrix} A*A* \\ A*C* \\ A*G* \\ \vdots \\ T*T* \end{matrix}$$

$$M_{(14|23)} = \begin{array}{c} \begin{matrix} *AA* & *AC* & *AG* & \cdots & *TT* \end{matrix} \\ \begin{pmatrix} p_{AAAA} & p_{AACA} & p_{AAGA} & \cdots & p_{ATTA} \\ p_{AAAC} & p_{AACC} & p_{AAGC} & \cdots & p_{ATTC} \\ p_{AAAG} & p_{AACG} & p_{AAGG} & \cdots & p_{ATTG} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{TAAT} & p_{TACT} & p_{TAGT} & \cdots & p_{TTTT} \end{pmatrix} \end{array} \begin{matrix} A**A \\ A**C \\ A**G \\ \vdots \\ T**T \end{matrix}$$

tenen en general rang màxim (16). És a dir, els menors 5×5 de la matriu $M_{(12|34)}$ són invariants filogenètics per a la topologia (12|34).

Gràcies a aquest teorema, només és necessari estudiar el rang de les matrius de *flattening* per a les tres topologies candidates, estimant la probabilitat com la freqüència relativa[1]. La que tingui rang 4 correspondrà a la topologia correcta, mentre que les altres dues topologies donaran lloc a matrius de rang 16. Ara bé, la propietat de tenir rang 4 no és gens robusta, i petites perturbacions en les entrades de la matriu la poden fer passar a rang 16, fent que la topologia correcta esdevingui indistingible de les altres. Per evitar això, en lloc de calcular directament el rang, el que es fa és estudiar la distància de les matriu d'estudi al conjunt de les matrius de rang ≤ 4. Aquí és on els dos mètodes que comparem, `SplitSup`[4] i `Erik+2`[3], difereixen.

El programa `SplitSup` realitza el càlcul esmentat utilitzant la distància de Frobenius, i normalitza el resultat per la norma (Frobenius) de la matriu inicial. Això és equivalent a calcular

$$Score_{SS} = \frac{\sqrt{\sigma_5^2 + \cdots + \sigma_{16}^2}}{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_{16}^2}},$$

on $\sigma_1, \ldots, \sigma_{16}$ són els valors singulars de la matriu, en ordre decreixent. El valor obtingut, l'anomenem l'*score* de la topologia en questió. El resultat esperat és que l'*score* de la topologia correcta sigui molt més baix que els *scores* de les altres dues topologies.

El programa `Erik+2`, en canvi, realitza dos processos. En primer lloc, multiplica cada columna per un factor adequat de tal manera que les seves columnes sumin 1, calcula la distància (sense normalitzar) d'aquesta nova matriu a les de rang ≤ 4 i obté així un *score*. En segon lloc, aplica el mateix procés, però normalitzant per files en lloc de per columnes, obtenint així un segon *score*. L'*score* assignat a la topologia és la mitjana aritmètica d'aquests dos valors:

$$Score_{E+2} = \frac{Score_{col} + Score_{row}}{2}$$

De nou, esperem que l'*score* de la topologia correcta sigui considerablement menor als *scores* de les altres dues topologies.

II. ANÀLISI DELS DOS MÈTODES

Per analitzar amb profunditat la *performance* dels dos mètodes, no n'hi ha prou amb utilitzar dades reals, ja que no hi ha suficient material per a fer un estudi gaire significatiu. Per aquest motiu, hem decidit fer moltes simulacions per a obtenir alineaments de quatre espècies, que anomenem $1, 2, 3$ i 4 sota la topologia (12|34). És més, al conèixer explicitament l'arbre, la topologia i els paràmetres, podem esbrinar millor quines són les mancances i virtuts de cadascun del programes.

Els alineaments han estat creats utilitzant el programa `GenNon-h` [6]. Les seqüències generades corresponen a un arbre de quatre fulles amb diferents paràmetres de

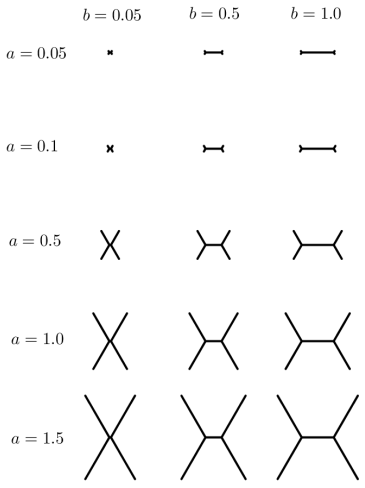

Figura 2: Els arbres estudiats, tots de quatre fulles, classificats segons a i b.

[1] Aquest és, de fet, l'estimador de màxima versemblança.

branca. Hem usat un model de dos paràmetres a i b que modelen respectivament a les longituds de les branques exteriors i central. Els valors que hem usat són els següents: $a \in \{0.05, 0.1, 0.5, 1.0, 1.5\}$; $b \in \{0.05, 0.5, 1.0\}$, que contempla una gran varietat de formes (veure Figura 2). Per a cada valor de a i b, hem generat 400 alineaments i hem calculat els *scores* per cadascun dels mètodes, utilitzant els programes pertinents. Per a cada a i b, hem realitzat un histograma on es mostren els scores de cada topologia (en total 15 histogrames per programa, és a dir, un total de 30), de manera que no tots poden ser mostrats en aquest treball.

En general, podem dir que els dos mètodes tenen els mateixos punts forts i febles: o bé tots dos distingeixen molt bé la topologia correcta, com en el cas $a = 0.05; b = 0.05$, o bé tots dos no la saben diferenciar de les altres, com en el cas $a = 1.5; b = 0.05$ (Figura 4). Aquest comportament sembla dependre fortament del paràmetre a, més inclús que del quocient a/b. Això es deu al fet que a's curtes significa poca mutació entre fulles, de manera que s'identifiquen millor com a fulles contigües. Per a a's grans, hi ha tantes mutacions que a ambdós programes els costa identificar-ne la procedència, i no són capaços de determinar categòricament la topologia correcta, encara que el resultat millora a mesura que augmenta la b: (per exemple, els histogrames per a $a = 1; b = 1$ es comencen a separar.

Pel que fa a les diferències entre els programes, veiem que, en primer lloc, el rang de valors que prenen els *scores* no és comparable: L'*score* de l'`SplitSup` està sempre entre 0 i 1 per construcció, mentre que el de l'`Erik+2` no té aquesta fita. És per aquest motiu que hem realitzat els histogrames del dos programes per separat. Un altre fet a remarcar és que, generalment, els *scores* de les males topologies de l'`SplitSup` són més dispersos que no pas els de l'`Erik+2`, cosa que fa que aquest últim sigui lleugerament més fiable.

III. CONTRAST AMB DADES REALS

A continuació, avaluem els dos mètodes de reconstrucció filogenètica sobre un conjunt d'espècies del qual ja es coneix l'arbre filogenètic. Els especímens en qüestió són 8 espècies de llevat: *C. albicans* (1), *S. bayanus* (2), *S. castellii* (3), *S. cerevisiae* (4), *S. kluyveri* (5), *S. kudriavzevii* (6), *S. mikatae* (7) i *S. paradoxus* (8). El seu arbre filogenètic és conegut [7] el mostrat la Figura 1. Degut al fet que ara tractem amb dades reals, hem decidit modelar l'evolució real com a una barreja d'evolucions independents, però sota el mateix arbre evolutiu. Si admetem m d'aquestes evolucions, llavors les matrius de *flattening* esdevenen sumes de m matrius. Així, la topologia correcta tindrà com a molt rang $4m$ per la desigualtat triangular i serà distingible de les de rang complet si $m \leq 3$. Les limitacions de l'algoritme ens han posat doncs una fita superior per a la m, però això no és gaire problema ja que des de la biologia es pensa que l'evolució de la vida sol seguir un patró amb una m baixa [1]. En qualsevol cas, nosaltres hem calculat els scores utilitzant $m = 3$ per als dos programes. Els programes agafen cada quartet de l'arbre i avaluen les tres topologies possibles. Els programes retornen uns pesos que sumen 1 i són inversament proporcionals als scores, de manera que s'espera que la topologia correcta en cada cas obtingui un pes alt comparat amb els altres. Heus aquí les gràfiques dels pesos que hem obtingut (Figura 3). Per a cada quartet, el rectangle negre simbolitza el pes associat a la topologia correcta, mentre que els rectangles gris i blanc mostren els pesos de les altres topologies.

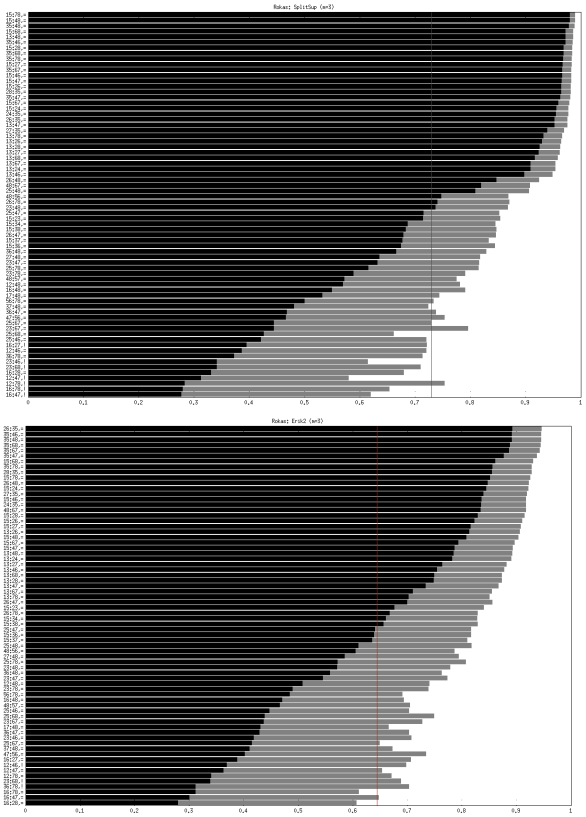

Figura 3: Al costat de cada quartet, hem posat el símbol `=` si el programa ha identificat correctament la topologia, mentre que hem posat el símbol `!` si no l'ha encertada.

Com es pot observar, el programa `Erik+2` ha realitzat menys errors (3 en front a 7) que l'`SplitSup`. Tot i així, els errors comencen en ambdós programes per sota d'un pes d'aproximadament 0.4. A més, l'`SplitSup` té molta més confiança quan dóna el resultat correcte (es pot observar com els que estan per sobre de la mitjana ho estan per un bon tros), cosa que l'`Erik+2` manca.

IV. CONCLUSIONS

Mitjançant les simulacions, hem vist que els dos mètodes actuen de manera semblant, però tenen proble-

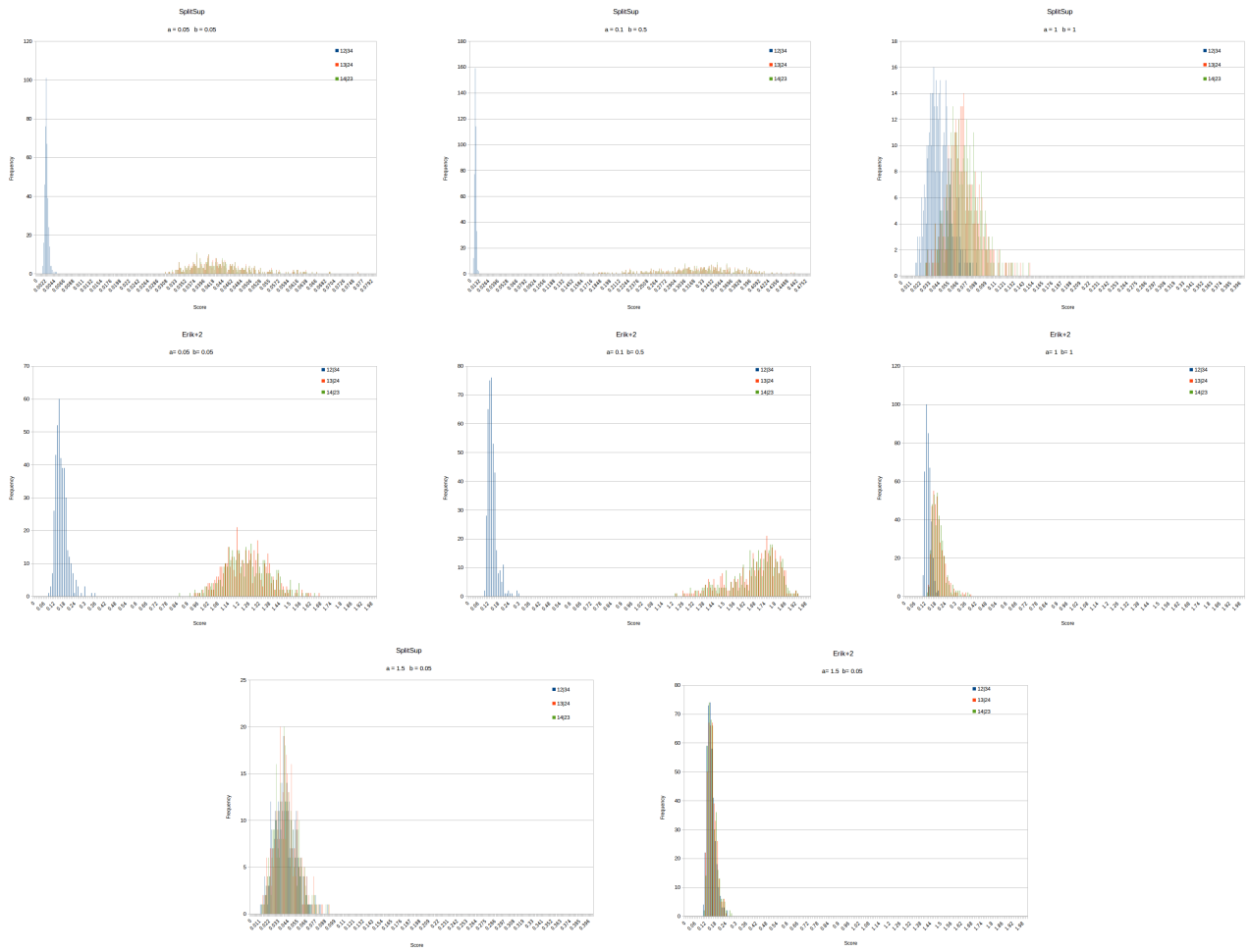

Figura 4: Vuit dels histogrames obtinguts amb els programes `SplitSup` i `Erik+2`. Primera fila: `SplitSup` amb $a = 0.05, b = 0.05$; $a = 0.1, b = 0.5$; $a = 1, b = 1$; d'esquerra a dreta. Segona fila, el mateix amb `Erik+2`. Tercera fila, `SplitSup` amb $a = 1.5, b = 0.05$ i `Erik+2` amb els mateixos valors. El color blau correspon a la topologia correcta.

mes amb els arbres amb branques massa llargues. Tot i no tenir suport compacte l'`Erik+2` té una distribució d'*scores* més concentrada que la de l'`SplitSup` cosa que pot permetre, a la llarga, saber si un *score* és bo o dolent sense haver de fer simulacions per a obtenir els histogrames en què es basa aquest treball. De fet, estudiar la distribució d'aquests *scores* per a aquest fi era un dels apartats que volíem inclure en aquest treball, però el temps i l'espai no ens ho han permès.

Finalment, també hem vist, utilitzant les dades reals, que els dos mètodes funcionen correctament sempre que el pes de la topologia guanyadora superi 0.4. En cas que el pes no predomini, el mètode pot induir a errors.

[1] M. Casanellas. *Técnicas algebraicas para la evolución de las especies*. La Gaceta de la RSME, Vol. 15, pp. 521-536, 2012

[2] N. Eriksson. *Tree Construction using Singular Value Decomposition*, 2005

[3] J. Fernández-Sánchez, M. Casanellas. *Invariant Versus Classical Quartet Inference When Evolution is Heterogeneous Across Sites and Lineages*, 2015

[4] E. S. Allman, et al. *Split scores: a tool to quantify phylogenetic signal in genome scale data*. Systematic Biology, doi:10.1093/sysbio/syw103, `http://arxiv.org/abs/1608.00942`, 2016.

[5] C. Sanderson and R. Curtin. *Armadillo: a template-based C++ library for linear algebra*. Journal of Open Source Software, Vol. 1, pp. 26, 2016.

[6] A.M. Kedzierska i M. Casanellas. *GenNon-h: Simulating multiple sequence alignments under the non-homogeneous DNA models*, BMC Bioinformatics 12:216, 2012.

[7] A. Rokas, et al. *Genome-scale approaches to resolving incongruence in molecular phylogenies*. Nature 425, 2003.