Butlletí de la Societat Catalana de Matemàtiques
Vol. 25, núm. 2, 2010. Pàg. 145–169. DOI: 10.2436/20.2002.01.30

Invariants, una aproximació a la filogenètica des de l'àlgebra*

Jesús Fernández-Sánchez

> This oak tree and me, we're made of the same stuff. If you go back far enough you'll find that we have a common ancestor. That's why our chemistry is so alike.
>
> **Carl Sagan, "Cosmos"**

> Phylogenetics needs mathematicians. It needs them for handling large data sets. It needs them for developing new methods, and analysing old ones. It needs them because mathematics is the art of being systematic.
>
> **David Bryant**

Resum En els darrers anys, una nova aproximació a la reconstrucció filogenètica basada en tècniques provinents de l'àlgebra s'ha estat consolidant. Fixat un model evolutiu per a un conjunt donat d'espècies, els *invariants* són relacions algebraiques satisfetes per les distribucions teòriques de nucleòtids d'aquestes espècies. En aquest article s'exposa com es poden fer servir els invariants per implementar algoritmes de reconstrucció filogenètica i s'explica com l'eficiència d'aquests algoritmes es veu beneficiada per resultats teòrics provinents de la geometria algebraica i la representació de grups.

Paraules clau: reconstrucció filogenètica, model evolutiu, varietat algebraica.

Classificació MSC2010: 14Q15, 92D15, 92D20.

* Aquesta exposició correspon a la conferència impartida per l'autor a la XII Trobada de la Societat Catalana de Matemàtiques, que va tenir lloc a Barcelona, el 5 de juny de 2009.

1 Introducció

Segons la *teoria de l'evolució* de Darwin (i Wallace), la selecció natural és el mecanisme bàsic a partir del qual es produeix l'evolució. Però per tal que la selecció sigui possible hi ha d'haver una variabilitat prèvia en l'estructura genètica de les espècies. Aquesta variabilitat genètica s'introdueix en el nivell molecular, en el DNA dels individus, a través de canvis aleatoris que es produeixen quan les molècules es copien per donar lloc a les noves generacions. En funció d'aquests canvis, els descendents poden ser més, menys o igualment viables que els pares. Es creu que gran part d'aquests canvis són selectivament neutres, i, per tant, són preservats en les successives generacions. De fet, el DNA d'un gen particular continua mutant de generació en generació, acumulant gradualment més diferències respecte a la seva forma ancestral fins a donar lloc a *gens homòlegs*. D'aquesta manera, espècies diferents que provenen d'un ancestre comú tenen un genoma semblant, però no idèntic. Les semblances apunten cap a l'ancestre comú, mentre que les diferències ens parlen de la divergència evolutiva dels descendents.

La semblança dels mecanismes moleculars dels organismes suggereix fortament que tots els organismes de la Terra tenen un ancestre comú, de manera que qualsevol conjunt d'espècies està relacionat evolutivament. Aquestes relacions s'anomenen *relacions filogenètiques* i es representen mitjançant un *arbre filogenètic* (vegeu les figures 1 i 2). D'una manera o altra, els arbres filogenètics s'han fet servir durant cent quaranta anys, però la recerca estadística, computacional i algorítmica relacionada es desenvolupa des de fa només quaranta anys. La recerca filogenètica té com a principal objectiu deduir aquests arbres a partir dels organismes existents i depèn fortament de la identificació de caràcters homòlegs entre un conjunt de tàxons. Tradicionalment, aquests caràcters podien ser característiques físiques dels organismes (tant vius com fossilitzats) com, per exemple, si els organismes són unicel·lulars o pluricel·lulars, de sang freda o calenta, etc. Actualment, podem «llegir» el genoma de les espècies amb relativa facilitat, així que sorgeix una pregunta natural: *Podem reconstruir les relacions filogenètiques entre un grup d'espècies comparant seqüències de nucleòtids dels seus gens?*

En l'article pioner [48], Zuckerkandl i Pauling ja mostraven que cadenes moleculars proporcionen conjunts de caràcters que aporten una gran quantitat d'informació evolutiva. És natural esperar que les espècies que tenen el genoma més semblant siguin les que estan més a prop evolutivament, però amb això no n'hi ha prou per deduir les relacions filogenètiques. Necessitem, primer de tot, precisar com mesurem el grau de semblança entre les cadenes: en aquest sentit, els *models evolutius* (models matemàtics dels processos de mutació del DNA) són de gran ajuda i es poden fer servir per fer deduccions sobre la història evolutiva a partir de dades reals.

D'altra banda, per tal de deduir relacions filogenètiques entre espècies a partir de cadenes de DNA, necessitem disposar aquestes cadenes d'una manera especial amb l'objectiu de posar de manifest les seves diferències i sem-

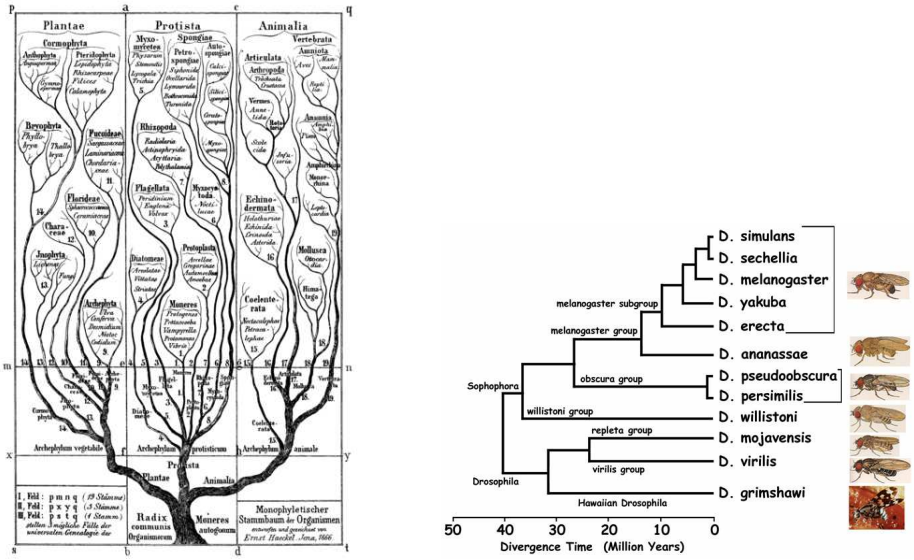

FIGURA 1: Dos exemples d'arbres filogenètics. A l'esquerra, tenim l'*Arbre de la vida* dibuixat pel biòleg E. Haeckel en la seva obra «Morfologia general dels organismes» (1866). L'arbre de la dreta representa les relacions filogenètiques entre algunes espècies del gènere *Drosophila*, així com el temps de divergència estimat.

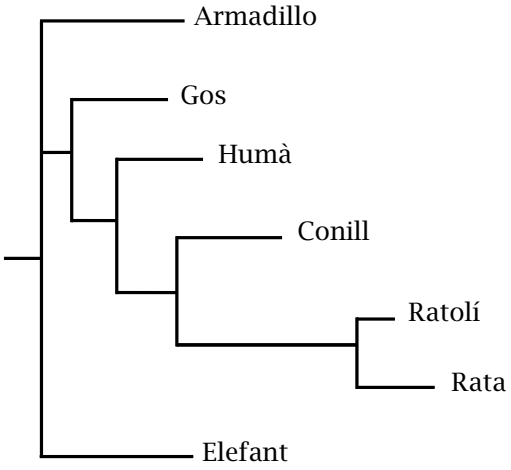

FIGURA 2: Arbre filogenètic de set espècies de mamífers obtingut a partir d'un alineament d'una part de la regió HOXA (regió ENCODE ENm010, vegeu [14] i genome.ucsc.edu/encode) mostrat parcialment en el text. L'arbre s'ha obtingut fent servir el paquet `dnaml` del programari PHYLIP [23].

blances, i així identificar regions de similitud (de tipus estructural, funcional o evolutiu). D'aquesta disposició especial en diem un *alineament*: les cadenes s'acostumen a representar com files d'una taula, on es poden inserir *gaps* (forats) entre les bases per indicar insercions/supressions de nucleòtids, de manera que cada columna de l'alineament representa el resultat de l'evolució sobre una mateixa posició del genoma de l'ancestre comú. L'alineament òptim és aquell que minimitza el nombre de substitucions, insercions o supressions. A tall d'exemple, mostrem parcialment un alineament de set espècies de mamífers:

```
Humà       CCCCGGTGTACTCTAACCACTGAAG-CGGCCGTGTCGGGGACTCACGCGCTTCCCATTCAGCTCTGGATCTGGAACTGGCCCCT
Ratolí     CCCCGTCGCCT-TGATCATTTAAACGGGCCCTGTAGCAGGCTAGCT--ATCCTATACATTTCTGGGCCTGGAGCTGGCCTCA
Rata       CCCCGTCACCCCATGATCGTTTAAGCGGGCCCTGTAGCAGTCTAGGT--GTCCTATTCATTTCTGGACATGGAGCTGGCCTCA
Conill     ACCTGGCGTGCGGTGACCACTGAAGGGGGCTCTGTCGGGACCTCACTCTCATCCCATACAGCTCTGGACCTGGAGCTGGCCTCA
Gos        CCTCCAGGCACTCTAACCATTAAGG-GGGCC-TGTCGGGGCCTCAGGCTCTTCCCATTCAGCTCGGGACCTGGAGCTG---
Armadillo  CCTCCGTGCTCTCTGACCACTAAGG-GGGCCCTGTCGGGGCCTCAAGCTCTTCCAGTCCAGCTGTGGACCTGGAGCGGGCCCCA
Elefant    CCCTGGGGCGCTCTGACCACTGAGA-AGCCTCGTCGGGGCTTCAAGCTCTTCCCCTTCAGC-CTGGACCTGCAGCTTGCCCCA
```

L'arbre filogenètic de la figura 2 ha estat obtingut a partir d'aquest alineament fent servir el paquet `dnaml` del programari PHYLIP [23]. Per a més informació sobre alineaments, vegeu [17, Ch.6].

Mètodes de reconstrucció filogenètica

A l'hora d'inferir relacions filogenètiques, els tàxons que volem relacionar són normalment espècies que encara existeixen. Coneixem (o podem aspirar a conèixer) el genoma d'aquestes espècies, però no disposem d'informació directa del genoma dels ancestres d'aquestes espècies. De fet, ni tan sols sabem quines espècies són aquests ancestres perquè no coneixem la topologia de l'arbre.

Existeix un bon nombre de mètodes de reconstrucció filogenètica. Un primer grup són els mètodes *basats en distància* que pretenen construir l'arbre a partir d'una taula de *distàncies* entre tàxons. En funció del model evolutiu que assumim, existeixen diferents fórmules per trobar aquestes distàncies a partir d'un alineament donat. Entre aquests mètodes trobem el `UPGMA`,[1] l'*algoritme de Fitch-Margoliash* i el popular mètode de *Neighbor-Joining* (vegeu [41]). Un altre gran grup són els mètodes de *màxim de parsimònia* (vegeu [26]), que busquen els arbres filogenètics que minimitzen el nombre de mutacions ocorregudes durant el procés evolutiu implicat. La tercera gran classe són els mètodes de *màxim de versemblança*. Un cop especificat un model evolutiu i un arbre filogenètic que relaciona un grup d'espècies, la probabilitat que les dades obtingudes (l'alineament) hagin estat generades és la *versemblança* del model. Els mètodes de *màxim de versemblança* trien l'arbre que maximitza aquesta probabilitat (vegeu [22]).

Els *invariants* suposen una aproximació radicalment diferent al problema de la reconstrucció filogenètica. Donat un alineament, els invariants són relacions algebraiques satisfetes per les freqüències esperades de cada columna possible en l'alineament d'acord amb un model evolutiu i un arbre filogenètic

1 Inicials de Unweighted Pair Group Method with Arithmetic Mean.

prèviament fixats. Des de la seva introducció per Cavender i Felsenstein [13] i Lake [36], hi ha hagut diferents intents per obtenir tots els invariants (vegeu, per exemple, [44, 25]), però no ha estat fins fa poc que els geòmetres algebraics s'han interessat per la qüestió i han aconseguit trobar-los tots per certs models evolutius [2, 12, 45]. Tot i que els mètodes basats en invariants han demostrat ser útils en genòmica comparativa [42], durant alguns anys ha imperat la creença que els invariants no són eficients per a la reconstrucció filogenètica ja que requereixen alineaments molt llargs per inferir-ne les relacions filogenètiques correctament (vegeu [32] i [31]). Però com Felsenstein explica a [24], els invariants mereixen més atenció pel que «poden aportar en el futur». Estudis recents mostren mètodes prometedors basats en invariants. Una introducció pràctica a la teoria dels invariants es pot trobar al capítol 22 de [24], mentre que [39] proporciona una bona introducció a les aplicacions de l'estadística algebraica (i, en particular, dels invariants filogenètics) a la biologia computacional.

Hi ha diverses motivacions per a fer servir els invariants en la reconstrucció filogenètica. En primer lloc, se sap que la topologia de l'arbre es pot recuperar a partir de les dades observades (diem que és *identificable*) per mètodes basats en invariants [3], de manera que el seu ús per a la reconstrucció d'arbres és consistent (vegeu [29, 24]). Un altre motiu és la despesa computacional d'altres mètodes, com el màxim de versemblança, que passen per l'estimació total de la topologia d'arbre, les longituds de branca i la matriu de taxes. Una de les virtuts de treballar amb invariants és que podem treballar més o menys directament amb les dades observades, en comptes d'intentar trobar els valors dels paràmetres que millor s'hi adapten. D'altra banda, els models evolutius en què es basen els invariants permeten tractar dades no homogènies (vegeu la secció 2.2). De fet, se sap que alguns conjunts de dades biològiques requereixen matrius de taxes diferents per als diferents llinatges. És essencial, doncs, tenir al nostre abast mètodes filogenètics que permetin tractar amb aquest tipus de dades [27, 47]. Els invariants interessants en la inferència filogenètica són aquells que depenen de la topologia de l'arbre: en diem *invariants filogenètics*. Però aquests invariants són interessants per moltes altres raons. Des del punt de vista teòric, s'han fet servir per respondre qüestions sobre la identificabilitat dels models (per ex., [3, 37]). D'altra banda, l'estudi de la geometria al voltant dels invariants també ha portat a problemes dignes d'atenció des del punt de vista de les matemàtiques [19, 5, 15].

L'objectiu de l'article és presentar un resum d'una part del treball realitzat en col·laboració amb Marta Casanellas [8, 9, 10]. El títol respon a la meva intenció en escriure'l. La filogenètica és un camp extens de recerca, amb un ample ventall de tècniques que sovint donen lloc a conclusions diferents. No existeix un mètode de reconstrucció que funcioni millor que els altres en tots els casos, però és possible trobar mètodes més adients que altres en funció de les dades. La filogenètica admet aproximacions diferents, i l'ús de l'àlgebra a través dels «invariants» n'és una més.

2 Preliminars

2.1 Arbres filogenètics

Un arbre filogenètic T és un graf connex i finit, sense cicles, els nodes del qual representen espècies o altres entitats biològiques. El conjunt de nodes de T el representem per $N(T)$ i en distingim les fulles $L(T)$, que representen espècies actuals, i els nodes interiors $\mathrm{Int}(T)$, que representen entitats ancestrals. D'aquesta manera: $N(T) = L(T) \cup \mathrm{Int}(T)$. Les branques de T les denotem per $E(T)$ i representen processos evolutius. Un *arbre amb arrel* és un arbre amb un node distingit, l'*arrel*, que indueix una orientació en les branques. Quan hi és present, l'arrel representa l'ancestre comú més proper a totes les espècies que apareixen a les fulles de l'arbre. El *grau* d'un node és el nombre de branques que hi incideixen. Treballarem amb arbres *trivalents*, és a dir, arbres pels quals cada node interior (excepte potser l'arrel) té grau 3. Aquesta hipòtesi és necessària en alguns resultats i sovint és el cas d'interès principal en filogenètica (vegeu [4]). La *topologia d'un arbre* és la seva topologia com a subespai del pla, tenint en compte que les fulles estan etiquetades. Per a $n \geq 3$, denotarem per $\mathcal{T}_n$ el conjunt de les topologies possibles de n fulles, trivalents, etiquetades i sense arrel. Per exemple, la figura 3 mostra les tres topologies d'arbre a $\mathcal{T}_4$.

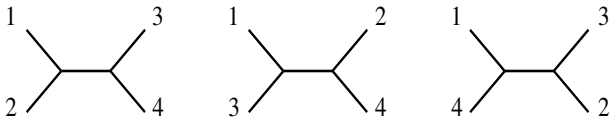

FIGURA 3: Les tres topologies possibles a $\mathcal{T}_4$ (arbres trivalents sense arrel i amb 4 fulles).

2.2 Models evolutius algebraics

Sigui T un arbre filogenètic trivalent amb n fulles i amb arrel. A cada node de l'arbre considerem una variable aleatòria que pren valors en un espai finit d'estats. Tenim, doncs, un model gràfic en el sentit de [18] (consulteu [39, secció 1.5] per a una introducció als models gràfics). En el nostre cas, suposem que aquest espai d'estats és el mateix per a totes les variables i igual al conjunt dels nucleòtids, que representem per $\Sigma = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. Més precisament, treballarem amb un model de Markov ocult (HMM) on les variables a les fulles són les variables *observades* del model, que denotem per X_i, mentre que les variables als nodes interiors són les variables *ocultes*, que denotem per Y_j (vegeu la figura 4).

A cada branca e de T considerem una *matriu de substitució* S_e, les entrades de la qual són les probabilitats d'observar mutacions entre el node pare i el

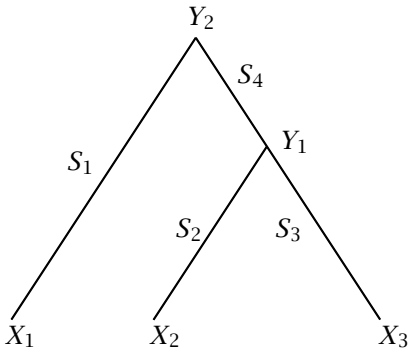

FIGURA 4: Model gràfic consistent en un arbre filogenètic amb variables aleatòries associades als nodes que prenen valors en $\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$: les variables X_1, X_2, X_3 són observades i les variables Y_1, Y_2 són ocultes.

node fill de la branca. Així, les matrius de substitució tenen la forma:

$$S = \begin{array}{c} \\ \mathsf{A} \\ \mathsf{C} \\ \mathsf{G} \\ \mathsf{T} \end{array} \begin{pmatrix} \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{T} \\ P(\mathsf{A}|\mathsf{A}) & P(\mathsf{C}|\mathsf{A}) & P(\mathsf{G}|\mathsf{A}) & P(\mathsf{T}|\mathsf{A}) \\ P(\mathsf{A}|\mathsf{C}) & P(\mathsf{C}|\mathsf{C}) & P(\mathsf{G}|\mathsf{C}) & P(\mathsf{T}|\mathsf{C}) \\ P(\mathsf{A}|\mathsf{G}) & P(\mathsf{C}|\mathsf{G}) & P(\mathsf{G}|\mathsf{G}) & P(\mathsf{T}|\mathsf{G}) \\ P(\mathsf{A}|\mathsf{T}) & P(\mathsf{C}|\mathsf{T}) & P(\mathsf{G}|\mathsf{T}) & P(\mathsf{T}|\mathsf{T}) \end{pmatrix}.$$

Suposarem també la hipòtesi habitual d'independència evolutiva dels caràcters, de manera que, per exemple, cada columna d'un alineament donat és una mostra independent i idènticament distribuïda del model evolutiu (*i. i. d. hypothesis*). Si suposem que l'evolució s'esdevé sota un procés de Markov en temps continu (vegeu [38]), cada branca e de T té associada una longitud t_e i una matriu Q_e les entrades de la qual són les taxes instantànies d'evolució al llarg de la branca. La matriu de substitució corresponent és $S_e = e^{Q_e t_e}$, que és la forma que pren la solució de l'equació diferencial:

$$S_e'(t) = S_e(t)Q_e \qquad S_e(0) = \mathrm{Id}\,.$$

En aquest cas, els paràmetres del model són les entrades de les matrius Q_e juntament amb les longituds de branca $\{t_e\}_{e\in E(T)}$.

Per raons computacionals, els mètodes més comuns de reconstrucció filogenètica es basen en models *homogenis*: models que presenten la mateixa matriu de taxes a totes les branques de l'arbre. Atès que l'evolució no succeeix sempre al mateix ritme [47], els models homogenis no són del tot realistes i desenvolupar mètodes per solucionar aquest inconvenient és un focus important de recerca (cf. [46, 28]; vegeu [1] per la relació entre models algebraics i models homogenis i el «problema dels rosegadors» relacionat). Els models evolutius *algebraics* no presenten aquest problema, ja que prenen com a paràmetres directament les entrades de les matrius de substitució. Si M és un model algebraic, denotem per Par_T^M l'espai de paràmetres corresponent. Cada

paràmetre és una certa entrada d'una matriu de substitució i, com a tal, és una probabilitat i està subjecte a certes restriccions. En funció de les simetries que presenten les matrius de substitució, distingim els diferents models algebraics.

Alguns exemples. En el model algebraic de *Kimura amb 3 paràmetres* (K81, vegeu [35]) les matrius de substitució presenten l'estructura següent:

$$S_e = \begin{array}{c} \\ \mathsf{A} \\ \mathsf{C} \\ \mathsf{G} \\ \mathsf{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{T} \end{array} \\ \begin{pmatrix} a_e & b_e & c_e & d_e \\ b_e & a_e & d_e & c_e \\ c_e & d_e & a_e & b_e \\ d_e & c_e & b_e & a_e \end{pmatrix} \end{array}.$$

Si en aquest model suposem que $b_e = d_e$, obtenim el model de *Kimura amb 2 paràmetres* (K80, vegeu [34]); si suposem que $b_e = c_e = d_e$, en resulta el model de *Jukes-Cantor* (JC69, vegeu [33]).[2] Tots aquests models assumeixen la hipotèsi addicional que la distribució a l'arrel és uniforme (*i. e.* $\pi_{\mathsf{A}} = \pi_{\mathsf{C}} = \pi_{\mathsf{G}} = \pi_{\mathsf{T}} = 1/4$). Els tres models anteriors s'anomenen *models de grup* perquè les matrius de substitució són compatibles amb l'estructura del grup additiu $H = (\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, +)$: si identifiquem els nucleòtids $\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}$ amb els elements $(0,0), (1,0), (0,1), (1,1)$ de H, podem entendre aquestes matrius com a funcions $f_e : H \to \mathbb{C}$ que satisfan $f_e(\mathsf{X} - \mathsf{Y}) = S_e(\mathsf{X}, \mathsf{Y})$.

D'altra banda, si no suposem cap estructura especial en les matrius de substitució, obtenim el model *general de Markov* (GMM). Per a més informació sobre models algebraics i models evolutius en general, vegeu [39, secció 4.5].

2.3 Geometria algebraica

La geometria algebraica és, abans de res, l'estudi de sistemes d'equacions polinomials en diverses variables i dels corresponents conjunts de solucions, que anomenem *conjunts algebraics*. Una *varietat algebraica* és un espai definit localment per equacions algebraiques. Les tècniques que s'utilitzen en geometria algebraica provenen fonamentalment de l'àlgebra abstracta i, especialment, de l'àlgebra commutativa. Aquesta interacció entre l'àlgebra i la geometria permet trobar respostes algebraiques a problemes geomètrics, d'una banda, i respostes geomètriques a problemes algebraics, de l'altra.

Donat un anell de polinomis sobre $\mathbb{C}$, el *teorema de la base* de Hilbert implica que tot ideal es pot generar a partir d'un nombre finit de polinomis.[3] Si ens restringim a l'estudi dels conjunts algebraics en un espai afí sobre $\mathbb{C}$, el *teorema dels zeros* de Hilbert (*Hilbertscher Nullstellensatz*) estableix un diccionari entre el llenguatge de la geometria i el de l'àlgebra, que tradueix «conjunt

2 Kimura va proposar el model K80 com una generalització del model JC69, que permet diferents taxes de mutació per a les *transicions* i les *transversions* (vegeu [34]).

3 Un ideal és una subestructura algebraica, tancada per la suma i pel producte amb elements de l'anell.

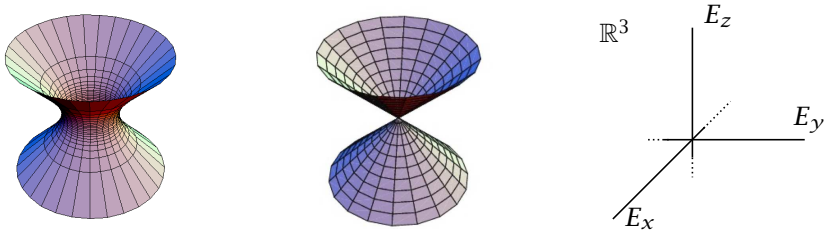

FIGURA 5: L'hiperboloide d'un full $X^2 + Y^2 - Z^2 = 1$ (esquerra) no té punts singulars; el con $X^2 + Y^2 - Z^2 = 0$ (centre) presenta un punt singular en $(0, 0, 0)$. Els eixos de coordenades en $\mathbb{R}^3$ (dreta) són un exemple de varietat que no és intersecció completa local en $(0, 0, 0)$.

algebraic» per «ideal radical»: a cada conjunt algebraic V li assignem l'ideal I_V format per les equacions polinomials que satisfan tots els punts del conjunt i, a la inversa, a cada ideal I de l'anell de polinomis li assignem el conjunt algebraic V_I format pels punts que anul·len tots els polinomis de l'ideal. D'aquesta manera, els conjunts algebraics *irreductibles* en l'espai afí $\mathbb{C}^n$ es tradueixen pels ideals *primers* de $\mathbb{C}[X_1, \ldots, X_n]$ i els punts de $\mathbb{C}^n$ es tradueixen pels ideals *maximals*.

Farem servir la paraula *varietat* com a sinònim de conjunt algebraic, irreductible o no. Hi ha un parell de conceptes en geometria algebraica que tindran un paper important en la secció 4: el concepte de *punt singular* i el d'*intersecció completa local*. De manera intuïtiva, un punt és *singular* si l'espai tangent a la varietat no està definit (amb precisió, si l'ideal maximal associat al punt requereix més generadors que el valor de la dimensió de la varietat; vegeu la figura 5). Se segueix que, fora dels seus punts singulars, tota varietat es pot definir localment per tantes equacions com la codimensió de la varietat. Diem, per tant, que, fora dels punts singulars, tota varietat és localment una *intersecció completa*. Un exemple senzill d'una varietat que no és localment una intersecció completa és la unió dels tres eixos en l'espai afí de dimensió 3 (vegeu [30, Lecture 11]).

3 Invariants filogenètics

Donat un model evolutiu (algebraic) M i un arbre filogenètic T amb arrel r i n fulles, la probabilitat d'observar nucleòtids $\mathsf{x}_1, \ldots, \mathsf{x}_n$ a les fulles de l'arbre ve donada per una expressió algebraica en els paràmetres del model:

$$p_{\mathsf{x}_1 \ldots \mathsf{x}_n} = 1/4 \sum_{\{(\mathsf{x}_v)_{v \in Int(T) \setminus \{r\}} \mid \mathsf{x}_v \in \Sigma\}} \prod_{e \in E(T)} (S_e)_{\mathsf{x}_{p(e)}, \mathsf{x}_{f(e)}}$$

on $p(e)$ i $f(e)$ denoten els nodes pare i fill de la branca e i suposem que si $e = e_l$ és una branca terminal, aleshores $\mathsf{x}_{f(e)} = \mathsf{x}_l$. D'aquí resulta que cada arbre T

determina una aplicació *polinomial* en el símplex de probabilitat $\triangle^{4^n-1}$:

$$\begin{aligned} \varphi_T^M : \mathrm{Par}_T^M &\longrightarrow \triangle^{4^n-1} \\ \theta = (\theta_1, \ldots, \theta_d) &\mapsto (p_{\mathsf{AA\ldots A}}, p_{\mathsf{AA\ldots C}}, p_{\mathsf{AA\ldots G}}, \ldots, p_{\mathsf{TT\ldots T}}). \end{aligned} \tag{3.1}$$

Per tal de poder treballar amb varietats algebraiques, oblidarem l'espai Par_T^M i suposarem l'aplicació φ_T^M definida en tot l'espai afí $\mathbb{C}^d$, on d és el nombre de paràmetres del model:

$$\varphi_T^M \colon \mathbb{C}^d \longrightarrow \mathbb{C}^{4^n}.$$

Denotarem per V_T^M la varietat algebraica definida per la clausura de Zariski de la imatge d'aquesta aplicació. Equivalentment, V_T^M és la varietat algebraica més petita que conté aquesta imatge. Observem que el model algebraic ha esdevingut considerablement més gran que el model estadístic que resulta de considerar l'aplicació (3.1), ja que ara les entrades $p_{\mathsf{x}_1\ldots\mathsf{x}_n}$ poden ser negatives o, fins i tot, complexes. Tot i que això ens permet fer servir eines i resultats algebraics, també pot esdevenir un desavantatge (vegeu [20, secció 7]). En qualsevol cas, no hem d'oblidar que des del punt de vista biològic els punts de la varietat que ens interessen són reals i estan sobre el símplex $\triangle^{4^n-1}$.

Fins ara hem suposat que els arbres tenen arrel. A partir d'ara, però, treballarem amb arbres sense arrel: el motiu és que, en general i tal com s'explica en [7, 9], la posició de l'arrel no és una dada *identificable*. *Grosso modo*, la qüestió de la identificabilitat consisteix a saber si les dades observades a les fulles de l'arbre contenen prou informació per determinar la topologia i els paràmetres del model. Això significa que existeix una única topologia i un únic conjunt de paràmetres que poden explicar les dades observades. En el cas d'arbres amb arrel, la dimensió de les fibres de φ_T^M és positiva, la qual cosa vol dir que existeix una infinitat de paràmetres a Par_T^M que donen lloc a les mateixes dades, i ens hem de restringir a arbres sense arrel si volem fibres discretes.

En la secció anterior, hem explicat els models evolutius associats als arbres amb arrel. Expliquem breument els models d'arbres sense arrel. Sigui T un arbre sense arrel i triem un node r a T, que prendrem com a arrel, i obtenim així un arbre amb arrel T^r. Assignem matrius de substitució S_e a les branques de T orientades per r. Si ara prenem un segon node $r' \neq r$ com a arrel, i matrius de substitució S'_e definides per $S'_e = S_e^t$ quan l'orientació induïda per r' en e és contrària a la induïda per r, i $S'_e = S_e$ quan és la mateixa, es pot veure que les varietats associades a T^r i $T^{r'}$ coincideixen; així que podem pensar que el model és independent de la posició de l'arrel. Val a dir també que per a molts models algebraics, com ara els models de grup, les matrius de substitució són simètriques i, per tant, la parametrització (3.1) no depèn de l'orientació de l'arbre o de la posició de l'arrel. Vegeu [2, 16] per a detalls sobre la relació entre arbres amb arrel i sense.

Els invariants filogenètics van ser introduïts en un parell d'articles apareguts a final de 1987, un d'aquests escrit per J. Cavender i J. Felsenstein [13] i l'altre per J. Lake [36], per tal d'estudiar l'adequació dels models evolutius a les dades i per estudiar i deduir divisions evolutives entre espècies.

1 DEFINICIÓ Fixat un model evolutiu algebraic M, l'*ideal filogenètic* associat a M i a un arbre T és l'ideal (dins de l'anell de polinomis amb 4^n indeterminades) corresponent a la varietat algebraica V_T^M, és a dir,

$$I_T^M = \{\text{equacions satisfetes pels punts de } V_T^M\}.$$

Els *invariants* de T (sota el model M) són els elements de I_T^M. Diem que un invariant $f \in I_T^M$ és *filogenètic* si $f(p) \neq 0$ per a algun punt $p \in V_{T'}^M$ on $T' \neq T$ és un arbre de n fulles. Els invariants filogenètics són, doncs, invariants específics d'una certa topologia i poden ser utilitzats per tal de deduir relacions filogenètiques entre espècies. Els invariants que no són filogenètics no aporten informació rellevant per a la reconstrucció filogenètica, però precisament per això poden ser molt útils a l'hora d'ajustar un model evolutiu a unes dades donades.

Un dels principals objectius d'aquesta línia de recerca és entendre com utilitzar els invariants per seleccionar el millor arbre donades unes dades (que es presenten en forma d'alineament). Cal senyalar, però, que hi ha una colla d'inconvenients inherents a l'ús dels invariants que haurem de tenir en compte per tal de minimitzar-ne l'impacte (vegeu [20]):

a) Atès que les varietats estan immerses a $\mathbb{C}^{4^n}$, el nombre d'indeterminades dels invariants depèn exponencialment del nombre d'espècies, i avaluar els invariants en un punt pot ser extremadament complicat quan n creix. D'altra banda, el nombre de topologies d'arbre possibles amb n fulles, així com la codimensió de les varietats corresponents, també depenen exponencialment de n. Els invariants són realment útils per tractar problemes quan involucren un nombre elevat d'espècies?

b) Els models estadístics no són varietats algebraiques complexes; només tenen sentit estadístic aquells punts de les varietats sobre el símplex de probabilitat. Aquests punts formen, doncs, conjunts reals i semialgebraics. Podem fer servir aquesta informació semialgebraica per tal de millorar els mètodes basats en invariants?

Tanmateix també hi ha sèrie de característiques dels invariants que fan especialment interessant i prometedor el seu ús:

A. Els mètodes basats en invariants permeten fer servir diferents matrius de taxes per a cada branca (models *no homogenis*). En [8] es pot veure que per dades homogènies i sota determinades condicions, aquests mètodes donen lloc a millors resultats que altres basats en models homogenis.

B. Els invariants permeten comprovar aspectes locals dels arbres (vegeu la secció 5). Sovint es plantegen qüestions relatives a la validesa d'una branca concreta en un arbre; en termes filogenètics, això pot respondre, per exemple, a preguntar-se si el ximpanzé és més a prop de l'esser humà o del goril·la.

Ús dels invariants

Com es desprèn de l'apartat *a*) anterior, no és factible fer servir programes d'àlgebra computacional per trobar tots els invariants associats a arbres amb un nombre elevat de fulles (fins i tot, en el cas dels models més senzills, com el model de Jukes-Cantor). Necessitem, doncs, resultats teòrics que proporcionin algoritmes per trobar els invariants necessaris.

El resultat següent ha estat provat en casos particulars per E. Allman i J. Rhodes [4], B. Sturmfels i S. Sullivant [45], M. Casanellas i S. Sullivant [12], però han estat J. Draisma i J. Kuttler els qui han demostrat una versió més general del resultat, vàlida per a models equivariants.[4]

2 TEOREMA ([16]) *Sigui T un arbre filogenètic de n espècies que evoluciona sota un model equivariant M. Existeix un algoritme que permet obtenir un sistema de generadors de l'ideal I_T^M a partir dels invariants d'un arbre de 3 fulles sota el mateix model juntament amb certes equacions associades a les branques de T.*

Les *equacions* que apareixen en l'enunciat del teorema estan definides per restriccions de rang de certes matrius que seran definides en la secció 5. Val la pena observar que mentre aquestes restriccions de rang són relativament fàcils d'obtenir, no passa el mateix amb els invariants associats als arbres de 3 fulles. De fet, encara no es coneixen els invariants associats als arbres de 3 fulles sota alguns models algebraics, com el model general de Markov o el *strand symmetric model* [12]. En aquests casos el resultat anterior no es pot portar a la pràctica.

La idea bàsica per fer servir els invariants en problemes d'inferència filogenètica és la següent. Un alineament de n espècies dóna lloc a una distribució empírica de probabilitat representada per un punt $\hat{p} \in \mathbb{C}^{4^n}$: les coordenades d'aquest punt són les freqüències relatives d'aparició de cada columna possible en l'alineament (obviem aquelles columnes que contenen algun *gap*).

Si f és un invariant d'una certa topologia d'arbre sota un cert model d'evolució, és d'esperar que $f(\hat{p}) \approx 0$ si (i en general només si) l'alineament prové d'aquest arbre i d'aquest model. Més precisament, si $\hat{p}_N$ representa la distribució empírica d'un alineament de longitud N, aleshores $\lim_{N\to\infty} \mathbb{E}(f(\hat{p}_N)) = 0$ (on $\mathbb{E}$ denota l'*esperança*). Aquest fet ens suggereix una possible estratègia per reconstruir arbres filogenètics a partir d'un alineament de n espècies fent servir els invariants (vegeu la figura 6):

1. Fixat un model M d'evolució convenient per a les dades, triem un conjunt d'invariants $\mathcal{F}_T^M$ per a cada arbre T amb n fulles.
2. Per a cada arbre T, avaluem els invariants de $\mathcal{F}_T^M$ en el punt $\hat{p}$ obtingut a partir de l'alineament.
3. Triem l'arbre T que minimitza un cert *score* calculat a partir de les avaluacions anteriors (en algun sentit que cal concretar).

4 Els *models equivariants* seran introduïts en la secció 5 d'aquest article. De moment, és suficient assenyalar que els models de Jukes-Cantor, de Kimura 2 i 3 paràmetres i el model general de Markov en són exemples.

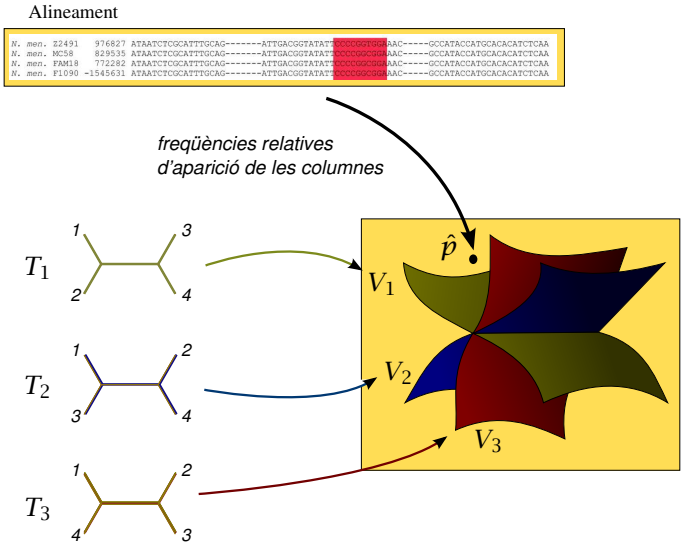

FIGURA 6: Donat un alineament de 4 espècies, les freqüències relatives de les diferents columnes de l'alineament són les coordenades d'un punt $\hat{p}$ a $\mathbb{C}^{4^4}$; per la seva banda, cada topologia possible a $\mathcal{T}_4$ determina una varietat algebraica.

Com ja hem assenyalat abans, cadascun d'aquests passos pot presentar dificultats. La tria d'un model adient a les dades pot ser difícil i, com és habitual en els processos de modelització, cal buscar un model de compromís entre el realisme biològic (que ens podria portar a centenars de paràmetres) i la necessària utilitat estadística del model. D'altra banda, la quantitat de polinomis on triar és infinita i el nombre de topologies d'arbres possibles creix exponencialment en funció del valor de n. Finalment, cal definir un *score* a partir dels invariants, que farem servir per a comparar els arbres. La idea és que aquest *score* reflecteixi la «distància» del punt $\hat{p}$ a la topologia correcta.

EXEMPLE En [11] s'introdueix un mètode de reconstrucció filogenètica basat en invariants que segueix l'estratègia anterior pel model `K81` i alineaments de 4 espècies. Un cop hem triat un sistema minimal de generadors $\mathcal{F}_T^{\mathtt{K81}}$ dels ideals filogenètics de les tres topologies a $\mathcal{T}_4$, prenem el valor

$$S_{\hat{p}}(T) = \sum_{f \in \mathcal{F}_T^{\mathtt{K81}}} |f(\hat{p})$$

com a *score* de la topologia T. El mètode tria la topologia que minimitza aquest valor. En la secció 6, veurem algunes simulacions realitzades amb aquest mètode.

4 La geometria del model K81

En aquesta secció, ens centrarem en el model K81 i en la seva geometria. Comencem amb una observació que implicarà una petita reducció en les dimensions del nostre espai ambient. Amb les notacions de la pàgina 154, si P és un punt sobre alguna de les varietats V_T^{K81}, és ben sabut que $p_{\mathsf{x}_1\ldots\mathsf{x}_n} = 0$ sempre que $\sum_i \mathsf{x}_i \neq \mathsf{A}$ (on aquesta suma té lloc al grup H introduït a la pàgina 152). Cadascuna d'aquestes equacions dóna lloc a un invariant lineal que no és filogenètic. Per tant, podem considerar que les varietats V_T^{K81} estan submergides en un espai afí de dimensió 4^{n-1}. Per exemple, si $T \in \mathcal{T}_4$, la varietat $V_T^{\mathsf{K81}} \subset \mathbb{C}^{64}$ té codimensió 48. D'altra banda, un conjunt minimal de generadors de l'ideal d'aquesta varietat consta de 8.002 polinomis de graus 2, 3 i 4 (vegeu [45]), que és un nombre molt gran en comparació amb el valor de la codimensió. El fet que aquesta diferència creixi en funció de n ens suggereix que un possible mètode de reconstrucció filogenètica basat en invariants que utilitzés un sistema complet de generadors de I_T seria impracticable en problemes que involucressin moltes fulles. D'altra banda, des del punt de vista de l'inferència filogenètica, no és necessari considerar un sistema complet de generadors de I_T, sinó només invariants que defineixin la varietat V_T en un obert que contingui les dades biològiques.

Aquest és el nostre punt de partida. En [9, teorema 3.7] provem el resultat següent:

3 TEOREMA *Sigui T un arbre amb n fulles.*

a) *La dimensió de la varietat $V_T^{\mathsf{K81}} \subset \mathbb{C}^{4^{n-1}}$ és $6n-8$.*

b) *Els punts de la varietat amb significat biològic*[5] *estan continguts en un obert dens de la varietat que pot ser definit per $4^{n-1}-6n+8$ invariants.*

La demostració de l'apartat *b*) passa per la comprovació que no hi ha punts singulars entre els punts de V_T^{K81} amb significat biològic (vegeu [9, corol·lari 3.12]). Llavors és suficient aplicar el resultat general segons el qual tota varietat és localment intersecció completa fora dels punts singulars.

En [9] presentem un algoritme recursiu que proporciona una col·lecció de polinomis que generen l'ideal de la varietat en un obert on la varietat és intersecció completa. Concretament, fixat un arbre $T \in \mathcal{T}_n$, l'algoritme genera un conjunt de

$$\begin{cases} 12(4^{n-3}-1) \text{ quàdriques} \\ 4^{n-2}-6n+20 \text{ quàrtiques.} \end{cases}$$

Sense entrar en els detalls de l'algoritme, val la pena notar que els polinomis obtinguts depenen d'una sèrie de tries relatives a la reconstrucció de l'arbre T a partir dels seus subarbres de 3 fulles. Tenint en compte aquest fet, pot ser interessant dissenyar estratègies per tal que les equacions resultants d'aplicar l'algoritme presentin el major nombre possible de simetries.

5 Vegeu [9] per a una definició precisa de punt amb *significat biològic*.

EXEMPLE En el cas de 4 fulles, l'algoritme genera 36 quàdriques (expressades en coordenades de Fourier)[6]

$$q_{CCCC}q_{AAAA} - q_{CCAA}q_{AACC}, \quad q_{GGCC}q_{AAAA} - q_{GGAA}q_{AACC}, \quad q_{TTCC}q_{AAAA} - q_{TTAA}q_{AACC},$$
$$q_{CCGG}q_{AAAA} - q_{CCAA}q_{AAGG}, \quad q_{GGGG}q_{AAAA} - q_{GGAA}q_{AAGG}, \quad q_{TTGG}q_{AAAA} - q_{TTAA}q_{AAGG},$$
$$q_{CCTT}q_{AAAA} - q_{CCAA}q_{AATT}, \quad q_{GGTT}q_{AAAA} - q_{GGAA}q_{AATT}, \quad q_{TTTT}q_{AAAA} - q_{TTAA}q_{AATT},$$
$$q_{ACAC}q_{CACA} - q_{ACCA}q_{CAAC}, \quad q_{GTAC}q_{CACA} - q_{GTCA}q_{CAAC}, \quad q_{TGAC}q_{CACA} - q_{TGCA}q_{CAAC},$$
$$q_{ACGT}q_{CACA} - q_{ACCA}q_{CAGT}, \quad q_{GTGT}q_{CACA} - q_{GTCA}q_{CAGT}, \quad q_{TGGT}q_{CACA} - q_{TGCA}q_{CAGT},$$
$$q_{ACTG}q_{CACA} - q_{ACCA}q_{CATG}, \quad q_{GTTG}q_{CACA} - q_{GTCA}q_{CATG}, \quad q_{TGTG}q_{CACA} - q_{TGCA}q_{CATG},$$
$$q_{AGAG}q_{GAGA} - q_{AGGA}q_{GAAG}, \quad q_{CTAG}q_{GAGA} - q_{CTGA}q_{GAAG}, \quad q_{TCAG}q_{GAGA} - q_{TCGA}q_{GAAG},$$
$$q_{AGCT}q_{GAGA} - q_{AGGA}q_{GACT}, \quad q_{CTCT}q_{GAGA} - q_{CTGA}q_{GACT}, \quad q_{TCCT}q_{GAGA} - q_{TCGA}q_{GACT},$$
$$q_{AGTC}q_{GAGA} - q_{AGGA}q_{GATC}, \quad q_{CTTC}q_{GAGA} - q_{CTGA}q_{GATC}, \quad q_{TCTC}q_{GAGA} - q_{TCGA}q_{GATC},$$
$$q_{ATAT}q_{TATA} - q_{ATTA}q_{TAAT}, \quad q_{CGAT}q_{TATA} - q_{CGTA}q_{TAAT}, \quad q_{GCAT}q_{TATA} - q_{GCTA}q_{TAAT},$$
$$q_{ATCG}q_{TATA} - q_{ATTA}q_{TACG}, \quad q_{CGCG}q_{TATA} - q_{CGTA}q_{TACG}, \quad q_{GCCG}q_{TATA} - q_{GCTA}q_{TACG},$$
$$q_{ATGC}q_{TATA} - q_{ATTA}q_{TAGC}, \quad q_{CGGC}q_{TATA} - q_{CGTA}q_{TAGC}, \quad q_{GCGC}q_{TATA} - q_{GCTA}q_{TAGC},$$

i 12 quàrtiques

$$q_{AAAA}q_{TTAA}q_{CGTA}q_{GCTA} - q_{CCAA}q_{GGAA}q_{ATTA}q_{TATA}, \quad q_{CACA}q_{TGCA}q_{ATTA}q_{GCTA} - q_{ACCA}q_{GTCA}q_{CGTA}q_{TATA},$$
$$q_{GGAA}q_{TTAA}q_{ACCA}q_{CACA} - q_{AAAA}q_{CCAA}q_{GTCA}q_{TGCA}, \quad q_{CCAA}q_{TTAA}q_{AGGA}q_{GAGA} - q_{AAAA}q_{GGAA}q_{CTGA}q_{TCGA},$$
$$q_{ACCA}q_{TGCA}q_{CTGA}q_{GAGA} - q_{CACA}q_{GTCA}q_{AGGA}q_{TCGA}, \quad q_{GAGA}q_{TCGA}q_{ATTA}q_{CGTA} - q_{AGGA}q_{CTGA}q_{GCTA}q_{TATA},$$
$$q_{AAAA}q_{AATT}q_{TACG}q_{TAGC} - q_{AACC}q_{AAGG}q_{TAAT}q_{TATA}, \quad q_{CACA}q_{CATG}q_{TAAT}q_{TAGC} - q_{CAAC}q_{CAGT}q_{TACG}q_{TATA},$$
$$q_{AAGG}q_{AATT}q_{CAAC}q_{CACA} - q_{AAAA}q_{AACC}q_{CAGT}q_{CATG}, \quad q_{AACC}q_{AATT}q_{GAAG}q_{GAGA} - q_{AAAA}q_{AAGG}q_{GACT}q_{GATC},$$
$$q_{CAAC}q_{CATG}q_{GACT}q_{GAGA} - q_{CACA}q_{CAGT}q_{GAAG}q_{GATC}, \quad q_{GAGA}q_{GATC}q_{TAAT}q_{TACG} - q_{GAAG}q_{GACT}q_{TAGC}q_{TATA}.$$

que defineixen una intersecció completa local que coincideix amb la varietat $V_{T_1}^{\mathsf{K81}}$ associada a l'arbre T_1 de la figura 6 en els punts amb significat biològic.

En vista dels resultats d'aquesta secció i en relació amb el mètode per a 4 espècies esmentat al final de la secció 3, és suficient avaluar els 48 invariants allistats a l'exemple 2 en comptes de treballar amb els 8.002 invariants necessaris per a generar l'ideal. Per a un nombre arbitrari d'espècies n, caldrà avaluar tants invariants com indiqui la codimensió de la varietat corresponent, és a dir, $4^{n-1} - 6n + 8$. Això implica una millora important en l'eficiència del mètode (vegeu les simulacions corresponents a aquesta millora del mètode en la secció 6).

5 Models evolutius equivariants

Des del punt de vista de la reconstrucció filogenètica i fixada una topologia d'arbre T, estem especialment interessats en els invariants *filogenètics* de T (*i. e.*, invariants específics de T) i no tant en els elements de l'ideal I_T en general. L'objectiu d'aquesta secció és, doncs, obtenir un conjunt (el més petit

6 Sota el model K81, la transformada discreta de Fourier indueix canvis lineals de coordenades en l'espai de paràmetres i alhora en $\mathbb{C}^{4^n}$. En els nous sistemes de coordenades, que anomenem paràmetres i coordenades de Fourier respectivament, la parametrització φ_T de (3.1) és monomial (vegeu [21]).

possible) d'invariants filogenètics que ens permetin determinar la topologia correcta per a un conjunt observat de dades. Aquests invariants vindran donats per condicions de rang de certes matrius que introduïm a continuació.

5.1 Biparticions, *splits* i *flattenings*

Fixat un conjunt qualsevol S, una *bipartició* de S és una descomposició $S = A \cup B$ on A i B són disjunts. Aquesta descomposició la denotem per $A \mid B$. Donat un arbre filogenètic T, és clar que cada branca interior indueix una bipartició del conjunt de les fulles: n'hi ha prou amb prendre les fulles dels dos subarbres de T que obtenim en eliminar la branca (vegeu la figura 7). D'aquesta bipartició en diem l'*split de T associat a la branca e*. Així, una bipartició de $L(T)$ és un *split* de T o no en funció de la topologia de T. El teorema clàssic de Buneman ens diu que podem recuperar la topologia d'un arbre a partir del conjunt dels seus *splits* (vegeu [6]). Donat un alineament (que suposem obtingut a partir d'un arbre T) i una bipartició del conjunt d'espècies $L(T)$, l'objectiu ara és assignar-hi condicions (algebraiques) de manera que ens permetin reconèixer si la bipartició prové d'un *split* de T. Per fer-ho, necessitem recordar el concepte de *flattening*.

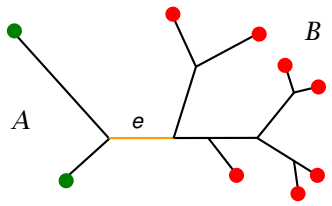

FIGURA 7: En eliminar la branca e obtenim de manera natural una bipartició $A \mid B$ de les fulles de l'arbre T. En diem l'*split associat a la branca e*.

Comencem amb un exemple senzill. Sigui T un arbre amb 4 fulles $\{v_1, v_2, v_3, v_4\}$ i considerem la bipartició de $L(T)$ donada per $A = \{v_1, v_2\}$, $B = \{v_3, v_4\}$. Si $P = (p_{XYZT})_{X,Y,Z,T \in \Sigma}$ és un punt a $\mathbb{C}^{4^4}$, el *flattening de P relatiu a $A \mid B$* és la matriu $4^2 \times 4^2$ obtinguda en ordenar les coordenades de P de manera que les files i les columnes quedin indexades per les obsevacions possibles a les fulles de A i de B, respectivament:

$$Flat_{A|B}(P) = \begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & \dots & p_{AATT} \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & \dots & p_{ACTT} \\ p_{AGAA} & p_{AGAC} & p_{AGAG} & \dots & p_{AGTT} \\ \dots & \dots & \dots & \dots & \dots \\ p_{TTAA} & p_{TTAC} & p_{TTAG} & \dots & p_{TTTT} \end{pmatrix}.$$

En general, per a un nombre arbitrari de fulles, donada una bipartició $A \mid B$ de $L(T)$, denotem per a i b els cardinals de A i B, respectivament (de manera que

$n = a + b$). Si P és un punt a $\mathbb{C}^{4^n}$, el *flattening de P relatiu a $A \mid B$* és la matriu $4^a \times 4^b$

$$Flat_{A|B}(P)$$

obtinguda en ordenar les coordenades de P en files i columnes d'acord amb aquesta bipartició.

A continuació, recordarem breument alguns fets bàsics de la teoria de representacions de grups. La utilització de la representació de grups en el context filogenètic és una de les grans aportacions del treball de J. Draisma i J. Kuttler a [16].

5.2 Representacions de grups

La teoria de representacions de grups és una branca important de l'àlgebra que permet reduir un bon nombre de problemes de teoria de grups a problemes més senzills d'àlgebra lineal. És una àrea amb aplicacions importants en altres ciències, com per exemple la física, on permet descriure com afecta el grup de simetries d'un sistema físic a les solucions d'un conjunt d'equacions que descriuen el sistema. En el context filogenètic, ens permetrà tractar tota una sèrie de models algebraics a la vegada sense haver d'entrar en les particularitats de cadascun d'aquests.

Donat un grup finit G, les seves representacions descriuen G com el grup de transformacions lineals d'un espai vectorial. Més precisament, una *representació* (lineal) de G és un morfisme de grups $\rho : G \to \mathrm{GL}(V)$ on V és un espai vectorial de dimensió finita. Denotem per V^G el subespai dels vectors invariants per l'acció de G. Diem que la representació és *irreductible* si V no conté cap subespai propi i invariant sota l'acció de G. Donades dues representacions de G, $\rho_i : G \to \mathrm{GL}(V_i), i = 1, 2$, un morfisme *$G$-equivariant* $f : V_1 \to V_2$ és una aplicació lineal compatible amb l'acció de G: $f \circ \rho_1(g) = \rho_2(g) \circ f$ per a tot $g \in G$. Denotem per $\mathrm{Hom}_G(V_1, V_2)$ el grup de morfismes G-equivariants de V_1 a V_2. Diem que ρ_1 i ρ_2 són *equivalents* si existeix un isomorfisme G-equivariant $f : V_1 \to V_2$. És ben sabut que mòdul aquesta noció d'equivalència, tot grup té un nombre finit de representacions irreductibles (vegeu [43]).

Ens limitarem a considerar subgrups G del grup $\mathfrak{S}_4$ de permutacions de 4 elements. Denotem per $W = \langle \mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T} \rangle_{\mathbb{C}}$ l'espai vectorial sobre $\mathbb{C}$ generat pel conjunt (ordenat) $\Sigma = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, i considerem el producte escalar en W que fa d'aquest conjunt una base ortonormal. La representació *canònica* $\rho : \mathfrak{S}_4 \to \mathrm{GL}(W)$ ve donada per la permutació dels elements de Σ. Donat un subgrup $G \subset \mathfrak{S}_4$ i una topologia d'arbre T, el *model equivariant* associat a G i a T és el model evolutiu algebraic que resulta d'imposar que les matrius de substitució siguin elements de $\mathrm{Hom}_G(W, W)$ expressats en la base Σ. D'aquesta manera, pel model equivariant associat a G, l'espai de paràmetres corresponent a T és $\mathrm{Par}_T^G = \prod_{e \in E(T)} \mathrm{Hom}_G(W, W)$.

EXEMPLE Si prenem $G = \langle (\mathsf{AC})(\mathsf{GT}), (\mathsf{AG})(\mathsf{CT}) \rangle$, és fàcil veure que el model resultant és el model K81 estudiat en la secció anterior:

$$S = \begin{array}{c} \\ \mathsf{A} \\ \mathsf{C} \\ \mathsf{G} \\ \mathsf{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{A} & \mathsf{C} & \mathsf{G} & \mathsf{T} \end{array} \\ \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix} \end{array}.$$

Per a $G = \mathfrak{S}_4$, les matrius invariants per l'acció del grup són les matrius de Jukes-Cantor. Observem que com més gran és el grup $G \subset \mathfrak{S}_4$ més simetries presenta el model equivariant corresponent i, en canvi, com més petit és G, més paràmetres té el model. En el cas extrem $G = \{id\}$, qualsevol matriu serà trivialment invariant per l'acció del grup i obtenim el model general de Markov GMM (vegeu [10] per a més detalls).

5.3 Invariants de branca

Fent ús del llenguatge de tensors, podem reescriure alguns dels conceptes introduïts en la secció 3. Per exemple, l'isomorfisme

$$\mathrm{Hom}_G(V_1, V_2) \cong (V_1 \otimes V_2)^G$$

ens permet pensar l'espai de paràmetres Par_T^G com un espai de tensors. A partir d'ara, identificarem els punts P de $\mathbb{C}^{4^n}$ amb tensors φ_P de $\otimes^n W$. Si denotem $\otimes_A W = \otimes_{i=1}^a W$ i $\otimes_B W = \otimes_{i=1}^b W$ els espais d'observacions possibles en A i en B respectivament, el *flattening* de P relatiu a $A \mid B$ correspon a la imatge de P via l'isomorfisme

$$\otimes^n W \cong \mathrm{Hom}(\otimes_A W, \otimes_B W), \tag{5.1}$$

i la identificació d'aquest espai amb l'espai de matrius $4^a \times 4^b$ (fent servir les bases $\{\mathsf{x}_1 \otimes \cdots \otimes \mathsf{x}_a \mid \mathsf{x}_i \in \Sigma\}$ de $\otimes_A W$ i $\{\mathsf{x}_1 \otimes \cdots \otimes \mathsf{x}_b \mid \mathsf{x}_i \in \Sigma\}$ de $\otimes_B W$). Observem que la representació canònica de G sobre W indueix una representació sobre les potències tensorials $\otimes^j W$ i que, per a una topologia d'arbre fixada T, els tensors sobre la varietat $V_T^G = \overline{\{\varphi_T(S) \mid S \in \mathrm{Par}_T^G\}} \subset \otimes^n W$ són invariants per aquesta acció, de manera que $V_T^G \subset (\otimes^n W)^G$.

En aquesta exposició, només necessitarem alguns resultats elementals de la teoria de representacions. El *teorema de Maschke* de descomposició en components isotípics (vegeu [43, Ch. 2]) implica que per a cada $j \geq 1$ existeix una descomposició única

$$\otimes^j W \cong \oplus_{i=1}^s N_i \otimes \mathbb{C}^{n_i(j)}, \quad n_i(j) \geq 0 \tag{5.2}$$

on $N_1, N_2, \ldots, N_s$ són les representacions irreductibles de G. Quan $j = 1$, dels enters positius $n_i := n_i(1)$ en direm les *multiplicitats* de W. Donada una topologia d'arbre $T \in \mathcal{T}_n$, una bipartició $A \mid B$ de $L(T)$ i mantenint la notació ja

introduïda, la descomposició (5.2) juntament amb el *lema de Schur* de teoria de representacions (vegeu [43]) impliquen que

$$(\otimes^n W)^G \cong \oplus_{i=1}^s \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}^{n_i(a)}, \mathbb{C}^{n_i(b)}).$$

Aquest isomorfisme ens permet «descompondre» cada tensor $\varphi \in (\otimes^n W)^G$ en una col·lecció $(f_1, f_2, \ldots, f_s)$ d'aplicacions lineals, $f_i : \mathbb{C}^{n_i(a)} \to \mathbb{C}^{n_i(b)}$. D'aquesta col·lecció en diem el *thin flattening de φ relatiu a $A \mid B$* i la denotem per $Tf_{A|B}(\varphi)$.

El resultat clau demostrat a [10] tradueix en termes d'equacions algebraiques el fet que una bipartició donada $A \mid B$ de $L(T)$ sigui un *split* de T. Aquestes equacions resulten d'imposar que els rangs de les aplicacions lineals a $Tf_{A|B}(\varphi) = (f_1, \ldots, f_s)$ no siguin més grans que les multiplicitats corresponents de W. Amb precisió,

4 Teorema ([10]) *Sigui $A \mid B$ una bipartició de $L(T)$ i $P \in V_T^G$ un punt genèric. Amb les notacions ja introduïdes, $A \mid B$ és un* split *de T si i només si*

$$\mathrm{rang}\,(f_i) \leq n_i, \quad i = 1, \ldots, s.$$

Aquest resultat corregeix i generalitza a qualsevol model equivariant el teorema 19.5 de [19], enunciat per al model general de Markov ($G = \{id\}$). Si $A \mid B$ és l'*split* induït per una branca $e \in E(T)$ i $Tf_{A|B}\varphi = (f_1, \ldots, f_s)$, definim els *invariants de la branca e* com la col·lecció d'equacions Z_e donada pels menors d'ordre $n_i + 1$ de les matrius associades a f_i, $i = 1, 2, \ldots, s$.

5 Teorema ([10]) *Sota un model equivariant, els invariants de branca són invariants filogenètics.*

A més, per a cada topologia $T \in \mathcal{T}_n$, existeix un obert dens $U_T \subseteq V_T$ de manera que, si $p \in \bigcup_{T \in \mathcal{T}_n} U_T$, llavors p pertany a V_{T_0} si i només si p és anul·lat per tots els invariants de branca de T_0.

Així doncs, fixada una topologia T_0 i un punt $P \in \bigcup_T V_T$ genèric, per saber si $P \in V_{T_0}$ és suficient avaluar els invariants donats per les condicions de branca de la topologia T_0.

Exemple Considerem el subgrup $G = \langle (\mathsf{AC})(\mathsf{GT}), (\mathsf{AG})(\mathsf{CT}) \rangle$, corresponent al model K81. Per aquest grup, hi ha 4 representacions irreductibles $\{N_i\}_{i=1,2,3,4}$, i la descomposició en components isotípics de W és $W \cong N_1 \oplus N_2 \oplus N_3 \oplus N_4$. Així doncs, $n_j = 1$ per a $j = 1, 2, 3, 4$. Si $Tf_{A|B}(\varphi) = (f_1, f_2, f_3, f_4)$ i expressem la matriu de cada f_j en les coordenades de Fourier, obtenim

$$S_1 = \begin{pmatrix} q_{\mathsf{AAAA}} & q_{\mathsf{AACC}} & q_{\mathsf{AAGG}} & q_{\mathsf{AATT}} \\ q_{\mathsf{CCAA}} & q_{\mathsf{CCCC}} & q_{\mathsf{CCGG}} & q_{\mathsf{CCTT}} \\ q_{\mathsf{GGAA}} & q_{\mathsf{GGCC}} & q_{\mathsf{GGGG}} & q_{\mathsf{GGTT}} \\ q_{\mathsf{TTAA}} & q_{\mathsf{TTCC}} & q_{\mathsf{TTGG}} & q_{\mathsf{TTTT}} \end{pmatrix} \quad S_2 = \begin{pmatrix} q_{\mathsf{ACAC}} & q_{\mathsf{AACA}} & q_{\mathsf{AAGT}} & q_{\mathsf{AATG}} \\ q_{\mathsf{CAAC}} & q_{\mathsf{CACA}} & q_{\mathsf{CAGT}} & q_{\mathsf{CATG}} \\ q_{\mathsf{GTAC}} & q_{\mathsf{GTCA}} & q_{\mathsf{GTGT}} & q_{\mathsf{GTTG}} \\ q_{\mathsf{TGAC}} & q_{\mathsf{TGCA}} & q_{\mathsf{TGGT}} & q_{\mathsf{TGTG}} \end{pmatrix}.$$

$$S_3 = \begin{pmatrix} q_{\mathtt{AGAG}} & q_{\mathtt{AGCT}} & q_{\mathtt{AGGA}} & q_{\mathtt{AGTC}} \\ q_{\mathtt{CTAG}} & q_{\mathtt{CTCT}} & q_{\mathtt{CTGA}} & q_{\mathtt{CTTC}} \\ q_{\mathtt{GAAG}} & q_{\mathtt{GACT}} & q_{\mathtt{GAGA}} & q_{\mathtt{GATC}} \\ q_{\mathtt{TCAG}} & q_{\mathtt{TCCT}} & q_{\mathtt{TCGA}} & q_{\mathtt{TCTC}} \end{pmatrix} \quad S_4 = \begin{pmatrix} q_{\mathtt{ATAT}} & q_{\mathtt{ATCG}} & q_{\mathtt{ATGC}} & q_{\mathtt{ATTA}} \\ q_{\mathtt{CGAT}} & q_{\mathtt{CGCG}} & q_{\mathtt{CGGC}} & q_{\mathtt{CGTA}} \\ q_{\mathtt{GCAT}} & q_{\mathtt{GCCG}} & q_{\mathtt{GCGC}} & q_{\mathtt{GCTA}} \\ q_{\mathtt{TAAT}} & q_{\mathtt{TACG}} & q_{\mathtt{TAGC}} & q_{\mathtt{TATA}} \end{pmatrix}$$

En aquest cas, els invariants de branca són els menors d'ordre 2 d'aquestes matrius. En un punt genèric de $V_T^{\mathtt{K81}}$, són suficients 36 d'aquests menors per garantir que el rang d'aquestes matrius és menor o igual que 1, 9 menors per a cada matriu. Per exemple, podem prendre les quàdriques de l'exemple de la pàgina 159. El teorema 4 ens diu que n'hi ha prou amb avaluar aquestes quàdriques per determinar la topologia de l'arbre correcte.

6 Algunes simulacions

En aquesta secció presentem algunes simulacions realitzades per a estudiar l'eficiència del mètode presentat al final de la secció 3. Hem generat els alineaments sota el model `K81` fent servir el programari `SeqGen` (vegeu [40]). Les matrius de substitució adopten la forma $S = e^{Qt}$, on t és la longitud de branca i Q és la matriu de taxes, que pren la forma

$$Q = \begin{pmatrix} \cdot & \gamma & \alpha & \beta \\ \gamma & \cdot & \beta & \alpha \\ \alpha & \beta & \cdot & \gamma \\ \beta & \alpha & \gamma & \cdot \end{pmatrix} \qquad \text{on } \cdot = -\alpha - \beta - \gamma.$$

Fem servir l'aproximació de Huelsenbeck donada a [31] per a arbres amb 4 fulles: considerem dos paràmetres a i b que representen certes longituds de branca de l'arbre i simulem dades per a cada parell (a, b). El paràmetre a assigna la longitud de la branca interior i de dues branques exteriors, mentre que el paràmetre b assigna la longitud de les altres dues branques (vegeu la figura 8). Tots dos paràmetres a i b varien des de $0,01$ fins a $0,75$ en increments de 0,02, i simulem 1.000 alineaments per a cada parell (a, b).

La figura 9 mostra els resultats obtinguts pel mètode considerant alineaments de longituds 100, 500 i 1.000 que evolucionen sota una matriu de taxes Q amb paràmetres $\gamma = 0,1$, $\alpha = 3,0$ i $\beta = 0,5$. A la primera fila de la figura apareixen els resultats quan hem fet servir els 8.002 generadors de l'ideal (versió original del mètode, proposada a [11]), mentre que a la segona i a la tercera fila hem utilitzat els 48 invariants que apareixen a l'exemple 2 i les 36 quàdriques corresponents als invariants de l'*split* dels arbres, respectivament. Val la pena observar que quan només fem servir 48 o 36 invariants (segona i tercera fila de la figura 9) el mètode funciona millor que quan fem servir un sistema complet de generadors (a més de ser notablement més ràpid). Una explicació possible és que tots els invariants de més que avaluem aporten un *soroll* que no ajuda a l'hora d'inferir l'arbre correcte.

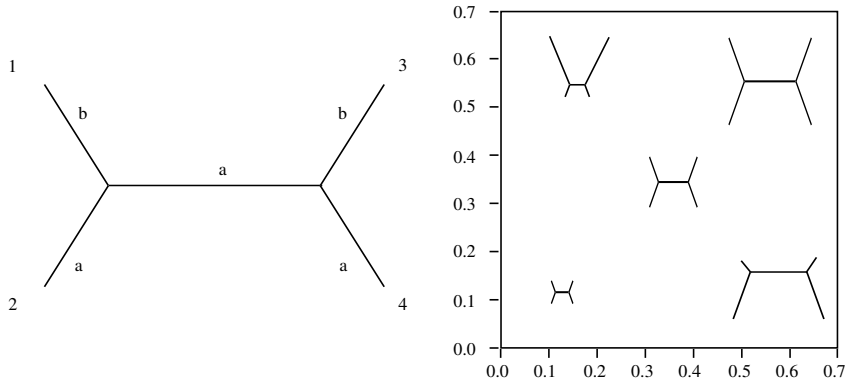

FIGURA 8: El paràmetre a assigna la longitud de la branca interior i de dues branques exteriors oposades; el paràmetre b assigna la longitud de les altres dues branques. En les simulacions realitzades, tots dos paràmetres varien des de 0,01 fins 0,75 en increments de 0,02.

Aquestes gràfiques es poden comparar també amb [31, figura A2]. Tot i que aquesta comparació està esbiaixada, cal assenyalar que, fins i tot amb dades homogènies, el mètode dels invariants amb 8.002 generadors es comporta millor que molts dels mètodes considerats en [31]. El mètode funciona clarament millor que el mètode de Lake, que només treballa amb invariants lineals. I per a alineaments de longitud a partir de 500, els resultats obtinguts són millors que el *Neighbor-Joining* (vegeu [31, figura A2]).

Per a longitud 1.000, l'eficiència del mètode dels invariants és similar a l'obtinguda per a longitud 10.000 per molts altres mètodes estudiats a [31]. Veiem, doncs, que per a reconstruir l'arbre correcte, són necessàries moltes menys dades en el mètode dels invariants que en molts altres mètodes (cf. [29]).

El lector interessat pot trobar a [8] més estudis comparatius del mètode dels invariants presentats aquí amb els mètodes de *Neighbor-Joining* [41] i *Màxim de versemblança* [22] per a dades homogènies (vegeu també [11, 7]).

Agraïments

Vull expressar el meu agraïment a la Societat Catalana de Matemàtiques per haver-me convidat a participar activament en la XII Trobada Matemàtica.

Aquest treball ha estat realitzat amb el suport del projecte d'investigació MTM2009-14163-C02-02 del Ministeri de Ciència i Innovació i del projecte 20009SGR-1284 de la Generalitat de Catalunya.

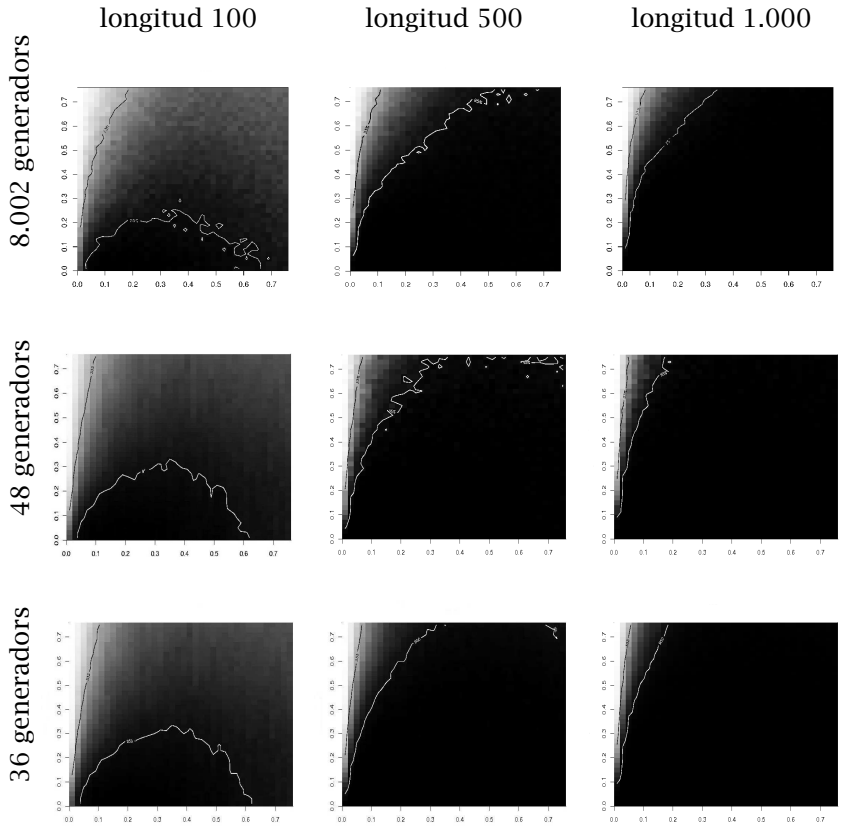

FIGURA 9: Aquestes nou gràfiques representen l'èxit obtingut a l'hora de reconstruir l'arbre correcte en l'espai de paràmetres de la figura 8. Les regions de color negre corresponen a parells (a, b) per als quals l'arbre ha estat correctament estimat el 100 % de les vegades, mentre que les regions blanques corresponen als valors (a, b) per als quals el mètode no ha predit mai l'arbre correcte; les tonalitats de color gris indiquen àrees de probabilitat intermèdia. La isoclínica del 95 % és dibuixada en blanc, mentre que la del 33 % és dibuixada en negre. Les dades han estat generades sota el model homogeni de `K81` fent servir una matriu de taxes amb paràmetres $\gamma = 0,1$, $\alpha = 3,0$ i $\beta = 0,5$. Agrupades per files, de dalt a baix, les gràfiques mostren els resultats obtinguts fent servir un sistema complet de generadors de l'ideal (algoritme original), els 48 invariants llistats en l'exemple de la secció 4 i les 36 quàdriques de l'exemple de la secció 5.

Referències

[1] Al-Aidroos, J.; Snir, S. «Analysis of point mutations in vertebrate genomes». A: *Algebraic statistics for computational biology*. Nova York: Cambridge Univ. Press, 2005, 375–386.

[2] Allman, E.; Rhodes, J. «Quartets and parameter recovery for the general Markov model of sequence mutation». *AMRX Appl. Math. Res. Express*, 4 (2004), 107–131.

[3] Allman, E.; Rhodes, J. «The identifiability of tree topology for phylogenetic models, including covarion and mixture models». *J. Comput. Biol.*, 13 (2006), 1101–1113.

[4] Allman, E.; Rhodes, J. «Phylogenetic ideals and varieties for the general Markov model». *Adv. in Appl. Math.*, 40 (2007), 127–148.

[5] Buczyńska, W.; Wiśniewski, J. A. «On geometry of binary symmetric models of phylogenetic trees». *J. Eur. Math. Soc. (JEMS)*, 9 (3) (2007), 609–635.

[6] Buneman, P. «The recovery of trees from measures of dissimilarity». A: *Mathematics in the Archaeological and Historical Sciences*. Edinburgh: Edinburgh University Press, 1971, 387–395.

[7] Casanellas, M. «Models algebraics en filogenètica». *Butl. Soc. Catalana Mat.*, 21 (2) (2006), 213–228, 248 (2007).

[8] Casanellas, M.; Fernández-Sánchez, J. «Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees». *Mol. Biol. Evol.*, 24 (1) (2007), 288–293.

[9] Casanellas, M.; Fernández-Sánchez, J. «Geometry of the Kimura 3-parameter model». *Adv. in Appl. Math.*, 41 (2008), 265–292.

[10] Casanellas, M.; Fernández-Sánchez, J. «Relevant phylogenetic invariants of evolutionary models». (Apareixerà al *J. Math. Pures Appl.* (9))

[11] Casanellas, M.; Garcia, L.; Sullivant, S. «Catalog of small trees». A: Pachter, L.; Sturmfels, B. [ed.]. *Algebraic statistics for computational biology*, cap. 15. Cambridge: Cambridge Univ. Press, 2005.

[12] Casanellas, M.; Sullivant, S. «The strand symmetric model». A: Pachter, L.; Sturmfels, B. [ed.]. *Algebraic statistics for computational biology*, cap. 16. Cambridge: Cambridge Univ. Press, 2005.

[13] Cavender, J.; Felsenstein, J. «Invariants of phylogenies in a simple case with discrete states». *J. Classification*, 4 (1987), 57–71.

[14] Consortium, E. P. «The encode (encyclopedia of DNA elements) project». *Science*, 306 (2004), 636–640.

[15] Cox, D.; Sidman, J. «Secant varieties of toric varieties». *J. Pure Appl. Algebra*, 209 (3) (2007), 651–669.

[16] Draisma, J.; Kuttler, J. «On the ideals of equivariants tree models». *Math. Ann.*, 344 (2009), 619–644.

[17] Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. *Biological sequence analysis*. Cambridge: Cambridge Univ. Press, 1998. (ISBN 0-521-52586-1)

[18] Elizalde, S., «Combinatorics and biology: inference functions and sequence alignment». *Butl. Soc. Catalana Mat.*, 21 (1) (2006), 39–52, 157–158.

[19] Eriksson, N. «Tree construction using singular value decomposition». A: Pachter, L.; Sturmfels, B. [ed.]. *Algebraic Statistics for computational biology*, cap. 19. Cambridge: Cambridge University Press, 2005, 347–358.

[20] Eriksson, N. «Using invariants for phylogenetic tree construction». A: *Emerging applications of algebraic geometry*. Nova York: Springer, 2009. IMA Vol. Math. Appl., 149, 89–108.

[21] Evans, S.; Speed, T. «Invariants of some probability models used in phylogenetic inference». *Ann. Statist.*, 21 (1993), 355–377.

[22] Felsenstein, J. «Evolutionary trees from DNA sequences: a maximum likelihood approach». *J. Mol. Evol.*, 17 (1981), 368–376.

[23] Felsenstein, J. «Phylip – phylogeny inference package (version 3.6)». *Cladistics*, 5.

[24] Felsenstein, J. *Inferring phylogenies*. Sinauer Associates, Inc., 2003.

[25] Ferreti, V.; Sankoff, D. «Phylogenetic invariants for more general evolutionary models». *J. theor. Biol.*, 147–162.

[26] Fitch, W. M. «Toward defining the course of evolution: minimum change for a specific tree topology». *Syst. Zool.*, 20 (1971), 406–416.

[27] Galtier, N.; Gouy, M. «Inferring pattern and process: maximum likelihood implementation of a non-homogeneous model of DNA sequence evolution for phylogenetic analysis.» *Mol. Biol. Evol.*, 154 (4) (1998), 871–879.

[28] Gascuel, O.; Guindon, S. «Modelling the variability of evolutionary processes». A: *Reconstructing evolution*. Oxford: Oxford Univ. Press, 2007, 65–107.

[29] Hagedorn, T.; Landweber, L. «Phylogenetic invariants and geometry». *J. theor. Biol.*, 205 (2000), 365–376.

[30] Harris, J. *Algebraic geometry*. Nova York: Springer, 1995. Un primer curs, reimpressió corregida de l'original de 1992. (Graduate Texts in Mathematics, 133)

[31] Huelsenbeck, J. «Performance of phylogentic methods in simulation». *Syst. Biol.*, 44 (1995), 17–48.

[32] Jin, L.; Nei, M. «Limitations of the evolutionary parsimony method of phylogenetic analysis.» *Mol. Biol. Evol.*, 7 (1990), 82–102.

[33] Jukes, T.; Cantor, C. «Evolution of protein molecules.» *In Mammalian Protein Metabolism*, (1969), 21–132.

[34] Kimura, M. «A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences». *J. Mol. Evol.*, 16 (1980), 111–120.

[35] KIMURA, M. «Estimation of evolutionary sequences between homologous nucleotide sequences». *Proc. Nat. Acad. Sci., USA*, 78 (1981), 454–458.

[36] LAKE, J. «A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony». *Mol. Biol. Evol.*, 4 (1987), 167–191.

[37] MATSEN, F. «Phylogenetic mixtures on a single tree can mimic a tree of another topology». *Systematic Biology*, 56 (2007), 767–775.

[38] NORRIS, J. R. *Markov chains*, Cambridge: Cambridge Univ. Press, 1998. (Camb. Ser. Stat. Probab. Math., 2. Reimpressió de l'original de 1997)

[39] PACHTER, L.; STURMFELS, B. [ed.]. *Algebraic Statistics for computational biology*. Cambridge: Cambride Univ. Press, 2005. ISBN 0-521-85700-7.

[40] RAMBAUT, A.; GRASSLY, N. «Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees». *Comput. Appl. Biosci.*, 13 (1997), 235–238.

[41] SAITOU, N.; NEI, M. «The neighbor joining method: a new method for reconstructing phylogenetic trees». *Mol. Biol. Evol.*, 4 (4) (1987), 406–425.

[42] SANKOFF, D.; BLANCHETTE, M. «Phylogenetic invariants for genome rearrangements». *J. Comput. Biol.*, 6 (1999), 431–445.

[43] SERRE, J. *Linear representations of finite groups*. Nova York: Springer, 1977. [Traduït de la segona edició francesa per Leonard L. Scott.] (Graduate Texts in Mathematics, v. 42)

[44] STEEL, M.; SZÉKELY, L.; ERDŐS, P.; WADDELL, P. «A complete family of phylogenetic invariants for any number of taxa under Kimura's 3st model». *N. Z. J. Bot.*, 31 (1993), 289–296.

[45] STURMFELS, B.; SULLIVANT, S. «Toric ideals of phylogenetic invariants». *J. Comput. Biol.*, 12 (2005), 204–228.

[46] YANG, Z.; ROBERTS, D. «On the use of nucleic acid sequences to infer early branchings in the tree of life». *Mol. Biol. Evol.*, 12 (1995), 451–458.

[47] YANG, Z.; YODER, A. D. «Estimation of the transition/transversion rate bias and species sampling». *J. Mol. Evol.*, 48 (1999), 274–283.

[48] ZUCKERKANDL, E.; PAULING, L. «Molecular disease, evolution and genetic heterogeneity». *Horizons in Biochemistry*, (1962), 189–225.

DEPARTAMENT DE MATEMÀTICA APLICADA I
UNIVERSITAT POLITÈCNICA DE CATALUNYA
AVINGUDA DIAGONAL, 647
08028 BARCELONA
jesus.fernandez.sanchez@upc.edu