

## Models algebraics en filogenètica\*

MARTA CASANELLAS

Mathematics is biology's next microscope,  
only better; biology is mathematics' next  
physics, only better.

Joel E. Cohen

**Resum** En aquest article fem una introducció a les aplicacions de la geometria algebraica en filogenètica. Gràcies al fet que gran part dels models evolutius usats en filogenètica corresponen a varietats algebraiques, l'ideal associat a aquestes varietats pot ser usat per a donar un nou enfocament a la inferència filogenètica.

**Paraules clau:** invariants filogenètics, model evolutiu, aplicació racional.

**Classificació MSC2000:** 14M25, 92D15, 92D20.

### 1 Introducció

En els darrers anys s'ha demostrat que dues disciplines tan distants com la geometria algebraica i la genètica comparativa poden beneficiar-se d'un treball conjunt. Aquestes dues àrees es poden connectar gràcies a l'estadística algebraica, que parteix del fet que un gran nombre de models estadístics usats en biologia corresponen a varietats algebraiques. L'estadística algebraica s'ha desenvolupat en els últims anys i s'ha aplicat a l'estudi de taules de contingència, disseny d'experiments... (vegeu [16]). Recentment, s'han unit esforços de biòlegs i geòmetres algebraics per incidir en les aplicacions de

---

\* Aquesta exposició correspon a la xerrada impartida per la mateixa autora a la Novena Trobada de la Societat Catalana de Matemàtiques, que va tenir lloc a Vic el 22 d'abril de 2006.

l'estadística algebraica a la biologia computacional. Els fruits d'aquest esforç els podem trobar, per exemple, al llibre *Algebraic statistics for computational biology* [15].

En aquest article pretenem fer una introducció a la relació entre la geometria algebraica i la genètica, o, més concretament, la filogenètica. Com veurem, l'ús de la geometria algebraica en la filogenètica pot aportar noves tècniques i nous resultats a aquesta branca de la biologia.

La filogenètica parteix de la base que tots els organismes de la Terra provenen d'un ancestre comú. Així, totes les espècies de la Terra tenen relacions ancestrals entre si que s'anomenen *filogènies*. Les filogènies es poden representar mitjançant un arbre denominat *arbre filogenètic*. En particular, la *filogenètica* es dedica a inferir aquests arbres a partir dels genomes<sup>1</sup> de les espècies que trobem actualment.

L'arbre filogenètic es dibuixa com un arbre invertit: les fulles de l'arbre representen les espècies actuals i les situem a baix de tot; l'arrel de l'arbre filogenètic es troba al node de dalt de tot i representa l'ancestre comú de totes les espècies que formen l'arbre, i els nodes interiors representen espècies ancestrals comunes de les espècies que se'n deriven. La longitud de les branques representa la *distància evolutiva* entre les espècies que uneix la branca o, dit d'una altra manera, el nombre de mutacions que s'han produït entre el node pare i el node fill.

Per exemple, si coneixem el genoma —o una part seva (normalment s'usen només els gens)— de l'espècie humana, del ximpanzé i del goril·la, podem preguntar-nos quin és l'arbre filogenètic d'aquestes espècies. Tenim tres possibilitats representades en la figura 1.

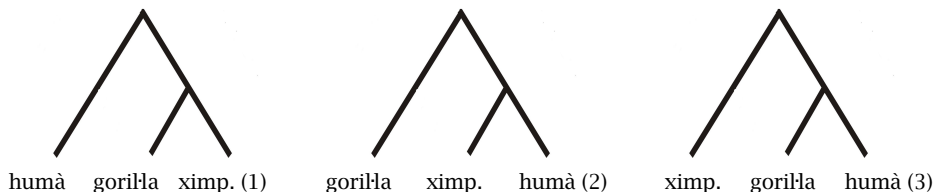


FIGURA 1: Els tres possibles arbres de tres fulles.

Fins als anys vuitanta es creia que l'arbre correcte era el primer, però, des que s'han fet estudis usant els genomes de les espècies, hi ha proves que l'arbre que ha donat lloc a l'espècie humana és el segon (vegeu [2]). En aquest article presentarem, des d'un punt de vista esbiaixat vers la geometria algebraica, el tipus d'estudis en filogenètica que han donat lloc a aquestes conclusions.

<sup>1</sup> El genoma d'una espècie eucariota és el contingut d'ADN que trobem en el nucli de les cèl·lules. Les unitats bàsiques que componen l'ADN són els *nucleòtids*. Hi ha quatre nucleòtids diferents: A denota *adenina*, C *citocina*, G *guanina* i T *timina*. En el nostre context, donar el genoma d'una espècie equival a donar una seqüència ordenada de caràcters en les lletres A, C, G, T.

En la secció següent farem una introducció als models estadístics més usats en filogenètica i n'estudiarem la relació amb la geometria algebraica. En la secció 3 presentarem els resultats de geometria algebraica que permeten donar un nou enfocament a la inferència filogenètica. Finalment, a la secció 4 donarem mètodes que demostren que a la pràctica sí que és factible usar les tècniques de geometria algebraica per a la reconstrucció d'arbres filogenètics.

## 2 Models algebraics en filogenètica

### 2.1 Models estadístics d'evolució

Per a representar el procés evolutiu de la formació d'espècies donarem un model estadístic basat en les mutacions que es produeixen en l'ADN, on suposarem que:

- a) Els arbres són *binaris*, és a dir, del node arrel en surten dues branques i cada branca es divideix en dues més fins a arribar a les fulles.
- b) L'evolució d'una espècie només depèn del node immediatament superior.
- c) Les mutacions ocorren aleatòriament i la probabilitat que es produeixi una mutació és sempre positiva.
- d) Suposarem donat un alineament de les seqüències d'ADN de les espècies. A causa de processos de mutació, duplicació i supressió de parts de l'ADN, les seqüències de les diferents espècies tenen parts idèntiques, parts que s'assemblen i parts que no es poden comparar. A més a més, les parts idèntiques o comparables no tenen cap raó per a trobar-se en el mateix lloc del genoma (els genomes de diferents espècies tenen diferents longituds i diferent nombre de cromosomes i de gens). És per això que abans d'estudiar les relacions entre les espècies ens interessa saber quines parts de l'ADN són comparables i quines parts dels genomes de les diferents espècies es corresponen entre si. Tot això es recull en un bon alineament de les seqüències d'ADN que volem considerar. En el nostre exemple, suposarem que partim de l'alineament següent de seqüències:

<i>humà</i>	AACCTTCGAGGCTTACCGCTG
<i>gorilla</i>	AACGTCTATGCTCACCGATG
<i>ximpanzé</i>	AAGGTCGATGCTCACCGATG

- e) Les diferents posicions de la cadena d'ADN evolucionen de la mateixa manera i independentment de les altres posicions.

Com que suposem que totes les posicions de la cadena d'ADN evolucionen de la mateixa manera, per a cada posició considerarem el mateix model evolutiu probabilístic que representarem a l'arbre de la manera següent. Enumerem els vèrtexs de l'arbre d'esquerra a dreta i de baix a dalt i anomenem  $t_i$  la longitud de la branca que puja des del vèrtex  $i$ . El node arrel s'anomenarà  $r$ . A cada

vèrtex  $i$  de l'arbre hi posem una variable aleatòria discreta que pren valors a  $\{A, C, G, T\}$ . Les variables aleatòries  $X_i$  a les fulles de l'arbre seran *variables observades* (perquè l'alineament ens dóna observacions del vector aleatori  $X = (X_1, X_2, X_3)$ ); les dels nodes interiors seran ocultes (perquè no en tindrem cap observació) i les anomenarem  $Y_i$ :

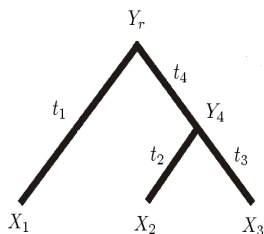


FIGURA 2: Model estadístic en un arbre de tres fulles.

Seguint un procés de Markov, a cada branca li associem una matriu  $S_i$  les entrades de la qual són les probabilitats  $P(x|y, t_i)$  que un nucleòtid  $y$  en el node pare muti a un nucleòtid  $x$  en el node fill al llarg d'una branca de longitud  $t_i$ :

$$S_i = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} P(A|A, t_i) & P(C|A, t_i) & P(G|A, t_i) & P(T|A, t_i) \\ P(A|C, t_i) & P(C|C, t_i) & P(G|C, t_i) & P(T|C, t_i) \\ P(A|G, t_i) & P(C|G, t_i) & P(G|G, t_i) & P(T|G, t_i) \\ P(A|T, t_i) & P(C|T, t_i) & P(G|T, t_i) & P(T|T, t_i) \end{pmatrix} \end{matrix}.$$

Aquestes probabilitats són desconegudes per a nosaltres i seran els *paràmetres* del model. Les matrius  $S_i$  s'anomenen *matrius de substitució*.

Suposarem que la distribució de nucleòtids en l'arrel  $\pi_A = P(Y_r = A)$ ,  $\pi_C = P(Y_r = C)$ ,  $\pi_G = P(Y_r = G)$ ,  $\pi_T = P(Y_r = T)$  és coneguda. La hipòtesi ( $\nu$ ) ens diu que la probabilitat que del node arrel de l'arbre (1) de la figura 1 s'hagi evolucionat a les seqüències d'humà, gorilla i ximpanzé donades en l'alineament és igual a

$$p_{AAA}^4 * p_{CCG} * p_{TGG} * p_{TTT}^3 * p_{CCC}^4 * p_{GTG} * p_{GTT} * p_{GGG}^3 * p_{TCC} * p_{CAA},$$

on  $p_{x_1x_2x_3} = P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ . Dit d'una altra manera, la hipòtesi ( $\nu$ ) ens diu que les columnes de l'alineament evolucionen de manera independent.

Si consideram el procés de Markov en l'arbre de la figura 2, la probabilitat d'observar els nucleòtids  $x_1, x_2, x_3$  a les fulles s'expressa en funció de les entrades de les matrius de substitució de la manera següent:

$$p_{x_1x_2x_3} = \sum_{y_r, y_4 \in \{A, C, G, T\}} \pi_{y_r} S_1(x_1, y_r) S_4(y_4, y_r) S_2(x_2, y_4) S_3(x_3, y_4). \quad (1)$$

Tenim diferents models estadístics d'evolució segons la forma que tinguin les matrius de substitució i segons la distribució de nucleòtids a l'arrel. Seguidament els presentem en la seva versió *algebraica* (vegeu la secció 2.2).

**Models de grup.** Els models més usats en filogenètica són els *models de grup* (a la secció següent veurem per què s'anomenen així). En aquests models la distribució de nucleòtids a l'arrel és uniforme i les matrius de substitució són de la forma

$$S_i = \begin{pmatrix} a_i & b_i & c_i & d_i \\ b_i & a_i & d_i & c_i \\ c_i & d_i & a_i & b_i \\ d_i & c_i & b_i & a_i \end{pmatrix},$$

on  $a_i + b_i + c_i + d_i = 1$ . La justificació biològica per a aquest model rau en el fet que, a causa de certes propietats químiques, l'adenina i la guanina són *purines* i la citosina i la timina són *pirimidines*. Aquest model (anomenat *Kimura 3-paràmetres* [13]) vol reflectir el fet que les *transicions* (és a dir, mutacions de purina a purina o pirimidina a pirimidina) són més freqüents que les *transversions* (mutacions de purina a pirimidina o a la inversa). Casos particulars d'aquest model són el *Kimura 2-paràmetres* [12] (només usa un paràmetre per a les transicions i un altre per a les transversions, és a dir,  $b_i = d_i$ ) i el Jukes-Cantor [11] (és el model més simple, ja que no distingeix entre transicions i transversions, és a dir,  $b_i = c_i = d_i$ ).

**Strand symmetric model.** En aquest model no se suposa que la distribució de nucleòtids sigui uniforme, sinó que  $\pi_A = \pi_T$  i  $\pi_C = \pi_G$ . Aquest model s'adapta més a la realitat de les seqüències que s'usen normalment per a inferir filogènies, que són *seqüències codificants* (gens o parts de gens). Les regions codificants contenen més C, G's que no pas A, T's i, a causa de la simetria de la doble cadena d'ADN, s'ha comprovat en diversos estudis que  $\pi_A \sim \pi_T$  i  $\pi_C \sim \pi_G$ . En aquest model suposarem que  $\pi_A = \pi_T$  i  $\pi_C = \pi_G$ . D'altra banda, si una A muta a una C, aleshores en la cadena d'ADN complementària es produeix una mutació de T a G perquè una A (respectivament C) sempre va enllaçada amb una T (respectivament G). Així es natural requerir que

$$\begin{aligned} P(A|A) &= P(T|T), & P(C|A) &= P(G|T), & P(G|A) &= P(C|T), \\ P(T|A) &= P(A|T), & P(A|C) &= P(T|G), & P(C|C) &= P(G|G), \\ P(G|C) &= P(C|G), & P(T|C) &= P(T|G). \end{aligned}$$

Aquest model només requereix aquestes igualtats i, per tant, les matrius de substitució són de la forma

$$S_i = \begin{pmatrix} a_i & b_i & c_i & d_i \\ e_i & f_i & g_i & h_i \\ h_i & g_i & f_i & e_i \\ d_i & c_i & b_i & a_i \end{pmatrix}.$$

Notem que els models de grup són un cas particular d'aquest. Aquest model algebraic va ser introduït a [6].

**Model de Markov general.** Aquest és el model més general possible. No es requereix res sobre la distribució a l'arrel i les matrius de substitució són genèriques:

$$S_i = \begin{pmatrix} a_i & b_i & c_i & d_i \\ e_i & f_i & g_i & h_i \\ j_i & k_i & l_i & m_i \\ n_i & o_i & p_i & q_i \end{pmatrix}.$$

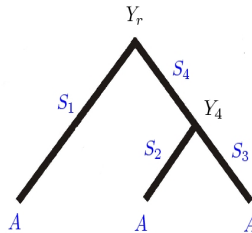
Un model amb més paràmetres sempre s'adequa més a la realitat, però augmentar el nombre de paràmetres pot augmentar moltíssim la complexitat dels càlculs necessaris per a inferir filogènies.

## 2.2 Invariants filogenètics

Els models evolutius que hem descrit són models *algebraics*. S'anomenen així perquè les probabilitats conjuntes de les variables observades s'expressen com a funció polinòmica en els paràmetres (vegeu l'equació (1)). Així, un model evolutiu algebraic de  $d$  paràmetres lliures sobre un arbre  $T$  de  $n$  fulles és descrit per l'aplicació polinomial següent:

$$\begin{aligned} \varphi : \mathbb{R}^d &\rightarrow \mathbb{R}^{4^n} \\ \theta = (\theta_1, \dots, \theta_d) &\mapsto (p_{AA\dots A}, p_{AA\dots C}, p_{AA\dots G}, \dots, p_{TT\dots T}) \end{aligned} \quad (2)$$

Per exemple, si a l'arbre següent considerem el model de Jukes-Cantor,



obtenim (substituint en l'equació (1)):

$$p_{AAA} = \frac{1}{4}(a_1 a_4 a_2 a_3 + 3b_1 b_4 a_2 a_3 + 3b_1 a_4 a_2 a_3 + 3a_1 b_4 a_2 a_3 + 6b_1 b_4 b_2 b_3), \quad (3)$$

on  $a_i + 3b_i = 1$ . Observem que, per a qualsevol dels models descrits,  $p_{x_1 x_2 x_3}$  és un polinomi homogeni de grau igual al nombre de branques de l'arbre.

En el cas del model de Jukes-Cantor, per a un arbre de tres fulles tenim definida una aplicació polinòmica

$$\begin{aligned} \varphi : \mathbb{R}^4 &\rightarrow \mathbb{R}^{64} \\ (a_1, a_2, a_3, a_4) &\mapsto (p_{AAA}, p_{AAC}, p_{AAG}, \dots, p_{TTT}). \end{aligned}$$

De fet, com que parlem de probabilitats, el model estadístic és definit a  $\Delta^1 \times \Delta^1 \times \Delta^1 \times \Delta^1$  i la imatge de l'aplicació polinomial ha de caure dins del simplex de dimensió 63,  $\Delta^{63}$ .

L'aplicació polinomial (2) parametriza un obert d'una *varietat algebraica*. Recordem que una varietat algebraica és el conjunt de punts que són solució d'un sistema d'equacions polinomials:  $V = Z(f_1, \dots, f_r)$ ,  $f_1, \dots, f_r \in k[x_1, \dots, x_n]$ , on  $k$  és un cos. La imatge d'una aplicació polinomial no és en general una varietat algebraica, però sempre podem considerar la seva *clausura*, és a dir, la menor varietat algebraica que la conté. Les varietats algebraiques són els tancats d'una topologia en  $k^n$  anomenada *topologia de Zariski*. A partir d'ara, quan parlem de la imatge d'una aplicació polinomial  $\varphi$ , ens referirem a la seva clausura. Així  $Im(k^d)$  denotarà la menor varietat algebraica que conté  $\varphi(k^d)$ . Quan el cos  $k$  és infinit, es pot veure que aquesta varietat algebraica és irreductible.

Per a poder estudiar una varietat algebraica ens interessa conèixer els generadors del seu *ideal*. Recordem que, donat un subconjunt  $X$  de  $k^n$ , l'ideal de  $X$  és el conjunt de polinomis que s'anul·len sobre tots els punts de  $X$ ,

$$I(X) = \{f(x_1, \dots, x_n) \mid f \in k[x_1, \dots, x_n], f(p) = 0 \forall p \in X\}.$$

Es pot demostrar fàcilment que  $I(X)$  és un ideal de l'anell de polinomis  $k[x_1, \dots, x_n]$ . El teorema de la base de Hilbert ens diu que aquest anell és noetherià i, per tant, tot ideal és finitament generat. Així existeixen generadors  $g_1, \dots, g_s \in k[x_1, \dots, x_n]$  tals que  $I(X) = (g_1, \dots, g_s)$ .

A causa del teorema dels zeros de Hilbert, sovint ens interessa considerar les nostres varietats dins de  $\mathbb{C}^n$ . De fet, en els casos en què es pugui, és molt més pràctic considerar els models estadístics com a aplicacions racionals des d'un producte d'espais projectius en un altre espai projectiu.

Tornant a l'aplicació polinomial (2) definida per un model evolutiu, ens interessa trobar els generadors de l'ideal de la clausura de la imatge.

**1 DEFINICIÓ** *Sigui  $V$  la clausura de la imatge de l'aplicació  $\varphi$  associada a un arbre  $T$  de  $n$  fulles i a un model evolutiu  $M$ . Els polinomis de  $I(V)$  s'anomenen invariants algebraics. Els polinomis de  $I(V)$  que no són en l'ideal  $I(V')$  corresponent a un altre arbre  $T'$  de  $n$  fulles sota el mateix model  $M$  s'anomenen invariants filogenètics.*

Els invariants filogenètics permeten distingir entre diferents arbres i, per tant, poden ser usats per a inferir l'arbre filogenètic d'espècies actuals. Els invariants filogenètics van ser introduïts per biòlegs i han estat usats per a estudiar l'adequació del model escollit o bé per a deduir divisions ancestrals entre grups d'espècies. Concretament van ser introduïts per Lake [14] i independentment per Cavender i Felsenstein [7]. No ha estat fins fa un parell d'anys que els matemàtics, o més precisament els geomètres algebraics, s'han interessat per aquesta aplicació de la geometria algebraica.

EXEMPLE En el cas d'un arbre  $T$  de tres fulles sota el model de Jukes-Cantor, les igualtats següents (que es dedueixen fàcilment de la simetria de les matrius de substitució en aquest model) donen lloc a invariants algebraics:

$$\begin{array}{ll}
 p_{AAA} = p_{CCC} = p_{GGG} = p_{TTT} & 4 \text{ termes} \\
 p_{AAC} = p_{AAG} = p_{AAT} = \dots = p_{TTG} & 12 \text{ termes} \\
 p_{ACA} = p_{AGA} = p_{ATA} = \dots = p_{TGT} & 12 \text{ termes} \\
 p_{CAA} = p_{GAA} = p_{TAA} = \dots = p_{GTT} & 12 \text{ termes} \\
 p_{ACG} = p_{ACT} = p_{AGT} = \dots = p_{CGT} & 24 \text{ termes}
 \end{array}$$

Un altre invariant que trobem fàcilment és  $\sum_{x_1, x_2, x_3} p_{x_1 x_2 x_3} - 1 = 0$ .

Aquests 60 invariants algebraics són polinomis de l'ideal per a qualsevol arbre de tres fulles sota el model Jukes-Cantor i, per tant, no són invariants filogenètics. Tenim tres arbres possibles de tres fulles (vegeu la figura 1) i l'única cosa que els distingeix sota el model de Jukes-Cantor és un polinomi de grau 3. Per a cadascun dels tres arbres possibles l'ideal és generat pels 60 invariants lineals que hem donat i per un polinomi de grau 3 que es pot calcular usant un programa d'àlgebra computacional (per exemple, el SINGULAR [10]). Això ens defineix una varietat de dimensió 3 a  $\mathbb{R}^{64}$  (que viu, de fet, a  $\mathbb{R}^4$ ).

Per a arbres de quatre o més fulles és impossible usar un programa d'àlgebra computacional per a trobar els generadors de l'ideal (fins i tot sota models senzills com Jukes-Cantor). Cal, doncs, provar resultats teòrics que ens permetin trobar els generadors de l'ideal (vegeu la secció 3).

A la pàgina web <http://bio.math.berkeley.edu/ascb/chapter15/> hi ha descrits els invariants algebraics per a arbres de fins a cinc fulles i sota diferents models evolutius.

Anomenem  $\rho_{x_1 \dots x_n}$  la freqüència relativa de la  $n$ -ples  $x_1, \dots, x_n$  en l'alineament donat. En l'exemple de Jukes-Cantor proposat en les pàgines 215 i 219, seguint l'arbre (1) de la figura 1 i l'alineament donat, tenim:

$$\begin{array}{llllll}
 \rho_{AAA} = 4/20, & \rho_{CAA} = 1/20, & \rho_{CCC} = 4/20, & \rho_{CCG} = 1/20, & \rho_{GGG} = 3/20 \\
 \rho_{GTC} = 1/20, & \rho_{GTT} = 1/20, & \rho_{TCC} = 1/20, & \rho_{TCG} = 1/20, & \rho_{TTT} = 3/20
 \end{array}$$

i és 0 en els altres casos.

En el cas hipotètic que un conjunt de seqüències haguessin evolucionat seguint un arbre i un model evolutiu dels descrits aquí, tindríem que els invariants filogenètics corresponents s'anullarien quan els avaluéssim en les freqüències relatives  $\rho_{x_1 \dots x_n}$ . Dit d'una altra manera,  $(\rho_{AA\dots A}, \dots, \rho_{TT\dots T})$  seria un punt de la nostra varietat algebraica. A la pràctica, les seqüències no han evolucionat seguint un model evolutiu i, per tant, en avaluar els invariants filogenètics en les freqüències relatives de l'arbre correcta, obtindrem valors només *propers* a zero.



NOTA Acabem de veure que el model de Jukes-Cantor per a un arbre arrelat de tres fulles ens defineix una varietat algebraica de dimensió 3 a  $\mathbb{R}^{64}$ . En particular en aquest model els paràmetres no són *identificables*; és a dir, donat un punt de la varietat, la seva antiimatge per  $\varphi$  no consta d'un únic punt  $(a_1, a_2, a_3, a_4) \in \mathbb{R}^4$ , sinó de tota una corba. En els models no algebraics, a l'hora d'inferir un arbre filogenètic cal inferir també quina és la longitud de les branques de l'arbre. Si en el model que considerem els paràmetres no són identificables, això implica que no podem trobar totes les longituds de les branques, sinó la suma d'algunes. En particular, no podem conèixer quina és la longitud de les branques que van des de l'arrel fins als altres nodes. Per a tenir models identificables, cal parlar d'arbres sense arrel i és per això que en la secció 4 considerem només arbres sense arrel.

### 3 Invariants per als models basats en grups i generalitzacions

Per a obtenir l'ideal de les varietats algebraiques corresponents als models de grup (Jukes-Cantor i Kimura 2 i 3-paràmetres) ha estat molt útil fer un canvi de coordenades en les indeterminades  $p_{x_1 \dots x_n}$ . El canvi proposat a [8] és una transformada de Fourier discreta tal com descrivim a continuació.

Pensem els caràcters A, C, G, T com a elements del grup  $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ . Així podem veure les matrius de substitució com a certes funcions sobre el grup:

$$S_i(g, h) = f^i(h - g), \quad g, h \in G$$

(és fàcil deduir com és  $f$  segons el model) i d'aquesta manera podem escriure les probabilitats conjuntes a les fulles també com a funcions sobre  $G \times G \times \dots \times G$ :

$$p(g_1, \dots, g_m) = \frac{1}{4} \sum \prod_{b \in \text{branques}} f^b(g_{p(b)} - g_{f(b)}),$$

on la suma és sobre tots els possibles valors de les variables en els nodes interns de l'arbre (si  $b$  és una branca de l'arbre,  $p(b)$  denota el node pare de  $b$  i  $f(b)$  el node fill). Recordem que, si tenim una funció  $f: G \rightarrow \mathbb{C}$ , la seva transformada de Fourier discreta  $\hat{f}: \hat{G} = \text{Hom}(G, \mathbb{C}^*) \rightarrow \mathbb{C}$  és donada per

$$\hat{f}(\chi) = \sum_{g \in G} \chi(g) f(g), \quad \chi \in G^*.$$

Una de les propietats més útils de la transformada de Fourier és que, si tenim una convolució de dues funcions  $(f_1 * f_2)(g) = \sum_{h \in G} f_1(h) f_2(g - h)$ , la seva transformada és el producte de transformades:  $\widehat{f_1 * f_2} = \widehat{f_1} \cdot \widehat{f_2}$ . Així es va poder demostrar el teorema següent:

2 TEOREMA (EVANS-SPEED [8]) *Per a un model basat en grup sobre un arbre  $T$ , la transformada de Fourier de la distribució conjunta  $p(g_1, \dots, g_n)$  té la forma*

$$q(\chi_1, \dots, \chi_m) = \prod_{b \in \text{branques}} \widehat{f}^b \left( \prod_{l \in \text{fulla per sota de } b} \chi_l \right).$$

Per tant, una expressió com (3) passa a ser una expressió monomial en les coordenades de Fourier. Dit d'una altra manera, en el cas dels models de grup, obtenim una parametrització monomial de la nostra varietat algebraica. És conegut que, si tenim una varietat algebraica donada per una parametrització monomial, aleshores el seu ideal és generat per binomis. Això fa que en aquestes noves coordenades de Fourier sigui molt fàcil calcular l'ideal de la varietat. Des del punt de vista de la geometria algebraica, aquestes varietats donades per parametritzacions monomials es coneixen com a *varietats tòriques* i han estat àmpliament estudiades.

3 TEOREMA (STURMFELS-SULLIVANT [17]) *Per als models de grup, l'ideal corresponent a un arbre filogenètic qualsevol és generat per binomis de grau com a màxim 4 en les coordenades de Fourier. A més a més, l'ideal d'un arbre arbitrari es pot descriure a partir de l'ideal d'un arbre de tres fulles sense arrel.*

La segona part del teorema dóna un algorisme de càlcul per als invariants filogenètics d'arbres de qualsevol nombre de fulles sota un model de grup.

Per a l'*strand symmetric model* un canvi de coordenades de Fourier no porta a una parametrització monomial, però en [6] hem considerat una transformada de Fourier *generalitzada* que ens ha permès trobar invariants filogenètics per a un arbre arbitrari a partir dels invariants d'un arbre de tres fulles sense arrel. La transformada que efectuem es pot explicar breument de la manera següent.

Considerem que les variables aleatòries als nodes de l'arbre prenen valors en parells  $\binom{j}{i}$ , on  $j$  pertany a un grup  $G$ ,  $i \in \{0, \dots, l\}$ , i  $l$  és un natural qualsevol. En el cas que ens ocupa considerem  $A = \binom{0}{0}$ ,  $G = \binom{0}{1}$ ,  $T = \binom{1}{0}$ ,  $C = \binom{1}{1}$ ,  $G = \mathbb{Z}/(2)$ ,  $l = 2$ . D'aquesta manera les matrius de substitució satisfan certes simetries. En efecte, si anomenem  $S_{i_1, i_2}^{j_1, j_2}$  la probabilitat  $P\left(\binom{j_2}{i_2} \mid \binom{j_1}{i_1}\right)$ , aleshores tenim

$$S_{i_1, i_2}^{j_1, j_2} = S_{i_1, i_2}^{k_1, k_2} \text{ si } j_1 - j_2 = k_1 - k_2.$$

Això permet considerar aquest model com una extensió dels models basats en grups ja que, denotant per  $p_{i_1 \dots i_n}^{j_1 \dots j_n}$  la probabilitat d'observar els estats  $\binom{j_1}{i_1}, \dots, \binom{j_n}{i_n}$  a les fulles de l'arbre, obtenim que, per a  $i_1, \dots, i_n$  fixats,  $p_{i_1 \dots i_n}^{j_1 \dots j_n}$  són donades per un *group-based model*. Aleshores, la transformada de Fourier que considerem serà respecte a aquest model de grup i és el que ens ha permès demostrar el resultat següent.

4 TEOREMA ([6]) *Per a l'strand symmetric model, si coneixem els invariants d'un arbre de tres fulles sense arrel, aleshores podem descriure els invariants d'un arbre qualsevol.*

En aquest cas és difícil descriure els invariants d'un arbre de tres fulles sense arrel (és un problema massa complex per als programes d'àlgebra computacional). Des del punt de vista de la geometria algebraica, és un problema interessant perquè es tracta de trobar l'ideal d'una projecció de la varietat de rectes secants de la varietat de Segre  $\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3 \hookrightarrow \mathbb{P}^{63}$ . Recentment, però, hem pogut descriure'n els generadors de grau  $\leq 5$  a partir de certes relacions binomials de matrius en les coordenades de Fourier generalitzades.

Per al model de Markov general també s'han aconseguit resultats anàlegs als teoremes 3 i 4 (vegeu [1]).

#### 4 Usant els invariants filogenètics

Els mètodes usats generalment en inferència filogenètica són el *neighbor-joining*, el *maximum parsimony* i un algorisme basat a trobar l'arbre i els paràmetres que satisfan el *màxim de versemblança*. Podeu trobar una explicació entenedora de tots a [9]. El primer no es basa en models evolutius i, per tant, no té en compte mutacions que no donen lloc a substitucions. Per exemple, al llarg d'una branca es pot produir una mutació d'una A a una G i després un altre cop a A. Aquest procés no és reflectit en les seqüències finals, però en canvi aquesta possibilitat sí que es recull en considerar models evolutius probabilístics. És per això que els algorismes de reconstrucció filogenètica basats en models evolutius són més fiables. El cost computacional de l'algorisme de màxim de versemblança, però, no permet usar aquest algorisme per a inferir arbres d'un gran nombre d'espècies. En canvi, l'algorisme *neighbor-joining* és àmpliament usat per a inferir filogènies de moltes espècies.

En aquesta secció proposem un mètode per a inferir arbres d'espècies usant els invariants filogenètics. Els resultats exposats en aquesta secció fan referència a arbres de quatre fulles, sense arrel. El model evolutiu considerat és el Kimura 3-paràmetres i els invariants filogenètics els vam calcular seguint els resultats de [17]. Aquests invariants es poden trobar a la pàgina web:

<http://bio.math.berkeley.edu/ascb/chapter15/>

Concretament, els generadors minimalis són 144 binomis de grau 2, 1.984 de grau 3 i 5.874 de grau 4.

Suposem que tenim quatre espècies i, per tant, tenim tres arbres possibles sense arrel (o, dit en el llenguatge de la filogenètica, tenim tres «topologies» possibles si mirem el graf etiquetat); vegeu la figura 3.

**Algorisme.** Partim d'un alineament de les quatre seqüències d'ADN  $s_1, s_2, s_3, s_4$ . Comptem les freqüències de cada 4-plet AAAA, AAAC, ..., TTTT segons l'arbre  $T_1$ . A partir d'aquestes freqüències  $\rho_{x_1 x_2 x_3 x_4}^{T_1}$  trobem fàcilment les freqüències de les 4-plets en els altres dos arbres  $T_2$  i  $T_3$ . Substituïm les indeterminades  $p_{x_1 x_2 x_3 x_4}$  pels valors de les freqüències  $\rho_{x_1 x_2 x_3 x_4}^T$  en cada polinomi invariant  $f$  i per a cada arbre  $T$ , i anomenem aquest valor  $s_f^T$ . Per fer-ho passem

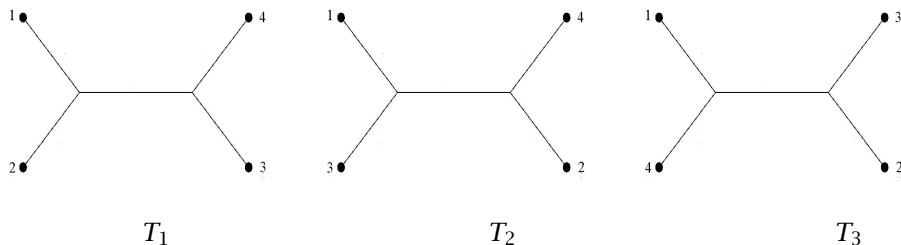


FIGURA 3: Els tres arbres de quatre fulles sense arrel.

primer a coordenades de Fourier, ja que en aquest cas l'avaluació del polinomi és molt més senzilla perquè es tracta d'un binomi. A partir d'aquests valors  $\{s_f^T\}_f$ , donem una puntuació a cada arbre:  $s(T) = \sum_f |s_f^T|$ . El nostre algorisme escull aleshores la topologia d'arbre que té menor puntuació.

L'algorisme és consistent en el sentit que, donades suficients dades provinents del model, recuperem l'arbre correcte amb probabilitat 1 (ja que els invariants s'han d'anul·lar).

Des del punt de vista de cost computacional, és impensable calcular els invariants filogenètics d'un arbre de qualsevol nombre de fulles. És per això que aquest mètode només ha estat desenvolupat per a arbres de quatre fulles. Donats els resultats obtinguts, però, creiem que pot ser un bon punt de partida per a mètodes d'inferència filogenètica basats en *quartets* ([9, capítol 12]). A continuació presentem un estudi de l'eficàcia d'aquest mètode. Uns altres estudis es poden trobar a [3], [5] i [6].

Usant el programa *evolver* del paquet PAML [18], hem generat seqüències seguint el model de Kimura 2-paràmetres per a l'arbre  $T_1$  (notem que aquest programa no permet generar seqüències sota el model de Kimura 3-paràmetres). A continuació descrivim les proves efectuades i els resultats obtinguts.

Per a cada  $n \in \{100, 200, \dots, 1.000\}$  vam generar mil alineaments de seqüències de longitud  $n$  que evolucionen sota el model Kimura 2-paràmetres en un arbre de quatre fulles amb longituds de branca distribuïdes uniformement entre 0 i 1. En la figura 4 es pot veure el percentatge d'arbres reconstruïts correctament usant tres mètodes diferents: el mètode basat en invariants presentat aquí, el *neighbor-joining* i el màxim de versemblança.

Veiem així que el mètode d'invariants supera el *neighbor-joining* i és compatible amb el mètode del màxim de versemblança. Quin és, doncs, l'avantatge d'usar mètodes basats en geometria algebraica? Intentarem explicar-ho breument a continuació.

Ja hem mencionat que els mètodes que considerem aquí són *algebraics*. En la versió no algebraica (és a dir, la versió usada pels biòlegs), les matrius de transició  $S_i$  són solucions d'una equació diferencial: si  $Q$  és una matriu fixada que representa les raons de canvi d'un nucleòtid a un altre en un instant,

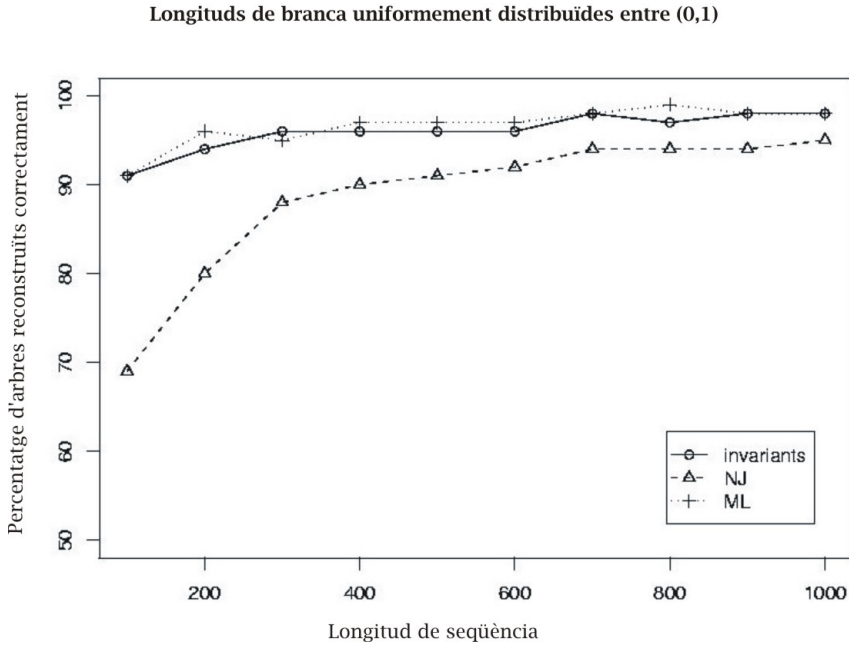


FIGURA 4: Percentatge d'arbres reconstruïts correctament amb longituds de branca uniformement distribuïdes entre 0 i 1. Els mètodes usats són el mètode d'invariants presentat aquí (invariants), el *neighbor-joining* (NJ) i el màxim de versemblança (ML).

aleshores la matriu de substitucions  $S(t)$  al cap d'un temps  $t$  satisfà

$$S'(t) = S(t)Q \quad i \quad S(0) = Id.$$

Per tant, en cada branca tenim  $S_i = \exp(Q \cdot t_i)$ . En els models que es consideren, normalment, en biologia la matriu  $Q$  és la mateixa per a totes les branques de l'arbre (model *homogeni*). Hi ha proves que no totes les espècies evolucionen amb les mateixes proporcions de canvis [19], però suposar que les matrius  $Q$  són diferents a cada branca (i, per tant, per a les diferents espècies de l'arbre) implicaria considerar un model amb un nombre excessiu de paràmetres. En canvi, en els models algebraics considerats aquí, els paràmetres del model són les entrades de les matrius  $S_i$ , cosa que implica que permetem que les matrius  $Q$  siguin diferents en cada branca. Això és el que s'anomena un model *nohomogeni*. Remarquem que en les simulacions de la figura 4 s'havia escollit un model homogeni per a generar els alineaments i, per tant, no s'ha afavorit el mètode basat en geometria algebraica.

Per exemple, els mètodes de màxim de versemblança i *neighbor-joining* aplicats a un model homogeni obtenen que la filogenia entre les espècies gos (*Canis familiaris*), humana (*Homo sapiens*), ximpanzé (*Pan troglodytes*), rata (*Rattus*

*norvegicus*), ratolí (*Mus musculus*) i pollastre (*Gallus gallus*) és la representada en la figura 5.

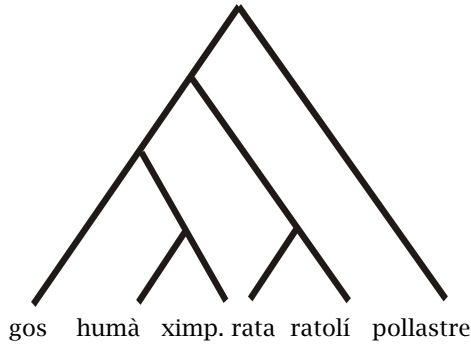


FIGURA 5: Arbre incorrecte.

Tot i que morfològicament l'espècie humana és més propera al gos que als rosegadors, l'arbre de la figura 5 no és l'arbre que es considera correcte. Donat que la raó de mutació en els rosegadors és més alta que en les altres espècies considerades, els mètodes basats en models homogenis tendeixen a situar-los en una posició incorrecta. En canvi, un mètode basat en invariants filogenètics reconstruiria l'arbre que s'estima com a cert (vegeu la figura 6).

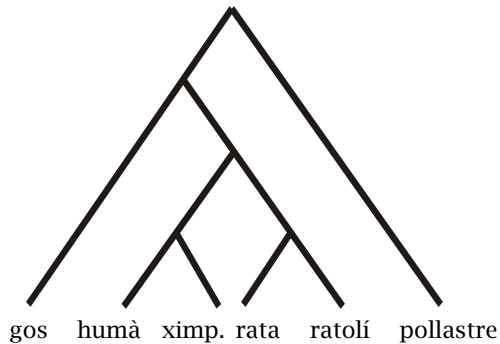


FIGURA 6: Arbre biològicament correcte.

**Agraïments** Agraïixo als organitzadors de la Novena Trobada de la Societat Catalana de Matemàtiques haver-me donat l'oportunitat de fer-hi una xerrada. Agraïixo també a la Facultat de Matemàtiques i Estadística de la Universitat Politècnica de Catalunya, i en especial a S. Xambó, haver-me permès incloure aquí part del que es va publicar al seu *Butlletí de la FME* [3]. Aquesta recerca ha

estat finançada pel Ministeri de Ciència i Tecnologia, Programa Ramón y Cajal, i BFM2003-06001.

## Referències

- [1] ALLMAN, E.; RHODES, J. «Phylogenetic ideals and varieties for the general Markov model» (2004). [Preprint, es pot trobar a la pàgina web <http://arxiv.org/abs/math.AG/0410604>]. *Advances in Applied Mathematics* [en premsa].
- [2] BARRY, D.; HARTIGAN, J. «Statistical analysis of hominoid molecular evolution». *Statistical Science*, 2 (1987), 191-210.
- [3] CASANELLAS, M. «Genètica i geometria algebraica». A: FME [ed.]. *Conferències FME. Volum II: Curs Einstein*. 2005, 299-316.
- [4] CASANELLAS, M.; FERNÁNDEZ-SÁNCHEZ, J., «Performance of a new invariants method on homogeneous and non-homogeneous quartet trees». *Mol. Biol. Evol.* [en premsa].
- [5] CASANELLAS, M.; GARCIA, L.; SULLIVANT, S. «Catalog of small trees». A: PACHTER, L.; STURMFELS, B. [ed.]. *Algebraic statistics for computational biology*, Cambridge University Press, 2005, cap. 15.
- [6] CASANELLAS, M.; SULLIVANT, S. «The strand symmetric model». A: PACHTER, L.; STURMFELS, B. [ed.]. *Algebraic statistics for computational biology*, Cambridge University Press, 2005, cap. 16.
- [7] CAVENDER, J.; FELSENSTEIN, J. «Invariants of phylogenies in a simple case with discrete states». *J. Classification*, 4 (1987), 57-71.
- [8] EVANS, S.; SPEED, T. «Invariants of some probability models used in phylogenetic inference». *The Annals of Statistics*, 21 (1993), 355-377.
- [9] FELSENSTEIN, J. *Inferring phylogenies*. Sinauer Associates, 2003.
- [10] GREUEL, G.; PFISTER, G.; SCHOENEMANN, H. «Singular: a computer algebra system for polynomial computations». 2003. Es pot trobar al web <http://www.singular.uni-kl.de/>.
- [11] JUKES, T.; CANTOR, C. «Evolution of protein molecules». A: MUNRO, H. [ed.]. *Mammalian protein metabolism*. Nova York: Academic Press, 1969, 21-32.
- [12] 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.
- [13] KIMURA, M. «Estimation of evolutionary sequences between homologous nucleotide sequences». *Proc. Nat. Acad. Sci., USA*, 78 (1981), 454-458.
- [14] LAKE, J. «A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony». *Mol. Biol. Evol.*, 4 (1987), 167-191.
- [15] PACHTER, L.; STURMFELS, B. [ed.]. *Algebraic statistics for computational biology*. Cambridge University Press, 2005. ISBN 0-521-85700-7.

- [16] PISTONE, G.; RICCOMAGNO, E.; WYNN, H. *Algebraic statistics: Computational commutative algebra in statistics*. Chapman & Hall: CRC Press, 2001.
- [17] STURMFELS, B.; SULLIVANT, S. «Toric ideals of phylogenetic invariants». *J. Comput. Biol.*, 12 (2005), 204–228.
- [18] YANG, Z. «PAML: a program package for phylogenetic analysis by maximum likelihood». *CABIOS*, 15 (1997), 555–556.
- [19] YANG, Z.; YODER, A. D. «Estimation of the transition/transversion rate bias and species sampling». *J. Mol. Evol.*, 48 (1999), 274–283.

DEPARTAMENT DE MATEMÀTICA APLICADA I  
UNIVERSITAT POLITÈCNICA DE CATALUNYA  
AVINGUDA DIAGONAL, 647  
08028-BARCELONA  
marta.casanellas@upc.edu