Genètica i geometria algebraica

M. Casanellas*
Dpt. Matemàtica Aplicada I
Universitat Politècnica de Catalunya
Av. Diagonal 647. 08028-Barcelona.
marta.casanellas@upc.edu

1 Introducció

En aquest escrit pretenem explicar la relació entre la geometria algebraica i la genètica, o més concretament la filogenètica. Aquestes disciplines, que aparentment són molt distants una de l'altra, es poden relacionar mitjançant una branca de l'estadística anomenada estadística algebraica.

La filosofia que hi ha en la base de l'estadística algebraica parteix del fet que molts models estadístics són varietats algebraiques. Aquesta disciplina s'ha començat a desenvolupar en els últims anys i s'ha aplicat a l'estudi de taules de contingència, disseny d'experiments.... En el cas de la biologia, per exemple, la coneguda equació d'equilibri de Hardy-Weinberg sobre la distribució d'al·lels en una població defineix una varietat algebraica. Exemples més complexos els trobem en filogenètica, que és la branca de la biologia que estudia les relacions ancestrals entre diferents espècies. En aquests camp ens trobem que els models evolutius probabílistics d'arbres filogenètics corresponen també a varietats algebraiques.

Un bon tractat sobre estadística algebraica el podem trobar al llibre [PRW00]. Les aportacions de l'estadística algebraica a la biologia computacional que s'han fet fins aquest moment es troben recollides al llibre "Algebraic statistics for computational biology" [PS05].

Recentment els esforços de biòlegs i geòmetres algebraics s'estan unint per a donar noves aplicacions de l'estadística algebraica a la biologia i més concretament a la biologia computacional. Com veurem, l'ús de la geometria algebraica en la filogenètica pot aportar noves tècniques i nous resultats a aquesta branca de la biologia. Aquesta interacció entre les matemàtiques i la biologia deixa entreveure que les dues disciplines poden beneficiar-se d'un treball conjunt.

En la secció següent presentem el tipus de models estadístics que seran del nostre interès. En la secció tres expliquem la relació entre aquests models estadístics i

*Amb el suport del Ministeri de Ciència i Tecnologia, Programa Ramón y Cajal, i BFM2003-06001

la geometria algebraica. Seguidament en la secció 4 ens restringim a l'aplicació de l'estadística algebraica en la filogenètica. Primer de tot descrivim quins són els models evolutius que ens permeten treballar amb la geometria algebraica i a continuació descrivim les tècniques de la geometria algebraica que són útils per a la inferència d'arbres filogenètics. En l'última secció presentem un mètode per a inferir arbres filogenètics usant la geometria algebraica.

2 Models paramètrics

Un model estadístic és una família de distribucions de probabilitats en un espai mostral Ω. En el nostre cas ens interessarà estudiar models estadístics paramètrics on les distribucions vinguin donades com a polinomis en els paràmetres. En tal cas parlarem de *model estadístic algebraic*.

Suposem que tenim una variable aleatòria discreta X i que el cardinal de Ω és N. Si en el model estadístic tenim d paràmetres reals $\theta = (\theta_1, \ldots, \theta_d)$, podem presentar-lo com una aplicació:

$$\begin{array}{rccc} \varphi: & \mathbb{R}^d & \longrightarrow & \mathbb{R}^N \\ \theta = (\theta_1, \ldots, \theta_d) & & \mapsto & (\varphi_1(\theta), \ldots, \varphi_N(\theta)) \end{array}$$

on $\varphi_i(\theta)$ és la probabilitat de l'esdeveniment i donats els paràmetres $\theta = (\theta_1, \ldots, \theta_d)$, és a dir, $\varphi_i(\theta) = Prob(X = i \,|\, \theta)$.

En general φ està definida en un subconjunt $U \subset \mathbb{R}^d$ i la seva imatge està dins del símplex $(N-1)$-dimensional de $\mathbb{R}^N$. Ens interessarà el cas en què $\varphi_i(\theta)$ sigui un polinomi en $\theta_1, \ldots, \theta_d$.

Exemple 2.1 Considerem X la variable aleatòria que representa el llançament d'una moneda dues vegades de forma independent. Aleshores si θ_1 és la probabilitat de treure "cara" en el llançament de la moneda i $\theta_2 = 1 - \theta_1$ tenim que $Prob(X = \{cara, cara\}) = \theta_1^2$, $Prob(X = \{creu, creu\}) = \theta_2^2$ i $Prob(X = \{cara, creu\}) = 2\theta_1\theta_2$. Aquest model estadístic el podem representar com l'aplicació polinomial:

$$\begin{array}{rccc} \varphi: \mathbb{R}^2 & & \longrightarrow & \mathbb{R}^3 \\ \theta = (\theta_1, \theta_2) & & \mapsto & (\theta_1^2, 2\theta_1\theta_2, \theta_2^2). \end{array}$$

Per a que aquest model estadístic tingui sentit cal que (θ_1, θ_2) estigui en el símplex 1-dimensional de $\mathbb{R}^2$ i en tal cas tindrem que la imatge $\varphi(\theta_1, \theta_2)$ també està en el símplex 2-dimensional de $\mathbb{R}^3$.

En la següent secció estudiarem la relació entre les aplicacions polinomials i la geometria algebraica.

3 Geometria algebraica i estadística algebraica

En la secció anterior hem presentat un tipus de models estadístics. Ara veurem quina relació hi ha entre aquests models estadístics i la geometria algebraica.

Grosso modo podem dir que la geometria algebraica és la part de les matemàtiques on s'estudien els conjunts de solucions de sistemes d'equacions polinomials:

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ \vdots \\ f_r(x_1, \ldots, x_n) = 0 \end{cases}$$

$f_1, \ldots, f_r \in k[x_1, \ldots, x_n]$ són polinomis en n variables amb coeficients en un cos k.

Definició 3.1 El conjunt de solucions $V = Z(f_1, \ldots, f_r) \subseteq k^n$ s'anomena *varietat algebraica*.

Donat un subconjunt X de k^n també podem preguntar-nos per tots els polinomis que s'anul.len en tots els punts de X:

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

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

Definició 3.2 Donat un subconjunt X de k^n, anomenem l'*ideal de* X al següent ideal de l'anell de polinomis

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

Per a poder estudiar una varietat algebraica ens interessa conèixer els generadors del seu ideal.

Exemple 3.3 En l'exemple 2.1 considerat en la secció anterior volem esbrinar si la imatge de l'aplicació

$$\begin{array}{ccc} \varphi : \mathbb{R}^2 & \longrightarrow & \mathbb{R}^3 \\ \theta = (\theta_1, \theta_2) & \mapsto & (\theta_1^2, 2\theta_1\theta_2, \theta_2^2). \end{array}$$

és una varietat algebraica i trobar generadors de $I(im\varphi)$. En aquest exemple es pot veure fàcilment que $V = im(\varphi)$ sí que és una varietat algebraica ja que és el conjunt de solucions de l'equació $y^2 - 4xz = 0$, on x, y, z són les coordenades de $\mathbb{R}^3$. Així, $V = Z(y^2 - 4xz)$. Observem però que un punt (x, y, z) d'aquesta varietat V correspon a una distribució de probabilitats del model estadístic de l'exemple 2.1 si, i només si, $(x, y, z) \in V \cap \Delta^2$ on Δ^2 és el símplex 2-dimensional de $\mathbb{R}^3$. És

a dir, els models estadístics algebraics corresponen a subconjunts de les varietats algebraiques.

En aquest cas l'ideal de V està generat pel polinomi y^2-4xz, $I(V)=(y^2-4xz)$. En general si V és el conjunt de zeros dels polinomis d'un ideal $J=(f_1,\ldots,f_r)$, $V=Z(f_1,\ldots,f_r)$, en considerar $I(V)$ no retrobem necessàriament l'ideal J. Per exemple, si $J=(x^2)$ aleshores $V=\{(x,y,z)\in\mathbb{R}^3|x^2=0\}=\{(x,y,z)\in\mathbb{R}^3|x=0\}$ i $I(V)=(x)$, que és diferent de l'ideal generat per x^2. L'explicació d'aquest fenomen rau en el teorema dels zeros de Hilbert [CLO97].

La imatge d'una aplicació polinomial no és en general una varietat algebraica, però sempre podem considerar la seva *clausura algebraica*, és a dir, la menor varietat algebraica que el 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 φ ens referirem a la seva clausura algebraica. 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.

Degut al teorema fonamental de l'àlgebra sovint ens interessa considerar les nostres varietats dins de $\mathbb{C}^n$. De fet, en els casos que es pugui, és molt més pràctic considerar els models estadístics com aplicacions des d'un producte d'espais projectius en un altre espai projectiu. En l'exemple anterior φ ens defineix també una aplicació

$$\begin{array}{ccc} \phi:\mathbb{P}^1_{\mathbb{C}} & \longrightarrow & \mathbb{P}^2_{\mathbb{C}} \\ \theta=[\theta_1:\theta_2] & \mapsto & [\theta_1^2:2\theta_1\theta_2:\theta_2^2] \end{array}$$

Notem que en definir l'aplicació φ en l'exemple 2.1 no hem substituït θ_2 per $1-\theta_1$ i això ens ha permès pensar-la com una aplicació entre espais projectius. L'objectiu de no fer aquesta substitució és que així φ ve definida per monomis en els paràmetres θ_1,θ_2 i és més senzill trobar l'ideal de la varietat imatge. En tal cas l'ideal està generat per binomis i la varietat imatge és una *varietat tòrica*.

4 Estadística algebraica en la filogenètica

Es creu que tots els organismes de la Terra provénen d'un ancessor comú. Així, totes les espècies de la Terra tenen relacions ancestrals entre elles i aquestes relacions s'anomenen filogènies. Les filogènies es poden representar mitjançant un arbre que s'anomena *arbre filogenètic*. La filogenètica es dedica a inferir aquests arbres a partir de les seqüències dels genomes[1] de les espècies que trobem en l'actualitat.

[1]El genoma és el contingut d'ADN que hi ha en una cèl.lula. Aquest contingut d'ADN està repartit en els diferents cromosomes del nucli de la cèl.lula. Cada cromosoma és una mol.lècula gegant d'ADN que està formada per unes unitats bàsiques anomenades *nucleòtids*. Hi ha quatre tipus diferents de nucleòtids: `A` denota *adenina*, `C` *citosina*, `G` *guanina* i `T` *timina*. Degut a certes propietats químiques, l'adenina i la guanina són *purines* i la citosina i la timina són *pirimidines*. La cadena d'ADN té una estructura de cadena doble (la *doble hèlix*) que preserva una simetria entre les dues cadenes que la formen: l'adenina (respectivament la citosina) d'una de les cadenes

Per exemple si coneixem el genoma (o part d'ell) de l'espècie humana, la rata i el ratolí podem preguntar-nos quin és l'arbre filogenètic d'aquestes espècies. Tenim tres possibilitats:

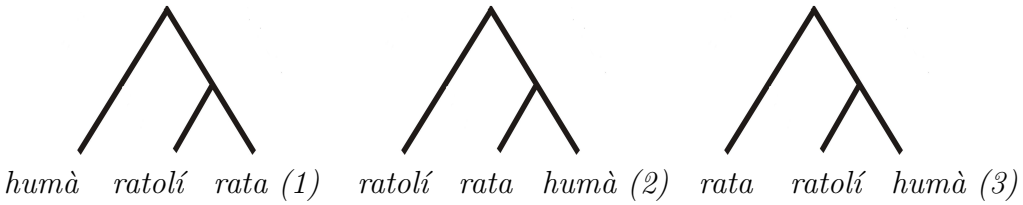


Figura 1: Els tres possibles arbres de tres fulles

L'arbre filogenètic es dibuixa com una arbre invertit: les fulles de l'arbre representen les espècies actuals i les posem a baix de tot; l'arrel de l'arbre filogenètic es troba al node de dalt de tot i representa l'ancessor comú de totes les espècies que formen l'arbre; els nodes interiors representen espècies ancestrals comuns 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.

4.1 Models evolutius algebraics

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

(i) Els arbres són *binaris*, és a dir, del node arrel en surten dues branques i cada branca es divideix en dues més fins arribar a les fulles.

(ii) L'evolució d'una espècie només depèn del node immediatament superior.

(iii) Les mutacions ocorren aleatòriament i la probabilitat que es produeixi una mutació és sempre positiva.

(iv) Suposarem donat un alineament de les seqüències d'ADN de les espècies. Degut a 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 per què trobar-se en el mateix lloc del genoma (els genomes de diferents espècies tenen diferents longituds i diferent nombre de

sempre s'enllaça amb la timina (resp. guanina). Donar el genoma d'una espècie equival a donar, per a tots els cromosomes, la seqüencia de caràcters en les lletres `A,C,G,T` que forma una de les dues cadenes. La seqüència d'ADN del genoma humà no va ser completament coneguda fins a finals de 2004 i avui en dia es coneix la seqüència d'ADN només d'unes quantes dotzenes d'espècies animals vertebrades i invertebrades.

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 elles. Tot això es recull en un bon alineament de les seqüències d'ADN que volem considerar. En el nostre exemple, suposarem que partim del següent alineament de seqüències:

humà	AACTTCGAGGCTTACCGCTG
ratolí	AACGTCTATGCTCACCGATG
rata	AAGGTCGATGCTCACCGATG

(v) 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 en l'arbre de la següent forma. Enumerem els vèrtexos de l'arbre d'esquerra a dreta i de baix a dalt i anomenem t_i a 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 $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. 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:

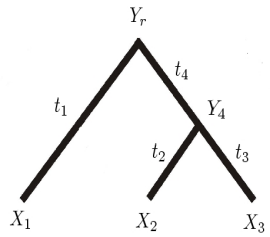


Figura 2: Model estadístic en un arbre de tres fulles

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:

$$S_i = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} P(\mathtt{A}|\mathtt{A}, t_i) & P(\mathtt{C}|\mathtt{A}, t_i) & P(\mathtt{G}|\mathtt{A}, t_i) & P(\mathtt{T}|\mathtt{A}, t_i) \\ P(\mathtt{A}|\mathtt{C}, t_i) & P(\mathtt{C}|\mathtt{C}, t_i) & P(\mathtt{G}|\mathtt{C}, t_i) & P(\mathtt{T}|\mathtt{C}, t_i) \\ P(\mathtt{A}|\mathtt{G}, t_i) & P(\mathtt{C}|\mathtt{G}, t_i) & P(\mathtt{G}|\mathtt{G}, t_i) & P(\mathtt{T}|\mathtt{G}, t_i) \\ P(\mathtt{A}|\mathtt{T}, t_i) & P(\mathtt{C}|\mathtt{T}, t_i) & P(\mathtt{G}|\mathtt{T}, t_i) & P(\mathtt{T}|\mathtt{T}, t_i) \end{pmatrix} \end{array}$$

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

Aquest model és un cas particular del que s'anomenen models de Markov ocults. Suposarem que la distribució de nucleòtids en l'arrel $\pi_{\mathtt{A}} = P(Y_r = \mathtt{A})$, $\pi_{\mathtt{C}} = P(Y_r = \mathtt{C})$, $\pi_{\mathtt{G}} = P(Y_r = \mathtt{G})$, $\pi_{\mathtt{T}} = P(Y_r = \mathtt{T})$ és coneguda. La hipòtesi (v) ens diu que la probabilitat que del node arrel de l'arbre (1) s'hagi evolucionat a les seqüències d'humà, rata i ratolí donades en l'alineament és igual a

$$p^4_{\mathtt{AAA}} * p_{\mathtt{CCG}} * p_{\mathtt{TGG}} * p^3_{\mathtt{TTT}} * p^4_{\mathtt{CCC}} * p_{\mathtt{GTG}} * p_{\mathtt{GTT}} * p^3_{\mathtt{GGG}} * p_{\mathtt{TCC}} * p_{\mathtt{CAA}}$$

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 (v) ens diu que les columnes de l'alineament evolucionen de forma independent. La probabilitat d'observar els nucleòtids x_1, x_2, x_3 en les fulles de l'arbre de la figura 2 s'expressa en funció de les entrades de les matrius de substitució de la següent manera:

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

Tenim diferents models probabilístics d'evolució segons la forma que tinguin les matrius de substitució i segons la distribució de nucleòtids en l'arrel. Aquí en descrivim uns quants ordenats de menor a major complexitat.

- **Model de Jukes-Cantor.** La probabilitat que un nucleòtid muti a un altre és sempre la mateixa i és molt més petita que la probabilitat que un nucleòtid es quedi invariant:

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

  Tenim dos paràmetres per branca, a_i, b_i que han de satisfer $a_i + 3b_i = 1$, $a_i \geq 0$, $b_i \geq 0$. Per a aquest model la distribució de nucleòtids en l'arrel se suposa uniforme, és a dir, $\pi_{\mathtt{A}} = \pi_{\mathtt{C}} = \pi_{\mathtt{G}} = \pi_{\mathtt{T}} = \frac{1}{4}$.

- **Kimura 2 paràmetres.** Aquest model reflecteix 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 l'inversa). A tal efecte les matrius de substitució d'aquest model tenen la següent forma:

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

  Per a aquest model la distribució de nucleòtids en l'arrel també se suposa uniforme.

- **Kimura 3 paràmetres.** Aquest model és una mica més general que l'anterior perquè afegeix un nou paràmetre per als dos tipus de transicions possibles. Les matrius de substitució són de la forma

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

  Per a aquest model la distribució de nucleòtids en l'arrel també se suposa uniforme.

- **Strand symmetric model.** En aquest model no se suposa que les distribució de nucleòtids és uniforme sinó que $\pi_{\mathtt{A}} = \pi_{\mathtt{T}}$ i $\pi_{\mathtt{C}} = \pi_{\mathtt{G}}$. 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 degut a la simetria de la doble cadena, s'ha comprovat en diversos estudis que $\pi_{\mathtt{A}} \sim \pi_{\mathtt{T}}$ i $\pi_{\mathtt{C}} \sim \pi_{\mathtt{G}}$. En aquest model suposarem que $\pi_{\mathtt{A}} = \pi_{\mathtt{T}}$ i $\pi_{\mathtt{C}} = \pi_{\mathtt{G}}$. 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` (resp. `G`). Així es natural requerir que

$$P(\mathtt{A}|\mathtt{A}) = P(\mathtt{T}|\mathtt{T}), P(\mathtt{C}|\mathtt{A}) = P(\mathtt{G}|\mathtt{T}), P(\mathtt{G}|\mathtt{A}) = P(\mathtt{C}|\mathtt{T}),$$

$$P(\mathtt{T}|\mathtt{A}) = P(\mathtt{A}|\mathtt{T}), P(\mathtt{A}|\mathtt{C}) = P(\mathtt{T}|\mathtt{G}), P(\mathtt{C}|\mathtt{C}) = P(\mathtt{G}|\mathtt{G}),$$

$$P(\mathtt{G}|\mathtt{C}) = P(\mathtt{C}|\mathtt{G}), P(\mathtt{T}|\mathtt{C}) = P(\mathtt{T}|\mathtt{G}).$$

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

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

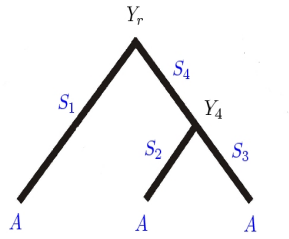
  Notem que tots els models descrits anteriorment són un cas particular d'aquest. Aquest model algebraic va ser introduït a [CS05].

- **Model de Markov general.** Aquest és el model més general possible. No es requereix res sobre la distribució en 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 parametres 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.

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. Per exemple, si en el següent arbre hi posem



el model de Jukes-Cantor, obtenim (substituint en l'equació (1)):

$$p_{\texttt{AAA}} = \frac{1}{4}(a_1a_4a_2a_3 + 3b_1b_4a_2a_3 + 3b_1a_4a_2a_3 + 3a_1b_4a_2a_3 + 6b_1b_4b_2b_3), \qquad (2)$$

on $a_i + 3b_i = 1$. Observem que per a qualsevol dels models descrits, $p_{x_1x_2x_3}$ 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

$$\begin{array}{rcl} \varphi : \mathbb{R}^4 & \longrightarrow & \mathbb{R}^{64} \\ (a_1, a_2, a_3, a_4) & \mapsto & (p_{\texttt{AAA}}, p_{\texttt{AAC}}, p_{\texttt{AAG}}, \ldots, p_{\texttt{TTT}}) \end{array}$$

Els altres models algebraics definits anteriorment es poden descriure també com a aplicacions polinòmiques. Per a un arbre T de n fulles i un model evolutiu algebraic de d paràmetres lliures tindrem la següent aplicació polinomial:

$$\begin{array}{rcl} \varphi : \mathbb{R}^d & \longrightarrow & \mathbb{R}^{4^n} \\ \theta = (\theta_1, \ldots, \theta_d) & \mapsto & (p_{\texttt{AA...A}}, p_{\texttt{AA...C}}, p_{\texttt{AA...G}}, \ldots, p_{\texttt{TT...T}}) \end{array}$$

Observem que hi ha molts arbres diferents de n fulles i per tant en donar un model evolutiu com una aplicació polinomial caldrà especificar quin és l'arbre T que estem considerant.

4.2 Invariants filogenètics

Acabem de veure que els models evolutius algebraics són un cas particular de models estadístics algebraics. Per tant ens podem preguntar per la clausura algebraica de la imatge de l'aplicació polinomial corresponent:

$$\begin{array}{rcl} \varphi : \mathbb{R}^d & \longrightarrow & \mathbb{R}^{4^n} \\ \theta = (\theta_1, \ldots, \theta_d) & \mapsto & (p_{\texttt{AA...A}}, p_{\texttt{AA...C}}, p_{\texttt{AA...G}}, \ldots, p_{\texttt{TT...T}}) \end{array}$$

Definició 4.1 Sigui V la clausura algebraica de l'aplicació φ associada a un arbre T de n fulles i a un model evolutiu M. Els polinomis de $I(V)$ s'anomenen *invariants algebraics*. Aquells polinomis de $I(V)$ que no estan 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 [Lak87] i independentment per Cavender i Felsenstein [CF87]. No ha estat fins fa un parell d'anys que els matemàtics, o més precisament els geòmetres algebraics, s'han interessat per aquesta aplicació de la geometria algebraica.

Exemple 4.2 En el cas d'un arbre T de 3 fulles sota el model de Jukes-Cantor les següents igualtats (que es dedueixen fàcilment de la simetria de les matrius de substitució en aquest model) donen lloc a invariants algebraics:

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

Un altre invariant que trobem fàcilment és $\sum_{x_1,x_2,x_3} p_{x_1x_2x_3} - 1 = 0$.

Aquests 60 invariants algebraics són polinomis de l'ideal per a qualsevol arbre de 3 fulles sota el model Jukes-Cantor i per tant no són invariants filogenètics. Tenim 3 arbres possibles de 3 fulles (vegeu la figura 1) i l'únic que els distingeix sota el model de Jukes-Cantor és un polinomi de grau 3. Per a cadascun dels tres arbres possibles l'ideal està 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` [GPS03]). Això ens defineix una varietat de dimensió 3 a $\mathbb{R}^4$.

Per a arbres de 4 o més fulles és impossible usar un programa d'àlgebra computacional per a trobar els generadors de l'ideal (fins i tot sota el model Jukes-Cantor). Cal doncs, provar resultats teòrics que ens permetin trobar els generadors de l'ideal (vegeu la secció 4.3).

A la pàgina web `http://bio.math.berkeley.edu/ascb/chapter15/` hi ha descrits els invariants algebraics per a arbres de fins a 5 fulles i sota diferents models evolutius.

Anomenem $\rho_{x_1\ldots x_n}$ a la freqüència relativa de la n-tuple $x_1, \ldots, x_n$ en l'alineament. En l'exemple proposat en les pàgines 5 i 6, seguint l'arbre (1) i l'alineament

donat tenim que:

$$\rho_{\mathtt{AAA}} = 4/20, \rho_{\mathtt{CAA}} = 1/20, \rho_{\mathtt{CCC}} = 4/20, \rho_{\mathtt{CCG}} = 1/20, \rho_{\mathtt{GGG}} = 3/20$$
$$\rho_{\mathtt{GTG}} = 1/20, \rho_{\mathtt{GTT}} = 1/20, \rho_{\mathtt{TCC}} = 1/20, \rho_{\mathtt{TGG}} = 1/20, \rho_{\mathtt{TTT}} = 3/20$$

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 s'anul.larien quan els avaluéssim en les freqüències relatives $\rho_{x_1 \ldots x_n}$. Dit d'una altra manera, $(\rho_{\mathtt{AA\ldots A}}, \ldots, \rho_{\mathtt{TT\ldots T}})$ seria un punt de la nostra varietat algebraica. En els casos reals, les seqüències no han evolucionat seguint un model evolutiu i per tant en avaluar els invariants filogenètics en les freqüències relatives de l'arbre correcte obtindrem valors propers a zero. Per inferir l'arbre filogenètic caldrà decidir a quina de les varietats algebraiques corresponent als diferents arbres és més proper el punt $(\rho_{\mathtt{AA\ldots A}}, \ldots, \rho_{\mathtt{TT\ldots T}})$.

Nota 4.3 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 $\mathbb{R}^4$. En particular en aquest model els paràmetres no són *identificables*, és a dir, donat un punt de la varietat la seva antiimatge per φ 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 estem considerant els paràmetres no són identificables, això implica que no podem trobar totes les longituds de les branques sinó la suma d'algunes d'elles. 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 5 considerem només arbres sense arrel.

4.3 Transformada de Fourier

Per a obtenir l'ideal de les varietats algebraiques corresponents als models de Jukes-Cantor i Kimura (2 i 3 paràmetres) ha estat molt útil fer un canvi de coordenades en les indeterminades $p_{x_1 \ldots x_n}$. El canvi proposat a [ES93] és una transformada de Fourier discreta tal i com describim a continuació.

Pensem els caràcters `A,C,G,T` com a elements del grup $G = \mathbb{Z}_2 \times \mathbb{Z}_2$. La coordenada de Fourier $q_{x_1 \ldots x_n}$ s'obté de les coordenades originals com:

$$q_{x_1 \cdots x_n} = \frac{1}{4^n} \sum_{y_1, \ldots, y_n} \chi^{x_1}(y_1) \cdots \chi^{x_n}(y_n) p_{x_1 \cdots x_n} \tag{3}$$

on χ^i és el caràcter del grup (és a dir, un element de $\hat{G} = Hom(\mathbb{Z}_2 \times \mathbb{Z}_2, \mathbb{C}^*)$) associat a l'element i del grup (usant l'isomorfisme entre G i $\hat{G}$). Dit d'una altra manera,

$\chi^i(j)$ és l'entrada (i,j) de la següent matriu:

	A	C	G	T
A	1	1	1	1
C	1	-1	1	-1
G	1	1	-1	-1
T	1	-1	-1	1

.

Fent aquest canvi de coordenades i el corresponent canvi de coordenades en els paràmetres del model, es pot veure que les coordenades de Fourier s'expressen com a monomis en els nous paràmetres ([ES93], [SS05]) en el cas que estiguem considerant models de Jukes-Cantor o Kimura. Així, una expressió com (2) passa a ser una expressió monomial en les coordenades de Fourier. Dit d'una altra manera, en el cas de Jukes-Cantor i Kimura (2 i 3 paràmetres) 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 està generat per binomis. Això fa que en aquestes noves coordenades de Fourier sigui molt fàcil calcular l'ideal de la varietat. A [SS05] es dóna un algoritme de càlcul per als invariants filogenètics d'arbres de qualssevol nombre de fulles sota el model de Jukes-Cantor o Kimura (2 o 3 paràmetres).

Teorema 4.4 ([SS05]) *Per als models Jukes-Cantor i Kimura (2 i 3 paràmetres), l'ideal corresponent a un arbre filogenètic qualsevol està generat per binomis de grau com a màxim 4 en les coordenades de Fourier.*

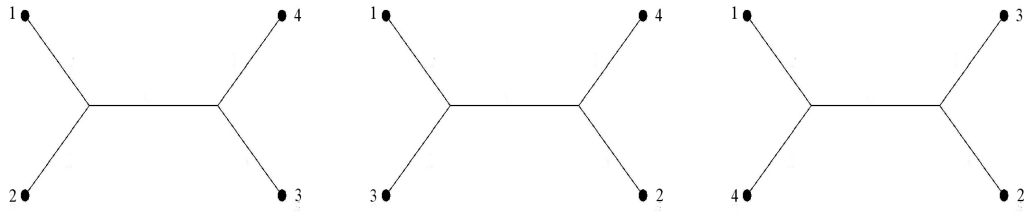
Per al "strand symmetric model" un canvi de coordenades de Fourier no porta a una parametrització monomial però en [CS05] hem considerat una transformada de Fourier "generalitzada" que ens ha permès trobar invariants filogenètics per a un arbre arbitrari.

Teorema 4.5 ([CS05]) *Per al "strand symmetric model", si coneixem els invariants d'un arbre de 3 fulles sense arrel, aleshores podem descriure els invariants d'un arbre qualsevol.*

5 Reconstrucció d'arbres filogenètics

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 4 fulles, sense arrel, i es poden trobar a [CGS05]. Vam usar el model evolutiu Kimura 3 paràmetres i els invariants filogenètics els vam calcular seguint els resultats de [SS05]. Aquests invariants es poden trobar a la pàgina web:

`http://bio.math.berkeley.edu/ascb/chapter15/`



T_1 T_2 T_3

Figura 3: Els tres arbres de 4 fulles sense arrel

Suposem que tenim 4 espècies i per tant tenim 3 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.

Algoritme: Partim d'un alineament de les quatres seqüències d'ADN s_1, s_2, s_3, s_4. Comptem les freqüències de cada 4-uple `AAAA, AAAC, ...,TTTT` segons l'arbre T_1. A partir d'aquestes freqüències $\rho^{T_1}_{x_1x_2x_3x_4}$ trobem fàcilment les freqüències de les 4-uples ens els altres dos arbres T_2 i T_3. Substituïm les indeterminades $p_{x_1x_2x_3x_4}$ pels valors de les freqüències $\rho^{T}_{x_1x_2x_3x_4}$ en cada polinomi invariant f i per a cada arbre T, i anomenem aquest valor s^T_f. Per fer-ho passem 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 $\{s^T_f\}_f$, donem una puntuació a cada arbre: $s(T) = \sum_f |s^T_f|$. El nostre algoritme escull aleshores la topologia d'arbre que té menor puntuació.

Hem provat l'eficàcia d'aquest algoritme per a diferents conjunts de seqüències. Usant el programa *evolver* del paquet `PAML` [Yan97], hem generat seqüències seguint el model de Kimura 2 paràmetres per a l'arbre T_1. A continuació descrivim les proves efectuades i els resultats obtinguts.

Vam generar arbres de 4 fulles amb longituds de branca distribuïdes uniformement entre 0 i 1 i vam fer 600 tests per a seqüències de longitud 1000, 2000, ..., 10000. El percentatge d'arbres resconstruïts correctament es pot veure en la figura 4. Hem observat que el nostre mètode falla majoritàriament quan la branca interior de l'arbre és significativament més curta que les altres, de longitud aproximadament un 10% de la mitjana de les altres longituds.

D'altra banda, també vam provar el nostre mètode generant arbres amb longituds de branca seguint una distribució normal de mitjana fixada μ. Seguint l'article [JWMV03], vam escollir els valors 0.25, 0.05 i 0.005 per a μ i desviació estàndar 0.1μ. En aquest cas vam fer proves per a seqüències de longitud entre 50 i 10000, però només mostrem els resultats per a seqüències de longitud com a màxim 1000 perquè per a longituds majors inferim sempre l'arbre correcte. Per a cada longitud de seqüència dins del conjunt $\{50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ vam generar longituds de branca distribuïdes segons una normal de mitjana μ usant el paquet estadístic `R` [RG96]. Amb aquestes longituds de branca vam generar amb

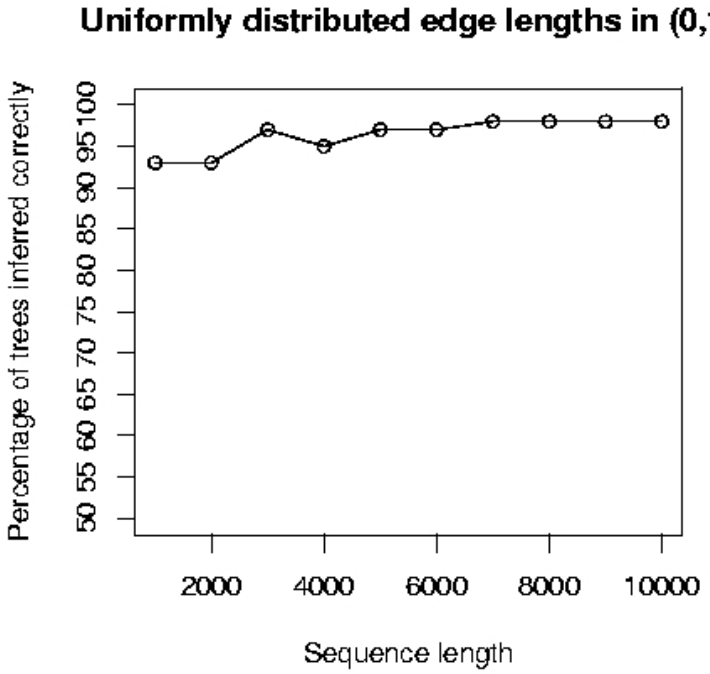


Figura 4: Percentatge d'arbres reconstruïts correctament amb longituds de branca uniformement distribuïdes entre 0 and 1.

l'*evolver* 10 conjunts de seqüències i vam aplicar l'algoritme a cadascun dels alineaments obtinguts. Vam repetir aquest procés 10 vegades, generant així un total de 100 alineaments per a cada mitjana i longitud de seqüències. Els resultats obtinguts es recullen a la figura 5. Observem que per a $\mu = 0.25$ o $\mu = 0.5$ és suficient consi-

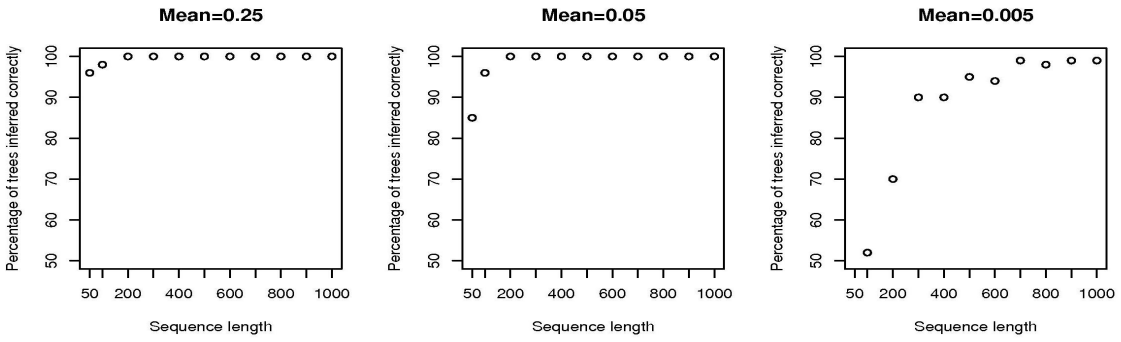


Figura 5: Percentatge d'arbres reconstruïts correctament amb longituds de branca seguint una normal de mitjana 0.25, 0.05 i 0.005.

derar seqüències de longitud 200 per a què l'algoritme sigui un 100% efectiu. Vam també considerar una mitjana menor $\mu = 0.0005$ però en aquest cas només vam obtenir una eficiència superior al 90% per a seqüències de longitud major o igual que 3000.

El mètode proposat aquí no és ni de bon tros l'únic mètode que usi invariants filogenètics per a l'inferència d'arbres d'espècies. Fins ara només s'havien usat invariants lineals i quadràtics amb resultats clarament decebedors (és evident que l'envolvent lineal d'una varietat no n'és una bona aproximació). Al capítol 19 del llibre [PS05] es dóna un altre algoritme d'ús d'invariants per al "general Markov model". Els següents passos en la interacció entre la geometria algebraica i la filogenètica comportaran la creació de nous algoritmes d'ús dels invariants per a arbres arbitraris i la comparació amb altres mètodes ja usats pels biòlegs.

Referències

[CF87] J. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71, 1987.

[CGS05] M. Casanellas, L.D. Garcia, and S. Sullivant. Catalog of small trees. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for computational biology*, chapter 15. Cambridge University Press, 2005.

[CLO97] D Cox, J Little, and D O'Shea. *Ideals, Varieties, and Algorithms*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1997. An introduction to computational algebraic geometry and commutative algebra.

[CS05] M. Casanellas and S. Sullivant. The strand symmetric model. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for computational biology*, chapter 16. Cambridge University Press, 2005.

[ES93] S Evans and T Speed. Invariants of some probability models used in phylogenetic inference. *The Annals of Statistics*, 21:355–377, 1993.

[GPS03] GM Greuel, G. Pfister, and H. Schoenemann. Singular: A computer algebra system for polynomial computations. Available at `http://www.singular.uni-kl.de/`, 2003.

[JWMV03] K St. John, T Warnow, B Moret, and L Vawter. Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor joining. *Journal of Algorithms*, 48:174–193, 2003.

[Lak87] J.A. Lake. A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, 4:167–191, 1987.

[PRW00] G Pistone, E Riccomagno, and HP Wynn. *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman & Hall/CRC, December 2000.

[PS05] L. Pachter and B. Sturmfels, editors. *Algebraic Statistics for computational biology*. Cambride University Press, October 2005. ISBN-10 0521857007.

[RG96] I Ross and R Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[SS05] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12:204–228, 2005.

[Yan97] Z Yang. PAML: A program package for phylogenetic analysis by maximum likelihood. *CABIOS*, 15:555–556, 1997.