]Filogenètica i Geometria Algebraica

[

Dionís Remón

4 de setembre de 2007

4 de setembre de 2007

Introducció

L'objectiu d'aquest treball és donar una breu introducció al que és un nou horitzó per a la geometria algebraica. Recentment han hagut avanços en filogenètica que han posat de manifest que certs ideals, provinents de varietats algebraiques, poden resultar utils alhora d'inferir arbres filogènetics, i de retruc, essencials per a conèixer com s'ha donat l'evolució de les espècies.

Darrera d'això esta l'Estadística Algebraica. i la relació consisteix en que molts models estadístics tenen associada una varietat algebraica. 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 i altres tasques pròpies de l'estadística. Les aportacions de l'estadística algebraica a la biologia computacional que s'han fet fins aquest moment es troben recollides al llibre de L. Patcher i B. Sturmfels [Pat05].

El present treball treball gira entorn de l'article de B. Sturmfels i S. Sullivant [Stu05]. Aquest article dóna un algoritme de construcció d'invariants algebrics que permeten inferir arbres filogenètics que seran els invariant que s'anomenara filogenètics.

El treball esta format pels 4 blocs següents.

El primer esta format pels capítols 1 i 2. El capítol 1 és un petit compendi de biologia bàsica que ens donarà certes definicions i conceptes de biologia relacionat amb el tema que tractarem. Aquests textos estan extrets bàsicament de documents de la web

i de llibres de divulgació sobre biologia. El capítol 2 dóna una breu introducció històrica a la filogenètica així com un breu resum sobre les tècniques matemàtiques emprades actualment, en aquest context.

El segon bloc esta format pels capítols 3 i 5 i formen la base d'aquest treball. En aquests capítols es farà una construcció dels invariants filogenètics per a cert tipus arbres, tal i com ve explicat en de B. Sturmfels i S. Sullivant, [Stu05]. En la part final del capítol 5 es donarà una breu introducció al altres algoritmes per al càlcul d'invariants. En tots dos capítols s'explicitaran tots els càlculs per a un exemple concret per tal de fer totes les definicions i teoremes donats més entenedors.

El tercer bloc, esta format pel capítol 4. Tracta sobre bases de Gröbner i varietats tòriques, conceptes que són clau per a entendre els resultats del capítol 5. La referència principal d'aquest capítol és el llibre de R. Fröberg [Fro98].

En l'últim bloc, que corresponent a l'últim capítol, és donaran uns exemples de com s'aplica la teoria explicada en la inferència d'arbres filogenètics. Els exemples provenen dels articles, de M. Casanelles, [Cas06] i de l'article de L. Patcher, B. Sturmfels [Pat07].

Índex

Capítol 1

Breu lliçó de biologia

En aquest capítol donarem una breu introducció sobre la genètica, des del punt de vista de la biologia, per tal de justificar alguns fets matemàtics més endavant.

1.1 Sobre la genètica

Tots els organismes vius de la Terra tenen alguna cosa en comú. Disposen d'un sistema d'informació, el qual es responsable del control de les funcions de cada organisme i també del manteniment de les característiques que distingeixen les espècies entre si d'una generació a la següent.

1.1.1 Definició. Al conjunt de la informació que distingeix les espècies l'anomenarem informació gènica.

1.1.2 Definició. El lloc on es troba aquesta informació s'anomena cromosoma, i esta situat dintre de cada una de les cel·lules.

Si estem parlant d'organismes procariotes (per exemple virus i bactèries), els cromosomes estan la regió de l'interior de la

Animal	Número cromosomes
Mosca	5
Ordi	14
Blat	42
Llebre	46
Humà	46
Pollastre	78

Taula 1.1: Dotació cromosòmica de diverses espècies

cel·lula. La resta d'organismes s'anomenen eucariotes i tenen un nucli dins de la cel·lula on els se situen els cromosomes. D'ara endavant només tractarem els organismes eucariotes.

1.1.3 Definició. El conjunt format per tota la informació genètica d'una espècie s'anomena genoma.

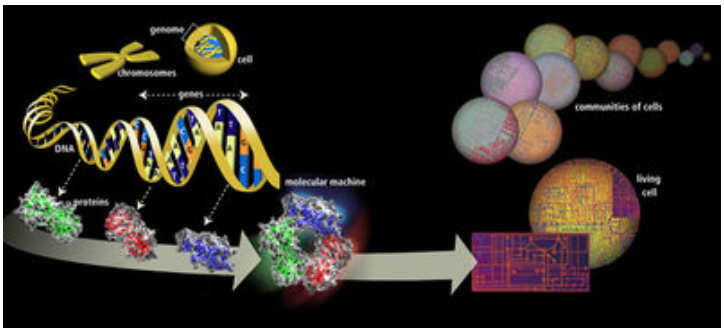


Fig. 1.1: Esquema del genoma

Aquesta informació es localitza en un nombre fix de cromosomes, que constitueix la dotació cromosòmica.

En el cas dels humans, es tenen 46 cromosomes agrupats en 23 parelles: 22 autosomes (iguals en els homes que en les dones) i dos cromosomes sexuals, denotats per X i Y. En la taula 1.1 podem observar la quantitat de cromosomes d'algunes espècies.

Els cromosomes estan formats per àcid desoxiribonucleic (ADN), la substància en què esta codificada la informació genètica i els

histones, que és una substància que manté unit el ADN en forma de cadena i que permet modificar la posició de la cadena per a adaptar-les a llur funció.

El ADN és una gran molècula composta per desoxiribosa, àcid fosfòric i bases nitrogenades. El conjunt format per una molècula de desoxiribosa, un àcid fosfòric i una base nitrogenada és un nucleòtid.

Al costat d'una cadena de nucleòtids se'n situa una altra que esta formada per subunitats d'àcid fosfòric i de desoxiribosa, amb les bases corresponents però en posició antiparal·lela (veure figura 1.2). Això vol dir que queden confrontats l'extrem del carboni 5 d'una cadena amb l'extrem corresponent al carboni 3 de l'altra cadena.

Totes dues cadenes s'uneixen formant una doble cadena, mitjançant ponts d'hidrogen que s'estableixen entre les bases nitrogenades. Aquestes parelles de bases són complementaries. Més endavant direm que vol dir que són complementaries.

Finalment, tot el conjunt està enrotllat en espiral i constitueix la doble hèlix (cf. [Wat53]). La descoberta d'aquest fet donar l'any 1962 el premi Nobel de medicina a James Watson i Francis Crick. Cada una de les hèlix és complementaria del altra.

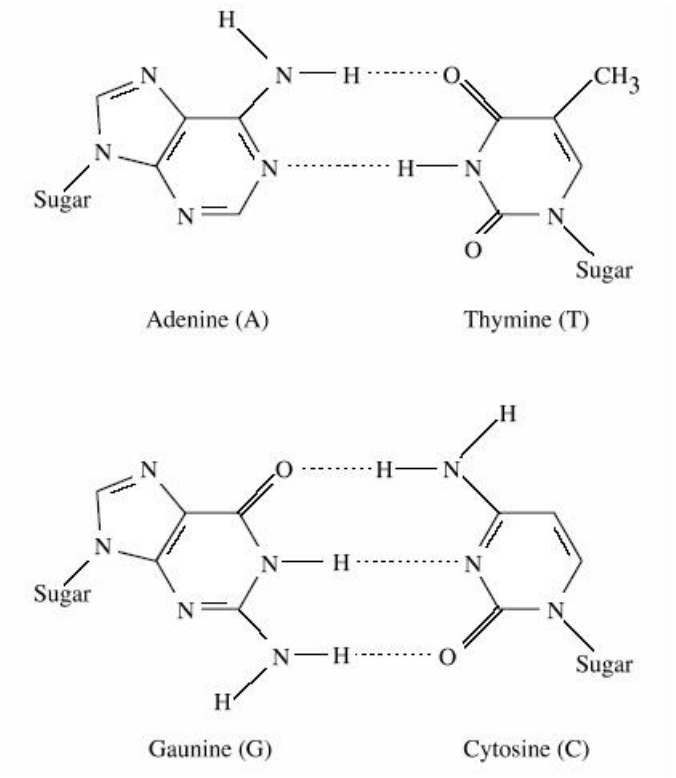


Fig. 1.2: Dibuix d'una cadena de DNA

Hi han quatre bases nitrogenades diferents, que es denoten amb les lletres {`A`, `C`, `G`, `T`}. La `A` correspon a l'adenina, la `C` a la citosina, la `G` a la guanina i la `T` a timina. L'adenina i la guanina són les purines, i la citosina i la timina són les pirimidines. Direm que la base `A` es la complementaria de `T` i viceversa, i de manera anàloga `C` amb `G`.

La informació gènica que conté el ADN depèn de l'ordre en què es troben aquestes bases nitrogenades. Les paraules de l'idioma del ADN es formen amb tres de les bases anteriors. Cada un d'aquest conjunt de bases s'anomena triplet i codifica un aminoàcid. Hi ha 20 aminoàcids diferents. A més hi ha tres triplets `TAA`, `TAG`, `TGA` que són especials; en lloc de codificar un aminoàcid, indiquen el lloc on acaba la proteïna.

TTT $\mapsto$ Phe	TCT $\mapsto$ Ser	TAT $\mapsto$ Tyr	TGT $\mapsto$ Cys
TTC $\mapsto$ Phe	TCC $\mapsto$ Ser	TAC $\mapsto$ Tyr	TGC $\mapsto$ Cys
TTA $\mapsto$ Leu	TCA $\mapsto$ Ser	TAA $\mapsto$ *stop*	TGA $\mapsto$ *stop*
TTG $\mapsto$ Leu	TCG $\mapsto$ Ser	TAG $\mapsto$ *stop*	TGG $\mapsto$ Trp
CTT $\mapsto$ Leu	CCT $\mapsto$ Pro	CAT $\mapsto$ His	CGT $\mapsto$ Arg
CTC $\mapsto$ Leu	CCC $\mapsto$ Pro	CAC $\mapsto$ His	CGC $\mapsto$ Arg
CTA $\mapsto$ Leu	CCA $\mapsto$ Pro	CAA $\mapsto$ Gln	CGA $\mapsto$ Arg
CTG $\mapsto$ Leu	CCG $\mapsto$ Pro	CAG $\mapsto$ Gln	CGG $\mapsto$ Arg
ATT $\mapsto$ Ile	ACT $\mapsto$ Thr	AAT $\mapsto$ Asn	AGT $\mapsto$ Ser
ATC $\mapsto$ Ile	ACC $\mapsto$ Thr	AAC $\mapsto$ Asn	AGC $\mapsto$ Ser
ATA $\mapsto$ Ile	ACA $\mapsto$ Thr	AAA $\mapsto$ Lys	AGA $\mapsto$ Arg
ATG $\mapsto$ Met	ACG $\mapsto$ Thr	AAG $\mapsto$ Lys	AGG $\mapsto$ Arg
GTT $\mapsto$ Val	GCT $\mapsto$ Ala	GAT $\mapsto$ Asp	GGT $\mapsto$ Gly
GTC $\mapsto$ Val	GCC $\mapsto$ Ala	GAC $\mapsto$ Asp	GGC $\mapsto$ Gly
GTA $\mapsto$ Val	GCA $\mapsto$ Ala	GAA $\mapsto$ Glu	GGA $\mapsto$ Gly
GTG $\mapsto$ Val	GCG $\mapsto$ Ala	GAG $\mapsto$ Glu	GGG $\mapsto$ Gly

Taula 1.2: Codi genètic

1.1.4 Definició. Un gen és un fragment de ADN que conté informació per a un tret característic d'una determinada espècie, també anomenat caràcter hereditari.

El genoma humà té aproximadament 25.000 gens, encara que no se sap el nombre exacte. La dificultat de descobrir nous gens rau en la gran quantitat d'elements del genoma que no constitueixen un gen. Només un 5% del genoma és gènic, és a dir, és funcional.

Les diferències entre genomes d'individus d'una població són petites i principalment degudes a events de recombinació (procés pel qual dos copies dels cromosomes paterns s'uneixen en la descendència). Però el que estudiarem en aquest treball són les diferències entre espècies.

Hi han varis motius pels quals es poden explicar aquestes diferències. Al llarg del temps les cadenes han sofert canvis degut

a les mutacions. Les mutacions més freqüents corresponent a les permutacions de les bases. Aquestes permutacions de les bases són de dos tipus. Les transicions són intercanvis d'una base purina (`A` $\leftrightarrow$ `G`) o d'una base pirimida (`C` $\leftrightarrow$ `T`). Les tranversions són intercanvis entre bases purines i pirimides. Encara que aquestes últimes són possibles els mecanismes moleculars fan que les transicions siguin més freqüents.

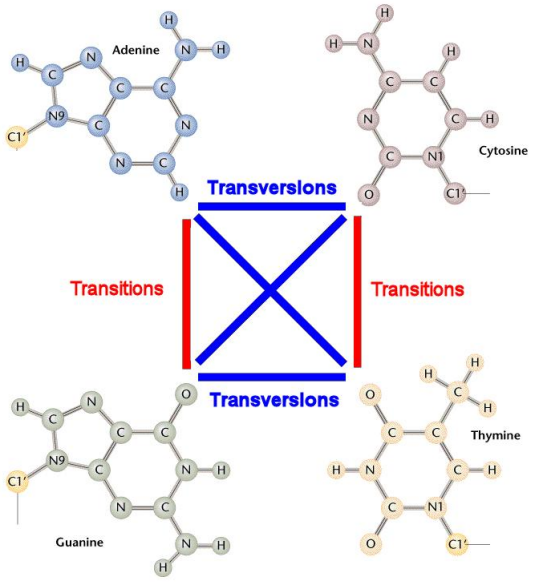


Fig. 1.3: Esquema de les transversions i transicions

També hi ha mutacions de caràcter més massiu que afecten a varies bases alhora. Es poden resumir en els quatre següents:

- *Reordenament genòmic.* Comparant cromosomes d'espècies, es pot veure en alguns casos que llargs segments de la cadena d'ADN han estat revertits (p.e. `AGG` passa a `GGA`), o bé intercanviats amb el corresponent fragment de la doble hèlix (p. e. `AGG` passa a `TCC`), o altres events de major magnitud, pel que fa a la quantitat de bases implicades com poden ser fusions entre cromosomes.

- *Duplicacions i pèrdues.* Alguns genomes han sofert una duplicació total. Habitualment, això repercuteix en una pèrdua de gens ja que els gens redundants perden funcionalitat, encara que en alguns casos donen lloc a alguna funcionalitat nova.

- *Expansió parasitària.* Alguns elements del genoma són repetitius, consisteixen en elements els quals poden duplicar-se i reintegrar-se en el genoma.

- *Punt de mutació.* Les seqüències de ADN muten, i en les regions no funcionals aquestes mutacions es van acumulant durant temps i poden esdevenir funcionals.

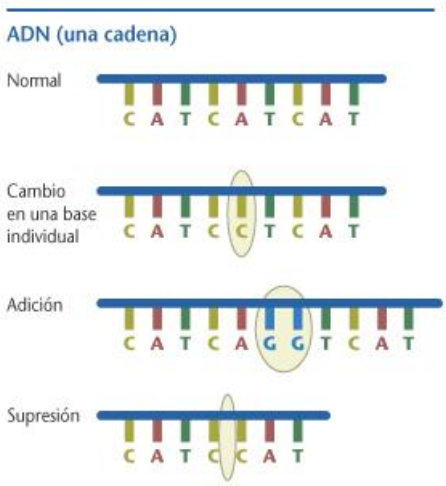


Fig. 1.4: Resum d'algunes mutacions massives.

Capítol 2

Filogenètica

En aquest capítol una explicació de que és la filogenètica, dels seus orígens, i de les tècniques que s'empren.

2.1 Introducció històrica

La filogenètica és una branca de la biología sistemàtica. L'objecte d'estudi de la biologia sistemàtica és la diversitat biològica en tots els nivells, és a dir, des de molécules i gens, fins a espècies i ecosistemes. Entre les nombroses disciplines d'aquests camp destaquen, per a l'objectiu d'aquest treball, les ciències que s'apliquen per a explicar fenòmens o funcions biológiques. En particular ens centrarem en la filogènetica.

La primera vegada que sonà el terme filogènia, fou en l'obra del biòleg i filòsof alemany Ernst Hæckel. Nascut a Postdam l'any 1824 estudià medicina en les universitats de Berlí, Wurzburg i Viena. Fou catedràtic de zoologia en la Universitat de Jena on morí l'any 1909. Les contribucions de Hæckel a la zoología foren una barreja de investigació i especulació. Fou un fervent evolucionista. Les seves idees al respecte foren recollides

en l'obra *Generelle Morphologie der Organismen* on formulà la seva teoria de la recapitulació que deia, a grans trets, que el desenvolupament d'un embrió de cada espècie repeteix el desenvolupament evolutiu d'aquesta espècie totalment.

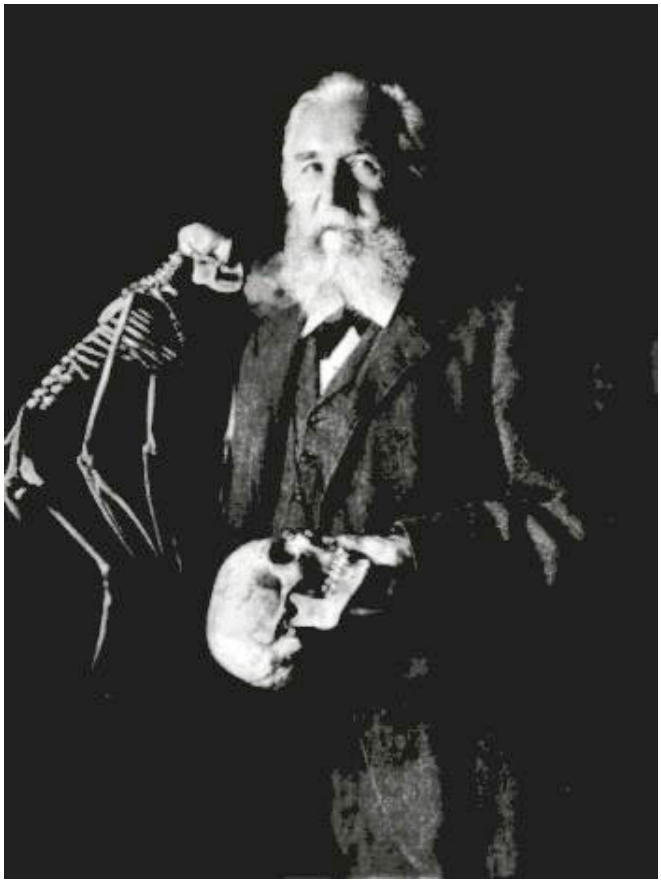


Fig. 2.1: Ernest Heinrich Philipp August Haeckel (1824 - 1909)

En aquesta obra dividí el camp de la morfologia, en dos subcamps, l'anatomía i la morfogènia. Aquesta última en dos, la ontogènia i la filogènia. La primera fa referència a la història del desenvolupament del individu i la segona fa referència a una classificació que reflexa la història evolutiva d'una especie o grup.

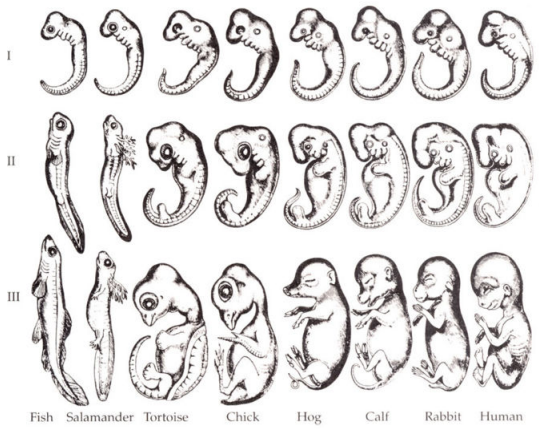


Fig. 2.2: Dibuixos per a justificar la th. de la recapitulació

La reconstrucció d'una filogènia es basa en la búsqueda de caràcters compartits per a grups d'organismes. Una "cronologia"de l'aparació d'aquests caràcters ens porta a la construcció d'un arbre filogenètic del grup en qüestió, en que les espècies ancestrals són representades com a punts de ramificació del arbre.

Actualment s'entén que la filogènia expressa les relacions d'una espècie o grup com a resultat de l'evolució deguda a la descendència i a la transformació de caràcters hereditaris. La sistemàtica filogenètica té les següents idees bàsiques:

- Intentar recuperar les relacions filogenètiques (genealògiques) entre grups d'organismes.
- Produïr classificacions que reflexen exactament aquestes relacions genealògiques.

La comprensió de la filogenètica és important per a predir les característiques dels organismes mitjançant les característiques del seus similars. Entendre les relacions dels organismes amb altres espècies és clau per a entendre les seves característiques.

2.1.1 Exemples. També és important per a prevenir comparacions inadequades basades en relacions no existents. Per exemple, durant anys fou habitual usar *Euglena* com a mòdel de sistema unicel·lular per a estudiar la fotosíntesi, però aquest organisme no es relaciona amb les plantes i per tant la part de la seva cel·lula que fa la seva fotosíntesis no té res a veure amb el cloroplast de les plantes.

2.2 Matemàtiques i filogenètica

La genètica, la filogenètica i les matemàtiques es poden combinar en l'objectiu de conèixer caràcters comuns de les espècies. Per exemple una bona dada per a conèixer pot ser, determinar a nivell genètic una característica comú dels vertebrats. Hi ha moltes espècies de vertebrats diferents des de l'home fins la tonyina. Existeixen una sèrie de gens que tenen la funció de "vertebrar"l'organisme. Mitjançant el coneixement dels codis genètics d'algunes espècies de vertebrats (en el cas que s'ha estdiat en aquest treballs les espècies són concretament la dels chimpanzés, rates, ratolins, gossos, pollastres, granotes, peixos zebra, peixos fugú i el tetraodons) i gràcies a càlculs sobre la freqüència d''aparició de les bases s'arribat a la següent conjectura:

2.2.1 Conjetura. (Significat de la Vida) *La seqüencia de 42 bases*

`TTTAATTGAAAGAAGTTAATTGAATGTGAAAATGATCAACTAGG`

estaba present en el genoma del antecesor de tots els vertebrats, i s'ha conservat completament fins els nostres dies (i. e., ninguna de les bases ha mutat, ni han hagut insersión ni eliminacions).

En l'últim capítol donarem una breu explicació de com s'ha arribat a provar aquesta conjectura.

De fet, aquesta conjectura va ser formulada la primavera del 2004 i actualment ha estat àmpliament superada en trobar una seqüència de 145 bases presents en tots els genomes de les espècies de vertebrats estudiades.

En resum, és possible doncs que ben aviat coneguem el codi genétic del *Myllokunmingia*, la primera espècie de vertebrat de la qual es té actualment, restes fossils mitjançant l'unió de tres branques de la ciencia aparentment tan distanciades.

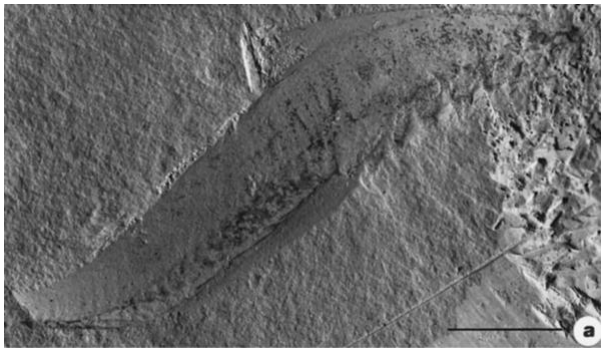


Fig. 2.3. *Myllokunmingia fengjiaoa*, primer vertebrat amb restes fòsils

2.3 Metodologia de la filogenètica.

Tal com hem vist, l'objectiu de la filogenètica és agrupar organismes en espècies i aquestes al seu torn en grups taxonòmics més grans.

De forma clàssica aquestes relacions venen representades gràficament en forma dels anomenats arbres.

2.3.1 Definició. Un arbre filogenètic (d'ara endavant només arbre) és una representació gràfica de les relacions filogenètiques entre les espècies. Un arbre filogènetic també es pot anomenar cladograma.

Un arbre filogenètic esta format per nodes i branques. Cada branca té dos nodes. De cada node pot sortir algun nombre de branques. Direm que un node es terminal, si d'ell només en surt una branca. En cas contrari direm que és un node interior. Un node que no és interior, s'anomena fulla. Dintre dels nodes interior, pot haver-hi un node distingit, anomenat node arrel. El node arrel indica una certa orientació en les branques.

Hi ha dos tipus d'arbre filogenètics diferents:

2.3.2 Definició. Es diu que un arbre filogenètic és arrelat si es considera un node arrel. En cas contrari s'anomena arbre filogenètic no arrelat.

En cas dels arbres arrelats, són necessaris per estudiar els ancestres comuns a les espècies, així com l'aparició dels caràcters.

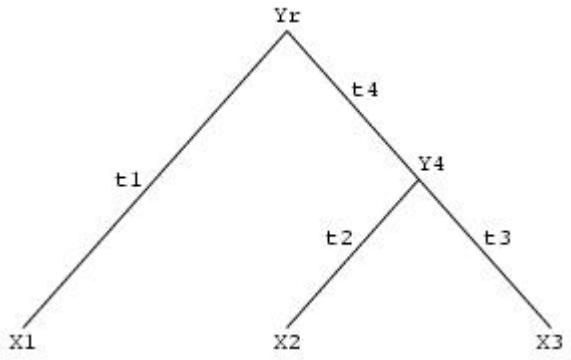


Fig. 2.4.: Exemples d'arbre arrelat

S'usen arbres sense arrels quan el que ens interesa és destacar, més que la relació amb els ancestres comuns, les relacions entre si. Per exemple, davant el problema de classificar quatre espècies segons característiques comunes, on no ens importen els ancestres comuns, tenim tres possibles solucions, que anomenem topologies i que estan representades en la figura 2.4.

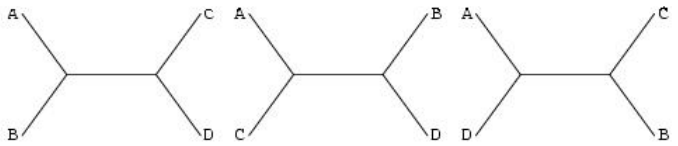


Figura 2.4. Topologies possible en un arbre de quatre fulles sense arrel.

2.3.3 Definició. Anomenem topologia d'un arbre a una ordenació de les fulles.

Per tal de classificar les espècies, antigament es compilaven les seves dades a comparar i es feia un estudi "multivariant mental", i se seleccionava una estructura de classificació, sense tenir en compte la viabilitat de les topologies desestimades.

La descripció matemàtica dels arbres en termes de teoria de grafs i teoria de probabilitat ha permès quantificar l'estructura de cada arbre possible així com mesurar certes propietats i comparar-les amb els arbres alternatius per a la mateixa col·lecció d'unitats de mostreig. Així doncs podem entendre la filògenica es pot resumir doncs en la búsqueda de caràcters i en l'anàlisis d'aquest.

En aquesta branca hi ha dos criteris principals alhora d'inferir arbres filogenètics, els métodes de parsimonia i els métodes probabístics.

2.3.1 Mètodes de parsimonia.

Aquests métodes permeten un examen més intuïtiu de la posible evolució dels caràcters. Quan es conten el total de canvis en cada topologia segons el ordre d'estats especificats pel model, s'obté el nombre de pasos com a estimació de la longitud total del arbre. Hi ha varis tipus de models de parsimonia, però en qualsevol cas la topologia que es considera l'òptima és aquella que recull el menor nombre de pasos posible.

La familia de mètodes de parsimonia sorgeixen en tant en quant no tots els canvis en les topologies tenen el mateix pes.

2.3.2 Métodes probabilístics.

Aquests métodes examinen que de bé explica un arbre les dades observades. En principi, cada arbre possible implica diferents probabilitats per a les diferents configuracions particulars de dades. La pregunta de la estimació estadística és: quina és la probabilitat de que les dades observades corresponguin a les dades predites per a una hipòtesi o model particular? El procediment de selecció d'un arbre filogenètic adequat per a les dades observades és un procediment comparable a la selecció d'un valor promig x com la millor estimació de la mitjana μ.

Dintre d'aquest camp es troba el que estudiarem en aquest treball. Per a fer una breu introducció indicarem que hi ha una branca recent de les matemàtiques que estudia la relació entre la probabilitat i la geometria algebraica, que s'anomena estadística algebraica.

En el nostre cas cada topologia de l'arbre la entendrem com un punt en un cert espai projectiu i el més adequat serà el més proper a una certa varietat.

Capítol 3

Models algebraics en filogenètica

3.1 Definicions i conceptes bàsics

En aquest apartat definirem els conceptes bàsics. Primer de tot farem unes suposicions per tal de simplificar el model de manera que sigui tractable de manera tant algebraica com computacional. Per a començar suposarem les cinc sentències següents:

1. El arbres seran binaris, és a dir, del node arrel sortiran dues branques i aquesta es dividirà en dues més fins arribar a les fulles. Més endavant obviarem aquesta condició.

2. L'evolució de l'especie depèn només del node anterior.

3. Les mutacions ocorren aleatòriament i la probabilitat que es produeixin és positiva.

4. Suposarem donat un alineament de les seqüències d'ADN de les espècies. A causa de processos de mutació i d'altres

explicats en el primer capítol, les seqüències de les diferents espècies tenen parts idèntiques, part que s'assemblen i parts que no es poden comparar. A més les parts comparables no tenen perquè estar en el mateix lloc del genoma. És per això que abans d'estudiar parts de l'ADN ens interessarà saber quines parts dels genomes de les espècies es corresponent entre si.

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

Sigui T un arbre qualsevol amb n fulles i m branques. Enumerem els nodes de l'arbre (incloent les fulles) d'esquerra a dreta i de baix a dalt i anomenem t_i a la branca que puja des del node i. El node arrel s'anomenarà r. A cada node 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", les dels nodes interiors seran ocultes (perquè no en tindrem cap observació) i les anomenarem Y_i.

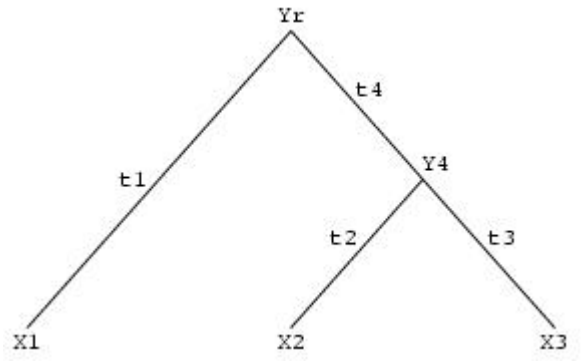


Fig. 3.1.: Exemples d'arbre de tres fulles

A cada branca li associem una matriu S_i les entrades de la qual són les probabilitats $P(x|y, t_i)$, de que un nucleòtid x en

el node pare muti a un nucleòtid y en el node fill al llarg d'una branca de longitud t_i:

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

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

La probabilitat d'observar els nucleòtids $x_1, x_2, ..., x_n$ en les fulles s'expressa en funció de les entrades de les matrius de substitució de la següent manera:

$$p_{x_1 x_2 \cdots x_n} = \sum_{\{(x_v)_{v\in N(T)} | x_v \in \sigma\}} \prod_{e\in E(T)} S(x_{p(t_i)}, x_{g(t_i)}), \qquad (3.1)$$

on, $E(T)$ denota el conjunt de branques de l'arbre T, σ, el conjunt de nodes interns de l'arbre T, on $p(t_i)$ denota el nucleotid del node pare, i $g(t_i)$ el nucleòtid del node fill. Tenim diferents models estadístics d'evolució segons la forma que tinguin les matrius de substitució i segons la distribució de nucleòtids en l'arrel.

3.1.1 Models de grup.

Els models més usats en filogenètica són els *models de grup*. En aquests models la distribució de nucleòtids en 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 $a_i + b_i + c_i + d_i = 1$ i

$$\begin{array}{llll} a_i & = P(\mathtt{A} \to \mathtt{A}) & = P(\mathtt{C} \to \mathtt{C}) & = P(\mathtt{G} \to \mathtt{G}) & = P(\mathtt{T} \to \mathtt{T}) \\ b_i & = P(\mathtt{A} \to \mathtt{C}) & = P(\mathtt{C} \to \mathtt{A}) & = P(\mathtt{G} \to \mathtt{T}) & = P(\mathtt{T} \to \mathtt{G}) \\ c_i & = P(\mathtt{A} \to \mathtt{G}) & = P(\mathtt{G} \to \mathtt{A}) & = P(\mathtt{C} \to \mathtt{T}) & = P(\mathtt{T} \to \mathtt{C}) \\ d_i & = P(\mathtt{A} \to \mathtt{T}) & = P(\mathtt{T} \to \mathtt{A}) & = P(\mathtt{C} \to \mathtt{G}) & = P(\mathtt{G} \to \mathtt{C}) \end{array}$$

La justificació biològica per a aquest mòdel (anomenat Kimura de 3 paràmetres) tal com hem vist en la primera secció, vol reflectir el fet que les transicions tenen probabilitat diferent de donar-se que les transversions, en aquest cas són més freqüents les transicions que les transversions. Casos particulars d'aquest model són el model de Kimura de 2 paràmetres (només un paràmetre per a les transcicions un altre per a les transversions, és a dir $b_i = d_i$) i Jukes-Cantor (és el model més simple ja que només distingeix si hi ha mutació o no hi ha, és a dir $b_i = c_i = d_i$).

3.1.2 Strand symmetric model.

En aquest model no se suposa que la 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 realitat, i són els que s'usen normalment per a inferir seqüències codificants (gens o parts de gens). Les regions codificants contenen més $\mathtt{C}, \mathtt{G}$ que no pas $\mathtt{A}, \mathtt{T}$. S'ha comprovat en diversos estudis que $\pi_{\mathtt{C}} \sim \pi_{\mathtt{G}}$ i $\pi_{\mathtt{A}} \sim \pi_{\mathtt{T}}$. D'altra banda, si $\mathtt{A}$ muta a $\mathtt{C}$, aleshores, en la cadena d'ADN complementaria es produeix una mutació de $\mathtt{T}$ a $\mathtt{G}$. Així es natural requerir que les matrius de substitució siguin de la forma

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

3.1.3 Altres models.

En aquest grup podem incloure el 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 paràmetres sempre s'adequa més a la realitat però augmentar el nombre de paràmetres provoca que la complexitat dels càlculs sigui molt més gran alhora d'inferir filogènies. En aquest cas els càlculs són totalment inviables, i tampoc s'obtindrien millores considerables.

3.2 Invariants filogènetics

Els mòdels descrits en la secció anterior són models algebraics. S'anomenen així perquè les probabilitats conjuntes de les variables observades s'expressen com a funció polinòmica en els paràmetres. Així, un model evolutiu algebraic de d paràmetres lliures sobre un arbre T de n fulles ve descrit per la següent aplicació polinomial

$$\begin{array}{rccl} \varphi: & k^d & \to & k^{4^n} \\ & \theta & \mapsto & (p_{\mathtt{AA}\cdots\mathtt{A}}, p_{\mathtt{AA}\cdots\mathtt{C}}, \cdots, p_{\mathtt{TT}\cdots\mathtt{T}}). \end{array}$$

3.2.1 Observació. Donada l'aplicació anterior observem que la n correspon al nombre de fulles de l'arbre, la d correspon al nombre de paràmetres (del model escollit) i φ depèn de l'arbre que estiguem estudiant.

La topologia de cada arbre ens donarà un punt de l'espai d'arribada.

Aquesta aplicació parametritza 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

$$V = Z(f_1, ..., f_r)$$

on $f_1, ..., f_r \in k[x_1, ..., x_n]$, on k és un cos. La imatge d'una aplicació polinomial no és en general la varietat algebraica que la conté. Les varietats algebraiques són els tancats de la topologia de Zariski de k^n. A partir d'ara quan parlem de la imatge d'una aplicació polinomial φ, ens referirem a la seva clausura. Així $\text{im}(k^d)$ 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 de definició. 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 X,

$$I(X) = \{f(x_1, ..., x_n) | f \in k[x_1, ..., x_n], f(p) = 0 \text{ per a tot } \ p \in X\}.$$

És molt senzill de demostrar que $I(X)$ és un ideal de l'anell de polinomis $k[x_1, ..., x_n]$. 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 $g_1, ..., g_s \in k[x_1, ..., x_n]$ tals que $I(X) = (g_1, ..., g_s)$.

3.2.2 Exemples. En el cas del model de Jukes-Cantor tindrem que $d = 2$. En el cas de Kimura 2 de paràmetres, respectivament 3 de paràmetres, tindrem que $d = 3$, respectivament $d = 4$.

3.2.3 Definició. Sigui V la clausura de la imatge 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.

Es pot observar que els invariants $I(V)$ depenen només de l'aplicació ϕ, és a dir, de l'arbre. Així doncs, podríem tenir que cada arbre tingues els seus propis invariants filogenètics i lineals. Però degut a la construcció d'aquests, resulta que molts d'ells coincideixen.

Per a obtenir l'ideal de les varietats algebraiques corresponents als models de grups en serà molt útil fer un canvi de coordenades en les indeterminades $p_{x_1 \ldots x_n}$. Pensem els caràcters $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ com elements del grup $\mathbb{Z}/(2) \times \mathbb{Z}/(2)$. Així podem veure les matrius de substitució com a certes funcions sobre el grup:

$$S_i(g,h) = f^i(h-g), \quad g,h \in G \tag{3.2}$$

(és pot deduir com és f segons el model) i d'aquesta manera podem escriure les probabilitats conjuntes a les fulles com a funcions de $G \times G \times \cdots \times G$:

$$p(g_1, ..., g_n) = \frac{1}{4} \sum \prod_{b \in E(T)} f^b(g_{p(b)} - g_{f(b)})$$

on la suma és sobre tots els el 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 donada una funció $f : G \to k$, on G denota un grup finit (en cas contrari seria la transformada de Fourier habitual), la seva transformada de Fourier discreta $\hat{f} : \hat{G} = Hom(G, k^*) \to k$ ve donada per

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

Una de les propietats més útils de la transformada de Fourier és que si tenim una convolució de dues

$$(f_1 * f_2)(g) = \sum_{h \in G} f_1(h) f_2(g-h),$$

la seva transformada és el producte de transformades: $\widehat{f_1 * f_2} = \hat{f}_1 \cdot \hat{f}_2$. Així es va poder demostrar el següent teorema:

3.2.4 Teorema. (Evans-Speed) *Per a un model basat en grup sobre un arbre T la transformada de Fourier de la distribució conjunta $p(g_1, ..., g_m)$ té la forma*

$$q(\chi_1, ..., \chi_n) = \prod_{b \in E(T)} \widehat{f^b} \left(\prod_{l \in \Lambda(b)} \chi_l \right), \tag{3.3}$$

on $E(T)$ denota el conjunt de branques de l'arbre T i $\Lambda(b)$ el conjunt de fulles per sota de la branca b.

Demostració. Cf. [Eva93]. □

Per tant l'expressió polinomial de la probabilitat conjunta 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 esta generat per binomis. Això fa que en aquestes noves coordenades de Fourier sigui fàcil de 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.

El que farem més endavant serà donar un algoritme per al càlcul d'un nombre mínim de generadors de l'ideal $I(V)$, o en l'ideal $I(V)$ mòdul els invariants lineals. Resulta que aquest algoritme es pot resumir en el teorema següent:

3.2.5 Teorema. (Sturmfels-Sullivant) *Per als models de grup, l'ideal corresponent a un arbre filogenètic qualsevol està generat per binomis de grau com a màxim 4 en les coordenades de Fourier. A més, l'ideal d'un arbre arbitrari es pot descriure a partir de l'ideal d'un arbre de tres fulles sense arrel.*

Demostració. Cf. Capítol 5. □

3.3 Model de Jukes-Cantor

El que farem en aquesta secció serà desenvolupar el model de Jukes-Cantor, per tal d'introduir els diferents conceptes que intervenen en la filogenètica, explicats en les seccions anteriors. En particular desenvoluparem l'arbre de tres fulles arrelat. Per a no complicar l'exemple, farem el cas que la probabilitat de les diferents bases en el node arrels siguin iguals.

Considerem l'arbres de tres fulles donat per la figura 3.3.

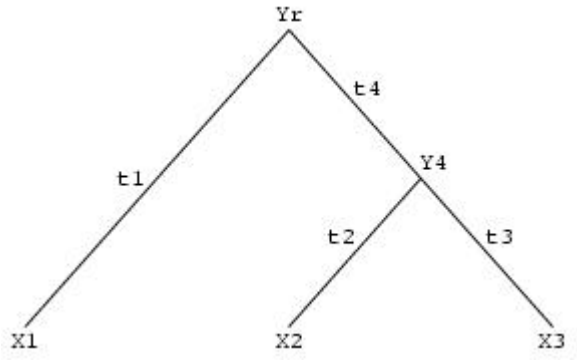


Fig. 3.2: Arbre de tres fulles trivalent

Com hem indicat en la secció anterior, les matrius de substitució del model de Jukes-Cantor són de la forma

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

on S_i és la matriu en cada la branca t_i, segons la figura anterior. Venen a representar un model força senzill que només distingeix el fet de mutar o no muntar. Es pot deduir del dit abans que $a_i + 3b_i = 1$, per tant $b_i = (1 - a_i)/3$.

Suposarem que la distribució dels nucleòtids és uniforme en

l'arrel, és a dir $\pi_{\mathtt{A}} = \pi_{\mathtt{C}} = \pi_{\mathtt{G}} = \pi_{\mathtt{T}} = \frac{1}{4}$ El càlculs de les probabilitats serien

$$p_{\mathtt{AAA}} = \frac{1}{4}(a_1(a_4a_3a_2 + 3b_4b_3b_2)) + \frac{3}{4}(b_1(2b_4b_3b_2 + a_4b_3b_2 + b_4a_3a_2)).$$

Es pot observar que aquesta fórmula és invariant per nucleòtid, és a dir,

$$p_{\mathtt{AAA}} = p_{\mathtt{CCC}} = p_{\mathtt{GGG}} = p_{\mathtt{TTT}}.$$

A aquests valors de probabilitats els anomenarem p_{123} fent referència a que és la probabilitat que les tres bases siguin iguals. De manera anàloga calculem p_{12}, p_{13}, p_{23} i p_{dis}, on l'últim valor és la probabilitat que totes le bases siguin diferents.

$$\begin{aligned}
p_{12} &= \tfrac{1}{4}(a_1(a_4a_3b_2 + 2b_4b_3b_2 + b_4b_3a_2)) + \\
&\quad \tfrac{1}{4}(b_1(a_4b_3a_2 + b_4a_3b_2 + 2b_4b_3b_2)) + \\
&\quad \tfrac{1}{2}(b_1(a_4b_3b_2 + b_4a_3b_2 + b_4b_3a_2)),
\end{aligned}$$

$$\begin{aligned}
p_{13} &= \tfrac{1}{4}(a_1(a_4b_3a_2 + b_4a_3b_2 + 2b_4b_3b_2)) + \\
&\quad \tfrac{1}{4}(b_1(a_4a_3b_2 + b_4b_3a_2 + 2b_4b_3b_2)) + \\
&\quad \tfrac{1}{2}(b_1(b_4b_3a_2 + b_4a_3b_2 + 2b_4b_3b_2)),
\end{aligned}$$

$$\begin{aligned}
p_{23} &= \tfrac{1}{4}(b_1(a_4a_3a_2 + 3b_4b_3b_2) + \\
&\quad \tfrac{1}{4}(a_1(a_4b_3b_2 + b_4a_3a_2 + 2b_4b_3b_2)) + \\
&\quad \tfrac{1}{2}(b_1(a_4b_3b_2 + b_4a_3a_2 + 2b_4b_3b_2)),
\end{aligned}$$

$$\begin{aligned}
p_{dis} &= \tfrac{1}{4}(a_1(a_4b_3b_2 + b_4a_3b_2 + b_4b_3a_2 + b_4b_3b_2)) + \\
&\quad \tfrac{1}{4}(b_1(a_4a_3b_2 + 2b_4b_3b_2 + b_4b_3a_2)) + \\
&\quad \tfrac{1}{4}(b_1(a_4b_3a_2 + 2b_4b_3b_2 + b_4a_3b_2)) + \\
&\quad \tfrac{1}{4}(b_1(a_4b_3b_2 + b_4b_3b_2 + b_4a_3b_2 + b_4b_3a_2))
\end{aligned}$$

Ara usarem el `SINGULAR` per a calcular els invariants filogenètics. El software `SINGULAR`, és un programa dissenyat per l'equip amb el mateix nom del Departament de Matemàtiques de la Universitat de Kaiserslautern, sota la direcció de Gert-Martin Greuel, Gerhard Pfister i Hans Schönemann. Escrivim les instruccions següents:

```
> ring r = 0, (t1,t2,t3,t4), lp;
ring s = 0, (x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,
x14,x15,x16,x17,x18,x19,x20,x21,x22,x23,x24,x25,x26,x27,
x28,x29,x30,x31,x32,x33,x34,x35,x36,x37,x38,x39,x40,x41,
x42,x43,x44,x45,x46,x47,x48,x49,x50,x51,x52,
x53,x54,x55,x56,x57,x58,x59,x60,x61,x62,x63,x64), lp;

>setring r; poly e1 = t1*(1-t1)/3 + t2*(1-t2);
poly e2 = t1*(1-t2)/3 + t2*(1-t1) + 2*t2*(1-t2)/3;
poly c0 =t3;
poly c1 = (1-t3)/3;
poly d0 = t4; poly d1 = (1-t4)/3;

> poly e0 = e1; poly e1 = e2;
poly p123 = e0*c0*d0 + 3*e1*c1*d1;
poly p12 = 3*e0*c0*d1 + 3*e1*c1*d0 + 6*e1*c1*d1;
poly p13 = 3*e0*c1*d1 + 3*e1*c0*d1 + 6*e1*c1*d1;
poly p23 = 3*e1*c0*d0 + 3*e0*c1*d1 + 6*e1*c1*d1;
poly pdis = 6*e1*c1*d0 + 6*e1*c0*d1 + 6*e0*c1*d1
+ 6*e1*c1*d1;
> map f=s,  p123, p123, p123, p123, p12, p12, p12,
p12, p12, p12, p12, p12,  p12,p12, p12, p12, p13,
p13, p13, p13, p13, p13, p13, p13,  p13, p13, p13,
 p13, p23, p23, p23, p23, p23, p23, p23, p23,  p23,
 p23, p23, p23, pdis, pdis, pdis, pdis, pdis, pdis,
 pdis, pdis, pdis, pdis, pdis, pdis, pdis, pdis,
 pdis, pdis, pdis, pdis, pdis, pdis,  pdis, pdis,
 pdis, pdis;
```

i obtenim la següent successió de invariants

```
> ideal i0 = 0; setring s; preimage(r,f,i0);
_[1]=x63-x64; _[2]=x62-x64;
_[3]=x61-x64;_[4]=x60-x64;
_[5]=x59-x64; _[6]=x58-x64;
_[7]=x57-x64; _[8]=x56-x64;
```

```
_[9]=x55-x64; _[10]=x54-x64;
_[11]=x53-x64; _[12]=x52-x64;
_[13]=x51-x64; _[14]=x50-x64;
_[15]=x49-x64; _[16]=x48-x64;
_[17]=x47-x64; _[18]=x46-x64;
_[19]=x45-x64; _[20]=x44-x64;
_[21]=x43-x64; _[22]=x42-x64;
_[23]=x41-x64; _[24]=x39-x40;
_[25]=x38-x40; _[26]=x37-x40;
_[27]=x36-x40; _[28]=x35-x40;
_[29]=x34-x40; _[30]=x33-x40;
_[31]=x32-x40; _[32]=x31-x40;
_[33]=x30-x40; _[34]=x29-x40;
_[35]=x27-x28; _[36]=x26-x28;
_[37]=x25-x28; _[38]=x24-x28;
_[39]=x23-x28; _[40]=x22-x28;
_[41]=x21-x28; _[42]=x20-x28;
_[43]=x19-x28; _[44]=x18-x28;
_[45]=x17-x28; _[46]=x15-x16;
_[47]=x14-x16; _[48]=x13-x16;
_[49]=x12-x16; _[50]=x11-x16;
_[51]=x10-x16; _[52]=x9-x16;
_[53]=x8-x16; _[54]=x7-x16;
_[55]=x6-x16; _[56]=x5-x16;
_[57]=24*x4*x16*x28-12*x4*x16*x64+24*x4*x28*x40-
24*x4*x28*x64+6*x4*x64^2-8*x16^2*x28+4*x16^2*x64-
32*x16*x28^2+24*x16*x28*x40+24*x16*x28*x64-
8*x16*x40^2-8*x16*x40*x64-4*x16*x64^2+
16*x28^2*x64-8*x28*x40^2-8*x28*x40*x64-
10*x28*x64^2+4*x40^2*x64+4*x40*x64^2+x64^3;
_[58]=x3-x4; _[59]=x2-x4; _[60]=x1-x4.;
```

Aquests són els 60 invariants de l'arbre filogènetic. Es pot observar que tots llevat del 57, estan describint-nos el conjunt d'igualtats trivials donades per l'aplicació. En aquest cas l'invariant filogenètic és l'ideal [57]. És un polinomi de grau tres amb 19

termes.

Per a observar la diferència entre els arbres filogenètics, escrivim a continuació els invariants algebraics del arbre de tres fulles de la forma següent,

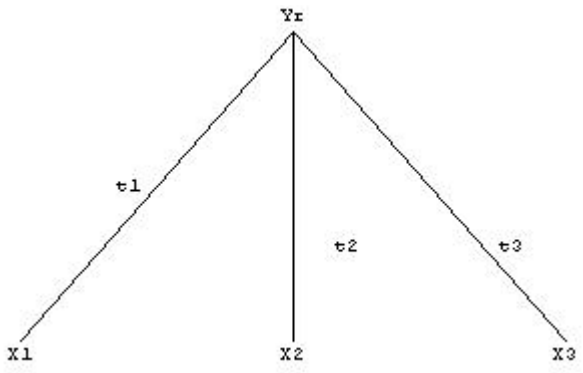


Figura 3.4: Arbre garfi de tres fulles

i ens dona la següent relació d'invariants filogenètics:

```
_[1]=x63-x64 _[2]=x62-x64 _[3]=x61-x64
_[4]=x60-x64 _[5]=x59-x64
_[6]=x58-x64 _[7]=x57-x64 _[8]=x56-x64
_[9]=x55-x64 _[10]=x54-x64
_[11]=x53-x64 _[12]=x52-x64
_[13]=x51-x64 _[14]=x50-x64
_[15]=x49-x64 _[16]=x48-x64
_[17]=x47-x64 _[18]=x46-x64
_[19]=x45-x64 _[20]=x44-x64
_[21]=x43-x64 _[22]=x42-x64
_[23]=x41-x64 _[24]=x39-x40
_[25]=x38-x40 _[26]=x37-x40
_[27]=x36-x40 _[28]=x35-x40
_[29]=x34-x40 _[30]=x33-x40
_[31]=x32-x40 _[32]=x31-x40
_[33]=x30-x40 _[34]=x29-x40
```

```
_[35]=x27-x28 _[36]=x26-x28
_[37]=x25-x28 _[38]=x24-x28
_[39]=x23-x28 _[40]=x22-x28
_[41]=x21-x28 _[42]=x20-x28
_[43]=x19-x28 _[44]=x18-x28 _[45]=x17-x28
_[46]=4*x16^2*x28+4*x16^2*x40+8*x16^2*x64+
4*x16*x28^2+8*x16*x28*x40+24*x16*x28*x64-
x16*x28+4*x16*x40^2+24*x16*x40*x64-x16*x40+
32*x16*x64^2-6*x16*x64+4*x28^2*x40+
8*x28^2*x64+4*x28*x40^2+24*x28*x40*x64-
x28*x40+32*x28*x64^2-6*x28*x64+
8*x40^2*x64+32*x40*x64^2-
6*x40*x64+32*x64^3-11*x64^2+x64
_[47]=x15-x16 _[48]=x14-x16
_[49]=x13-x16 _[50]=x12-x16
_[51]=x11-x16 _[52]=x10-x16
_[53]=x9-x16 _[54]=x8-x16 _[55]=x7-x16
_[56]=x6-x16 _[57]=x5-x16
_[58]=x3-x4; _[59]=x2-x4; _[60]=x1-x4.
```

Ara l'invariant filogenètic resulta ser el [46] i com abans, els altres invariants ens diuen les igualtats trivials de les variables.

En aquest cas hem aconseguit calcular els invariants filogenètics directament de les funcions i usant un programa con el `SINGULAR`.

Degut a la complexitat dels càlculs que s'han de realitzar, l'exemple de quatre fulles és l'únic que es pot calcular directament sense usar el canvi a coordenades de Furier. Si fem aquest canvi resultarà una base de Gröbner, i, aleshores, els càlculs seran factibles per dimensió molt més alta, i. e., per a un major nombre d'espècies.

Per a arbres amb més de quatre fulles necessitem usar un canvi de variables que faci més senzill el càlcul d'aquest invariants. Per això usarem la transformada de Fourier que s'obté del teorema 3.2.4 demostrat a l'article [Eva93].

Per a veure un exemple de la transformada de Fourier reprendrem l'exemple del primer arbre filogenètic. Entendrem $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ com un grup isomorf a $\mathbb{Z}/(2) \times \mathbb{Z}/(2)$ mitjançant l'aplicació

$$\begin{array}{ccc} \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} & \to & \mathbb{Z}/(2) \times \mathbb{Z}/(2) \\ \mathtt{A} & \mapsto & (0,0) \\ \mathtt{C} & \mapsto & (1,0) \\ \mathtt{G} & \mapsto & (0,1) \\ \mathtt{T} & \mapsto & (1,1) \end{array}$$

i d'aquesta manera dotem el conjunt de les bases d'estructura de grup.

Ara tractem de calcular la funció f de la qual hem fet referència en la fórmula (3.2). Recordem segons hem vist en el capítol 2, la matriu S_{ij} ve donada per la forma

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

i en el cas del model de Jukes-Cantor la matriu de substitució ve donada com

$$S = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix},$$

on els paràmetres a i b compleixen que $a + 3b = 1$. Per tal que compleixi la igualtat

$$S(i,j) = f(i-j),$$

deduïm que $f(0) = f(i-i) = S(i,i) = a$ i $f(x) = b$ altrament. Tindrem

$$f^i(x) = \begin{cases} a_i & \text{si } x = 0 \\ b_i & \text{cc.} \end{cases},$$

on el subíndex ens indicarà la branca en la que estem treballant.

Hem d'observar que els caràcters del grup $\mathbb{Z}/(2)\times\mathbb{Z}/(2)$ venen donats per la següent taula:

χ	A	G	C	T
1	1	1	1	1
ϕ	1	-1	1	-1
ψ	1	1	-1	-1
$\phi\psi$	1	-1	-1	1

Si en la fórmula provinent del teorema d'Evans-Speed

$$q(\chi_1,...,\chi_m) = \prod_{b\in E(T)} \widehat{f^b}\left(\prod_{l\in\Lambda(b)}\chi_l\right), \tag{3.4}$$

on $E(T)$ denota el conjunt de les branques de l'arbre T i $\Lambda(b)$ el conjunt de fulles per sota de la branca b, substituïm de forma adequada obtenim que canvi de coordenades per a q_{0000} ve donat per

$$\begin{aligned} q_{0000} &= \hat{f}^1(\chi_1)\hat{f}^2(\chi_1)\hat{f}^3(\chi_1)\hat{f}^4(\chi_1) \\ &= (a_1+3b_1)(a_2+3b_2)(a_3+3b_3)(a_4+3b_4) \\ &= a_1a_2a_3a_4 + 3a_2a_3a_4b_1 + 3a_1a_3a_4b_2 + \\ &\quad 9a_3a_4b_1b_2 + 3a_1a_2a_4b_3 + 9a_2a_4b_1b_3 + \\ &\quad 9a_1a_4b_2b_3 + 27a_4b_1b_2b_3 + 3a_1a_2a_3b_4 + \\ &\quad 9a_2a_3b_1b_4 + 9a_1a_3b_2b_4 + 27a_3b_1b_2b_4 + \\ &\quad 9a_1a_2b_3b_4 + 27a_2b_1b_3b_4 + 27a_1b_2b_3b_4 + \\ &\quad 81b_1b_2b_3b_4 \\ &= p_{123} + p_{12} + p_{13} + p_{23} + p_{dis}. \end{aligned}$$

En els altres casos, podem aconseguir aquestes igualtats,

$$\begin{aligned}
q_{0011} &= (a_1+3b_1)(a_2+3b_2)(a_3-b_3)(a_4-b_4)\\
&= p_{123}-\tfrac{1}{3}p_{12}-\tfrac{1}{3}p_{13}+p_{23}-\tfrac{1}{3}p_{dis}\\
q_{1101} &= (a_1-b_1)(a_2-b_2)(a_3+3b_3)(a_4-b_4)\\
&= p_{123}-\tfrac{1}{3}p_{12}+p_{13}-\tfrac{1}{3}p_{23}-\tfrac{1}{3}p_{dis}\\
q_{1110} &= (a_1-b_1)(a_2-b_2)(a_3-b_3)(a_4-d_4)\\
&= p_{123}+p_{12}-\tfrac{1}{3}p_{13}-\tfrac{1}{3}p_{23}-\tfrac{1}{3}p_{dis}\\
q_{1111} &= (a_1-b_1)(a_2-b_2)(a_3-b_3)(a_4-b_4)\\
&= p_{123}-\tfrac{1}{3}p_{12}-\tfrac{1}{3}p_{13}-\tfrac{1}{3}p_{23}+\tfrac{1}{3}p_{dis}.
\end{aligned}$$

Després d'aquest canvi l'invariant [57] del primer model. Per tant la matriu de canvi de base queda de la següent forma:

$$\begin{pmatrix}
1 & 1 & 1 & 1 & 1\\
1 & -\frac{1}{3} & -\frac{1}{3} & 1 & -\frac{1}{3}\\
1 & -\frac{1}{3} & 1 & -\frac{1}{3} & -\frac{1}{3}\\
1 & 1 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3}\\
1 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{1}{3}
\end{pmatrix}.$$

En les noves coordenades tindrem que l'invariant [45] correspon a

$$q_{0011}q_{1110}q_{1101}-q_{0000}q_{1111}^2.$$

Capítol 4

Bases de Gröbner, varietats tòriques i ideals tòrics.

En aquest capítol donarem les definicions i conceptes bàsics referents a les base de Gröbner i les varietat tòriques. En serà molt útil ja que l'algoritme de construcció dels invariants filogenètics, ens donarà una base de Gröbner.

4.1 Bases de Gröbner.

4.1.1 Definició. Sigui k un cos i $A = k[x_1, ..., x_n]$. Un element $x_1^{i_1} \cdots x_n^{i_n}$ en A s'anomena monomi i un element $cx_1^{i_1} \cdots x_n^{i_n}$ amb $c \in k \setminus \{0\}$ s'anomena terme.

Denotarem per $\mathcal{M}$ el conjunt de tots els monomis en A. Introduïm la definició d'una relació d'ordre $\prec$ sobre $\mathcal{M}$.

4.1.2 Definició. Es defineix $\prec$ com un ordre de relació d'ordre admissible sobre $\mathcal{M}$ si és un ordre total què és compatible amb

la multiplicació de monomis, i. e.

1. Per a qualsevol parell de monomis m, n tenim $m \prec n$ o $n \prec m$ o $m = n$.
2. Si $m_1 \prec m_2$ i $m_2 \prec m_3$ llavors $m_1 \prec m_3$.
3. $1 \prec m$ per a qualsevol monomi $m \neq 1$.
4. Si $m_1 \prec m_2$ llavors $mm_1 \prec mm_2$ per a qualsevol monomi m.

4.1.3 Definició. L'ordre lexicogràfic, lex. Aquí $x_1^{i_1} \cdot \cdots \cdot x_n^{i_n} \prec x_1^{j_1} \cdot \cdots \cdot x_n^{i_n}$ si es dóna que $i_1 = j_1$,...,$i_k = j_k$, $i_{k+1} < j_{k+1}$ per a algun k. En altres paraules $m_1 \prec m_2$ si la primera variable amb exponent diferent en m_1 i m_2 té exponent més baix en m_1.

4.1.4 Definició. L'ordre grau lexicogràfic, deglex. Aquí $m_1 = x_1^{i_1} \cdot \cdots \cdot x_n^{i_n} \prec x_1^{j_1} \cdot \cdots \cdot x_n^{j_n} = m_2$ si es dóna que $\deg(m_1) = i_1 + \cdots + i_n < j_1 + \cdots + j_n = \deg(m_2)$ o si $\deg(m_1) = \deg(m_2)$ i $m_1 \prec m_2$ en Lex.

4.1.5 Definició. L'ordre lexicogràfic revertit, degrevlex. Aquí $m_1 = x_1^{i_1} \cdot \cdots \cdot x_n^{i_n} \prec x_1^{j_1} \cdot \cdots \cdot x_n^{j_n} = m_2$ si $\deg(m_1) < \deg(m_2)$ en cas que $\deg(m_1) = \deg(m_2)$ i $i_n = j_n$, $i_{n-1} = j_{n-1}$, ..., $i_k = j_k$, $i_{k-1} > j_{k-1}$ per algun k. En altres paraules $m_1 \prec m_2$ si $\deg(m_1) < \deg(m_2)$ o si $\deg(m_1) = \deg(m_2)$ i la última variable amb exponent diferent en m_1 i m_2 té exponent més alt en m_1.

4.1.6 Observació. Existeixen molts més ordres.

Per tal de definir el concepte de bases de Gröbner cal primer un resulta previ.

4.1.7 Lema. (Lema de Dickson) *Cada ideal monomial, i. e. generat per monomis, en l'anells de polinomis $k[x_1, ..., x_n]$, sobre el cos k, és finitament generat.*

DEMOSTRACIÓ. Apliquem inducció sobre n. Donat que cada ideal en $k[x]$ és principal, el lema és cert. Suposem que és cert per $n-1$ variables, i sigui $\mathfrak{a}$ un ideal monomial en $k[x_1, ..., x_n]$. Sigui

$$\mathfrak{b}_j = (\mathfrak{a} : \langle x_n^j \rangle) \cap k[x_1, ..., x_n] = \langle S_j \rangle.$$

Atès que $\mathfrak{b}_j$ és un ideal en $k[x_1, ..., x_{n-1}]$, podem triar S_j finit. Tenim $\mathfrak{b}_0 \subseteq \mathfrak{b}_1 \subseteq \cdots$. Es segueix que $\cup \mathfrak{b}_j$ és un ideal $\mathfrak{b}$ en $k[x_1, ..., x_{n-1}]$ i d'aquí finitament generat, $\mathfrak{b} = \langle S \rangle$. Si $m \in \mathfrak{a}$ és un monomi llavors $m = m' x_n^k$ per algun monomi $m' \in k[x_1, ..., x_{n-1}]$ i per algun k. Donat que $m' x_n^k \in \mathfrak{a}$ tenim $m' \in \mathfrak{a} : \langle x_n^k \rangle$, per tant $m \in \langle x_n^k S_k \rangle$. D'aquesta manera $S' = S_0 \cup x_n S_1 \cup x_n^2 S_2 \cup \cdots$ és un conjunt generat per $\mathfrak{a}$. Per a algun r tenim que $S_k = S_r$ si $k \geq r$, per tant $S_0 \cup x_n S_1 \cup \cdots \cup x_n^r S_r$ és un conjunt finit de generadors per a $\mathfrak{a}$. □

El lema de Dikinison no és més que un cas particular del teorema de les bases de Hilbert. En realitat podria ser un corol·lari d'aquest. En aquest cas la prova esta feta de forma independent del teorema. Per tant queda provat un mètode constructiu de les bases, en el cas d'un ideal generat per monomis.

Sigui $f \in k[x_1, ..., x_n]$, $f \neq 0$, i suposem $\prec$ un ordre admissible dels monomis en A. Llavors f es pot escriure de manera única escrivint $f = c_1 m_1 + \cdots c_N m_N$ amb monomis $m_1 \succ m_2 \succ \cdots m_N$ i $c_i \neq 0$, $i = 1, ..., N$.

4.1.8 Definició. Es defineix els suport de f com

$$\text{supp}(f) = \{m_i \mid i = 1, ..., N\}.$$

Es defineix el monomi líder de f com $\text{lm}(f) = m_1$.

Es defineix el terme líder de f com $\text{lt}(f) = c_1 m_1$.

Es defineix el coeficient líder de f com $\text{lc}(f) = c_1$.

4.1.9 Exemples. Els conceptes anteriors, depenen de l'ordre. Per a veure això considerem el polinomi

$$f = x_1^5 + x_1^3x_2^3x_3 + x_1^4x_2x_3^2.$$

Aleshores, es té que

	$lt(f)$
$lex:$	x_1^5
$deglex:$	$x_1^4x_2x_3^2$
$degrevlex:$	$x_1^3x_2x_3$

4.1.10 Exemples. Sigui $f = 3x_1^2x_2 + 2x_1 - x_2^2$. Respecte l'ordre lex, tindrem: $\text{supp}(f) = \{x_1^2x_2, x_1, x_2^2\}$, $\text{lm}(f) = x_1^2x_2$, $\text{lt}(f) = 3x_1^2x_2$ i $\text{lc}(f) = 3$.

4.1.11 Lema. *Sigui $\prec$ un ordre admissible de monomis de A. Llavors un conjunt de monomis no buit S té un element més petit, i.e. hi ha un $m_0 \in S$ tal que $m_0 \prec n$ per qualsevol $n \in S$, $n \neq m_0$.*

Demostració. Sigui $\mathfrak{a} = \langle S \rangle$ l'ideal generat per S. Pel lema de Dickson, $\mathfrak{a}$ és finitament generat. Prenem $\mathfrak{a} = \langle m_1, ..., m_r \rangle$. Donat que qualsevol monomi en S és un multiple de algun m_i, $m = m_i m'$, per algun i i algun m', podem triar m_0 el polinomi més petit en $\{m_1, ..., m_r\}$ donat que qualsevol multiple de m_i, és mes gran que m_i. □

Podem escriure aquest lema de la següent manera

4.1.12 Corol·lari. *Qualsevol successió decreixent de monomis en un ordre admissible $\succ$ és finit.*

L'algoritme de divisió per a una variable ve donat per l'algoritme d'Euclides, ampliament conegut i estudiat i sabem que proporciona una única solució.

El següent pas serà calcular un algoritme de divisió per a un anell de polinomis en varies variables. El primer que notem és que encara que no es digui explícitament en $k[x]$ tenim l'ordre $1 \prec x \prec x^2 \prec \cdots$. Es pot observar que en $k[x]$ els ordres lex, deglex i degrevlex coincideixen.

En multiples variables podrem prendre moltes eleccions d'ordres. Però per a aconseguir un bon algoritme haurem de triar algun ordre admissible, com els vistos abans, però tenint en compte que les diferents tries d'ordres són bones per diferents propòsits. Un segon obstacle és que els ideals no són principals. Això significa que, en general, és necessari fer una bona tria del conjunt de generadors. Aquesta és la idea de les bases de Gröbner.

Primer definim un algoritme de divisió. Farem la divisió del polinomi f en l'ideal $g = (g_1, ..., g_s)$. Primer considerem que g està generat per un sol polinomi, és a dir $g = (g_1)$. Fixem un ordre de monomis. Definirem el residu de f respecte a g_1. Si $f = 0$ aleshores el residu és 0. En altre cas, sigui $m = \mathrm{lt}(g_1)$ i prenem $S_f = \{n \in \mathrm{supp}(f) \mid m \text{ divideix } n\}$. Sigui p_0 l'element més gran en S_f i prenem el coeficient de p_0 en f sent c_{p_0}. Sigui

$$f_1 = f - \frac{(c_{p_0} p_0)}{(\mathrm{lc}(g_1)m)} g_1.$$

Sigui $S_{f_1} = \{n \in \mathrm{supp}(f_1) \mid m \text{ divideix } n\}$.

4.1.13 Lema. *Si p_1 és l'element més gran en S_{f_1} llavors $p_1 \prec p_0$.*

Demostració. D'això se segueix del fet que $\mathrm{supp}(f_1) \subseteq (\mathrm{supp}(f) \cup ((p_0/m)g_1)) \setminus \{p_0\}$. Si $p_1 \in \mathrm{supp}(f) \setminus \{p_0\}$ llavors òbviament $p_1 \prec p_0$. Tenim $\mathrm{lm}((p_0/m)g_1)) = (p_0/m)lm(g_1) = (p_0/m)m = p_0$. Per tant també si tenim $p_1 \in \mathrm{supp}((p_0/m)g_1)$ tindrem $p_1 \prec p_0$. □

Continuant aquest procediment podrem definir $f_2 = f_1 - f - \frac{(c_{p_1}p_1)}{(\mathrm{lc}(g_1)m)}g_1$ i $p_2 = max(S_{f_2})$. La successió $p_0 \succ p_1 \succ p_2 \succ \cdots$ és finit d'acord amb el corol·lari anterior, per tant després d'un nombre finit de passos no hi hauran elements en $\mathrm{supp}(f_N)$ divisibles per m. Tenim $f_N = f - hg_1$ per algun polinomi h, i definim f_N el residu de f respecte g_1.

Ara formulem el procés per a qualsevol nombre d'elements en l'ideal quocient. Sigui $(g_1, ..., g_s)$. Es defineix $\mathrm{rem}(f, I)$ com el residu de f en l'ideal I. El residu es pot calcular recursivament de la següent manera. Si $\mathrm{lm}(g_1)$ divideix $\mathrm{lm}(f)$, sigui $\mathrm{rem}(f, (g_1)) = \mathrm{rem}(f - ng_1, (g_1))$, on n és un terme triat de manera que $\mathrm{lt}(f) = \mathrm{lt}(ng_1)$, Si $\mathrm{lm}(g_1)$ no divideix $\mathrm{lm}(f)$, prenem $\mathrm{rem}(f, (g_1)) = \mathrm{lt}(f) + \mathrm{rem}(f - \mathrm{lt}(f), (g_1))$. Això dóna una definició recursiva de $\mathrm{rem}(f)$ respecte a g_1. Ara prenem en $(g_1, ..., g_s)$ una successió ordenada de polinomis no nuls i f un polinomi. Recursivament definim $\mathrm{rem}(f, (g_1, ..., g_s)) = \mathrm{rem}(f - ng_k, (g_1, ..., g_s))$, on k és el índex més petit tal que $\mathrm{lm}(g_k)$ divideix $\mathrm{lm}(f)$ i n és un terme triat tal que $\mathrm{lt}(f) = \mathrm{lt}(ng_k)$. Si no $\mathrm{lm}(g_i)$ divideix $\mathrm{lm}(f)$ definirem $\mathrm{rem}(f, (g_1, ..., g_s)) = (\mathrm{lt}(f) + \mathrm{rem}(f - \mathrm{lt}(f)), (g_1, ..., g_s))$. Observem que el procés és finit.

Per a una successió general $(g_1, ..., g_s)$ el residu no té un bon comportament. Per exemple el resultat depèn de l'ordre dels elements en la llista.

4.1.14 Lema. *Sigui* $f, g_1, ..., g_s \in k[x_1, ..., x_n]$. *Aleshores*

$$f - \mathrm{rem}(f, (g_1, ..., g_s)) \in \langle g_1, ..., g_s \rangle.$$

En particular, si $\mathrm{rem}(f, (g_1, ..., g_s)) = 0$ *llavors* $f \in \langle g_1, ..., g_s \rangle$.

4.1.15 Observació. El nostre objectiu és triar un conjunt generadors $g'_1, ..., g'_N$ per $\mathfrak{a} = \langle g_1, ..., g_s \rangle$, tal que el residu d'un polinomi respecte a $(g'_1, ..., g'_N)$ és únic en el sentit que només depen de l'ideal. Per tant volem aconseguir que

$$\mathrm{rem}(f_1, (g'_1, ..., g'_N)) = \mathrm{rem}(f_2, (g'_1, ..., g'_N))$$

i això passa si i només si, si $f_1 - f_2 \in \mathfrak{a}$.

4.1.16 Definició. Sigui $\mathfrak{a}$ un ideal en $k[x_1, ..., x_n]$. Un conjunt d'elements $\{g_1, ..., g_s\}$ en $\mathfrak{a}$ tal que $\langle \mathrm{lm}(g_1), ..., \mathrm{lm}(g_s) \rangle = l(\mathfrak{a})$ s'anomena bases de Gröbner per $\mathfrak{a}$.

A continuació el que farem serà comprovar que aquesta definició és bona en el sentit que compleix el requerit en la observació.

4.1.17 Lema. *Si $\{g_1, ..., g_s\}$ és una base de Gröbner per $\mathfrak{a}$, llavors $\langle g_1, ..., g_s \rangle = \mathfrak{a}$.*

Demostració. Clarament $\langle g_1, ..., g_s \rangle \subseteq \mathfrak{a}$ atès que $g_i \in \mathfrak{a}$ per tot i. Sigui $f \in \mathfrak{a}$. Llavors $\mathrm{lm}(f) \in \langle \mathrm{lm}(g_1), ..., \mathrm{lm}(g_s) \rangle$, per tant $\mathrm{lm}(f - ng_k) \prec \mathrm{lm}(f)$ per algun g_k i algun terme n. Com que $f - ng_k \in \mathfrak{a}$, obtenim de manera recursiva que $f \in \langle g_1, ..., g_s \rangle$. □

4.1.18 Teorema. *Sigui $\{g_1, ..., g_s\}$ una base de Gröbner per l'ideal $\mathfrak{a} = \langle g_1, ..., g_s \rangle$. Llavors* $\mathrm{rem}(f_1, (g_1, ..., g_s)) = \mathrm{rem}(f_2(g_1, ..., g_s))$ *si i només si $f_1 - f_2 \in \mathfrak{a}$. En particular* $\mathrm{rem}(f, (g_1, ..., g_s)) = 0$ *si i només si $f \in \mathfrak{a}$.*

Demostració. Suposem $\mathrm{rem}(f_1, G) = \mathrm{rem}(f_2, G)$, on $G = (g_1, ..., g_s)$. Com que $f_i - \mathrm{rem}(f_i, G) \in \mathfrak{a}$, $i = 1, 2, ...$ pel 4.1.14, tindrem que $f_1 - \mathrm{rem}(f_1, G) - (f_2 - \mathrm{rem}(f_2, G)) = f_1 - f_2 \in \mathfrak{a}$. Suposem $f_1 - f_2 \in \mathfrak{a}$. Llavors $\mathrm{rem}(f_1, G) - \mathrm{rem}(f_2, G) = (f_2 - \mathrm{rem}(f_2, G)) - (f_1 - \mathrm{rem}(f_1, G)) + (f_1 - f_2) \in \mathfrak{a}$ ja que $f_i - \mathrm{rem}(f_i, G) \in \mathfrak{a}$, $i = 1, 2$. Però $\mathrm{rem}(f_i, G)$, $i = 1, 2$ és una combinació lineal de monomis fora del conjunt $l(\mathfrak{a})$. Si tinguéssim $\mathrm{rem}(f_1, G) \neq \mathrm{rem}(f_2, G)$ aleshores tindríem que $\mathrm{lm}(\mathrm{rem}(f_1, G) - \mathrm{rem}(f_2, G) \notin \mathfrak{a}$. Però $\mathrm{rem}(f_1, G) - \mathrm{rem}(f_2, G) \in \mathfrak{a}$ i per tant tindríem $\mathrm{lm}(\mathrm{rem}(f_1, G) - \mathrm{rem}(f_2, G)) \in \mathfrak{a}$, però això ens dóna una contradicció. □

4.2 Ideals tòrics i varietats tòriques

Sigui $k[x_1, ..., x_n]$ un anell de polinomis. Fixem $\mathcal{A} = \{a_1, ..., a_n\}$ de $\mathbb{Z}^d$. Cada vector a_i s'identifica amb un monomi t^{a_i} en l'anell de polinomis de Laurent $k[t^{\pm 1}] := k[t_1, ..., t_d, t_1^{-1}, ..., t_d^{-1}]$. Considerem l'aplicació

$$\begin{array}{rccl} \pi: & \mathbb{N}^n & \to & \mathbb{Z}^d \\ & (u_1, ..., u_n) & \mapsto & u_1 a_1 + \cdots + u_n a_n. \end{array} \tag{4.1}$$

La imatge de π és el conjunt

$$N\mathcal{A} = \{\lambda_1 a_1 + \cdots + \lambda_n a_n : \lambda_1, ..., \lambda_n \in \mathbb{N}\}.$$

L'aplicació π aixeca a un homomorfisme de semigrups d'àlgebres, de la següent forma:

$$\begin{array}{rccl} \hat{\pi}: & k[x_1, ..., x_n] & \to & k[t^{\pm 1}] \\ & x_i & \mapsto & t^{a_i}. \end{array} \tag{4.2}$$

4.2.1 Definició. El nucli de $\hat{\pi}$ es denota $I_{\mathcal{A}}$ i s'anomena l'ideal tòric de $\mathcal{A}$.

Clarament, l'ideal $I_{\mathcal{A}}$ és un ideal primer, i per tant la seva varietat afí $V(I_{\mathcal{A}})$ de zeros en k^n és irreductible. Aquesta és la clausura de Zariski del conjunt de punts $(t^{a_1}, ..., t^{a_n})$, in $t \in (k^*)^d$. El grup multiplicatiu $(k^*)^d$ s'anomena el tor algebraic de dimensió d.

4.2.2 Definició. Una varietat de la forma $V(I_{\mathcal{A}})$ és una varietat tòrica afí.

4.2.3 Lema. *L'ideal tòric $I_{\mathcal{A}}$ esta generat com a k-espai vectorial pel conjunt de binomis*

$$\{x^u - x^v : u, v \in \mathbb{N} \ \ amb \ \ \pi(u) = \pi(v)\}.$$

DEMOSTRACIÓ. Un binomi $x^u - x^v$ està en $I_{\mathcal{A}}$ si i només si $\pi(u) = \pi(v)$. És suficient provar que cada polinomi en $I_{\mathcal{A}}$ és una combinació lineal d'aquest binomis. Fixem un ordre $\prec$ sobre $k[X]$. Suposem $f \in I_{\mathcal{A}}$ no pot estar escrit com una combinació de binomis de k. Triem un polinomi f amb la propietat que x^u és el terme mínim respecte l'ordre $\prec$. Quan expandim $f(t^{a_i}, ..., t^{a_n})$ obtenim que val zero. En particular, el terme $t^{\pi(u)} = \hat{\pi}(x^u)$ s'ha de cancelar en aquesta expansió. Per tant hi ha algun altre monomi $x^v \prec x^u$ que apareix en f tal que $\pi(u) = \pi(v)$. També el polinomi $f' := f - x^u + x^v$ no pot estar escrit com una combinació de binomis de k en $I_{\mathcal{A}}$. Però el fet que el monomi més petit de f' sigui més petit que el de f, cosa que és contradicció. □

En el cas dels invariants filogenètics hem construït una aplicació semblant a la de la fórmula (4.2), però en un espai més senzill encara ja que els exponents són enters positius. Per tant, estem en el cas que tenim la funció φ, definida com

$$\begin{array}{rccl} \varphi: & k[\theta_1, ..., \theta_d] & \to & k[x_1, ..., x_{4^n}] \\ & \{\theta_i\}_{i=1}^d & \mapsto & (p_{\mathtt{AA\cdots A}}, p_{\mathtt{AA\cdots C}}, \cdots, p_{\mathtt{TT\cdots T}}). \end{array}$$

i volem calcular un conjunt de generadors per a la clausura algebraica de la imatge.

Capítol 5

Algoritme de construcció d'invariants filogenètics

En aquest capítol, provarem el teorema que ens donarà un algoritme de construcció dels invariants filogenètics. En la part final d'aquest capítol, construirem el un exemple, és a dir passarem de l'arbre més senzill (el de dos fulles) als següent (el de tres fulles), refent les construccions explicites en cada pas.

5.1 Etiquetatge de branques i invariants lineals

En aquesta secció buscarem una manera de calcular els polinomis $q_{g_1,\ldots,g_m} - q_{h_1,\ldots,h_m}$ que estan en el nucli de l'apliació

$$\begin{array}{rccl} \varphi: & k[\theta_1,\ldots,\theta_d] & \to & k[x_1,\ldots,x_{4^n}] \\ & \{\theta_i\}_{i=1}^d & \mapsto & (p_{\mathtt{AA}\cdots\mathtt{A}}, p_{\mathtt{AA}\cdots\mathtt{C}}, \cdots, p_{\mathtt{TT}\cdots\mathtt{T}}). \end{array}$$

i establirem un sistema de coordenades per a treballar mòdul aquests invariants. L'eina que usarem seran les funcions d'etiquetatge.

Partim de l'equació del canvi de coordenades

$$q(\chi_1, ..., \chi_n) = \prod_{b \in E(T)} \widehat{f^b} \left(\prod_{l \in \Lambda(b)} \chi_l \right). \tag{5.1}$$

En cas que les distribucions en l'arrel no són uniformes tindrem que

$$q(\chi_1, ..., \chi_n) = \widehat{\pi} \prod_{b \in E(T)} \widehat{f^b} \left(\prod_{l \in \Lambda(b)} \chi_l \right). \tag{5.2}$$

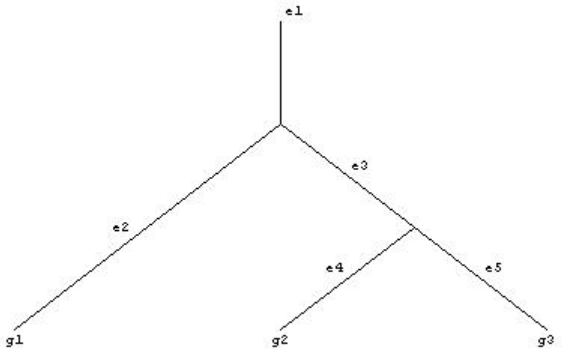
Aleshores per tal de no distingir entre les funcions $\hat{\pi}$ i $\hat{f}$, afegim una branca extra a l'arrel de T per a aconseguir un nou arbre T amb $m+1$ fulles. Notarem T al nou arbre per tal de no carregar la notació. Sigui $E(T)$ el conjunt de branques. Associem un conjunt de paràmetres a cada $e \in E(T)$, movent els paràmetres directament a la branca sobre del node.

Donat elements del grup $(g_1, ..., g_m)$ associats fixats a les m fulles del arbre T, assignem a cada branca de T un element del grup fixat com:

$$g(e) = \sum_{v \in \Lambda(e)} g_v,$$

on $\Lambda(e)$ denota el conjunt de fulles per sota de la branca e.

5.1.1 Exemples. En el nostre exemple això fa que l'arbre quede de la forma de la figura 5.1.



5.1. Arbre de tres fulles amb la branca extra

Amb aquesta notació l'equació (5.1) pot ésser escrita com

$$q_{g_1,\ldots,g_m} \mapsto \prod_{e \in E(T)} f^{(e)}(g(e)), \tag{5.3}$$

eliminant la distinció especial del node arrel. Si les variables $f^{(e)}(g)$ son totes diferents llavors no hi ha invariants lineals (com passa en el model de Kimura 3-parametres). Permetrem, doncs, la possibilitat que $f^{(e)}(g) = f^{(e)}(g')$ per a diferents elements de grup $g, g' \in G$. D'acord amb aquest fet, introduïm les funcions d'etiquetatge. Sigui $\mathcal{L}$ un conjunt finit d'etiquetes.

5.1.2 Definició. Una funció d'etiquetatge és una funció

$$L : G \to \mathcal{L}$$

tal que $f^{(e)}(g) = f^{(e)}(g')$ si i només si $L(g) = L(g')$.

Pel que resta, suposarem que la funció d'etiquetatge del arbre és la mateixa per a totes les branques.

Per a cada branca e del arbre T i cada etiqueta $l \in \mathcal{L}$, introduïm una indeterminada a_l^e. L'anell de polinomis en aquestes variables es denota $k[a_l^{(e)}]$. De forma similar, $k[q_{g_1 \ldots g_m}]$ és l'anell

de polinomis generat per les coordenades de Fourier. Desitgem estudiar l'homomorfismes d'anells

$$\begin{array}{lcl} k[q_{g_1\ldots g_m}] & \to & k[a_l^{(e)}], \\ q_{g_1..g_m} & \mapsto & \prod_{e\in E(T)} a^{(e)}_{L(g(e))}. \end{array} \tag{5.4}$$

El nucli d'aquesta aplicació és el l'ideal tòric d'invariants filogenètics en la transformada de Fourier de probabilitats. Es denota aquest ideal per $I_{T,L}$, suprimint la depèndencia del grup G per tal de no carregar la notació. D'aquesta descripció deduïm l'estructura dels invariants filogènetics.

5.1.3 Proposició. (Invariants Lineals) *El espai vectorial generat pels invariants lineals de l'ideal $I_{T,L}$ esta generat per totes les diferencies $q_{g_1\ldots g_m} - q_{h_1\ldots h_m}$ on $L(g(e)) = L(h(e))$ per totes les branques e de T.*

Demostració. Atès que $I_{T,L}$ és un ideal tòric, té un espai vectorial amb bases consisten amb els binomis $q^u - q^v$. En particular el subespai lineal en $I_{T,L}$ és generat per les diferencies de les incògnites $q_{g_1\ldots g_m} - q_{h_1\ldots h_m}$. Tals diferencies estan en $I_{T,L}$ si i només si

$$\prod_{e\in E(T)} a^{(e)}_{L(g(e))} = \prod_{e\in E(T)} a^{(e)}_{L(h(e))}.$$

Ja que les variables $a_l^{(e)}$ són tots diferents, això passa quan

$$L(g(e)) = \sum_{v\in\Lambda(e)} g_v$$

coincideix $L(h(e)) = \sum_{v\in\Lambda(e)} h_v$ per a tot $e \in T$. □

5.1.4 Exemples. En el nostre exemple, tenim que els polinomis lineals del nucli són aquells que tenen la mateixa etiqueta, així doncs tenim que $q_{CTT} - q_{GCC}$, serà del nucli d'aquesta primera aplicació.

Introduïm ara coordenades per al polinomi

$$k[q_{g_1\ldots g_m} : (g_1, ..., g_m) \in G^m]$$

mòdul l'ideal generat pels invariants lineals en $I_{T,L}$.

5.1.5 Definició. Sigui L una funció d'etiquetatge $L : G \to \mathcal{L}$. Aleshores aquesta funció indueix la funció

$$\begin{array}{rccc} L^T : & G^m & \to & \mathcal{L}^{E(T)}, \\ & (g_1, ..., g_m) & \mapsto & (L(g(e)))_{e\in E(T)}, \end{array} \tag{5.5}$$

Denotem $\mathrm{im}(L^T)$ la imatge d'aquesta aplicació. Anomenem $\mathrm{im}(L^T)$ el conjunt de les etiquetes consistents del arbre T.

5.1.6 Definició. Direm que dos elements són de la mateixa classe, si la diferència entre ells és un invariant lineal. De forma equivalent direm que un element de $x \in G^n$ és de la mateixa classe y si $L^T(x) = L^T(y)$.

Per a cada $\lambda \in \mathrm{im}(L^T)$, introduïm una nova variable q_λ. Aquests polinomis seràn generadors d'un nou anell de polinomis.

Per tant, amb aquest nou anells podem descompondre l'aplicació (5.3) com a composició de les aplicacions

$$\begin{array}{ccc} k[q_{g_1\ldots g_m} : (g_1, ..., g_m) \in G^m] & \to & k[q_\lambda : \lambda \in \mathrm{im}(L^T)] \\ q_{g_1\ldots g_m} & \mapsto & q_{L^T(g_1,\ldots,g_m)}, \end{array} \tag{5.6}$$

i la següent aplicació monomial la qual ja no té invariants lineals en el seu nucli:

$$\begin{array}{ccc} k[q_\lambda : \lambda \in \mathrm{im}(L^T)] & \to & k[a_l^{(e)} : e \in E(T), l \in \mathcal{L}] \\ q_\lambda & \mapsto & \prod_{e\in E(T)} a^{(e)}_{\lambda(e)}. \end{array} \tag{5.7}$$

Observem que el nucli de l'aplicació (5.6) correspon als invariants lineals que hem vist en la proposició 5.1.3 i que per tant, el nucli de (5.7) correspondrà als invariants filogenètics.

Per tant, el nostre objectiu és determinar el nucli de l'aplicació monomial (5.7). El nucli és un ideal tòric $I_{T,L}$ mòdul invariants lineals. Usem el mateix símbol per a denotar el nucli de (5.7).

Donat que el nostre objectiu serà donar un algoritme recursiu per al càlcul dels invariants filogenètics, hem de construir algun objecte matemàtic que tingui la funció d'enganxar els diferents arbres. Aquest objecte seran les funcions amistoses.

5.1.7 Definició. Fixada una funció d'etiquetatge $L : G \to \mathcal{L}$ sobre el grup G. Per a $m \geq 3$ considerem el conjunt

$$Z = \left\{ (g_1, ..., g_m) \in G^m : \sum_{i=1}^{m-1} g_i = g_m \right\}.$$

Considerem l'aplicació induïda $\widetilde{L} : Z \subseteq G^m \to \mathcal{L}^m$ i denotem per π_i la projecció $\pi_i : G^m \to G$ sobre la i-ésima coordenada. La funció L s'anomena m-amistosa si, per cada $l = (l_1, ..., l_m) \in \widetilde{L}(Z) \subseteq \mathcal{L}^m$,

$$\pi_i(\widetilde{L}^{-1}(l)) = L^{-1}(l_i) \qquad \text{per a tot } i = 1, ..., m. \tag{5.8}$$

Notem que la inclusió " $\subseteq$ " sempre val. Però per gran part de les funcions d'etiquetatge, serà una inclusió estricta.

5.1.8 Lema. *Per a $m = 2$, el conjunt Z és el conjunt d'assignacions permeses del grup d'elements de les branques dels arbre no arrelats $T = K_{1,m}$.*

Demostració. Cal provar que (g_1, g_2, g_3) són les etiquetes de l'arbre garfi de dos fulles. Però, esta clar ja que la tercera component correspon al node arrel i el valor resulta ser la suma de les fulles que hi ha per sota de l'arrel, i. e., $g_1 + g_2$. Per tant es prova el que volíem. □

5.1.9 Exemples. Sigui $G = \mathbb{Z}/(4)$ i $\mathcal{L} = \{0, 1, 2\}$. Llavors la funció d'etiquetatge L definida per

$$L(0) = 0,\ L(1) = 1\ L(2) = L(3) = 2$$

no és una 3-amistosa perquè $L^{-1}(2) = \{2, 3\}$ estrictament continguda en $\pi_3(\widetilde{L}^{-1}((1, 1, 2))) = \pi_3((1, 1, 2)) = \{2\}$.

La definició de m-amistosa garantitza que si una etiqueta particular λ ve d'una assignació del grup d'elements, llavors alguna elecció d'un grup d'elements a una branca particular e la qual consta amb λ i e pot ser estesa a una assignació que es consisteix amb λ sobre totes les branques de $K_{1,m}$.

5.1.10 Exemples. Prenem $G = \mathbb{Z}/(2) \times \mathbb{Z}/(2)$ i les etiquetes $\mathcal{L} = \{0, 1, 2\}$. La funció d'etiquetatge per al model de Kimura de 2 paràmetres ve definit per

$$L((0, 0)) = 0, \quad L((0, 1)) = 1, \quad L((1, 0)) = L((1, 1)) = 2.$$

5.1.11 Lema. *El funcions d'etiquetatge que són 3-amistoses són amistoses.*

Demostració. Mostrarem que una funció que és 3-amistosa i m-amistosa és també $(m + 1)$-amistosa. Prenem $l \in \widetilde{L}(Z)$. Mostrem que $\pi_{m+1}(\widetilde{L}^{-1}(l)) = L^{-1}(l_{m+1})$. Sigui $l' = (L(g_1 + g_2), L(g_3), ..., L(g_{m+1}))$ on $(g_1, ..., g_{m+1}) \in \widetilde{L}^{-1}(l)$. Ja que L és m-amistosa, a cada $h_{m+1} \in L^{-1}(l_{m+1})$ se li pot assignar un element del grup $h' = (h'_2, h_3, ..., h_{m+1})$. A més, L és 3-amistosa per tant hi ha alguna assignació d'elements del grup (h_1, h_2, h'_2) que satisfà l'etiquetatge $(L(g_1), L(g_2), L(g_1 + g_2))$. Però llavors $h = (h_1, h_2, h_3, ..., h_{m+1})$ té com a imatge $\pi_{m+1}(h) = h_{m+1}$ tal com volíem provar. □

Aquest lema diu que comprovar si una funció és d'etiquetatge es pot fer simplement amb uns càlculs finits. El punt clau per

a estudiar funcions d'etiquetatge és que les etiquetes consistents poden enganxar-se, i per tant podrem enganxar arbres garfis de forma que podrem obtenir qualsevol arbre.

Sigui e una branca interior de l'arbre T. Denotem per $T_{e,-}$ l'arbre obtingut de T prenent la branca e i totes les branques per sota e. Denotem per $T_{e,+}$ l'arbre obtingut de T prenen la branca e i totes les branques de T no incloses en $T_{e,-}$. Llavors tenim el següent:

5.1.12 Lema. *Sigui λ^- i λ^+ etiquetes consistents de $T_{e,-}$ i $T_{e,+}$, respectivament; i. e. $\lambda^- \in \mathrm{im}L^{T_{e,-}}$ i $\lambda^+ \in \mathrm{im}L^{T_{e,+}}$. Denotem per $\lambda(e) := L(g(e))$. Suposem a més que $\lambda^-(e) = \lambda^+(e)$. Llavors les etiquetes λ de T obtingudes de λ^- i λ^+ per branques etiquetades de T apropiades és consistent, i. e., $\lambda \in \mathrm{im}(L^T)$.*

Demostració. Ja que λ^+ i λ^- són consistents, hi ha algunes assignacions de grups d'elements de les branques $T_{e,+}$ i $T_{e,-1}$ que venen de $(L^{T_{e,+}})^{-1}(\mathrm{im}(L^{T_{e,+}}))$ i $(L^{T_{e,-}})^{-1}(\mathrm{im}(L^{T_{e,-}}))$. Construirem una assignació de grups d'elements de les branques de T que pertanyen a $(L^{T_{e,+}})^{-1}(\mathrm{im}(L^{T_{e,+}}))$. Primer, prenem alguna assignació la qual és compatible amb λ^+ de $T_{e,+}$. Aquesta assigna algun grup d'elements a la branca e. Sigui v el vertex que no sigui fulla de $T_{e,-}$ incident a e. Com que L és amistosa i λ^- és consistent, existeix una assignació d'elements del grup per a totes les altres branques incidents a v la qual es compatible amb $\lambda^-(e)$ o és localment consistent. Per inducció sobre el número de nodes interiors de $T_{e,-}$ construïm una assignació global del grup d'elements a les branques de T. □

5.2 Relació entre els invariants filogenètics i les varietats tòriques

A continuació considerem l'ideal $I_{T,L}$ el nucli de l'aplicació (5.7) i construirem els generadors minimals i bases de Gröbner per $I_{T,L}$ a partir d'informació local de l'arbre. Aquestes bases de Gröbner són una llista de binomis $q^u - q^v$, en les indeterminades q_l les quals estan indexats per les etiquetes consistents $l \in \mathrm{im}(L^T)$.

Per a simplificar les notacions, representarem $M = q_{l_1} \cdot \cdots \cdot q_{l_d}$ com una matriu $d \times |E(T)|$:

$$M = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_d \end{bmatrix}.$$

Aquesta matriu amb entrades en $\mathcal{L}$ l'anomenarem taula. Les columnes d'una taula estan indexades per les branques del arbre T considerat. El nombre de files de M és el grau d del monomi.

Dos taules representen el mateix monomi si estan relacionats per permutacions de files. Binomis $q^u - q^v$ en les indeterminades q_l són representats com a diferencies formals $M - M'$ de taules. Notem que es pot comprovar un binomi donat $M - M'$ esta en l'ideal tòric $I_{T,L}$.

5.2.1 Observació. Sigui M i M' dos taules $d \times |E(T)|$ amb entrades en L. Llavors el binomi $M - M'$ està en el ideal $I_{T,L}$ si i només si, es donen les següents condicions:

- cada fila de M i cada fila M' és un etiquetatge consistent per a l'arbre de T,
- per a cada branca $e \in E(T)$, el multiset de etiquetes en la columna e és la mateixa en M i en M'.

Construïm ara els binomis que constitueixen que constituiran les bases de Gröbner de $I_{T,L}$. Sigui e una branca interior de T, i siguin $T_{e,+}$ i $T_{e,-}$ els dos subarbres del lema 5.1.12. Després del reetiquetatge les branques de T, cada taules M poden ser escrites en tres grups de columnes,

$$M = \begin{bmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ \vdots & \vdots & \vdots \\ l_d & m_d & l_d \end{bmatrix}$$

on les columnes de l'esquerra (amb entrades l_i) corresponent a les branques $T_{e,-} \setminus \{e\}$, la del mig correspon a la branca e, i la de la dreta correspon a les branques $T_{e,+} \setminus \{e\}$.

5.2.2 Lema. *Sigui (l_1, m, n_1) i (l_2, m, n_2) un etiquetatge consistent de T. Llavors el binomi quadràtic*

$$g = \begin{bmatrix} l_1 & m & n_1 \\ l_2 & m & n_2 \end{bmatrix} - \begin{bmatrix} l_1 & m & n_2 \\ l_2 & m & n_1 \end{bmatrix} \tag{5.9}$$

esta en el ideal tòric $I_{T,L}$.

Demostració. Els etiquetatges (l_1, m) i (l_2, m) són consistents per al subarbre $T_{e,-}$, i els etiquetatges (m, n_1) i (m, n_2) ho són pel subarbre $T_{e,+}$. Pel lema 5.1.12 l'etiquetatge (l_1, m, n_2) i (l_2, m, n_1) són consistents pel l'arbre gran. Per tant $g \in I_{T,L}$. □

5.2.3 Definició. Es denota per $Quad(e, T)$ el conjunt de tots els binomis quadràtics construïts segons el lema 5.2.2.

Els binomis quadràtics en $Quad(e, T)$ es poden expressar com a determinants 2×2 o menors de $|\mathcal{L}|$ matrius. De fet, per a cada element $m \in \mathcal{L}$, construïm la matriu M_n la qual conté totes les coordenades $q_{(l,m,n)}$ la etiqueta del qual sobre la branca e és m. Les files de les matrius són indexades per etiquetatges consistents

sobres les branques en $T_{e,-} \setminus \{e\}$, i les columnes són indexades per etiquetatges consistents sobre les branques en $T_{e,+} \setminus \{e\}$. D'aquesta manera, qualsevol submatriu 2×2 de M_n té la forma

$$\begin{pmatrix} q_{(l_1,m,n_1)} & q_{(l_1,m,n_2)} \\ q_{(l_2,m,n_1)} & q_{(l_2,m,n_2)} \end{pmatrix}$$

i el seu determinant és precisament l'equació (5.9).

Considerem ara un binomi arbitrari en l'ideal $I_{T,L}$. Té la forma

$$h = \begin{bmatrix} l_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l_d & m_d & n_d \end{bmatrix} - \begin{bmatrix} l'_1 & m'_1 & n'_1 \\ \vdots & \vdots & \vdots \\ l'_d & m'_d & n'_d \end{bmatrix}$$

on les m_i i m'_i són etiquetes senzilles corresponents a les branques e, la l_i i l'_i són etiquetatges consistents de $T_{e,-} \setminus \{e\}$, i el n_i i n'_i són etiquetatges consistents de $T_{e,+} \setminus \{e\}$. Notem que ja que el binomi h pertany a $I_{T,f}$, el multiset d'etiquetes les quals apareixen sobre la branca e ha de ser el mateix en ambdós termes de h. Per tant, després de les reordenacions de les columnes de la taula, podem escriure

$$h = \begin{bmatrix} l_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l_d & m_d & n_d \end{bmatrix} - \begin{bmatrix} l'_1 & m_1 & n'_1 \\ \vdots & \vdots & \vdots \\ l'_d & m_d & n'_d \end{bmatrix}.$$

Cada binomi en $I_{T,L}$ restringeix a un binomi en $I_{T_{e,-},L}$ i a un binomi $I_{T_{e,+},L}$. Concretament, si h és el binomi anterior, llavors el següent binomi esta en $I_{T_{e,-},L}$:

$$h|_{T_{e,-}} = \begin{bmatrix} l_1 & m_1 \\ \vdots & \vdots \\ l_d & m_d \end{bmatrix} - \begin{bmatrix} l'_1 & m_1 \\ \vdots & \vdots \\ l'_d & m_d \end{bmatrix}.$$

De forma similar, esborrant la primera columna produeix un binomi $h|_{T_{e,+}}$ en $I_{T_{e,+},L}$. Ara farem la construcció inversa, és a dir, a partir del binomis en $I_{T_{e,-},L}$ i $I_{T_{e,+},L}$ podem extendre esteses a binomis en $I_{T,L}$.

5.2.4 Lema. *Sigui g un binomi en $I_{T_{e,-},L}$ escrit en notació de taula com*

$$g = \begin{bmatrix} l_1 & m_1 \\ \vdots & \vdots \\ l_d & m_d \end{bmatrix} - \begin{bmatrix} l'_1 & m_1 \\ \vdots & \vdots \\ l'_d & m_d \end{bmatrix}.$$

Sigui $n_1, ..., n_d$ seqüències d'etiquetes tals que cada (m_i, n_i) és un etiquetatge consistent de $T_{e,+}$. Llavors

$$g^* = \begin{bmatrix} l_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l_d & m_d & n_d \end{bmatrix} - \begin{bmatrix} l'_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l'_d & m_d & n_d \end{bmatrix}.$$

és un binomi de $I_{T,L}$.

Demostració. Restringint les dos taules als arbres $T_{e,-}$ i $T_{e,+}$ veiem que el multiset d'etiquetatges les quals apareixen sobre cada branca són les mateixes. De fet tenim

$$g^*|_{T_{e,+}} = g \qquad g^*|_{T_{e,+}} = 0.$$

Hem de comprovar que cada (l_i, m_i, n_i) i (l'_i, m_i, n_i) són unes etiquetes consistents sobre T. Pel lema 5.1.12 implica això, per què (l, m_i) i (l'_i, m_i) són consistents sobre $T_{e,-}$ i (m_i, n_i) són consistents $T_{e,+}$. □

5.2.5 Definició. Sigui $\mathcal{B}$ una col·lecció de binomis en $I_{T_{e,-}}$. Es defineix per $Ext(\mathcal{B} \to T)$ el conjunt de tots el binomis g^* on g recorre $\mathcal{B}$ i $n_1, ..., n_d$ recorre totes les possibles successions d'etiquetes com el lema 5.2.4. De forma similar, definim $Ext(T \leftarrow \mathcal{B})$, per a alguna col·lecció de binomis $\mathcal{B}$ en $I_{T_{e,+}}$.

5.2.6 Teorema. *Sigui T un arbre amb un etiquetatge amistós $L : G \to \mathcal{L}$, i sigui e una branca interior de T. Suposem que $\mathcal{B}_-$ és un conjunt de generadors binomials per a $I_{T_{e,-}}$, i $\mathcal{B}_+$ un*

conjunt de generadors binomials per a $I_{T_{e,-}}$. Llavors el següent conjunt genera l'ideal tòric $T_{T,L}$:

$$Ext(\mathcal{B}_- \to T) \cup Ext(T \leftarrow \mathcal{B}_+) \cup Quad(e,T). \tag{5.10}$$

A més, si $\mathcal{B}_-$ és una base de Gröbner per a $I_{T_{e,-}}$ i $\mathcal{B}_+$ és una base de Gröbner per a $I_{T_{e,-}}$, llavors existeix un ordre en $k[q_\lambda : \lambda \in \mathrm{im}(L^T)]$ tal que el conjunt 5.10 és una base de Gröbner per $I_{T,L}$.

Demostració. Provem primer la segona afirmació respecte les bases de Gröbner. Necessitem especificar les ordres. Sigui $\prec_-$ algun ordre en $k[q_\lambda : \lambda \in im(L^{T_{e,-}})]$ tal que $\mathcal{B}_-$ és una base de Gröbner per a $I_{T_{e,-},L}$ i sigui $\prec_+$ algun ordre en $k[q_\lambda : \lambda \in im(L^{T_{e,+}})]$ tal que $\mathcal{B}_+$ és una base de Gröbner per a $I_{T_{e,+},L}$.

Es defineix ara un ordre lexicogràfic invers $\prec_Q$ el qual fa que $Quad(e,T)$ sigui una base de Gröbner per al ideal que genera. Per a fer això, primer prenem un ordre total $\prec_1$ sobre les etiquetes de la branca e, llavors prenem un ordre total $\prec_2$ sobre les etiquetes consistents de $im(L^{T_{e,-}})$ i $\prec_3$ sobre $im(L^{T_{e,+}})$ les quals són refinaments de $\prec_1$. L'ordre revlex $\prec_Q$ s'obté declarant $q_{\lambda_1} \prec_Q q_{\lambda_2}$ si i només si

$$\lambda_1^- \prec_2 \lambda_2^- \quad \text{o} \quad (\lambda_1^- = \lambda_2^- \quad \text{i} \quad \lambda_1^+ \prec_3 \lambda_2^+).$$

Construïm un ordre producte sobre l'anell de polinomis $k[q_\lambda : \lambda \in \mathrm{im}(L^T)]$ com segueix. Si M i M' són monomis (taules amb columnes indexades per $E(T)$), llavors $M \prec_T M'$ si i només si

1. $M|_{T_{e,-}} \prec_- M'|_{T_{e,-}}$, o
2. $M|_{T_{e,-}} = M'|_{T_{e,-}}$ i $M|_{T_{e,+}} \prec_+ M'|_{T_{e,+}}$, o
3. $M|_{T_{e,-}} = M'|_{T_{e,-}}$ i $M|_{T_{e,+}} = M'|_{T_{e,+}}$ i $M \prec_Q M'$.

L'objectiu és demostrar que el conjunt (5.10) és una base de Gröbner respecte l'ordre $\prec_T$, i.e., que el terme líder de cada binomi g en $I_{T,L}$ sigui divisible pel terme líder d'algun binomi del conjunt (5.10). Per a probar això considerem un binomi arbitrari en el nostre ideal tòric:

$$g = M' - M = \begin{bmatrix} l_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l_d & m_d & n_d \end{bmatrix} - \begin{bmatrix} l'_1 & m'_1 & n'_1 \\ \vdots & \vdots & \vdots \\ l'_d & m'_d & n'_d \end{bmatrix}.$$

Suposem que M' és el terme líder de g. Hi ha precisament tres maneres diferents de que això passi, d'acord amb el tres casos en la definició de $\prec_T$. Cada cas serà analitzat per separat.

Cas 1: Suposem que $M|_{T_{e,-}} \prec_- M'|_{T_{e,-}}$. Llavors $g|_{T_{e,-}}$ és un binomi no zero en $I_{T_{e,-},L}$ i $M'|_{T_{e,-}}$ és el seu terme líder. Com que $\mathcal{B}_-$ és una base de Gröbner, existeix un binomi $h = N' - N \in \mathcal{B}_-$ terme líder del qual, N', divideix M'. Llevat reordenació les files de M' i M, podem suposar que

$$h = N' - N = \begin{bmatrix} l_1 & m_1 \\ \vdots & \vdots \\ l_i & m_i \end{bmatrix} - \begin{bmatrix} l''_1 & m_1 \\ \vdots & \vdots \\ l''_i & m_i \end{bmatrix} \qquad \text{per algun} \quad i \leq d.$$

Aquí $l_1, ..., l_i$ i $m_1, ..., m_i$ són les mateixes etiquetes que apareixen en M'. Ara considerem el binomi $h^* \in Ext(\mathcal{B}_- \to T)$ obtinguda ajuntant les etiquetes $n_1, ..., n_i$:

$$h^* = (N')^* - N^* = \begin{bmatrix} l_1 & m_1 & n_1 \\ \vdots & \vdots & \vdots \\ l_i & m_i & n_i \end{bmatrix} - \begin{bmatrix} l''_1 & m_1 & n_i \\ \vdots & \vdots & \vdots \\ l''_i & m_i & n_i \end{bmatrix}.$$

La taula $(N')^*$ és el terme líder de h^* amb respecte a $\prec_T$, i $(N')^*$ divideix com volíem.

Cas 2: Suposem $M|_{T_{e,-}} = M'|_{T_{e,-}}$ i $M|_{T_{e,+}} \prec_T M'|_{T_{e,+}}$. Llavors pel mateix argument com en el Cas 1, deduïm que hi ha un binomi $h^* \in Ext(T \leftarrow \mathcal{B}_+)$ que té com a terme líder M'.

Cas 3: Suposem que $M|_{T_{e,-}} = M'|_{T_{e,-}}$ i $M|_{T_{e,+}} = M'|_{T_{e,+}}$ i $M \prec_Q M'$. L'única manera que això passi és si existeix un parell de files en $M', (l_1, m, n_1)$ i (l_2, m, n_2), tal que $(l_1, m) \prec_1 (l_2, m)$ i $(m, n_1) \succ_2 (m, n_2)$. Però llavors el binomi $h \in Quad(e, T)$ donat per

$$h = N' - N = \begin{bmatrix} l_1 & m & n_1 \\ l_2 & m & n_2 \end{bmatrix} - \begin{bmatrix} l_1 & m & n_2 \\ l_2 & m & n_1 \end{bmatrix}$$

amb terme líder N', i aquest terme divideix el terme M' del binomi g'.

D'aquesta tres casos junts estableixen la segona afirmació: el conjunt (5.10) és una base de Gröbner per al ideal $I_{T,L}$. A més, per a unes bases de Gröbner $\mathcal{B}_-$ i $\mathcal{B}_+$, tenim la igualtat següent d'ideals

$$I_{T,L} = \langle Ext(\mathcal{B}_- \to T)\rangle + \langle Ext(\mathcal{B}_+ \leftarrow T)\rangle + \langle Quad(e, T)\rangle.$$

Podem acabar observant que en aquesta equació, podem reemplaçar $Ext(\mathcal{B}_- \to T)$ pel conjunt que el genera $\langle Ext(\mathcal{B}_- \to T)\rangle$. En aquest cas, però $Ext(\mathcal{C}_- \to T)$ generarà $\langle Ext(\mathcal{B}_- \to T)\rangle$ sempre que $\mathcal{C}_-$ sigui un conjunt generat per $I_{T_{e,-},L}$. Una afirmació similar és vàlida per a $I_{T_{e,+},L}$. Això completa la demostració de la primera afirmació. □

Sigui v un node interior de l'arbre T, i sigui $e_1, ..., e_c$ les branques de T incidents a v. Denotem per T_{v,e_i} el subarbre $T_{e_i,-}$ o $T_{e_i,+}$ el qual té v com una fulla. Donada una etiqueta fixa l per la branca e_i, denotem $\mathrm{im}(L^{T_{v,e_i}}, l)$ el conjunt de totes les etiquetes consistents de T_{v,e_i} la qual té d'etiqueta e_i igual l. Denotem per T_v el subarbre de T amb només el node interior v i la branca $e_1, ..., e_c$. Notem que T_v és un arbre garfi $K_{1,c}$.

5.2.7 Lema. *Sigui g un binomi del ideal de l'arbre garfi T_v escrit en notació de taula com*

$$g = \begin{bmatrix} l_1^1 & \cdots & l_1^c \\ \vdots & \vdots & \vdots \\ l_d^1 & \cdots & l_d^c \end{bmatrix} - \begin{bmatrix} m_1^1 & \cdots & m_1^c \\ \vdots & \vdots & \vdots \\ m_d^1 & \cdots & m_d^c \end{bmatrix} \in I_{T_v,L}.$$

Per a cada fila i i columna j, considerem els etiquetatges $L_i^j \in \mathrm{im}(L^{T_{v,e_j}}, l_i^j))$ i $M_i^j \in L_i^j \in \mathrm{im}(L^{T_{v,e_j}}, l_i^j))$ amb la propietat que el multiset $\{L_i^h\}_{j=1}^d$ és igual al multiset $\{M_i^j\}_{j=1}^d$. Llavors el binomi

$$g^* = \begin{bmatrix} L_1^1 & \cdots & L_1^c \\ \vdots & \vdots & \vdots \\ L_d^1 & \cdots & L_d^c \end{bmatrix} - \begin{bmatrix} M_1^1 & \cdots & M_1^c \\ \vdots & \vdots & \vdots \\ M_d^1 & \cdots & M_d^c \end{bmatrix},$$

pertany al ideal tòric $I_{T,L}$ de l'arbre T.

Demostració. Atès que L és amistosa, cada fila en la taula és un etiquetatge consistent. Restringim a cada subarbre hi ha el mateix multiset d'etiquetes. Per tant,el binomi g^* és en $I_{T,L}$. □

5.2.8 Definició. Sigui $\mathcal{B}$ qualsevol conjunt de binomis en el arbre ideal garfí $I_{T_v,L}$. Denotem per $Ext(\mathcal{B} \to T)$ el conjunt de tots binomis g^* obtinguts per aplicació de la construcció en 5.2.7 als binomis $g \in \mathcal{B}$.

5.2.9 Teorema. (Estructura local dels invariants) *Sigui T un arbre amb una funció d'etiquetatge amistosa $L : G \to \mathcal{B}$. Per a cada node interior v de l'arbre T, denotem $\mathcal{B}_v$ un conjunt de generadors binomials de $I_{T_v,L}$. Llavors el següent conjunt de binomis genera l'ideal $I_{T,L}$ de tots els invariants filogenètic de T:*

$$\bigcup_v Ext(\mathcal{B}_v \to T) \cup \bigcup_e Quad(e, T). \tag{5.11}$$

La primera unió és sobre els nodes interiors de T. La segona unió és sobre les branques interiors de T.

Demostració. Procedim per hipòtesis inducció sobre el nombre de nodes interiors de T. Si hi ha només un, llavors l'afirmació és una tautològica. Suposem que hi ha $m \geq 2$ nodes interiors. Llavors existeix un nodes interior el qual és incident a només un altre node interior u. Sigui e la branca que conecta

v amb u. L'arbre $T_{e,-}$ té $m-1$ nodes interiors, i l'arbre $T_{e,+}$ té només un node. Per inducció, el corresponent ideal té com a conjunt de generador, el conjunt de la forma (5.11). L'aplicació del teorema 5.2.6 produeix un conjunt de generadors per $I_{T,L}$ el qual és més gran que el llistat de polinomis del conjunt descrit en (5.11). Afirmem que cada binomi en la diferència (5.10) $\setminus$ (5.11) està en l'ideal generat per (5.11). De fet, cada binomi de (5.10) $\setminus$ (5.11) difereix d'un de (5.11) per intercanvi d'algunes etiquetes en les columnes corresponents a l'arbre $T_{v,e}$. Un tal intercanvi pot ocórrer només quan l'etiqueta de la branca e és la mateixa per a cada fila de la taula involucrada en el intercanvi. Per tals intercanvis (o successions d'intercanvis) poden ser realitzades afegit-hi un nombre multiple de binomis quadràtics en $Quad(e, T)$. □

5.3 Jukes-Cantor per a un arbre sense arrel de quatre fulles

En aquesta secció anem a fer els càlculs i algoritmes descrits en les dos seccions anteriors per tal de veure amb un exemple tot el procés.

Per tal de calcular els invariants de l'arbre de quatre fulles, i per a poder aplicar el procediment explicant en seccions anteriors el primer pas serà estudiar l'arbre més senzill possible que correspon al l'arbre de dos fulles, amb distribució no uniforme o de tres fulles amb distribució uniforme. Aquest arbre corresponent a l'arbre $K_{1,2}$ representat en la figura 5.2.

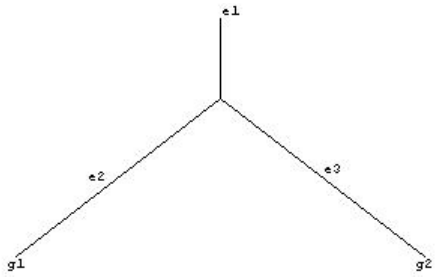


Fig. 5.2: Arbre de dos fulles amb la branca extra

Primer de tot hem d'assignar un valor com a element d'un grup finit a les bases,

$$\begin{array}{ccc} \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\} & \to & \mathbb{Z}/(2)\times\mathbb{Z}/(2) \\ \mathtt{A} & \mapsto & (0,0) \\ \mathtt{C} & \mapsto & (1,0) \\ \mathtt{G} & \mapsto & (0,1) \\ \mathtt{T} & \mapsto & (1,1). \end{array}$$

Amb aquesta assignació i considerant el model de Jukes-Cantor el conjunt d'etiquetes és senzillament $L = \{0,1\}$, i la funció d'etiquetatge corresponent és

$$\begin{array}{cccc} L: & G & \to & \mathcal{L} \\ & \mathtt{A} & \mapsto & 0 \\ & \mathtt{X} \neq A & \mapsto & 1. \end{array}$$

Primer de tot comencem estudiant l'arbre de la figura 5.2. Assignem a cada branca l'element del grup que li pertoca, és a dir,

$$g(e) = \sum_{v\in\Lambda(e)} g_v,$$

on $\Lambda(e)$ denota el conjunt de fulles per sota de e. En el nostre cas

$$\begin{array}{ll} g(e1) & = g_1 + g_2 \\ g(e2) & = g_1 \\ g(e3) & = g_2. \end{array}$$

Notarem que $a_1a_2a_3$ al valor de l'etiqueta del arbre, corresponent a la branca i. Així doncs la funció d'etiquetatge ens quedarà

$$\begin{array}{rccc} L^T: & G\times G & \to & L\times L\times L \\ & \texttt{AA} & \mapsto & 000 \\ & \texttt{AM} & \mapsto & 101 \\ & \texttt{MA} & \mapsto & 110 \\ & \texttt{MM} & \mapsto & 011 \\ & \texttt{MN} & \mapsto & 111, \end{array}$$

on M i N són una base diferent de $\texttt{A}$ i $M \neq N$. Observem que aquestes són els únics 5 casos que podem tenir. Per tant en aquest cas, el conjunt d'etiquetes consistents és

$$\mathrm{im}(L^T) = \{000, 011, 101, 110, 111\}.$$

Ara hem de buscar els successius invariants lineals. Segons la proposició 5.1.3 els polinomis lineals de l'ideal $I_{T,L}$ són de la forma $q_{g_1g_2g_3} - q_{h_1h_2h_3}$ amb $L^T(g_1g_2) = L^T(h_1h_2)$. Per exemple, $q_{\texttt{CC}}, q_{\texttt{GG}}$ i $q_{\texttt{TT}}$ compleixen que $L^T(CC) = L^T(GG) = L^T(TT)$. Per tant correspondran els invariants lineals següents:

$$\begin{aligned} z_1 &= q_{CC} - q_{GG} \\ z_2 &= q_{CC} - q_{TT} \\ z_3 &= q_{GG} - q_{TT} = z_2 - z_1. \end{aligned}$$

En aquest cas, en ser l'arbre més senzill, construir arbres per la dreta i per la esquerra no té sentit, però es pot comprovar que l'ideal dels invariants filogenètics està format per

$$I_{K_{1,2},T} = \langle q_{000}q_{111}^2 - q_{011}q_{101}q_{110} \rangle,$$

ja que aquest polinomi compleix l'observació 5.2.1.

Procedim ara a calcular alguns dels invariants filogenètics de l'arbre dibuixat en la figura 5.3, a partit del calculat fins ara.

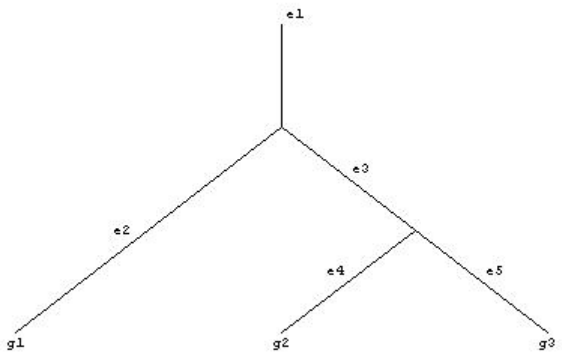


Fig. 5.3: Arbre de dos fulles amb la branca extra

Aquest arbre es pot interpretar de dos maneres diferents: com a arbre de tres fulles amb arrel, o bé com a arbre de binari balancejat de 4 fulles amb distribució uniforme en les bases del node arrel. Per a veure les etiquetes consistents del nostre arbre enganxem segons el lema 5.1.11. Ajuntem les branques per la branca e3.

Agafem les etiquetes que acaben amb 0 i per un altre costat les que hi comencem,

$$\begin{array}{c|c} 000 & 000 \\ 110 & 011 \end{array}$$

i enganxant segons el lema tenim

$$\begin{array}{c} 00000 \\ 11000 \\ 00011 \\ 11011 \end{array}$$

i de les etiquetes que acaben amb 1 i per un altre costat les que hi comencen

$$\begin{array}{c|c} 011 & 110 \\ 101 & 101 \\ 111 & 111 \end{array}$$

i enganxant segons el lema tenim

01110
01101
01111
10110
10101
10111
11110
11101
11111.

Aquestes etiquetes són el conjunt d'etiquetes consistents per al nou arbre. Es pot comprovar mitjançant el procediment constructiu que l'afirmació és certa. A cada una de les combinacions de 3 bases li assignem una etiqueta mitjançant l'aplicació

$$\begin{array}{rccl} L^T: & G^3 & \to & L^5 \\ & g_1g_2g_3 & \mapsto & (L(g(e1)), L(g(e2)), L(g(e3)), L(g(e4)), L(g(e5))) \end{array}$$

on la funció g pren els següents valor en cada una de les branques de l'arbre T.

$$\begin{aligned} g(e1) &= g_1 \\ g(e2) &= g_1 + g_2 + g_3 \\ g(e3) &= g_2 + g_3 \\ g(e4) &= g_2 \\ g(e5) &= g_3. \end{aligned}$$

Amb aquest passos podem veure que les etiquetes obtingudes amb el primer procediment són les mateixes que en el segon.

El següent pas serà calcular els generadors de l'ideal tòric dels invariants filogenètics. Primer de tot descartem els invariants

Classe	Etiqueta	Parametrització	Exemple de bases
1	00000	$a_0^{(1)}a_0^{(2)}a_0^{(3)}a_0^{(4)}a_0^{(5)}$	AAA
2	00011	$a_0^{(1)}a_0^{(2)}a_0^{(3)}a_1^{(4)}a_1^{(5)}$	ATT
3	01101	$a_0^{(1)}a_1^{(2)}a_1^{(3)}a_0^{(4)}a_1^{(5)}$	AAT
4	01110	$a_0^{(1)}a_1^{(2)}a_1^{(3)}a_1^{(4)}a_0^{(5)}$	ATA
5	01111	$a_0^{(1)}a_1^{(2)}a_1^{(3)}a_1^{(4)}a_1^{(5)}$	ATC
6	10101	$a_1^{(1)}a_0^{(2)}a_1^{(3)}a_0^{(4)}a_1^{(5)}$	TAT
7	10110	$a_1^{(1)}a_0^{(2)}a_1^{(3)}a_1^{(4)}a_0^{(5)}$	TTA
8	10111	$a_1^{(1)}a_0^{(2)}a_1^{(3)}a_1^{(4)}a_1^{(5)}$	TGC
9	11000	$a_1^{(1)}a_1^{(2)}a_0^{(3)}a_0^{(4)}a_0^{(5)}$	TAA
10	11011	$a_1^{(1)}a_1^{(2)}a_0^{(3)}a_1^{(4)}a_1^{(5)}$	TGG
11	11101	$a_1^{(1)}a_1^{(2)}a_1^{(3)}a_0^{(4)}a_1^{(5)}$	TAC
13	11110	$a_1^{(1)}a_1^{(2)}a_1^{(3)}a_1^{(4)}a_0^{(5)}$	TCA
12	11111	$a_1^{(1)}a_1^{(2)}a_1^{(3)}a_1^{(4)}a_1^{(5)}$	TCT

Taula 5.1: Taula de classes amb parametritzacions

lineals. Considerem l'aplicació

$$\begin{array}{lcl} k[q_{g_1g_2g_3}] & \to & k[a_0^{(1)}, a_1^{(1)}, a_0^{(2)}, a_1^{(2)}, a_0^{(3)}, a_1^{(3)}, a_0^{(4)}, a_1^{(4)}, a_0^{(5)}, a_1^{(5)}] \\ q_{AAA} & \mapsto & a_0^{(1)}a_0^{(2)}a_0^{(3)}a_0^{(4)}a_0^{(5)} \\ q_{AAC} & \mapsto & a_0^{(1)}a_1^{(2)}a_1^{(3)}a_0^{(4)}a_1^{(5)} \\ q_{CTT} & \mapsto & a_1^{(1)}a_1^{(2)}a_0^{(3)}a_1^{(4)}a_1^{(5)} \\ q_{GCC} & \mapsto & a_1^{(1)}a_1^{(2)}a_0^{(3)}a_1^{(4)}a_1^{(5)} \\ q_{TGC} & \mapsto & a_1^{(1)}a_0^{(2)}a_1^{(3)}a_1^{(4)}a_1^{(5)} \end{array} \tag{5.12}$$

Primer de tot calculem $Quad(e, T)$ per l'única branca en el interior del nostre arbre, és a dir, per la 3. Farem servir la notació amb el número de classe corresponent. Es pot observar que aquest corresponent al invariants menors 2×2 provinent de

les matrius

$$M_0 = \begin{pmatrix} q_1 & q_2 \\ q_9 & q_{10} \end{pmatrix}$$
$$M_1 = \begin{pmatrix} q_7 & q_6 & q_8 \\ q_4 & q_3 & q_5 \\ q_{13} & q_{11} & q_{12} \end{pmatrix},$$

fent els càlculs queda

$$\begin{array}{lll} q_1q_{10} - q_2q_9, & q_3q_7 - q_4q_6, & q_3q_8 - q_5q_6, \\ q_4q_8 - q_5q_7, & q_3q_{13} - q_4q_{11}, & q_3q_{12} - q_5q_{11}, \\ q_4q_{12} - q_5q_{13}, & q_6q_{13} - q_7q_{11}, & q_6q_{12} - q_8q_{11}, \\ q_7q_{12} - q_8q_{13}. & & \end{array}$$

Per a calcular la resta segons (5.11) queda calcular els conjunts $Ext(\mathcal{B}_v \to T)$, per a tots els nodes interiors, en el nostre cas en tenim dos i tenim que pel vertex de l'esquerra ens queden els invariants següents

$$\begin{array}{lll} q_1q_{11}q_{11} - q_3q_6q_9, & q_1q_{11}q_{13} - q_3q_7q_9, & q_1q_{11}q_{12} - q_3q_8q_9, \\ q_1q_{13}q_{11} - q_4q_6q_9, & q_1q_{13}q_{13} - q_4q_7q_9, & q_1q_{13}q_{12} - q_4q_8q_9, \\ q_1q_{12}q_{11} - q_5q_6q_9, & q_1q_{12}q_{13} - q_5q_7q_9, & q_1q_{12}q_{12} - q_5q_8q_9, \\ q_2q_{11}q_{11} - q_3q_6q_{10}, & q_2q_{11}q_{13} - q_3q_7q_{10}, & q_2q_{11}q_{12} - q_3q_8q_{10}, \\ q_2q_{13}q_{11} - q_4q_6q_{10}, & q_2q_{13}q_{13} - q_4q_7q_{10}, & q_2q_{13}q_{12} - q_4q_8q_{10}, \\ q_2q_{12}q_{11} - q_5q_6q_{10}, & q_2q_{12}q_{13} - q_5q_7q_{10}, & q_2q_{12}q_{12} - q_5q_8q_{10}. \end{array}$$

De l'altre node interior obtenim

$$\begin{array}{lll} q_1q_5q_5 - q_3q_4q_2, & q_1q_5q_8 - q_3q_7q_2, & q_1q_5q_{12} - q_3q_{13}q_2, \\ q_1q_8q_5 - q_6q_4q_2, & q_1q_8q_8 - q_6q_7q_2, & q_1q_8q_{12} - q_6q_{13}q_2, \\ q_1q_{12}q_5 - q_{11}q_4q_2, & q_1q_{12}q_8 - q_{11}q_7q_2, & q_1q_{12}q_{12} - q_{11}q_{13}q_2, \\ q_9q_5q_5 - q_3q_4q_{10}, & q_9q_5q_8 - q_3q_7q_{10}, & q_9q_5q_{12} - q_3q_{13}q_{10}, \\ q_9q_8q_5 - q_6q_4q_{10}, & q_9q_8q_8 - q_6q_7q_{10}, & q_9q_8q_{12} - q_6q_{13}q_{10}, \\ q_9q_{12}q_5 - q_{11}q_4q_{10}, & q_9q_{12}q_8 - q_{11}q_7q_{10}, & q_9q_{12}q_{12} - q_{11}q_{13}q_{10}. \end{array}$$

5.4 Altres models de grups

En aquesta secció es donaran els càlculs explícits per al model de Kimura de 3 paràmetres per a l'arbre garfi de dos fulles. En aquest cas la funció d'etiquetatge esdevé la identitat.

En aquest cas sigui $n = 2$ i $G = \mathbb{Z}/(2) \times \mathbb{Z}/(2)$. Identifiquem el grup d'element amb els nucleotits com abans. Llavors l'ideal I_G és un ideal en l'anell de polinomis

$$k[q_{AA}, q_{AG}, q_{AC}, q_{AT}, ..., q_{TA}, q_{TG}, q_{TC}, q_{TT}].$$

Es pot procedir com abans i tindrem que els polinomis que estan en el nucli, estan mínimament generats per les 16 cúbiques

$$\begin{array}{ll}
q_{AA}q_{CT}q_{TG} - q_{AG}q_{CA}q_{TT}, & q_{AA}q_{GT}q_{TC} - q_{AC}q_{GA}q_{TT}, \\
q_{AC}q_{CT}q_{TA} - q_{AT}q_{CA}q_{TC}, & q_{AC}q_{GG}q_{TA} - q_{AA}q_{GC}q_{TG}, \\
q_{AG}q_{CC}q_{TA} - q_{AA}q_{CG}q_{TC}, & q_{AG}q_{GC}q_{CA} - q_{AC}q_{GA}q_{CG}, \\
q_{AG}q_{GT}q_{CC} - q_{AC}q_{GG}q_{CT}, & q_{AG}q_{GT}q_{TA} - q_{AT}q_{GA}q_{TG}, \\
q_{AT}q_{CC}q_{TG} - q_{AC}q_{CG}q_{TT}, & q_{AT}q_{GA}q_{CC} - q_{AA}q_{GC}q_{CT}, \\
q_{AT}q_{GG}q_{CA} - q_{AA}q_{GT}q_{CG}, & q_{AT}q_{GG}q_{TC} - q_{AG}q_{GC}q_{TT}, \\
q_{GA}q_{CC}q_{TG} - q_{GG}q_{CA}q_{TC}, & q_{GC}q_{CT}q_{TG} - q_{GT}q_{CG}q_{TC}, \\
q_{GG}q_{CT}q_{TA} - q_{GA}q_{CG}q_{TT}, & q_{GT}q_{CC}q_{TA} - q_{GC}q_{CA}q_{TT}.
\end{array}$$

i les 18 quàrtiques

$$\begin{array}{ll}
q_{AA}q_{AT}q_{TG}q_{TC} - q_{AG}q_{AC}q_{TA}q_{TT}, & q_{AA}q_{GG}q_{CT}q_{TC} - q_{AG}q_{GA}q_{CC}q_{TT}, \\
q_{AA}q_{GT}q_{CC}q_{TG} - q_{AC}q_{GG}q_{CA}q_{TT}, & q_{AA}q_{GT}q_{CT}q_{TA} - q_{AT}q_{GA}q_{CA}q_{TT}, \\
q_{AC}q_{AT}q_{GA}q_{GG} - q_{AA}q_{AG}q_{GC}q_{GT}, & q_{AC}q_{GA}q_{CC}q_{TA} - q_{AA}q_{GC}q_{CA}q_{TC}, \\
q_{AC}q_{GA}q_{CT}q_{TG} - q_{AG}q_{GT}q_{CA}q_{TC}, & q_{AC}q_{GT}q_{CG}q_{TA} - q_{AT}q_{GC}q_{CA}q_{TG}, \\
q_{AG}q_{AT}q_{CA}q_{CC} - q_{AA}q_{AC}q_{CG}q_{CT}, & q_{AG}q_{GC}q_{CC}q_{TG} - q_{AC}q_{GG}q_{CG}q_{TC} \\
q_{AG}q_{GC}q_{CT}q_{TA} - q_{AT}q_{GA}q_{CG}q_{TC}, & q_{AG}q_{GG}q_{CA}q_{TA} - q_{AA}q_{GA}q_{CG}q_{TG}, \\
q_{AT}q_{GG}q_{CC}q_{TA} - q_{AA}q_{GC}q_{CG}q_{TT}, & q_{AT}q_{GG}q_{CT}q_{TG} - q_{AG}q_{GT}q_{CG}q_{TT}, \\
q_{AT}q_{GT}q_{CC}q_{TC} - q_{AC}q_{GC}q_{CT}q_{TT}, & q_{CC}q_{CT}q_{TA}q_{TG} - q_{CA}q_{CG}q_{TC}q_{TT}, \\
q_{GA}q_{GT}q_{CG}q_{CC} - q_{GG}q_{GC}q_{CA}q_{CT}, & q_{GG}q_{GT}q_{TA}q_{TC} - q_{GA}q_{GC}q_{TG}q_{TT}.
\end{array}$$

Per a quatre fulles seguint aquest algoritme tindriem 8002 polinomis de graus 2, 3 i 4.

5.5 Altres algoritmes

Un altre algoritme de construcció per a un conjunt de generadors de l'ideal filogenètic és el proposat per l'article [Cas07].

Usant el mètode que hem exposat en l'apartat anterior es pot comprovar que per al model de Kimura 3 parametres, $d = 4$, s'obtenen, amb un arbre de quadrivalent, amb $n = 4$ s'obtenen 8002 polinomis de grau 2,3 i 4. Aquest nombre és molt elevat tenint en compte que la codimensió de la varietat és 48. Encara que a més, com més incrementa el nombre d'espècies, el conjunt de generadors augmenta exponencialment. El que fa que per a un nombre molt gran el mètode proposat sigui ineficient per a inferir tals arbres.

Donen també un algoritme per calcular els generadors. Aquest, resulten en binomis de grau 2 i 4 per a qualsevol nombre de fulles n.

En el cas un arbre de quatre fulles balancejat tindrem com a invariants les 36 quàdriques

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

i les 12 quàrtiques

$$
\begin{aligned}
q_{\mathtt{AAAA}}q_{\mathtt{ATTA}}q_{\mathtt{TCGA}}q_{\mathtt{TGCA}} &- q_{\mathtt{ACCA}}q_{\mathtt{AGGA}}q_{\mathtt{TATA}}q_{\mathtt{TTAA}},\\
q_{\mathtt{CCAA}}q_{\mathtt{CTGA}}q_{\mathtt{TATA}}q_{\mathtt{TGCA}} &- q_{\mathtt{CACA}}q_{\mathtt{CGTA}}q_{\mathtt{TCGA}}q_{\mathtt{TTAA}},\\
q_{\mathtt{AGGA}}q_{\mathtt{ATTA}}q_{\mathtt{CACA}}q_{\mathtt{CCAA}} &- q_{\mathtt{AAAA}}q_{\mathtt{ACCA}}q_{\mathtt{CGTA}}q_{\mathtt{CTGA}},\\
q_{\mathtt{ACCA}}q_{\mathtt{ATTA}}q_{\mathtt{GAGA}}q_{\mathtt{GGAA}} &- q_{\mathtt{AAAA}}q_{\mathtt{AGGA}}q_{\mathtt{GCTA}}q_{\mathtt{GTCA}},\\
q_{\mathtt{CACA}}q_{\mathtt{CTGA}}q_{\mathtt{GCTA}}q_{\mathtt{GGAA}} &- q_{\mathtt{CCAA}}q_{\mathtt{CGTA}}q_{\mathtt{GAGA}}q_{\mathtt{GTCA}},
\end{aligned}
$$

$$
\begin{aligned}
q_{\mathtt{GGAA}}q_{\mathtt{GTCA}}q_{\mathtt{TATA}}q_{\mathtt{TCGA}} &- q_{\mathtt{GAGA}}q_{\mathtt{GCTA}}q_{\mathtt{TGCA}}q_{\mathtt{TTAA}},\\
q_{\mathtt{AAAA}}q_{\mathtt{AATT}}q_{\mathtt{TACG}}q_{\mathtt{TAGC}} &- q_{\mathtt{AACC}}q_{\mathtt{AAGG}}q_{\mathtt{TAAT}}q_{\mathtt{TATA}},\\
q_{\mathtt{CACA}}q_{\mathtt{CATG}}q_{\mathtt{TAAT}}q_{\mathtt{TAGC}} &- q_{\mathtt{CAAC}}q_{\mathtt{CAGT}}q_{\mathtt{TACG}}q_{\mathtt{TATA}},\\
q_{\mathtt{AAGG}}q_{\mathtt{AATT}}q_{\mathtt{CAAC}}q_{\mathtt{CACA}} &- q_{\mathtt{AAAA}}q_{\mathtt{AACC}}q_{\mathtt{CAGT}}q_{\mathtt{CATG}},\\
q_{\mathtt{AACC}}q_{\mathtt{AATT}}q_{\mathtt{GAAG}}q_{\mathtt{GAGA}} &- q_{\mathtt{AAAA}}q_{\mathtt{AAGG}}q_{\mathtt{GACT}}q_{\mathtt{GATC}},\\
q_{\mathtt{CAAC}}q_{\mathtt{CATG}}q_{\mathtt{GACT}}q_{\mathtt{GAGA}} &- q_{\mathtt{CACA}}q_{\mathtt{CAGT}}q_{\mathtt{GAAG}}q_{\mathtt{GATC}},\\
q_{\mathtt{GAGA}}q_{\mathtt{GATC}}q_{\mathtt{TAAT}}q_{\mathtt{TACG}} &- q_{\mathtt{GAAG}}q_{\mathtt{GACT}}q_{\mathtt{TAGC}}q_{\mathtt{TATA}}.
\end{aligned}
$$

Evidentment aquest algoritme és més eficient que l'anterior.

Capítol 6

Exemples pràctics sobre filogenètica.

En aquest capítol donarem dos exemples generals d'aplicació de la teoria donada en les seccions anteriors a la biologia. El primer té la finalitat de probar que els vertebrats ténen una série de bases en comú, que no s'han arribat per casualitat, sinò que totes elles venen d'un ancestre comú. L'exemple esta estret de l'àrticle [**?**].

El segon té un caràcter divulgatiu i tracta d'explicar de com és pot inferir un arbre filogenètic mitjançant invariants filogenètics. Esta extret de l'article [**?**].

6.1 Sobre la conjectura el "significat de la vida".

En aquesta secció veurem una aplicació pràctica de com funciona els estudis sobre filogenètica. Descriurem aquí els passos per provar la conjectura anomenada "el significat de la vida".

6.1.1 Conjetura. (Significat de la Vida) *La seqüencia de 42 bases*

TTTAATTGAAAGAAGTTAATTGAATGTGAAAATGATCAACTAGG

estaba present en el genoma del antecesor de tots els vertebrats, i s'ha conservat completament fins els nostres dies (i. e., ninguna de les bases ha mutat, ni han hagut insersión ni eliminacions).

El que s'ha fet aquí és provar que la probabilitat que totes les cadenes s'hagin donat de forma aleatòria. Per a provar la conjectura s'hauria de donar que la probabilitat és zero.

El procés de demostració en aquest cas consisteix en reconstruir un arbre mitjançant eines estadístiques i considerar uns valors de l'espai de paràmetres que en vindran a reflectir el fet que les mutacions s'hagin donat per casualitat.

Demostració.

Pas 1: Aconseguir els genomes. El National Center for Biotechnology Information (NCBI)

$$http://www.ncbi.nlm.nih.gov$$

manté una base de dades pública anomenada GENBANK que contè totes les seqüències de genoma públiques de tot el món. El grans centres de seqüenciació que reben fons públics generalment se'ls obliga a dipositar files de seqüències en aquesta base de dades. La base de dades GENBANK ha passat de contenir 680.000 parells de bases quan ba començar en 1982, i va anar creixent fins a 49 milions l'any 1990. Actualment hi ha 44 mil millions de parells de DNA en GENBANK.

Els 10 genomes de les espècies de vertebrats que estudiarem en aquesta conjectura, encara que no estan penjades en la seva totalitat, la part dels gens que estan involucrats si hi són. Per

tant, ja que tenim seqüències comparables podem dur a terme l'estudi.

Pas 2: Anotar els gens. Anomenem anotar un gen a descriure la funcionalitat d'aquest.

Per a respondre la nostra pregunta necessitem conèixer quins són els gens que volem estudiar dins del genoma de cada espècie. Alguns genomes tenen funcionalitats que s'han derivat experimentalment, però usualment s'usen certs algoritmes. Aquestes anotacions són reconstruides per grans centres tals com UC Santa Cruz

$$http://genome.ucsc.esu/$$

així com autors individuals de programes. Queda obert el problema d'anotar exactament genomes.

Pas 3: Alineament. Es tracta aquí el procès de donar un alineament de les seqüències d'ADN.

Els mètodes actuals per alinear genomes complets estan basats bàsicament en certs mètodes. Encara que a la pràctica no és possible alinear seqüències que continguin milers de milions, ni tan sols milions de parells de bases amb els algoritmes directament, existeixen subrutines que implementen estratègies d'alineament més complexes les quals inicialment identifiquen regions més petites per a l'alineament posterior dels genomes sencers.

Pas 4: buscant DNA neutral. Per a calcular la probabilitat de que una certa subseqüència es conserva entre genomes, és necessari estimar la taxa neutral de evolució. Això és, estimar paràmetres per a un model d'evolució de parells de bases en genoma que no estigui sota selecció i que muta més aleatòriament. Com que regions neutrals són difícils de identificar a priori, habitualment s'usen substitucions sinònimes dels triplets a estudiar.

Pas 5: Buscar una métrica. Voldríem idealment usar tècniques de versemblança per a reconstruir un arbre T mitjançant la longitud de les branques, això és, a cada branca li

podem assignar un nombre que indiqui pròximitat

Una aproximació és intentar usar una aproximació mitjançant mètodes de maxima versemblança, però això és difícil de dur a terme degut a la complexitat de les equacions de versemblança, fins i tot per al model de Jukes-Cantor, amb només 10 espècies.

Una aproximació alternativa és estimar les distàncies de les espècies per parelles usant certa fórmula per mesurar distàncies que depen del mòdel.

La fórmula resultant dóna una métrica en un conjunt de les espècies que s'estiguin estudiant. En el cas del nostre problema segons el model de Jukes-Cantor, la mètrica vindrà donada per la fórmula

$$\delta_{12} = -\frac{3}{4} \cdot \log\left(1 - \frac{4k}{3n}\right)$$

amb n la longitud de la cadena i k el nombre de diferències.

El resultat de aplicar la mètrica de Jukes-Cantor a la nostra conjectura ens dóna la taula 6.1.

	gg	hs	mm	pt	rn	cf	dr	tn	tr	xt
gg	–	0.831	0.928	0.831	0.925	0.847	1.321	1.326	1.314	1.121
hs	–	–	0.414	0.013	0.411	0.275	1.296	1.274	1.290	1.166
mm	–	–	–	0.413	0.176	0.441	1.256	1.233	1.264	1.218
pt	–	–	–	–	0.411	0.275	1.291	1.267	1.288	1.160
rn	–	–	–	–	–	0.443	1.255	1.233	1.258	1.212
cf	–	–	–	–	–	–	1.300	1.251	1.269	1.154
dr	–	–	–	–	–	–	–	1.056	1.067	1.348
tn	–	–	–	–	–	–	–	–	0.315	1.456
tr	–	–	–	–	–	–	–	–	–	1.437

Taula 6.1. Resultant de la mètrica.

Pas 6: Construint un arbre.

Un cop hem trobat una mètrica per a l'arbre, el que es fa és trobar un valor per a les probabilitats. En aquest cas usarem les parametritzacions que hem fet de la varietat i usarem el punt de l'espai de sortida trobat en el pas 5 i es trobara el valor de les probabilitats.

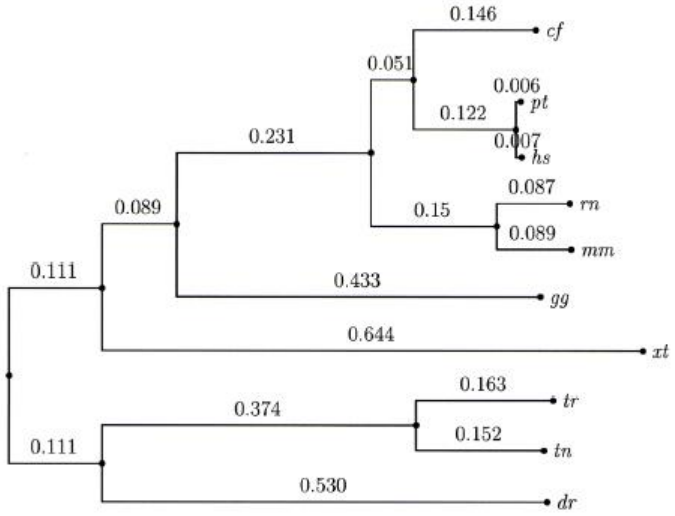


Fig 6.1. Arbre resultant de la mètrica.

En aquest punt volem remarcar que la figura 6.1 és un un punt sobre una varietat algebraica.

Pas 7: Concluint el resultat. Ara estem donant un punt específic en una varietat que representa el model de Jukes-Cantor. Recordem que aquesta varietat esta en un espai projectiu de dimensió $4^{10} - 1$.

Nosaltres estem interesats en 4 coordenades específiques del punt que representa l'arbre, concretament.

$$p_{\texttt{AAAAAAAAAA}} = p_{\texttt{CCCCCCCCCC}} = p_{\texttt{GGGGGGGGGG}} = p_{\texttt{TTTTTTTTTT}}.$$

Aquest punt és el d'interès ja que són els que ens mostren quina és la probabilitat que totes les bases siguin iguals.

Tal com s'ha explicant, aquesta expressió és un polinomi multilineal. Quan l'evaluem en els parametres que es dedueixen de la longitud de les branques, trobem que

$$p_{\texttt{AAAAAAAAAA}} = 0.009651...$$

Tornant a la demostració de la conjectura, tenim que això implica que

6.1.2 Proposició. *Suposant que la distribució donada pel model de J-C sobre la figura A2.1, la probabilitat d'observar una seqüència de longitud 42 intacta en una determinada localització en els deu genomes dels vertebrats estudiats dintre d'un entorn neutral és igual a* $PX = (0.038694)^{40} = 4.3 \cdot 10^{-60}$.

Aquest càlcul no té en compte el fet que la seqüència pot ocorre en una posició arbitraria del genoma en qüestió. Per ajustar això podem multiplicar el nombre PX per la longitud de la cadena dels genomes. El genoma humà conté aproximadament 2.8 mil millions de nucleotids, per tant és raonable pensar per acabar que la probabilitat d'observar una seqüència de longitud 42 intacta en algun lloc en els 10 vertebrats és aproximadament

$$2.8 \cdot 10^9 \times 4.3 \cdot 10^{-60} \approx 10^{-50}.$$

Aquesta probabilitat és un nombre molt petit per tractar-se només d'una casualitat. □

6.2 Inferint arbre filogenètics

En l'exemple anterior hem tractat un cas on només hem fet anar les parametritzacions de la varietat. De fet no hem usat els invariants filogenètics. El que farem en aquesta secció és donar un exemple en que si inferim un arbre.

Suposem que estem estudiant un arbre amb quatre espècies. Per tant serà un arbre com els estudiats en seccions anteriors. Siguin s_1, s_2, s_3, s_4, les seqüències d'ADN. Comptem les freqüències de cada 4-ples segons les diferents topologies de l'arbre, les anomenem $\rho^{T_1}_{x_1x_2x_3x_4}$. Repetim el procés per als arbre T_2 i T_3. Un cop obtingut aquests valors de les freqüències els

substituïm en les variables $p_{x_1x_2x_3x_4}$ pels valors de les freqüències $\rho^T_{x_1x_2x_3x_4}$ en cada invariant filogenètic f i per a cada arbre T. Anomenem aquest valor s^T_f. Per a facilitar els càlculs és molt útil passar els valors a les coordenades de Fourier, ja que en aquest cas l'avaluació del polinomi ja que es tracta d'un binomi. A partir d'aquests valors $\{s^T_f\}_f$, donem una puntuació a cada arbre

$$s(T) = \sum_f |s^T_f|.$$

L'algoritme escollirà aleshores la topologia d'arbre que té menor puntuació.

Bibliografia

[Cas07] Casanellas, M.; Fernández-Sánchez, J.: Geometry of the Kimura 3-parameter model. arXiv:math/AG/0702834v1.

[Cas06] Casanelles, M.: Models algebraics en filogenètica. *But. Cat. Sci* 21 (2006), no 2, 213-228.

[DLu05] De Luna, E.; Guerrero, J.A.; Chew-Taracena, T.: Sistemática biológica: avances y direcciones en la teoría y en los métodos de la reconstrucción fologénetica. Artículo de Revisión. *Hidrobiológica* 15 (2005), 341-370.

[Eri04] Eriksson, N.; Ranestad, K.; Sturmfels, B.; Sullivant, S.: Phylogenetic Algebraic Geometry. arXiv:math.AG/0407033.

[Fro98] Fröberg, R.: *An Introduction to Gröbner bases*. Pure and Applied Mathematics. Series of text Monographs, and tracts. Wiley Intersciense. ISBN: 0-471-97442-0.

[Pat07] Patcher, L.; Sturmfels, B.: The Mathematics of Phylogenomics. *SIAM*, Vol.49, No.1,pp.3-31.

[Pat05] Patcher, L.; Sturmfels, B. eds: *Algebraic Statistics for Computational Biology*. Cambridge University

Press, New York, 2005. xii+420 pp. ISBN: 978-0-521-85700-0.

[Shu99] Shu, D-G.; Luo, H-L.; Conway Morris, S.; Zhang X-L.; Hu, S-X.; Chen, L.; Han, J.; Zhu, M.; Li, Y.; Chen, L-Z.: Lower Cambrian vertebrates from south China. *Nature* 402 (1999), 42-46.

[Sin03] Singular. Software per a realitzar calculs especialitzat en àlgebra commutativa, geometria algebraica i teoria de singularitats.

http://www.singular.uni-kl.de/index.html.

[Ste95] Stell, M.; Fu, Y.: Classifying and Counting Linear Phylogenetic Invariants for the Jukes-Cantor Model. *Journal of Computational Biology* 2 (1995), 39-47.

[Stu05] Sturmfels, B.; Sullivant, S.: Toric Ideal of Phylogenetic Invariants. *Journal of Computational Biology* 12 (2005), 204-228.

[Eva93] Envans, S.; Speed, T.: Invariants of some probability models used in phylogenetic inference. *The Annals of Statistics* 21 (1993), 355-377.

[Wat53] Watson, J.; Crick, F.: A structure for Deoxydribose Nucleic Acid. *Nature* 171 (1953), 964-967.