Premi Albert Dou 2012

El premi Albert Dou fou instituït per la Societat Catalana de Matemàtiques l'any 2010, i té periodicitat biennal. S'ofereix a l'autor d'un treball que contribueixi a fer visible la importància de la matemàtica en el nostre món, a transmetre el coneixement matemàtic a un públic més ampli que els mateixos especialistes, i a promoure tot el que pugui ajudar a l'extensió del prestigi de la matemàtica a la nostra societat. La dotació del premi és de dos mil cinc-cents euros.

En la passada convocatòria de 2012 va ser guardonat el professor Arturo Valdívia pel seu treball titulat "Matemática financiera en tiempos de crisis". Per als lectors interessats en aquest tema, aquest treball es publicarà properament al Butlletí de la Societat Catalana de Matemàtiques.

Societat Catalana de Matemàtiques

Convocatòries de 2013

La Societat Catalana de Matemàtiques ha convocat enguany una edició més, dins el cartell de premis de l'IEC, del premi

Premi Évariste Galois

Instituït l'any 1962 i convocat per cinquantaunena vegada, s'ofereix a un treball d'investigació, bibliogràfic o d'assaig, sobre matemàtiques. Està dotat amb mil euros (1.000 €).
Termini per a la presentació de candidatures: 2 de desembre de 2013 a les 13 h.

Més informació: `http://scm.iec.cat`

Guardonats en la convocatòria 2013

- El Premi Évariste Galois ha estat atorgat a Esther Ibáñez Marcelo pel treball "Expectation maximization per a arbres filogenètics".

Aquest premi fou lliurat el 22 d'abril passat en l'acte de lliurament de premis i borses d'estudi de l'Institut d'Estudis Catalans.

Ressenyes d'obres guardonades

Esther Ibáñez Marcelo, «Expectation maximization per a arbres filogenètics». Premi Évariste Galois 2013

La *reconstrucció filogenètica* consisteix a intentar reconstruir el procés evolutiu que ha donat lloc a les diferents espècies actuals. Les relacions ancestrals es representen en un arbre $\mathcal{T}$ anomenat arbre filogenètic. Una bona referència en filogenètica és el llibre *Inferring Phylogenies* de Felsenstein [F04]. Una manera de dur a terme la reconstrucció és modelant l'evolució de les cadenes d'ADN mitjançant processos de Markov, bé a temps continu o bé a temps discret. En els models a temps continu la matriu $S(t)$ de substitució de nucleòtids en un temps t ve donada per $S(t) = e^{Qt}$, on Q és la matriu de raons de mutació instantània (*rate matrix*). En la majoria de models continus se suposa que la matriu Q és constant en tot el procés evolutiu al llarg de l'arbre (procés homogeni).

Els models a temps discret equivalen a cadenes de Markov ocultes (*HMM*) sobre arbres filo-

genètics i són en certa manera més generals que els de temps continu. Aquests models permeten diferents matrius de transició (*rate matrix*) en diferents llinatges, o dades altrament anomenades *no homogènies*. Seguint un procés de Markov, associem una matriu θ^e a cada branca e. Les entrades de θ^e són les probabilitats $P(x|y,e)$ que un nucleòtid y en el node pare sigui substituït per un nucleòtid x en el node fill al llarg del procés evolutiu representat per la branca e. La longitud de la branca e representa la distància evolutiva entre espècies.

Aquestes probabilitats són per a nosaltres desconegudes, són els paràmetres del model. Segons l'estructura que presentin aquestes matrius obtenim diferents models d'evolució (model estadístic).

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

Segons les restriccions que es posen a les matrius de transició tenim diferents models evolutius (per exemple, GMM és el model més general, i Kimura 3-paràmetres és un dels més usats).

Hi ha estudis que avalen que moltes de les dades biològiques usades en filogenètica no satisfan la hipòtesi d'homogeneïtat. És necessari doncs, un mètode de reconstrucció filogenètica (tant de reconstrucció de topologia (possibles relacions ancestrals entre espècies, (vegeu fig. 1) com de longitud de branca o en general de tots el paràmetres que intervenen en el model) que consideri models de Markov a temps discret. Per als models en temps continu els mètodes més usats són màxima versemblança (ML) [F22] i *neighbor-joining* (NJ) [SN87].

En aquest treball hem desenvolupat el mètode d'Expectation Maximization (EM) [DLR77] per estimar el màxim de versemblança en arbres filogenètics sobre models a temps discret. Aquest algorisme en cada iteració estima els estats ocults i calcula els estimadors de màxima versemblança sota el model plenament observat, d'aquesta forma la versemblança augmenta a cada pas.

Per fer-ho s'ha demostrat quins són els estimadors de màxima versemblança per als paràmetres d'un procés de Markov a temps discret plenament observat en arbres de n fulles, tant per a un model de Kimura 3-paràmetres (K3) com per a un model general de Markov (GMM). Usant aquests estimadors implementem l'algorisme EM per a un model de Markov ocult sobre arbres filogenètics de 4 fulles amb els models esmentats.

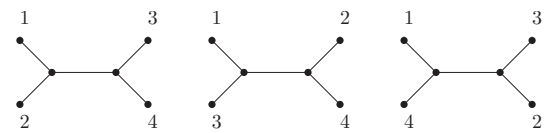


Figura 1: Les tres possibles topologies per a un arbre de 4 fulles sense arrel. $1, 2, 3, 4$ representen espècies actuals, mentre que els nodes interiors són espècies ancestrals.

Hem fet estudis de precisió del mètode per a 4 fulles (vegeu fig. 1), estudis de l'èxit en reconstrucció de topologies i comparació amb altres mètodes de reconstrucció filogenètica (*neighbor-joining*, ML i un mètode basat en invariants).

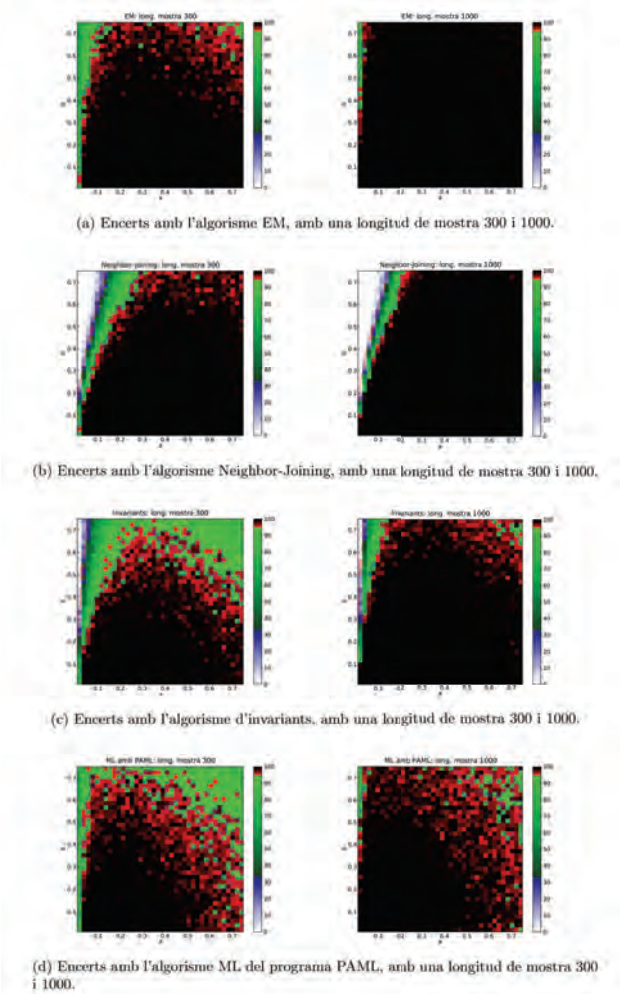
Els resultats d'aplicar l'algorisme *expectation maximization* desenvolupat han estat testats amb diferents mètodes: intervals de confiança i test χ^2 i, d'aquesta forma se n'ha comprovat la capacitat d'estimació.

Per a l'algorisme EM i tant amb un model Kimura 3-paràmetres com amb un model general de Markov, s'ha estudiat l'evolució de l'error en la longitud de branca segons la longitud de la mostra. Per exemple, sota un model Kimura 3-paràmetres, per a longitud de mostra 1000 hi ha més d'un 80% de paràmetres estimats amb un error menor de $0,03$ i si prenem alineaments de longitud 6000 obtenim un 100% de paràmetres estimats amb un error menor de $0,001$.

S'ha programat en Python l'algorisme EM per a arbres de 4 fulles, tant sota el model K3 com el model GMM. També hem comparat l'algorisme EM, amb l'algorisme d'estimació de màxima versemblança que es fa servir en el programari PAML per a models a temps continu amb raó de mutació constant al llarg de l'arbre, amb l'algorisme de *neighbor-joining* i un algorisme basat en invariants filogenètics [CF07].

Sota un espai d'arbres de 4 fulles sense arrel fixat, (en aquest cas hem fet servir l'espai proposat per Huelsenbeck [H95]), s'han fet una sèrie de simulacions, generant dades so-

ta models a temps discret i aplicant els diferents mètodes de reconstrucció filogenètica per avaluar-ne l'eficiència. Amb les simulacions i tests de reconstrucció de topologia s'ha pogut veure la dependència de l'eficiència del mètode segons les longituds de branca sota una espai d'arbres fixat.



(a) Encerts amb l'algorisme EM, amb una longitud de mostra 300 i 1000.

(b) Encerts amb l'algorisme Neighbor-Joining, amb una longitud de mostra 300 i 1000.

(c) Encerts amb l'algorisme d'invariants, amb una longitud de mostra 300 i 1000.

(d) Encerts amb l'algorisme ML del programa PAML, amb una longitud de mostra 300 i 1000.

Figura 2: Resum resultats: comparació amb diversos algorismes. Negre: 100% encerts, blanc: 0.

A la gràfica de la fig. 2 es mostra l'eficiència de l'algorisme EM i els altres mètodes esmentats per a la reconstrucció de la topologia d'arbres de 4 fulles. Cada punt de cada figura es correspon a 100 realitzacions de reconstrucció filogenètica, i el nombre d'encerts amb la topologia real de les dades es comptabilitza segons el color: negre 100% encerts, blanc 0%, del negre al vermell es troben punts que tenen com a mínim un 95% d'encerts, la gamma de verds conté del 95% al 33% (que equivaldria a triar una topologia a l'atzar) i els blaus fins al blanc seria del 33% al 0% d'encerts. Els paràmetres a i b que es troben als eixos de cada figura corresponen a longitud de les branques de l'espai d'arbres on estem treballant.

D'aquesta forma es veu que el millor algorisme per a la reconstrucció de topologia és l'*expectation maximization* i el que presenta pitjors resultats seria el *neighbor-joining*. Cal dir però, que tant el *neighbor-joining* amb distància Kimura 3-paràmetres com l'ML estan basats en models continus amb raons de mutació constant en tot l'arbre, i en canvi, els hem avaluat sobre dades no homogènies. També volem remarcar que el mètode basat en invariants és el que més millora respecte a l'augment de la longitud de la mostra.

Els principals problemes a l'hora de diferenciar topologies es troben en la zona coneguda com la zona de Felsenstein, on la branca interior i dues exteriors oposades presenten una longitud molt petita en comparació a les altres dues.

Finalment volem dir que el mètode EM és molt precís, tot i que més costós, per tant podria ser un bon mètode per a usar dins d'altres mètodes que no necessitin recórrer tot l'espai d'arbres, per exemple, mètodes basats en quartets.

Referències

[CF07] M. Casanellas; J. Fernández-Sánchez. "Performance of a new invariants method on homogeneous and non-homogeneous quartet trees". *Molecular Biology and Evolution* **24** (2007), n. 1, 288–293.

[DLR77] A.P. Dempster; N.M. Laird; D.B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society.* Series B (Methodological) **39** (1977), n. 1, 1–38.

[F04] J. Felsenstein. *Inferring Phylogenies*, Sinauer Associates, Inc., 2004.

[F22] R.A. Fisher. "On the mathematical foundations of theoretical statistics". *Phil. Trans.* A, **222** (1922), 309–368.

[H95] J.P. Huelsenbeck. "Performance of phylogenetic methods in simulation". *Systematic Biology* **44** (1995), 17–48.

[SN87] N. Saitou; M. Nei. "The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees". *Mol. Biol. Evol.* **4** (1987), n. 4, 406–425.

Borses Ferran Sunyer i Balaguer

Convocatòries de 2013

fundació FERRAN SUNYER I BALAGUER
Institvt d'Estvdis Catalans

Premi Ferran Sunyer i Balaguer de Matemàtiques

- Ofert a una monografia escrita en anglès que exposi els resultats més destacats d'una de les àrees de les matemàtiques en la qual s'hagin produït avenços recentment. L'obra ha de tenir un mínim de cent cinquanta fulls, no pot estar subjecta a *copyright* i no ha d'haver estat sotmesa a cap empresa editorial per ser publicada.
- La dotació del premi és de 15.000 €, i l'obra guanyadora serà publicada en la col·lecció «Progress in Mathematics», de l'editorial Birkhäuser Verlag.

  Termini per a la presentació de candidatures: 2 de desembre de 2013 a les 13 h.

Premi Matemàtiques i Societat

- Ofert a autors de reportatges o activitats en qualsevol llengua, de caràcter generalista, sobre qualsevol aspecte de les matemàtiques (ensenyament, recerca, divulgació, presència en la societat), produïts als Països Catalans en els dotze mesos anteriors a la data de resolució.

  Termini d'admissió de candidatures: 28 de febrer de 2014 a les 13 hores.

Borses Ferran Sunyer i Balaguer

- Ofertes als millors projectes d'estudi o de recerca matemàtica relacionats amb la tesi doctoral. Els sol·licitants han de ser estudiants de doctorat de matemàtiques d'una universitat dels Països Catalans, en el tram final de la tesi doctoral.
- L'objectiu d'aquestes borses és reforçar la formació en recerca dels estudiants premiats mitjançant l'estada d'entre un i tres mesos d'estudi o de recerca en una institució fora de l'àmbit geogràfic de la universitat d'origen.

Les sol·licituds s'han de trametre abans del dia 28 de febrer de 2014 a les 14 hores. La resolució de la convocatòria es farà durant la segona quinzena de març de 2014.

Més informació: `http://ffsb.iec.cat`