Détection de sélection darwinienne sur un gène par une approche sans vraisemblance

Aude Grelaud, Christian P. Robert, François Rodolphe

Détection de sélection darwinienne sur un gène par une approche sans vraisemblance

Aude GRELAUD, Christian P. ROBERT & François RODOLPHE

Department of Statistics and Biostatistics, Busch Campus, Rutgers University, 110 Frelinghuysen road, Piscataway, NJ 08854, USA
agrelaud@rutgers.stat.edu

CEREMADE, Université Paris Dauphine, Place Maréchal Delattre de Tassigny, 75016 Paris Cedex
et CREST-LS, Avenue P. Larousse, Malakoff, France
xian@ceremade.dauphine.fr

INRA, unité MIG, Domaine du Vilvert, 78350 Jouy en Josas, France
francois.rodolphe@jouy.inra.fr

Si l'on considère un vecteur d'observations $\mathbf{x}^0 = (x_1, \cdots, x_n)$ associé à un modèle statistique de vraisemblance $f(\cdot|\boldsymbol{\theta})$, où le paramètre $\boldsymbol{\theta}$ a pour distribution a priori $\pi(\boldsymbol{\theta})$, l'objectif en statistique bayésienne est de déterminer la distribution a posteriori $\pi(\boldsymbol{\theta}|\mathbf{x}^0)$. Rubin (1984) faisait remarquer qu'un algorithme de rejet permet parfois de simuler exactement sous cette distribution sans jamais évaluer la vraisemblance.

Cette idée a ensuite été exploitée par Pritchard (1999), mais cette approche reste gourmande en temps de calcul si le taux d'acceptation est faible. Pour remédier à cela, des procédures de simulation non exactes ont été mises en place. Ces algorithmes, aussi connus sous le nom d'algorithmes ABC (Approximate Bayesian Computation), sont particulièrement intéressants quand le modèle considéré est complexe (Beaumont et al., 2002), mais sont encore souvent associés à un temps de calcul important.

Il a été démontré que combiner un algorithme ABC avec des techniques de Monte-Carlo, Monte-Carlo séquentiel (Del Moral et al., 2009) ou Population Monte-Carlo (Beaumont et al., 2009), permet de réduire le temps de calcul.

En génétique des populations, les modèles sont habituellement complexes ce qui rend le calcul de vraisemblance difficile. Cependant, des mécanismes de simulation d'observations existent, ce qui a rendu les méthodes sans vraisemblance populaires dans ce domaine.

Nous nous intéressons ici aux effets de la sélection darwinienne sur un gène. Celle-ci peut favoriser, ou non, l'apparition de nouvelles séquences. Les données utilisées sont des séquences homologues, c'est-à-dire des séquences qui codent pour un même gène chez des espèces proches. La généalogie reliant celles-ci induit une dépendance représentée par un arbre phylogénétique. Celui-ci est inconnu et doit donc être traité comme un paramètre du modèle.

Dans la littérature, les méthodes de détection de sélection darwinienne reposent sur l'estimation de la séquence de pressions de sélections à partir d'un alignement multiple

de ces séquences (Yang et al., 2000 ; Wong et al., 2005 ; Huelsenbeck et al., 2006). Cet indicateur peut être vu comme le ratio du taux de mutations non-synonymes divisé par le taux de mutations synonymes. Dans notre approche, nous considérons directement ces différents taux de mutations, le premier variant d'un site à l'autre alors que le second est constant sur la séquence.

Quand nous évaluons les effets de la sélection darwinienne sur un gène, notre travail est d'estimer certains paramètres d'un modèle d'évolution de séquences,

$$(\theta_{NS,1}, \cdots, \theta_{NS,L}, \theta_S) \in \boldsymbol{\theta} = (\theta_{NS,1}, \cdots, \theta_{NS,L}, \theta_S, \beta),$$

plus exactement de déterminer leur distribution a posteriori, les autres, notés β étant considérés comme des paramètres de nuisance.

L'arbre phylogénétique fait partie de ceux-ci. Il est représenté ici par un arbre de coalescence; nous pouvons en effet montrer que nous pouvons nous placer dans le cadre du modèle de Moran (1958). Le premier avantage de ce choix est que simuler sous un modèle de coalescence est relativement simple (Kingman, 1982; Hein et al., 2005) et nous proposons ici un mécanisme de simulation pour notre modèle d'évolution de séquence. De plus, dans ce cas l'unité de temps est donnée par le jeu de données et non plus fixée arbitrairement. Ceci nous permet de comparer les résultats d'un gène à l'autre.

Sous ce modèle, évaluer la vraisemblance est possible, mais avec un temps de calcul important. Comme un méchanisme de génération des données est disponible, utiliser une approche sans vraisemblance parait particulièrement intéressant. Notre procédure d'estimation repose sur l'algorithme ABC-SMC (Del Moral et al., 2009).

Nous évaluons les performances de notre approche par une étude de simulation puis nous présentons les résultats obtenus pour le gène HIV-1 *env*.

Mots clés : Méthodes bayesiennes, biostatistique.

Some recent methods based on Bayesian simulations have provided ways of evaluating approximately posterior distributions without computing likelihood functions.

Given a dataset $\mathbf{x}^0 = (x_1, \cdots, x_n)$ associated with the sampling distribution $f(\cdot|\boldsymbol{\theta})$, and under a prior distribution $\pi(\boldsymbol{\theta})$ on the parameter $\boldsymbol{\theta}$, the aim is to sample from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}^0)$. Rubin (1984) noticed that using rejection techniques sometimes enable to sample exactly from the target distribution. This idea was then exploited by Pritchard (1999), but this appproach can be time consuming if the acceptance rate is too low. As an alternative, some procedures that sample approximatively from the posterior distribution have been developed.

These likelihood free algorithms, also known as ABC algorithms (Approximate Bayesian Computation), are of particular interest when considering complex models (Beaumont et al., 2002) but are still time consuming.

It has been shown that combining ABC with Monte-Carlo techniques such that Sequential Monte Carlo algorithm (Del Moral et al., 2009) or Population Monte Carlo (Beaumont et al., 2009) can reduce the computational time.

In Population genetics, models are usually complex and evaluating the likelihood function can be computationnaly prohibitive. Meanwhile, data can be simulated and rendered likelihood free approaches popular in this area. Our interest here is on he effects of the darwinian selection on a gene. It can favor, or not, the apparition of new sequences.

Dataset consists in a multiple alignment of homolog sequences, in other words sequences coding for the same gene in different species. The relationship between these species can be represented by a phylogenetic tree, which is unknown and thus belongs to the parameters of the model.

In the litterature, methods for detecting darwinian selection at the molecular level rely on estimating the sequence of selective pressure on an alignment of protein coding sequences (Yang et al., 2000 ; Wong et al., 2005 ; Huelsenbeck et al., 2006). This indicator can be viewed as the non-synonymous mutations rate, $\theta_{NS,l}$, $l = 1, \cdots, L$, L being the length of the sequence, divided by the synonymous mutations rate θ_S. In our approach, we consider directly both mutation rates, the first one being site specific and the other constant along the sequence.

When evaluating the effects of darwinian selection on a gene, the aim is to estimate a few parameters of a model of sequence evolution,

$$(\theta_{NS,1}, \cdots, \theta_{NS,L}, \theta_S) \in \boldsymbol{\theta} = (\theta_{NS,1}, \cdots, \theta_{NS,L}, \theta_S, \beta),$$

the remaining ones, denoted by β, being considered as nuisance parameters. These include a phylogenetic tree represented as a coalescent tree, which corresponds to the model of Moran (1958). The first advantage is that simulating under a coalescent model is easy and we propose a generating mecanism under our model of sequence evolution (Kingman, 1982; Hein et al., 2005). Then, in a coalescent framework, the time unit is given by the set of sequences, not choosen arbitrary. This allows to compare results from one gene to another.

Under this model, evaluating the likelihood is possible, but computationally intensive. As a generating mecanism is available, using likelihood free inference is thus of particular interest. Our estimation procedure relies on the ABC-SMC algorithm (Del Moral et al., 2009).

We evaluate the performance of the method on simulated data and show the results obtained on the HIV-1 *env* gene.

Key words: Bayesian methods, biostatistics.

Bibliographie

[1] Beaumont, M.A., and Zhang, W., and Balding, D. (2002) Approximate Bayesian Computation in Population Genetics, Genetics.
[1] Beaumont, M.A. and Cornuet, J.-M. and Marin, J.-M. and Robert, C.P. (2009) Adaptive approximate Bayesian computation, Biometrika.
[2] Del Moral, P. and Doucet, A. and Jasra, A. (2009) An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation, Annals of Applied Statistics.
[3] Hein, J. and Schierup, M. and Wiuf, C. (2005) *Gene, Genealogies, Variation and Evolution*, Oxford University press, Oxford.
[4] J. Huelsenbeck and S. Jain and S. Frost and S. Pond (2006) A Dirichlet process model for detecting positive selection in protein-coding DNA sequences, Proceedings of the National Academy of Sciences of the United States of America.
[5] Kingman, J. (1982) The coalescent, Stochastic processes and their Applications.
[6] Moran, P. (1958) Random processes in genetics, Proceedings of the Cambridge Philosophical Society.
[7] Pritchard, J.K. and Seielstad, M.T. and Perez-Lezaun, A. and M. W. Feldman (1999) Population growth of human Y chromosome: a study of Y chromosome microsatellites, Molecular Biology and Evolution.
[8] Rubin (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician, Annals of Statistics.
[9] Z. Yang and R. Nielsen and N. Goldman and A.-M. Krabbe Pedersen (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites, Genetics.
[10] W. Wong and Z. Yang and N. Goldman and R. Nielsen (2005) Bayes empirical Bayes inference of amino acid sites under positive selection, Molecular Biology and Evolution.