Application d'une méthode spectrale utilisant des séries temporelles pour estimer les paramètres de sélection d'un SNP dans un modèle de diffusion de Wright Fisher

Cyriel PARIS, Bertrand SERVIN, Simon BOITARD

GenPhySE, INRA, Toulouse, France

28 avril 2017

Objectifs de la thèse

Un approche récente

Utiliser des données génomiques d'époques différentes pour préciser notre connaissance de phénomènes évolutifs. L'information contenue dans un échantillonnage au cours du temps est (a priori) plus importante que dans un simple échantillonnage au temps présent.

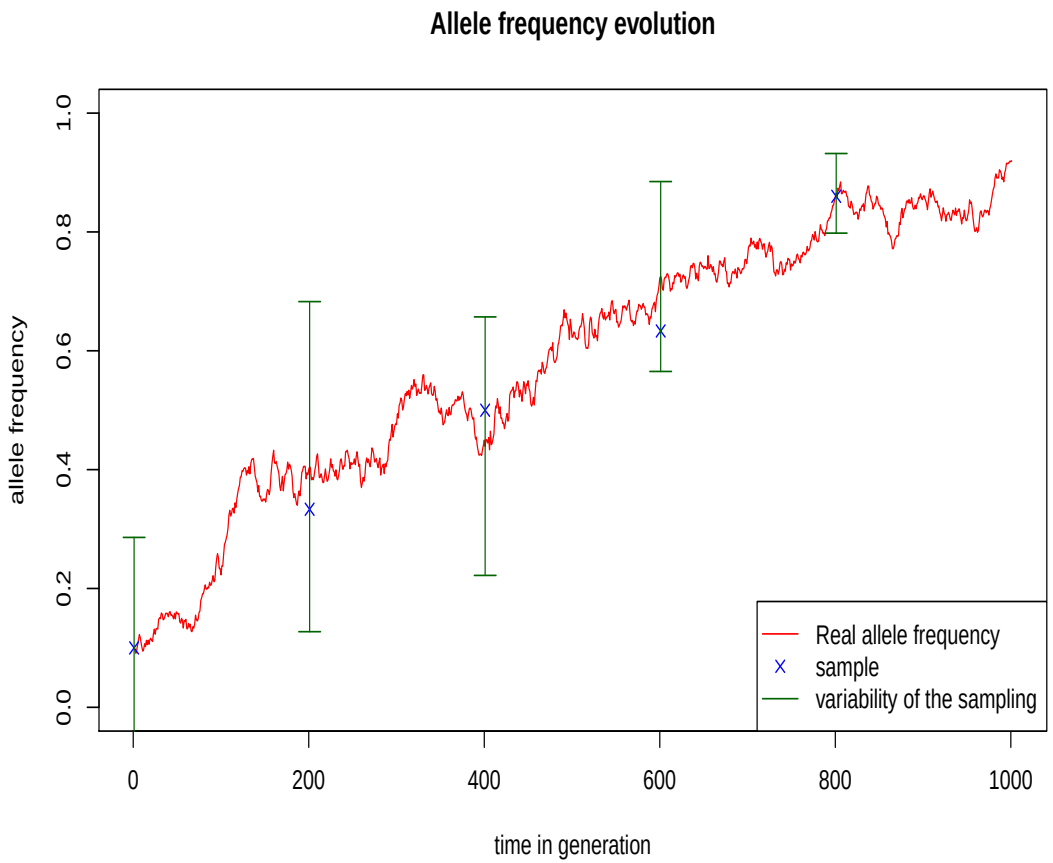
2 questions de recherche

- Histoire démographique des populations (changement de taille...)
- Détection de régions génomiques sous sélection

2 types de données

- Données à court terme (cryobanques)
- Données à long terme (ADN ancien)

Idée des séries temporelles



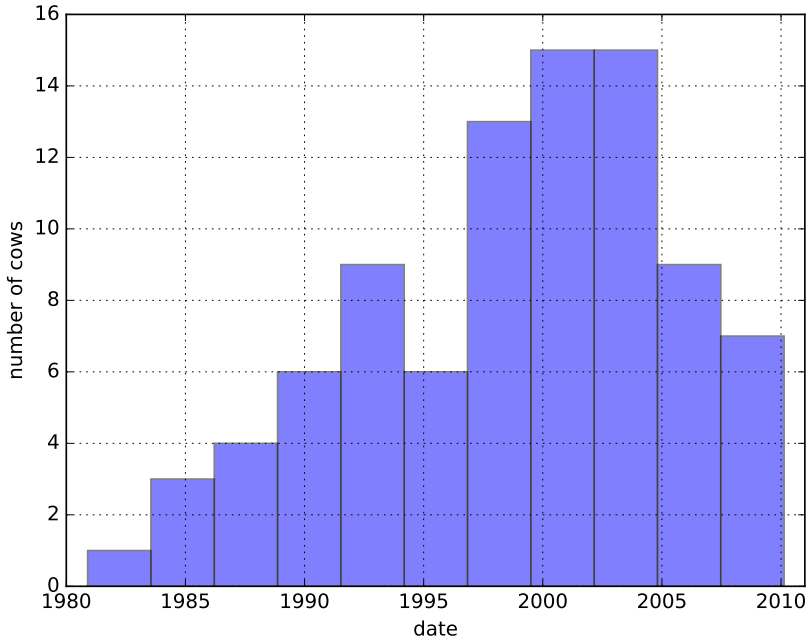
Concept des méthodes s'appuyant sur des séries temporelles

Tenir compte de l'incertitude de l'échantillonnage dans la détermination de la trajectoire. Déterminer alors s'il est probable qu'il y ait eu sélection.

Exemple de données temporelles à court terme

Données à analyser

87 animaux au total (vaches de la race "Asturianas de los valles") pour lesquels $50,000$ SNP ont été génotypés. Données fournies par Susana Dunner (UCM).



Modèle de diffusion

Notations

- X_t : fréquence allélique à l'instant t
- $P(X_t = x | X_0 = p) = f(t, x, p)$

Equation de Kolmogorov Backward

$$\frac{\partial f}{\partial t} = \frac{1}{2} p(1-p) \frac{\partial^2 f}{\partial p^2} + \frac{1}{2}[\alpha(1-p) - \beta p] \frac{\partial f}{\partial p} + 2p(1-p)[\sigma_1(1-2p) + \sigma_2 p] \frac{\partial f}{\partial p}$$

Publications utilisant cette modélisation

- Bollback *et al.* (2008) : résolution numérique
- Song *et al.* (2012) : résolution analytique
- Steinrucken *et al.* (2014) : calcul de vraisemblance

Opérateur différentiel

Définition de l'opérateur L

$$L = \frac{1}{2}p(1-p)\frac{\partial^2}{\partial p^2} + \frac{1}{2}(\alpha(1-p) - \beta p)\frac{\partial}{\partial p} + 2p(1-p)(\sigma_1(1-2p) + \sigma_2 p)\frac{\partial}{\partial p}$$

Si f est une fonction propre pour L de valeur propre λ et si $\frac{\partial f}{\partial t} = \lambda f$, Alors f est solution de l'équation de diffusion.

Existence des vecteurs propres

Définitions

- $\bar{\sigma}(p) = 4\sigma_1 p(1-p) + 2\sigma_2 p^2$
- $\pi(p) = \mathrm{e}^{\bar{\sigma}(p)} p^{\alpha-1}(1-p)^{\beta-1}$
- $\langle f, g \rangle_\pi = \int_0^1 f(x)g(x)\pi(x)\mathrm{d}x$

Propriétés

- L est auto adjoint pour $\langle , \rangle_\pi$ (Song *et al.* 2012)
- Il existe une base hilbertienne de fonctions propres B_n associées aux valeurs propres Λ_n telles que $0 \geq \Lambda_0 > \Lambda_1 > \Lambda_2 > \cdots$ et $\lim_{n\to\infty} \Lambda_n = -\infty$ (Karlin and Taylor 1981)

Détermination de la solution

Une famille orthogonale connue

$K_n(p) = \mathrm{e}^{-\frac{\bar{\sigma}(p)}{2}} R_n^{\alpha,\beta}(p)$, où $(R_n^{\alpha,\beta})_{n\in\mathbb{N}}$ est la famille des polynomes de Jacobi modifiés.

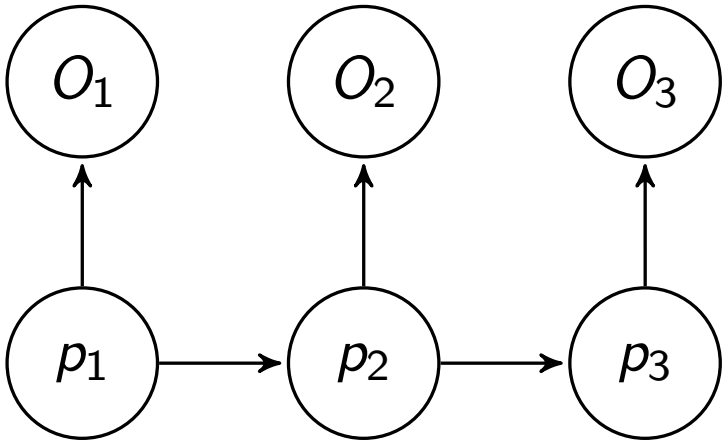
- $(R_n^{\alpha,\beta})_{n\in\mathbb{N}}$ est une famille de polynômes bien étudiée
- K_n est une base hilbertienne donc $B_n(p) = \sum_{m=0}^{\infty} w_{n,m} K_m(p)$
- $w_{n,m}$ s'obtient comme solution de $Mw_n = \Lambda_n w_n$ où M est une matrice connue ne dépendant que des paramètres de l'EDP

Synthèse

En considérant la condition au bord $f(0, x, p) = \delta(x - p)$, la solution s'écrit :

$$f(t, x, p) = \sum_{n=0}^{\infty} \pi(x) \mathrm{e}^{\Lambda_n t} \frac{B_n(p) B_n(x)}{\langle B_n, B_n \rangle_\pi}$$

Modèle de Markov Caché (HMM) - Présentation



Modelisation du HMM

- n_i : taille de l'échantillon i
- émission : $O_i \sim \mathcal{B}(n_i, p_i)$
- transition : $p_\Theta(t_i - t_{i-1}, p_i, p_{i-1})$

Définitions et propriétés (algorithme forward)

- $f_k(x)\mathrm{d}x = \mathbb{P}_\Theta(O_{[1:k]}, X(t_k) \in \mathrm{d}x)$
- $\mathbb{P}_\Theta(O_{1:K}) = \int_0^1 f_K(x)\mathrm{d}x$ (vraisemblance)
- $f_k(x) = \binom{n_k}{O_k} x^{O_k}(1-x)^{n_k - O_k} \int_0^1 f_{k-1}(p) p_\Theta(t_k - t_{k-1}, x, p)\mathrm{d}p$

Modèle de Markov Caché (HMM) - Application

Définition et notation

$$f_k(x) = \pi(x) \sum_{n=0}^{\infty} b_{k,n} B_n(x) \quad \text{et} \quad b_k = (b_{k,0}, b_{k,1}, \dots)$$

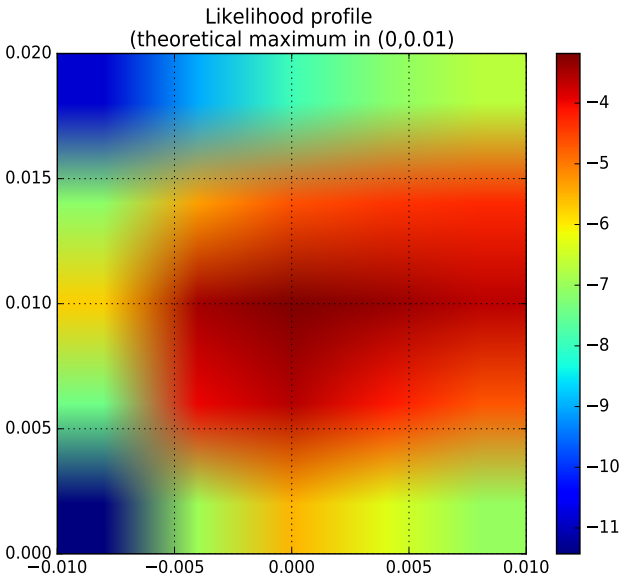
Algorithme

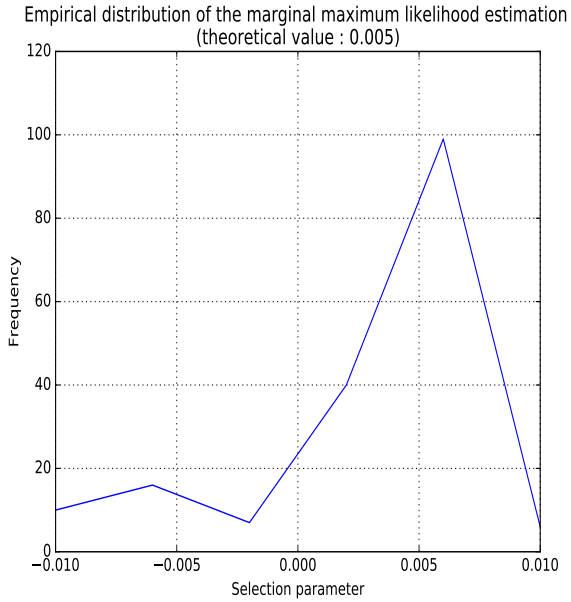
- b_0 est à fixer selon l'état initial du modèle
- $b_k = b_{k-1} \exp(-\Lambda(t_k - t_{k-1})) W G^{O_k} (I - G)^{n_k - O_k} W^{-1}$
- $\mathbb{P}_\Theta(O_{1:K}) = \frac{\langle B_0, B_0 \rangle_\pi}{B_0(0)} b_{K,0}$
- Λ est la matrice diagonale des valeurs propres de l'opérateur
- W est la matrice des vecteurs propres B_n exprimés dans la base K_n
- G est une matrice obtenue à l'aide de propriétés sur les polynômes de Jacobi

Premiers résultats sur des simulations

Chercher le maximum de la vraisemblance

On calcule la vraisemblance sur une grille de paramètres et on cherche les paramètres qui la maximisent.





Distribution de l'estimation

La distribution du maximum de vraisemblance est localisée autour de la vraie valeur

Futur

Sur ce sujet

- Analyser des données réelles
- Comparer avec les méthodes n'utilisant que des échantillons actuels
- Inclure la corrélation entre deux locus dans la modélisation.
- Généraliser cette méthode à d'autres modèles.

Sur l'histoire démographique d'une population

- Trouver (ou concevoir) une modélisation permettant d'utiliser les séries temporelles pour estimer les changements de taille d'une population

Merci de votre attention !