Détection de balayages sélectifs à l'aide de chaînes de Markov cachées.

Simon Boitard, Andreas Futschik, Christian Schlötterer

INRA, LGC, Toulouse
Université de Vienne (Autriche)

Outline

Contexte

- Une majorité de sites neutres, quelques uns sous sélection.
- Balayage sélectif (selective sweep) : un allèle sélectionné passe d'une fréquence très faible à une fréquence de 1 (fixation) dans une population.
- Laisse des traces sur les fréquences alléliques des sites voisins.
- Détection des sites sous sélection au niveau d'une seule population.

Type de données

- n séquences d'ADN de longueur L issues d'une même population.
- **Modèle à infinité de sites** $\longrightarrow$ au plus deux alléles distincts par site
  0 = ancestral, 1 = dérivé.

A-A-C-G-**G**-G-T-A-**T**-C-G- ....

A-A-C-G-**G**-G-T-A-**A**-C-G- ....

A-A-C-G-**C**-G-T-A-**T**-C-G- ....

Type de données

- n séquences d'ADN de longueur L issues d'une même population.
- **Modèle à infinité de sites** $\longrightarrow$ au plus deux alléles distincts par site
  0 = ancestral, 1 = dérivé.

0-0-0-0-**1**-0-0-0-**0**-0-0- ....

0-0-0-0-**1**-0-0-0-**1**-0-0- ....

0-0-0-0-**0**-0-0-0-**0**-0-0- ....

Type de données

- n séquences.

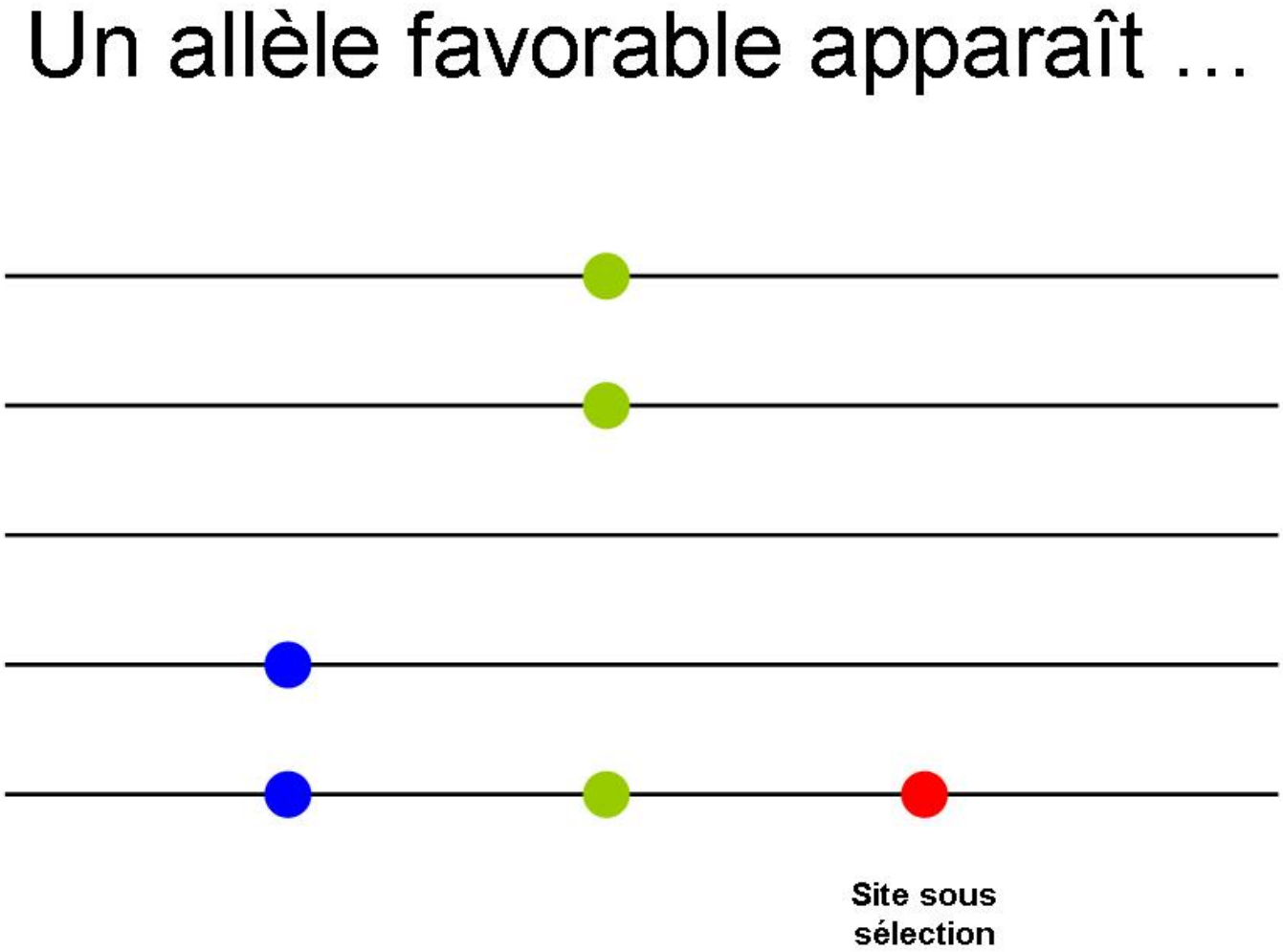
$$0-0-0-0-1-0-0-0-0-0-0-\ldots$$
$$0-0-0-0-1-0-0-0-1-0-0-\ldots$$
$$0-0-0-0-0-0-0-0-0-0-0-\ldots$$

- x_i nombre d'allèles 1 (dérivés) au site i.

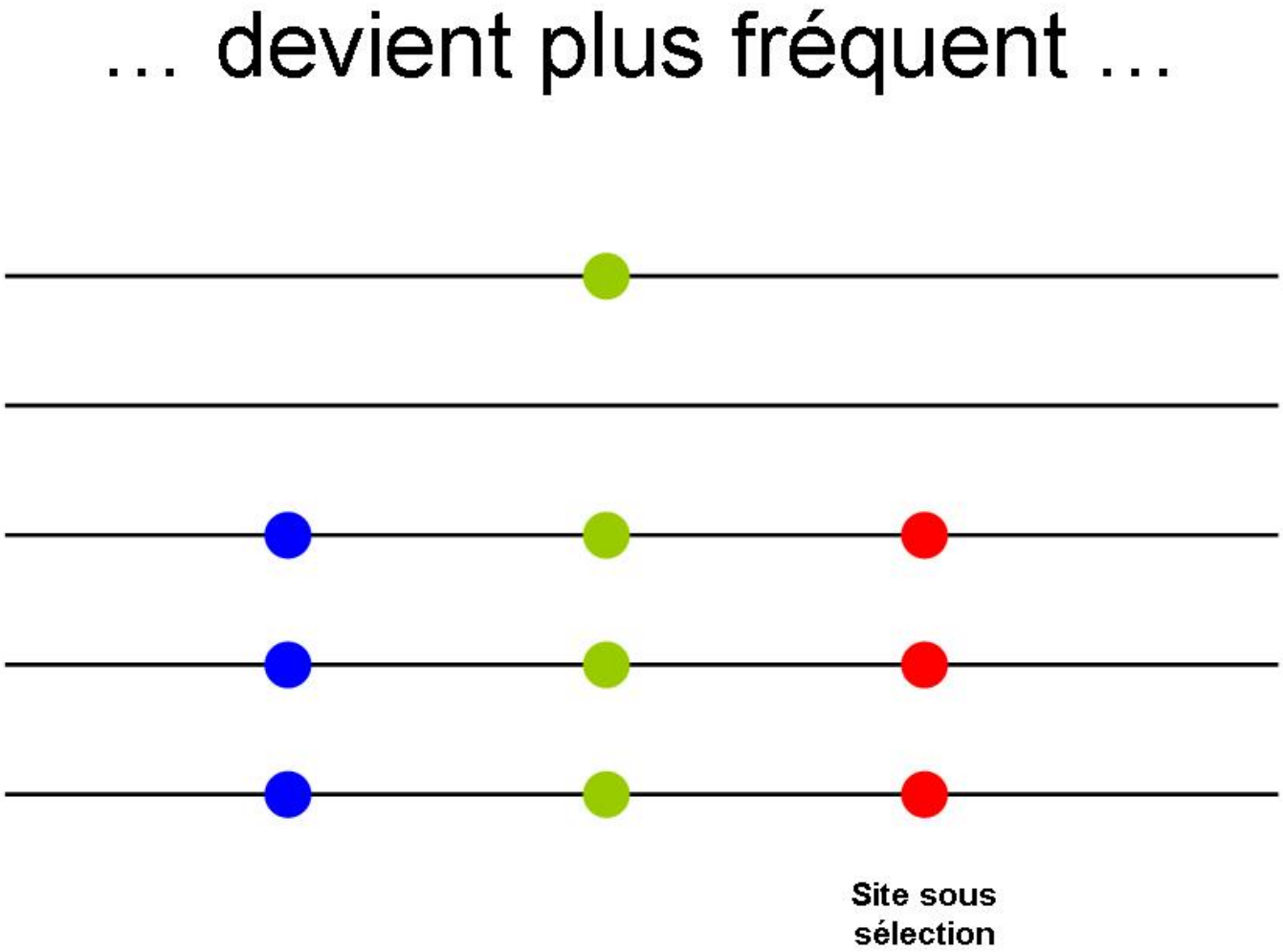
$$0-0-0-0-2-0-0-0-1-0-0-\ldots$$

- Distribution de probabilité de X_i dépend du modèle d'évolution supposé.

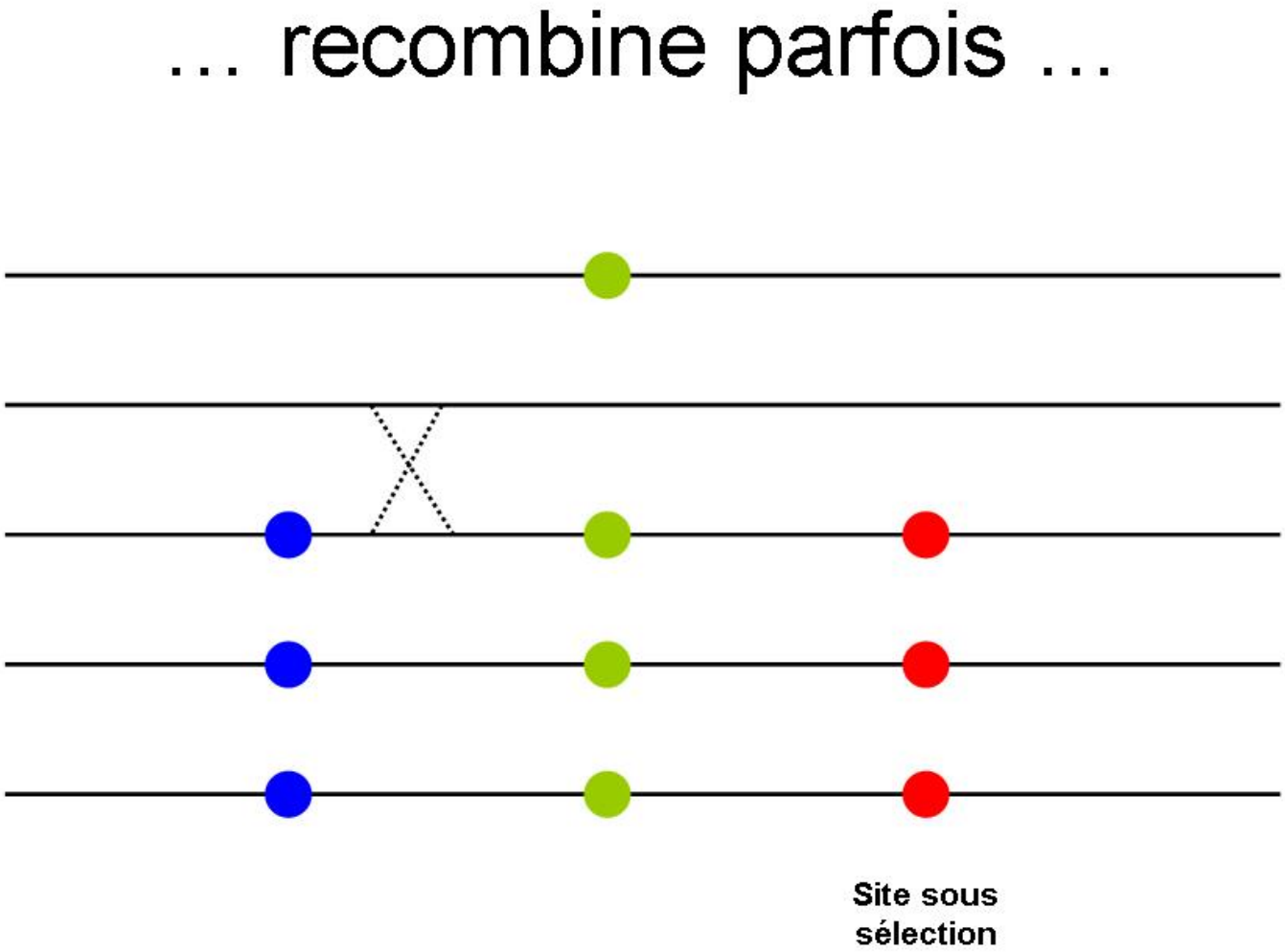
Balayage sélectif



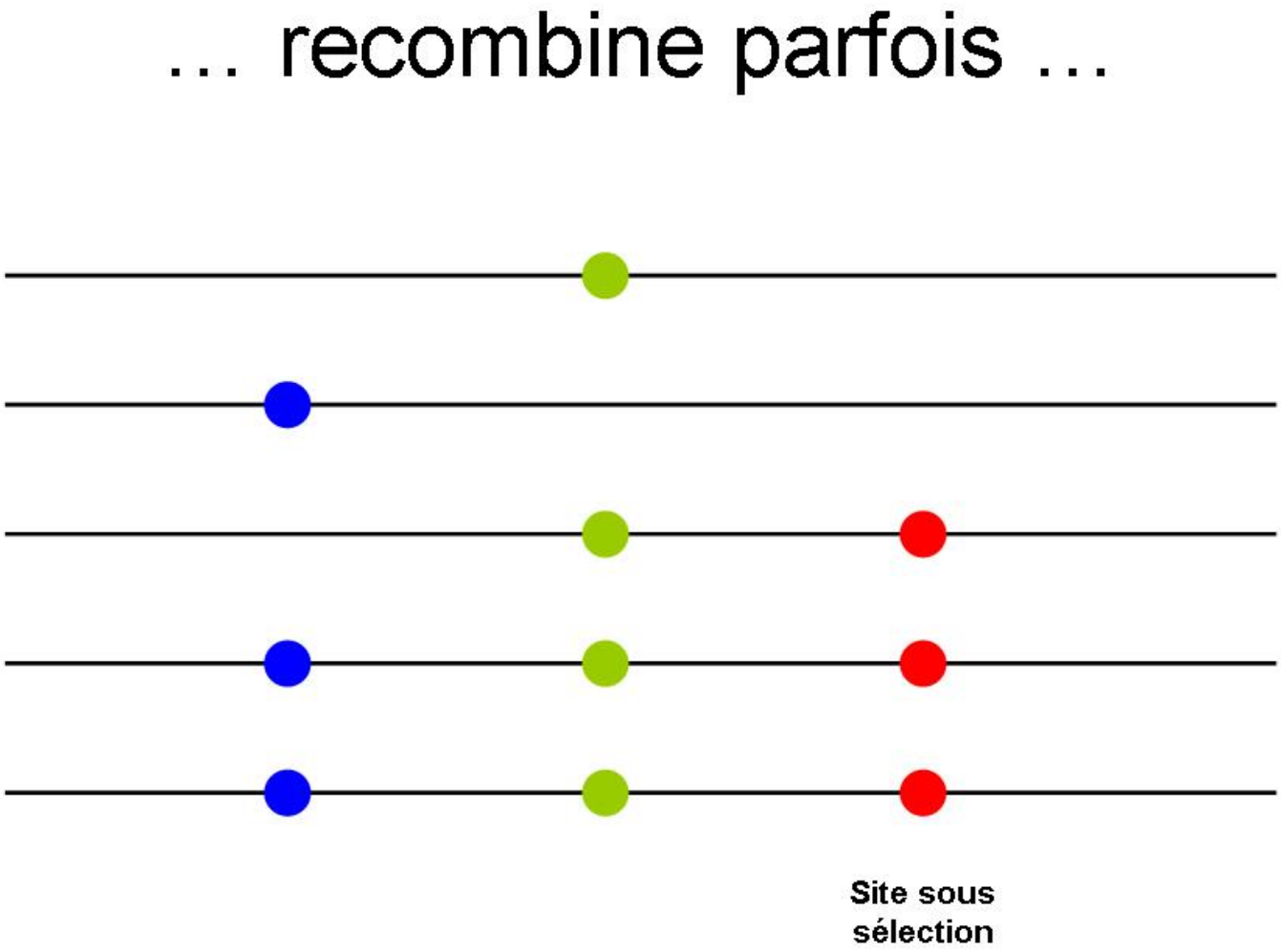
Balayage sélectif



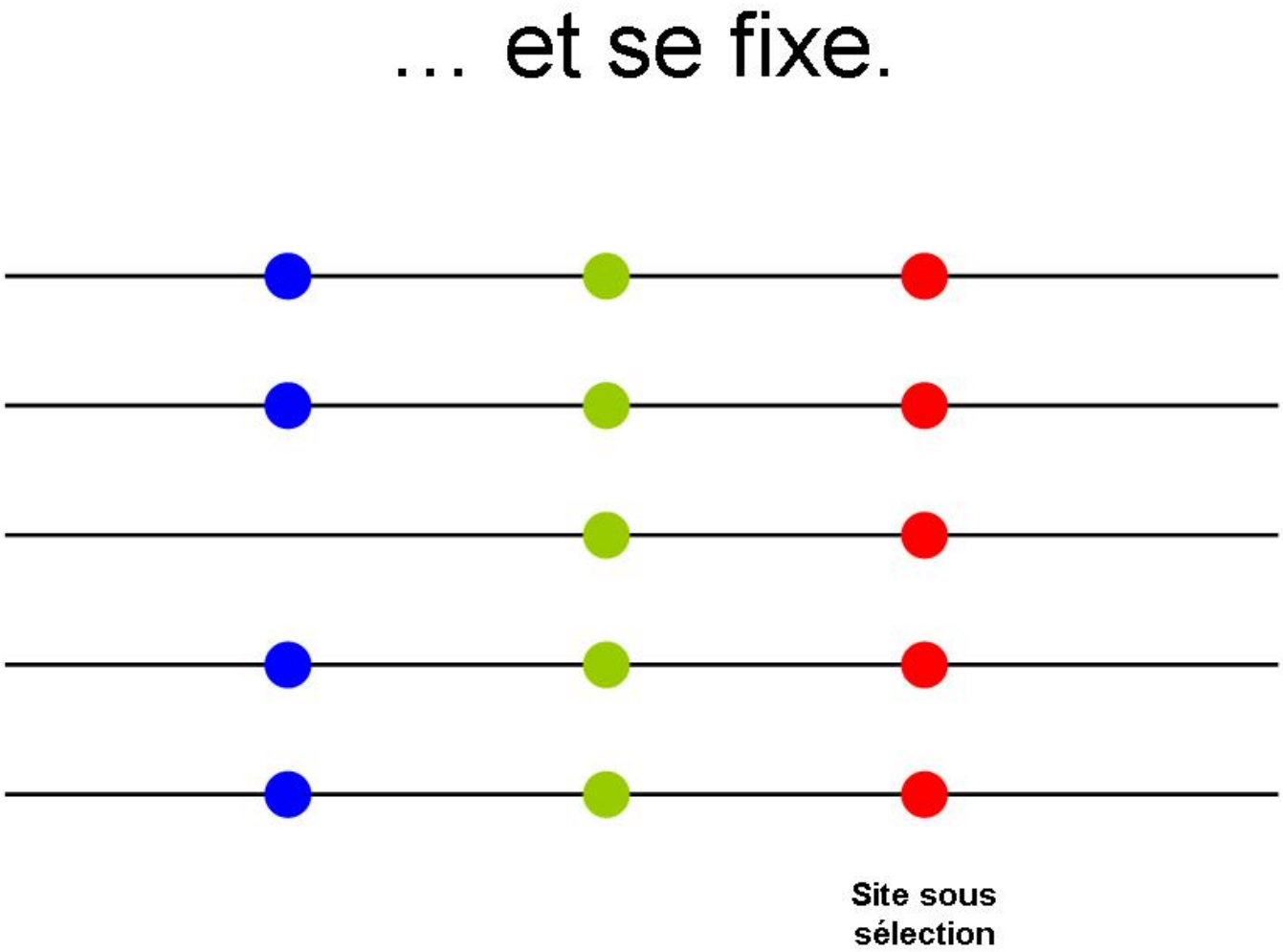
Balayage sélectif



Balayage sélectif



Balayage sélectif



Effet d'un balayage sélectif sur les fréquences alléliques

A proximité d'un site sous sélection :

- Densité des sites polymorphes diminue (Kaplan et al. 1989),
- Probabilité des fréquences extrèmes (hautes et basses) augmente (Braverman et al. (1995), Fay et Wu (2000)),

par rapport à ce que l'on attend sous un modèle neutre.

Loi de probabilité de X_i

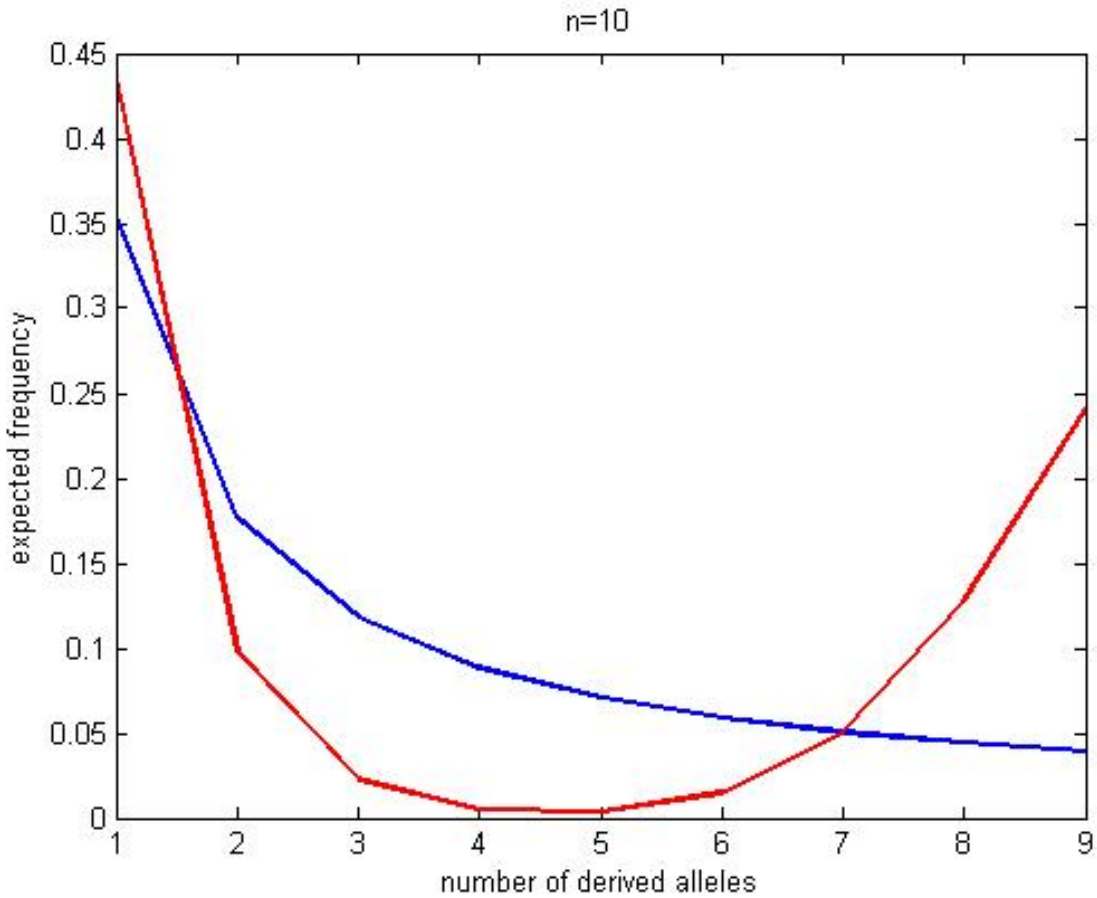


Figure: Loi de probabilité du nombre d'allèles mutés (X_i) à un site neutre (i) loin d'un site sélectionné (courbe bleue) (ii) proche d'un site sélectionné (courbe rouge)

Détection par maximum de vraisemblance

- Vraisemblance composite :

$$\mathcal{L}(j,s) = \mathbb{P}(X_1 = x_1, \ldots, X_L = x_l \mid j, s) \approx \prod_{i=1}^{L} \mathbb{P}(X_i = x_i \mid j, s)$$

  oú $\mathbb{P}(. \mid j, s)$ suppose qu'il y a eu un évènement de sélection d'intensité s au site j. La distance $j - i$ a une influence.
- Maximum de vraisemblance $\hat{j}, \hat{s}$.
- $\mathcal{L}_0 \approx \prod_{i=1}^{L} \mathbb{P}_0(X_i = x_i)$, $\mathbb{P}_0$ sous un modèle neutre.
- Rejet de H_0 : "neutralité" pour H_1 : "sélection" si $\frac{\mathcal{L}(\hat{j},\hat{s})}{\mathcal{L}_0} > \lambda$.
- Kim et Stephan (2002), Nielsen *et al* (2005), Li et Stephan (2006) ...

Méthode de Nielsen *et al* (2005)

- **Hypothèse :** la plupart des sites sont neutres
  $\longrightarrow \mathbb{P}_0$ estimé à partir des données :

$$\hat{\mathbb{P}}_0(X_i = k) = \frac{L_k}{L}, \quad k = 0, \ldots, n-1$$

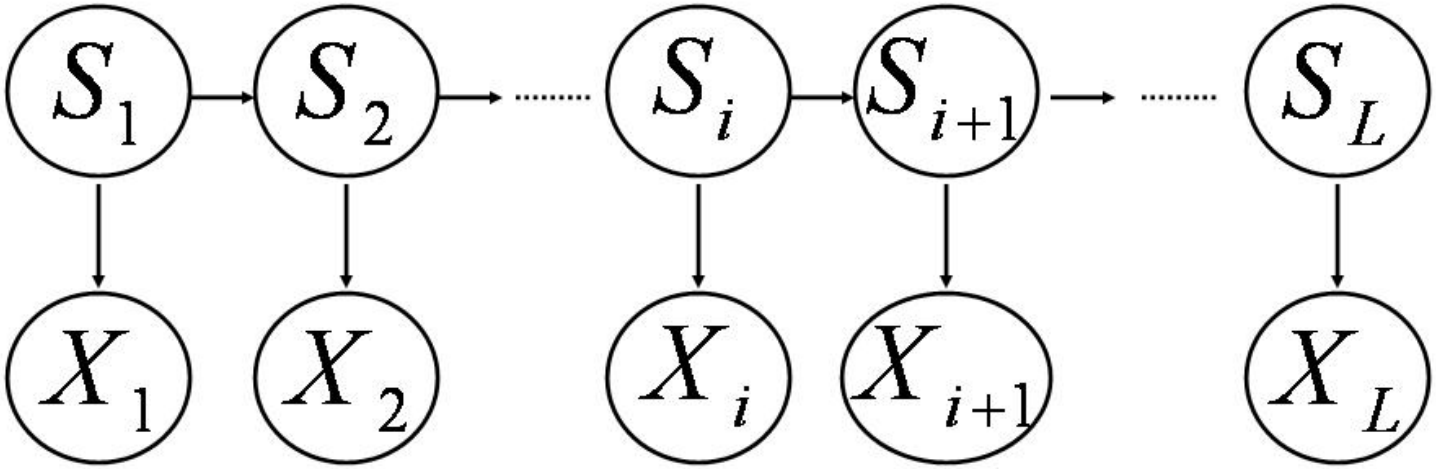
  L_k nombre de sites avec k alleles 1.
- $\mathbb{P}(. \mid j, s)$ déduit de $\mathbb{P}_0$ grâce à un modèle simplifié de coalescent sous sélection. Un paramètre α tient compte des effets combinés de $j - i$ et s.

1 Détection de balayages sélectifs

2 Un modèle de chaîne de Markov cachée
- Méthodes
- Résultats

Modèle

Etats cachés : "Neutre", "Intermédiaire", "Sélection"



Etats observés : nombres d'allèles dérivés

Objectif : prédire la chaîne d'états cachés S sachant la chaîne d'états observés X.

Motivations

- S_i représente la généalogie des n séquences observés au site i.
- La structure Markovienne de S permet de tenir compte de la corrélation entre les généalogies aux différents sites de la séquence.
- La loi de X_i est entièrement déterminée par la généalogie au site i (donc par S_i).

Paramètres

- Probabilités d'émission comme Nielsen *et al* (2005).
  deux valeurs distinctes de α pour les états "Intermédiaire" et "Sélection".
- Matrice de transition :

$$\begin{pmatrix} 1-p & p & 0 \\ p/2 & 1-p & p/2 \\ 0 & p & 1-p \end{pmatrix}$$

  p est utilisé pour contrôler le taux d'erreur sous H_0 (neutralité).
- Prédiction de S par l'algorithme de Viterbi.

Exemple

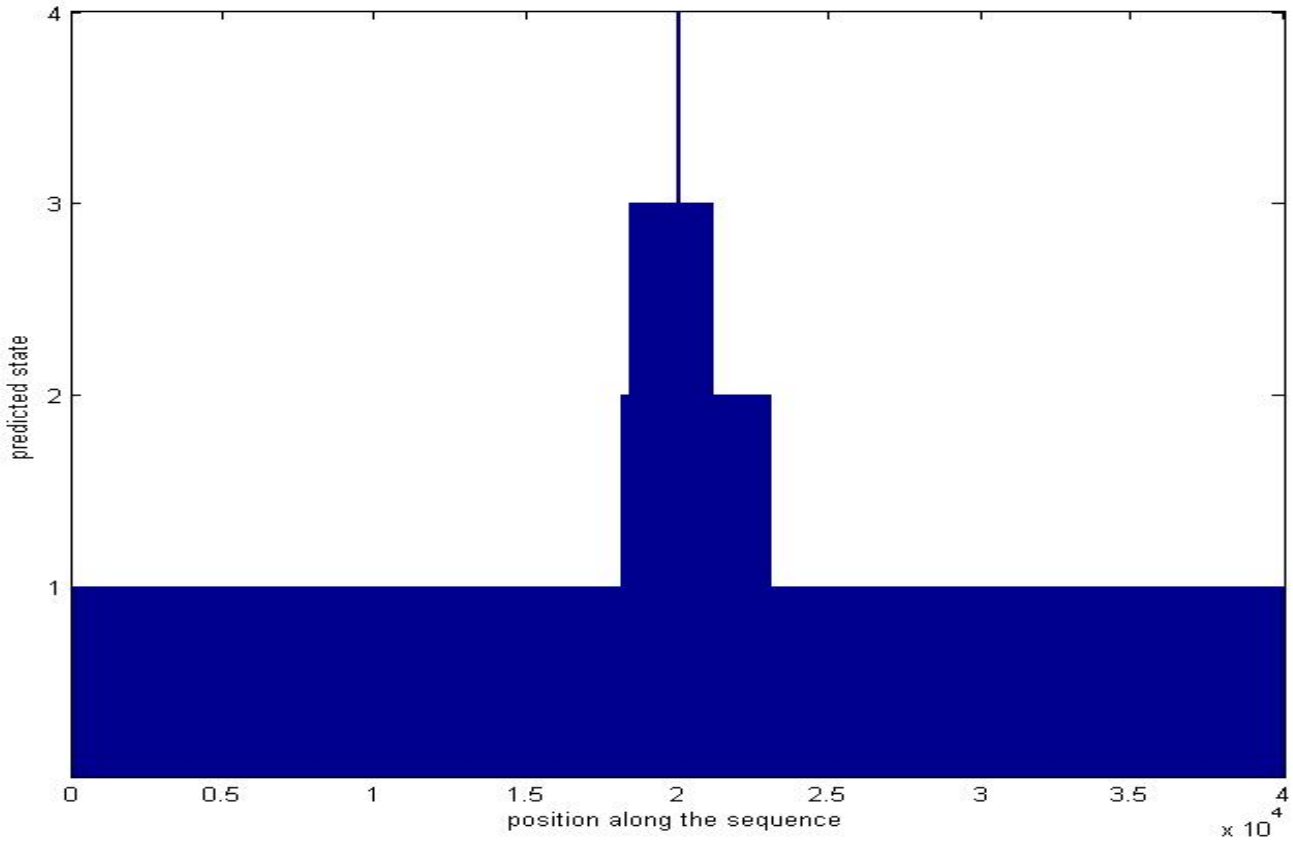


Figure: Etats cachés prédits le long de la séquence pour un échantillon simulé sous un scénario de balayage sélectif.. 1 = "Neutre", 2 = "Intermédiaire", 3 = "Sélection", 4 = vrai site sous sélection.

Puissance de détection

Echantillons de taille $n = 30$ et de longueur $L = 100kb$.
Mutation $\theta = 4N\mu = 0.005$, recombinaison $\rho = 4Nr = 0.02$.
Balayage sélectif récent ($\tau/N = 0.002$) d'intensité $\alpha = 2Ns = 300$.
5% d'erreur de type I.

	HMM	HMM-SEG	SF
Puissance de détection	0.97	0.84	0.94
quand de la sélection est détectée			
Nombre de fenêtres de sélection*	1.13	1.13	X
Taille d'une fenêtre de sélection*	5.99 kb	5.71 kb	X
Site sous sélection inclus?	1.00	0.81	X
Erreur d'estimation sur j^*	1.48 kb	2.45 kb	3.20 kb

* : en moyenne

Conclusions et perspectives

- Méthodologie alternative pour la détection de balayages sélectifs récents.
- Bonne puissance de détection, y compris pour sélection faible, et localisation précise du site sous sélection.
- Peu sensible à la démographie (croissance, goulot d'étranglement).
- Développements futurs : estimation des paramètres par EM, prise en compte des taux de mutation variables le long de la séquence.

Bibliographie

- Kim, Y. and W. Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-777, 2002.
- Nielsen, R., Williamson, S., Kim, Y, Hubisz, M. J., Clark, A. G., and Bustamante, C.. Genomic scans for selective sweeps using SNP data. Genome Research 1566-1575, 2005.
- Spencer, C.C.A. and Coop, G.. SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20:3673-3675.