Conseil scientifique du projet CIMENT-GRID

Lundi 13 Janvier 2003

Méthodes de Monte Carlo par chaînes de Markov en génétique des populations

Olivier François

TIMC - IMAG

en collaboration avec S. Manel (LECA) et M. Blum (TIMC)

Des perspectives scientifiques ouvertes

- De nouveaux moyens pour l'étude de la génétique des populations : couplage de l'explosion de l'information au niveau moléculaire et de méthodes statistiques "computationnelles" (bayésiennes).
- Étude de la structure des populations, des flux de gènes, de l'évolution de l'ADN, de la cartographie des gènes, etc
- Applications : Écologie, Biologie de la conservation, Génétique du cancer, Épidémiologie moléculaire, etc

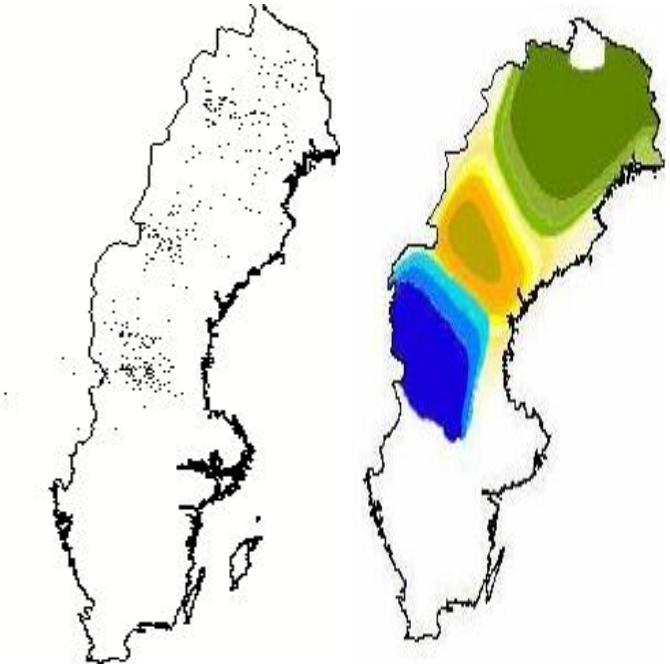
Un exemple en biologie de la conservation

Collaboration avec S. Manel (leca), A. Cercueil (timc), J. Swenson (NLH, Norvège)

- Espèce *Ursus arctos* (Ours brun de Scandinavie) : 516 individus séquencés, 19 marqueurs de type *STR* (Short Tandem Repeats)
- Population décimée jusqu'en 1930, protégée depuis, et reconstituée à partir de 4 foyers maternels originels bien connus géographiquement.
- Objectif : Estimer l'évolution de la structure de la population et la dispersion spatiale des individus et des gènes de génération en génération.

Structure génétique de la population

- Méthode de phylogénie fondée sur le modèle de mutation IAM (infinité d'allèles) dont la pertinence a été testée grâce à l'estimation MCMC des flux de gènes : `migrate` (Beerli et al., PNAS 2001)
- Méthode bayésienne d'estimation de mélange : `structure` (Pritchard et al., Genetics 2000)



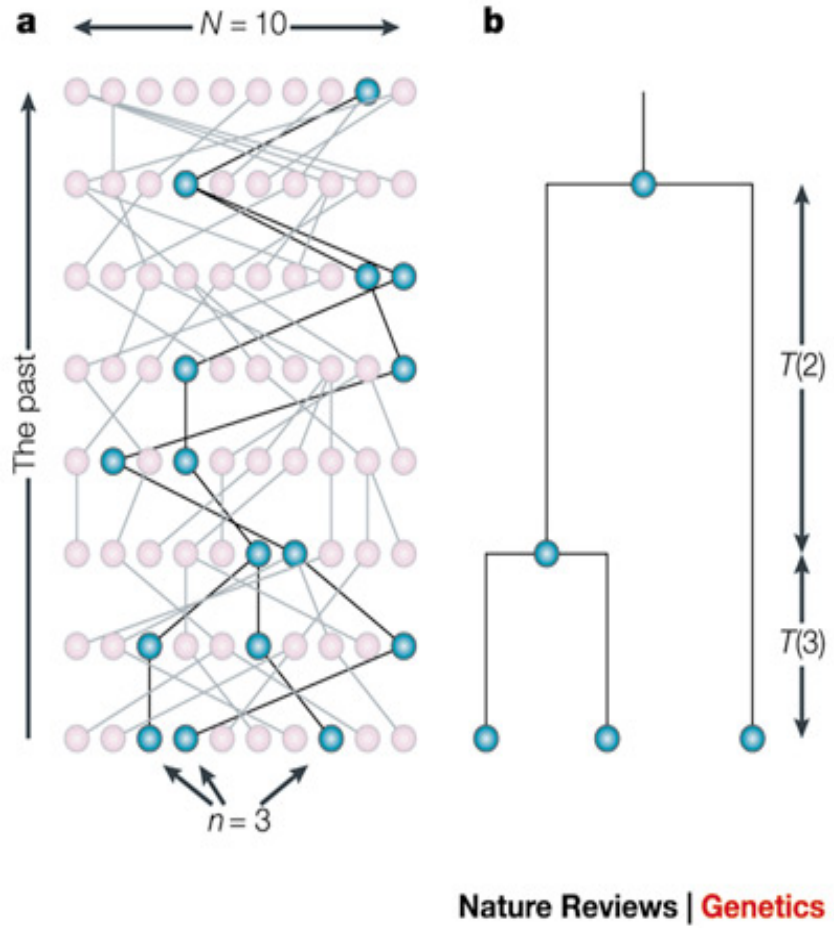
Modélisation statistique

- Modèle aléatoire de généalogie décrivant les relations de parenté entre individus : le *coalescent* (Kingmann, 1982).
- Événements structurants : mutations (μ), dispersion spatiale (σ) décrits par un paramètre scalaire

$$\theta = 2N\mu \qquad \text{ou} \qquad \theta = N\sigma^2$$

- La véritable généalogie n'est pas connue : Modèle à données manquantes

Arbre de coalescence



Arbre de coalescence de $n = 3$ individus dans une population de $N = 10$ individus.

Estimation par une méthode de maximum de vraisemblance

- Calcul de la vraisemblance

$$L(\theta) = p(D/\theta)$$

  D représente l'ensemble des observations moléculaires.

- Difficulté : le calcul direct de la vraisemblance est infaisable

$$L(\theta) = \int_{\mathcal{T}} p(D/T)p(T/\theta)dT$$

- Nécessité d'intégrer sur l'ensemble de toutes les généalogies $\mathcal{T}$.

- Calcul de l'intégrale $\longrightarrow$ Méthode de Monte Carlo sophistiquée.

Principe de la méthode de Monte Carlo

- Calcul approché de la vraisemblance : *Méthode d'Importance Sampling*

$$L(\theta) \propto \frac{1}{T} \sum_{t=1}^{T} \frac{p(T^{(t)}/\theta)}{p(T^{(t)}/\theta_0)}$$

où θ_0 est une valeur de référence (estimation grossière)

- $p(T/\theta)$ est calculé de façon exacte selon la *loi du coalescent* (N grand)

$$p(T/\theta) = \frac{1}{\theta^{n-1}} \exp(-\sum_{k=1}^{n-1} \lambda_k \frac{z_k}{\theta}).$$

- Les généalogies $T^{(t)}$ sont simulées selon la loi d'importance

$$q(T) = p(T/D;\theta_0) \propto p(D/T)p(T/\theta_0).$$

Simulation de la loi d'importance

- Méthode de *Métropolis-Hastings* : Markov chain Monte Carlo (MCMC)
- On fait évoluer de manière itérative un arbre initial en ne modifiant qu'un noeud interne à chaque transition et en recalculant les temps de coalescence conditionnels.
- *Méthode de rejet* fondée sur le calcul de $p(D/T)$ (programmation dynamique).
- Il est difficile de contrôler la vitesse de convergence de cette méthode théoriquement.

Échantillon d'arbres tirés selon le modèle du coalescent

Lors d'un échantillonnage aléatoire des généalogies, certaines (rares) vont contribuer à la vraisemblance, mais un grand nombre risque d'être improbable. Cela justifie de diriger la méthode MC vers les généalogies qui ont le plus de chance d'être réalisées.

Logiciels

- `genetree` : Estimation des taux de mutation pour les séquences (taille de population variable)
- `lamarc, fluctuate, migrate` : taille efficace, taux de mutation et migration entre sous-populations, grande variété de modèles pour les données
- `batwing` : taux de migration entre sous-populations avec séparation (split) de la population ancestrale

Utilisation des grappes/fermes de PC

- Taille des jeux de données : 10-100 marqueurs pour 100-1000 individus
- Temps de calcul actuel pour un run (pc 2.4Ghz, 1024Kram) : 2 à 20 jours.
- Parallélisation possible : un arbre généalogique par marqueur (équilibre de liaison)
- Difficultés de calibrage propre aux algorithmes MCMC (convergence)
- Nécessité de nombreuses simulations pour valider les modèles d'evolution et les estimateurs

Estimation de la dispersion spatiale chez Ursus arctos

- Développement d'une méthode de pseudovraisemblance (calcul rapide)
- Utilisation de grappes de PCs pour la correction du biais de la méthode (simulations)
- Estimation du paramètre de dispersion mère/fille $\sigma \approx 8 - 11$ km

Conclusions

- Pour utiliser ou développer de nouveaux logiciels, des moyens de calcul intensif sont indispensables : données réelles, validation des estimateurs, validation des modèles
- Similitude avec les méthodes de Monte Carlo pour les systèmes de particules : exploration de grands espaces d'états (généalogies), transition de phase pour les algorithmes, etc
- Multi-disciplinarité : génétique des populations / statistique / informatique