

Journées Statistiques du Sud 2010

Mèze, 23 - 25 juin 2010

Mercredi 23 juin 2010 matin

- 08h30 - 09h00 : Accueil des participants et présentation des Journées
- 09h00 - 10h30 : **Olivier François**
Population genetics models I
Modèles de génétique des populations I
- 10h30 - 11h00 : Pause
- 11h00 - 12h00 : **Nicolas Verzelen**
What is the minimal sample size to analyze a microarray experiment?
Quelle taille minimale d'échantillon pour analyser des données microarray ?

Les récents progrès en acquisition de données génomiques permettent désormais d'étudier l'ensemble des gènes comme un système complexe. À partir de données d'expression de gènes, on cherche par exemple à inférer les interactions entre gènes. Le coût relativement élevé de la collecte des données fait que la taille n de l'échantillon est beaucoup plus faible que le nombre p de gènes étudiés. On parle de problème d'estimation en grande dimension. Un grand nombre de travaux ont été consacrés à des procédures d'estimation du graphe d'interaction entre gènes dont les propriétés de convergence restent valables en grande dimension. On s'intéressera ici au problème symétrique de savoir si pour une taille d'échantillon n et un nombre p de gènes donnés, il est possible ou non de "bien" estimer le graphe. D'un point de vue mathématique, cette question correspond à établir des risques minimax d'estimation pour un modèle de régression linéaire dont le design suit une loi gaussienne.

Mercredi 23 juin 2010 après-midi

- 14h00 - 15h00 : **Olivier Gimenez**

Hidden process models for population dynamics

Modèles à structure cachée pour la dynamique des populations

Les modèles de capture-recapture sont utilisés depuis longtemps pour estimer les paramètres démographiques de populations animales, végétales et humaines avec des applications en écologie, évolution, épidémiologie et sociologie. Le développement de ces modèles a récemment bénéficié de l'introduction de structures cachées et de la modélisation hiérarchique. En se focalisant sur le problème de l'hétérogénéité individuelle, nous illustrons ces progrès sur 3 cas d'étude en biologie de la conservation et écologie évolutive.

- 15h00 - 16h00 : **Simona Grusea**

Assessing the significance of conserved genomic regions

Test de significativité pour des régions génomiques conservées

In this talk I will present a statistical test for detecting significant conserved genomic regions (or gene clusters) between two different species. We will focus on gene clusters found by the "reference-region" approach. First, we will consider only the proximity of the genes in the clusters. I will present a compound Poisson approximation for computing probabilities involved in the significance test. In a second time, we will be interested also in the gene order. We propose three measures, based on the transposition distance in the permutation group, for quantifying the exceptionality of the order of the genes in the clusters. We have obtained analytic expressions for the distributions of these measures for a random permutation, i.e. under the null hypothesis of random gene order in the genomes. In the remaining time I will speak about some results on the distribution of the reversal distance in the case of a random genome.

- 16h00 - 16h30 : Pause

- 16h30 - 18h00 : **Stéphane Robin**

Some statistical models for genomic analysis I

Méthodes statistique pour l'analyse de données génomiques I

Jeudi 24 juin 2010 matin

- 09h00 - 10h30 : **Stéphane Robin**
Some statistical models for genomic analysis II
Méthodes statistique pour l'analyse de données génomiques II
- 10h30 - 11h00 : Pause
- 11h00 - 12h30 : **Olivier François**
Population genetics models II
Modèles de génétique des populations II

Jeudi 24 juin 2010 après-midi

- 14h00 - 15h00 : **Raphaël Leblois - François Rousset**
Inference of demographic parameters using importance sampling algorithms
Estimation de paramètres démographiques par algorithmes d'échantillonnage préférentiel

Likelihood inferences of demographic parameters from genetic data have been slowly developing due to the difficulties of computing likelihoods under most genetic models of interest. Some recent advances in this field use importance sampling algorithms developed by Griffiths et al. I will present such IS algorithms and then describe their performances under two different scenarios : (i) inference of migration rates under isolation by distance models and (ii) detection and characterization of a past change in population size. I will shows that IS algorithms are very efficient for equilibrium and nearly-equilibrium models and in such situations, likelihood surfaces are very well etimated. On the other hand, estimation of the likelihood for models under strong disequilibrium is much more computationally demanding and often very unprecise. However, comparison with alternative likelihood based inference methods indicate better precision for IS algorithms in almost all situations.

- 15h00 - 16h00 : **Pierre Ribereau**
Extreme values theory for environmental science
Théorie des valeurs extrêmes en sciences de l'environnement

Nous présenterons quelques problèmes statistiques liés à l'application de la théorie des valeurs extrêmes en sciences de l'environnement. Nous porterons une attention particulière à l'estimation non-paramétrique de la densité spectrale pour des variables à variation régulière multivariées. Ces techniques seront illustrées sur des données de concentration de gaz dans l'atmosphère.

- 16h00 - 16h30 : Pause
- 16h30 - 17h30 : **Sophie Schbath**
Statistics of motifs: importance in genome analysis and open problems
Statistique de mots en génomique: ce qui a été fait, ce qu'il reste à faire

L'identification de motifs d'ADN fonctionnels au sein des génomes est l'une des préoccupations du processus d'annotation des génomes. Elle repose en partie sur la prédiction de motifs candidats obtenue par des méthodes statistiques, prédictions qu'il faut ensuite valider expérimentalement. Ces méthodes statistiques exploitent généralement le caractère innattendu dit "exceptionnel" de certaines quantités directement liées aux occurrences de motifs le long d'une séquence d'ADN. Nous passerons en revue un certain nombre de ces quantités d'intérêt (comptage, présence, distance, temps d'attente, etc.) en indiquant les résultats statistiques sur leur distribution et les modèles dans lesquels ils sont obtenus, et nous soulèverons à chaque fois des questions de recherche non résolues ou à approfondir.

Vendredi 25 juin 2010 matin

- 09h00 - 10h30 : **Olivier François**
Population genetics models III
Modèles de génétique des populations III
- 10h30 - 11h00 : Pause
- 11h00 - 12h30 : **Stéphane Robin**
Some statistical models for genomic analysis III
Méthodes statistique pour l'analyse de données génomiques III

Mini-cours de Olivier François **Population genetics models** **Modèles de génétique des populations**

La génétique des populations est une discipline née dans la première partie du 20e siècle. Sa théorie est construite sur des modèles mathématiques, qui pendant longtemps, sont restés très en avance sur des données empiriques presque inexistantes. Depuis quelques années, la tendance s'est inversée et la discipline, devenue essentiellement empirique, doit faire face à un afflux de données sans précédent dans son histoire.

Dans ce mini-cours, nous montrerons comment les modèles mathématiques de la génétique des populations peuvent être utilisés dans des approches modernes "computationnelles" pour étudier les données de polymorphisme génétique intra-spécifique. Le parti-pris est d'illustrer quelques exemples fondamentaux d'inférence bayésienne pour des questions classiques de la génétique des populations : la sélection naturelle, la dérive génétique et les processus démographiques, la structure génétique des populations et leur métissage.

Séance 1 : nous présenterons une méthode de criblage bayésien pour détecter les régions du génome responsables de l'adaptation locale des organismes à leur environnement. Après une brève introduction aux techniques de simulation par chaînes de Markov, nous décrirons les principes de l'algorithme de Beaumont et Balding, puis une application à des populations de drosophile.

Séance 2 : nous présenterons une introduction au modèle du coalescent ainsi qu'aux méthodes dites "Approximate Bayesian Computation" (ABC). Nous montrerons comment ces méthodes de Monte Carlo sont utilisées pour estimer la taille efficace des populations à partir des données de séquences d'ADN. Nous discuterons aussi les extensions des méthodes ABC reposant sur des techniques d'apprentissage statistique.

Séance 3 : nous présenterons les modèles de "clustering" bayésien pour l'inférence de la structure génétique des populations à partir de données multilocus. Après un rappel des concepts de déséquilibre (Hardy-Weinberg et liaison), nous présenterons l'algorithme de Pritchard, Stephens et Donnelly. Ensuite, nous présenterons l'extension de ce modèle dans un contexte spatialement explicite, permettant une modélisation fine de la variation du métissage des génomes le long de gradients géographiques.

Bibliographie :

- Beaumont M. et Balding D. (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*. 13: 969-980.
- Beaumont M. et Rannala B. (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics* 5: 251-261.
- Csilléry K. et al. (2010) Approximate Bayesian Computation (ABC) in practice, *Trends in Ecology and Evolution*, online pub.
- Durand E. et al. (2009) Spatial inference of admixture proportions and secondary contact zones. *Mol. Biol. Evol.* 26:1963-1973.
- Marjoram P. et Tavaré S. (2006) Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet.* 7(10):759-70.
- Pritchard J. et al. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.

Mini-cours de Stéphane Robin

Some statistical models for genomic analysis

Méthodes statistique pour l'analyse de données génomiques

For many years, the analysis of genomic data has been raising original and interesting statistical issues. This course will try to give an overview of some of the recent ones, focusing on the specific models that need to be developed for each question.

Microarrays now constitute a common technique to evaluate DNA or RNA abundance within a cell or population of cells. **The first lecture** will focus on high-density arrays where probes are spread all along the genome, regardless of the gene positions. We will consider their use for the detection of new genes, protein-DNA interactions or chromosomal aberrations.

“Next generation sequencing” (NGS) technologies now tend to replace microarray in many applications. **The second lecture** will present the type of data provided by NGS, which strongly differs from microarrays. We will discuss which statistical questions need (or do not need) to be reconsidered. We will also introduce meta-genomic studies, which aim at providing a global picture of the species present in a given middle.

Because the living of a cell is basically governed by the interactions between its components, biological networks have become a central object of study in the recent years. **The third lecture** will introduce different kind of networks and discuss some statistical aspects regarding their inference, their evolution, their dynamics or the analysis of their topological properties.