Bayesian inference and variant identification in metagenomic data: subsampling methods for massive datasets.

Keywords: Hierarchical mixture models, inférence bayésienne, données massives, Markov Chain Monte Carlo, méthodes de sondage, sous-échantillonnage, vraisemblance approchée, métagénomique.

Mots-clefs: Modèles de mélange hiérarchiques, inférence bayésienne, données massives, Markov Chain Monte Carlo, méthodes de sondage, sous-échantillonnage, vraisemblance approchée, métagénomique.

La version française du sujet se débute en page 4.

Context

Metagenomic analyses involve several biological samples sharing multiple variants (strains) of a genome, each sample being a mixture of these variants with sample-specific relative abundances (some may be zero). For example, wastewater sampling across cities to monitor viral circulation (e.g. SARS-CoV-2) will reveal variants (Alpha, Delta, Omicron, ...) at differing proportions in each city. Statistically, we aim to infer (i) the genetic sequences of the variants (discrete variables) and (ii) their sample-wise proportions from raw sequencing reads corrupted by measurement error and sampling variability.

A hierarchical mixture model arises naturally: each sample s is a mixture over variants g with weights $\pi_{g,s}$, while all samples share the same unknown variant genome sequences τ_g. The generation of a read can be described naturally: one read originates from a single variant with probability $\pi_{\cdot,s_r}$, it covers a fragment of its sequence. Then, bases are potentially affected by sequencing noise, described by an error matrix ε (probability of observing a when the true base is b). This formalism cleanly separates structural parameters (τ, π) from the observation model (ε), which is valuable for statistican analysis and evaluating subsampling effects.

The problem is high-dimensional, with discrete and continuous components, simplex constraints, read errors and subsampling uncertainty; frequentist point estimators with asymptotic approximations can be unreliable. Bayesian inference naturally integrates all sources of uncertainty, hierarchical sharing of variants, and produces posterior distributions instead of point estimators. Markov Chain Monte Carlo (MCMC) methods allow efficient exploration of these complex distributions.

A computational bottleneck is the full read likelihood: aggregating over all reads and covered positions becomes prohibitive for large numbers of reads and long genomes. The idea for this internship is to study principled data subsampling (reads and/or positions) inspired by survey sampling, then Bayesian inference using the reduced data. The central question is both theoretical and practical: can we control the approximation error of the likelihood (or posterior functionals), and how to design sampling plans that minimise error for fixed cost?

Sampling methods

Markov Chain Monte Carlo (MCMC) methods are ubiquitous in a wide range of scientific and industrial fields such as population genetics, phylogenetics, cosmology, computational chemistry, epidemiology, quantitative finance and statistical learning. They have become an essential tool for inference in complex models where analytical approaches are impractical. The Monte Carlo Markov chain approach is a fascinating methodology that has profoundly influenced scientific practice by offering a flexible framework to explore highly structured or multimodal distributions that are often inaccessible analytically. Its principle rests on constructing a chain whose stationary distribution coincides with the desired posterior; posterior expectations are then approximated by empirical averages along the simulated trajectory.

In our context, the goal is to sample from the posterior distribution of the parameters $\theta = (\tau, \pi, \varepsilon)$, where τ denotes the discrete variant sequences and π their continuous proportions lying on simplices. This is challenging because of high dimensionality, strong dependence between τ and π, multimodality, and computational cost dominated by the number and length of reads. Subsampling reduces cost but injects noise into Metropolis–Hastings acceptance ratios; careful calibration is required to avoid deteriorating acceptance rates or biasing the target.

A starting point exists with the Gibbs algorithm proposed by [Quince et al., 2017], designed for aggregated counts rather than individual reads, yielding a coarse approximation to the true likelihood. Although not very

efficient in our setting, it provides a useful structural backbone. The team has already developed several enhanced variants incorporating modern techniques, which the intern will be able to leverage.

More broadly, the field of MCMC methods is highly dynamic, supported by a large community and fueled by numerous recent advances [Fearnhead et al., 2024]. For instance, Delayed acceptance rapidly filters proposals using a cheap surrogate before computing the full likelihood, lowering per-iteration cost without altering the target. For the continuous part π, non-reversible dynamics such as Zig-Zag or Bouncy Particle, applied after a reparameterisation (e.g. stick-breaking), exploit log-likelihood differentiability and can improve mixing. Block updates or collapsed strategies for τ reduce the effective dimension and favour exploration of nearby modes.

This combination of ideas illustrates the flexibility and conceptual depth of MCMC: the methods adapt to complex constraints, integrate refinements for scalability, and retain theoretical rigour.

Model and likelihood description

More precisely, after aligning the measured gene fragments to known positions for a given gene (and discarding fragments that are too short or with too many gaps), the data consist of read fragments: sequences over $\{A, C, G, T\}$ indexed by a subset of the positions of the gene under study.

We assume that reads originate from a set of G variants. For convenience, encode variant sequences by binary indicators: $\tau_{v,g,b} = 1$ if the nucleotide at position v of variant g is b, and 0 otherwise. Let $\varepsilon_{b,a}$ be the probability of observing a instead of b (sequencing error). For a read r from sample s_r, let $\mathbf{v}_r$ be the set of positions covered by this fragment; for $v \in \mathbf{v}_r$ and a nucleotide a, define $x_{r,v,a} = 1$ if base a is observed at position v in read r, and 0 otherwise.

Single variant ($G = 1$). With indicator sequence $\tau_{1,v,b} \in \{0, 1\}$, the probability that read r contains nucleotide a at position v is:

$$P(X_{r,v,a} = 1 \mid \varepsilon, \tau) = \sum_{b \in \mathscr{A}} \tau_{1,v,b}\, \varepsilon_{b,a}.$$

Likelihood of one read (single variant), obtained by considering all positions covered by read r:

$$\mathscr{L}_r^{(G=1)}(\varepsilon, \tau) = \prod_{v \in \mathbf{v}_r} \prod_{a \in \mathscr{A}} P(X_{r,v,a} = x_{r,v,a} \mid \varepsilon, \tau).$$

For G variants with mixture weights π_{g,s_r} depending on the sample s_r of read r, we obtain the usual mixture likelihood, reflecting that each read comes from variant g with probability vector $\pi_{\cdot,s_r}$:

$$\mathscr{L}_r(\varepsilon, \pi_{\cdot,s_r}, \tau) = \sum_{g=1}^{G} \pi_{g,s_r}\, \mathscr{L}_r^{(G=1)}(\varepsilon, \tau_g).$$

The full likelihood over all reads $r \in \mathscr{R}$ is

$$\mathscr{L}\left(\varepsilon, \pi_{1:G,1:S}, \tau_{v,1:G}; \mathscr{R}\right) = \prod_{r \in \mathscr{R}} \mathscr{L}_r(\varepsilon, \pi_{\cdot,s_r}, \tau).$$

The large number of reads and the length of fragments make this evaluation costly, which motivates considering read subsampling.

Likelihood under subsampling

Computing the likelihood on only a subset of reads accelerates runtime. If this subset $\mathscr{R}^\star \subset \mathscr{R}$ is drawn at random and independently of the read values, the likelihood of the selected reads naturally writes

$$\mathscr{L}^\star(\theta; x^\star) = \prod_{r \in \mathscr{R}^\star} \mathscr{L}_r(x_r, \theta).$$

This can reduce computational complexity and speed up posterior approximation, at the price of losing information carried by the unselected data. To limit this loss, one could envision randomising selection so as to preserve certain marginal counts, thereby reducing the variance due to sampling. Indeed, not all reads carry the same information: reads that cover polymorphic sites discriminating between variants, that are long, and of good quality

(e.g. unambiguous alignment) provide a strong signal about g and τ. Conversely, short reads or reads redundant over conserved or repeated regions contribute little to the likelihood. It is therefore natural to prioritise some reads over others in the subsample.

However, in that case the selection is *endogenous*. The observed-data likelihood must integrate the full likelihood over all possible values of the unselected reads and can be written

$$\int_{x^\dagger} \left(\left(\prod_{r \in \mathscr{R}^\star} \mathscr{L}_r(x_r, \theta) \times \prod_{r \notin \mathscr{R}^\star} \mathscr{L}_r(x_r^\dagger, \theta) \right) \times P(\mathscr{R}^\star \mid x^\star, x^\dagger) \right) \mathrm{d}x^\dagger,$$

where $x^\dagger$ denotes the unselected reads and $P(\mathscr{R}^\star \mid x^\star, x^\dagger)$ the probability of selecting $\mathscr{R}^\star$ conditional on the complete read values.

In practice, this likelihood is intractable. An alternative is to use the inverse-probability-weighted pseudo log-likelihood:

$$\sum_{r \in \mathscr{R}^\star} P(r \in \mathscr{R}^\star \mid x)^{-1} \log \mathscr{L}_r(x_r, \theta),$$

whose conditional expectation given the reads equals the full log-likelihood.

Questions:

— How to choose inclusion probabilities to minimise inferential error for fixed cost?
— Can we select informative positions within reads, and how to choose them?
— Quantify the distance between full posterior and pseudo-likelihood induced posterior?
— How to stabilise the pseudo-likelihood?

Indicative bibliography

— Biological variant identification problem: Christopher Quince, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. DESMAN: A new tool for de novo extraction of strains from metagenomes. Genome Biology, 18(1):181, September 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1309-9
— Tillé, Yves. Sampling algorithms. Springer, 2006.
— Rubin, D. B. Inference and missing data. Biometrika 63(3), 1976.
— Recent Monte Carlo methods: Paul Fearnhead, Christopher Nemeth, Chris J. Oates, and Chris Sherlock. Scalable monte carlo for bayesian learning, 2024. URL `https://arxiv.org/abs/2407.12751`

Required skills

Probability/statistics (hierarchical models, estimation), asymptotic analysis, MCMC, programming (Python/R/Julia).

Supervision and location

Supervision: Anne-Laure ABRAHAM (MaIAGE, INRAE), Daniel BONNÉRY (IGN), Guillaume KON KAM KING (MaIAGE, INRAE). Work on site at MaIAGE (INRAE, Jouy-en-Josas). Access to Migale computing resources. Remote work possible 1–3 days/week. Housing possibly available on site. Internship stipend per current regulations; partial transportation expenses reimbursement. **Contacts:** anne-laure.abraham@inrae.fr, daniel.bonnery@ign.fr, guillaume.konkamking@inrae.fr

References

Paul Fearnhead, Christopher Nemeth, Chris J. Oates, and Chris Sherlock. Scalable monte carlo for bayesian learning, 2024. URL `https://arxiv.org/abs/2407.12751`.

Christopher Quince, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. DESMAN: A new tool for de novo extraction of strains from metagenomes. Genome Biology, 18(1):181, September 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1309-9.

Inférence bayésienne et identification de variants dans les données métagénomiques: méthodes de sous-échantillonnage pour les données massives.

Contexte

Les analyses métagénomiques portent sur plusieurs échantillons biologiques partageant plusieurs variants (ou "souches") d'un génome, chaque échantillon étant un mélange de ces variants avec des abondances relatives qui varient d'un échantillon à l'autre (certaines abondances pouvant être nulles). Par exemple, on peut imaginer des prélèvements réalisés dans différentes villes d'un pays pour surveiller la circulation d'un virus (e.g. SARS-CoV-2) : chaque ville peut abriter plusieurs variants (Alpha, Delta, Omicron, ...) à des proportions différentes. L'objectif statistique est d'inférer (i) la séquence génétique des variants (variables discrètes) et (ii) leurs proportions par échantillon à partir d'observations brutes issues du séquençage du génome, bruitées par des erreurs de lecture et par la nature aléatoire du prélèvement.

Ce problème se formalise naturellement par un modèle hiérarchique de mélange : chaque échantillon s est vu comme un mélange de composantes — les variants — où chaque composante correspond à une séquence discrète (le génome τ_g du variant) et le poids $\pi_{g,s}$ représente l'abondance relative du variant g dans l'échantillon s. La partie "hiérarchique" tient au fait que les mêmes composantes (les mêmes séquences τ_g) sont partagées entre tous les échantillons, tandis que seuls les poids $\pi_{g,s}$ varient d'un échantillon à l'autre (et peuvent être nuls si un variant est absent). Les données de séquençage se présentent sous la forme d'un grand nombre de *reads*, fragments des génomes des variants potentiellement corrompus par une erreur de mesure. La génération d'un read s'exprime naturellement : pour un read issu de l'échantillon s, on tire d'abord un variant g selon $\pi_{\cdot,s}$, on prélève un fragment de sa séquence τ_g, puis la lecture est affectée par un bruit de séquençage modélisé par une matrice d'erreur ε (probabilité de lire a alors que la vraie base est b). Ce formalisme dissocie clairement les paramètres de structure (τ et π) du modèle d'observation (ε), ce qui facilite l'analyse statistique et l'étude de l'impact du sous-échantillonnage.

La complexité du problème — variables discrètes et continues, contraintes de type simplexe, erreurs de lecture et incertitude liée au sous-échantillonnage — rend les approches classiques peu adaptées. Une méthode fréquentiste fournirait des estimateurs ponctuels et des approximations asymptotiques souvent peu fiables dans un espace de paramètres multimodal et fortement corrélé. L'inférence bayésienne, en revanche, offre un cadre naturel pour intégrer toutes les sources d'incertitude, modéliser la hiérarchie des paramètres (variants partagés entre échantillons, proportions spécifiques à chaque échantillon), et produire des distributions a posteriori plutôt que des estimations uniques. Cette flexibilité est essentielle pour quantifier la variabilité des résultats, comparer des modèles et prendre en compte le bruit induit par le sous-échantillonnage. De plus, les méthodes de Monte Carlo par chaînes de Markov (MCMC) permettent d'explorer efficacement ces distributions complexes, rendant possible une inférence robuste.

Un véritable goulot d'étranglement computationnel est le calcul de la vraisemblance complète des reads : celle-ci requiert d'agréger sur tous les fragments observés et toutes les positions du génome couvertes, ce qui devient prohibitivement coûteux lorsque le nombre de reads et la longueur des génomes sont importants. L'idée étudiée dans ce stage est d'effectuer, en s'inspirant des méthodes développées dans le cadre des mathématiques pour les sondages, un sous-échantillonnage systématique (tirage aléatoire de fragments et/ou de positions) afin de réduire la charge de calcul, puis d'utiliser cette version réduite pour l'inférence bayésienne. La question centrale est théorique et pratique : peut-on obtenir des garanties contrôlées sur la qualité de l'approximation de la vraisemblance (ou des estimateurs postérieurs) obtenue par sous-échantillonnage, et quel plan de sondage minimise l'erreur pour un coût donné ?

Méthodes d'échantillonnage

Les méthodes de Monte Carlo par chaînes de Markov (MCMC) sont omniprésentes dans des domaines scientifiques et industriels variés, tels que la génétique des populations, la phylogénie, la cosmologie, la chimie computationnelle, l'épidémiologie, la finance quantitative et l'apprentissage statistique. Elles constituent un outil incontournable pour l'inférence dans des modèles complexes, où les approches analytiques sont impraticables. L'approche par chaînes de Markov de type Monte Carlo (MCMC) est une méthode fascinante, qui a profondément marqué la pratique scientifique en offrant un cadre flexible pour explorer des distributions complexes, souvent inaccessibles par des méthodes analytiques. Son principe repose sur la construction d'une chaîne dont la loi stationnaire coïncide avec la loi a posteriori recherchée, permettant d'approximer des espérances par des moyennes le long de la trajectoire simulée.

Dans notre contexte, il s'agit d'échantillonner dans la distribution a posteriori des paramètres $\theta = (\tau, \pi, \varepsilon)$, où τ désigne les séquences discrètes des variants et π leurs proportions continues, soumises à des contraintes de type simplexe. Cette tâche est difficile en raison de la dimension élevée, des dépendances fortes entre τ et π, de la multimodalité et du coût de calcul dominé par le volume des reads. Le sous-échantillonnage réduit ce coût mais introduit du bruit dans les ratios d'acceptation, ce qui exige un calibrage précis pour éviter une dégradation du taux d'acceptation ou un biais sur la cible.

Une base de départ existe avec l'algorithme Gibbs proposé par [Quince et al., 2017], conçu pour des comptages agrégés plutôt que pour des reads individuels, ce qui constitue une approximation grossière de la vraisemblance. Cet algorithme n'est pas très efficace dans notre cadre, mais il fournit une structure utile. L'équipe a déjà développé plusieurs variantes améliorées, intégrant des techniques modernes, sur lesquelles l'étudiant pourra s'appuyer sans repartir de zéro.

Plus largement, le domaine des méthodes MCMC est extrêmement dynamique, porté par une large communauté et enrichi par de nombreuses avancées récentes [Fearnhead et al., 2024]. Par exemple, l'acceptation différée permet de filtrer rapidement les propositions avant d'évaluer la vraisemblance complète, réduisant ainsi le temps par itération sans altérer la cible. Pour la partie continue π, des dynamiques non réversibles comme Zig-Zag ou Bouncy Particle, appliquées à une reparamétrisation (par exemple stick-breaking), exploitent la différentiabilité de la log-vraisemblance et améliorent le mélange. Enfin, des mises à jour par blocs ou des stratégies ≪collapsées≫ pour τ diminuent la dimension effective et favorisent l'exploration des modes voisins.

Cette combinaison d'idées illustre la flexibilité et la profondeur conceptuelle des méthodes MCMC : elles s'adaptent à des contraintes complexes, intègrent des raffinements pour la scalabilité et conservent une rigueur théorique.

Description du modèle et de la vraisemblance

Plus précisément, les données métagénomiques se présentent (après un traitement consistant à aligner les fragments de gènes mesurés sur les positions connues d'un gène donné, et à écarter les fragments trop courts ou avec trop de trous), sous forme de séquences de lettres parmi A, C, G et T indexées par un sous ensemble de l'ensemble des positions du gène étudié.

On suppose que les fragments de gène mesurés proviennent d'un ensemble de G variants. Par commodité, on encode les séquences par des variables binaires, on note $\tau_{v,g,b} = 1$ si le nucléotide à la position v du variant g est b, 0 sinon. On note $\varepsilon_{b,a}$ la probabilité de d'observer a au lieu de b lors de la mesure d'un nucléotide (le bruit de séquençage). Pour un *read* r, obtenu d'un échantillon s_r, on note $\mathbf{v}_r$ l'index des positions disponibles pour ce fragment, et pour $v \in \mathbf{v}_r$ et un nucléotide a, on note $x_{r,v,a} = 1$ si le nucléotide mesuré pour ce fragment à cette position est a, 0 sinon.

Considérons un seul variant (G=1) avec séquence indicatrice $\tau_{1,v,b} \in \{0,1\}$, la probabilité qu'un read r contienne le nucléotide a en position v est :

$$P(X_{r,v,a} = 1 \mid \varepsilon, \tau) = \sum_{b \in \mathscr{A}} \tau_{1,v,b}\, \varepsilon_{b,a}. \tag{1}$$

La vraisemblance d'un read est obtenue, pour un seul variant, en considérant l'ensemble des positions couvertes par le read $\mathbf{r}$:

$$\mathscr{L}_{\mathbf{r}}^{(G=1)}(\varepsilon,\tau) = \prod_{v\in\mathscr{V}_r}\prod_{a\in\mathscr{A}} P(X_{r,v,a} = x_{r,v,a} \mid \varepsilon,\tau) \tag{2}$$

Pour G variants et un mélange π_{g,s_r} dépendant de l'échantillon s_r du read r, on obtient une vraisemblance de classique pour un modèle de mélange, qui traduit le fait que chaque read provient d'un variant g avec une probabilité $\pi_{\cdot,s_r}$:

$$\mathscr{L}_{\mathbf{r}}(\varepsilon,\pi_{1:G,s_r},\tau_{v,1:G}) = \sum_{g=1}^{G} \pi_{g,s_r}\mathscr{L}_{\mathbf{r}}^{(G=1)}(\varepsilon,\tau_g) \tag{3}$$

La vraisemblance complète s'obtient en agrégeant sur l'ensemble des reads $r \in \mathscr{R}$:

$$\mathscr{L}\left(\varepsilon,\pi_{1:G,1:S},\tau_{v,1:G};\mathscr{R}\right) \;=\; \prod_{\mathbf{r}\in\mathscr{R}} \mathscr{L}_{\mathbf{r}}(\varepsilon,\pi_{1:G,s_r},\tau_{v,1:G}).$$

Le grand nombre de reads et la longueur des fragments rendent le calcul de cette vraisemblance coûteux, et encourage à considérer la possibilité de sous-échantillonner les reads.

Vraisemblance en cas de sous-échantillonnage

Ne calculer la vraisemblance que sur un sous-échantillon des reads accélère le temps de calcul. Si ce sous-échantillon $\mathscr{R}^\star \subset \mathscr{R}$ est tiré aléatoirement et indépendamment des valeurs des reads, la vraisemblance des reads sélectionnés s'écrit naturellement : $\mathscr{L}^\star(\theta;x^\star) = \prod_{r\in\mathscr{R}^\star}\mathscr{L}_r(x_r,\theta)$. Cette vraisemblance peut permettre réduire la complexité des calculs et accélérer l'approximation de la loi a posteriori des paramètres d'intérêt, au prix de la perte de l'information apportée par les données non sélectionnées. Afin de perdre le moins d'information possible, il serait possible de sélectionner aléatoirement les reads de façon à préserver des comptages marginaux, afin de réduire la variance due à la sélection des reads. En pratique, tous les reads n'apportent pas la même information : ceux qui recouvrent des sites polymorphes discriminants entre variants, sont longs et de bonne qualité (e.g. alignement non ambigu) fournissent un signal fort sur g et τ. À l'inverse, des reads courts, redondants sur des régions conservées ou répétées contribuent peu à la vraisemblance. Il est donc naturel de privilégier certains reads plutôt que d'autres dans le sous-échantillonnage.

Toutefois, dans ce cas la sélection des reads est *endogène*, le calcul de la vraisemblance nécéssite d'intégrer la vraisemblance complète pour toutes les valeurs possibles des reads non sélectionnés et s'écrit :

$$\int_{x^\dagger}\left(\left(\prod_{r\in\mathscr{R}^\star}\mathscr{L}_r(x_r,\theta)\times\prod_{r\notin\mathscr{R}^\star}\mathscr{L}_r(x_r^\dagger,\theta)\right)\times\left(P(\mathscr{R}^\star \mid x^\star,x^\dagger)\right)\right)\,\mathrm{d}x^\dagger.$$

Ici, $x^\dagger$ désigne les reads non sélectionnés, et $P(\mathscr{R}^\star \mid x^\star,x^\dagger)$ la probabilité de sélectionner le sous-échantillon $\mathscr{R}^\star$ conditionnellement aux valeurs complètes des reads.

En pratique, cette vraisemblance est incalculable. Une alternative consiste à utiliser la pseudo log vraisemblance, donnée par :

$$\sum_{r\in\mathscr{R}^\star} P(\{r\in\mathscr{R}^\star \mid x\})^{-1}\log\mathscr{L}_r(x_r,\theta),$$

dont l'espérance conditionnellement aux reads est la log vraisemblance complète des observations.

Les questions intéressantes à explorer dans ce cadre sont les suivantes :

— Comment choisir les probabilités d'inclusion des reads pour minimiser l'erreur d'inférence pour un coût de calcul donné ?
— Peut-on sélectionner, au sein de chaque read, les positions les plus informatives pour l'inférence, et comment choisir ces positions ?
— Peut-on quantifier la distance entre la loi a posteriori complète et celle induite par la pseudo-vraisemblance ?
— Comment utiliser des techniques de stabilisation de la pseudo-vraisemblance ?

Ces divers points seront étudiés théoriquement dans des cas simples, puis mis en œuvre et testés sur des données simulées et réelles.

Bibliographie indicative

— Un article de biologie sur le problème d'identification de variants dans les données métagénomiques : Christopher Quince, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. DESMAN : A new tool for de novo extraction of strains from metagenomes. Genome Biology, 18(1) :181, September 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1309-9

— Tillé, Yves. Sampling algorithms. New York, NY : Springer New York, 2006.

— Donald B. Rubin. Inference and missing data. Biometrika, Volume 63, Issue 3, December 1976, Pages 581–592, https ://doi.org/10.1093/biomet/63.3.581

— Un livre sur les méthodes de Monte Carlo récentes : Paul Fearnhead, Christopher Nemeth, Chris J. Oates, and Chris Sherlock. Scalable monte carlo for bayesian learning, 2024. URL `https://arxiv.org/abs/2407.12751`

Compétences requises

Probabilités/statistiques (modèles hiérarchiques, estimation), analyse asymptotique, méthodes Monte Carlo Markov Chain, programmation (Python/R/Julia).

Encadrement et lieu

Encadrement par Anne-Laure ABRAHAM (MaIAGE, INRAE), Daniel BONNÉRY (IGN), Guillaume KON KAM KING (MaIAGE, INRAE). Travail local au laboratoire MaIAGE (Mathématiques et Informatique Appliquées du Génome à l'Environnement), INRAE, Jouy-en-Josas. Accès aux ressources de calcul de la plateforme Migale. Télétravail possible 1 à 3 jours par semaine. Hébergement sur site potentiellement disponible. Gratification selon le tarif horaire en vigueur et remboursement partiel du Passe Navigo.

Contacts

anne-laure.abraham@inrae.fr, daniel.bonnery@ign.fr, guillaume.konkamking@inrae.fr

Références

Paul Fearnhead, Christopher Nemeth, Chris J. Oates, and Chris Sherlock. Scalable monte carlo for bayesian learning, 2024. URL `https://arxiv.org/abs/2407.12751`.

Christopher Quince, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. DESMAN : A new tool for de novo extraction of strains from metagenomes. Genome Biology, 18(1) :181, September 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1309-9.