

# Approches variationnelles pour l'analyse des séquences génomiques évolutives

Amine Remita

Direction de recherche: Pr. Abdoulaye Baniré Diallo

Université du Québec à Montréal

1<sup>er</sup> mai 2024  
Doctorat en informatique

# Aperçu

- Introduction
  - Contexte
- Hypothèse et objectifs
- Contributions principales
  - Évaluation des modèles linéaires de classification
  - Modèle génératif bayésien variationnel
  - Modèle variationnel d'inférence phylogénétique
- Conclusion et directions futures

# Séquences génomiques

## Importance des séquences génomiques

- Diversité biologique
- Évolution des espèces
- Annotation fonctionnelle et structurelle
- Détection des pathogènes
- Conception de vaccins
- Résistance aux médicaments
- Surveillance épidémiologique

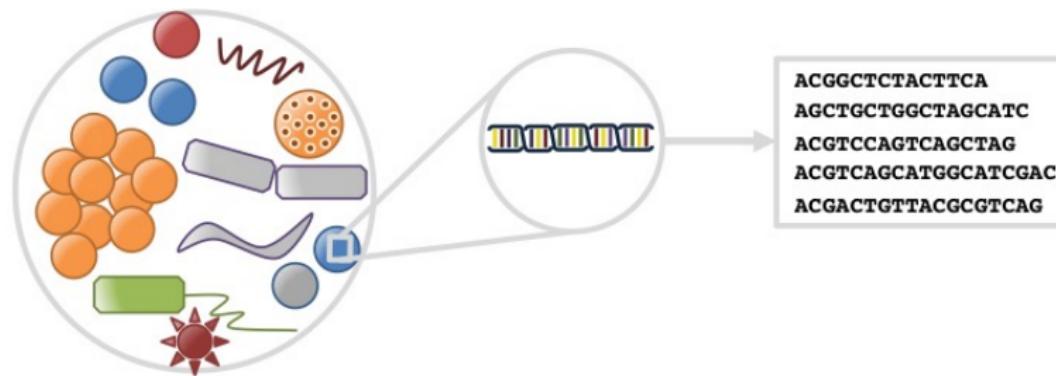


# Séquences génomiques

**Génome** : ensemble de matériel génétique de chaque organisme. Il est encodé dans des séquences d'ADN ou d'ARN.

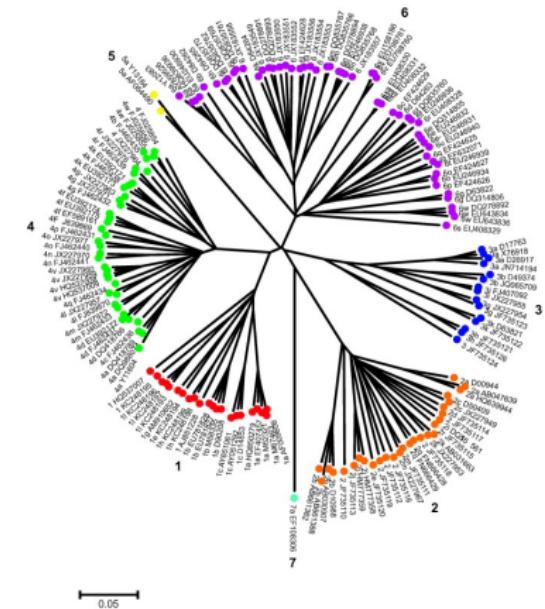
**Séquence** : enchaînement de symboles appartenant à un alphabet fini.

- ADN/ARN :  $\in \{A, C, G, T, U\}$



## Séquences évolutives

- Séquences évoluent le long d'une **phylogénie**.  
⇒ Histoire évolutive des séquences conceptualisée à travers d'un arbre (relations ancestrales)

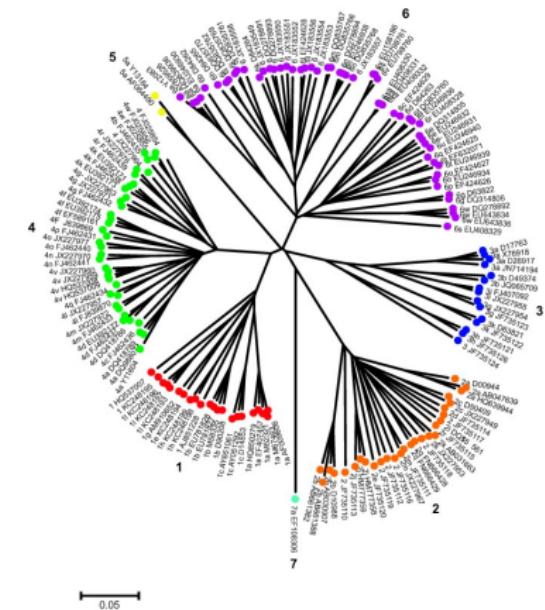


## Phylogénie du virus de l'hépatite C

(Jackowiak et al., 2014)

# Séquences évolutives

- Séquences évoluent le long d'une **phylogénie**.
  - ⇒ Histoire évolutive des séquences conceptualisée à travers d'un arbre (relations ancestrales)
- Mécanismes de modifications génomiques (Brown, 2002) :
  - ① Mutations ponctuelles : **substitution, indel d'un nucléotide**
  - ② Mutations à petite échelle : insertion/suppression de fragments
  - ③ Mutations à grande échelle : réarrangements génomiques



Phylogénie du virus de l'hépatite C

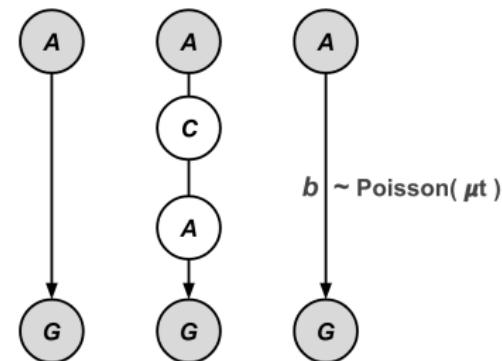
(Jackowiak et al., 2014)

# Modèle probabiliste d'évolution

- Modéliser les **mutations cachées** au cours du temps
  - Processus stochastique
- **Chaîne de Markov à temps continu**
  - Matrice de substitutions **Q**
  - Fréquences relatives  $\pi$
  - Matrice des probabilités de transition  
 $P(b) = e^{Qb}$

(Jukes and Cantor, 1969; Tavaré et al., 1986)

- Distance entre deux séquences
  - Similarité/divergence
  - Comparaison des séquences



Observation/  
supposition

Mutations  
cachées

Modélisation  
probabiliste

# Alignement de séquences

- Superposition des séquences en supposant que chaque position aie un ancêtre commun

(Durbin et al., 1998)

$x^1$  ATGTGTGAAACCTGTCGTG  
 $x^2$  ATGA-TGAGCCTGTGGTC  
 $x^3$  TTGAGTGAACCTG-GGTC  
 $x^4$  TTGAGTGAACGTGTGGTC

# Inférence évolutive

- Inférence de *paramètres* qui peuvent expliquer l'**histoire évolutive** d'un ensemble de **séquences liées**.

Étant donné :

- Alignement de séquences **X**

$x^1$  ATGTGTGAACCTGTCGTG  
 $x^2$  ATGA-TGAGCCTGTGGTC  
 $x^3$  TTGAGTGAACCTG-GGTC  
 $x^4$  TTGAGTGAACGTGTGGTC

Inférer :

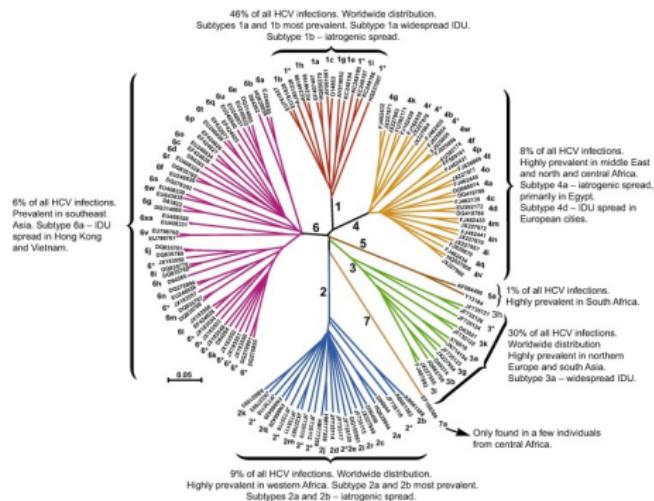
- Paramètres évolutifs  $\Theta_{\text{evo}}$ 
  - Phylogénie  $\tau$
  - Longueurs de branches **b**
  - Paramètres du modèle de substitution (K80, HKY85, GTR, etc.)
  - États ancestraux **a**, etc.

→ Modéliser  $P(\Theta_{\text{evo}} | X)$

# Séquences divergentes → diversité phénotypique

## Phénotypes

- Taxonomie
  - Génotypes et sous-types
- Pathogénicité et virulence
- Résistance aux médicaments
- Affinité vis-à-vis de l'hôte



Génotypes du virus de l'hépatite C

(Bukh, 2016)

# Classification des séquences

- **Assigner** une séquence inconnue à un groupe de séquences connues en fonction de leurs **caractéristiques**.

Étant donné :

- Ensemble de séquences  $\mathbf{X}$  classées en  $\mathbf{C}$

$x^1$  ATGTGTGAACCTGTCGTG

$x^2$  ATGATGAGCCTGTGGTC

$x^3$  TTGAGTGAACCTGGGTG

$x^4$  TTGAGTGAACGTGTGGTC

Inférer :

- Paramètres du classifieur  $\Theta_{\text{clf}}$ 
  - Entraînement du classifieur
  - Modéliser  $P(\Theta_{\text{clf}} | \mathbf{X}, \mathbf{C})$

Prédire :

- les classes  $\mathbf{C}'$  des nouvelles séquences  $\mathbf{X}'$ 
  - Modéliser  $P(\mathbf{C}' | \mathbf{X}'; \Theta_{\text{clf}})$

# Limitations des méthodes de classification des séquences

## Limitations :

- Absence d'un **modèle d'évolution explicite** dans l'algorithme de classification
- Classification dans un **espace fermé de classes** (A. Remita et al., 2017)

## Contraintes :

- Séquences homologues et de tailles comparables (Solis-Reyes et al., 2018; Lebatteux, A. M. Remita, and Diallo, 2019)
- Taux de faux positifs relativement élevé (A. Remita et al., 2017)

# Hypothèse initiale

- Amélioration de la classification des séquences biologiques :
  - ① Considérer les caractéristiques évolutives des séquences
  - ② Redéfinir la problématique de classification dans un espace ouvert de classes (open-set classification)

# Objectif global

- Intégration d'informations évolutives dans les modèles d'apprentissage statistique
  - ⇒ Modélisation de  $P(\Theta_{\text{evo}}, \Theta_{\text{clf}}, \mathbf{C}, \mathbf{X})$

# Objectifs spécifiques

- ① Évaluation exhaustive de modèles statistiques dans la classification des séquences
  - ② Conception d'un modèle génératif bayésien variationnel pour l'estimation de paramètres évolutifs et la génération de séquences
  - ③ Élaboration d'un modèle bayésien variationnel pour l'inférence de paramètres phylogénétiques
- 
- ★ Reproductibilité de la recherche
  - ★ Libre accès aux méthodes et aux résultats

## Objectif spécifique 1

# Évaluation exhaustive des modèles d'apprentissage statistique pour la classification des génomes de virus

- ★ Amine M. Remita et Abdoulaye Baniré Diallo (2019) Statistical Linear Models in Virus Genomic Alignment-free Classification: Application to Hepatitis C Viruses.  
Dans *IEEE International Conference on Bioinformatics and Biomedicine*
- ★ [https://github.com/maremita/slm\\_kgenomvir](https://github.com/maremita/slm_kgenomvir)

# Classificateurs des séquences virales et métagénomiques dans la littérature

- **Modèles génératifs** :  $P(\mathbf{C}, \mathbf{X}; \Theta_{\text{clf}})$

- Modèle bayésien naïf bayes :

RDP (Wang et al., 2007), NBC (Rosen et al., 2008)

- Markov d'ordre variable :

COMET (Struck et al., 2014)

- **Modèles discriminatifs** :  $P(\mathbf{C} | \mathbf{X}; \Theta_{\text{clf}})$

- Régression logistique :

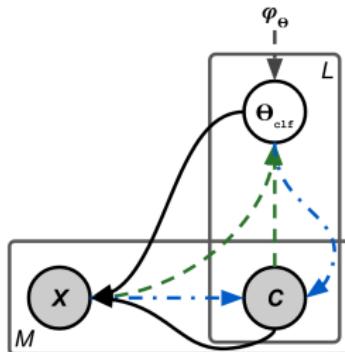
**VirFinder** (Ren et al., 2017),

**Kameris** (Solis-Reyes et al., 2018)

- Séparateurs à vaste marge (SVM) :

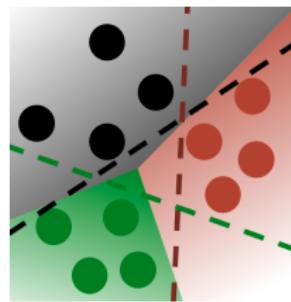
**PhyloPythiaS+** (Gregor et al., 2016),

**CASTOR-KRFE** (Lebatteux, A. M. Remita, and Diallo, 2019)



# Modèle linéaire

- Génère une **frontière de décision linéaire** pour séparer les classes
- Exprimé par une fonction  $f(\mathbf{w}^T \mathbf{x})$ 
  - $\mathbf{w}$  : vecteur de poids
  - $\mathbf{x}$  : vecteur représentant une séquence



# Comment calculer $\mathbf{x}$ ?

- **k-mers** : transformation d'une séquence en vecteur  $\mathbf{x}$  défini par le nombre d'occurrences de chaque mot  $u$ ; de taille  $k$  qui compose une séquence  $S$  (Vinga and Almeida, 2003).
- $\mathbf{x} = (x_1, x_2, \dots, x_m)$
- $x_i = \sum_{j=1}^{|S|-k+1} \mathbb{I}(u_i, S_{[j,j+k-1]})$

Séquences	ATGTGTG	CATGTG
Mots de tailles 3 (3-mers)	ATG TGT GTG TGT GTG	CAT ATG TGT GTG
Occurrences des mots	ATGTGTG	CATGTG
ATG	1	1
TGT	2	1
GTG	2	1
CAT	0	1

Adapté de Zielezinski et al., 2017

# Comment calculer les poids $\mathbf{w}$

- **Modèles génératifs :**

- $w_{i,c} = \log p(u_i | c)$
- Estimation par maximum de vraisemblance (MLE)
- Inférence Bayésienne : lissage additif (pseudocomptes  $\alpha$ )

- **Modèles discriminatifs :**

- Minimisation conjointe :

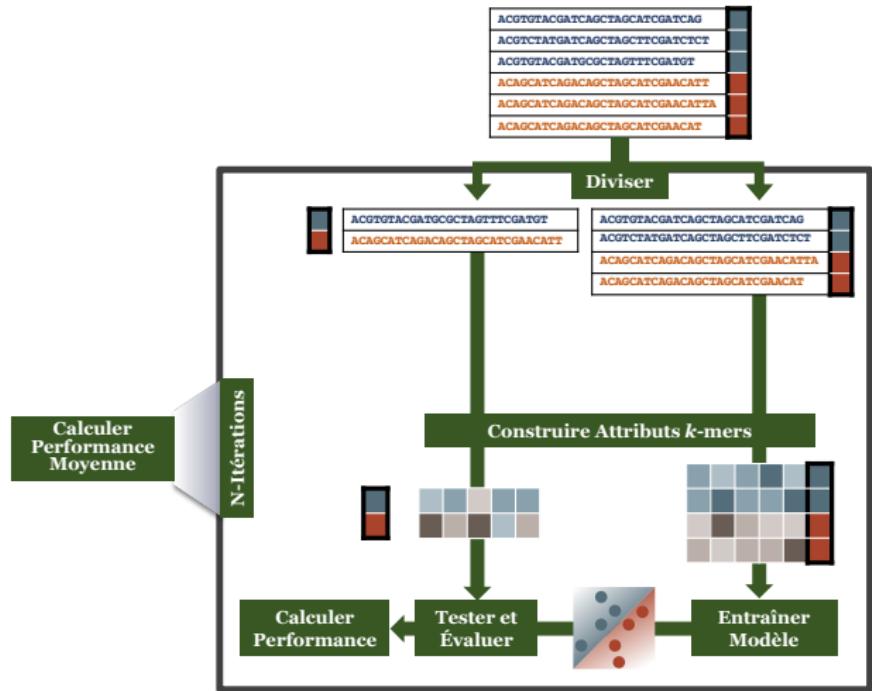
$$\min_{\mathbf{w}} \sum_{\mathbf{w}, c} \underbrace{L(\mathbf{w}, \mathbf{x}, c)}_{\text{Fonction de perte}} + \underbrace{\lambda R(\mathbf{w})}_{\text{Fonction de pénalité}}$$

# Cadre d'évaluation

- Type de classifieurs
  - Génératifs
  - Discriminatifs
- Inférence des paramètres  $\Theta_{\text{clf}}$ 
  - Maximum de vraisemblance
  - Inférence bayésienne
  - Régularisations L1 et L2
- Tâche de classification
  - Génotypage
  - Sous-typage
- Longueur des séquences
  - Complètes
  - Partielles
- Longueur des  $k$ -mers
  - 4 à 15

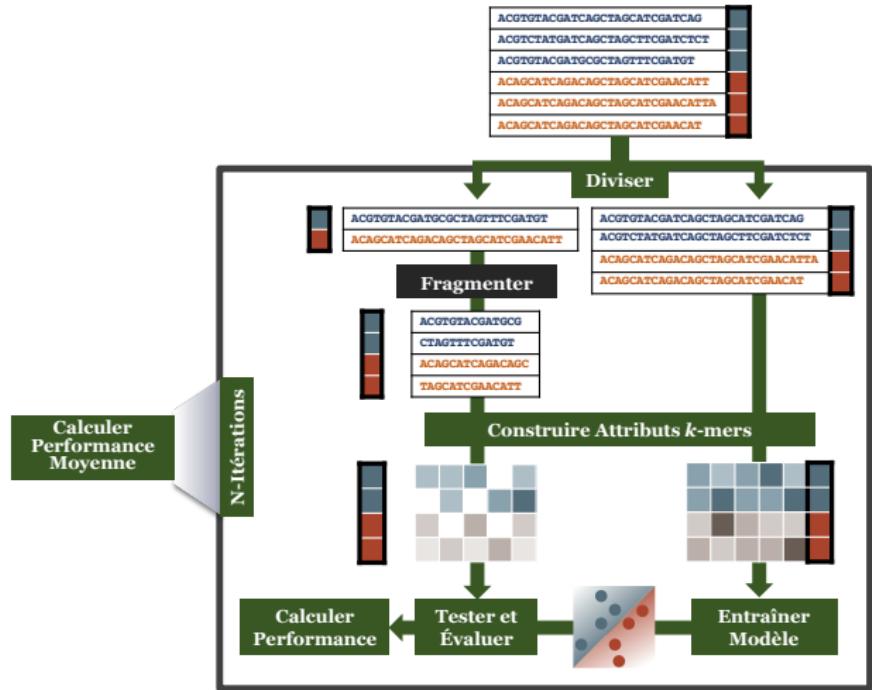
# Évaluation avec validation croisée

## Évaluation avec génomes complets



# Évaluation avec validation croisée

## Évaluation avec fragments génomiques



# Étude de cas : virus de l'hépatite C

- Données

- Génome : ARN simple brin sens-positif
- Taille :  $\sim 9600$  nucléotides
- **6 génotypes** (30% de divergence)
- **18 sous-types** (20% de divergence)

(Simmonds et al., 2005)

# Étude de cas : virus de l'hépatite C

- Données

- Génome : ARN simple brin sens-positif
- Taille :  $\sim 9600$  nucléotides
- **6 génotypes** (30% de divergence)
- **18 sous-types** (20% de divergence)

(Simmonds et al., 2005)

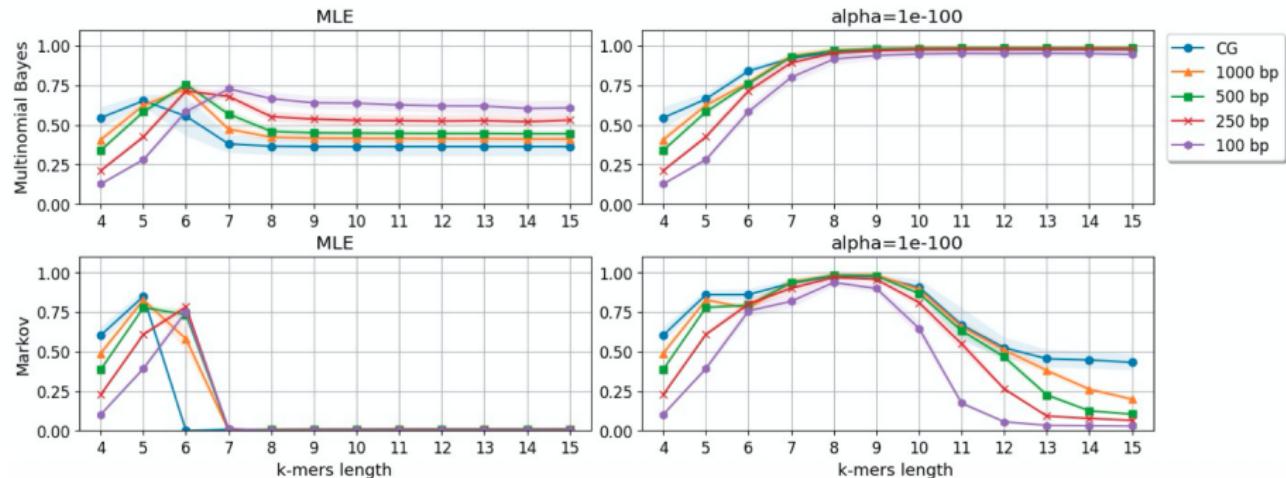
- Validation croisée à 5 itérations

- Mesures de performance

- Rappel
- Précision
- F-mesure

# Résultats et discussion

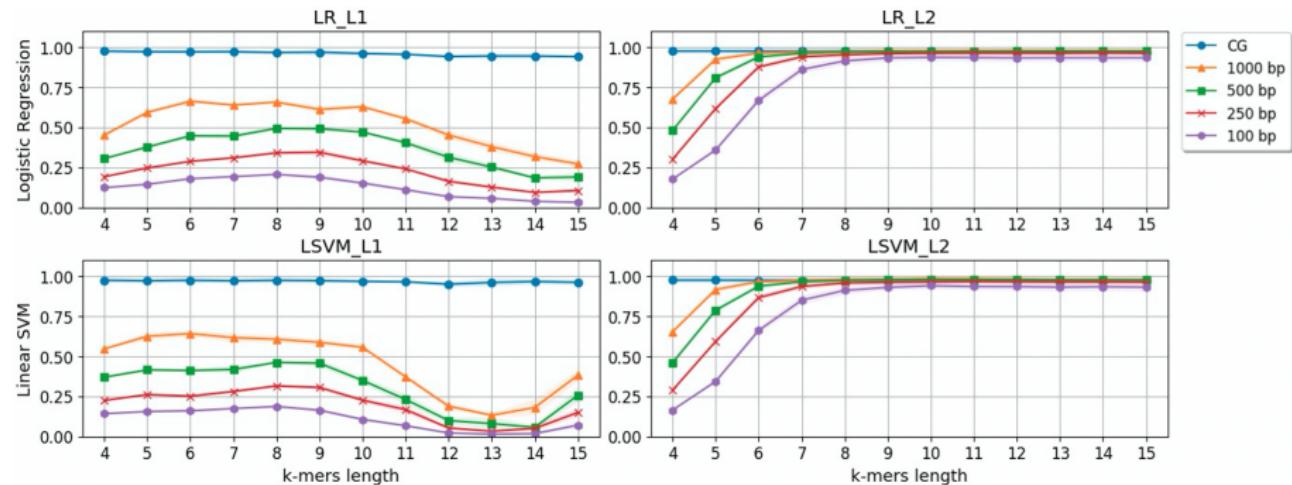
## Modèles génératifs : Bayes naïf multinomial et chaîne de Markov



- Estimation MLE
  - $k \leq 5$  : sur-ajustement
  - $k > 5$  : sous-ajustement
- Estimation bayésienne ( $\alpha = 1e - 100$ )
  - $k \leq 7$  : sur-ajustement
  - $k > 7$  : performance meilleure (sauf Markov quand  $k > 9$ )

# Résultats et discussion

## Modèles discriminatifs : Régression logistique et SVM linéaire



- Régularisation L1
  - $4 \leq k \leq 15$  : sur-ajustement

- Régularisation L2
  - $k \leq 7$  : sur-ajustement
  - $k > 7$  : performance meilleure

# Résultats et discussion

- Pas de supériorité entre les deux types de modèles linéaires
  - Performance dépend des variables de l'évaluation
  - Tous les modèles peuvent avoir F-mesure > 0.950 (sauf estimation MLE)
- Classification plus difficile aux niveaux taxonomiques bas
- Modèles avec matrice de poids dense sont convenables à classifier les fragments
  - Génératifs avec estimation bayésienne et discriminatifs L2
- Implémentation : `slm_kgenomvir`

## Objectif spécifique 2

# Modèle génératif bayésien variationnel pour l'estimation des paramètres évolutifs

- ★ Amine M. Remita et Abdoulaye Baniré Diallo (2022) EvoVGM: a Deep Variational Generative Model for Evolutionary Parameter Estimation.  
*The 13<sup>th</sup> ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*
- ★ <https://github.com/maremita/evoVGM>

# Inférence évolutive : rappel

## Étant donné :

- Alignement de séquences  $\mathbf{X}$

$\mathbf{x}^1$  ATGTGTGAACCTGTCGTG

$\mathbf{x}^2$  ATGA-TGAGCCTGTGGTC

$\mathbf{x}^3$  TTGAGTGAACCTG-GGTC

$\mathbf{x}^4$  TTGAGTGAACGTGTGGTC

## Inférer :

- Paramètres évolutifs  $\Theta_{\text{evo}}$ 
  - Phylogénie  $\tau$  et longueurs de branches  $\mathbf{t}$
  - Paramètres de modèle de substitution (K80, HKY85, GTR, etc.)
  - États ancestraux  $\mathbf{a}$
- Modéliser  $P(\Theta_{\text{evo}} | \mathbf{X})$

# Inférence Bayésienne

**Théorème de Bayes :**

$$\underbrace{p(\Theta | \mathbf{X})}_{A \text{ posteriori}} = \frac{\overbrace{p(\mathbf{X} | \Theta)}^{\text{Vraisemblance}} \overbrace{p(\Theta)}^{A \text{ priori}}}{\underbrace{p(\mathbf{X})}_{\text{Évidence}}}$$

# Inférence Bayésienne

**Théorème de Bayes :**

$$p(\Theta | \mathbf{X}) = \frac{p(\mathbf{X} | \Theta) p(\Theta)}{\underbrace{\int_{\Theta} p(\mathbf{X} | \Theta) p(\Theta) d\Theta}_{\text{Intractable}}}$$

# Inférence bayésienne computationnelle

Méthodes pour approximer la densité *a posteriori*  $p(\Theta | \mathbf{X})$  :

## ① Méthodes probabilistes

- Markov Chain Monte Carlo (MCMC)

(Metropolis et al., 1953; Hastings, 1970)

- **Inférence bayésienne variationnelle (IV)**

(Jordan et al., 1999; Bishop, 2006)

## ② Modèles génératifs profonds

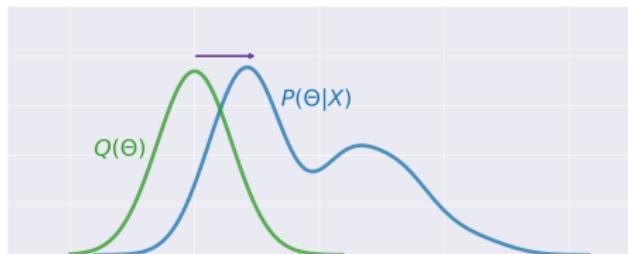
Optimisation par apprentissage

(Diederik P Kingma and Welling, 2014; Rezende, Mohamed, and Wierstra, 2014)

# Inférence variationnelle (IV)

- Se base sur **optimisation** rapide
- Utilise **distribution d'approximation**  $Q$  choisie à partir d'une famille tractable
- Minimise sa **divergence** avec l'*a posteriori* exact :

$$\hat{\phi} = \arg \min_{\phi} \text{KL}(Q_{\phi}(\Theta) \parallel P(\Theta | X))$$



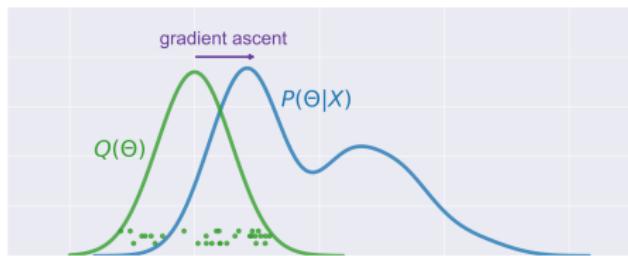
## IV: Limite inférieure de l'évidence (ELBO)

$$\begin{aligned}\mathcal{L}(\phi, \mathbf{X}) &= \mathbb{E}_{Q_\phi} [ p(\mathbf{X} | \Theta) ] - \text{KL}(Q_\phi(\Theta) \| P(\Theta)) \\ &\leq p(\mathbf{X})\end{aligned}$$

- Échantillonnage à partir de la densité approximative
- Mise à jour des paramètres de la densité approximative
  - Montée de coordonnées
  - Montée de gradient

⇒ S'adapte aux grandes données

⇒ Vraisemblance basée sur un modèle d'évolution explicite

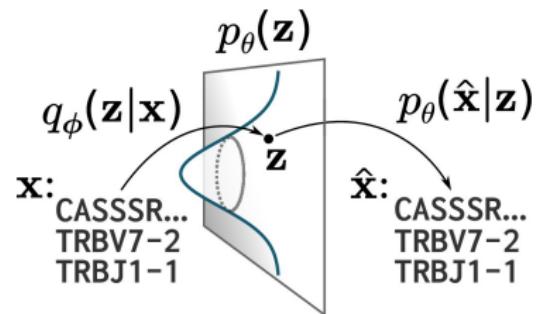
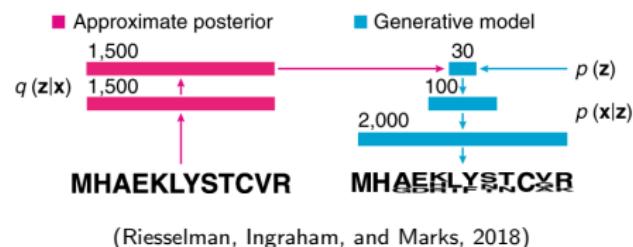


## IV en phylogénétique

- Inférence sur un **arbre fixe** (Fourment and Darling, 2019)
- Inférence des **longueurs de branches et taux d'évolution spécifiques au sites** (Dang and Kishino, 2019)
- Inférence simultanée d'**arbres avec longueurs de branches**  
(Zhang and Matsen IV, 2019; Zhang, 2020)
- Inférence simultanée d'**arbres, séquences ancestrales et longueurs de branches** (Koptagel et al., 2022)

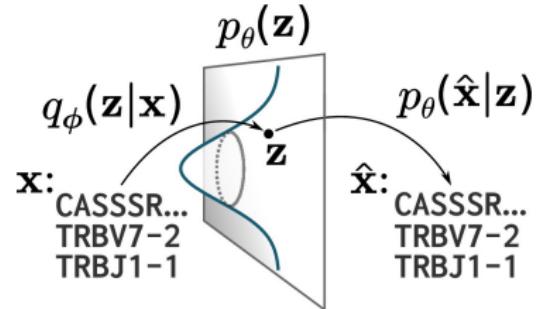
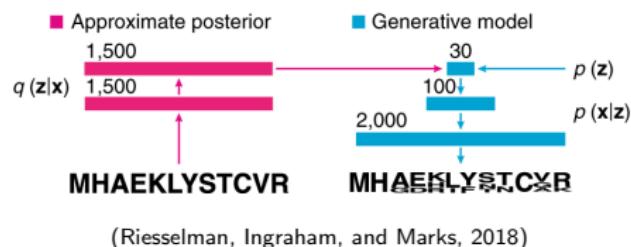
# Modèles variationnels génératifs profonds

- Encoder les données en variable latente  $\mathbf{Z}$ 
  - Auto-encodeurs variationnels  
(VAE, Diederik P. Kingma and Welling, 2019)
- $\mathbf{Z}$  est corrélée avec les relations évolutives sous-jacentes



# Modèles variationnels génératifs profonds

- Encoder les données en variable latente  $\mathbf{Z}$ 
  - Auto-encodeurs variationnels  
(VAE, Diederik P. Kingma and Welling, 2019)
- $\mathbf{Z}$  est corrélée avec les relations évolutives sous-jacentes
- + Efficaces pour trouver des motifs dans les données et des caractéristiques cachées
- + Adaptabilité aux grandes données (SGD)
- Absence de modèle évolutif explicite



# Modèle génératif variationnel des séquences évolutives

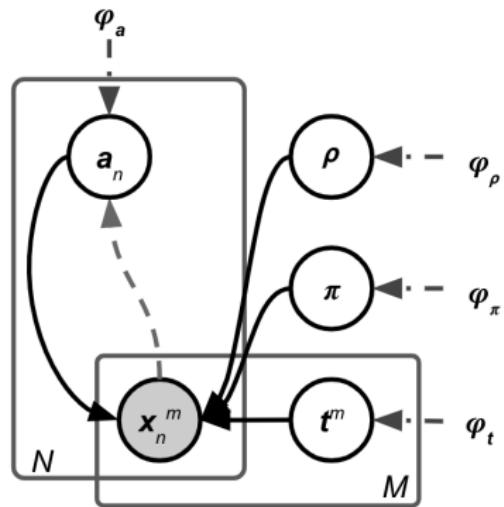
**Étant donnée :** Alignement de séquences  $\mathbf{X}$

## 1. Inférer :

- Parameters évolutifs  $\Theta_{\text{evo}}$

$$\approx Q_\phi(\Theta_{\text{evo}} | \mathbf{X}) = Q_\phi(\mathbf{a}, \mathbf{t}, \rho, \pi | \mathbf{X})$$

⇒ Encodeurs variationnels profonds



# Modèle génératif variationnel des séquences évolutives

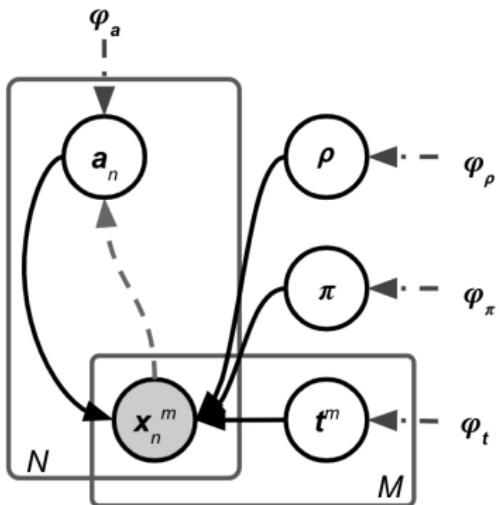
**Étant donnée :** Alignement de séquences  $\mathbf{X}$

## 1. Inférer :

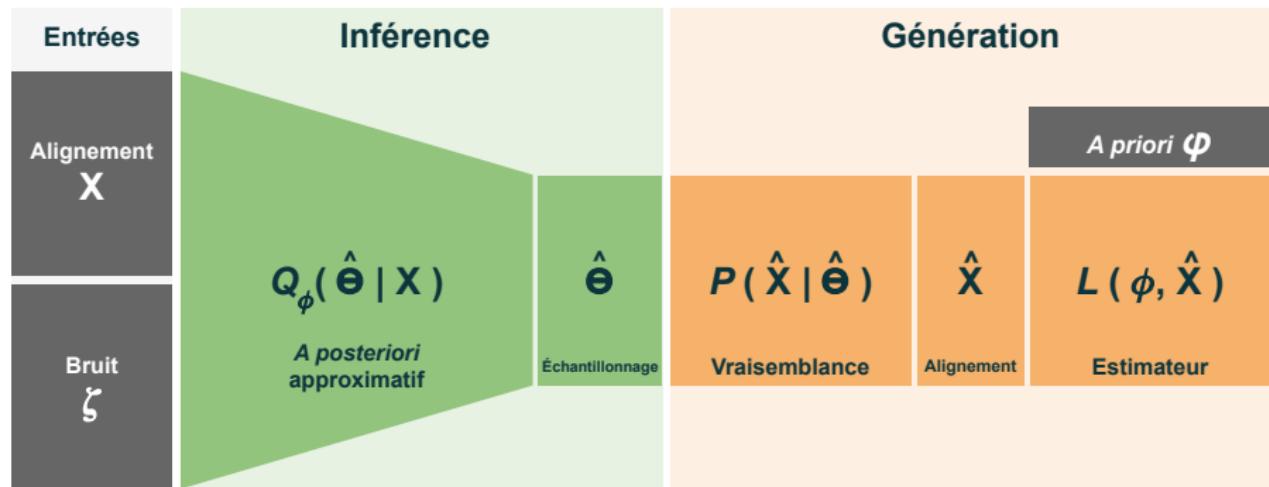
- Parameters évolutifs  $\Theta_{\text{evo}}$
- $$\approx Q_\phi(\Theta_{\text{evo}} | \mathbf{X}) = Q_\phi(\mathbf{a}, \mathbf{t}, \rho, \pi | \mathbf{X})$$
- $\Rightarrow$  Encodeurs variationnels profonds

## 2. Générer :

- Alignement de séquences  $\hat{\mathbf{X}}$
- $$\Rightarrow p(\mathbf{X}, \Theta_{\text{evo}}) = p(\mathbf{X} | \Theta_{\text{evo}}) p(\Theta_{\text{evo}})$$
- $\Rightarrow$  Modèle d'évolution explicite
- $\Rightarrow$  Parcours préfixe d'arbre

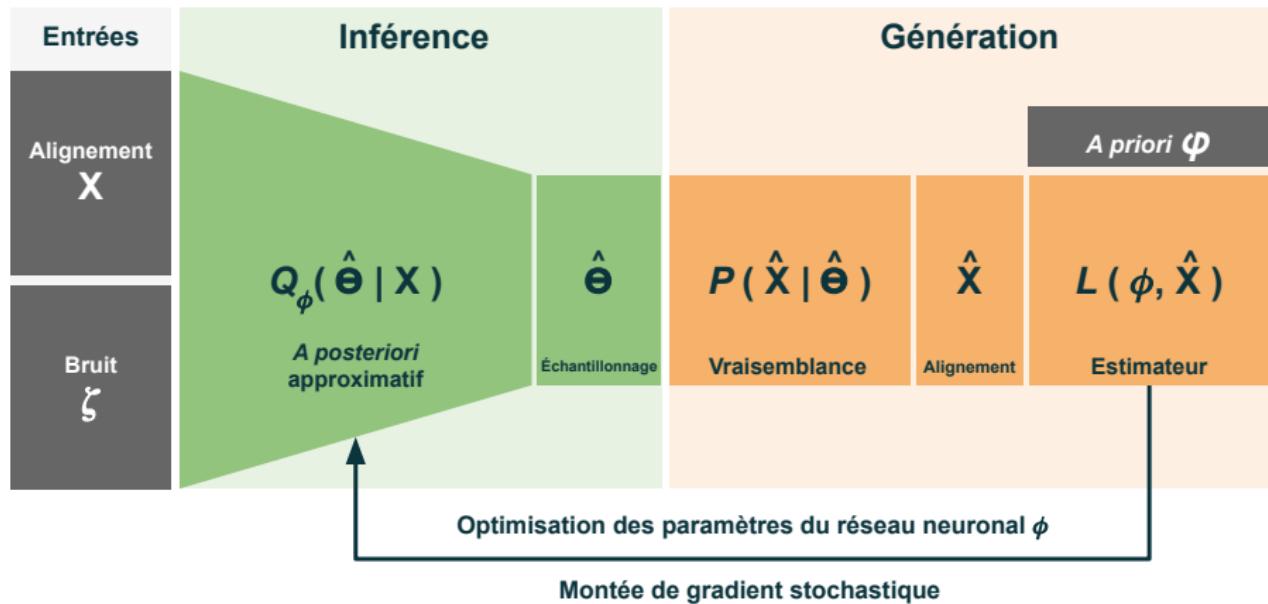


# EvoVGM : modèle génératif variationnel



# Ajustement des paramètres

$$\text{ELBO} : \mathcal{L}(\phi, \mathbf{X}) = \mathbb{E}_Q [ p(\mathbf{X} | \Theta) ] - \alpha_{\text{KL}} \text{KL}(Q(\Theta | \mathbf{X}) \| P(\Theta))$$



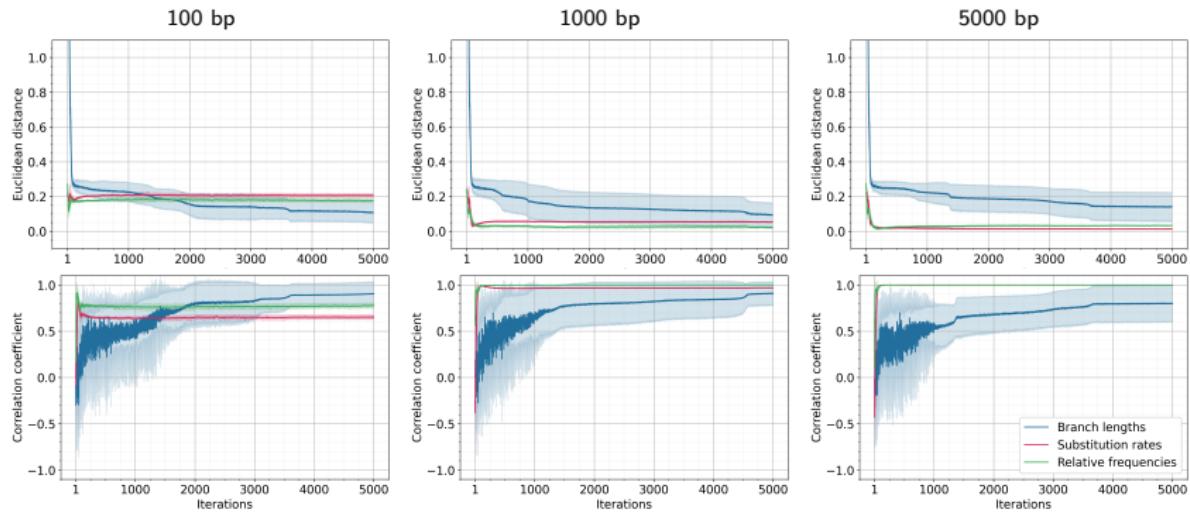
# Défis dans l'implémentation de EvoVGM

- Effondrement du postérieur  $\Rightarrow$  régularisation avec  $\alpha_{KL}$
  - Effondrement de certaines longueurs de branches
  - Dérive des états ancestraux vers un seul état
- $\Rightarrow$  Test de différents types de distributions d'approximation
- $\Rightarrow$  Réglage fin des hyperparamètres

# Évaluation d'EvoVGM

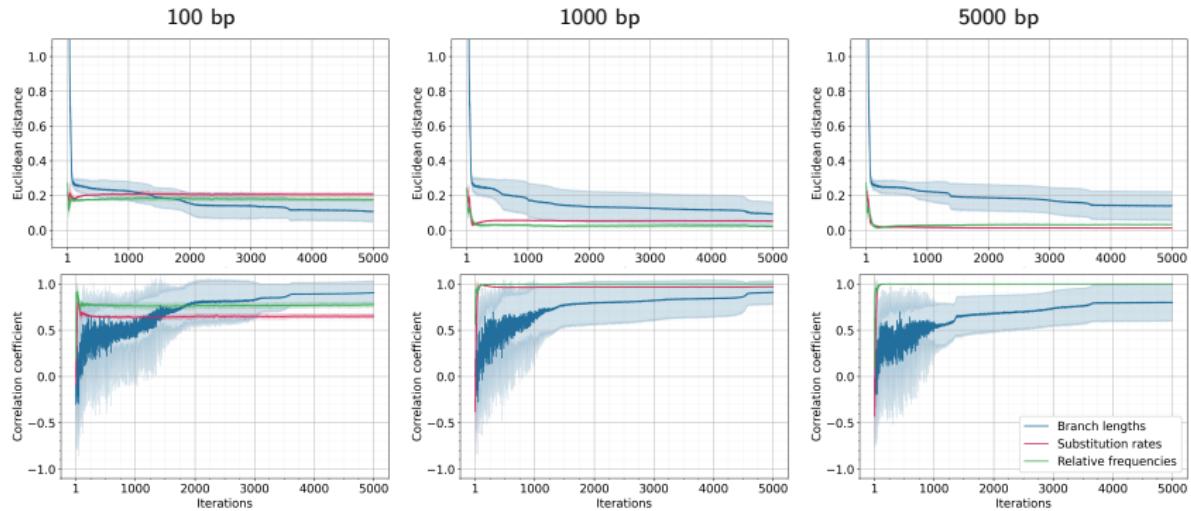
- *Consistance et efficacité*
  - Alignements de séquences **simulés**
  - Données d'**entraînement** et de **validation** ont été simulées en utilisant des *seeds* aléatoires différents.
- *Robustesse*
  - Alignement de séquences du **gène S du coronavirus** (Samson, Lord, and Makarenkov, 2022)
  - Comparaison avec **MrBayes** (Huelsenbeck and Ronquist, 2001)

# Estimation des paramètres sur données simulées



- EvoVGM implémenté avec le modèle de substitution GTR
- Alignements de 5 séquences de longueurs 100, 1000 et 5000 pb

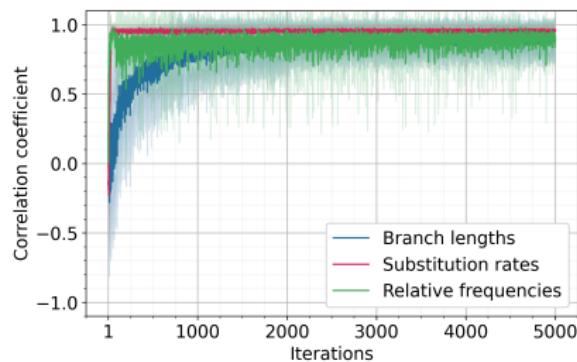
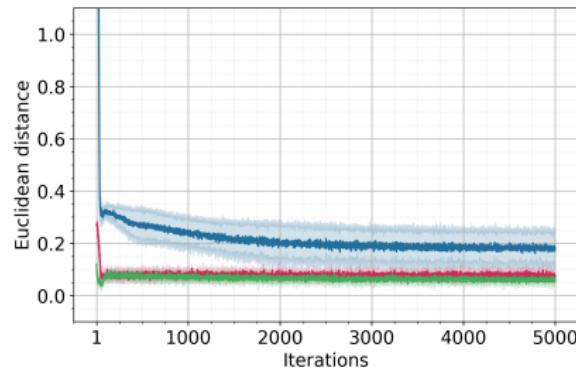
# Estimation des paramètres sur données simulées



- **Longueurs de branche** souffrent d'une **estimation à forte variance** et d'une **convergence lente**
- **Taux de substitution** et les estimations des **fréquences relatives** convergent **plus rapidement** avec **une faible variance**
- La précision est améliorée en utilisant un **nombre plus élevé de séquences** et des **alignements plus longs**

# Estimation des paramètres sur données réelle

Alignement de six séquences du gène S des coronavirus (3688 pb)



- Les estimations de **longueur de branche** diffèrent de celles de MrBayes avec une distance inférieure à 0.2 mais avec une **variance élevée**
- EvoVGM\_GTR estime les **taux de substitution** et les **fréquences relatives** avec des valeurs de **faible variance** et **plus proches** des estimations de MrBayes

# Limitations

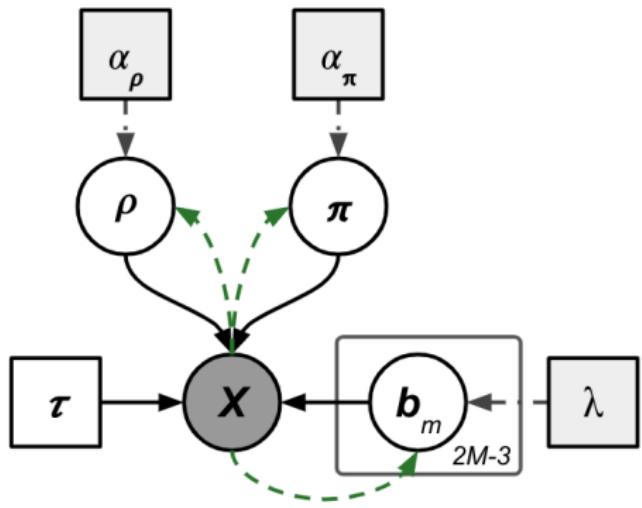
- Variance élevée dans l'estimation des longueurs de branches
  - ⇒ Utiliser la distribution composée Dirichlet-Gamma
- Améliorer l'inférence et la génération de séquences
  - ⇒ Considérer une topologie d'arbre binaire fixe
  - ⇒ Étudier l'influence des densités a priori sur l'inférence
  - ⇒ Permettre l'hétérogénéité entre les sites et les lignées

## Objectif spécifique 3

# Modèle bayésien variationnel profond pour l'inférence des paramètres phylogénétiques

- ★ Amine M. Remita, Golrokhsit Vitae et Abdoulaye Baniré Diallo (2023) Prior Density Learning in Variational Bayesian Phylogenetic Parameters Inference. *The 20<sup>th</sup> RECOMB Comparative Genomics Satellite Conference*. LNBI, volume 13883
- ★ <https://github.com/maremita/nnTreeVB>

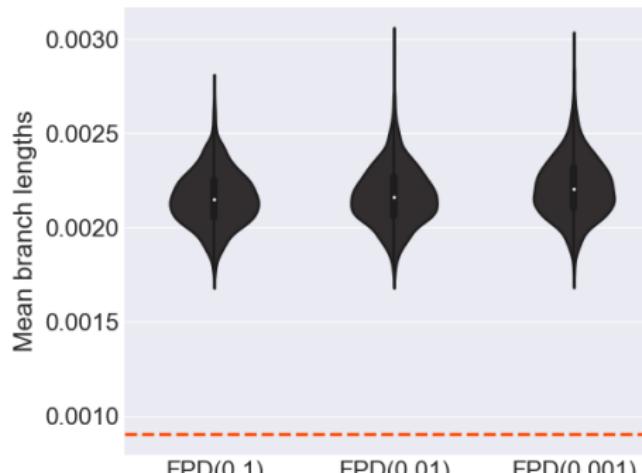
# VP-FPD: modèle variationnel avec des priors fixes



- Arbre fixe  $\tau$
- Modèle de substitution GTR  $\{\rho, \pi\}$
- $\omega = \{\mathbf{b}, \rho, \pi\}$
- $q_{\phi_i}(\omega_i) = \text{Normal}(\omega_i; \mu_i, \sigma_i)$
- Hyperparamètres *d'a priori* :
  - $\{\lambda, \alpha_\rho, \alpha_\pi\}$
  - Prédéfinis

# Rigidité de la densité *a priori*

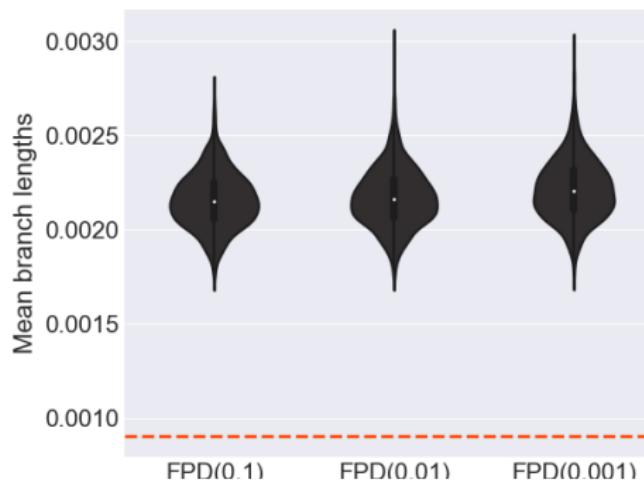
- Densité *a priori* fixée
  - Biaise l'approximation
  - Induit des probabilités élevées  
(Huelsenbeck, Larget, et al., 2002; Huelsenbeck and Ronquist, 2005; Nascimento, Reis, and Yang, 2017; Fabreti and Höhna, 2022)
- Sensibilité à la densité *a priori*
  - petits ensembles de données
  - Séquences similaires
  - Corrélation des paramètres



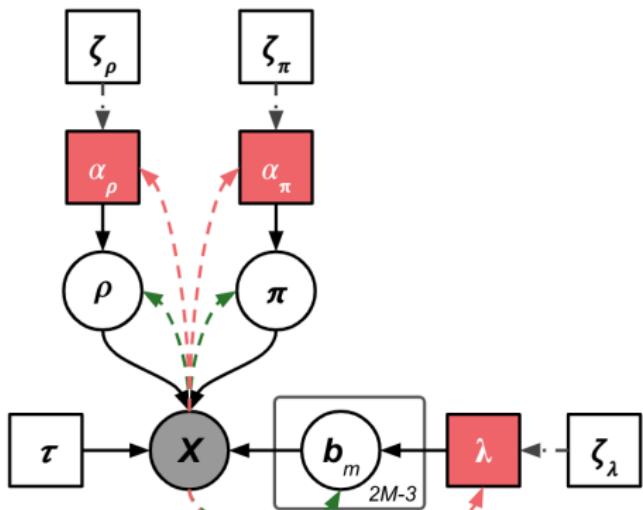
- Moyenne des longueurs de branches simulées : **0.001**

# Rendre l'*a priori* plus souple

- Améliorer l'approximation postérieure variationnelle
  - Rendre l'*a priori* **plus flexible**
  - Tirer parti de la structure du ELBO

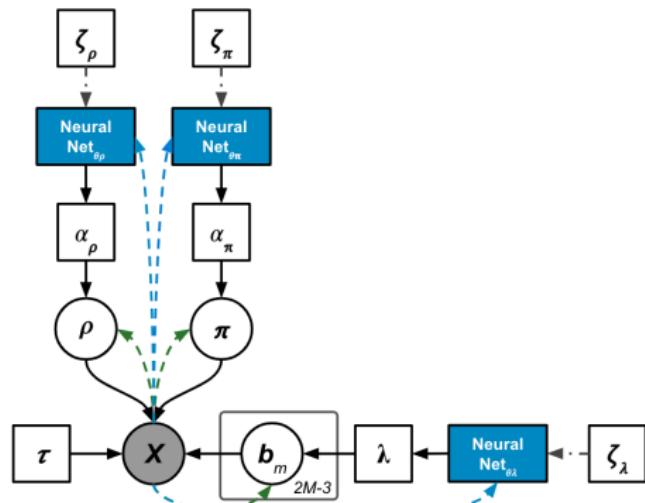


## VP-LPD: priors apprenables



- Arbre fixe  $\tau$
  - Modèle de substitution GTR  $\{\rho, \pi\}$
  - $\omega = \{b, \rho, \pi\}$
  - $q_{\phi_i}(\omega_i) = \text{Normal}(\omega_i; \mu_i, \sigma_i)$
  - Paramètres *a priori* adaptables :
- 
- $\theta = \{\lambda, \alpha_\rho, \alpha_\pi\}$
  - Optimisation avec montée du gradient
  - $\zeta_i$  Initialisation par échantillonnage uniforme

# VP-NPD: priors apprenables avec réseaux de neurones



- Arbre fixe  $\tau$
- Modèle de substitution GTR  $\{\rho, \pi\}$
- $\omega = \{\mathbf{b}, \rho, \pi\}$
- $q_{\phi_i}(\omega_i) = \text{Normal}(\omega_i; \mu_i, \sigma_i)$
- Paramètres *a priori* adaptables :
  - Générés avec des réseaux de neurones
  - $\theta = \{\text{poids, biais}\}$
  - Optimisation avec montée du gradient
  - $\zeta_i$  Initialisation par échantillonnage uniforme.

# Cadre d'évaluation

- Comparaison de VP-FPD, VP-LPD et VP-NPD
- Estimation des paramètres phylogénétiques
  - ① Longueurs de branches (modèle JC69)
  - ② Taux de substitutions (modèle GTR)
  - ③ Fréquences relatives (modèle GTR)

# Cadre d'évaluation

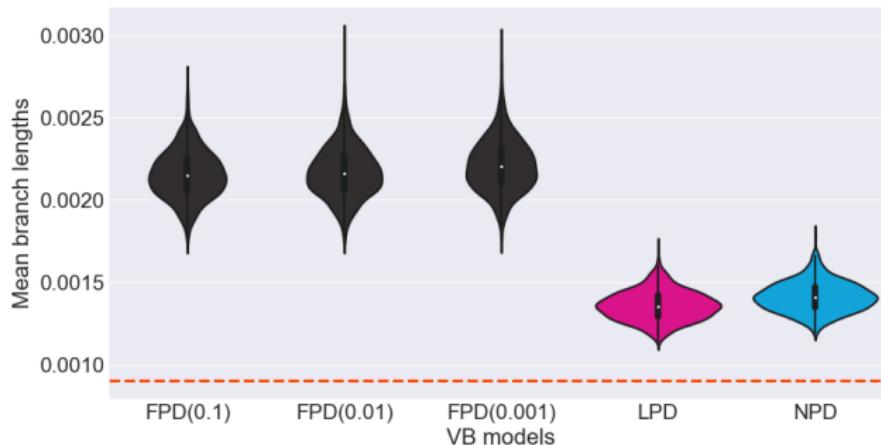
- Comparaison de VP-FPD, VP-LPD et VP-NPD
- Estimation des paramètres phylogénétiques
  - ① Longueurs de branches (modèle JC69)
  - ② Taux de substitutions (modèle GTR)
  - ③ Fréquences relatives (modèle GTR)
- Données simulées
  - Nombre de taxa: 16, **64** séquences
  - Longueur des alignements: **1000**, 5000 bp
  - Bootstrap : 100 jeux de données × 10 entraînements des modèles

# Cadre d'évaluation

- Comparaison de VP-FPD, VP-LPD et VP-NPD
- Estimation des paramètres phylogénétiques
  - ① Longueurs de branches (modèle JC69)
  - ② Taux de substitutions (modèle GTR)
  - ③ Fréquences relatives (modèle GTR)
- Données simulées
  - Nombre de taxa: 16, **64** séquences
  - Longueur des alignements: **1000**, 5000 bp
  - Bootstrap : 100 jeux de données × 10 entraînements des modèles
- Mesures de performance
  - Distance et corrélation entre
    - Vecteurs des paramètres réels et estimés

# Estimation des longueurs des branches

- Alignement : **64** taxa / **1000** bp
- Modèle : **JC69**
- Moyenne des branches : **0.001**



# Estimation des longueurs des branches

- VP-FPD utilise une densité *a priori* exponentielle avec moyenne **0.1** sur les branches

Moyenne LB	0.001	0.01	0.1
Séquences	plus similaires	↔	plus divergente

	Distance	Corrélation
	Moyenne LB	0.001
VP-FPD	0.0240(0.00)	0.4636(0.10)
VP-LPD	0.0152(0.00)	<b>0.5783(0.09)</b>
VP-NPD	<b>0.0151(0.00)</b>	0.5732(0.09)
	Moyenne LB	0.01
VP-FPD	0.0506(0.01)	0.8966(0.03)
VP-LPD	<b>0.0479(0.01)</b>	<b>0.9058(0.02)</b>
VP-NPD	<b>0.0479(0.01)</b>	0.9055(0.02)
	Moyenne LB	0.1
VP-FPD	<b>0.1773(0.02)</b>	0.9806(0.00)
VP-LPD	0.1790(0.02)	<b>0.9810(0.00)</b>
VP-NPD	0.1791(0.02)	<b>0.9810(0.00)</b>

# Estimation des taux de substitution

- VP-FPD utilise une densité *a priori* Dirichlet avec des hyperparamètres uniformes sur les taux de substitution

$\kappa$       0.25      1      4  
 Mutations   plus transversions    $\longleftrightarrow$    plus transitions

	Distance	Corrélation
$\kappa$ 0.25		
VP-FPD	0.0185(0.01)	0.9950(0.00)
VP-LPD	<b>0.0175(0.01)</b>	<b>0.9955(0.00)</b>
VP-NPD	<b>0.0175(0.01)</b>	0.9954(0.00)
$\kappa$ 1		
VP-FPD	0.0190(0.01)	0.5226(0.40)
VP-LPD	0.0181(0.01)	0.5360(0.40)
VP-NPD	<b>0.0178(0.01)</b>	<b>0.5407(0.39)</b>
$\kappa$ 4		
VP-FPD	0.0182(0.01)	0.9980(0.00)
VP-LPD	0.0176(0.01)	<b>0.9981(0.00)</b>
VP-NPD	<b>0.0173(0.01)</b>	<b>0.9981(0.00)</b>

# Discussion

- Inférence variationnelle pour l'estimation des paramètres phylogénétiques avec un **arbre fixe**
- Les modèles VP avec des priors fixés sont sujets aux biais
- Donner de la **souplesse** aux densités *a priori*
  - Deux stratégies d'optimisation : **VP-LPD** and **VP-NPD**
  - Apprendre les paramètres *a priori* en utilisant la montée de gradient.

# Discussion

- Inférence variationnelle pour l'estimation des paramètres phylogénétiques avec un **arbre fixe**
- Les modèles VP avec des priors fixés sont sujets aux biais
- Donner de la **souplesse** aux densités *a priori*
  - Deux stratégies d'optimisation : VP-LPD and VP-NPD
  - Apprendre les paramètres *a priori* en utilisant la montée de gradient.
- À considérer dans le futur
  - Capturer les corrélations entre les paramètres
  - Évaluer d'autres architectures de réseaux neuronaux
  - *A priori* adaptable des topologies d'arbre

# Tableau récapitulatif

	slm_kgenomvir	EvoVGM	nnTreeVB
Inférence	$P(\Theta_{\text{clf}}   \mathbf{C}, \mathbf{X})$	$P(\Theta_{\text{evo}}   \mathbf{X})$	$P(\Theta_{\text{evo}}   \mathbf{X}, \tau)$
Génération	$P(\mathbf{C}, \mathbf{X}, \Theta_{\text{clf}})$	$P(\mathbf{X}, \Theta_{\text{evo}})$	$P(\mathbf{X}, \Theta_{\text{evo}}   \tau)$
Prédiction	$P(\mathbf{C}   \mathbf{X}, \Theta_{\text{clf}})$	-	-

# Directions futures

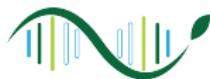
Hétérogénéité des sites et branches	Dépendance du contexte	Inférence d'arbre et apprentissage du prior	Classifieur évolutivement informé
Inférence $P(\Theta_{\text{evo}}, \mathbf{r}   \mathbf{X})$	$P(\Theta_{\text{evo}}   \mathbf{X}, \mathbf{h})$	$P(\Theta_{\text{evo}}, \tau   \mathbf{X})$	$P(\Theta_{\text{clf}}, \Theta_{\text{evo}}   \mathbf{C}, \mathbf{X})$
Génération $P(\mathbf{X}, \Theta_{\text{evo}}, \mathbf{r})$	$P(\mathbf{X}, \Theta_{\text{evo}}   \mathbf{h})$	$P(\mathbf{X}, \Theta_{\text{evo}}, \tau)$	$P(\mathbf{C}, \mathbf{X}, \Theta_{\text{clf}}, \Theta_{\text{evo}})$
Prédiction -	-	-	$P(\mathbf{C}   \mathbf{X}, \Theta_{\text{clf}}, \Theta_{\text{evo}})$

# Liste des publications

- [OS1] Remita et Diallo (2019) Statistical Linear Models in Virus Genomic Alignment-free Classification: Application to Hepatitis C Viruses. Dans *IEEE International Conference on Bioinformatics and Biomedicine*. BIBM 2019. San Diego, CA, USA, pp. 474-481
- [OS2] Remita et Diallo (2021) Evolutionary-based variational generative models for biological sequences. Dans *Joint meeting of the 29th Annual Conference on Intelligent Systems for Molecular Biology and the 20th European Conference on Computational Biology*. ISMB/ECCB 2021. MLCSB COSI: Machine Learning in Computational and Systems Biology
- [OS2] Remita et Diallo (2021) Sequence evolution modelling using variational self-supervised learning framework. Dans *The 2021 Society for Molecular Biology & Evolution meeting*. SMBEv2021
- [OS2] Remita et Diallo (2022) Toward Inferring Ancestral States and Evolutionary Parameters using a Variational Generative Model for Multiple Sequence Alignments. Dans *The 2022 ICML Workshop on Computational Biology*. Baltimore, Maryland, USA.
- [OS2] Remita et Diallo (2022) EvoVGM: a Deep Variational Generative Model for Evolutionary Parameter Estimation. Dans *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Association for Computing Machinery*. BCB '22. Northbrook, Illinois
- [OS3] Remita, Vitae et Diallo (2023) Prior Density Learning in Variational Bayesian Phylogenetic Parameters Inference. Dans *Proceedings of the 20th RECOMB Comparative Genomics Satellite Conference*. LNBI, volume 13883. Istanbul, Türkiye.
- [OS3] Remita, Vitae et Diallo (2023) Prior Density Learning in Variational Bayesian Phylogenetic Parameters Inference. Dans *The 15th Great Lakes Bioinformatics conference*. GLBIO. Montreal, Canada.
- [OS3] Remita et Diallo (2023) nnTreeVB: a neural network-based variational Bayesian framework for phylogenetic parameter inference. Dans *The 2023 Society for Molecular Biology & Evolution meeting*. SMBE23. Ferrara, Italy

# Remerciements ❤

- **Pr. Abdoulaye Baniré Diallo**
- Membres du jury
- Membres du laboratoire bioinformatique à l'UQAM
- Amis, famille et collègues au Canada et en Algérie



LABORATOIRE  
BIOINFORMATIQUE

UQÀM | Université du Québec  
à Montréal



Fonds de recherche  
Nature et  
technologies

Québec 

∞  
Calcul Québec