



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Soutenance de thèse

Monsieur Massimo MAIOLO

Titulaire d'un *Msc in Engineering* de la *Scuola Universitaria
Professionale della Svizzera Italiana (SUPSI)*, Suisse

Soutiendra en vue de l'obtention du grade de
Doctorat ès sciences de la vie (PhD)
de l'Université de Lausanne

sa thèse intitulée :

Progressive multiple sequence alignment with indel evolution

Directeur de thèse :

Monsieur le Professeur Christophe DESSIMOZ

Cette soutenance aura lieu le

Mardi 12 février 2019 à 10h30

Salle 321, Amphipôle, quartier UNIL-Sorge, 1015 Lausanne

L'entrée est publique

Prof. Niko GELDNER
Directeur de l'Ecole Doctorale

Progressive Multiple Sequence Alignment with Indel Evolution

Massimo Maiolo

Département de biologie computationnelle
Université de Lausanne

Le développement de la phylogénétique moléculaire et la disponibilité croissante de données moléculaires ont permis une compréhension sans précédent de l'évolution à des niveaux allant des gènes aux génomes offrant un grand potentiel pour découvrir de nouveaux mécanismes cellulaires et des relations auparavant inconnues. Une étude intégrée des phénomènes évolutifs peut révéler leurs mécanismes sous-jacents qui sont bien sûr les résultats de l'interaction de la phylogénie, de la forme, de la fonction et de l'évolution moléculaire. La pensée évolutive fournit un cadre puissant pour démêler les mécanismes biologiques sous-jacents, révélant ainsi et comprenant les modèles et les processus de la microévolution.

Une partie intégrante de la phylogénétique contemporaine est le développement de modèles mathématiques et de solutions algorithmiques efficaces pour s'attaquer aux problèmes de calcul de grande complexité liés à la construction de scénarios évolutifs, à l'inférence de modèles de coévolution de molécules, de voies, de systèmes de régulation et d'espèces et à l'assemblage de données sur les séquences et les arbres. Une hypothèse phylogénétique robuste est étroitement liée à la qualité de l'alignement des séquences. L'alignement de séquence dans une analyse phylogénétique consiste à évaluer les homologies : les résidus alignés doivent être homologues au sens évolutif. Ces homologies sont des hypothèses sur des événements évolutifs inconnus plutôt que des observations empiriques. Idéalement, la phylogénie et l'alignement devraient être déduits conjointement (les deux problèmes d'estimation sont NP-durs).

Dans cette thèse, nous présentons pour la première fois une méthode MSAs progressif selon un modèle évolutif explicite de substitution, d'insertion et de suppression (ensemble appelé "indel"). La modélisation de l'évolution de l'indel à l'aide du processus de Poisson Indel réduit la complexité de calcul d'exponentielle à linéaire (Bouchard-Côté & Jordan 2013). Notre nouvelle méthode, permet d'inférer des MSAs progressifs sous un modèle évolutif explicite d'indels. En ce qui concerne l'inférence de MSA, il s'agit de la première avancée méthodologique majeure depuis la mise au point du PRANK qui tient compte de la phylogénie.

A chaque nœud interne de l'arbre-guide, nous alignons les deux MSAs des nodes-enfants par l'approche du maximum de vraisemblance. Cela n'a jamais été fait auparavant. Notre méthode est hautement compétitive par rapport à la méthode PRANK, qui tient compte de la phylogénie, et surpasse de loin les autres logiciels d'alignement populaires. En plus de la haute qualité, notre méthode permet d'inférer des interstices significatifs du point de vue phylogénétique comme alternative à PRANK, tout en produisant des alignements d'une longueur similaire. Notre algorithme est d'une complexité temporelle polynomiale et peut être parallélisé, ce qui le rend adapté aux analyses à grande échelle.

Jusqu'à présent, seules les méthodes bayésiennes mettent en œuvre des modèles évolutifs d'indel, mais se limitent à quelques séquences en raison des algorithmes d'échantillonnage MCMC très coûteux. Avec cette thèse, nous ouvrons de nouvelles voies dans l'analyse des séquences génomiques. Par exemple, des caractéristiques plus réalistes du modèle pourraient être mises en œuvre, ce qui permettrait de faire varier les taux d'indentation le long des séquences et au fil du temps. De plus, en incluant l'estimation de l'incertitude d'alignement, il sera possible d'étendre notre outil pour inclure l'estimation conjointe des MSA et des arbres dans un cadre fréquentiste sous un modèle unique explicite et solide de l'évolution des indel.