Détection phylogénétique de sites protéiques associés à un phénotype à l'échelle génomique

Louis Duchemin

Résumé

Les différences de phénotypes — les caractéristiques physiques, physiologiques et fonctionnelles — entre espèces sont une manifestation de variations qui se sont produites dans leur génome, à l'échelle moléculaire. Ces changements à long terme sont la conséquence de l'interaction entre processus de différentes natures. La mutation, communément considérée comme aléatoire, est le moteur initial de la diversification, et se produit à l'échelle d'un individu. Au fil des générations, un nouveau variant génétique se diffuse au sein d'une population sous l'action combinée d'un processus non-adaptatif, la dérive génétique, et de la sélection qui favorise ou non sa transmission selon l'avantage reproductif que ce variant procure. L'adaptation du vivant à un environnement changeant émerge de la combinaison de ces processus, à partir de laquelle son immense diversité se déploie.

Les espèces actuelles, et donc leurs génomes, partagent une histoire commune de par leur descendance d'une même espèce ancestrale, qui s'est séparée au fil de l'accumulation de divergences entre populations. En associant les séquences génomiques issus d'une même séquence ancestrale, et en examinant leur divergence, il est possible d'interpréter les traces laissées par leur histoire évolutive pour la reconstruire en partie. Parmi les événements de modification du génome, je m'intéresse au cas des substitutions, c'est à dire des remplacements ponctuels à une position, au sein des gènes codants pour des protéines, dont la structure et la fonction peut en être modifiée et donc avoir un effet adaptatif. En confrontant le signal porté par ces substitutions à l'histoire d'un trait phénotypique on peut tenter de déceler une corrélation entre l'histoire évolutive d'un site codant et celle du phénotype. L'identification de telles corrélations pourrait être le signal qu'une position génotypique est impliquée dans l'émergence ou le maintien du phénotype considéré, et plus largement témoigner de son implication dans l'adaptation d'une espèce à un environnement donné.

De nombreux modèles du processus de substitution basés sur ce genre d'approches comparatives existent déjà et sont largement utilisés pour construire et améliorer nos connaissances de l'évolution moléculaire. Il est toutefois difficile de les appliquer à l'échelle génomique pour effectuer une détection systématique des sites associés à un phénotype, du fait de la quantité de données que cela représente et de la limitation de la puissance de calcul existante. Dans cette thèse, je cherche à proposer une solution pour permettre ce genre d'analyse à large échelle à moindre coût en temps, tout en préservant la qualité des prédictions obtenues.

Après des premières tentatives infructueuses d'adapter des modèles linéaires utilisés en GWAS à l'échelle des populations pour étudier les associations génotype-phénotype, pour les appliquer à l'échelle inter-espèces, j'ai identifié une approche qui semble constituer une solution satisfaisante. Celle-ci se base sur un modèle d'évolution des séquences protéiques — le produit de la traduction des séquences d'ADN — publié précédemment, mais dont le potentiel n'avait pas été bien reconnu. J'ai montré, sur la base de simulations, que l'implémentation que nous avons faite de ce modèle permet de déceler des changements dans la dynamique de substitution en association avec des variations du phénotype aussi bien que plusieurs modèles plus complexes et plus coûteux en calculs. Bien qu'elle ne soit peut-être pas plus rapide que d'autres implémentations de modèles phylogénétiques, ce qu'il faudrait évaluer, elle apparaît comme la plus rapide des méthodes dites "à profils" qui permettent d'estimer une direction pour la sélection.

Une partie de cette thèse est consacrée à détailler cette méthode, que nous appelons Pelican, son modèle, son implémentation et quelques unes de ses limites. Une stratégie alternative pour l'estimation des paramètres du modèle, en déportant les calculs sur GPU pour exploiter leur capacité

de parallélisme, est aussi explorée pour tenter d'améliorer la vitesse des analyses. J'ai également proposé une extension du modèle basée sur des phénotypes continus, et non plus catégoriels. Celle-ci demande encore davantage de travail pour évaluer sa validité. Enfin, j'ai cherché à identifier une manière de prédire les gènes associés à un phénotype à partir des prédictions individuelles réalisées à chacune des positions de leur séquence. C'est un objectif difficile à atteindre, du fait de problèmes statistiques inhérents à la méthode, mais j'ai identifié une approche qui permet d'exploiter les prédictions par sites avec une puissance suffisante, bien qu'elle puisse manquer de robustesse dans certains cas.

Afin de valider notre approche sur des données empiriques, je l'ai appliquée à des alignements de gènes de mammifère pour identifier des sites et des gènes associés à divers phénotypes discrets. Les prédictions obtenues, comparées aux annotations et à la littérature existantes, suggèrent que la méthode est capable d'identifier des sites associés au trait considéré de manière relativement fiable. Le résultat de ce travail est l'implémentation logicielle de Pelican, qui bien qu'elle soit encore à un stade précoce, propose une solution pour détecter des associations génotype-phénotype inter-espèces à l'échelle génomique.