Résumé

L'évolution moléculaire vise à caractériser les mécanismes à l'œuvre dans l'évolution des séquences, régie par un processus stochastique dont les principaux composants sont la mutation, la sélection et la dérive génétique. À long terme, ce processus stochastique se traduit par une histoire d'événements de substitutions le long des arbres d'espèces, induisant des motifs complexes de divergence moléculaire entre les espèces. En analysant ces divergences, les modèles de codons phylogénétiques visent à capturer les paramètres intrinsèques de l'évolution. Dans ce contexte, cette thèse s'est concentrée sur les modèles à codons phylogénétiques et sur la modélisation de l'interaction entre la mutation, la sélection et la dérive génétique dans les séquences d'ADN codant pour des protéines. Parce que la composition de ces séquences ne reflète pas le processus de mutation sous-jacent, mais son filtrage par sélection au niveau des acides aminés, une modélisation minutieuse est nécessaire pour démêler la mutation et la sélection. Ainsi, j'ai développé un modèle d'inférence phylogénétique dans lequel différents taux d'évolution donnent une représentation précise de la manière dont la mutation et la sélection s'opposent à l'équilibre. Deuxièmement, l'équilibre entre mutation et sélection est arbitré par la dérive génétique, qui est médiée par la taille efficace de la population, et ses changements le long d'une phylogénie peuvent être déduits des motifs de substitutions le long des lignées. J'ai ainsi développé un deuxième modèle d'inférence, reconstituant à la fois le paysage de fitness en chaque site, les tendances à long terme de taille efficace de population et les changements de taux de mutation le long de la phylogénie. Ce cadre bayésien a été testé sur des données simulées puis appliqué à des données empiriques. Les estimations de la variation de taille efficace de population correspondent à la direction attendue de la corrélation avec les traits d'histoire de vie ou les variables écologiques, bien que l'ampleur de la variation de la taille efficace de population estimée soit étroite. Afin de comprendre cette variation étroite de la taille efficace de population estimée, j'ai finalement développé un modèle théorique décrivant comment les changements à la fois de taille efficace de population ou du niveau d'expression de la protéine se traduisent par un changement du taux de substitution, sous l'hypothèse que les protéines sont sous sélection directionnelle pour maximiser leur stabilité conformationnelle. Cette réponse est déterminée en fonction des paramètres moléculaires de la biophysique des protéines, et implique une faible réponse du taux de substitution aux changements de niveau d'expression ou de taille efficace de population dans ce contexte. Ce travail démontre que les hypothèses faites sur la structure du paysage de fitness ont une importance critique sur la sensibilité des changements de vitesse d'évolution à des changements de variables écologiques ou moléculaires. Réciproquement, les observations empiriques des motifs de substitutions en réponse à des changements de variables moléculaires ou écologiques nous informent sur la structure sous-jacente du paysage de fitness. En se basant sur l'équilibre mutation-sélection et en intégrant explicitement la taille efficace de population, ce travail présente aussi un cadre conceptuel permettant de relier phylogénie et génétique des populations, dont certaines pistes d'unifications sont envisagées.

Résumé étendu

La théorie neutre de l'évolution a influencé notre compréhension de la génétique des populations et de l'évolution moléculaire. Au-delà des disputes et des controverses entre neutralisme et sélectionnisme, le consensus actuel est de considérer l'évolution des séquences génétiques comme un processus stochastique combinant mutation, sélection et dérive génétique. Les mutations sont source de diversité génétique. La sélection, quant à elle filtre cette diversité. Enfin, l'équilibre entre mutation et sélection est arbitré par la dérive génétique, déterminé par la taille efficace de population (N_e). Sur la longue durée évolutive, mutation, sélection et dérive génétique résultent en une accumulation de substitutions ponctuelles entre les espèces, qui dans les séquences codantes peuvent être soit synonymes, soit non synonymes. S'appuyant ainsi sur ces différences interspécifiques, telles qu'observées dans les alignements multiples de séquences d'ADN codant pour des protéines, l'objectif des modèles à codons phylogénétiques est de mieux caractériser et quantifier les processus mutationnels et sélectifs et de mieux comprendre leur articulation. Les modèles à codons sont toujours un domaine de recherche actif et se scindent en deux philosophies différentes. D'un côté, les modèles phénoménologiques visent à capturer l'effet net de la sélection s'exerçant sur toutes les mutations non synonymes au sein de la protéine, à travers un seul paramètre. De l'autre côté, des approches mécanistes ont pour objectif de capturer l'effet de la sélection sur chaque mutation non synonyme prise individuellement, ce qui requiert de modéliser explicitement le paysage du fitness sous-jacent. En l'état, cependant, de nombreuses questions restent ouvertes et les modèles actuels, qu'ils soient phénoménologiques ou mécanistes, présentent de nombreuses faiblesses. Les approches phénoménologiques n'articulent pas explicitement la relation entre mutation, sélection et dérive génétique, et pourraient encore être améliorées, tout en restant dans l'idée de ne pas modéliser explicitement le paysage sélectif dans ses détails. Quant aux approches mécanistes, dans leurs versions actuelles, elles font des hypothèses très fortes, telles que l'indépendance entre sites, un paysage de fitness fixe au cours du temps, mais aussi une taille efficace de population (N_e) constante le long de la phylogénie. Plus fondamentalement, il existe un certain vide à combler entre ces approches phénoménologiques et mécanistes, et de meilleures connexions conceptuelles et pratiques pourraient être établies entre elles.

Dans ce contexte, mon travail de thèse représente une tentative de démêler les interactions complexes entre mutation, sélection et dérive génétiques en construisant de nouveaux modèles à codons phylogénétiques, selon les deux approches, phénoménologiques et mécanistes. Au cours de ce travail, j'ai lié des idées théoriques à des données empiriques, en utilisant une combinaison d'approches analytiques, d'expériences de simulation de développements statistiques et informatiques utilisant les principes de l'inférence bayésienne par chaines de Markov Monte-Carlo. Les résultats sont divisés en trois manuscrits indépendants, sur le point d'être soumis à des journaux à comité de lecture.

Le premier article revient sur la question de l'équilibre entre biais de mutation et biais de sélection, et de comment cet équilibre doit être correctement formalisé dans le contexte des modèles à codons phénoménologiques. Parce que la composition des séquences d'ADN codant pour les protéines ne reflète pas le processus sous-jacent de mutation, mais son filtrage par sélection au niveau des acides aminés, une modélisation minutieuse est nécessaire pour démêler le processus de mutation et les biais nucléotidiques d'un côté, et la sélection d'un autre côté. Malheureusement, les modèles à codons phénoménologiques actuels, développés à l'origine pour estimer la pression de sélection s'exerçant sur les protéines, ne modélisent pas correctement cet équilibre mutation-sélection. En effet, ils utilisent le biais de composition nucléotidique observé comme proxy pour le biais mutationnel. En conséquence, ils ne fournissent pas une estimation précise du processus de mutation, même s'ils sont capables d'estimer de manière assez fiable la pression de sélection agissant sur les acides aminés. Pour résoudre ce problème, j'ai développé un modèle à codon phylogénétique dans lequel la pression de sélection n'est pas considérée comme un paramètre unique, mais comme un tenseur (95 paramètres libres). Le tenseur capture les faibles différences de pression de sélections dans différentes directions, ce qui donne une représentation précise de la manière dont la mutation et la sélection s'opposent à l'équilibre. Cette paramétrisation représente la forme paramétrique la plus simple, dans un contexte phénoménologique, capable de séparer les effets de la mutation et de la sélection de manière exacte, ou asymptotiquement exacte. Grâce à cela, cette approche de modélisation donne une estimation fiable du processus de mutation, tout en démêlant les pressions de sélection dans différentes directions. Ces développements offrent des outils qui permettront ultimement de mieux comprendre comment le processus mutation-mutation s'articule avec d'autres processus évolutifs impactant la composition nucléotidique, tels que la conversion génique biaisée (gBGC).

Si le premier manuscrit se focalise sur l'articulation entre mutation et sélection, l'équilibre entre ces deux forces est arbitré par la dérive génétique, qui à son tour est modulée par taille efficace de population (N_e). En conséquence, théoriquement, la variation de N_e le long d'une phylogénie peut être déduite de l'histoire des substitutions le long des lignées. Le deuxième manuscrit explore ainsi la question de la prise en compte des variations à long terme de la taille efficace de population (N_e) entre les espèces, dans le contexte d'un modèle à codons mécaniste. Les travaux présentés dans ce second manuscrit représentent la partie la plus intensive du travail de doctorat, en matière de modélisation, d'algorithmes de Monte-Carlo et de développement logiciel. J'ai ainsi développé un modèle à codons mécaniste reconstituant le paysage de fitness en chaque site, les tendances à long terme de la taille efficace de population et du taux de mutation le long de la phylogénie, à partir d'alignements d'ADN de séquences codantes. Simultanément, l'approche estime la corrélation entre les traits d'histoires de vie, le taux de mutation et la taille efficace de population, prenant explicitement en compte l'inertie phylogénétique. Ce modèle a été testé sur des données simulées, puis appliqué à des données empiriques chez les mammifères, les isopodes, les primates et les drosophiles. Les résultats sur données simulées et empiriques suggèrent qu'il existe des signaux per-

sistants dans les séquences d'ADN qui permettent de reconstruire l'histoire évolutive du N_e le long de la phylogénie. Par ailleurs, les variations de taille efficace de population inférées corrèlent avec les traits d'histoire de vie ou les variables écologiques d'une façon qui est attendue d'après les connaissances écologiques disponibles par ailleurs. Cependant, l'ampleur de la variation inférée de N_e à travers la phylogénie est plus étroite que prévu, si l'on compare en particulier aux estimés sur la base du polymorphisme.

Cette dernière observation, qui suggère une violation de certaines hypothèses du modèle, m'a amené à revoir la question de savoir comment la biophysique des protéines, et plus généralement l'épistasie, peut moduler quantitativement la réponse du processus évolutif moléculaire aux changements de la taille efficace de population. Ce dernier travail est présenté comme un troisième manuscrit. En effet, les hypothèses sur la structure sous-jacente du paysage de fitness peuvent avoir une grande influence sur la vitesse d'évolution des protéines, et tout particulièrement sur les changements de cette vitesse d'évolution après un changement de N_e. En plus de N_e, le niveau d'expression des protéines est un autre facteur majeur susceptible de moduler la vitesse d'évolution moléculaire. Les protéines fortement exprimées évoluent généralement moins vite, une corrélation prédite par les modèles biophysiques supposant que les protéines mal repliées sont toxiques et donc soumises à une sélection purificatrice. En conséquence, il convient d'articuler ensemble toutes ces corrélations entre la vitesse d'évolution, la taille efficace de population et le niveau d'expression, en rapport avec la structure du paysage de fitness sous-jacent. Pour ce faire, j'ai dérivé une approximation théorique de la réponse quantitative de vitesse d'évolution à des changements à la fois de N_e et du niveau d'expression, en fonction de la relation génotype-phénotype-fitness sous-jacente. Ce développement est généralement valide pour des traits phénotypiques additifs et une fonction de fitness concave, mais a été appliqué plus spécifiquement à un modèle biophysique dans lequel les protéines sont sous sélection directionnelle pour maximiser leur stabilité conformationnelle. Dans ce cas précis, le modèle prédit une réponse faible du taux d'évolution aux changements de N_e ou de niveau d'expression (qui sont interchangeable), un résultat corroboré par des simulations sous des modèles plus complexes. Sur la base de preuves empiriques, je propose que l'adéquation basée sur la stabilité conformationnelle puisse ne pas fournir un mécanisme suffisant pour expliquer l'amplitude des variations de la vitesse d'évolution observée empiriquement, entre protéines ou entre espèces, induites par les variations de niveau d'expression ou de taille efficace de population. D'autres aspects de la biophysique des protéines pourraient être explorés tels que la sélection pour limiter les interactions non spécifiques entre protéines. Ces aspects pourraient conduire à une réponse plus forte de la vitesse d'évolution aux changements de N_e. Plus généralement, ce travail offre des perspectives pour réduire l'écart entre les prévisions quantitatives des modèles biophysiques et les observations empiriques reliant la réponse de la pression de sélection aux changements de N_e et du niveau d'expression.

Pour conclure, ce travail est une tentative encourageante, quoiqu'encore inaboutie de construire des modèles intégrés d'évolution des séquences d'ADN codant pour les protéines. Ce travail réussit à consolider l'idée que les motifs de substitutions nous informent sur les fluctuations à long terme de la dérive génétique le long des branches et la sélection le long des séquences. Il démontre que les hypothèses faites sur la structure du paysage de fitness ont une importance critique sur la sensibilité des changements vitesse d'évolution à des changements de variables écologiques (N_e) ou de moléculaires (niveau d'expression des protéines). Réciproquement, les observations empiriques des motifs de substitutions en réponse à des changements de variables moléculaires ou écologiques nous informent sur la structure sous-jacente du paysage de fitness. En se basant sur l'équilibre mutation-sélection et en intégrant explicitement la taille efficace de population, ce travail présente aussi un cadre conceptuel permettant de relier phylogénie et génétique des populations, dont certaines pistes d'unifications sont envisagées. Enfin, je pense que cette thèse consolide les modèles théoriques sur lesquels se fonde l'évolution moléculaire et souligne les écueils à éviter, tout en donnant des perspectives pour le développement de méthodes d'inférence permettant d'intégrer différentes données empiriques et niveaux de complexité.