Avant-propos

Gilles DIDIER[1] et Stéphane GUINDON[2]
[1] *IMAG, CNRS, Université de Montpellier, Montpellier, France*
[2] *LIRMM, CNRS, Université de Montpellier, Montpellier, France*

L'évolution, dans son sens usuel en sciences de la vie qui est le phénomène par lequel les espèces vivantes se transforment au cours du temps, est un processus biologique essentiel. On peut même dire qu'il est le plus important puisque tous les autres phénomènes biologiques en découlent d'une façon ou d'une autre. L'évolution est ainsi à l'origine non seulement de l'extraordinaire diversité du vivant, mais a aussi façonné toutes les fonctions biologiques que l'on peut observer. Sa théorie générale date du XIX^{e} siècle avec les travaux de Darwin. Si cette théorie est aujourd'hui largement admise et partagée, son étude reste un domaine scientifique extrêmement vivant et fécond, qui s'est particulièrement développé ces dernières décennies, et est, en conséquence, vaste.

Cet ouvrage n'ambitionne naturellement pas d'en présenter tous les aspects et s'intéresse plus particulièrement aux nombreuses interactions entre la biologie, les mathématiques et l'informatique que son étude a suscitées. La raison principale pour laquelle les études évolutives sont autant consommatrices de modèles mathématiques et d'algorithmes vient certainement du fait que l'évolution biologique est un processus qui agit depuis plus de 4 milliards d'années et qui n'est pas, à de rares exceptions près (voir chapitre 11), directement observable à notre échelle de temps. On ne peut donc l'étudier qu'à partir des données qui nous sont accessibles aujourd'hui, c'est-à-dire les espèces actuelles et le registre fossile. Pour tester des hypothèses sur les mécanismes gouvernant l'évolution, il est généralement nécessaire d'exprimer celles-ci en termes de modèles mathématiques. Ces derniers sont des représentations simplifiées de la

Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stéphane GUINDON. © ISTE Editions 2022.

réalité biologique qui permettent de reconstruire (imparfaitement) l'histoire évolutive des espèces contemporaines (et ancestrales dans le cas des fossiles) que l'on observe. L'ajustement de ces modèles aux données permet d'évaluer leur pertinence et de valider ou réfuter les hypothèses initialement proposées. La conception des modèles et l'inférence de leurs paramètres soulèvent des questions mathématiques, informatiques et statistiques qui ont ouvert de nouveaux champs de recherche à la fois théoriques et appliqués dans ces domaines.

L'objet central dans l'étude de l'évolution est l'arbre du vivant. Ce dernier est tout d'abord une représentation naturelle de la diversification des espèces dans le sens où il décrit les relations de parentés entre espèces (ou même entre individus). Dans les modèles, nous verrons qu'il est interprété tantôt comme support de l'évolution, que l'on peut tenter de reconstruire à partir des données disponibles, tantôt comme représentation de la dépendance statistique entre les caractères portés par les espèces. L'arbre est également un objet théorique qui a été étudié en informatique et mathématiques, du point de vue de sa combinatoire notamment. Le chapitre 1 présente brièvement cet aspect avant de décrire différents modèles évolutifs amenant à des arbres et les probabilités qui leur sont associées sous ces modèles.

Si les arbres permettent de représenter le cadre dans lequel l'évolution prend place, cette dernière opère sur les différents caractères (un terme à prendre dans un sens très large ici) portés par les êtres vivants. C'est d'ailleurs à travers ces caractères que l'on peut l'étudier. Le « caractère » le plus utilisé dans ce cadre est le matériel génétique, c'est-à-dire des molécules/polymères dont on peut extraire la séquence des briques élémentaires sous la forme de (longs) mots sur des alphabets finis. En effet, le développement de la génétique, d'abord identifiée comme support de l'évolution, puis des techniques de séquençage de l'ADN ont révolutionné l'étude de l'évolution biologique en changeant la nature et en provoquant une explosion de la quantité des données exploitables dans ce domaine. Utiliser ces données (et d'autres) pour mieux comprendre l'évolution nécessite des ressources mathématiques et informatiques sans cesse renouvelées afin de faire face au perpétuel changement de la quantité et du type des données.

Le chapitre 2 présente les principaux modèles markoviens d'évolution des séquences d'ADN. Ces modèles, généralement considérés comme mécanistes, décrivent les processus évolutifs au niveau moléculaire, sur des périodes de temps suffisamment longues pour que la variabilité génétique intra-espèce soit négligeable par rapport à la variabilité interspécifique. La grande majorité de ceux-ci considèrent que les différentes positions le long des séquences génétiques évoluent indépendamment les unes des autres et suivent le même modèle markovien en temps continu. Ce même chapitre décrit également des modèles probabilistes permettant de prendre en compte la variabilité des vitesses d'évolution le long des séquences, un phénomène important d'un point de vue biologique, notamment en ce qui concerne l'évolution des parties codantes des génomes, celles-ci étant contraintes par la structure du code génétique.

Enfin, des modèles de même type que ceux utilisés pour les séquences d'ADN peuvent également être utilisés pour modéliser l'évolution de caractères discrets tels que la présence ou l'absence d'une caractéristique morphologique donnée (par exemple le nombre de doigts, etc.).

L'évolution concerne également les caractères physiques des espèces, notamment les caractères dits quantitatifs tels que la taille, le poids, etc. Si ceux-ci sont moins utilisés à des fins d'inférence phylogénétique, comprendre leur évolution est essentiel en biologie, par exemple pour tester des hypothèses à propos de relations morphométriques et allométriques en écologie. Les modèles d'évolution de caractères continus constituent également un outil pertinent pour détecter les traces éventuelles de la sélection naturelle sur l'évolution de caractères morphologiques. Le chapitre 3 présente en détail le cadre générique dans lequel ces modèles sont implémentés ainsi qu'une large gamme d'approches régulièrement utilisées pour modéliser de manière pertinente la corrélation entre caractères qui découle des liens de parentés évolutives entre espèces comparées.

Les modèles présentés dans les chapitres 2 et 3 supposent l'indépendance en probabilité des caractères considérés (entre sites des séquences, entre branches de la phylogénie). Si cette hypothèse d'indépendance permet d'éviter une explosion des temps de calcul et de la taille des modèles, elle ne peut être considérée comme réaliste dans de nombreux cas. Le chapitre 4 présente différentes approches permettant de mettre en évidence et d'étudier l'interdépendance des évolutions de plusieurs caractères, discrets ou continus. Tout comme pour les modèles présentés dans le chapitre 3, les modèles dits de co-évolution considèrent l'arbre phylogénétique comme un paramètre de nuisance qu'il est nécessaire de prendre en compte afin d'évaluer la part de corrélation entre caractères morphologiques qui n'est pas expliquée par les parentés évolutives.

Les séquences génétiques n'évoluent pas seulement par mutations, comme présenté dans le chapitre 2, mais changent parfois de manière plus radicale. À l'échelle du génome, notamment, l'évolution procède parfois par duplication ou inversion de pans entiers des chromosomes. Ces types de changements sont très riches en termes d'information sur les distances évolutives entre génomes. Ils offrent une vision plus globale de l'évolution que les « simples » modèles de substitutions ponctuelles entre nucléotides. Les réarrangements génomiques sont cependant plus difficiles à modéliser mathématiquement. Le chapitre 5 fait le point sur l'approche générale pour détecter ces réarrangements et reconstruire l'évolution à l'échelle des génomes.

Les cinq premiers chapitres de notre ouvrage donnent un tour d'horizon des modèles en évolution. Les quatre chapitres suivants illustrent comment certains de ces modèles sont mis à profit dans le cadre de l'inférence phylogénétique, c'est-à-dire la détermination des relations de parentés entre espèces ou individus et du temps depuis

lequel elles ont divergé. Plusieurs approches couramment utilisées pour répondre à cette question y sont présentées.

Une première approche pour reconstruire l'évolution d'un groupe d'espèces consiste à considérer une matrice de distance ou de dissimilarité, c'est-à-dire une mesure de la « ressemblance » entre paires d'espèces. On fait ici l'hypothèse que plus les espèces s'éloignent d'un point de vue évolutif, moins elles se ressemblent du point de vue de la mesure choisie. Sous cette hypothèse, l'arbre qui représente au mieux ces distances et que l'on va chercher à déterminer est proche de celui qui retrace leur évolution. La dissimilarité, ou distance, utilisée peut être calculée à partir de séquences génétiques, de caractères morphologiques ou autres. Bien qu'une distance résume drastiquement l'ensemble des caractères portés par les espèces, différentes méthodes présentées dans le chapitre 6 permettent de reconstruire des arbres d'évolution réalistes à partir de cette information. Ces approches sont même quasiment les seules applicables en pratique pour reconstruire les arbres comprenant un grand nombre d'espèces grâce à leur vitesse de calcul.

Si les méthodes présentées dans le chapitre 6 sont très rapides, elles ne considèrent pas directement l'évolution des espèces étudiées, ne la voyant qu'à travers la distance choisie. D'autres approches font directement intervenir des mécanismes ou des modèles d'évolution à des fins d'inférence phylogénique. L'une des premières à avoir été envisagées est la parcimonie qui repose sur le principe du rasoir d'Occam et recherche la phylogénie impliquant le moins d'événements évolutifs possible. La parcimonie a peu à peu été supplantée par des approches basées sur des modèles probabilistes tels que ceux présentés dans le chapitre 2. Une première façon d'utiliser de tels modèles pour l'inférence phylogénétique consiste à rechercher l'arbre maximisant la probabilité des données observées sous le modèle choisi. C'est ce qu'on appelle le *maximum de vraisemblance*. La méthode est assez proche dans son esprit de la parcimonie et ces deux approches sont décrites dans le chapitre 7.

Une autre manière d'utiliser les modèles probabilistes d'évolution de séquences pour l'inférence phylogénique est de se placer dans le cadre de l'échantillonnage bayésien où l'on ne cherche plus seulement à déterminer l'arbre maximisant la probabilité des données observées, mais à associer à tout arbre sa probabilité *a posteriori*, c'est-à-dire sa probabilité conditionnellement aux données observées. Cette approche permet de mieux prendre en compte l'incertitude inhérente au processus d'inférence que lorsqu'on utilise des approches par maximum de vraisemblance. Le chapitre 8 présente les principes généraux des approches bayésiennes par chaînes de Markov par Monte Carlo et leurs applications à la phylogénie.

Les approches bayésiennes ne sont pas les seules à pouvoir quantifier l'incertitude sur les arbres inférés par les différentes méthodes. Évaluer cette dernière est essentiel, car dans de nombreuses situations l'histoire évolutive est difficile à reconstruire, suscitant même des controverses parmi ceux qui l'étudient. Le chapitre 9 présente plusieurs

approches permettant de quantifier l'incertitude associée à chaque branche d'un arbre phylogénétique. La technique du *bootstrap* non paramétrique, bien connue des statisticiens, a longtemps été l'approche privilégiée. D'autres approches, plus rapides, ont récemment vu le jour au cours des deux dernières décennies. Celles-ci sont décrites dans le détail, offrant ainsi une vision d'ensemble des solutions actuelles permettant de quantifier les incertitudes dans les reconstructions d'arbres phylogénétiques.

Si cet ouvrage a été jusqu'ici consacré aux modèles et aux méthodes sur lesquels reposent les outils modernes nous permettant d'étudier l'évolution, les chapitres suivants illustrent comment ces approches sont utilisées pour faire avancer nos connaissances en biologie qui, à leur tour, peuvent parfois enrichir les analyses phylogénétiques.

Par exemple, considérer les phylogénies contenant des espèces fossiles est intéressant à plusieurs titres. Au-delà de leur intérêt paléontologique, les espèces fossiles sont essentielles à la datation des arbres phylogénétiques, plus précisément à celle de leurs nœuds internes représentant les spéciations. En effet, les modèles d'évolution présentés dans les chapitres 2 et 3 ne permettent pas de déterminer des durées d'évolution en temps « absolu ». Celles-ci ne sont déterminées que relativement à une vitesse d'évolution qui n'est pas identifiable sans point de calibration temporel (par exemple, la date d'un événement évolutif). D'un autre côté, les strates géologiques dans lesquelles sont trouvés les fossiles permettent de dater plus ou moins précisément la période pendant laquelle l'espèce associée vivait. Intégrer des espèces fossiles à une phylogénie permet donc d'y ajouter des informations temporelles et d'estimer notamment des dates de spéciation. Le chapitre 10 présente les enjeux et particularités de ce type de données et les difficultés à les intégrer dans une analyse phylogénétique.

Un autre domaine d'application des modèles phylogénétiques est celui de la phylodynamique présentée dans le chapitre 11. Il s'agit ici de retracer l'évolution d'agents pathogènes, virus ou bactéries, afin de comprendre la dynamique populationnelle sous-jacente. L'arbre phylogénétique ne décrit plus ici une suite de spéciations mais plutôt les événements de transmission de l'agent pathogène entre hôtes successifs. Le tempo auquel ces évènenements se produisent, lorsque celui-ci est analysé dans le cadre d'un modèle épidémiologique adéquat, est directement lié à la dynamique de l'épidémie. Le chapitre 11 donne une vue d'ensemble des approches statistiques modernes (et malheureusement d'actualité à l'heure où nous écrivons cette préface) mêlant phylogénétique et épidémiologie.

Enfin, il est à noter que dans les méthodes d'inférence phylogénétique présentées jusqu'ici (à l'exception du chapitre précédent), l'unité d'évolution est l'espèce. Retracer l'évolution des individus constituant une espèce, voire même une population, est un autre défi d'envergure. Le chapitre 12 donne un vaste aperçu des avancées permises par la généalogie des individus retracée à partir des données génétiques combinée aux modèles probabilistes en génétique des populations. De telles approches permettent

notamment d'inférer les variations de taille de populations ancestrales chez l'Homme et de reconstruire, à partir de la seule analyse de séquences génétiques géoréférencées, des événements migratoires survenus voici plusieurs milliers d'années.

L'objectif de cet ouvrage n'est pas de donner une vision exhaustive des techniques permettant d'étudier l'évolution ou leurs applications. Par exemple, les approches dites de phylogénomiques, qui visent à reconstruire l'évolution à partir de l'analyse de multiples gènes, ne sont pas présentées. De la même façon, les techniques de détection des traces laissées par la sélection naturelle au sein des séquences génétiques, ou encore celles de datation moléculaire, ne sont pas abordées dans les détails. Notre parti pris ici a été de nous concentrer sur certaines des approches les plus fréquemments utilisées à l'heure actuelle et d'en donner une description approfondie. Les exemples d'application de ces méthodes sont évidemment très nombreux et, là encore, nous ne donnons qu'un aperçu de ceux-ci. Nous espérons néanmoins que la lecture de cet ouvrage suscitera l'intérêt et la curiosité chez le lecteur peu familier des recherches sur l'évolution tout en permettant aux étudiants et chercheurs du domaine d'approfondir leurs connaissances sur ces questions.