Phylogénétique : quelles mesures de support pour les branches d'un arbre

Olivier Gascuel, Frédéric Lemoine

Chapitre 1

Phylogénétique : quelles mesures de support pour les branches d'un arbre

Olivier Gascuel[1,2] et Frédéric Lemoine[1]

1- ISYEB, UMR7205 (CNRS, MNHN, EPHE, SU, UA), Paris, FRANCE
2- Unité de Bioinformatique Evolutive, Institut Pasteur, Paris, FRANCE
olivier.gascuel@mnhn.fr

Résumé

Une fois une phylogénie estimée, se pose la question du support de ses branches et des inférences associées : Décrivent-elles la réalité du passé biologique ? Sont-elles fiables, reproductibles, robustes ? Ou à l'inverse peut-on conclure que le signal contenu dans les données ne permet pas de trancher entre plusieurs hypothèses alternatives ? De nombreuses approches ont été proposées pour répondre à ces questions. Nous décrivons dans ce chapitre les principales méthodes utilisées aujourd'hui en phylogénétique. Certaines sont locales et demandent peu de temps calcul. D'autres comme le bootstrap requièrent à l'inverse des calculs lourds mais donnent une vision plus globale. On distingue les méthodes paramétriques qui font explicitement appel à un modèle de l'évolution des séquences et se basent sur la fonction de vraisemblance ou la distribution a posteriori dans un cadre Bayésien, et les approches non-paramétriques qui reposent sur moins d'hypothèses explicites. On trouvera quatre parties dans ce chapitre : (1) supports locaux ; (2) bootstrap phylogénétique de Felsenstein et bootstrap de transfert ; (3) supports Bayésiens basés sur les probabilités a posteriori; (4) comparaison de ces approches et discussion.

1.1. Introduction

La première étape en phylogénie est de construire un arbre à partir de données. Dans la plupart des cas on se base aujourd'hui sur un alignement multiple de séquences homologues, ADN ou protéines, et c'est dans ce cadre que se place ce chapitre. Chaque séquence est associée à un taxon (*taxonomic unit*) représenté par une étiquette (par exemple le nom d'une espèce). L'arbre inféré comporte des nœuds internes ou 'nœuds ancestraux', des feuilles associées bijectivement aux taxons, et des longueurs de branche mesurées en nombre moyen de substitutions par site. La plupart des méthodes produisent des arbres non-enracinés et binaires (chaque nœud interne est de degré 3). A nouveau on se limitera à ce cadre pour plus de simplicité. Chaque branche d'un arbre définit une bipartition de l'ensemble des taxons : en retirant une branche on coupe en deux cet ensemble. Les branches externes séparent un taxon de tous les autres, on parle de bipartitions triviales, présentes dans toutes les phylogénies portant sur le même ensemble de taxons. A l'inverse les bipartitions et branches internes sont spécifiques de l'arbre considéré. Il y a une forme d'équivalence entre un arbre phylogénétique et les bipartitions qu'il définit : si ses bipartitions sont connues on peut reconstruire la topologie de l'arbre correspondant sans ambiguïté. C'est pour cette raison que lorsqu'on s'intéresse au support d'un arbre, on considère généralement le support de chacune de ses branches ou bipartitions. Voir Chapitre XX pour plus de détails sur les arbres et leurs propriétés.

La reconstruction d'un arbre à partir de séquences est une forme d'inférence statistique. Comme dans toute inférence de ce type on s'intéresse à la fiabilité des estimations obtenues à partir des données. Par exemple, dans un sondage politique on calcule en se basant sur les caractéristiques de l'échantillon une fourchette qui encadre le pourcentage estimé de votes pour tel ou tel candidat. Cette fourchette représente l'erreur liée à l'échantillonnage nécessairement limité d'une population très large. En phylogénie l'estimé est bien plus complexe, c'est un arbre et non une simple valeur numérique. Comme on s'intéresse aux bipartitions (non numériques) on parlera de valeurs de confiance ou de tests, plutôt que d'intervalle de confiance. A nouveau, une cause d'incertitude est liée à l'échantillonnage qui porte sur les séquences comme sur les sites de l'alignement. On verra que le nombre n de séquences et le nombre s de sites ont des impacts essentiellement opposés : des valeurs élevées de s et faibles de n tendent à produire de forts supports, et vice et versa. Le choix des séquences peut aussi avoir un fort impact, et l'alignement est une source d'incertitude, car il est le résultat d'un programme lui-même incertain et possiblement biaisé. Également, l'ensemble du processus repose fondamentalement sur un modèle plus ou moins explicite de l'évolution des séquences (Chapitre XX). L'arbre lui-même est un modèle. Le modèle global incluant l'arbre et les mécanismes substitutionnels est nécessairement simplificateur eu égard à la

complexité de l'évolution biologique. Pour cette raison, on ne peut généralement pas considérer que les valeurs de confiance associées aux inférences phylogénétiques nous permettent de dire avec une probabilité élevée que tel ou tel événement évolutif s'est bien produit dans le passé. En général, on pourra simplement affirmer que compte tenu de la méthode et du modèle considérés, le résultat de l'analyse est plus ou moins robuste et reproductible. Si le modèle et/ou la méthode sont inappropriés, on pourra avoir des supports élevés pour des inférences phylogénétiques erronées.

On a vu dans les Chapitres précédents (XX) qu'il y a essentiellement quatre approches pour inférer des arbres à partir de séquences : parcimonie, distances évolutives, maximum de vraisemblance, et chaine de Markov Monte-Carlo (MCMC) dans un cadre Bayésien. Seule la première (parcimonie) ne repose pas sur un modèle probabiliste explicite de l'évolution des séquences. Pour la parcimonie la seule approche standard pour calculer des supports de branche est le bootstrap non-paramétrique introduit en phylogénie par Felsenstein (1985). Cette approche est applicable à toutes les méthodes d'inférence d'arbre à partir d'alignements multiples de séquences. Elle est très souvent utilisée, sauf dans le cadre Bayésien où l'approche MCMC permet d'autres mesures basées sur les probabilités a posteriori d'observer ou non les bipartitions considérées (Chapitre XX). L'article de Felsenstein est l'un des plus cités de l'histoire des sciences, avec plus de 40 000 citations (Van Noorden et al. 2014). Mais cette méthode est très lourde en temps calcul car elle impose de relancer l'inférence d'arbre sur un grand nombre (typiquement de 100 à 1000) de pseudo-jeux de données ou « échantillons bootstrap ». Dans le cadre de la vraisemblance (lourde en temps calcul), des algorithmes ont été proposés pour accélérer le bootstrap, en combinant le ré-échantillonnage propre au bootstrap avec la recherche d'arbres optimaux au sens de la vraisemblance (Stamatakis et al. 2008; Minh et al. 2013), mais elles restent couteuses en temps calcul et n'offrent pas les mêmes garanties que la méthode standard. Cette lourdeur explique le succès des méthodes locales basées sur la vraisemblance et un modèle explicite d'évolution des séquences. Ces méthodes (Anisimova et Gascuel 2006 ; Anisimova et al. 2011) ne remettent pas en cause tout l'arbre comme le bootstrap, mais seulement les configurations topologiques locales autour de la branche d'intérêt. Leur temps calcul est faible et ne change pas significativement le temps nécessaire pour inférer un arbre au sens du maximum de vraisemblance.

Dans la suite on présente : (1.2) les méthodes locales rapides (*aLRT*) ; (1.3) le bootstrap de Felsenstein et une nouvelle version basée sur la distance de transfert (*Transfer Bootstrap Expectation* ; Lemoine et al. 2018), particulièrement adaptée pour les grands jeux de données avec peu de signal phylogénétique ; (1.4) les probabilités postérieures utilisées dans un cadre Bayésien, y compris une approche

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

locale rapide (*aBayes*) ; (1.5) une discussion et une comparaison de ces supports, qui reposent sur des bases très différentes et doivent être interprétés avec précaution.

1.2. Supports locaux : aLRT paramétrique et non-paramétrique

1.2.1. *Le test de branche nulle et ses limites*

Les méthodes locales trouvent leur source dans un test implémenté par Felsenstein dans plusieurs logiciels de la suite PHYLIP : le 'test de branche nulle' (*null branch test*). L'idée est simple et facile à mettre en œuvre dans le cadre du maximum de vraisemblance, mais aussi pour les méthodes de distance aux moindres carrés, qui s'apparentent au maximum de vraisemblance sous hypothèse de normalité.

Considérons un arbre T, tel que retourné par des logiciels comme PhyML (Guindon et Gascuel 2003 ; Guindon et al. 2010) ou RaxML (Stamatakis 2014) basés sur le maximum de vraisemblance, et une branche b de T dont on cherche à calculer le support. Comme ces logiciels sont heuristiques et que le problème de reconstruction d'arbre est NP-difficile (Chor et Tuller 2006), il y a peu de chance pour que T soit l'optimum global parmi tous les arbres possibles, mais c'est un optimum local au sens de la vraisemblance : ses branches sont de longueur optimale pour la topologie considérée, et on ne trouve pas de topologie au voisinage de T qui conduise à une meilleure vraisemblance que celle de T. La notion de voisinage topologique en phylogénie est définie par les 'mouvements topologiques'. Le plus simple de ces mouvements est 'l'échange de voisins les plus proches' (*nearest neighbor interchange* ou NNI). Dans un arbre binaire T une branche interne b définit 4 sous arbres, 2 à chaque extrémité de b, notés A, B, C et D (Fig. 1.1). A partir de la configuration courante T on peut réaliser deux NNI autour de b : l'échange de B et C et l'échange de B et D. Il y a donc trois configurations topologiques résolues autour de b (A avec C, A avec D, et la configuration courante issue de T où A est avec B). Il y a de plus une configuration irrésolue dite en étoile où la branche b est de longueur nulle, ce qui sera noté pour simplifier $b = 0$. Tous les logiciels raisonnables assurent que la configuration de meilleure vraisemblance est la configuration courante issue de T (optimum local) dont la vraisemblance est notée V_1. On note V_2 et V_3 les vraisemblances des deuxième et troisième meilleures configurations NNI autour de b, et V_0 la vraisemblance de l'arbre en étoile $b = 0$. Les log-vraisemblances sont notées LV_1, LV_2, LV_3 et LV_0. Ces quatre vraisemblances sont définies après optimisation par maximum de vraisemblance des longueurs de branche de T, une fois effectué le NNI correspondant ou la mise à zéro de b.

En termes de modèle on a ici des emboitements : chacune des trois configurations NNI autour de *b* contient la configuration en étoile obtenue en posant $b = 0$. Et chacune des quatre vraisemblances V est celle du maximum de vraisemblance de la configuration considérée, une fois les longueurs de branches optimisées. On a donc $V_1 \geq V_2 \geq V_3 \geq V_0$, la dernière inégalité venant de l'emboitement de la configuration en étoile dans les trois alternatives NNI.

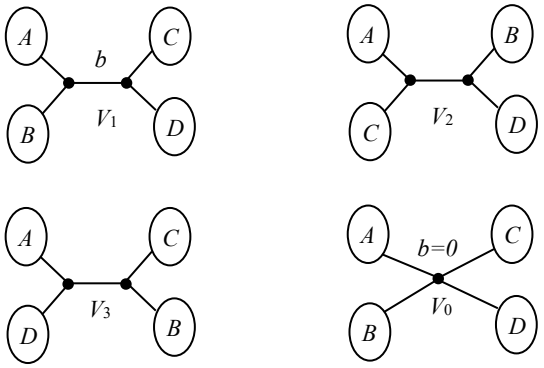

Figure 1.1. *Les trois configurations NNI autour de la branche d'intérêt b, de vraisemblance V_1, V_2, V_3. La première configuration (A avec B, V_1) est issue de l'arbre du maximum de vraisemblance T, dont on calcule les supports de branche. Les configurations NNI suivantes correspondent aux deuxième et troisième meilleures vraisemblances V_2, V_3. La configuration en étoile, notée $b=0$, a une vraisemblance V_0 .*

Le test de branche nulle compare LV_1 et LV_0 et effectue un test de rapport de vraisemblance en se basant sur la statistique $-2(LV_0 - LV_1)$. Sachant que la configuration en étoile est emboitée dans la configuration courante T autour de *b*, peut-on exclure l'hypothèse H0 selon laquelle la branche *b* est de longueur nulle ? Si on ne peut pas rejeter H0, on considère que la branche *b* n'est pas supportée. Dans l'hypothèse alternative H1, la branche est supportée avec un niveau de confiance égale à 1 moins la p-valeur calculée sous H0. Il s'agit d'un test très simple entre deux modèles emboités dont l'un a 1 degré de liberté de plus que l'autre. Une complication vient de ce que sous H0 on est sur la frontière des valeurs possibles pour la longueur de *b*, qui est nécessairement positive ou nulle. Du fait de cette contrainte/frontière la loi limite sous H0, utilisée pour calculer la p-valeur, est un mélange 50|50 entre les lois du χ^2 à 0 et 1 degrés de liberté (Anisimova et Gascuel 2006). Il faut de plus prendre en compte la multiplicité du test, puisqu'on compare

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

ici la configuration en étoile aux trois configurations alternatives. Une correction de Bonferroni est appliquée à la p-valeur avec un facteur 3.

Ce test a eu un certain succès du fait de sa rapidité et sa simplicité. Puis on s'est rendu compte sur de vraies données que la comparaison entre les configurations de vraisemblance LV_2 et LV_0, voire LV_3 et LV_0, donnait parfois des p-valeurs significatives. Autrement dit les trois configurations NNI autour de b pouvaient toutes les trois être supportées par les données, par comparaison avec la configuration en étoile. Le test ne répondait donc pas à la question posée, qui est de départager la configuration courante des deux configurations NNI alternatives. Le test de branche nulle est aujourd'hui totalement abandonné et n'est plus disponible dans les logiciels récents. Mais c'est une source d'inspiration comme on va le voir.

1.2.2. *Le test local aLRT, version paramétrique*

L'idée de ces tests est de se baser sur la statistique $-2\left(LV_2 - LV_1\right)$, qui compare la meilleure configuration NNI courante issue de T $\left(LV_1\right)$ à la deuxième meilleure configuration $\left(LV_2\right)$. Si l'écart est significatif, la branche b est supportée avec un niveau de confiance égal à 1 moins la p-valeur du test.

La difficulté est de calculer cette p-valeur. Dans une première version (Anisimova et Gascuel 2006) nous nous sommes basés sur l'inégalité $-2\left(LV_2 - LV_1\right) \leq -2\left(LV_0 - LV_1\right)$, issue de $LV_2 \geq LV_0$, et avons utilisé le même mélange de χ^2 avec une correction de Bonferroni que le test de branche nulle, pour calculer la p-valeur associée à cette nouvelle statistique $-2\left(LV_2 - LV_1\right)$. Si celle-ci est significative, celle du test de branche nulle l'est plus encore, mais on a maintenant la garantie de répondre à la bonne question qui est d'écarter les deux configurations NNI alternatives. Compte tenu de l'inégalité entre les deux statistiques, on s'attend à un test conservatif, du moins plus conservatif que le test de branche nulle. Ce test directement inspiré du rapport de vraisemblance a été nommé *approximate likelihood ratio test* ou aLRT.

Des comparaisons avec le bootstrap phylogénétique de Felsenstein ont montré que ce test n'est pas aussi conservatif qu'attendu et soutient parfois des branches qui sont écartées par le bootstrap, lui-même conservatif il est vrai (voir Fig. 1.2 pour un exemple). Cet écart à l'attendu vient de l'utilisation de la loi limite du χ^2, la log-vraisemblance en phylogénie ayant un comportement très particulier, avec des écarts considérables entre les log-vraisemblances de chaque site, sans qu'apparaisse de site moyen et de normalité de la loi. Aussi nous nous sommes tournés vers une approche non-paramétrique pour calculer la *p*-valeur du test.

1.2.3. *Le test local aLRT, version SH-like non-paramétrique*

Les modèles phylogénétiques font l'hypothèse de l'indépendance des sites de l'alignement multiple donné en entrée des méthodes de reconstruction. Aussi la log-vraisemblance des données est égale à la somme des log-vraisemblances de chaque site (Chapitre XX). On dispose donc pour chaque statistique LV_1 et LV_2 de la distribution des log-vraisemblances par site, que l'on va utiliser pour calculer la significativité de l'écart $(LV_1 - LV_2)$, de manière analogue à ce qu'on fait dans un test de comparaison de moyennes avec des observations appariées. Comme on ne peut pas faire d'hypothèse de normalité, on se base sur une application du bootstrap non-paramétrique, plutôt que sur un calcul d'écart type ou une autre approche paramétrique. Soit LV_{1i} et LV_{2i} les log-vraisemblances au site i des deux configurations étudiées. On a : $(LV_1 - LV_2) = \sum_{i=1,n}(LV_{1i} - LV_{2i})$ où n est le nombre de sites de l'alignement.

La méthode consiste à tirer avec remise n écarts $(LV_{1i} - LV_{2i})^*$, à sommer ces écarts puis à recentrer, suivant la formule :

$$(LV_1 - LV_2)^* = \left[\sum_{i=1,n}(LV_{1i} - LV_{2i})^*\right] - (LV_1 - LV_2).$$

Le recentrage obtenu en soustrayant le terme $(LV_1 - LV_2)$ permet de se conformer à l'hypothèse nulle selon laquelle les deux configurations seraient équivalentes et la différence observée entre leurs deux log-vraisemblances ne serait due qu'aux fluctuations des log-vraisemblances des sites. Ainsi l'espérance de l'écart $(LV_1 - LV_2)^*$ dans l'hypothèse nulle est bien égale à zéro. En répétant un grand nombre de fois ces opérations, on obtient la distribution des écarts bootstrap $(LV_1 - LV_2)^*$, à laquelle est comparée l'écart observé $(LV_1 - LV_2)$. Si cet écart tombe dans la queue de la distribution, avec une valeur typiquement plus grande que les quantiles à 95% ou 99%, alors l'écart est significatif et la branche est supportée. La valeur de support de la branche est donnée par la fonction de répartition empirique des écarts bootstrap $(LV_1 - LV_2)^*$.

Cette méthode s'apparente au test de Shimodaira et Hasegawa, dit test SH, pour comparer des phylogénies (Shimodaira et Hasegawa 1999), d'où le nom de support 'aLRT SH-like'. On utilise ici comme dans le test SH un ré-échantillonnage des log-vraisemblances par site. Ceci est particulièrement rapide car ces valeurs ont déjà été calculées. Il n'est pas nécessaire de repasser par une étape d'inférence phylogénétique, comme dans le bootstrap de Felsenstein (voir ci-dessous). On parle de bootstrap RELL (Kishino et al. 1990). La méthode se distingue du test SH par sa statistique de décision. Dans le test SH l'objectif est de lister les phylogénies qui sont statistiquement équivalentes à la phylogénie la plus vraisemblable. Ici l'objectif

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

est de comparer la meilleure phylogénie à la deuxième meilleure, mais on ne s'intéresse pas à la troisième configuration NNI.

Il y a cependant un coût calcul que nous n'avons pas évoqué jusqu'ici : en principe dans ces tests toutes les log-vraisemblances $(LV_1, LV_2, LV_3$ et $LV_0)$ doivent correspondre au maximum de vraisemblance en termes de longueurs de branche. C'est le cas pour LV_1, puisqu'il s'agit de la log-vraisemblance de l'arbre *T*. Mais les autres log-vraisemblances sont issues de mouvements topologiques NNI appliqués à *T*, ou de la mise à 0 de la branche *b* pour LV_0, et donc en principe il faudrait ré-optimiser toutes les longueurs des branches de l'arbre *T* ainsi transformé. Cela implique des calculs considérables. En pratique nous avons montré qu'il suffit d'optimiser 5 branches, la branche centrale et les 4 branches adjacentes (cf. Fig. 1.1), pour avoir une excellente approximation des log-vraisemblances autres que LV_1. C'est l'approche implémentée dans PhyML. Elle permet à la fois de calculer les supports de branche et de s'assurer que les configurations NNI retenues pour chaque branche sont bien localement optimales et meilleures que les configurations NNI alternatives.

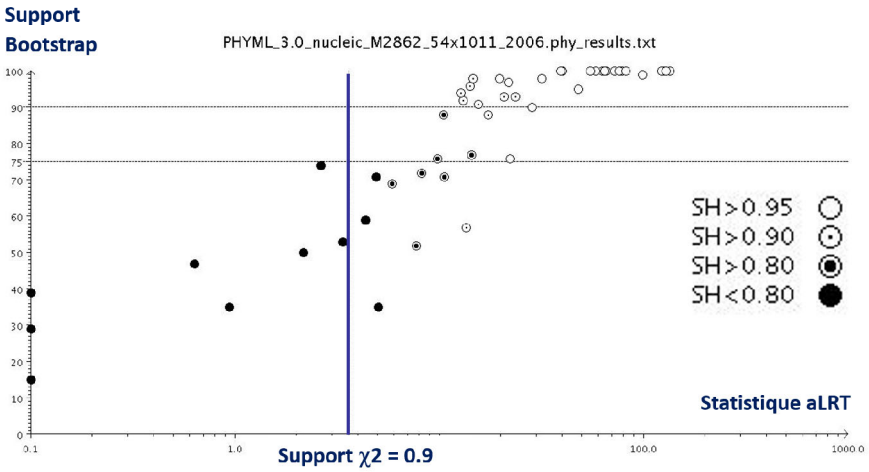

Figure 1.2. *Comparaison des supports bootstrap (axe vertical) et de la statistique aLRT (axe horizontal) sur un jeu de données de 54 séquences et 1011 sites nucléiques, avec un arbre inféré par PhyML. Chaque point correspond à une branche et ses différents supports. Le support aLRT à 90% en utilisant le mélange de χ^2 et une correction de Bonferroni est indiqué par un trait bleu. Le support SH-like est montré avec les différents types de points. Le signal dans ce jeu de données est globalement élevé et les supports sont bien corrélés.*

1.2.4 *Comparaison sur un exemple des supports aLRT et bootstrap*

La Figure 1.2 compare le bootstrap de Felsenstein et les supports aLRT paramétrique et non-paramétrique sur un alignement de 54 séquences d'ADN comportant 1011 sites, avec un signal phylogénétique relativement élevé (extrait de Treebase, référence M2862, cf. Guindon et al. 2010). Globalement, on observe une bonne corrélation entre les trois approches. Un support bootstrap supérieur à 75% (une valeur communément utilisée) et un support SH-like supérieur à 90% (comme classique pour un test) sont essentiellement équivalents, à une branche près dont le support bootstrap est à environ 60%. Entre les deux approches locales, basées sur la même statistique de décision $\left(LV_1 - LV_2\right)$, l'approche SH-like (non-paramétrique) est plus conservative que l'approche (paramétrique) reposant sur le χ^2. L'écart est plus grand encore entre le bootstrap, connu pour être conservatif, et la version paramétrique du support aLRT, avec une dizaine de branches ayant un support χ^2 supérieur à 90% mais un support bootstrap inférieur à 80%, et parfois de beaucoup. Globalement ces résultats militent pour l'utilisation du support aLRT SH-like, plutôt que sa version paramétrique. Ces tests locaux ayant un faible coût calcul, ils seront d'autant plus utilisés que le jeu de données est important et contient un grand nombre de taxons.

On voit aussi une caractéristique importante de certaines implémentations du bootstrap, corrigée autant que possible dans PhyML, mais apparente malgré tout dans la Figure 1.2 : alors que certaines branches ont une statistique aLRT nulle en raison d'une longueur de branche interne nulle, elles ont un support bootstrap non négligeable, ce qui est une forme de non-sens. En effet, une longueur nulle signifie qu'aucune mutation ne supporte la branche, et donc une telle branche ne devrait être supportée par aucune mesure et en particulier par le bootstrap. La raison de ce phénomène vient d'un déterminisme caché des algorithmes de construction d'arbres, qui tendent à construire le même arbre avec des jeux de données différents. Ce déterminisme est par exemple réduit lorsqu'on mélange l'ordre des taxons dans les alignements bootstrap (comme implémenté dans PhyML). D'autres facteurs plus profonds mais du même ordre rentrent sans doute en jeu. On verra (Fig. 1.3) que ceux-ci sont très sensibles dans les implémentations rapides du bootstrap.

1.3 Le bootstrap phylogénétique

1.3.1 Le bootstrap statistique

Le bootstrap est une méthode de statistique computationnelle aujourd'hui largement utilisée. Dans le contexte qui nous intéresse ici, c'est une approche non-

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

paramétrique (il existe aussi un "bootstrap paramétrique", bien différent). L'objectif est d'étudier la distribution d'une statistique S (par exemple une variance ou un quantile) estimée à partir d'un n-échantillon d'observations ou unités statistiques, noté $X^n = \left\{ X^n_{i=1,\ldots n} \right\}$. On note $S\left(X^n\right)$ la statistique calculée sur ce n-échantillon, qu'on suppose tiré aléatoirement et indépendamment dans la distribution théorique inconnue des observables. Lorsque le modèle est un tant soit peu complexe, on dispose difficilement de formules analytiques pour la distribution de $S\left(X^n\right)$ et on fait typiquement appel à la méthode du bootstrap, étudiée et popularisée par Bradley Efron à la fin des années 70 (Efron 1979 ; Efron et Tibshirani 1993). Le principe est simple : on tire avec remise n observations de X^n pour obtenir un n-échantillon bootstrap ou pseudo-échantillon $X^{n*} = \left\{ X^{n*}_{i=1,\ldots n} \right\}$, et on calcule $S\left(X^{n*}\right)$. En répétant cette opération un certain nombre de fois (typiquement 100 à 1 000), on obtient une distribution dont on peut, par exemple, calculer la variance ou un quantile, et on utilise ces valeurs pour quantifier la variabilité de $S\left(X^n\right)$. On fait ici l'hypothèse que la distribution théorique des observables est bien approximée par la distribution empirique définie par les observées X^n_i. On peut ainsi estimer non seulement la variabilité de $S\left(X^n\right)$, mais aussi le biais induit par l'usage d'un n-échantillon, en se basant sur l'écart $E\left(S\left(X^{n*}\right)\right) - S\left(X^n\right)$ pour estimer le biais $E\left(S\left(X^n\right)\right) - S\left(X\right)$, où $S\left(X\right)$ représente la valeur de la statistique sur l'ensemble des observables. C'est ce qu'on appelle le "plug-in principle". Il s'agit d'un principe général et non d'un théorème, qui ne peut être démontré sans faire des hypothèses très précises sur X, S etc. Cependant, il est clair que plus n est grand, plus l'approche est fondée car la distribution empirique des observables se rapproche alors de la distribution théorique.

1.3.2 Le bootstrap de Felsenstein

Le bootstrap de Felsenstein reprend ces grands principes. Les données sont ici un alignement multiple de taille n (nombre de sites) × s (nombre de séquences). Chaque site (ou colonne) de l'alignement est considérée comme une unité statistique. On fait l'hypothèse de l'indépendance des sites, comme dans le bootstrap RELL (voir plus haut) et dans de nombreux modèles et programmes de phylogénétique (cf. Chapitre XX). L'approche consiste à tirer avec remise n sites de l'alignement initial, reconstruire un arbre à partir de l'alignement bootstrap ainsi obtenu, et comparer l'arbre bootstrap T^* et l'arbre original T obtenu sur l'alignement initial. Cette opération est répétée un grand nombre de fois, 100 au minimum, 1 000 de préférence si les temps calcul sont raisonnables. En pratique, l'inférence de chaque arbre T^* peut nécessiter un temps calcul relativement long, mais il est possible de paralléliser l'inférence des différents arbres bootstrap et d'utiliser les approches

rapides (Stamatakis 2008 ; Minh et al. 2013), qui utilisent certains arbres trouvés au cours de la recherche topologique (cf. Chapitre XX) comme arbres bootstrap.

Dans ce cadre, le support d'une branche ou bipartition de l'arbre T, est tout simplement la proportion de fois où cette bipartition est présente parmi les arbres bootstrap T^*. Plus formellement, l'arbre T définit un ensemble de bipartitions $B = \{b\}$, où une bipartition b découpe l'ensemble des taxons en deux sous-ensembles disjoints non-vides. De même, T^* définit $B^* = \{b^*\}$. Pour une branche b de B et un arbre T^* on utilise l'indicatrice $I_{b \in B^*}$ qui vaut 1 si b est retrouvée à l'identique dans B^*, et 0 sinon. Le support de b est alors :

$$FBP(b) = E_{T^*}\left(I_{b \in B^*}\right),$$

où l'espérance (ou moyenne) E_{T^*} est calculée sur l'ensemble des arbres bootstrap dont on dispose, et où *FBP* signifie "Felsenstein bootstrap proportion".

Cette procédure se comprend bien intuitivement. En tirant par bootstrap des pseudo-alignements, on obtient des alignements analogues à l'alignement initial, à partir desquels on infère des arbres qui sont censés ressembler à l'arbre initial. S'ils en diffèrent nettement, le processus d'inférence apparaît instable et le support de l'arbre initial est nécessairement faible. On cherche ainsi à mesurer la robustesse de l'arbre initial et de ses branches. Dit autrement, on fait ici l'hypothèse que l'alignement initial est issu du tirage de n sites dans un alignement plus grand, typiquement à l'échelle génomique, et qu'en tirant avec remise un n-échantillon bootstrap on simule le tirage de n sites dans l'alignement "génomique". On mesure ainsi la reproductibilité de l'inférence de chacune des branches de l'arbre initial. Il est rapidement apparu que cette robustesse ou reproductibilité n'est pas le garant d'une probabilité élevée que la branche considérée corresponde à une vraie branche de l'arbre évolutif des espèces considérées. En effet, si la méthode de reconstruction a un biais (par exemple d'attraction des longues branches, cf. Chapitre XX), les arbres bootstrap seront frappés du même biais et une branche erronée issue de ce biais pourra être fortement supportée.

Cette interprétation simple des supports *FBP* en termes de robustesse a été la première. Dans les années 90, un débat passionné s'est engagé sur le sens mathématique des supports *FBP*, l'objectif étant de faire le lien entre les supports *FBP* et la valeur de confiance d'un test statistique au sens usuel (Hillis et Bull 1993 ; Felsenstein et Kishino 1993). Mais rien de simple ici, puisque pour commencer il faut définir les hypothèses nulle et alternative du test, ce qui ne tombe pas sous le sens lorsqu'on parle de branches et d'arbres. Des méthodes statistiques sophistiquées ont été proposées pour corriger *FBP* et se rapprocher de ce qu'on

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

entend communément par test, p-valeur, etc. (Efron et al. 1996 ; Susko 2010). Dans les années 2000 ces approches ont montré leurs limites (Susko 2009). Elles sont aujourd'hui abandonnées pour l'essentiel, et on est revenu à l'approche standard de Felsenstein et à la robustesse, avec l'idée que les supports *FBP* sont très conservatifs, et qu'une valeur supérieure ou égale à 70% constitue un support de branche solide (Soltis et Soltis 2003).

Dans les études récentes de phylogénomique (Chapitre XX) on combine plusieurs gènes, parfois des centaines voire des milliers, pour constituer un très grand alignement en termes de nombre de sites s. Ce très grand alignement est destiné à inférer l'arbre des espèces étudiées. Dans cette configuration, les supports de branche *FBP* sont pratiquement tous égaux à 1, en raison du nombre élevé d'unités statistiques. Ré-échantillonner de telles données ne change guère le résultat, on retrouve systématiquement les mêmes branches ou presque. Pour autant, les branches de l'arbre des espèces ne sont pas nécessairement correctes et peuvent refléter un biais de la méthode d'inférence. Nous reviendrons dans la partie Discussion sur ce point et les méthodes proposées dans ce cadre.

1.3.3 Le bootstrap de transfert

Dans le cas opposé d'alignements très grands en termes de nombre n de séquences, se pose le problème inverse : les supports de branche deviennent faibles voire très faibles, même si les données contiennent un signal phylogénétique fort. La Figure 1.3 donne un exemple de ce phénomène sur des données de VIH (Lemoine et al. 2018). On a ici plus de 9 000 séquences du gène *pol* ($s \approx 1\,050$). Les séquences ont été classées par sous-types à l'aide d'une méthode directe sans construction d'arbre (Schultz et al. 2009). Celle-ci, outre les sous-types standards du VIH, détecte une cinquantaine de séquences recombinantes dans le jeu de données. La phylogénie sépare remarquablement bien les sous-types, mais les clades qui leur correspondent sont pourtant très peu supportés par *FBP*. Par exemple, le clade correspondant au sous-type B contient toutes les séquences assignées au sous-type B (~3 500), plus 2 séquences recombinantes, et ce clade presque parfait ne reçoit qu'un support *FBP* de 3%. L'explication est simple. Elle est liée à ce qu'on appelle les « rogue taxa » ou taxons instables. Certaines séquences (typiquement recombinantes dans l'exemple du VIH) ont des positions instables dans l'arbre et changent de position dans les arbres bootstrap. Or, pour qu'une bipartition bootstrap b^* induite par T^* soit comptabilisée présente dans *FBP*, elle doit être rigoureusement identique à la bipartition b étudiée. Il suffit d'un seul taxon mal placé pour que b^* ne soit pas comptabilisée. Avec de grands jeux de données (au sens de n), la présence de tels taxons est presque inévitable. Elle peut avoir des causes multiples : recombinaison,

transfert horizontal, séquences erronée ou incomplète, erreur de reconstruction due aux approximations numériques et algorithmiques, etc.

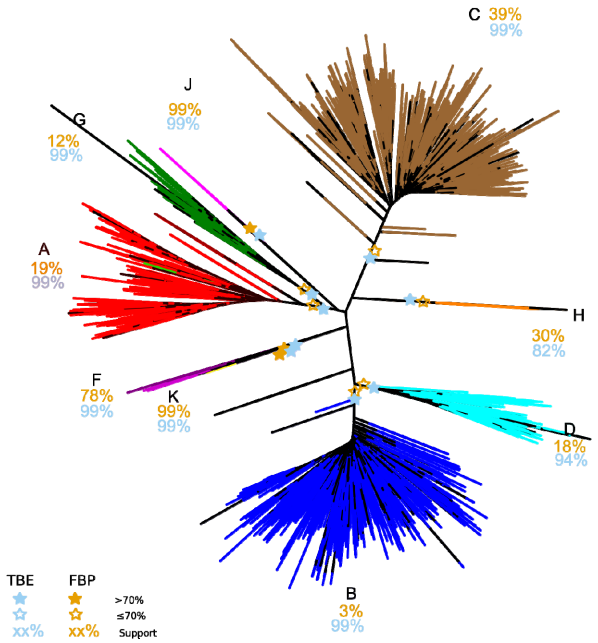

Figure 1.3. *Comparaison des supports de Felsenstein (*FBP*) et de transfert (*TBE*) sur une phylogénie portant sur 9147 souches du VIH-SIDA (logiciel FastTree, VIH-1 groupe M, gène pol). Les neuf sous-types du VIH1-M (A, B, C, D, F, G, H, J, K, dont les souches sont colorisées différemment, par exemple bleu pour le B) sont clairement séparés et leur branchement est conforme à celui des études antérieures. Ces neuf sous-types ont un support* TBE *élevé, souvent proche de 100%. A l'inverse le support FBP est plus faible, en particulier pour les sous-types les plus prévalents (A, B, C, D, G). Par exemple,* FBP *ne « voit » pas le cluster du sous-type B (majoritaire en Europe et Amérique du Nord,* FBP *= 3%) alors que celui-ci est très clairement identifié (il contient toutes les souches bleues plus deux recombinants en noir) et obtient un support de 99% avec* TBE.

L'idée qui préside au bootstrap de transfert est naturelle : au lieu d'utiliser une indicatrice en 0/1 comme dans le bootstrap de Felsenstein, on utilise une mesure quasi-continue dans $[0,1]$ de la présence de la branche étudiée b dans l'ensemble B^*

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

des bipartitions induites par T^*. Cette mesure repose sur la distance de transfert, introduite en classification et phylogénétique par (Day 1981 ; Lin et al. 2012). La distance de transfert $\delta(b,b^*)$ entre deux bipartitions b et b^* du même ensemble de taxons est égale au nombre minimum de taxons qu'il faut transférer d'un coté à l'autre de la bipartition b pour qu'elle devienne égale à b^*. Les rôles sont symétriques, on peut comptabiliser le nombre de taxons à transférer dans b^* pour qu'elle devienne égale à b, ou encore le nombre minimum de taxons qu'on doit retirer de l'ensemble de tous les taxons pour que les deux bipartitions deviennent égales. Une définition alternative est basée sur la distance de Hamming notée H. En ordonnant l'ensemble des taxons, on peut encoder une bipartition par une vecteur 0/1 de longueur n. On note v le vecteur codant pour b, et $\underline{v}$ le vecteur inverse où les 1 deviennent des 0 et vice et versa ; $\underline{v}$ code aussi pour b ; on note v^* le vecteur associé à b^*. On a alors :

$$\delta(b,b^*) = Min\left\{H(v,v^*), H(\underline{v},v^*)\right\}.$$

L'index de transfert $\phi(b,T^*)$ mesure la présence d'une bipartition b de T dans l'ensemble de bipartitions B^* induit par T^* (y compris les bipartitions triviales séparant un taxon des n-1 autres taxa). L'index de transfert est le minimum des distances de transfert pour toutes les bipartitions de B^* :

$$\phi(b,T^*) = Min_{b^*\in B^*}\left\{\delta(b,b^*)\right\}.$$

Cet index est normalisé dans [0,1] et moyenné (comme *FBP*) sur l'ensemble des arbres bootstrap. La normalisation repose sur le résultat suivant. Une bipartition b divise l'ensemble des taxons en deux sous-ensembles. On appelle p le cardinal du plus petit de ces deux sous-ensembles, qui est noté P $(p \le n-p)$. On a alors $\phi(b,T^*) \le p-1$. En effet, en prenant une bipartition triviale b^* séparant un taxon de P de tous les autres taxa, on a $\delta(b,b^*) = p-1$. La forme normalisée de l'index de transfert est égale $1-\phi(b,T^*)/(p-1)$, et le support de transfert *TBE* (*Transfert Bootstrap Expectation*) est défini par :

$$TBE(b) = E_{T^*}\left(1-\frac{\phi(b,T^*)}{p-1}\right).$$

Le support de transfert *TBE* ainsi défini appartient à $[0,1]$ par construction et possède nombre de bonnes propriétés qui en font un support de branche pertinent :

- $TBE(b) = 1$ si et seulement si b appartient à tous les arbres bootstrap T^* (tout comme *FBP*).

- $TBE(b) \approx 0$ lorsqu'il n'y a pas de signal dans les données (à nouveau comme *FBP*). Ce résultat asymptotique (Davila Felipe et al. 2019) établit que si les arbres T^* sont aléatoires, alors le support *TBE* tend vers 0 lorsque n augmente, et que cette convergence est rapide.

- $TBE(b) \geq FBP(b)$ lorsqu'on utilise le même ensemble d'arbres bootstrap. La différence entre les deux supports dépend de p et donc de la profondeur de la branche b. Lorsque p est une cerise $(p = 2)$, les deux supports sont égaux. A l'inverse, lorsque la branche est profonde et que p se rapproche de $n/2$, il y a souvent un grand écart entre les deux supports. La Figure 1.3 sur les sous-types du VIH en est une bonne illustration. Alors que *FBP* vaut 3% pour le sous-type B, *TBE* vaut 99%, comme on s'y attend intuitivement vu la quasi perfection de ce clade (cf. plus haut). La même observation se retrouve sur les autres sous-types. Il faut noter que chaque support a ici sa logique et que chacun a "raison" à sa manière : *FBP* en disant que le clade B est rarement retrouvé à l'identique, *TBE* en disant qu'on retrouve généralement le clade B de manière quasi-parfaite et qu'il y a un fort signal dans les données (ce à quoi *FBP* est aveugle).

- La propriété ci-dessus est souhaitable pour les grands jeux de données (en nombre de taxons n), mais il faut vérifier que *TBE* ne soutient pas de branches largement erronées. C'est ce que montre les résultats (Lemoine et al. 2018) sur de grands jeux de données simulées et sur des données de mammifères, pour lesquelles on a de bonnes connaissances taxonomiques, disponibles sur le site du NCBI (https://www.ncbi.nlm.nih.gov/taxonomy).

- *TBE* s'interprète naturellement comme une fraction de taxons à déplacer (ou retirer de l'étude). Il diffère en ceci radicalement de *FBP*, qui représente la fraction des arbres bootstrap en accord avec l'arbre de référence pour la branche considérée. Sur les exemples que nous avons étudiés (mammifères, HIV, arbres simulées), les taxons instables qui impactent les supports *TBE* se distinguent des autres et peuvent être remis en question. C'est par exemple le cas des séquences recombinantes du VIH, qui sont bien détectées par l'approche (Lemoine et al. 2018).

- Finalement, *TBE* se calcule rapidement, en $O(n^2)$ dans le pire cas pour comparer toutes les branches de T à toutes les branche d'un arbre bootstrap T^* donné. Le calcul est presque aussi rapide que pour *FBP* qui utilise du hachage pour accélérer les comparaisons. En pratique, pour ces deux supports le temps calcul est négligeable par rapport à celui nécessaire pour inférer l'ensemble des arbres bootstrap, même en utilisant des approches rapides (Stamatakis 2008 ; Minh et al. 2013).

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

1.3.4 Comparaison sur un exemple des supports bootstrap

Nous reprenons ici les données de la Figure 1.2, portant sur 54 séquences d'ADN d'environ 1 000 sites, avec un signal phylogénétique relativement élevé.

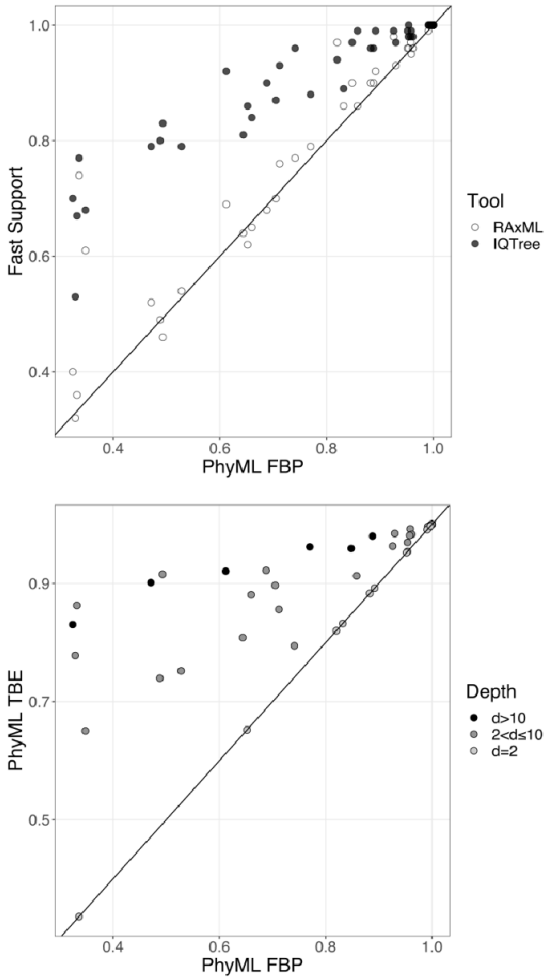

Figure 1.4. *En haut, comparaison des supports bootstrap de Felsenstein obtenus avec PhyML en suivant l'approche classique (*FBP), *avec les bootstrap rapides proposés par iQTree et RAxML, en combinant échantillonnage et recherche topologique. Chaque point représente une branche. En bas, comparaison des supports* FBP *avec les supports du bootstrap de transfert* TBE, *dans une implémentation standard (PhyML).*

Les méthodes rapides disponibles dans RAxML et iQTree (Stamatakis 2008 ; Minh et al. 2013) accélèrent les calculs des arbres bootstrap en ré-échantillonnant un nombre limité de fois les sites de l'alignement, et en proposant comme arbres bootstrap des arbres trouvés au cours des recherches topologiques sur les alignements bootstrap. On voit dans la Figure 1.4 que ces approches rapides n'aboutissent pas aux mêmes supports que l'approche classique, plus lente, où chaque arbre bootstrap est issu d'un alignement bootstrap indépendant. En particulier iQTree, très rapide, propose des supports de branche beaucoup plus élevés, là où on attendrait des supports quasi-identiques pour les différentes approches qui reposent toutes sur le même principe statistique.

Un autre défaut déjà signalé est le déterminisme caché de certains algorithmes, là où on s'attendrait à un comportement aléatoire en raison de l'absence de signal dans les données. Ainsi, une branche de longueur nulle ou quasi-nulle, donc supportée par aucune mutation, induit une trifurcation dont la résolution aléatoire devrait recevoir un support bootstrap d'environ 1/3. Trois branches ont des longueurs quasi-nulles pour ce jeu de données et pour les trois logiciels. Elles obtiennent bien des supports de ~1/3 avec PhyML qui randomise l'ordre des séquences pour chaque échantillon bootstrap, mais des supports bien plus élevés avec les bootstraps rapides de RAxML et iQTree (respectivement ~0.55 et ~0.65, en moyenne). Ce déterminisme caché explique vraisemblablement les supports bootstrap très élevés de iQTree, et dans une moindre mesure ceux de RAxML.

A l'inverse, lorsque l'on compare les supports du bootstrap de Felsenstein et du bootstrap de transfert, on ne s'attend pas aux mêmes résultats, les deux supports ayant des bases et des interprétations très différentes. Comme prévu, on voit dans la Figure 1.4 que les branches profondes tendent à être mieux supportées par *TBE*, tandis que les deux supports sont identiques sur les cerises qui correspondent à un sous-arbre ne comportant que 2 feuilles ($d = p = 2$, où p est la taille du plus petit sous-ensemble induit par la bipartition considérée, voir plus haut). En revenant aux interprétations de ces supports, sur une cerise il est équivalent de dire que la branche est incorrecte *(FBP*) que de dire qu'il faut déplacer 1 taxon de la cerise pour retrouver une bipartition (triviale) de T^* ; dans les deux cas le support vaut 0, alors qu'il vaut 1 si la cerise en question est retrouvée dans T^*. A l'inverse, sur les branches profondes $\left(2 \ll p \leq n/2\right)$ il peut suffire de déplacer 1 seul taxon pour rendre deux bipartitions identiques, auquel cas le support *TBE* sera élevé tandis que *FBP* sera nul.

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

1.4 Supports Bayésiens

1.4.1 Principe, utilisation de chaines de Markov Monte-Carlo

Les principes et algorithmes propres aux méthodes Bayésiennes sont décrits dans le Chapitre XX. Nous les résumons ici pour expliquer les bases sur lesquelles reposent les probabilités postérieures qui servent de support de branche dans le cadre Bayésien.

Les méthodes d'inférence discutées jusqu'ici estiment la topologie de l'arbre et les valeurs des paramètres numériques (longueurs de branches, modèle substitutionnel, etc.) par maximisation de la vraisemblance. Les méthodes Bayésiennes reposent sur la probabilité postérieure, qui est le produit de la vraisemblance par la probabilité a priori sur les paramètres du modèle. En phylogénie cet a priori peut porter sur : les paramètres numériques, par exemple les fréquences d'équilibre des nucléotides ou acides aminés ; les dates de certains nœuds ancestraux, lorsque l'objectif est de faire de la datation moléculaire ; sur l'arbre lui-même parfois, lorsqu'on estime que celui-ci résulte d'un processus bien identifié et modélisé, par exemple en épidémiologie ou en écologie. En l'absence d'information dans les données, les probabilités a priori et postérieures sont identiques. A l'inverse, une forte information dans les données minimise le rôle de l'a priori. En pratique, dans la plupart des problèmes de phylogénie on n'a pas ou peu d'a priori, et maximiser la vraisemblance ou la postérieure change très peu voire pas du tout les résultats. Concrètement, utiliser PhyML, RAxML ou MrBayes conduit essentiellement au même arbre dans la plupart des cas, dès lors qu'il y a un signal clair eu égard au nombre de taxons étudiés et que l'arbre inféré ne résulte pas de choix effectués au hasard par les programmes en question.

La différence entre les deux approches est ailleurs, du moins dans le cadre de la phylogénie. Les méthodes usuelles de vraisemblance procèdent par optimisation directe dite en « horizon limité » ou « gloutonne ». Cela signifie qu'elles partent d'un arbre initial et qu'elles le changent localement sur le plan topologique, ou plus globalement pour les longueurs de branches ; si le changement effectué augmente la vraisemblance, cet arbre devient l'arbre courant et le processus est répété ; s'il est impossible par ce mécanisme d'améliorer significativement l'arbre courant, celui-ci constitue le résultat qui est renvoyé à l'utilisateur (voir Chapitre XXX pour plus de détails).

Dans le cadre Bayésien tel qu'utilisé en phylogénie, on procède différemment en explorant plus systématiquement l'espace des arbres (il existe des domaines où on procède aussi par optimisation directe et gloutonne de la postérieure). Cette

exploration Bayésienne repose sur l'utilisation de chaines de Markov Monte Carlo. L'arbre courant subi une modification locale, par exemple un changement de longueur de branche ou un mouvement topologique de type NNI (Fig. 1.1) ; ce changement est accepté ou refusé en considérant le rapport (de Metropolis) des probabilités postérieures des deux configurations, ainsi que le rapport (de Hasting, lié à l'implémentation des changements et non aux données) de la probabilité du changement effectué sur la probabilité du changement inverse ; si le produit (de Metropolis-Hasting) des deux rapports a une valeur supérieure à 1, la nouvelle configuration améliore la configuration courante et elle est acceptée ; sinon, elle est acceptée avec une probabilité égale au rapport de Metropolis-Hasting. On est dans une logique proche du recuit simulé. En procédant ainsi on explore plus systématiquement l'espace des arbres et on évite les optima locaux si la chaine est suffisamment longue. La grande force de cette approche vient de ce que la fréquence d'un évènement particulier observé le long de la chaine approxime la postérieure de cet évènement, et cette approximation est d'autant plus précise que la chaine de Markov Monte-Carlo est longue.

En particulier, cet événement peut être la présence d'une branche dans l'arbre courant. Avec une chaine suffisamment longue, on sera donc à même de quantifier la postérieure de toutes les bipartitions de l'ensemble des taxons, notamment celles qui correspondent aux branches de l'arbre fourni en sortie de la méthode. Les probabilités postérieures ainsi estimées constituent les supports de branches usuels dans un cadre Bayésien. On les obtient en échantillonnant les arbres au sein de la chaîne, typiquement en considérant 1 arbre sur 1000 de manière à éviter les corrélations entre arbres consécutifs ou trop proches dans la chaine.

Une des forces de cette approche est que ces supports sont directement interprétables en tant que probabilité postérieure d'observer telle ou telle branche, y compris pour des branches qui n'appartiennent pas à l'arbre fourni en sortie de la méthode. Il y a cependant deux limites qui doivent être bien comprises. La première est que cette approche est profondément paramétrique : en changeant le modèle de substitution ou les probabilités a priori, on peut changer considérablement les supports de branche. Il importe donc de faire des études de sensibilité au modèle et aux a priori. Une autre limite est que les chaines de Markov Monte Carlo sont en pratique trop courtes pour avoir des garanties solides de convergence vers la postérieure. Il faut donc répéter plusieurs fois l'algorithme avec des points de départ différents, utiliser des chaines « chaudes » pour accélérer la convergence, etc., et vérifier ainsi la stabilité des résultats.

Une littérature abondante (Douady et al. 2003, par exemple) traite des différences entre les supports Bayésiens et le bootstrap de Felsenstein. Le consensus

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

est qu'en raison du caractère paramétrique et de chaines de Markov Monte-Carlo nécessairement limitées en longueur, les supports Bayésiens sont généralement libéraux et tendent à supporter à l'excès certaines branches. Ils s'opposent en ceci au bootstrap de Felsenstein, considéré comme conservateur. De nombreux auteurs produisent les deux supports avec l'idée que la « vérité se situe entre les deux », ce qui est un peu simpliste car ces deux supports ne reposent pas sur les mêmes bases et ne mesurent pas la même chose. Mais c'est une approche raisonnable, de même que celle qui consiste à combiner bootstrap et aLRT, chacun corrigeant les défauts de l'autre (voir plus haut).

1.4.2 Support Bayésien local

L'idée du support Bayésien local, noté aBayes pour "approximate Bayes" (Anisimova et al. 2011), est directement inspirée du support aLRT (1.2.2). Les trois configurations NNI de la Figure 1.1 ont pour vraisemblances V_1, V_2, V_3, où V_1 est la vraisemblance la plus élevée correspondant à la branche inférée dont on évalue le support. Il s'agit bien de vraisemblances $\left(V_i\right)$, et non de log-vraisemblances $\left(LV_i\right)$ comme dans aLRT, et celles-ci reposent à nouveau (cf. 1.2.2) sur une optimisation locale de 5 longueurs de branche : la branche d'intérêt et ses 4 voisines (Fig. 1.1). Les trois vraisemblances V_1, V_2, V_3 une fois normalisées peuvent s'interpréter comme des probabilités postérieures, car elles incorporent les probabilités a priori sur les données (typiquement les probabilités de chaque base ou chaque acide aminé, cf. Chapitre XXX). Aussi, une approximation du support Bayésien est donnée par le ratio :

$$aBayes\left(b\right) = \frac{V_1}{V_1 + V_2 + V_3}.$$

Comme pour aLRT ce support se calcule remarquablement vite, sans beaucoup impacter le temps calcul total nécessaire à l'estimation de la topologie de l'arbre et des longueurs de branche. Dans PhyML, ces calculs locaux sont aussi utilisés pour améliorer la topologie de l'arbre inféré, lorsque qu'une alternative NNI apparaît meilleure que la configuration courante.

1.4.3 Comparaison des supports Bayésiens sur un exemple

Nous reprenons à nouveau le jeu de données des Figures 1.2 et 1.4, portant sur un alignement de 54 séquences d'ADN de longueur 1011, avec un signal phylogénétique relativement élevé.

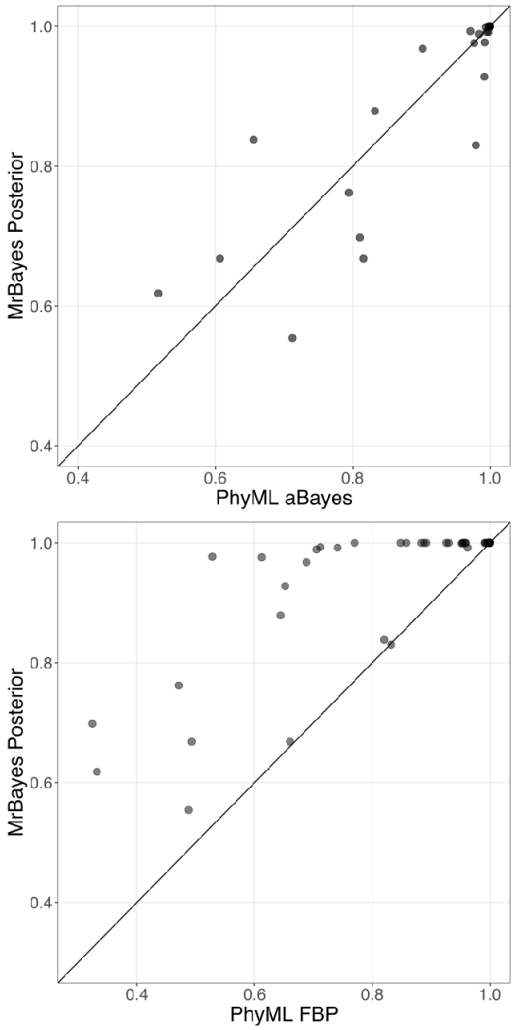

Figure 1.5. *En haut, comparaison du support local Bayésien approché (PhyML aBayes) et des postérieures Bayésiennes (MrBayes Posterior). Chaque point représente une branche. En bas, comparaison du bootstrap de Felsenstein (PhyML FBP) et des postérieures Bayésiennes (MrBayes Posterior).*

On voit (Fig. 1.5) que les supports approchés aBayes (PhyML) sont proches des postérieures Bayésiennes (MrBayes), sans biais particulier et avec une corrélation élevée. Ces supports approchés sont pourtant nettement plus rapides à calculer, et peuvent être utilisés sur de très grands arbres comportant des milliers voire des

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

dizaines de milliers de feuilles, ce qui est juste impossible pour les méthodes Bayésiennes. On voit aussi (Fig. 1.5, en bas), que les supports Bayésiens sont nettement plus libéraux que le bootstrap de Felsenstein. Ce point est rapporté dans de nombreux articles de la littérature, qui indiquent aussi une grande sensibilité des postérieures Bayésiennes au choix de modèle aux a priori (Yang et Rannala 2005). En changeant le modèle ou les a priori, même de peu, on peut obtenir non seulement un arbre différent, mais des supports très différents sur les mêmes branches (ou bipartitions).

1.5 Discussion

Nous avons présenté trois grandes approches et leurs variantes pour mesurer les supports des branches d'une phylogénie. La première leçon à garder en mémoire est que ces approches reposent sur des principes mathématiques différents et que les supports qu'elles produisent ne peuvent être interprétés de la même manière. Les supports locaux (aLRT, aBayes) indiquent si la configuration NNI retenue autour de la branche d'intérêt est significativement meilleure que les deux alternatives. Le bootstrap de Felsenstein nous dit si en ré-échantillonnant les données, la branche d'intérêt à une probabilité élevée d'être retrouvée. Le bootstrap de transfert indique s'il existe un nombre élevé ou restreint de taxons instables vis-à-vis de la branche en question, et nous permet d'identifier les taxons les plus instables. Les supports Bayésiens donnent la probabilité postérieure de la branche étudiée, qui dépend de la quantité d'information contenue dans les données en faveur de cette branche. Dans tous les cas il convient de prendre des précautions, car tous ces supports dépendent du modèle substitutionnel, nécessairement simplificateur. Certaines méthodes y sont très sensibles, en particulier les approches Bayésiennes. Egalement, les biais des méthodes et des algorithmes font que des supports importants peuvent être apportés à des branches erronées, par exemple en raison de l'attraction des longues branches.

Il est de plus en plus commun dans les publications de fournir pour chaque branche les supports de différentes méthodes, typiquement aLRT, bootstrap de Felsenstein et probabilité a posteriori. C'est une bonne pratique, qui permet pour une part de combler les lacunes de certaines méthodes, par exemple le support excessif des postérieures Bayésiennes, ou le déterminisme caché de certains algorithmes d'inférence qui induit des supports bootstrap excessifs, ou encore le caractère trop local de aLRT et aBayes. Lorsque ces supports sont faibles, le bootstrap de transfert permet de vérifier si cela est dû à quelques taxons instables, ou si tout simplement il n'y a pas de signal dans les données.

Ces différents supports ont des temps calcul très différents. Une bonne approche est de conduire les études préliminaires (choix des taxons et des séquences,

alignement, nettoyage des sites, choix de modèle, etc.) avec une approche locale rapide (aLRT ou aBayes), puis lorsque l'étude et les données se précisent, de passer à une méthode plus complète et couteuse en temps calcul (bootstrap et/ou Bayésienne), qui viendra compléter les premiers supports locaux.

Nous avons évoqué les très grands jeux de données. Lorsque le nombre n de taxons est grand mais que le nombre s de sites est modéré, le bootstrap de transfert est surement une bonne approche. A l'inverse, typiquement dans les études phylogénomiques associant de nombreux gènes, s est très grand, de l'ordre du million parfois, alors que n est modéré. Dans ce cas, on s'intéresse au biais induit par la méthode d'inférence, plutôt qu'à l'erreur d'échantillonnage qui est quasi-nulle. On est dans une problématique de sensibilité au modèle (Wang et al. 2019), plutôt que de variabilité statistique. Certains auteurs (Burleigh et al. 2006) ont proposé de bootstraper les gènes, plutôt que les sites de l'alignement global issu de la concaténation des alignements propres à chaque gène. Cette approche permet de quantifier l'impact du choix des gènes utilisés pour la reconstruction de l'arbre des espèces. D'autres auteurs ont proposé de mesurer sur chaque branche de l'arbre des espèces, l'accord des arbres de gènes quant à la branche considérée (Zhou et al. 2020), remettant ainsi en cause l'existence d'un arbre unique et prenant en compte des phénomènes comme le transfert horizontal ou la duplication et la perte de gènes.

Les perspectives de recherche dans ce domaine sont nombreuses. On observe peu d'évolution depuis pas mal d'année concernant les grands principes et les algorithmes d'inférence d'arbres. Les implémentations s'améliorent, les meilleures idées sont combinées, si bien que les logiciels se ressemblent et sur des données courantes produisent des résultats similaires. Le champ statistique du support de branche est plus ouvert. C'est un domaine complexe, comme l'a montré le débat des années 90 sur l'interprétation des supports bootstrap. Ce débat n'est toujours pas clos et les questions posées restent ouvertes et d'actualité. D'autres questions se posent naturellement, par exemple sur : l'utilisation de la distance de transfert dans un cadre Bayésien ; la mise en place de support semi-locaux mais rapides, qui généraliseraient l'approche NNI (utilisée dans aLRT et aBayes, Fig. 1.1) en regardant plus loin dans l'arbre ; la quantification simple et rapide des biais de modèle et de méthode, plutôt que l'évaluation des erreurs d'échantillonnage dont l'impact est de plus en faible avec les grands jeux de données disponibles aujourd'hui.

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

Références

Anisimova, M., & Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology*, *55*(4), 539–552. https://doi.org/10.1080/10635150600755453

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., & Gascuel, O. (2011). Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology*, *60*(5), 685–699. https://doi.org/10.1093/sysbio/syr041

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics, 7*(1), 1-26. http://www.jstor.org/stable/2958830

Burleigh, J. G., Driskell, A. C., & Sanderson, M. J. (2006). Supertree Bootstrapping Methods for Assessing Phylogenetic Variation among Genes in Genome-Scale Data Sets. *Systematic Biology*, *55*(3), 426–440. https://doi.org/10.1080/10635150500541722

Chor, B., & Tuller, T. (2006). Finding a maximum likelihood tree is hard. *Journal of the ACM*, *53*(5), 722–744. https://doi.org/10.1145/1183907.1183909

Dávila Felipe, M., Domelevo Entfellner, J.-B., Lemoine, F., Truszkowski, J., & Gascuel, O. (2019). Distribution and asymptotic behavior of the phylogenetic transfer distance. *Journal of Mathematical Biology*, *79*(2), 485–508. https://doi.org/10.1007/s00285-019-01365-0

Day, W. H. E. (1981). The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, *1*(3), 269–287. https://doi.org/10.1016/0165-4896(81)90042-1

Douady, C. J. (2003). Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability. *Molecular Biology and Evolution*, *20*(2), 248–254. https://doi.org/10.1093/molbev/msg042

Efron, B., Halloran, E., & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, *93*(23), 13429. https://doi.org/10.1073/pnas.93.23.13429

Felsenstein, J., & Kishino, H. (1993). Is there Something Wrong with the Bootstrap on Phylogenies? A Reply to Hillis and Bull. *Systematic Biology*, *42*(2), 193–200. https://doi.org/10.1093/sysbio/42.2.193

Felsenstein, Joseph. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, *39*(4), 783. https://doi.org/10.2307/2408678

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Guindon, S., & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, *52*(5), 696–704. https://doi.org/10.1080/10635150390235520

Hillis, D. M., & Bull, J. J. (1993). An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology*, *42*(2), 182–192. https://doi.org/10.1093/sysbio/42.2.182

Kishino, H., Miyata, T., & Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, *31*(2), 151–160. https://doi.org/10.1007/BF02109483

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., & Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, *556*(7702), 452–456. https://doi.org/10.1038/s41586-018-0043-0

Minh, B. Q., Nguyen, M. A. T., & von Haeseler, A. (2013). Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution*, *30*(5), 1188–1195. https://doi.org/10.1093/molbev/mst024

Schultz, A.-K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., & Stanke, M. (2009). JpHMM: Improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Research*, *37*(suppl_2), W647–W651. https://doi.org/10.1093/nar/gkp371

Shimodaira, H., & Hasegawa, M. (1999). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution*, *16*(8), 1114–1116. https://doi.org/10.1093/oxfordjournals.molbev.a026201

Soltis, D. E., & Soltis, P. S. (2003). Applying the Bootstrap in Phylogeny Reconstruction. *Statistical Science*, *18*(2), 256–267. https://doi.org/10.1214/ss/1063994980

Version auteurs, parue dans
Modèles et méthodes pour l'évolution biologique,
coordonné par Gilles DIDIER et Stephane GUINDON.
ISTE Editions 2022

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Systematic Biology*, *57*(5), 758–771. https://doi.org/10.1080/10635150802429642

Susko, E. (2009). Bootstrap Support Is Not First-Order Correct. *Systematic Biology*, *58*(2), 211–223. https://doi.org/10.1093/sysbio/syp016

Susko, E. (2010). First-Order Correct Bootstrap Support Adjustments for Splits that Allow Hypothesis Testing When Using Maximum Likelihood Estimation. *Molecular Biology and Evolution*, *27*(7), 1621–1629. https://doi.org/10.1093/molbev/msq048

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, *57*, 1–436.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*(7524), 550–553. https://doi.org/10.1038/514550a

Wang, H.-C., Susko, E., & Roger, A. J. (2019). The Relative Importance of Modeling Site Pattern Heterogeneity Versus Partition-Wise Heterotachy in Phylogenomic Inference. *Systematic Biology*, *68*(6), 1003–1019. https://doi.org/10.1093/sysbio/syz021

Y. Lin, V. Rajan, & B. M. E. Moret. (2012). A Metric for Phylogenetic Trees Based on Matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(4), 1014–1022. https://doi.org/10.1109/TCBB.2011.157

Yang, Z., & Rannala, B. (2005). Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny. *Systematic Biology*, *54*(3), 455–470. https://doi.org/10.1080/10635150590945313

Zhou, X., Lutteropp, S., Czech, L., Stamatakis, A., Looz, M. V., & Rokas, A. (2020). Quartet-Based Computations of Internode Certainty Provide Robust Measures of Phylogenetic Incongruence. *Systematic Biology*, *69*(2), 308–324. https://doi.org/10.1093/sysbio/syz058