Méthodes de distances pour estimer le taux de substitution à partir d'un ensemble de séquences hétérochrones, application au virus de l'immunodéficience humaine (VIH)

La vitesse d'évolution (mesurée par le taux de substitution) des séquences est différente d'une espèce à l'autre. Ce taux peut être estimé à l'aide de séquences échantillonnées dans le temps, ou séquences hétérochrones, lorsque le nombre de substitutions accumulées entre ces séquences est significatif. Les virus sont des candidats idéals, car ils accumulent un nombre de substitutions important en seulement quelques années. L'utilisation de ce taux trouve de nombreuses applications biologiques, comme par exemple dater l'origine d'une épidémie ou d'une infection. Les méthodes présentées dans ce chapitre sont des méthodes de distances, rapides en temps de calcul, qui estiment le taux de substitution à l'aide de séquences hétérochrones uniquement, et en faisant l'hypothèse d'une horloge moléculaire stricte, comme, par exemple, TREBLE, sUPGMA ou encore les régressions linéaires Pairwise-Distance *et* Root-to-Tip*. Enfin, deux méthodes probabilistes, basées sur des principes différents, sont présentées succinctement.*

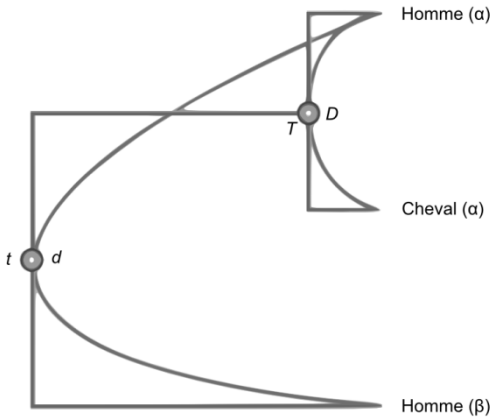
Sommaire

2.1 Introduction

Les organismes évolués, comme les mammifères, ont un processus de réplication de leur matériel génétique très sophistiqué, mais des erreurs de réplication surviennent souvent pendant ce processus (Reha-Krantz, 2010). Elles peuvent être dangereuses pour l'organisme car elles peuvent se produire sur un gène et le rendre inactif ou modifier sa fonction. Il existe certains mécanismes qui permettent de corriger ces erreurs, mais tous les organismes ne les possèdent pas (Roberts *et al*, 1988). Par exemple, ces mécanismes sont absents chez le virus de l'immunodéficience humaine (VIH) et donc, à l'intérieur d'un hôte, la population virale est constituée d'une multitude de variants génétiques changeant continuellement, quelque fois appelés des quasi-espèces (Domingo, 1998; Nowak, 1992). En partie pour cette raison, la vitesse d'évolution des organismes, identifiée par le taux de substitution, varie d'une espèce à l'autre. Cette vitesse est exprimée en nombre de substitutions par site et par unité de temps (généralement en années, jours ou générations).

En 1962 et 1965, Zuckerkandl et Pauling ont publié deux chapitres de livre fondamentaux sur la vitesse évolutive des protéines (Zuckerkandl & Pauling, 1965, 1962). Leur objectif était d'estimer la date de divergence de différentes globines. Pour cela, ils ont fait l'hypothèse d'une horloge moléculaire stricte, c'est-à-dire qu'ils ont supposé que la vitesse évolutive est constante au cours du temps et uniforme chez toutes les espèces étudiées. Cette hypothèse, ou une alternative, est essentielle pour estimer la vitesse évolutive (Kumar, 2005; Bromham & Penny, 2003). Par exemple, dans leur publication de 1962, Zuckerkandl et Pauling (1962) estimaient la date de divergence entre les hémoglobines α et β de l'homme. Ils disposaient des p-distances d et D définissant respectivement le nombre de différences observées entre les protéines de l'hémoglobine α et β de l'homme et le nombre de différences observées entre les protéines des hémoglobines α du cheval et de l'homme (Figure 9). Comme la vitesse évolutive est supposée constante et connaissant la date de divergence homme/cheval T sur la base d'estimations fossiles, ils ont estimé t à l'aide de la relation $d/t = D/T$. Ainsi, la date de divergence t entre les hémoglobines α et β de l'homme est estimée à $T(d/D)$. La vitesse évolutive dans cet exemple est donc égale à $D/2T$; le chiffre 2 venant du fait que la distance D correspond à la somme de la quantité évolutive séparant les deux espèces de leur ancêtre commun.

Figure 9. Illustration de la première utilisation d'une horloge moléculaire.
Dans leur papier de 1962, Zuckerkandl et Pauling (1962) estimaient la date de divergence t entre les hémoglobines α et β de l'homme. Pour cela, ils ont eu recourt à l'hypothèse de l'horloge moléculaire stricte qui stipule que la vitesse d'évolution est constante et uniforme. Comme la date de divergence T entre l'homme et le cheval était connue (d'après des estimations fossiles), ils ont pu estimer la date de divergence t sachant le nombre de substitutions D entre les séquences de l'hémoglobine α de l'homme et du cheval à l'aide de la relation $d/t = D/T$. Ainsi, la date t peut être estimée par $d(T/D)$.
Adaptation de Kumar (2005).



Dans cet exemple, les estimations du taux de substitution et de la date de divergence entre les hémoglobines α et β de l'homme n'étaient pas possibles sans l'information de la date de divergence entre les lignées de l'homme et du cheval. Autrement dit, les estimations nécessitent un point de calibration, limitant ainsi le nombre d'études similaires puisque les points de calibration sont généralement difficiles à obtenir et entachés d'erreurs. Toutefois, une autre source d'information temporelle peut servir à l'estimation du taux de substitution (et donc aux dates de divergence) : les dates d'échantillonnage des séquences. Mais pour qu'il soit possible d'estimer la vitesse évolutive à partir de séquences hétérochrones (séquences échantillonnées dans le temps ; à mettre en opposition avec les séquences isochrones, échantillonnées à la même date), il faut que l'accumulation de substitutions entre deux échantillons collectés à des moments différents soit significative. Les populations pour lesquelles des séquences hétérochrones peuvent être utilisées pour estimer le taux de substitution sont appelées des MEP (*measurably evolving populations*) (Drummond *et al*, 2003b). Ce terme désigne essentiellement des virus, organismes pour lesquels la vitesse évolutive est très importante et peut être mesurée à l'aide d'échantillons espacés dans le temps par seulement quelques années, comme le VIH ou le virus de la Dengue (Chen *et al*, 2011; Dunham & Holmes, 2007), ou, plus rare, des organismes dont on possède de l'ADN ancien (Lambert *et al*, 2002).

Les bases de données biologiques, et notamment celle du laboratoire national de Los Alamos sur le VIH, abondent en séquences hétérochrones. En effet, dans le cadre du VIH, le séquençage est une pratique routinière (Taylor *et al*, 2008) et, donc, des dates de prélèvement différentes sont associées aux séquences. Ces études renseignent généralement sur l'apparition de nouveaux sous-types ou formes recombinantes (Ng *et al*, 2011; Ibe *et al*, 2010), l'apparition de résistances aux traitements médicamenteux (Hanna & D'Aquila, 2001; Hirsch *et al*, 2000), l'apparition de nouvelles zoonoses

(Plantier *et al*, 2009; Damond *et al*, 2004) ou encore les stratégies de prévention comme, par exemple, la conception d'un vaccin (Gaschen *et al*, 2002). Les séquences hétérochrones, dont la quantité est en perpétuelle augmentation, sont donc des supports idéaux pour estimer le taux de substitution de virus et notamment celui du VIH.

La mesure du taux de substitution trouve de nombreuses applications biologiques. Par exemple, l'estimation de plusieurs taux de substitution différents au sein d'une même population est un indicateur dans la recherche de traitements efficaces contre les virus. Prenons le cas du VIH et supposons qu'un patient soit infecté par celui-ci. Des souches du VIH lui sont prélevées, et leur matériel génétique est séquencé en trois temps distincts t_0, t_1 et t_2 où t_1 représente la date à laquelle le patient a commencé un traitement contre le VIH, t_2 celle où le patient a été infecté et t_0 la date la plus récente, à laquelle le patient suit toujours son traitement (depuis t_1 donc). Pour pouvoir en déduire les taux de substitution, les intervalles de temps entre les dates d'échantillonnage doivent être suffisamment grands pour permettre une accumulation significative de substitutions. La comparaison entre les taux de substitution ω_2, correspondant à l'intervalle de temps où le patient n'a pas subi de traitement ($t_1 - t_2$), et ω_1, correspondant à l'intervalle de temps où le patient prend son traitement ($t_0 - t_1$), permet d'en déduire l'influence du traitement sur le virus. En effet, si la vitesse d'évolution du virus a subit une accélération ($\omega_1 > \omega_2$), alors le traitement est efficace contre la souche dominante du virus, car cette souche a tendance à disparaître pour en laisser apparaître de nouvelles, ayant une meilleure résistance au traitement, d'où une accélération de la vitesse évolutive. Dans le cas contraire ($\omega_1 \leq \omega_2$), le traitement n'a pas d'influence sur la souche dominante du virus. D'autres applications sont possibles, par exemple pour comparer les vitesses d'évolution des gènes les uns par rapport aux autres. Dans le cas de notre patient atteint par le VIH, cette pratique permettrait de savoir quels gènes le traitement doit cibler pour être efficace, c'est-à-dire ceux conservés car essentiels au virus (Hué *et al*, 2004). Le taux de substitution permet aussi de dater l'origine d'une épidémie ou d'une infection (Wertheim & Worobey, 2009; Korber *et al*, 2000), comme montré dans l'exemple du début. Les applications biologiques rendues possibles par la connaissance du taux de substitution sont donc nombreuses, et, grâce à l'accroissement considérable du nombre de séquences dans les bases de données biologiques, nous pouvons imaginer que certaines d'entre elles vont devenir routinières. Le besoin d'une méthode d'estimation précise et rapide se fait sentir, et pour ce faire les méthodes de distances ont de solides atouts, en raison de leur vitesse et de leur propriété de convergence asymptotique.

Nous présentons dans ce chapitre des méthodes de distances qui permettent d'estimer le taux de substitution à partir d'un ensemble de séquences hétérochrones, sans la connaissance de points de calibration, et sous l'hypothèse d'une horloge moléculaire stricte. Lorsqu'un (plusieurs) point(s) de

calibration existe(nt), l'information des temps de collecte n'est plus indispensable et l'estimation du taux de substitution ou des dates des ancêtres communs peut être faite pour n'importe quelle espèce, y compris à évolution lente (Xia & Yang, 2011; Sanderson, 1997). Nous présentons également deux méthodes probabilistes, chacune partant d'un principe différent (maximum de vraisemblance et bayésien), lourdes en temps de calcul, mais largement utilisées par la communauté scientifique. Mais avant cela, nous discutons de la différence entre taux de substitution synonyme et non synonyme, ainsi que des différents modèles d'horloge moléculaire, incluant notamment les horloges relâchées.

2.2 Taux de substitution synonyme et non synonyme

Dans la littérature, deux sortes de mutations sont distinguées : les mutations synonymes et les mutations non synonymes. Les mutations synonymes (ou silencieuses) sont des mutations qui n'induisent pas de changement d'acide aminé, tandis que les mutations non synonymes (non silencieuses) induisent un changement d'acide aminé. Cela est possible à cause de la redondance du code génétique. Par exemple, si la transversion C → A se produit en première position du codon GCC, codant une Alanine, alors ce codon sera traduit par une Thréonine, tandis que si elle se produit à la troisième position du codon, l'acide aminé traduit restera l'Alanine. De cette observation, découle deux taux de substitution différents : les taux de substitution synonyme et non synonyme et ils ne peuvent être estimés que sur les régions codantes du génome. Le taux de substitution synonyme (resp. non synonyme) est calculé à partir des seules mutations silencieuses (resp. non silencieuses). Généralement les mutations silencieuses se produisent sur le troisième nucléotide du codon et sont plus fréquentes que les mutations non silencieuses qui elles se produisent généralement sur les deux premiers nucléotides du codon (Gojobori *et al*, 1994, 1990). Lorsqu'aucun des deux termes (synonyme et non synonyme) n'est employé, le taux de substitution est calculé en comptant toutes les sortes de mutations (silencieuses ou non). Dans ce cas, il peut aussi être estimé sur les régions non codantes du génome.

2.3 Modèles d'horloge moléculaire

Il est communément admis que la vitesse d'évolution des séquences moléculaires n'est pas strictement uniforme et constante, mais qu'elle peut varier en fonction du temps (par exemple, lorsqu'une pression de sélection supplémentaire s'exerce sur un virus au moment du début d'un traitement) et/ou des lignées (Li & Tanimura, 1987). Ces variations ne sont pas considérées par le modèle d'horloge moléculaire stricte, mais s'en soustraire complètement est impossible. En effet, l'évolution est un processus complexe et la cause de plusieurs facteurs géographiques, géologiques, biologiques, sociologiques, etc. Imaginer une relation universelle entre la distance évolutive et le temps n'est

donc pas faisable (Bromham & Penny, 2003). Dans ce but, plusieurs modèles d'horloge moléculaire ont été proposés. Ils peuvent être regroupés en quatre catégories suivant une terminologie introduite par Rambaut (2000).

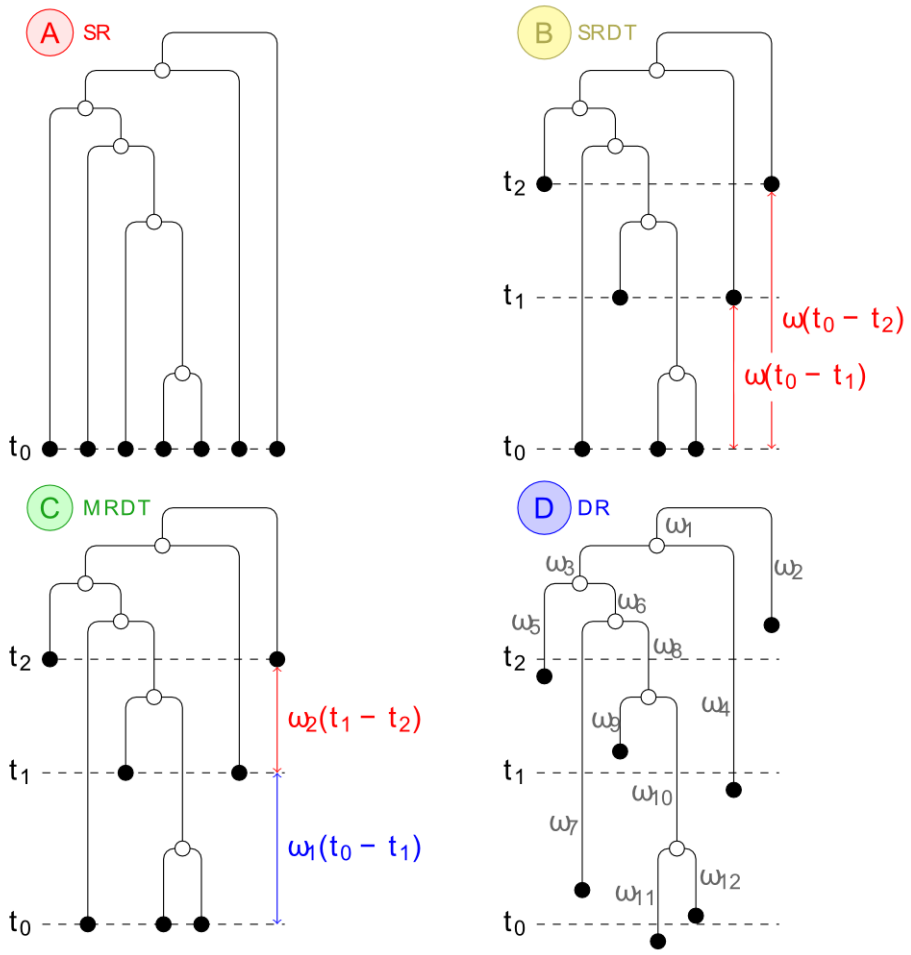
Le modèle *Single Rate* (SR) est le modèle standard (Figure 10A). Il fait l'hypothèse d'une horloge moléculaire stricte mais les séquences sont supposées être échantillonnées au même temps (séquences isochrones). Sinon les intervalles de temps qui séparent les dates de collecte doivent être négligeables par rapport à l'échelle de temps de l'arbre tout entier. Dans ce modèle, le taux de substitution peut uniquement être estimé à l'aide d'un (ou de plusieurs) point(s) de calibration (Xia & Yang, 2011).

Le modèle *Single Rate Dated Tips* (SRDT) fait toujours l'hypothèse d'une horloge moléculaire stricte, mais les séquences sont maintenant prélevées en des temps distincts (séquences hétérochrones) ; il est alors possible d'estimer le taux de substitution avec la connaissance des dates de collecte (Figure 10B) (Rambaut, 2000). Ce modèle est le plus couramment utilisé pour estimer le taux de substitution par des méthodes de distances.

Le modèle *Multiple Rates Dated Tips* (MRDT) suppose une horloge moléculaire relâchée par l'existence de plusieurs taux de substitution, un pour chaque intervalle de temps défini entre deux dates de prélèvement successives (Figure 10C) (Drummond *et al*, 2001). Ce modèle admet une approche alternative que nous distinguerons par la notation MRDT *alternative* (MRDTa). Ce dernier permet à l'utilisateur de choisir ses propres intervalles de temps. Notons toutefois qu'il est impossible d'estimer le taux de substitution lorsque le nombre d'intervalles de temps choisi par l'utilisateur est supérieur au nombre d'intervalles de temps obtenus avec les dates de collecte. De plus, comme les estimations des taux de substitution se font par rapport aux feuilles, il est nécessaire que chaque intervalle de temps contienne au moins une feuille. Donc le nombre maximum d'intervalle de temps est donnée par le nombre de dates de collecte moins un (un temps de collecte doit être utilisé comme référence). Typiquement, ce dernier modèle peut être utilisé pour connaître l'efficacité d'un traitement viral, en comparant sa vitesse évolutive avant le début du traitement et pendant celui-ci (cf. section 2.1).

Enfin, le modèle *Different Rate* (DR) suppose que chaque branche de l'arbre a un taux de substitution propre, ces taux pouvant être corrélés entre eux ou non (Figure 10D) (Rambaut, 2000; Felsenstein, 1981). Ce dernier modèle est le plus réaliste de tous, mais il est excessivement paramétré et insoluble en l'absence de corrélation ou contraintes fortes liant les taux. Les horloges moléculaires locales, c'est-à-dire des horloges moléculaires strictes spécifiques à certaines lignées, associées à une horloge moléculaire stricte globale, sont une variante à ce modèle (Yoder & Yang, 2000).

Figure 10. Illustrations des différents modèles d'horloge moléculaire.
La figure A montre le cas d'une phylogénie sous les contraintes du modèle SR (horloge moléculaire stricte et séquences isochrones). Cette phylogénie est ultramétrique, c'est-à-dire que toutes les séquences sont à égale distance de la racine. La figure B montre une phylogénie sous le modèle SRDT (horloge moléculaire stricte et séquences hétérochrones). La figure C une phylogénie sous le modèle MRDT (un taux de substitution par intervalle de temps entre dates de collecte successives et séquences hétérochrones) et la figure D une phylogénie sous le modèle DR (séquences hétérochrones avec un taux de substitution par branche ; dans cette figure l'écart à l'horloge reste faible).



2.4 Méthodes de distances estimant le taux de substitution sous le modèle SRDT

2.4.1 Premières méthodes

Les premières méthodes de distances permettant d'estimer la vitesse d'évolution sont relativement simples et s'appliquent généralement sur un groupe de deux à trois séquences au plus. À notre connaissance, Hahn *et al*. (1986) sont les premiers à avoir estimé le taux de substitution du VIH-1. Cette estimation est seulement faite à partir de deux séquences provenant d'un même patient, un enfant haïtien vivant en Floride et ayant eu une infection prénatale. Le taux de substitution $\hat{\omega}$ est estimé par la relation

$$\widehat{\omega} = \frac{\hat{d}}{2T}$$

où $\hat{d}$ est la distance évolutive estimée qui sépare les deux séquences, alors calculée sous le modèle JC69 (Jukes & Cantor, 1969), et T le temps écoulé depuis la divergence de leur ancêtre commun. Cette méthode a été préalablement décrite par Gojobori et Yokoyama (1985) mais appliquée à *Moloney murine sarcoma virus*, virus oncogène (pour les souris) de la même famille que le VIH-1. Bien que l'estimation du taux de substitution soit du même ordre de grandeur que celle admise aujourd'hui, plusieurs limites sont à relever. Premièrement, cette méthode suppose que le taux d'évolution est constant, c'est-à-dire que l'estimation du taux de substitution est faite sous l'hypothèse d'une horloge moléculaire stricte (Zuckerkandl & Pauling, 1962), hypothèse admise par de nombreuses autres méthodes, notamment par les méthodes de distances. Deuxièmement, la valeur du paramètre T ne peut être connue avec certitude, elle doit donc être estimée. Pour leurs séquences, Hahn *et al*. (1986) l'avaient estimée variant de une à cinq années. Ils proposaient alors un taux de substitution oscillant entre $1{,}58 \times 10^{-2}$ et $3{,}17 \times 10^{-3}$ substitutions par site et par année sur le gène *env* et entre $1{,}85 \times 10^{-3}$ et $3{,}70 \times 10^{-4}$ substitutions par site et par année sur le gène *gag*. Ces estimations sont donc très imprécises, car elles varient dans une fourchette de 1 à 5.

Pour contrer le problème dû à l'estimation de l'intervalle de temps entre le moment de divergence des séquences et le moment de collecte de celles-ci, nous devons utiliser des données temporelles connues. Li *et al*. (1988) proposent d'utiliser les dates de prélèvement des échantillons qui, elles, sont connues avec certitude. Pour les employer, nous devons toutefois utiliser une troisième séquence, servant d'*outgroup*, afin de mesurer la distance évolutive passée entre deux dates de prélèvement. En effet, le taux de substitution n'est pas égal à la distance évolutive entre deux échantillons divisée par l'intervalle de temps qui sépare leur date de prélèvement (Figure 11) (Drummond *et al.*, 2003). Cela produit une surestimation du taux de substitution, puisque la distance évolutive mesure le nombre de substitutions par site depuis leur divergence de leur ancêtre commun et qui a probablement existé bien avant leur date d'échantillonnage (Figure 11B). Notons que dans le cas où l'une des deux séquences est un ancêtre direct de l'autre, cette formule est exacte (Figure 11A), mais les cas sont rares.

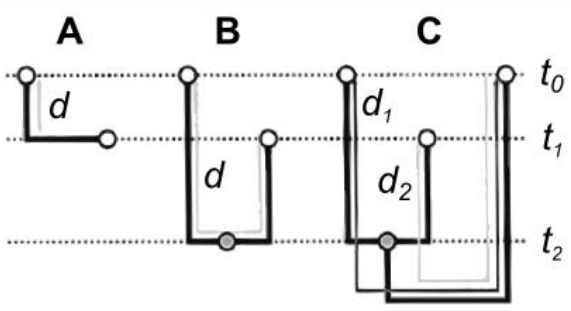
L'utilisation d'un *outgroup* permet donc d'obtenir la distance évolutive entre les deux dates de collecte (Figure 11C). Le choix de l'*outgroup* ne doit pas être fait au hasard, il doit être le plus proche possible des séquences d'intérêt afin d'obtenir une variance d'estimation faible. Soient trois séquences A, B et O où O réfère à l'*outgroup*. Les séquences A et B sont respectivement échantillonnées aux temps t_A et t_B, où t_A est plus récent que t_B, noté $t_B < t_A$, et $\hat{d}_{AO}$ et $\hat{d}_{BO}$ sont les distances

évolutives estimées (obtenues sous n'importe quel modèle) entre les séquences A et O, et, B et O respectivement. Alors le taux de substitution $\hat{\omega}$ vaut

$$\hat{\omega} = \frac{\hat{d}_{AO} - \hat{d}_{BO}}{t_A - t_B}.$$

En utilisant plusieurs séquences différentes comme *outgroup* et comme *ingroup*, dont notamment celles du jeune haïtien, Li *et al*. (1988) estiment un taux de substitution moyen à $5{,}9 \times 10^{-3}$ substitutions par site et par année sur le gène *env*. Avec cette méthode, Gojobori *et al*. (1994) estiment les taux de substitution synonyme et non synonyme du VIH-1 sur les gènes *env* et *gag*. Plusieurs souches y sont comparées et plusieurs estimations du taux de substitution synonyme et non synonyme sont présentées. En conclusion, ils retiennent que les taux de substitution synonyme et non synonyme sont respectivement de $26{,}0 \times 10^{-3}$ et $1{,}0 \times 10^{-3}$ substitutions par site et par année sur *gag* et respectivement de $35{,}5 \times 10^{-3}$ et $3{,}9 \times 10^{-3}$ substitutions par site et par année sur *env*. La différence entre les taux de substitution synonyme et non synonyme s'explique par le fait que les contraintes fonctionnelles appliquées sur le premier sont plus faibles que celles appliquées sur le second.

Figure 11. Relation entre distance évolutive et temps d'échantillonnage.
Schéma montrant la relation entre la distance évolutive et l'intervalle de temps qui sépare deux dates d'échantillonnage. Lorsqu'une souche est l'ancêtre commun d'une autre (figure A), la distance évolutive est proportionnelle au temps écoulé entre les deux dates de prélèvement et une estimation du taux de substitution est donnée en divisant la distance d par l'intervalle de temps $t_0 - t_1$, où t_0 est le temps le plus récent. Malheureusement, cela n'est pas le cas lorsqu'aucune des deux séquences n'est un ancêtre de l'autre (figure B). Dans ce cas, il est nécessaire d'utiliser un *outgroup* afin d'obtenir la distance évolutive $d = d_1 - d_2$ entre les deux temps de collecte t_0 et t_1 (figure C). Ainsi, le taux de substitution peut être estimé sur l'intervalle de temps entre t_0 et t_1 par $(d_1 - d_2)/(t_0 - t_1)$
Adaptation de Drummond *et al.* (2003a).



Bien que ces deux approches offrent des estimations cohérentes avec celles admises aujourd'hui (même ordre de grandeur), elles s'orientent vers une grande erreur type et ne peuvent être appliquées qu'à de petits jeux de données (Suzuki *et al*, 2000).

2.4.2 Les régressions linéaires simples

Le modèle de régression linéaire simple cherche à établir une relation linéaire entre une variable explicative $X = \{x_1, \cdots, x_n\}$ et une variable expliquée $Y = \{y_1, \cdots, y_n\}$, c'est-à-dire

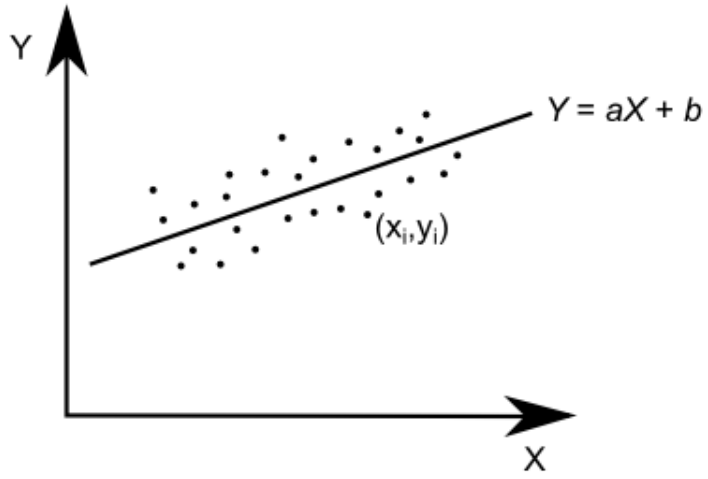
$$Y = aX + b + \varepsilon,$$

où les coefficients a et b sont les paramètres inconnus du modèle à estimer à l'aide des observations sur (X, Y). Le vecteur $\varepsilon = \{\varepsilon_1, \cdots, \varepsilon_n\}$ est le bruit associé au modèle (de moyenne nulle, c'est-à-dire $E[\varepsilon] = 0$), qui prend en compte le fait que la relation entre les variables X et Y n'est pratiquement jamais complètement expliquée par une droite. Afin de ne pas considérer cette erreur dans les notations, on note parfois

$$E[Y|X = x_i] = ax_i + b,$$

où $E[Y|X = x_i]$ représente la valeur moyenne de Y sachant la valeur x_i de X. Une régression linéaire peut être représentée par un graphique à deux dimensions sur lequel un nuage de points, de coordonnées (x_i, y_i), est approximé par une droite qui passe au plus près de ces points. Les coefficients de cette droite sont les paramètres a et b correspondant au modèle de régression linéaire.

Figure 12. Schéma représentant une régression linéaire.
Représentation graphique d'une régression linéaire. Chaque point (x_i,y_i) est représenté sur un graphique à deux dimensions et la droite qui passe au plus près de ces points est la régression linaire dont les coefficients (a et b) sont les paramètres du modèle.



L'estimation du taux de substitution à l'aide d'une régression linéaire ne peut être faite que sous le modèle SRDT, c'est-à-dire avec une horloge moléculaire stricte. Sous ce modèle, la variable Y est associée à la distance évolutive, la variable X au temps et le taux de substitution correspond donc au paramètre a. Sachant l'ensemble des points observés (temps, distance) le modèle cherche à établir une relation linéaire d'où découlera l'estimation du taux de substitution.

Une des faiblesses des modèles de régression linéaire est qu'ils supposent l'indépendance des observations (x_i, y_i) et donc, dans notre cas, des distances évolutives. Ce qui est faux puisque les séquences partagent une partie de leur histoire évolutive (Drummond *et al*, 2003a). Ce problème d'indépendance des données survient aussi dans plusieurs autres problèmes d'évolution, comme par

exemple dans les modèles d'évolution moléculaire qui supposent que les sites d'un alignement évoluent de manière indépendante (cf. Chapitre 1) (Morton & Clegg, 1995; Gutell *et al*, 1994). Les estimations résultant de ces méthodes doivent donc être interprétées avec précaution puisque l'utilisation de méthodes qui incorporent la notion d'indépendance peuvent induire des biais non prédictibles (Drummond *et al*, 2003a).

2.4.2.1 *Pairwise-Distance*

La régression linéaire *Pairwise-Distance* est introduite par Leitner et Albert (1999) dans le but de tester l'existence d'une horloge moléculaire stricte sur les gènes *env* et *gag* du VIH-1. Cette méthode se fonde sur un résultat de la génétique des populations qui dit qu'une population haploïde (resp. diploïde) de taille constante N_e partage un ancêtre commun à N_e générations dans le passé. Donc, deux séquences accumulent en moyenne $\Theta = 2N_e\omega_g$ (resp. $\Theta = 4N_e\omega_g$) mutations par site, où ω_g est le taux de substitution par site et par génération (Felsenstein, 2007; Rodrigo *et al*, 2007). Adapter ce résultat dans le cas où deux séquences i et j sont échantillonnées à des temps différents $t_i < t_j$, c'est-à-dire que t_j est plus récent que t_i, donne la relation linéaire

$$E[\hat{d}_{ij}] = \hat{\omega}(t_j - t_i) + \hat{\Theta},$$

où $\hat{\omega}$ est l'estimation du taux de substitution, $\hat{\Theta}$ une estimation de la diversité génétique des souches échantillonnées au temps t_i et $\hat{d}_{ij}$ la distance évolutive estimée entre les séquences i et j (Figure 13). Ainsi, la régression linéaire des variables $\hat{d}$ et des intervalles de temps d'échantillonnage fournit une estimation du taux de substitution ω et du paramètre Θ. La faiblesse de cette méthode est qu'elle suppose constante la distance génétique entre chaque paire de séquence prise au même temps, alors que celle-ci peut largement varier. Même si la méthode devient correcte lorsque le nombre de séquences est très important, elle est très largement sous-optimale dans la mesure où elle ignore totalement la phylogénie des séquences étudiées. Avec cette méthode, Leitner et Albert (1999) estiment le taux de substitution sur les gènes *gag* et *env* à $2{,}7 \pm 0{,}5 \times 10^{-3}$ substitutions par site et par année et à $6{,}7 \pm 2{,}1 \times 10^{-3}$ substitutions par site et par année respectivement.

2.4.2.2 *Root-to-tip*

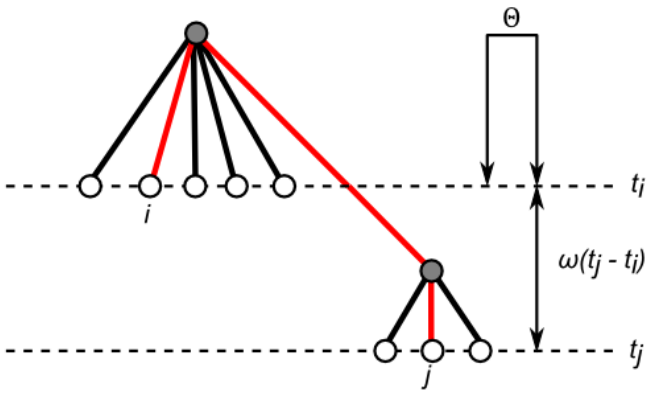
Cette méthode de régression linéaire est l'une des plus utilisées parce qu'elle permet d'estimer simultanément le taux de substitution ω et la date de l'ancêtre commun aux séquences t_{racine} (Drummond *et al*, 2003a). De ce fait, et contrairement à la régression *Pairwise-Distance*, cette méthode utilise une phylogénie enracinée des séquences étudiées, puis fait une régression linéaire entre les dates d'échantillonnage t_i de chaque séquence i avec la distance estimée $\hat{d}_{i,\text{racine}}$ qui sé-

pare la feuille représentant i de la racine (obtenue en additionnant les longueurs des branches de l'arbre situées sur le chemin de la feuille i jusqu'à la racine). Ainsi, le modèle linéaire (Figure 14) est

$$E\left[\hat{d}_{i,\text{racine}}\right] = \hat{\omega}(t_i - \hat{t}_{\text{racine}}),$$

où $\hat{\omega}$ et $\hat{t}_{\text{racine}}$ sont des estimations du taux de substitution et de la date de l'ancêtre commun aux séquences. L'intersection avec l'axe des abscisses donne l'estimation de t_{racine}, car, dans ce cas, on a $\hat{\omega}(t_i - \hat{t}_{\text{racine}}) = 0$, donc $t_i = \hat{t}_{\text{racine}}$ lorsque $\hat{\omega} \neq 0$. Avec cette méthode Korber *et al.* (2000) ont estimé, sur le gène *env*, la date de l'ancêtre commun aux souches appartenant au groupe du VIH-1 responsable de la pandémie actuelle (groupe M) à 1931 [1915-1941]. Leur estimation du taux de substitution est de $2{,}4 \times 10^{-3}$ $[1{,}8 \times 10^{-3}; 2{,}8 \times 10^{-3}]$ substitutions par site et par année. Sur le gène *gag*, ils estiment un taux de substitution à $1{,}9 \times 10^{-3}$ $[0{,}9 \times 10^{-3}$; $2{,}7 \times 10^{-3}]$ substitutions par site et par année et une date de l'ancêtre commun au VIH actuel à 1934 [1869 ; 1950].

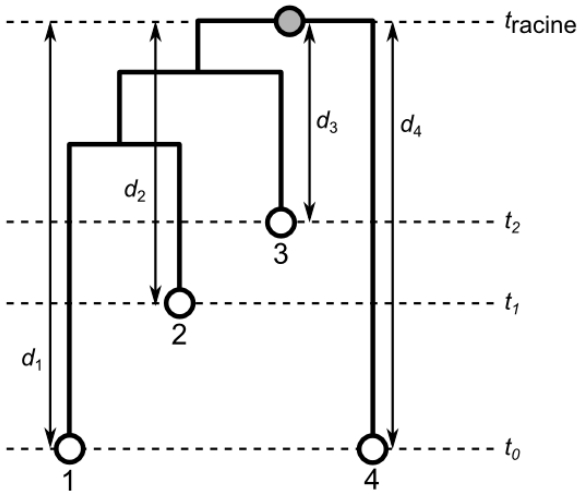
Figure 13. Modèle *Pairwise-Distance*.
Le modèle *Pairwise-Distance* suppose que la distance évolutive d_{ij}, séparant les souches i et j (en rouge), respectivement échantillonnées aux temps t_i et t_j (t_j est plus récent que t_i), est égale à la diversité génétique moyenne Θ entre chaque paire de séquences échantillonnées à t_i, plus la distance évolutive entre t_i et t_j (proportionnelle au taux de substitution ω à estimer). À savoir $d_{ij} = \omega(t_j - t_i) + \Theta$.



La connaissance de l'emplacement de la racine dans la phylogénie est donc primordiale pour utiliser cette méthode. Mais il est tout de même possible d'utiliser une phylogénie non enracinée. Dans ce cas, il est nécessaire de parcourir toutes les branches de la phylogénie afin de trouver l'emplacement optimal pour la racine. Par exemple, l'emplacement qui maximise le coefficient de corrélation de Pearson entre les dates de prélèvement t_i et les distances évolutives $\hat{d}_{i,\text{racine}}$, qui mesure la « qualité » de la régression linéaire. L'emplacement sur la phylogénie qui maximise ce coefficient est alors choisi comme racine et les paramètres sont estimés en fonction de cette racine. Cette méthode est mise en œuvre dans les versions antérieures à la version 1.3 du logiciel Path-O-Gen[2]. Depuis la version 1.3, Path-O-Gen localise l'emplacement optimal de la racine en minimisant la somme des résidus, c'est-à-dire l'écart des estimations à la droite de régression.

[2] http://tree.bio.ed.ac.uk/software/pathogen/

Figure 14. Modèle *Root-to-tip*.
Quatre souches (cercle blanc) sont échantillonnées à trois temps différents t_0, t_1 et t_2. Les valeurs d représentent la distance évolutive qui sépare chaque séquence i de la racine (cercle gris). Soit ω le taux de substitution, alors pour chaque séquence i, on a $d_i = \omega(t_i - t_{\text{racine}})$.



2.4.3 sUPGMA

Serial-Sample UPGMA (sUPGMA) est une méthode de distances d'inférence phylogénétique sous les hypothèses du modèle SRDT (Rodrigo *et al*, 2007; Drummond & Rodrigo, 2000). Elle est le prolongement de la méthode UPGMA (*unweighted pair grouping method with arithmetic means*, cf. Chapitre 1), qui est adaptée au modèle SR (Sokal & Michener, 1958). En effet, l'algorithme UPGMA construit une représentation où chaque feuille de l'arbre est à égale distance de la racine (cohérent si les séquences sont isochrones et si on suppose une horloge moléculaire stricte), c'est-à-dire une phylogénie ultramétrique ou un dendrogramme (Barthélemy & Guénoche, 1988). Or, le modèle SRDT implique que les feuilles sont échantillonnées à des dates différentes et doivent donc être à des distances différentes de la racine, en fonction de la date de collecte de ces dernières. Mais deux feuilles échantillonnées au même moment doivent se situer à la même distance de la racine. sUPGMA, mise en œuvre dans PEBBLE (Goode & Rodrigo, 2004), prend donc en considération les temps de collecte des feuilles dans le calcul de la phylogénie. Pour faire cela, quatre étapes principales sont nécessaires, sachant que la première étape est une méthode d'estimation du taux de substitution sous le modèle SRDT. Les autres étapes servent uniquement à calculer la phylogénie.

Estimation du taux de substitution

La première étape de la méthode sUPGMA consiste à estimer le taux de substitution relatif à l'ensemble des séquences. Soient p temps de prélèvement tels que le temps i est obtenu plus récemment que le temps $i+1$ ($i \in 1, \cdots p$). Soit $\hat{d}(m_i, n_j)$ la distance évolutive estimée entre la $i^{\text{ème}}$ séquence collectée à la date m et la $j^{\text{ème}}$ séquence collectée à la date n, avec $m \geq n$. Alors

$$\hat{d}(m_i, n_j) = \widehat{\Theta}_m + \widehat{\omega}(t_m - t_n) + \varepsilon_{m_i,n_j},$$

où t_k est la date du temps d'échantillonnage k, $\widehat{\omega}$ le taux de substitution à estimer et $\widehat{\Theta}_m$ la diversité génétique des séquences échantillonnées au temps m aussi à estimer. Les termes ε_{m_i,n_j} représentent les erreurs dues à l'estimation des distances évolutives. Il est possible d'exprimer ces équations à l'aide d'une notation matricielle. Soient D le vecteur contenant les estimations des distances évolutives, $\beta = \{\widehat{\Theta}_1, \cdots, \widehat{\Theta}_\mathrm{p}, \widehat{\omega}\}$ le vecteur des paramètres à estimer, et E le vecteur des erreurs, alors

$$D = M\beta + E$$

avec M la matrice telle que pour chaque ligne i et chaque colonne $j \leq p$

$$\left(M_{i,j}\right)_{m,n} = \begin{cases} 1 \text{ si } j = m \\ 0 \text{ sinon} \end{cases}$$

et $\left(M_{i,p+1}\right)_{m,n} = t_m - t_n$. Le vecteur des paramètres estimés $\beta = \{\widehat{\Theta}_1, \cdots, \widehat{\Theta}_\mathrm{p}, \widehat{\omega}\}$, qui minimise la somme des erreurs au carré $E^T E$, est alors donné par la méthode des moindres carrés :

$$\beta = (M^T M)^{-1} M^T D.$$

Cette méthode peut facilement être étendue au modèle MRDT (Drummond *et al*, 2001). Dans ce cas, il suffit de décomposer l'intervalle de temps $(t_m - t_n)$ en $(t_m - t_{m-1}) + \cdots + (t_{n+1} - t_n)$ et d'affecter à chaque intervalle de temps le taux de substitution correspondant. À l'inverse, une hypothèse simplificatrice est de supposer une diversité génétique constante quel que soit le temps d'échantillonnage. Cela revient à estimer qu'un seul paramètre Θ, au lieu d'un pour chaque temps de collecte. Dans ce cas, le modèle devient

$$\hat{d}_{ij} = \widehat{\Theta} + \widehat{\omega}(t_j - t_i)$$

et il est alors équivalent à la régression linéaire *Pairwise-Distance*.

Avec cette méthode les auteurs ont estimé le taux de substitution du VIH-1 sur des souches isolées chez un même patient, sur cinq temps d'échantillonnage couvrant 1 005 jours (Rodrigo *et al*, 1999). Leur estimation du taux de substitution sur le gène *env*, en considérant un paramètre Θ et un taux de substitution unique, est de $7{,}8 \times 10^{-6}$ $[-3{,}47 \times 10^{-6}; 3{,}87 \times 10^{-5}]$ substitutions par site et par jour. Ramenée à l'échelle des années, l'estimation est approximativement de 3×10^{-3} substitutions par site et par année.

Correction de la matrice de distances

Une fois le taux de substitution estimé, il est alors possible de corriger la matrice de distances $\hat{d}$ en ajoutant, à chaque distance estimée, la mesure manquante afin de voir i et j comme contemporains, c'est-à-dire que

$$\hat{d}'_{ij} = \hat{d}_{ij} + \hat{\omega}(t_0 - t_i) + \hat{\omega}(t_0 - t_j),$$

où t_i et t_j réfèrent au temps de collecte des souches i et j, et où le temps d'échantillonnage le plus récent est noté t_0. La mesure d' voit alors les séquences i et j comme contemporaines (c'est-à-dire échantillonnées au temps t_0).

Calcul de l'arbre à l'aide de UPGMA

Un arbre UPGMA ou WPGMA est calculé à partir de la mesure corrigée $\hat{d}'$ qui voit toutes les souches comme contemporaines, c'est-à-dire que toutes les souches doivent se situer à égale distance de la racine.

Modification de l'arbre UPGMA

L'arbre UPGMA ou WPGMA obtenu est ultramétrique, c'est-à-dire que toutes les feuilles sont à égale distance de la racine. Afin d'obtenir un arbre où chaque feuille collectée à un temps d'échantillonnage différent est à une distance différente de la racine, mais où toutes les feuilles d'un même temps d'échantillonnage sont à une même distance de la racine, il suffit de soustraire la mesure $\hat{\omega}(t_0 - t_i)$ à la longueur de la branche associée à la séquence i. De cette façon, la topologie obtenue respecte celle du modèle SRDT.

2.4.4 TREBLE

Tree and rate estimation by local evaluation (TREBLE) est une méthode estimant le taux de substitution à partir d'un ensemble de séquences hétérochrones et en faisant l'hypothèse d'une horloge moléculaire stricte, donc sous les hypothèses du modèle SRDT (Yang *et al*, 2007). Cette méthode utilise des triplets de séquences, c'est-à-dire que pour chaque triplet de séquences possible, vérifiant une certaine condition, un taux de substitution et sa variance sont estimés, puis elle calcule la moyenne des taux de substitution estimés sur chaque triplet, pondérée par l'inverse de la variance correspondante, afin d'obtenir un taux de substitution global, solution du problème.

Cette méthode part de l'observation que pour deux séquences données i et j, échantillonnées respectivement aux temps t_i et t_j, il existe une relation entre leur distance génétique estimée $\hat{d}_{ij}$, leur taux de substitution ω_{ij} et leur temps de collecte, telle que (Figure 15A)

$$\hat{d}_{ij} = \omega_{ij}(t_i + t_j - 2t_{ij}) + \varepsilon_{ij},$$

où t_{ij} réfère à la date de l'ancêtre commun aux souches i et j et ε_{ij} aux erreurs associées à l'estimation des distances évolutives $\hat{d}_{ij}$, négligées par la suite. Comme les paramètres ω_{ij} et t_{ij} sont inconnus, l'équation n'a pas de solution unique. Cet handicap peut être résolu en considérant une séquence supplémentaire k, échantillonnée au temps t_k, mais ayant une configuration topologique particulière avec les deux autres séquences i et j (Figure 15B). Considérant en plus cette troisième séquence et leur configuration géométrique, il est maintenant possible d'estimer le taux de substitution et les dates de leurs ancêtres communs par les équations

$$\hat{t}_{ik} = \hat{t}_{jk} = \frac{1}{2}\left[\frac{\hat{d}_{ik}t_j - \hat{d}_{jk}t_i}{\hat{d}_{ik} - \hat{d}_{jk}} + t_k\right],$$

$$\hat{t}_{ij} = \frac{1}{2}\left[(t_i + t_j) - \frac{\hat{d}_{ij}(t_i - t_j)}{(\hat{d}_{ik} - \hat{d}_{jk})}\right]$$

et

$$\widehat{\omega}_{ij|k} = \frac{\hat{d}_{ik} - \hat{d}_{jk}}{t_i - t_j}.$$

Le taux de substitution estimé (noté par $\widehat{\omega}$) est relatif aux séquences i et j sachant la séquence supplémentaire k (d'où la notation $\widehat{\omega}_{ij|k}$). Ce dernier dépend seulement des temps d'échantillonnage correspondants aux séquences i et j appelées « paire informative ». La séquence restante k est appelée *outgroup*. Cette formule est identique à celle proposée par Li *et al*. (1988), présentée à la section 2.4.1, c'est-à-dire qu'elle estime le taux de substitution à partir de la distance évolutive entre deux temps d'échantillonnage et, pour cette raison, elle nécessite l'utilisation d'un *outgroup*. Les dates des ancêtres communs étant aussi estimées, elles sont aussi notées $\hat{t}$. La topologie nécessaire pour de telles estimations n'est *a priori* pas connue et l'utilisation de triplets quelconques peut conduire à des estimations non valides. Ainisi, les estimations valides sont celles qui vérifient les conditions suivantes : $\widehat{\omega}_{ij|k} > 0$, $\hat{t}_{ij} \leq t_i$, $\hat{t}_{ij} \leq t_j$, $\hat{t}_{ik} \leq t_k$, $\hat{t}_{ik} \leq \hat{t}_{ij}$, en accord avec la configuration géométrique que doit présenter le triplet (Figure 15B).

Plusieurs sources d'erreur peuvent causer des biais dans l'estimation des taux de substitution $\widehat{\omega}_{ij|k}$. Par exemple, des erreurs inhérentes à l'estimation des distances évolutives dues aux substitutions cachées (cf. Chapitre 1). Afin d'augmenter la précision de l'estimation du taux de substitution global, Yang *et al*. (2007) proposent d'associer à chaque $\widehat{\omega}_{ij|k}$ une variance représentant la confiance associée à l'estimation. Suivant Rzhetsky et Nei (1995), la covariance entre les distances évolutives d_{ab} et d_{cd}, des séquences a, b, c et d, est égale à la variance de la longueur des branches partagées par les chemins reliant les séquences a à b et les séquences c à d, notée $d_{ab,cd}$. Soit

$$\mathrm{cov}(d_{ab}, d_{cd}) = \mathrm{var}(d_{ab,cd}).$$

Pour le triplet de la Figure 15B, il vient

$$\mathrm{cov}(d_{ik}, d_{jk}) = \mathrm{var}(d_{ik,jk}) = \mathrm{var}(d_{lk})$$

où l représente l'ancêtre commun des séquences i et j, c'est-à-dire celui au temps t_{ij}. Avec cette observation, la variance associée au taux de substitution estimé est

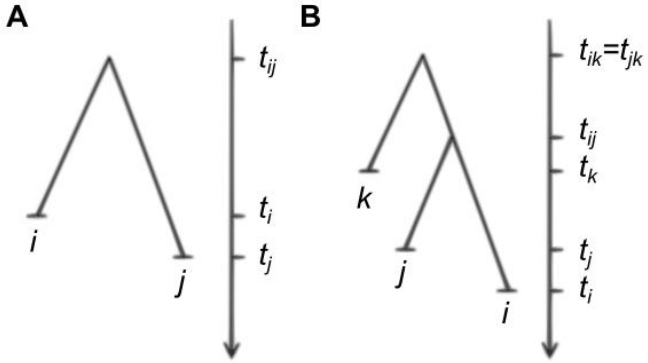
$$\begin{aligned} \mathrm{var}(\hat{\omega}_{ij|k}) &= \mathrm{var}\left(\frac{\hat{d}_{ik} - \hat{d}_{jk}}{t_i - t_j}\right) \\ &\approx \frac{\mathrm{var}(\hat{d}_{ik}) + \mathrm{var}(\hat{d}_{jk}) - 2\mathrm{cov}(\hat{d}_{ik}, \hat{d}_{jk})}{(t_i - t_j)^2} \\ &\approx \frac{\mathrm{var}(\hat{d}_{ij})}{(t_i - t_j)^2}. \end{aligned}$$

Remarquons que la variance est indépendante de l'*outgroup*. Elle est donc identique pour chaque paire informative quel que soit l'*outgroup* considéré. La plupart des modèles d'évolution moléculaire propose une formule analytique pour calculer la variance de la distance évolutive (Rzhetsky & Nei, 1995).Cependant, elle peut aussi être approximée par

$$\mathrm{var}(\hat{d}_{ij}) \approx \frac{\hat{d}_{ij}}{N},$$

où N est la longueur des séquences dans l'alignement (Gascuel, 2000; Bulmer, 1991).

Figure 15. Illustration et comportement d'une paire de séquence et d'un triplet de séquence.
La figure A montre deux séquences i et j respectivement échantillonnées aux temps t_i et t_j et divergent de leur ancêtre commun au temps t_{ij}. La figure B montre trois séquences i, j et k respectivement échantillonnées aux temps t_i, t_j et t_k. Les souches i et j divergent de leur ancêtre commun au temps t_{ij} et les souches k et i, ainsi que les souches k et j, divergent de leur ancêtre commun au temps $t_{ik} = t_{jk}$.
Adaptation de Yang *et al.* (2007).



Une fois la connaissance de toutes les paires informatives et des *outgroups* valides, TREBLE calcule pour chaque paire informative i et j une moyenne $\widehat{\omega}_{ij}$ des taux de substitution estimés avec chaque *outgroup*

$$\widehat{\omega}_{ij} = \frac{1}{|O_{ij}|} \sum_{k \in O_{ij}} \widehat{\omega}_{ij|k},$$

où O_{ij} est l'ensemble des *outgroups* retenus pour la paire informative i et j. Le taux de substitution global $\widehat{\omega}$ est alors donné par la moyenne pondérée de chaque $\widehat{\omega}_{ij}$

$$\widehat{\omega} = \frac{1}{W} \sum_{i,j} w_{ij} \widehat{\omega}_{ij},$$

avec $w_{ij} = 1/\mathrm{var}(\widehat{\omega}_{ij})$ et W= $\sum_{i,j} w_{ij}$.

Après avoir estimé le taux de substitution global, TREBLE propose de vérifier à nouveau la validité des *outgroups* associés à chaque paire informative mais en considérant cette fois-ci l'estimation globale du taux de substitution. Pour cela, il impose une contrainte sur les distances $\hat{d}_{jk}$, telle que

$$\widehat{\omega} t_i + \hat{d}_{jk} - \widehat{\omega} t_k - \hat{d}_{ij} > \varepsilon_{jk} - \varepsilon_{ij},$$

où les ε sont les erreurs provenant de l'estimation des distances évolutives. Vérifier cette contrainte, c'est vérifier que la moitié de la distance évolutive entre t_{ij} et t_{jk} est strictement positive, puisqu'on a l'égalité $\widehat{\omega} t_i + \hat{d}_{jk} - \widehat{\omega} t_k - \hat{d}_{ij} + 2\widehat{\omega}(t_{jk} - t_{ij}) = \varepsilon_{jk} - \varepsilon_{ij}$. Les erreurs ε sont des variables aléatoires inconnues distribuées suivant une loi normale d'espérance nulle (donc leur différence aussi). Ainsi, $(\varepsilon_{jk} - \varepsilon_{ij})/\sqrt{\mathrm{var}(\varepsilon_{jk} - \varepsilon_{ij})}$ suit une loi normale centrée réduite, avec $\mathrm{var}(\varepsilon_{jk} - \varepsilon_{ij}) \approx \mathrm{var}(\hat{d}_{ij})$ d'après les formules de variance ci-dessus. Soit Z_α la valeur correspondant du quantile α obtenue dans la table de la loi normale centrée réduite, avec α choisi par l'utilisateur. Alors la probabilité pour que $(\varepsilon_{jk} - \varepsilon_{ij}) < Z_\alpha \sqrt{\mathrm{var}(\hat{d}_{ij})}$ est $1 - \alpha$. Si $\widehat{\omega} t_i + \hat{d}_{jk} - \widehat{\omega} t_k - \hat{d}_{ij} > Z_\alpha \sqrt{\mathrm{var}(\hat{d}_{ij})}$, alors la probabilité pour que $\widehat{\omega} t_i + \hat{d}_{jk} - \widehat{\omega} t_k - \hat{d}_{ij} > (\varepsilon_{jk} - \varepsilon_{ij})$ est au moins $1 - \alpha$. Ainsi, les *outgroups* ne satisfaisant pas cette contrainte sont supprimés et la procédure complète est recommencée sans ces *outgroups*. Un nouveau taux de substitution global est alors estimé et ce dernier test répété. Et ceci jusqu'à stabilisation des *outgroups*.

Nous pouvons donc voir cette méthode comme une généralisation de la méthode proposée par Li *et al*. (1988), mais où les critères statistiques de sélection conservent uniquement les *outgroups* qui forment une configuration bien précise avec les paires de séquences informatives, donc ceux qui

permettent une bonne estimation du taux de substitution. Ces critères sont aussi une faiblesse de cette méthode parce qu'ils rejettent, en pratique, beaucoup de triplets et donc de l'information. Malgré les promesses de vitesse et de performance encourageantes, aucune application biologique concrète[3] n'a été faite avec cette méthode, outre celles des auteurs.

Une adaptation de cette méthode dans le cas où l'on considère trois clades différents dans la phylogénie, par exemple représentant chacun un sous-type différent au sein d'un même virus, est proposée par O'Brien *et al*. (2008). Cette dernière méthode estime la date de divergence entre deux clades (sachant le troisième) et le taux de substitution relatif aux deux clades considérés.

2.4.5 *TreeRate*

TreeRate est une méthode qui se base sur une phylogénie racinée pour estimer la distance génétique séparant deux groupes de séquences choisis par l'utilisateur (Maljkovic Berry *et al*, 2009, 2007). Elle permet aussi d'estimer le taux de substitution sous les hypothèses du modèle SRDT. Pour faire cela, l'utilisateur choisit préalablement deux collections de feuilles assignées respectivement au groupe T1 et T2 (Figure 16). Certaines feuilles peuvent être écartées de l'analyse. Elles sont alors considérées comme *outgroup*. Une moyenne $\bar{X}_1$ (respectivement $\bar{X}_2$) des distances de chaque feuille du groupe T1 (resp. T2) à la racine est calculée. Ainsi, la distance génétique qui sépare les deux groupes de feuilles est calculée comme la différence entre ces deux moyennes, soit $\Delta\hat{d} = \bar{X}_2 - \bar{X}_1$. Dès lors, il est possible d'estimer le taux de substitution ω entre ces deux groupes en s'aidant de Δt un intervalle de temps calculé à partir de $\bar{T}_1$ (resp. de $\bar{T}_2$) qui représente la moyenne des dates de collecte des feuilles de T1 (resp. T2), ainsi

$$\widehat{\omega} = \frac{\Delta\hat{d}}{\Delta t} = \frac{\bar{X}_2 - \bar{X}_1}{\bar{T}_2 - \bar{T}_1}.$$

Dans le cas d'une phylogénie non enracinée, *TreeRate* estime au préalable la position optimale de la racine suivant un test statistique. Plusieurs tests sont proposés par les auteurs, mais ils suggèrent que celui consistant à minimiser la somme des variances est le plus performant, c'est-à-dire à minimiser le terme

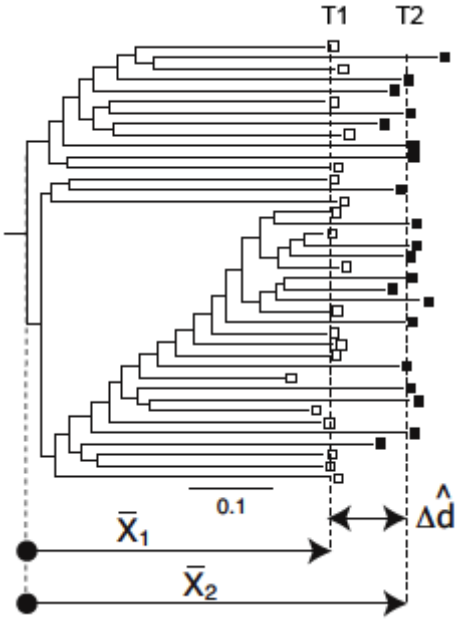
$$\sigma^2 = \left[\frac{1}{N_1}\sum_{i=1}^{N_1}\left(\hat{X}_i - \bar{X}_1\right)^2\right] + \left[\frac{1}{N_2}\sum_{j=1}^{N_2}\left(\hat{X}_j - \bar{X}_2\right)^2\right],$$

où N_i est le nombre de feuilles contenues dans le groupe Ti et X_i la distance de la feuille i à la racine. Cette méthode très simple n'est donc pas vraiment originale, puisque très proche de méthodes déjà

[3] D'après PubMed (http://www.ncbi.nlm.nih.gov/pubmed), consultée le 2 février 2012.

discutées plus haut (comme sUPGMA, TREBLE ou encore *Root-to-tip*). Elle est mise en œuvre sur le site web de la base de données sur le VIH du laboratoire national de Los Alamos[4]. Avec cette méthode, les auteurs montrent que sur la région *env* du génome du VIH-1 le taux de substitution du sous-type C est de $9{,}65 \times 10^{-3}$ [$8{,}88 \times 10^{-3}$; $10{,}4 \times 10^{-3}$] substitutions par site et par année alors que celui du sous-type A est de $16{,}9 \times 10^{-3}$ [$12{,}1 \times 10^{-3}$; $21{,}6 \times 10^{-3}$] substitutions par site et par année (à partir d'échantillons collectées en Afrique uniquement) (Maljkovic Berry *et al*, 2007).

Figure 16. Illustration de la méthode *TreeRate*.
La distance moyenne des feuilles du groupe T1 à la racine est noté $\bar{X}_1$, celle des feuilles du groupe T2 à la racine $\bar{X}_2$. La mesure $\Delta\hat{d} = \bar{X}_2 - \bar{X}_1$ représente la distance qui sépare les feuilles du groupe T1 et T2. Le taux de substitution $\hat{\omega}$ vaut $\Delta\hat{d}/\Delta t$, où Δt représente l'intervalle de temps entre la moyenne arithmétique des dates d'échantillonnage des feuilles de T1 et celle des feuilles de T2.
Source : http://www.hiv.lanl.gov/content/sequence/TREERATE/treerate_explanation.html



2.4.6 Méthode de Langley-Fitch

La méthode de Langley-Fitch (1974), mise en œuvre dans r8s (Sanderson, 2003), permet d'estimer simultanément le taux de substitution et les dates des ancêtres communs d'une phylogénie racinée sous les hypothèses du modèle SRDT. Plus tard, Sanderson (2002) propose la méthode *Penalized Likelihood* qui est une adaptation de la méthode Langley-Fitch au modèle DR, c'est-à-dire avec un taux de substitution pour chaque branche de la phylogénie, et une corrélation entre les taux des branches adjacentes. Ces deux méthodes suivent une approche de distance mais posent un modèle probabiliste sur le bruit et les (éventuelles) corrélations, si bien que ce sont aussi des méthodes probabilistes avec des temps de calcul plus importants que les méthodes précédentes, mais moins que ceux des méthodes pleinement probabilistes basées sur les caractères, comme BEAST par exemple.

[4] http://www.hiv.lanl.gov/content/sequence/TREERATE/combinedBranchlength.html

Supposons que la phylogénie a S nœuds internes représentés par un nombre entier de 0 à S, où 0 représente le nœud racine, et M feuilles représentées par les nombres entiers $S+1$ à $S+M$. L'âge du nœud k est noté t_k et $\mathrm{anc}(k)$ représente le nœud ancestral à k dans la phylogénie. Notons par ω le taux de substitution et par b_i la longueur de la branche $\big(i,\mathrm{anc}(i)\big)$. Les paramètres du modèle (SRDT) à estimer dans la méthode de Langley-Fitch sont donc $\Omega = \{t_0, \cdots, t_S, \omega\}$. On s'intéresse ici au nombre de substitutions par site sur des durées déterminées par les temps de prélèvement et sur la phylogénie. Le modèle Poissonien offre un cadre naturel.

Supposons que le nombre de substitutions par unité de temps et par site suit une loi de Poisson de moyenne ω. Alors le nombre de substitutions par site sur une branche $\big(i,\mathrm{anc}(i)\big)$ suit aussi une loi de Poisson mais de moyenne $\omega\big(t_{\mathrm{anc}(i)} - t_i\big)$. Autrement dit, la probabilité d'avoir b_i substitutions par site sur la branche i est $P(b_i|\,\omega\big[t_{\mathrm{anc}(i)} - t_i\big])$, avec $P(b|a) = a^b \exp(-a)/b!$. Le logarithme de la vraisemblance de l'arbre tout entier est donné par

$$\log L(\Omega|b_1, \cdots, b_{S+M}) = \sum_{k=1}^{S+M} \log P\big(b_k|\omega\big[t_{\mathrm{anc}(k)} - t_k\big]\big),$$

et les valeurs des paramètres Ω qui maximisent ce logarithme sont les estimateurs du maximum de vraisemblance. Maximiser cette expression ne pose pas de problème majeur et peut être réalisé par une approche standard.

2.5 Quelques méthodes pleinement probabilistes

Les méthodes probabilistes présentent un avantage certain en précision d'estimation par rapport aux méthodes de distances. Également, avec des méthodes probabilistes, les paramètres du modèle d'évolution peuvent être auto-estimés (cf. Chapitre 1). Toutefois, ces méthodes ne peuvent être appliquées qu'à des petits jeux de données (quelques centaines de séquences au plus), en raison des temps de calcul considérables qu'elles nécessitent. Dans cette section nous présentons brièvement deux méthodes probabilistes permettant d'estimer le taux de substitution à partir des hypothèses du modèles SRDT. La première méthode, *TipDate*, utilise le principe du maximum de vraisemblance et la seconde, BEAST, utilise une approche bayésienne (Drummond *et al*, 2012; Drummond & Rambaut, 2007; Rambaut, 2000). Cette dernière est actuellement la méthode de référence dans le domaine.

TipDate est une méthode développée par Rambaut (2000) qui permet d'estimer simultanément une phylogénie, les dates associées à chaque nœud interne de celle-ci, ainsi que le taux de substitution et cela sous les hypothèses du modèle SRDT. Plus tard, Drummond *et al*. (2001) l'adaptent au modèle MRDT. Cette méthode estime les dates des ancêtres communs et le taux de substitution en

remplaçant dans la procédure décrite par Felsenstein (1981), les longueurs de branche par le produit du taux de substitution et de l'intervalle de temps correspondant à cette branche (obtenu en soustrayant les dates associées aux nœuds adjacents). Les estimations de ces paramètres sont alors ceux qui maximisent la fonction de vraisemblance

$$L(\omega) = P(D|T, \omega, M),$$

où D représente l'alignement, ω est le taux de substitution, T la phylogénie (supposée suivre ici une horloge moléculaire stricte) et M les paramètres associés au modèle d'évolution.

BEAST (*bayesian evolutionary analysis by sampling trees*) est le logiciel d'estimation de taux de substitution le plus utilisé aujourd'hui (Drummond *et al*, 2012; Drummond & Rambaut, 2007; Drummond *et al*, 2002). Ce qui en fait son succès est sans doute les multiples services qu'il propose. Il est bien sûr possible d'y estimer le taux de substitution sous le modèle SRDT, mais ce logiciel donne aussi la possibilité d'utiliser d'autres modèles d'horloge moléculaire, comme par exemple des horloges moléculaires relâchées où les taux de substitution varient au niveau des nœuds internes (horloge moléculaire relâchée en exponentiel) ou le long des branches auxquelles ils sont associés (horloge moléculaire relâchée en log-normal) (Drummond *et al*, 2006). Il donne aussi la possibilité d'inférer une phylogénie mise à l'échelle temporelle (et donc il estime aussi les dates des ancêtres communs à chaque nœud) sous une large gamme de modèles d'évolution, ou d'obtenir le graphique représentant la taille effective de la population en fonction du temps $N_e(t)$ (cf. section 2.4.2.1) sous plusieurs modèles démographiques. Une option spéciale *BEAST (prononcée « star BEAST ») permet d'utiliser simultanément plusieurs régions d'un génome afin d'obtenir des résultats globaux (Heled & Drummond, 2010). Depuis peu, la reconstruction de caractères ancestraux, comme des régions géographiques est mise à disposition (Lemey *et al*, 2010, 2009a). Ce logiciel offre de multiples autres possibilités et les études l'utilisant sont très nombreuses, notamment en épidémiologie moléculaire. Citons en exemple, Dalai *et al*. (2009) qui estiment à $2{,}19 \times 10^{-3}$ [$1{,}83 \times 10^{-3}$; $2{,}56 \times 10^{-3}$] substitutions par site et par année le taux de substitution du VIH-1 sur le gène *pol*, avec le modèle d'horloge moléculaire stricte, et Bello *et al*. (2008) qui estiment à $1{,}5 \times 10^{-3}$ [$1{,}0 \times 10^{-3}$; $2{,}0 \times 10^{-3}$] et à $5{,}8 \times 10^{-3}$ [$3{,}8 \times 10^{-3}$; $7{,}8 \times 10^{-3}$] substitutions par site et par année le taux de substitution du VIH-1 sous le modèle d'horloge moléculaire stricte pour les gènes *pol* et *env* respectivement. Cependant, le point faible de ce logiciel est le temps de calcul considérable qu'il demande sur un jeu de données d'à peine quelques centaines de séquences. En effet, ce logiciel utilise le principe bayésien des chaînes de Markov par technique de Monte Carlo (MCMC) (cf. Chapitre 1), avec la variante de Metropolis-Hasting (Hastings, 1970; Metropolis *et al*, 1953), qui nécessite une quantité très importante de calculs pour approximer au mieux la distribution *a posteriori* des paramètres

d'intérêt à partir de données et d'une distribution *a priori* ou *prior*. De plus, cette *prior* en fait une méthode assez controversée puisqu'utilisée à tort, elle permet généralement d'obtenir des résultats souhaités.

Tableau 1. Récapitulatif des taux de substitution du VIH estimés par les différentes méthodes.
Les taux de substitution du VIH sont donnés en substitutions par site et par année. Les gènes sur lesquels le taux de substitution est estimé sont précisés et lorsque la méthode utilisée ne porte pas de nom particulier, la référence de l'article est donnée à la place. La liste des taux de substitution, triée par gène, correspond aux estimations citées dans le chapitre et n'est en rien exhaustive par rapport à la littérature.

Méthode	Gène	Taux de substitution ($\times 10^{-3}$)			Référence
			Min	Max	
Hahn *et al*. (1996)	*env*	-	3,17	15,80	Hahn *et al*. (1986)
Li *et al*. (1988)	*env*	5,90	-	-	Li *et al*. (1988)
Li *et al*. (1988)	*env*	35,50[a]	-	-	Gojobori *et al*. (1994)
Li *et al*. (1988)	*env*	3,90[b]	-	-	Gojobori *et al*. (1994)
Pairwise-Distance	*env*	6,70	4,60	8,80	Leitner et Albert (1999)
Root-to-tip	*env*	2,40	1,80	2,80	Korber *et al*. (2000)
sUPGMA	*env*	3,00	-1,34	14,89	Drummond et Rodriguo (2000)
TreeRate	*env*	9,65	8,88	10,40	Maljkovic Berry *et al*. (2007)
TreeRate	*env*	16,90	12,10	21,60	Maljkovic Berry *et al*. (2007)
BEAST	*env*	5,80	3,80	7,80	Bello *et al*. (2008)
Hahn *et al*. (1986)	*gag*	-	0,37	1,85	Hahn *et al*. (1986)
Li *et al*. (1988)	*gag*	26,00[a]	-	-	Gojobori *et al*. (1994)
Li *et al*. (1988)	*gag*	1,00[b]	-	-	Gojobori *et al*. (1994)
Pairwise-Distance	*gag*	2,70	2,20	3,20	Leitner et Albert (1999)
Root-to-tip	*gag*	1,90	0,90	2,70	Korber *et al*. (2000)
BEAST	*pol*	2,19	1,83	2,56	Dalai *et al*. (2009)
BEAST	*pol*	1,50	1,00	2,00	Bello *et al*. (2008)

[a]Taux de substitution synonyme
[b]Taux de substitution non synonyme

2.6 Conclusion

Nous présentons dans ce chapitre différentes méthodes qui permettent d'estimer le taux de substitution, c'est-à-dire la vitesse évolutive, sous les hypothèses du modèle SRDT (horloge moléculaire stricte et séquences hétérochrones), comme les régressions linéaires *Pairwise-Distance* et *Root-to-tip*, les méthodes de distances sUPGMA, TREBLE et *TreeRate*, la méthode probabiliste Langley-Fitch qui utilise une approche de distance et les méthodes pleinement probabilistes *TipDate* (vraisemblance) et BEAST (bayésien). Certaines de ces méthodes sont étendues à des modèles d'horloge moléculaire plus complexes, comme le modèle MRDT (sUPGMA ou *TipDate*) ou le modèle DR (Langely-Fitch ou BEAST) et d'autres nécessitent de l'information supplémentaire, comme un arbre enraciné (Langley-Fitch) ou l'intervention de l'utilisateur afin de considérer deux groupes de séquences à par-

tir desquels le taux sera estimé (*TreeRate*). Les estimations du taux de substitution du VIH données tout au long de ce chapitre permettent de se faire une bonne idée sur l'ordre de grandeur de celui-ci et ne peuvent en aucun cas être utilisées pour comparer la performance des méthodes entre elles, étant donné que les jeux de données sont différents les uns des autres (Tableau 1). Ces estimations suggèrent que la vitesse évolutive du VIH est plus élevée sur le gène *env* que sur les gènes *gag* et *pol*. En effet, le gène *env* code pour un précurseur des glycoprotéines gp120 et gp41 qui sont exposées à la surface du virion et mutent beaucoup afin de chercher à échapper au système immunitaire.

Dans cette thèse nous proposons une méthode de distances qui permet d'estimer rapidement le taux de substitution sous les hypothèses du modèle SRDT, tout en gardant une bonne précision d'estimation. Un des objectifs est de pouvoir analyser de très grands jeux de données afin de stabiliser les estimations du taux de substitution proposées dans la littérature, et cela sous un modèle donné.