N° D'ORDRE :

UNIVERSITÉ PARIS XI
UFR SCIENTIFIQUE D'ORSAY

THÈSE

présentée

pour obtenir le grade de

DOCTEUR EN SCIENCES

DE L'UNIVERSITÉ PARIS XI ORSAY

Spécialité : Mathématiques

Par

Ana ARRIBAS GIL

ESTIMATION DANS DES MODÈLES A VARIABLES CACHÉES : ALIGNEMENT DE SÉQUENCES BIOLOGIQUES ET MODÈLES D'ÉVOLUTION

Soutenue le 18 mai 2007 devant la commission d'examen :

Mme.	GASSIAT	Élisabeth	Directrice de thèse
M.	LAVIELLE	Marc	Président
M.	LUGOSI	Gábor	Examinateur
Mme.	POURSAT	Marie-Anne	Examinatrice
M.	ROBERT	Christian	Rapporteur

A mis padres

Remerciements

Je voudrais tout d'abord remercier Élisabeth Gassiat sans qui rien de cela n'aurait été possible. Merci Élisabeth pour ta générosité, ta patience, ta disponibilité et surtout pour la confiance que tu m'as accordée au cours de ces années.

Je tiens à remercier Jean-Claude Fort et Christian Robert pour avoir accepté de rapporter cette thèse. Merci à Marc Lavielle, Gábor Lugosi et Marie-Anne Poursat pour leur présence dans mon jury. Muchas gracias a Gábor Lugosi por haber venido especialmente para esta ocasión.

J'ai eu la chance de travailler pendant ma thèse avec Catherine Matias. Merci Catherine pour cette collaboration qui a été si enrichissante pour moi. Merci de m'avoir fait profiter de ta compétence et ton expérience. Merci surtout pour ton aide et tes conseils.

Je voudrais remercier Dirk Metzler et Jean-Louis Plouhinec avec qui j'ai pu approfondir l'aspect appliqué de ce travail. Thank you for your ideas and advice.

Je tiens à remercier Sylvie Mazan pour ses encouragements pendant mon stage de DEA. C'est sans doute grâce à elle que j'ai commencé cette thèse.

Merci aux personnes que j'ai côtoyées pendant mon séjour à Orsay et qui m'ont aidé d'une manière ou une autre. Merci spécialement à Yves Misiti qui m'a consacré son aide inestimable et beaucoup de son temps lors de ma première année de thèse, et qui m'a toujours montre sa sympathie.

Merci aux doctorants avec qui j'ai partagé toutes ces années. Je pense tout particulièrement à Neil, Graham, Bouthaina, Cristian, Héctor, Mario et Mélanie. Merci pour votre présence chaleureuse et votre soutien.

Merci aux amis, d'ici et là-bas, qui m'ont tant soutenu et encouragé. Gracias a Sabine, Daniela, Héctor, Cristian y Natalia, Carmen y Rocío, Sandra, Laura y Suni por haber estado siempre ahí.

Gracias a mis padres por su apoyo constante y su cariño.

Gracias a Lisandro, por tantas cosas...

Table des matières

Chapitre 1

Introduction

Sommaire

Cette thèse est consacrée à l'estimation paramétrique dans certains modèles d'alignement de séquences biologiques. Il s'agit d'étudier d'un point de vue statistique les modèles et les algorithmes d'estimation utilisés en bioinformatique pour l'alignement de séquences.

Dans cette introduction on va d'abord faire une présentation du problème de l'alignement de séquences biologiques. Dans une deuxième partie on décrira plus précisément les sujets abordés dans les différents chapitres de la thèse ainsi que les résultats obtenus.

1.1 L'alignement de séquences biologiques

Le terme séquence biologique fait référence à toute chaîne de caractères contenant l'information d'une macromolécule biologique, et plus particulièrement aux séquences d'ADN ou génomiques et aux séquences protéiques.

L'ADN (acide désoxyribonucléique) est une molécule que l'on retrouve dans tous les organismes vivants. On dit que c'est le support de l'hérédité car il constitue le génome des êtres vivants et se transmet en totalité ou en partie lors des processus de reproduction. La molécule d'ADN possède une structure en double hélice composée de deux brins enroulés l'un autour de l'autre. Chacun des brins de l'ADN est un enchaînement de nucléotides. Chaque nucléotide est composé d'un phosphate, d'un sucre et d'une base azotée. Ce qui différencie un nucléotide d'un autre sont les bases azotées puisque le sucre et le phosphate sont identiques. Il existe quatre bases azotées différentes : deux sont dites purines (la guanine G et l'adénine A), les deux autres sont pyrimidines (la cytosine C et la thymine T). Les bases azotées sont complémentaires deux à deux : l'adénine s'associant avec la thymine et la guanine avec la cytosine.

Un brin d'ADN est formé par la répétition ordonnée de ces quatre nucléotides. Le second brin d'ADN est complémentaire au premier, c'est à dire, chaque nucléotide est relié à son complémentaire sur le premier brin. L'ADN est donc défini par la séquence de nucléotides d'un des deux brins et peut être représenté comme une séquence des caractères A, C, G et T.

A l'intérieur des séquences d'ADN (ou génomiques) on retrouve les gènes, des morceaux de séquence qui correspondent à une unité d'information génétique transmise par un individu à sa descendance. Plus concrètement, un gène est toute portion d'ADN destinée à être traduite en protéine. Chez les eucaryotes, un gène est constitué d'une alternance de séquences codantes, appelées exons, et de séquences non codantes, les introns, qui seront éliminées avant la traduction en protéine. Dans les régions codantes l'enchaînement des quatre nucléotides doit coder les 20 acides aminés constitutifs de protéines. En fait l'information génétique s'exprime par triplets de nucléotides (appelés codons), à chaque

codon correspond un acide aminé. Le nombre de codons possibles (4^3) étant largement supérieur à 20, il existe des codons différents qui codent pour le même acide aminé : on dit que le code génétique est dégénéré.

Une séquence protéique est l'enchaînement des vingt types d'acides aminés ; cette séquence est classiquement représentée par une chaîne de caractères qui utilise un alphabet de vingt lettres.

Avec le développement spectaculaire de la biologie moléculaire ces derniers années, le séquençage, c'est à dire l'obtention des séquences d'ADN ou protéiques conformant le génome d'une espèce, s'est banalisé permettant l'accès à un très grand nombre de séquences biologiques. Le génome complet d'environ deux cents espèces (parmi lesquelles l'homme, le chien, la souris, le rat et plusieurs poissons) ainsi que de morceaux de génomes de plus de quatre cents procaryotes et 300 eucaryotes sont disponibles sur les bases publiques de séquences telles que Genbank pour des séquences génomiques (http ://www.ncbi.nlm.nih.gov/Genbank/) ou UniProt pour des séquences protéiques (http ://www.uniprot.org/).

Cet afflux de données a impulsé le développement de méthodes de comparaison de séquences. La comparaison de séquences est de loin la tâche informatique la plus fréquemment exécutée par les biologistes. Il s'agit de déterminer dans quelle mesure deux séquences, génomiques ou protéiques, se ressemblent.

Un premier but de la comparaison de séquences est d'inférer des connaissances sur une séquence à partir des connaissances attachées à une autre. Ainsi, si deux séquences génomiques sont très similaires, et si l'une est connue pour être codante, l'hypothèse que la seconde le soit aussi peut être avancée. De même, si deux séquences protéiques sont similaires, il est souvent fait l'hypothèse que les protéines correspondantes assument des fonctions semblables ; si la fonction de l'une est connue, la fonction de la seconde peut ainsi s'en déduire. En effet, si deux séquences protéiques sur deux espèces différentes se ressemblent, et que ces deux espèces ont évolué à partir d'une séquence ancestrale commune, on peut supposer que les deux séquences proviennent de la même séquence protéique ancestrale et donc les protéines associées conservent les mêmes propriétés.

La comparaison de séquences sert aussi à l'identification de motifs importants pour l'expression ou la fonction d'un gène. En effet si l'on retrouve un même motif sur des séquences de plusieurs espèces suffisamment éloignées ce n'est probablement pas par hasard mais parce qu'il a été conservé le long du processus évolutif chez toutes ces espèces. Une raison qui expliquerait cette conservation est que le motif joue un rôle important dans la séquence.

C'est également en comparant des séquences de génomes d'espèces actuelles qu'il est

possible de reconstruire des arbres phylogénétiques (voir explication un peu plus bas) qui rendent compte des relations évolutives entre les différentes espèces.

Le principe de comparaison, on vient de le voir, se justifie par des considérations sur le processus d'évolution qui agit sur les séquences génomiques. Des facteurs multiples sont à l'origine de modifications de la séquence génomique. Parmi ceux qu'on connaît on peut distinguer deux types : les mutations ponctuelles constituées essentiellement des substitutions, insertions et délétions (ou suppressions) au niveau de la séquence d'ADN (i.e. des nucléotides) ; et les réarrangements génomiques qui correspondent à des modifications, à l'échelle des gènes, dans l'organisation des génomes (parmi lesquelles les duplications d'une portion de chromosome, l'inversion de l'ordre des gènes dans un chromosome ou encore des transferts d'une portion de chromosome d'une espèce à une autre). Ces erreurs et ces mutations sont susceptibles de se propager au sein des populations. Ainsi, la séquence d'un génome d'une espèce, c'est à dire l'enchaînement des nucléotides qui composent les macromolécules d'ADN au sein de ses chromosomes, évolue dans le temps.

L'histoire des espèces peut être représentée par un arbre, dont les feuilles sont les espèces actuelles (voir Figure 1.1). Ce type d'arbre est connu sous le nom d'arbre phylogénétique. Chacun des nœuds de l'arbre représente l'ancêtre commun de ses descendants. Les longueurs des branches représentent les distances évolutives, en termes du temps d'évolution, entre les espèces. Deux espèces sont considérées d'autant plus proches que leur espèce ancestrale commune est récente.
On peut construire aussi des arbres phylogénétiques pour les séquences biologiques. Toutes les séquences d'une espèce n'évoluent pas de la même façon, donc la phylogénie des séquences ne correspond pas forcement à la phylogénie des espèces dont elles sont issues. Notamment, dans les arbres phylogénétiques de séquences, les longueurs des branches représentent les distances évolutives en termes du nombre de mutations espérées par site, ce sont donc des distances relatives (ceci sera détaillé par la suite). Deux gènes de deux espèces différentes et issus d'un même gène ancestral sont dits "homologues". De même, deux sites de deux séquences différentes issus du même site d'une séquence ancestrale commune sont dits "homologues".

Pour comparer deux séquences, la méthode la plus employée est l'alignement. Aligner deux séquences est un moyen de les comparer en mettant en évidence les ressemblances qui existent entre elles. Un alignement est représenté sous la forme d'une matrice à deux lignes, chaque ligne contenant une des séquences à aligner. Par exemple, un alignement

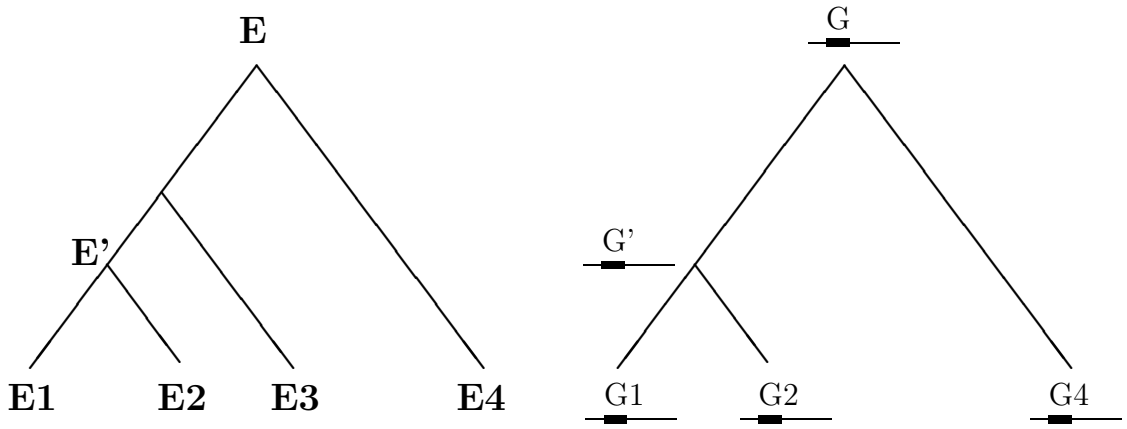


FIG. 1.1 – *Arbre phylogénétique des espèces (gauche) et arbre phylogénétique de séquences issues de ces mêmes espèces (droite). Les deux espèces E1 et E2 possèdent une espèce ancestrale commune (E') plus récente que l'espèce ancestrale E commune à E2 et E4. E1 et E2 seront donc considérées comme plus proches que E2 et E4. Les gènes G1 et G2, sont homologues car issus d'un gène ancestral commun G'. Il en est de même des gènes G2 et G4, dont le gène ancestral commun G est cependant plus lointain. On peut donc s'attendre à ce que les séquences de G1 et G2 se ressemblent plus que celles de G2 et G4. Le gène G n'a pas été conservé chez l'espèce E3 d'où la différence entre les deux arbres.*

des mots SCIENTIFIQUE et STATISTIQUE peut se représenter de la manière suivante :

$$\begin{array}{cccccccccccccc} S & C & I & E & N & - & - & T & I & - & F & I & Q & U & E \\ S & - & - & - & - & T & A & T & I & S & T & I & Q & U & E \end{array}$$

Quand il s'agit de séquences biologiques l'alignement est utilisé pour montrer la correspondance évolutive entre séquences différentes. Ceci dit, avec l'alignement on ne sait prendre en compte que les substitutions, insertions et délétions pour l'instant. Quand on met en face deux séquences biologiques ayant évolué à partir d'un ancêtre commun le problème de l'alignement consiste à retrouver les endroits où des substitutions, insertions ou délétions ont eu lieu. La mise en correspondance d'un caractère d'une séquence avec un *gap* (espace vide representé par le symbole -) dans l'autre séquence s'interprète soit comme une délétion du caractère dans la première séquence, soit comme une insertion d'un caractère dans la seconde, et rend compte d'une insertion ou d'une délétion d'un nucléotide dans la macromolécule d'ADN ou d'un acide aminé dans le polypéptide. La mise en correspondance de deux caractères différents rend compte d'une substitution.

Pour ce qui est de l'alignement de plus de deux séquences l'enjeu est le même, il s'agit

de les comparer pour retrouver des positions homologues sur plusieurs séquences et celles qui au contraire ont subi des mutations le long du processus d'évolution. Évidemment l'alignement multiple est un outil de comparaison plus puissant puisqu'il permet de détecter des ressemblances entre plus de deux séquences, mais aussi plus difficile à mettre en oeuvre comme on le verra par la suite.

Quant au type de séquences à aligner, génomiques ou protéiques, cela dépend de ce que l'on veut étudier. Si l'on veut comparer deux protéines une comparaison sur les séquences protéiques s'avère plus pertinente. D'un coté la dégénérescence du code génétique pourrait faire que l'on retrouve des différences entre deux protéines identiques si l'on compare leurs séquences génomiques. D'un autre coté, le fait de travailler sur un alphabet de taille 20 au lieu d'un alphabet de taille 4 permet des comparaisons plus fines. Mais on peut aussi s'intéresser à des séquences génomiques non codantes qui, comme on l'a montré récemment (voir [71] par exemple), peuvent jouer un rôle très important dans l'expression des gènes. Dans ce cas on alignera évidemment des séquences génomiques.
Toutes les méthodes d'alignement sont applicables aux deux types de séquences biologiques à condition de les utiliser avec l'alphabet et le processus d'évolution relatifs au type de séquence considéré.

Dans la suite on va d'abord présenter les méthodes d'alignement de séquences les plus courantes, les méthodes d'alignement par *score*, qui reposent sur des algorithmes de programmation dynamique. On parlera de leurs limitations et on justifiera l'introduction de modèles d'évolution dans l'alignement de séquences. On introduira ensuite le modèle d'évolution TKF ([76, 77]), pionnier dans la modélisation des insertions et délétions, ainsi que le modèle probabiliste sous-jacent, le modèle Markov caché pair. On présentera les généralisations du modèle TKF et finalement on parlera de l'alignement multiple de séquences.

1.1.1 Méthodes d'alignement par programmation dynamique

Ces méthodes consistent à assigner un coût ou *score* à chaque alignement possible entre deux séquences données et à choisir l'alignement avec le meilleur *score* à l'aide d'algorithmes de programmation dynamique.
Le *score* que l'on attribue à un alignement est la somme des *scores* individuels affectés à chaque paire de nucléotides alignés plus les termes de *score* pour chaque *gap*. L'ensemble des coûts de substitution est donné dans une matrice symétrique, dont la taille est la taille de l'alphabet, appelée matrice de substitution. Le coût des *gaps* est défini par celle qu'on appelle la fonction de pénalité de *gap*.

• Matrices de substitution :

Le *score* de la mutation de la lettre x par la lettre y, $s(x, y)$, s'interprète souvent comme le logarithme du rapport de vraisemblances suivant : les nucléotides x et y sont reliés (par un certain processus évolutif), contre ils sont indépendants. On notera $s(x, y) = \log\left(\frac{p(x,y)}{q_x q_y}\right)$. Intuitivement on espère que des conservations soient plus probables dans des séquences reliées qu'elles ne le sont par chance, et elles vont donc avoir un *score* positif, et de même pour les substitutions conservatives (celles dont les nucléotides ou acides aminés concernés sont "proches"). Au contraire, on espère trouver moins de substitutions non conservatives entre deux séquences reliées que entre deux séquences indépendantes, et elles auront donc un *score* négatif. D'une façon plus rigoureuse, pour pouvoir calculer ces probabilités il faut se donner un modèle probabiliste de substitution.
Pour ce qui est des séquences d'ADN, la plupart des modèles de substitution sont des modèles markoviens à temps continu réversibles dans le temps. Dans ce cas q_x est la probabilité stationnaire du caractère x, et $p(x, y)$ est q_x fois la probabilité de transition de x vers y. Parmi les modèles les plus courants on peut citer ceux de Jukes et Cantor [40], Felsenstein [21] ou encore Kimura [42] (pour plus de détails sur ces modèles se référer à [22]). Ceci dit, les méthodes d'alignement par *score* s'utilisent rarement sur des séquences d'ADN mais beaucoup plus sur des séquences protéiques.
Quand aux séquences protéiques, les matrices de substitution les plus utilisées sont les familles de matrices PAM [16] et BLOSUM [32]. Elles ne reposent pas directement sur un modèle probabiliste d'évolution. Avec des méthodes différentes dans les deux cas, le calcul de la probabilité d'observer une substitution dans des séquences reliées se fait essentiellement par comptage du nombre de fois que l'on observe la dite substitution dans des alignements de séquences proches qui servent comme échantillon d'apprentissage. Pour différents alignements de séquences plus ou moins proches on en déduit des matrices PAM et BLOSUM qui modélisent des distances évolutives plus ou moins longues.

• Fonction de pénalité de *gaps* :

Les insertions et délétions sont des événements évolutifs qui se produisent beaucoup moins souvent que les substitutions. Ainsi, dans les méthodes par *score* on pénalise les *gaps* en leur attribuant des *scores* négatifs. Les deux fonctions de pénalité de *gaps* les plus courantes sont la fonction linéaire

$$\gamma(g) = -gd$$

et la fonction affine

$$\gamma(g) = -d - (g-1)e$$

où g est la longueur du *gap*, d est le coût d'ouverture de *gap* et e est le coût d'extension de *gap*. En général le coût d'extension de *gap* est inférieur au coût d'ouverture de *gap*.

Comme pour les *scores* de substitution il existe une interprétation probabiliste des fonctions de *gaps* en terme du rapport de vraisemblances du *gap* sur le modèle où les séquences seraient reliées et sur le modèle où elles seraient indépendantes. Si les séquences sont reliées alors la probabilité d'avoir un *gap* de longueur g serait $f(g)\prod_{i\in G} q_i$ où f est une fonction qui ne dépend que de la longueur du *gap* et les indices $i \in G$ correspondent aux caractères dans le *gap*. La probabilité du *gap* dans le modèle indépendant serait $\prod_{i\in G} q_i$ et donc on aurait $\gamma(g) = \log(f(g))$.

Une fois choisie une fonction de *score* (combinaison de modèle de substitution et fonction de pénalité de *gaps*) on peut rechercher l'alignement optimal (celui avec le plus grand *score*). Considérons l'exemple suivant. On cherche le meilleur alignement global entre les séquences génomiques $X = ACCTGA$ et $Y = ACGT$. Pour ceci on se donne une fonction de pénalité de *gaps* linéaire avec coût d'ouverture de *gap* $d = 1$ et la matrice de substitution suivante :

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Avec une telle fonction de *score* il est évident que l'alignement optimal est parmi ceux qui ont un nombre de *gaps* minimal. En effet, les alignements optimaux pour ces deux séquences sont

$$\begin{array}{cccccc} A & C & C & T & G & A \\ A & C & G & T & - & - \end{array} \qquad \begin{array}{cccccc} A & C & C & T & G & A \\ A & C & - & - & G & T \end{array} \qquad \begin{array}{cccccc} A & C & C & T & G & A \\ A & - & C & - & G & T \end{array}$$

avec un même *score* de 1.

L'algorithme de référence pour l'alignement global de deux séquences est celui de Needleman et Wunsch [62]. Soient $x_{1:n}$ et $y_{1:m}$ les séquences à aligner. L'algorithme consiste à créer un tableau S de dimension $n \times m$ où $S(i,j)$ est le *score* du meilleur alignement des sous séquences $x_{1:i}$ et $y_{1:j}$. L'algorithme s'initialise avec $S(0,0) = 0$ et la matrice S se remplit de la façon suivante :

$$S(i,j) = \max \begin{cases} S(i-1,j-1) + s(x_i, y_j) \\ S(i-1,j) - d \\ S(i,j-1) - d \end{cases}$$

En même temps que l'on calcule $S(i,j)$ on garde en mémoire le sous alignement depuis lequel on a atteint le maximum, ce qui nous permet de retrouver le chemin du meilleur alignement à la fin de l'algorithme.
Ceci est la formulation originale de l'algorithme qui ne considère que des fonctions de pénalité de *gaps* linéaires. Sa généralisation à des fonctions de *gaps* affines nécessite trois matrices $n \times m$, M, I et D qui contiennent les *scores* des meilleurs sous alignements qui terminent par un *match*, une insertion dans la séquence y et une délétion dans la séquence x respectivement. Les coefficients (i,j) se calculent de la façon suivante :

$$\begin{aligned} M(i,j) &= \max \begin{cases} M(i-1,j-1) + s(x_i, y_j) \\ I(i-1,j-1) + s(x_i, y_j) \\ D(i-1,j-1) + s(x_i, y_j) \end{cases} \\ I(i,j) &= \max \begin{cases} M(i,j-1) - d \\ I(i,j-1) - e \end{cases} \\ D(i,j) &= \max \begin{cases} M(i-1,j) - d \\ D(i-1,j) - e \end{cases} \end{aligned} \tag{1.1}$$

Cette formulation ne permet pas d'avoir des insertions suivies par des délétions ni des délétions par des insertions. En fait, ceci n'arrive pas sur le chemin optimal si $-d-e$ est inférieur au plus petit *score* associé à une substitution.

L'alignement local, c'est à dire la détermination des sous séquences homologues parmi un ensemble de longues séquences est aussi important pour l'analyse de séquences biologiques. L'algorithme de Smith et Waterman [70], basé sur celui de Needleman et Wunsch, donne une réponse à ce problème. Il s'agit d'une optimisation du *score* sur l'ensemble des sous séquences possibles. Cet algorithme reste la méthode standard pour l'alignement local, néanmoins il s'avère lent pour la comparaison d'une séquence avec toute une base de données.

L'enjeu majeur de l'alignement de deux séquences est de savoir dans quelle mesure elles se ressemblent. Dans le cadre des méthodes d'alignement par *score* il s'agit de déterminer la significativité du *score* d'un alignement, c'est à dire de pouvoir décider à partir de la valeur du *score* si l'alignement met en évidence des relations biologiques entre des séquences liées ou si il est seulement le meilleur alignement entre deux séquences complètement indépendantes. Ceci est fait en calculant la probabilité d'obtenir un *score* supérieur à la valeur observée, sous l'hypothèse que les séquences soient indépendantes. Dans le cas de l'alignement local sans *gaps*, Karlin et Altschul [41] ont donné la loi des valeurs extrêmes du *score* de deux séquences indépendantes. Dans le cas de l'alignement avec *gaps* on n'a

pas de résultat équivalent à celui de Karlin et Altschul. Le travail théorique le plus complet dans le cas de l'alignement avec *gaps* est celui de Grossman et Yakir [28] qui donnent une borne supérieure pour la loi des valeurs extrêmes du *score*. D'un point de vue heuristique Mott et Tribbe [61] étudient des approximations de cette loi et suggèrent qu'elle aurait un comportement similaire à celle de l'alignement sans *gaps*.

Dans le cadre de recherches de similitudes entre une nouvelle séquence et des séquences répertoriées dans des bases de données, les méthodes les plus utilisées sont les heuristiques BLAST [2] et FASTA [51, 64]. BLAST est une méthode heuristique pour trouver les meilleurs alignements locaux de plus grands *scores* entre une séquence donnée appelée séquence *query* et une banque de séquences. Il est important de noter que BLAST ne permet pas des insertions ou délétions mais l'algorithme permet de trouver plusieurs régions similaires (qui peuvent être alignées à la séquence *query* sans *gaps*) pour une même séquence de la banque. L'algorithme permet aussi de localiser presque toutes les régions similaires dont le *score* dépasse une valeur seuil (qui est déterminée comme une p-valeur de la loi des valeurs extrêmes du *score* à partir de résultats de [41]). L'algorithme FASTA est aussi orienté vers la comparaison d'une séquence *query* contre une banque de séquences. La différence avec BLAST est que FASTA permet les insertions et délétions. FASTA cherche d'abord les régions les plus denses en similitudes (sans *gaps*) pour après relier entre elles (en permettant des *gaps*) celles qui ont un plus grand *score* si la pénalité des *gaps* ajoutés ne dépasse pas une certaine valeur. Finalement, l'algorithme cherche l'alignement local optimal entre ces régions repérées dans la séquence *query* et les séquences de la base de données.

Malgré leur utilisation massive, toutes ces méthodes d'alignement présentent la même carence : le manque de critères objectifs pour le choix des fonctions de *score*. Pour ce qui est des processus de substitution, on doit se donner les paramètres des processus markoviens (dans le cas des séquences génomiques) ou choisir une matrice de substitution (dans le cas des séquences protéiques), ce qui est équivalent à faire des hypothèses a priori sur le processus d'évolution subi par les séquences. Quand aux fonctions de pénalité de *gaps*, même si elles n'ont pas de sens évolutif, le choix des paramètres induit aussi un biais dans l'alignement.
Cette faiblesse a conduit les chercheurs à utiliser deux approches dans l'usage de ces algorithmes. La première consiste à choisir arbitrairement des fonctions de *score* pour produire un premier alignement. Si celui-ci est "esthétiquement" satisfaisant on arrête le processus, dans le cas contraire on ajuste le *score* jusqu'à obtenir un alignement qui soit "esthétiquement" satisfaisant. La deuxième approche consiste à toujours utiliser la

même fonction de *score* pour tous les alignements. Même si cette deuxième technique est moins subjective que la première, le choix initial de la fonction de *score* reste arbitraire. Puisque le processus d'évolution est ce qui crée les différences entre les séquences, on devrait considérer les alignements de séquences dans le contexte de l'évolution et donner un sens évolutif à tous les paramètres de *score*, notamment à ceux qui modélisent les insertions et délétions. Par ailleurs il est fondamental d'estimer la valeur des paramètres à partir des séquences observées en même temps que l'on aligne les séquences, pour éviter tout biais dans l'analyse.
C'est en réponse à cette problématique qu'en 1986 Bishop et Thompson [9] décrivent une technique d'alignement de deux séquences basée sur un modèle évolutif. Son approche consiste à estimer par maximum de vraisemblance la distance évolutive entre les séquences à partir d'un modèle simple et produire des alignements dont la probabilité a posteriori est grande. Ce travail a inspiré l'article de Thorne, Kishino et Felsenstein (1991) [76] qui étend l'approche du maximum de vraisemblance à un modèle rigoureux d'évolution pour des séquences d'ADN (connu comme le modèle TKF) qui est devenu le modèle évolutif de référence des insertions et délétions.

1.1.2 Le modèle d'évolution TKF et le modèle Markov caché pair

Le modèle TKF [76] présente l'évolution de séquences d'ADN comme un processus qui agit en deux temps. Dans un premier temps une séquence subit un processus d'insertion-délétion qui agit de façon homogène sur toutes les positions de la séquence avec indépendance du nucléotide particulier présent à chaque position, donnant naissance à une nouvelle séquence. Dans un deuxième temps, et conditionnellement au résultat du processus d'insertion-délétion, un processus de substitution agit sur les deux séquences.
Le modèle TKF décrit seulement le processus insertion-délétion et peut se combiner avec n'importe quel processus de substitution. Cependant, le processus insertion-délétion est réversible dans le temps (se référer à l'Annexe pour les détails) donc pour que le modèle d'évolution final conserve cette propriété il faut choisir un modèle de substitution aussi réversible dans le temps. La propriété de réversibilité dans le temps, bien que biologiquement peu réaliste, est intéressant puisqu'elle permet de traiter deux séquences comme si l'une était l'ancêtre de l'autre au lieu de devoir sommer sur toutes les séquences ancestrales possibles des deux séquences.

Le modèle TKF utilise une représentation des séquences en terme de séquences de *liens* invisibles auxquels les caractères sont associés. L'introduction du concept artificiel de *lien* sert à rendre le modèle d'évolution plus compréhensible. Il existe deux types différents

de *liens*, des *liens mortels* (notés par $\star$) qu'on placera à droite de chaque caractère de la séquence, et un *lien immortel* (noté par $\bullet$) qui est placé tout au début de la séquence, à gauche de tous les caractères. Ainsi une séquence quelconque sera représentée par

$$\bullet B \star B \star B \star B \star B \star B \star B \star B \star \ldots$$

où B dénote n'importe quelle nucléotide. On ne s'intéresse pas à la nature particulière de chaque caractère puisque le processus insertion-délétion agit de façon homogène sur tous les caractères. En effet le processus d'insertion-délétion se décrit en termes d'un processus de naissance et mort sur les *liens*. Chaque *lien* évolue indépendamment de tous les autres *liens* ; la naissance ou mort d'un *lien* ne modifie pas la probabilité de naissance et mort des autres *liens*. Tous les *liens* produisent des naissances de *liens mortels* (qu'on placera par consensus à droite du *lien* original) avec un taux $\lambda > 0$. Quand une naissance se produit on associe au nouveau *lien* un caractère (dont la nature sera déterminée par le processus de substitution). Tous les *liens mortels* sont frappés par un processus de mort avec un taux $\mu > \lambda$. Quand un *lien* meurt, son caractère associé est supprimé. Un exemple du résultat d'un tel processus est l'alignement suivant :

$$\begin{array}{cccccccccc} \bullet & - & - & B\star & - & B\star & - & - & - & B\star\ldots \\ \bullet & B\star & B\star & - & B\star & B\star & B\star & B\star & B\star & B\star\ldots \end{array}$$

L'existence du *lien immortel* ainsi que le fait de considérer le taux de délétion μ supérieur au taux d'insertion λ permet d'avoir à l'équilibre une distribution pour la longueur des séquences qui auraient évolué selon ce modèle. En effet à l'équilibre la longueur d'une séquence suit une loi géométrique de paramètre λ/μ, c'est à dire que la probabilité qu'une séquence ait n caractères est $(1-\lambda/\mu)(\lambda/\mu)^n$, pour $n \geq 0$.

On remarque que puisque les *liens* évoluent tous de façon indépendante il suffit de comprendre comment agit le processus sur chacun d'entre eux. Pour la description des probabilités d'insertion et délétion sur chaque *lien* on utilise les notations de [74]. Soit $p_n^H(t)$ la probabilité qu'un *lien mortel* ait survécu après un temps t et qu'il ait eu n descendants (parmi lesquels lui-même). Ici H veut dire homologue. Soit $p_n^N(t)$ la probabilité qu'un *lien mortel* n'ait pas survécu après un temps t et qu'il ait eu n descendants (N veut dire non homologue). Soit finalement $p_n^I(t)$ la probabilité que le *lien immortel* ait eu n descendants après un temps t parmi lesquels lui-même (I veut dire immortel). La description du processus de naissance et mort sur chaque *lien* donne lieu à des équations différentielles sur chacun des termes $p_n^H(t)$, $p_n^N(t)$ et $p_n^I(t)$ (voir [76] ou [74]). Après résolution de ces

équations on obtient

$$\begin{aligned} p_n^H(t) &= e^{-\mu t}[1-\lambda\beta(t)][\lambda\beta(t)]^{n-1}\ n>0 \\ p_n^N(t) &= [1-e^{-\mu t}-\mu\beta(t)][1-\lambda\beta(t)][\lambda\beta(t)]^{n-1}\ n>1;\ \ p_0^N(t)=\mu\beta(t) \\ p_n^I(t) &= [1-\lambda\beta(t)][\lambda\beta(t)]^{n-1}\ n>0 \end{aligned} \tag{1.2}$$

avec $\beta(t)=\frac{1-e^{(\lambda-\mu)t}}{\mu-\lambda e^{(\lambda-\mu)t}}$.

Un des avantages du modèle TKF, et sans doute la raison pour laquelle il est devenu populaire, est que l'alignement de deux séquences peut se formuler comme une chaîne de Markov à trois états (insertion, délétion et *match* ou conservation). On parle d'alignement mais c'est en fait l'alignement sans spécification des caractères à chaque position des séquences qui est une chaîne de Markov. Par la suite on utilisera souvent le mot alignement pour désigner le résultat du processus d'insertion-délétion

On va voir ceci en suivant les idées de [30]. Considérons un *lien mortel*. Soit U_t la variable aléatoire qui vaut 1 si le *lien* a survécu après le temps t et 0 si il est mort après ce temps. Puisque le taux de mort est μ on a

$$\mathbb{P}(U_t=1)=e^{-\mu t}\quad et\quad \mathbb{P}(U_t=0)=1-e^{-\mu t}.$$

Soit N_t la variable aléatoire "nombre de descendants du *lien* après le temps t". On a

$$\begin{aligned} \mathbb{P}(N_t=n|U_t=1) &= \frac{p_n^H(t)}{e^{-\mu t}}=[1-\lambda\beta(t)][\lambda\beta(t)]^{n-1} && n>0 \\ \mathbb{P}(N_t=n|U_t=0) &= \frac{p_n^N(t)}{1-e^{-\mu t}}=\begin{cases} 1-\kappa(t) & n=0 \\ \kappa(t)[1-\lambda\beta(t)][\lambda\beta(t)]^{n-1} & n\geq 1 \end{cases} \end{aligned} \tag{1.3}$$

où $\kappa(t)=1-\frac{\mu\beta(t)}{1-e^{(-\mu t)}}$. Ceci implique que

$$\begin{aligned} \mathbb{P}(N_t\geq n+1|N_t\geq n,U_t=1) &= \lambda\beta(t) && n\geq 0 \\ \mathbb{P}(N_t\geq n+1|N_t\geq n,U_t=0) &= \begin{cases} \kappa(t) & n=0 \\ \lambda\beta(t) & n\geq 1. \end{cases} \end{aligned} \tag{1.4}$$

Alors la probabilité d'avoir une insertion après une autre insertion ($\mathbb{P}(N_t\geq n+1|N_t\geq n)$, $n\geq 1$) est toujours $\lambda\beta(t)$, la probabilité d'avoir une insertion après un *match* ($\mathbb{P}(N_t\geq 1|N_t\geq 0,U_t=1)$) est aussi $\lambda\beta(t)$ et la probabilité d'avoir une insertion après une délétion ($\mathbb{P}(N_t\geq 1|N_t\geq 0,U_t=0)$) est $\kappa(t)$. Donc la probabilité d'avoir une insertion ne dépend que de l'état immédiatement précédent. Pour les deux autres états, *match* et délétion, ceci est aussi vrai à cause de l'indépendance entre les *liens*. La probabilité d'avoir un *match* (respectivement une délétion) est la probabilité d'avoir un nouveau *lien* dans la première séquence (λ/μ), fois la probabilité de ne plus avoir d'insertions dans le *lien* précédent

($(1-\lambda\beta(t))$ ou $(1-\kappa(t))$) et fois la probabilité d'avoir le nouveau *lien* conservé (resp. pas conservé) dans la deuxième séquence. On obtient la matrice de transition suivante :

$$
\begin{array}{c}
 \\
\texttt{Start} \\
\begin{smallmatrix}\texttt{B}\\\texttt{B}\end{smallmatrix} \\
\begin{smallmatrix}\texttt{B}\\\texttt{-}\end{smallmatrix} \\
\begin{smallmatrix}\texttt{-}\\\texttt{B}\end{smallmatrix}
\end{array}
\begin{array}{c}
\begin{smallmatrix}\texttt{B}\\\texttt{B}\end{smallmatrix} \qquad\qquad \begin{smallmatrix}\texttt{B}\\\texttt{-}\end{smallmatrix} \qquad\qquad \begin{smallmatrix}\texttt{-}\\\texttt{B}\end{smallmatrix} \qquad\qquad \texttt{End} \\
\left(\begin{array}{cccc}
(1-\lambda\beta(t))(\frac{\lambda}{\mu})\alpha(t) & (1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t)) & \lambda\beta(t) & (1-\lambda\beta(t))(1-\frac{\lambda}{\mu}) \\
(1-\lambda\beta(t))(\frac{\lambda}{\mu})\alpha(t) & (1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t)) & \lambda\beta(t) & (1-\lambda\beta(t))(1-\frac{\lambda}{\mu}) \\
(1-\kappa(t))(\frac{\lambda}{\mu})\alpha(t) & (1-\kappa(t))(\frac{\lambda}{\mu})(1-\alpha(t)) & \kappa(t) & (1-\kappa(t))(1-\frac{\lambda}{\mu}) \\
(1-\lambda\beta(t))(\frac{\lambda}{\mu})\alpha(t) & (1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t)) & \lambda\beta(t) & (1-\lambda\beta(t))(1-\frac{\lambda}{\mu})
\end{array}\right)
\end{array}
\tag{1.5}
$$

où $\alpha(t) = e^{-\mu t}$. La ligne `Start` correspond à la probabilité initiale de chacun des états. Elle est calculée à partir des probabilités du *lien immortel*. Par exemple, la probabilité de commencer l'alignement par un *match* est la probabilité que le *lien immortel* n'ait plus de descendants que lui même ($p_t^I(1)$) fois la probabilité d'avoir un nouveau *lien* dans la première séquence (λ/μ) qui est en plus conservé dans la deuxième séquence ($\alpha(t)$). L'état `End` est un état absorbant auquel on arrive quand il n'y plus d'insertions ($1-\lambda\beta(t)$) ni de nucléotides dans la première séquence ($1-\lambda/\mu$).
Le temps d'évolution (ou distance évolutive entre les séquences) t se mesure en fonction du nombre de mutations par site entre deux séquences plutôt qu'en unités réelles de temps. Par exemple, sous le modèle TKF, μt est le nombre espéré de délétions par site pour deux séquences à distance t. Cependant cette distance est relative et n'a de sens que si on la compare à d'autres distances évolutives, donc pour l'alignement de seulement deux séquences t sera toujours fixé à 1.

Notons $\{\varepsilon_s\}_{s\geq 1}$ la chaîne de Markov de l'alignement. Si on choisit un processus de substitution markovien avec matrice de transition $p(\cdot,\cdot)$ (pour un temps $t=1$) et loi stationnaire q, la vraisemblance de deux séquences $x_{1:n}$ et $y_{1:m}$ conditionnellement à un alignement de longueur ℓ, noté $e_{1:\ell}$, est

$$
\mathbb{P}_\theta(x_{1:n},\, y_{1:m} \,|\, \varepsilon = e) = \prod_{i=1}^{\ell} q_{x_{n_i}}^{\mathbb{1}\{e_i=\begin{smallmatrix}\texttt{B}\\\texttt{-}\end{smallmatrix}\}}\, q_{y_{m_i}}^{\mathbb{1}\{e_i=\begin{smallmatrix}\texttt{-}\\\texttt{B}\end{smallmatrix}\}}\, \left(q_{x_{n_i}} p(x_{n_i}, y_{m_i})\right)^{\mathbb{1}\{e_i=\begin{smallmatrix}\texttt{B}\\\texttt{B}\end{smallmatrix}\}} \tag{1.6}
$$

où n_i dénote le caractère dans la séquence x correspondant à la position i de l'alignement et de même pour la séquence y et m_i, et θ contient tous les paramètres du modèle. Ceci veut dire que la probabilité de chacun des nucléotides de la séquence x est donnée par la loi stationnaire du processus de substitution et que quand on a une insertion le type de nucléotide est aussi tiré selon cette loi. On remarque que, conditionnellement à l'alignement, le processus de substitution est indépendant site par site.

Par rapport à ce dernier point, on voudrait signaler l'existence de modèles de substitution qui ne rentrent pas dans ce cadre mais qui sont biologiquement beaucoup plus réalistes. En effet il s'agit de modèles de substitution à contexte dépendant, c'est à dire, la substitution d'un caractère par un autre dépend des caractères qui l'entourent dans la séquence (voir par exemple [10, 38]). Il existe des méthodes récentes d'alignement par *score* à contexte dépendant (voir [26]), mais elles présentent un inconvénient majeur : le *score* d'un alignement dépend de l'ordre dans lequel les substitutions apparaissent (par exemple le *score* de l'alignement $\begin{matrix} a & b & c & d \\ a & b' & c' & d \end{matrix}$ dépend de si on a fait $abcd \to ab'cd \to ab'c'd$ ou $abcd \to abc'd \to ab'c'd$). Dans un cadre probabiliste, les modèles de substitution à contexte dépendant sont utilisés jusqu'à présent dans l'alignement sans *gaps* mais ils n'ont pas encore été combinés avec des processus d'insertion-délétion. L'utilisation de ces modèles de substitution dans le cadre du modèle TKF serait donc une approche très intéressante qu'on envisage d'étudier.

En revenant sur le modèle TKF, la loi jointe des séquences et un alignement s'écrit

$$\mathbb{P}_\theta(x_{1:n},\, y_{1:m},\, \varepsilon = e) = \mathbb{P}_\theta(\varepsilon_1 = e_1) \prod_{i=2}^{\ell} \mathbb{P}_\theta(\varepsilon_i = e_i | \varepsilon_{i-1} = e_{i-1})\, \mathbb{P}_\theta(x_{1:n},\, y_{1:m} \,|\, \varepsilon = e). \quad (1.7)$$

La vraisemblance des séquences $x_{1:n}$ et $y_{1:m}$ est simplement la somme de cette quantité sur tous les alignements possibles des deux séquences.

Le modèle d'alignement issu du modèle d'évolution TKF correspond à ce qui a été appelé le modèle Markov caché pair. Le modèle Markov caché pair (ou pair-HMM de l'anglais pair Hidden Markov Model) a été introduit par Durbin *et al.* en 1998 ([19]) dans le cadre de l'alignement de deux séquences biologiques. Ce modèle est défini comme une chaîne de Markov cachée à trois états (insertion, délétion et *match*, plus un état absorbant `End`) qui émet deux suites de variables aléatoires qui correspondent aux deux séquences observées. La chaîne de Markov cachée correspond elle à l'alignement. Conditionnellement à l'alignement, les caractères émis à différentes positions de l'alignement sont indépendants. La loi jointe des observations et une trajectoire du processus caché se calcule comme en (1.7) où q et p seraient les lois d'émission. Dans la définition du pair-HMM donnée par Durbin *et al.* [19] la chaîne de Markov ne permet pas des transitions entre les insertions et les délétions, mais une définition plus générale est possible si on supprime cette contrainte. C'est avec cette définition générale qu'on peut dire que l'alignement sous le modèle TKF est un pair-HMM.

La seule différence entre les modèles de Markov cachés classiques (HMMs) et les pair-HMMs est l'émission de deux séquences au lieu d'une seule. Dans la pratique les deux

modèles ne sont pas très différents et les algorithmes qui permettent de sommer sur tous les chemins cachés possibles (Forward) ou de trouver le chemin caché le plus probable a posteriori (Viterbi) sont aussi valables pour les pair-HMMs.

Cette interprétation du modèle TKF conduit à des procédures d'alignement qui, en utilisant les algorithmes pour pair-HMM (voir [19]), sont aussi rapides que les méthodes de programmation dynamique basées sur des *scores* avec l'avantage d'avoir un sens évolutif dans les paramètres qu'interviennent et de pouvoir les estimer en même temps que l'on cherche l'alignement. En effet un schéma d'alignement par *score* correspond exactement à un modèle pair-HMM si on considère une fonction de *gap* affine (pour une fonction de *gap* linéaire ça correspond à un modèle plus simple où les variables aléatoires de l'alignement seraient indépendantes). Cependant, pour une fonction de *gap* plus complexe que la fonction affine on n'a plus de correspondance entre le pair-HMM et l'alignement par *score*. On peut résumer les relations entre les trois types de modèles avec le diagramme suivant :

$$\text{TKF} \subset \text{pair-HMM} \subset \textit{scores}$$

(en négligeant le fait que les méthodes par *score* ne considèrent pas en général des transitions entre les insertions et les délétions).
Si on revient sur la fonction de *gap* affine, qui reste quand même la plus fréquemment utilisée, et si on considère l'interprétation probabiliste des *scores* évoquée dans la section précédente, on peut constater que la log-vraisemblance d'un alignement des séquences $x_{1:n}$ et $y_{1:m}$ (1.7) et le *score* de cet alignement sont égaux sauf pour la constante additive $-\log(\prod_i^n q_{x_i} \prod_j^m q_{y_j})$. Aussi, l'additivité du schéma de *score* correspond à l'indépendance des observations émises à des instants différents du processus caché dans le pair-HMM.
Par rapport aux algorithmes, on retrouve que l'algorithme Viterbi pour les pair-HMM (avec la définition de Durbin *et al.* [19]) est exactement celui décrit dans (1.1). L'algorithme Forward, qui calcule la vraisemblance de deux séquences dans un pair-HMM, est similaire à l'algorithme Viterbi sauf que maintenant on somme sur tous les alignements possibles au lieu de chercher un alignement optimal (on remplace donc les max par des sommes). Si on note M l'état match, I l'état insertion, D l'état délétion, π_{ij} la probabilité de transition de i à j et $\pi(\cdot)$ la loi initiale, alors l'algorithme Forward calcule pour chaque i de 1 à n et pour chaque j de 1 à m :

$$\begin{aligned} P(i,j,M) &= p(x_i,y_j)[P(i-1,j-1,M)\pi_{MM} \\ &\qquad +P(i-1,j-1,I)\pi_{IM} + P(i-1,j-1,D)\pi_{DM}] \\ P(i,j,I) &= q_{y_j}[P(i,j-1,M)\pi_{MI} + P(i,j-1,I)\pi_{II} + P(i,j-1,D)\pi_{DI}] \\ P(i,j,D) &= q_{x_i}[P(i-1,j,M)\pi_{MD} + P(i-1,j,I)\pi_{ID} + P(i-1,j,D)\pi_{DD}] \end{aligned} \quad (1.8)$$

La vraisemblance des séquences $x_{1:n}$ et $y_{1:m}$ est donc $P(n,m,M)+P(n,m,I)+P(n,m,D)$.

La question de la validité de l'alignement se pose aussi dans le cadre des pair-HMMs. Il ne s'agit plus de déterminer la significativité d'un seul alignement (car on ne travaille pas avec des alignement optimaux) mais plutôt de décider si deux séquences peuvent être considérées comme étant reliées par un modèle d'évolution pair-HMM ou si au contraire elles sont indépendantes. Un travail en collaboration avec Elisabeth Gassiat et Catherine Matias est en cours pour mettre en place un test du rapport de vraisemblance pour séparer ces deux hypothèses.

En revenant sur le modèle d'évolution TKF, signalons qu'il a inspiré de nombreux travaux dans le cadre de l'estimation des paramètres d'évolution dans un modèle pair-HMM (voir [31] et [56] par exemple) ainsi que de nouveaux modèles évolutifs d'insertion et délétion que nous présentons maintenant.

1.1.3 Les généralisations du modèle TKF

Le principal inconvénient du modèle TKF (qu'on notera dorénavant TKF91) est que les insertions et délétions ne peuvent se produire que nucléotide par nucléotide. C'est pourquoi, Thorne, Kishino et Felsenstein (1992) [77] ont étendu leur modèle au modèle TKF92 qui permet l'insertion et la délétion de plusieurs nucléotides à la fois. On considère la séquence en terme de *liens* comme dans le modèle TKF91 mais maintenant chaque *lien* est associé non pas à un seul nucléotide mais à un fragment de nucléotides dont la longueur suit une loi géométrique de paramètre γ. Ainsi une séquence quelconque sera représentée par

- $\boxed{\text{B B B}} \star \boxed{\text{B}} \star \boxed{\text{B B}} \star \boxed{\text{B}} \star \boxed{\text{B}} \star \boxed{\text{B B B B}} \star \ldots$

où les boîtes représentent les fragments. Quand un *lien* est supprimé au taux μ tout son fragment associe est supprimé. Quand un nouveau *lien* naît, un fragment de nucléotides (dont la longueur est tirée selon la loi géométrique de paramètre γ) est inséré à son coté. La matrice de transition de l'alignement est maintenant :

$$
\begin{array}{c}
\\
\texttt{Start} \\
\begin{smallmatrix}\texttt{B}\\\texttt{B}\end{smallmatrix} \\
\begin{smallmatrix}\texttt{B}\\\texttt{-}\end{smallmatrix} \\
\begin{smallmatrix}\texttt{-}\\\texttt{B}\end{smallmatrix}
\end{array}
\begin{array}{c}
\begin{smallmatrix}\texttt{B}\\\texttt{B}\end{smallmatrix} \qquad\qquad \begin{smallmatrix}\texttt{B}\\\texttt{-}\end{smallmatrix} \qquad\qquad \begin{smallmatrix}\texttt{-}\\\texttt{B}\end{smallmatrix} \qquad\qquad \texttt{End} \\
\begin{pmatrix}
(1-\lambda\beta)(\frac{\lambda}{\mu})\alpha & (1-\lambda\beta)(\frac{\lambda}{\mu})(1-\alpha) & \lambda\beta & (1-\lambda\beta)(1-\frac{\lambda}{\mu}) \\
\frac{1}{\gamma}(1-\lambda\beta)(\frac{\lambda}{\mu})\alpha+1-\frac{1}{\gamma} & \frac{1}{\gamma}(1-\lambda\beta)(\frac{\lambda}{\mu})(1-\alpha) & \frac{1}{\gamma}\lambda\beta & \frac{1}{\gamma}(1-\lambda\beta)(1-\frac{\lambda}{\mu}) \\
\frac{1}{\gamma}(1-\kappa)(\frac{\lambda}{\mu})\alpha & \frac{1}{\gamma}\lambda\beta+1-\frac{1}{\gamma} & \frac{1}{\gamma}\kappa & \frac{1}{\gamma}(1-\kappa)(1-\frac{\lambda}{\mu}) \\
\frac{1}{\gamma}(1-\lambda\beta)(\frac{\lambda}{\mu})\alpha & \frac{1}{\gamma}(1-\lambda\beta)(\frac{\lambda}{\mu})(1-\alpha) & \frac{1}{\gamma}\lambda\beta+1-\frac{1}{\gamma} & \frac{1}{\gamma}(1-\lambda\beta)(1-\frac{\lambda}{\mu})
\end{pmatrix}
\end{array}
$$

Le modèle TKF92 n'a jamais été aussi populaire que le modèle TKF91. Les raisons sont probablement la nécessité d'un paramètre en plus pour modéliser la longueur des fragments insérés ou supprimés et le fait que les nucléotides qui ont une fois été insérés ensembles ne peuvent être supprimés qu'ensemble.

En réponse à ces limitations Metzler présente en 2003 [55] un nouveau modèle d'insertion et délétion par fragment (FID en anglais). En effet pour ne pas agrandir la taille de l'espace de paramètres il considère un seul taux commun pour les insertions et les délétions au lieu de deux considérés normalement. Ceci s'appuie sur le fait que dans la pratique, malgré la contrainte dans les modèles TKF qui impose que le taux de délétion soit supérieur au taux d'insertion, les deux taux s'avèrent très proches. En effet, la distribution stationnaire de la longueur des séquences sous le modèle TKF91 est $(1-\lambda/\mu)(\lambda/\mu)^n$ pour $n \geq 0$ ce qui privilégie les séquences courtes. Si on travaille avec des séquences longues il faut alors supposer que les deux taux sont égaux. Ceci entraîne qu'il n'existe plus de distribution stationnaire pour la longueur de séquence et que les séquences observées doivent être considérés comme des sous séquences de séquences beaucoup plus longues extraites entre des positions homologues connues.
Quand on fait tendre μ vers λ alors $\beta = \frac{1-e^{(\lambda-\mu)}}{\mu-\lambda e^{(\lambda-\mu)}}$ devient $\frac{1}{1+\lambda}$. La matrice de transition de l'alignement s'écrit dans ce modèle de la façon suivante :

$$\begin{array}{c} \\ \texttt{B} \\ \texttt{B} \\ \texttt{B} \\ \texttt{-} \\ \texttt{-} \\ \texttt{B} \end{array} \begin{array}{c} \begin{array}{ccc} \texttt{B} & \texttt{B} & \texttt{-} \\ \texttt{B} & \texttt{-} & \texttt{B} \end{array} \\ \begin{pmatrix} 1-\frac{1+\lambda-e^{-\lambda}}{\gamma(1+\lambda)} & \frac{1-e^{-\lambda}}{\gamma(1+\lambda)} & \frac{\lambda}{\gamma(1+\lambda)} \\ \frac{\lambda e^{-\lambda}}{\gamma(1-e^{-\lambda})(1+\lambda)} & \frac{\gamma(1+\lambda)-1}{\gamma(1+\lambda)} & \frac{1-(1+\lambda)e^{-\lambda}}{\gamma(1-e^{-\lambda})(1+\lambda)} \\ \frac{e^{-\lambda}}{\gamma(1+\lambda)} & \frac{1-e^{-\lambda}}{\gamma(1+\lambda)} & \frac{\gamma(1+\lambda)-1}{\gamma(1+\lambda)} \end{pmatrix} \end{array}$$

où γ joue le même rôle que dans le modèle TKF92. Il n'y a plus les états `Start` et `End` puisque on suppose que les séquences observées sont des sous séquences d'autres séquences plus longues. Maintenant la probabilité d'avoir un nouveau *lien* est toujours 1. A la différence des modèles TKF91 et TKF92, le modèle FID produit une chaîne de Markov stationnaire.
En ce qui concerne l'impossibilité du modèle TKF92 de diviser des fragments qui ont été insérés, ce problème apparaît aussi pour le FID. Ceci dit Metzler compare dans son papier [55] son modèle à un modèle plus général sans cette contrainte et montre que les résultats sont similaires avec un coût computationnel beaucoup plus grand pour ce dernier. Néanmoins il y a d'autres travaux qui insistent sur la nécessité de permettre des

insertions et délétions de taille variable à n'importe quel endroit de la séquence, comme celui de Miklós *et al.*(2004) [60].

1.1.4 L'alignement multiple

L'alignement multiple est l'extension naturelle à l'alignement de deux séquences. Son intérêt par rapport à ce dernier est évident. D'un coté l'alignement multiple permet de réaliser des comparaisons plus fines entre les séquences et d'obtenir des similarités à plusieurs niveaux. L'alignement multiple contient aussi de l'information sur l'histoire évolutive d'un ensemble de séquences ce qui va nous permettre de construire des arbres phylogénétiques. En effet, si plusieurs séquences ont évolué à partir d'un même ancêtre commun, l'information qu'elles partagent permet, en les alignant, de retracer le processus d'évolution suivi par ces séquences (voir Figure 1.2).

```
SEQ1 : A A C G T A T _ _ _ G G C C
SEQ2 : A A C G A A T _ _ _ C G C G
SEQ3 : A A C _ _ A T T A A G G _ _
SEQ4 : A A C _ _ A T T A A G A _ _
```

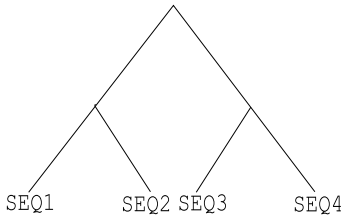


FIG. 1.2 – *L'alignement multiple de ces quatre séquences suggère un arbre phylogénétique comme celui à droite.*

A la différence de l'alignement de deux séquences qui sont alignées pour savoir dans quelle mesure elles se ressemblent, l'alignement multiple se fait en général avec des séquences dont ont sait a priori qu'elles sont liées.

Il existe plusieurs méthodes d'alignement multiple. On peut envisager la généralisation des méthodes par *score* utilisées dans l'alignement de deux séquences. Il existe des travaux dans ce sens (voir [19] pour un aperçu), mais ce n'est pas une approche très utilisée car il existe des ambiguïtés dans la définition des *scores* pour plus de deux séquences. De plus, les algorithmes de programmation dynamique deviennent lents quand le nombre de séquences augmente (la complexité algorithmique est $O(2^k L^k)$ pour k séquences de longueur moyenne L).

La famille des méthodes la plus courante pour l'alignement multiple est *l'alignement progressif*. Il consiste à construire un alignement multiple à partir d'alignements de deux séquences. Initialement on choisit les deux séquences qui se ressemblent le plus et on les aligne par une méthode de *score*. Successivement on choisit soit une nouvelle séquence soit

un autre alignement de deux séquences et on l'aligne à l'alignement déjà obtenu, jusqu'à ce qu'on ait aligné toutes les séquences. L'alignement de deux alignements se fait de la façon suivante : on cherche le meilleur alignement de deux séquences entre toutes les séquences du premier groupe et toutes les séquences du deuxième groupe ; cet alignement optimal déterminera l'alignement entre les deux groupes. Pour l'alignement d'une séquence à un alignement la méthode est la même. Il existe plusieurs algorithmes différents d'alignement progressif. Les différences entre eux reposent sur l'ordre et la façon choisis pour aligner les séquences et les fonctions de *score* utilisées. L'alignement obtenu n'est optimal en aucun sens. De plus il est très sensible à des changements dans l'ordre d'alignement des séquences. Par contre la méthode donne en général des résultats raisonnables (si on commence par aligner en premier les séquences qui se ressemblent le plus) et a l'avantage d'être très rapide. D'ailleurs la méthode d'alignement multiple la plus populaire, ClustalW [33, 75], est une méthode d'alignement progressif.

Une autre approche pour l'alignement multiple de séquences biologiques a été développée au cours des années quatre-vingt-dix. C'est celle initiée par Baldi *et al.*[5] et Krogh *et al.*[45] où les modèles de Markov cachés sont utilisés pour faire des alignements multiples et des tests d'homologie sur des familles de séquences bien connues. Ces modèles sont connus sous le nom de *profile*-HMMs d'après les *profils* standard (des structures similaires mais sans connotation probabiliste introduites par Gribskov *et al.* [27]). La technique consiste à créer une chaîne de Markov cachée avec N états *match*, N états délétion et $N + 1$ états insertion (voir Figure 1.3). Les états *match* correspondent à des positions homologues, c'est à dire issues d'une même position ancestrale (dans l'alignement (1.10) les cinq premières colonnes correspondent à des positions homologues et les deux dernières à des insertions). Leur nombre N est déterminé à partir de l'information disponible sur la famille de séquences. Les probabilités des transitions et des émissions sont aussi définies en fonction des séquences à aligner. Il s'agit donc d'un problème de sélection de modèle qui se résoud de la façon suivante : on a un ensemble de séquences d'apprentissage et on choisi N, en général comme étant la longueur moyenne des séquences (il est raisonnable que la séquence ancestrale et les séquences observées aient une longueur équivalente, voir le Chapitre 3 pour plus de détails), ou bien à partir de connaissances a priori sur les séquences ; ensuite on estime les paramètres du modèle par maximum de vraisemblance sur l'ensemble d'apprentissage.
Une fois le modèle choisi le HMM fournit une distribution de probabilité sur l'espace des séquences. Pour chaque séquence on peut donc donner le chemin le plus probable a posteriori avec l'algorithme de Viterbi. L'alignement multiple se construit alors en mettant ensemble les alignements les plus probables pour chacune des séquences. Ainsi, les positions des séquences alignées au même état *match* du *profile* HMM seront placées dans la

même colonne de l'alignement multiple.
Le principal inconvénient de cette approche c'est qu'elle suppose que les séquences sont générées indépendamment par le modèle quand en fait elles sont bien dépendantes et reliées par un processus commun d'évolution.

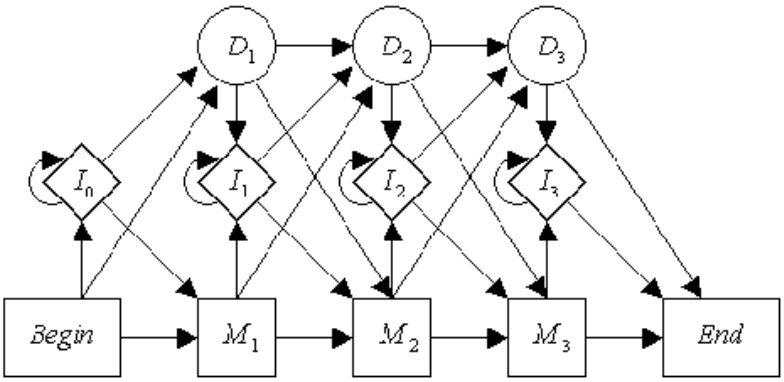


FIG. 1.3 – Profile *HMM avec trois états* match.

Comme c'était déjà le cas pour l'alignement de deux séquences, on retrouve dans les méthodes d'alignement multiple habituelles des carences importantes. D'un coté le processus d'évolution, qui joue un rôle encore plus important quand il s'agit de comparer plus de deux séquences, n'est pas pris en compte pour l'obtention des alignements. D'un autre coté le choix du modèle et des paramètres a priori, avec indépendance des séquences à aligner, introduit des biais dans la procédure d'alignement. C'est pourquoi il semble important d'étendre le modèle TKF91 aux alignements multiples. Ceci nécessite de prendre en compte non seulement le processus évolutif des insertions, délétions et mutations mais aussi la phylogénie des séquences à traiter.
On se donne alors les séquences à aligner et l'arbre phylogénétique qui les relie. Les séquences à aligner se placent dans les feuilles de l'arbre. La racine de l'arbre correspond à une séquence ancestrale commune à toutes les séquences. Le chemin qui relie la racine de l'arbre à une feuille représente l'évolution de la séquence ancestrale dans le temps et à travers une série de séquences intermédiaires donnant lieu à la séquence observée en question. On suppose que le même processus d'évolution agit sur toutes les branches de l'arbre, dans notre cas le processus TKF91. Une hypothèse importante est que le processus évolutif agit indépendamment dans chacune des branches de l'arbre, c'est à dire que chaque séquence évolue de façon indépendante vers chacun de ses descendants. Dans ce cadre, l'alignement multiple consiste à mettre dans la même colonne les caractères homo-

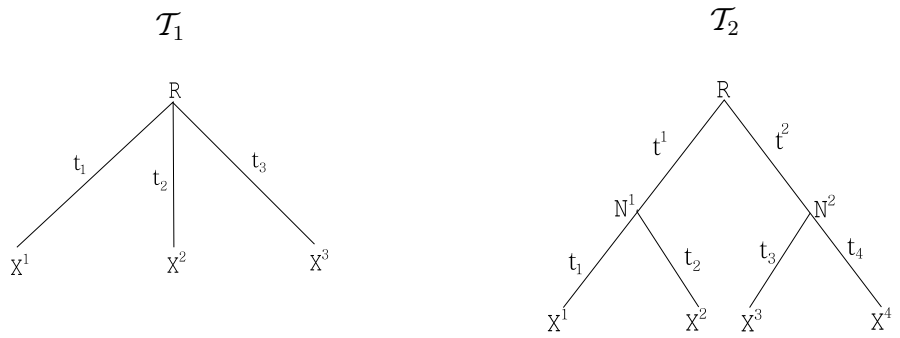


FIG. 1.4 – *Deux arbres phylogénétiques.*

logues, c'est à dire les caractères issus du même caractère ancestral. Il s'agit en fait de regrouper les alignements de la racine avec chacune des feuilles pour former l'alignement multiple.

Considérons l'exemple suivant. On a trois séquences X^1, X^2 et X^3 reliées par l'arbre $\mathcal{T}_1$ de la Figure 1.4. Si la racine, R, a évolué dans chacune des trois séquences de la façon suivante

$$\begin{array}{cccccccc} \mathrm{R}: & A & C & C & T & G & A & - \\ X^1: & A & C & G & - & G & A & T \end{array} \qquad \begin{array}{cccccccc} \mathrm{R}: & A & C & C & T & G & A \\ X^2: & A & C & - & T & G & G \end{array}$$

$$\begin{array}{cccccccc} \mathrm{R}: & A & C & C & T & G & A & - \\ X^3: & A & C & C & T & G & A & A \end{array}$$

alors l'alignement multiple est donné par

$$\begin{array}{cccccccc} X^1: & A & C & G & - & G & A & T & - \\ X^2: & A & C & - & T & G & G & - & - \\ X^3: & A & C & C & T & G & A & - & A \end{array} \tag{1.9}$$

Quand on a un arbre plus complexe, comme $\mathcal{T}_2$ dans la Figure 1.4, il faut aussi prendre en compte tous les alignements entre les nœuds internes de l'arbre. Ainsi, si le processus évolutif a donné

$$\begin{array}{ccccc} \mathrm{R}: & A & C & T & G \\ N^1: & A & C & - & G \end{array} \qquad \begin{array}{ccccc} \mathrm{R}: & A & C & T & - & G \\ N^2: & A & C & T & A & C \end{array}$$

$$\begin{array}{ccccc} N^1: & A & C & G & - \\ X^1: & A & C & G & T \end{array} \qquad \begin{array}{ccccc} N^2: & A & C & T & A & C \\ X^3: & A & - & T & A & G \end{array}$$

$$\begin{array}{cccc} N^1: & A & C & G \\ X^2: & A & - & G \end{array} \qquad \begin{array}{cccccc} N^2: & A & C & T & A & G & - \\ X^4: & A & C & T & A & C & G \end{array}$$

l'alignement multiple de ces quatre séquences sera

$$\begin{array}{lccccccc} X^1: & A & C & - & - & G & T & - \\ X^2: & A & - & - & - & G & - & - \\ X^3: & A & - & T & A & G & - & - \\ X^4: & A & C & T & A & C & - & G \end{array} \tag{1.10}$$

Les évolutions selon chaque branche sont indépendantes donc, conditionnellement aux séquences aux nœuds internes, la vraisemblance des séquences observées se calcule, par exemple dans ce dernier cas, comme

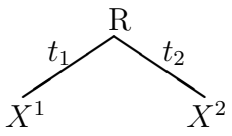
$$\mathbb{P}_\theta(R)\mathbb{P}_{\theta(t^1)}(N^1|R)\mathbb{P}_{\theta(t^2)}(N^2|R)\mathbb{P}_{\theta(t_1)}(X^1|N^1)\mathbb{P}_{\theta(t_2)}(X^2|N^1)\mathbb{P}_{\theta(t_3)}(X^3|N^2)\mathbb{P}_{\theta(t_4)}(X^4|N^2)$$

avec

$$\mathbb{P}_{\theta(t^1)}(N^1|R) = \frac{\mathbb{P}_{\theta(t^1)}(N^1,R)}{\mathbb{P}_\theta(R)} = \frac{\mathbb{P}_{\theta(t^1)}(N^1,R)}{(1-\frac{\lambda}{\mu})(\frac{\lambda}{\mu})^{|R|}\prod_{i=1}^{|R|} q_{r_i}} \tag{1.11}$$

et de même pour les autres couples de séquences, où $\mathbb{P}_{\theta(t^1)}(N^1,R)$ est la vraisemblance d'observer les séquences N^1 et R dans modèle TKF91 pour l'alignement de deux séquences, $|R|$ est la longueur de la séquence $R = r_1 \dots r_{|R|}$, et q est la loi stationnaire du processus de substitution choisi. Maintenant les paramètres du modèle TKF91 dépendent de la distance évolutive entre chaque paire de séquences, c'est à dire des longueurs de branche.
Dans la pratique on ne connaît ni les séquences aux nœuds internes ni le résultat du processus d'évolution. L'alignement multiple peut donc être vu comme un processus caché qui émet des caractères dans k séquences (k étant le nombre de séquences à aligner).

Les travaux qui étendent le modèle TKF91 aux alignements multiples sont récents. En 2000 Steel et Hein [72] présentent un algorithme pour calculer la vraisemblance d'un ensemble de séquences reliées par un arbre phylogénétique en forme d'étoile (arbre qui n'a pas d'autres nœuds internes que la racine), qui a été ensuite généralisé et amélioré (voir [29] et [59]). Ces travaux cependant ne rendent pas compte de la structure cachée sous-jacente à l'alignement multiple et les algorithmes reposent sur des calculs récursifs complexes. En effet, quand on prend en compte la structure cachée, la généralisation des algorithmes pour les pair-HMMs aux alignements multiples dévient évident. Holmes et Bruno [36] et Hein *et al.* [30] ont montré comment construire une HMM multiple à partir du modèle TKF91 pour un arbre phylogénétique quelconque. De la même façon qu'un pair-HMM émet des caractères dans deux séquences à partir de trois états cachées, un HMM multiple est une chaîne de Markov qui émet des caractères dans k séquences à partir de $2^k - 1$ états cachés (un pour chaque sous-ensemble non vide de l'ensemble des k séquences). Prenons un exemple. Considérons l'arbre phylogénétique suivant



qui représente le cas le plus simple d'alignement multiple. En fait, sous le modèle TKF91, aligner les séquences sous cet arbre est équivalent à aligner directement X^1 et X^2 (voir l'Annexe pour les détails), mais ce cas sert quand même à illustrer la construction des HMM multiples. Il y a deux types d'états cachés, ceux qui représentent un caractère dans la séquence ancestrale et sa conservation ou disparition dans les séquences observées, c'est à dire, des positions homologues

$$\begin{array}{cccc} R & R & R & R \\ B & B & - & - \\ B & - & B & - \end{array}$$

et ceux qui représentent les insertions dans les séquences observées

$$\begin{array}{ccc} (R) & (R) & (R) \\ (B) & (-) & B \\ B & B & (B) \end{array}$$

Ces derniers ne correspondent pas exactement à des colonnes dans l'alignement multiple. En effet, une insertion dans la séquence X^1 dépend de la dernière position de l'alignement où un événement évolutif en X^1 a eu lieu ; si entre les deux on a une insertion dans la séquence X^2 on n'a plus de dépendance markovienne. C'est pourquoi les états insertion *gardent en mémoire*, entre parenthèses, les derniers événements sur toutes les séquences. On a donc plusieurs états différents qui représentent la même colonne dans l'alignement multiple, par exemple dans ce cas les deux premiers états représentent une insertion dans X^2, mais le premier indique que la dernière position sur X^1 était une conservation et le deuxième indique que la dernière position sur X^1 était une délétion. Sur l'alignement on ne fera pas de différence entre ces deux états, mais pour le calcul des probabilités de transition on a besoin de cette distinction. De plus on établit un ordre pour écrire les insertions, d'abord celles dans X^2 et après celles dans X^1 (pour un arbre général, on commence par les séquences les plus éloignées de la racine, en terme de nombre de nœuds intermédiaires). En effet, puisque les insertions dans les différentes séquences sont des évenements totalement indépendants, l'ordre dans lequel on les écrit entre deux positions homologues de l'alignement n'a aucune importance. La façon la plus simple de le faire est donc d'écrire toutes les insertions d'une même séquence ensembles.

Maintenant, les probabilités de transition entre deux états se calculent facilement à partir de celles du modèle TKF91 pour l'alignement de deux séquences données dans (1.5).

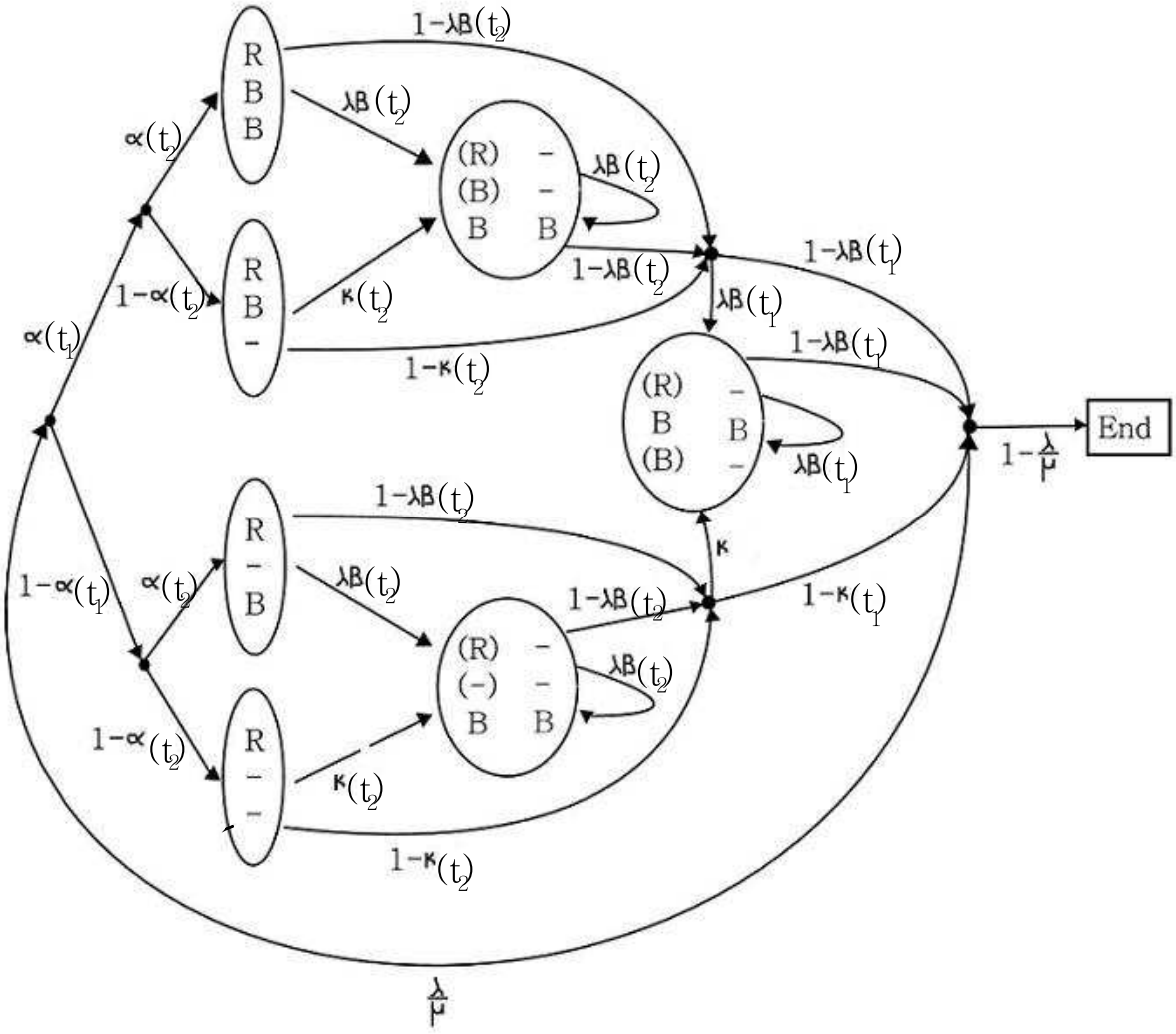


FIG. 1.5 – *Chaîne de Markov de l'alignement sous le modèle TKF91 pour un arbre en étoile avec 2 feuilles. Pour les états correspondant à des insertions, il est donné aussi la représentation de l'état en termes des colonnes de l'alignement multiple.*

Par exemple, la probabilité d'aller de $\begin{matrix}(R)\\(-)\\B\end{matrix}$ à $\begin{matrix}(R)\\B\\(B)\end{matrix}$ est la probabilité de ne plus avoir d'insertions dans X^2, $(1-\lambda\beta(t_2))$ (car si on passe à un état représentant une insertion dans la première séquence c'est parce qu'on a fini d'écrire les insertions dans la deuxième séquence), fois la probabilité d'avoir une insertion dans X^1 après une délétion, $\kappa(t_1)$. Dans la Figure 1.5 on retrouve la représentation graphique de cette chaîne de Markov avec toutes les probabilités de transition entre les états.

Les algorithmes habituels pour les pair-HMMs, notamment l'algorithme de Viterbi et l'algorithme Forward, s'appliquent aussi aux HMM multiples (voir [19]). Cependant ces algorithmes (qui sont les mêmes que ceux utilisés pour l'alignement multiple par *score*) ont une complexité algorithmique de $O(2^k L^k)$ (pour k séquences de longueur moyenne L) ce qui les rend vite inutilisables. Il est donc important de trouver des algorithmes plus efficaces. Quelques travaux récents s'avancent dans ce sens (voir par exemple [53]). Finalement, le grand enjeu de l'alignement multiple est de pouvoir le combiner avec la construction des arbres phylogénétiques. C'est à dire, au lieu d'aligner des séquences à arbre fixé, estimer les alignements et les arbres en même temps. Des travaux sur ce sujet sont ceux de Hein *et al.* [52] et Metzler *et al.* [25].

1.2 Contributions

La suite de cette thèse est composée de trois chapitres constitués d'articles indépendants. C'est pourquoi il sont écrits en anglais et rappellent quelques remarques générales sur l'alignement de séquences déjà évoquées dans cette introduction. Voici une présentation en français de chacun d'eux.

1.2.1 Un modèle d'alignement de séquences avec des taux d'évolution variables selon les sites

Dans le premier chapitre de cette thèse nous nous sommes intéressés à la construction d'un modèle d'évolution des insertions et délétions permettant la variabilité des paramètres d'évolution. Ceci est un travail en collaboration avec Dirk Metzler (bioinformaticien au FB Informatik und Mathematik, J.W. Goethe-Universität, Frankfurt am Main) et Jean-Louis Plouhinec (biologiste au CNRS, Institut de Transgènose, Orléans), soumis à *IEEE Transactions on Computational Biology and Bioinformatics*.

Le cadre biologique de ce travail est l'identification de motifs conservés dans des régions non codantes des séquences d'ADN. Il est connu que les régions non codantes contiennent les signaux de régulation qui contrôlent l'expression génique. Il serait donc intéressant de pouvoir identifier des plages de vitesse d'évolution homogène le long de la séquence (les vitesses d'évolution les plus lentes correspondant à des contraintes de sélection plus importantes et donc, potentiellement, à des signaux de régulation).

Des processus d'évolution avec des paramètres d'évolution variables existaient déjà

mais toujours pour la modélisation des substitutions (voir [23] par exemple). Le point fort de ce travail est d'étendre la variabilité au processus d'insertion-délétion.

Le modèle proposé est un modèle d'insertion et délétion par fragments basé sur le modèle TKF92, avec l'avantage de permettre deux types de comportements évolutifs le long d'une séquence, une évolution *rapide* et une évolution *lente*. Les régions d'évolution rapide sont divisées en fragments qui évoluent selon le modèle TKF92 avec des paramètres $\lambda < \mu$ (les taux d'insertion et délétions) et $\gamma_2 > 1$ (la taille moyenne des fragments). Les régions *lentes* ont une longueur géométrique de moyenne $\gamma_1 < \gamma_2$ et ce sont des régions très conservées qui ne subissent que des substitutions. Les autres paramètres du modèle sont les taux de substitution α_1 pour les régions *lentes* et α_2 pour les régions *rapides* avec $\alpha_1 < \alpha_2$.

L'alignement de deux séquences ayant évolué selon ce modèle est un pair-HMM à quatre états, les états *match*, insertion et délétion pour les régions *rapides* et seulement l'état *match* pour les régions *lentes*. Ceci nous permet d'utiliser l'algorithme Forward pour pair-HMMs pour le calcul de la vraisemblance du modèle et l'échantillonnage d'alignements.

On propose deux approches d'estimation : une approche bayésienne dans laquelle à partir de lois a priori non-informatives sur les paramètres on génère des alignements et des paramètres d'évolution distribués selon la loi jointe a posteriori ; et une approche par maximum de vraisemblance pour l'estimation des paramètres d'évolution. Pour l'estimation bayésienne on utilise une méthode MCMC (Markov Chain Monte Carlo), plus particulièrement l'échantillonnage de Gibbs. La loi des alignements sachant les paramètres est calculée par l'algorithme Forward. On introduit une étape de Metropolis-Hastings à chaque itération pour la génération des paramètres sachant l'alignement, dont la loi n'est pas connue.
Pour l'estimation par maximum de vraisemblance on utilise l'algorithme SAEM [17] qui est une version stochastique de l'algorithme EM [18]. L'algorithme SAEM nécessite aussi à chaque itération la génération d'un nouvel alignement. Ceci est très coûteux algorithmiquement car l'algorithme Forward a une complexité de $O(n \times m)$, où n et m sont les longueurs des séquences observées.
Pour accélérer nos algorithmes d'estimation on propose de réaliser à chaque itération un échantillonnage partiel de l'alignement. Ceci consiste à choisir un bout de l'alignement de l'itération précédente et de re-échantillonner seulement cette partie de l'alignement. La procédure qu'on propose génère une chaîne de Markov uniformément ergodique dans

l'espace des alignements et sa loi stationnaire est la loi des alignements conditionnellement aux valeurs des paramètres. Cette propriété garantit la convergence de nos algorithmes d'estimation (voir [46] pour le SAEM et [67] pour l'échantillonnage de Gibbs).

Nous présentons des applications de notre modèle à des données simulées et réelles. On obtient de bonnes estimations sur les données simulées et des résultats encourageants en ce qui concerne la détection de motifs conservés sur les données réelles.

1.2.2 Estimation paramétrique dans le modèle Markov caché pair

Le chapitre 2 de cette thèse est consacré à l'étude de la consistance des estimateurs bayesiens et par maximum de vraisemblance dans le modèle Markov caché pair. Ceci est un travail en collaboration avec Elisabeth Gassiat et Catherine Matias (chargée de recherche au CNRS, Génopole, Evry) qui a été accepté par *Scandinavian Journal of Statistics*.

Depuis l'introduction des pair-HMMs dans le cadre de l'alignement de deux séquences biologiques, des techniques et des algorithmes d'estimation bayésienne et par maximum de vraisemblance ont été développés pour estimer les paramètres de ces modèles. Cependant, il n'existe aucun résultat théorique sur les propriétés statistiques de ces méthodes d'estimation. C'est pourquoi nous nous sommes intéressées à l'étude de la consistance des estimateurs bayésien et MLE.
Même si les algorithmes des pair-HMMs reposent sur ceux des HMMs classiques, les deux modèles sont très différents d'un point de vue théorique. Dans les pair-HMMs les observations sont composées de deux séquences au lieu d'une seule, mais aussi la longueur du processus caché n'est pas observée. Les propriétés statistiques des estimateurs dans les pair-HMMs doivent donc être établies en dehors du cadre des résultats connus pour les HMMs.

On considère ici une définition du pair-HMM plus générale que celle donnée dans [19]. Soit $\{\varepsilon_t\}_{t\geq 1}$ une chaîne de Markov stationnaire avec espace d'états $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$ et matrice de transition π. Avec cette notation $(1,0)$ correspond à une délétion, $(0,1)$ à une insertion et $(1,1)$ à un *match*. Cette chaîne génère la marche aléatoire $Z_t = \sum_{1\leq s\leq t} \varepsilon_s$ pour $t \geq 0$ et $Z_0 = (0,0)$ à valeurs dans $\mathbb{N}\times\mathbb{N}$. A l'instant t, les deux composantes N_t et M_t de cette marche aléatoire ($Z_t = (N_t, M_t)$) représentent la longueur de chacune des séquences émises par le processus caché jusqu'à l'instant t.

Conditionnellement au chemin caché les observations sont émises de la façon suivante

(voir Figure 1.6) : à l'instant t,

- si $\varepsilon_t = (1,0)$, une variable aléatoire X est émise dans la première séquence selon une loi f sur $\mathcal{A}$;
- si $\varepsilon_t = (0,1)$, une variable aléatoire Y est émise dans la deuxième séquence selon une loi g sur $\mathcal{A}$;
- si $\varepsilon_t = (1,1)$, un couple de variables aléatoires (X,Y) est émis, X dans la première séquence et Y dans la deuxième selon une loi jointe h sur $\mathcal{A} \times \mathcal{A}$.

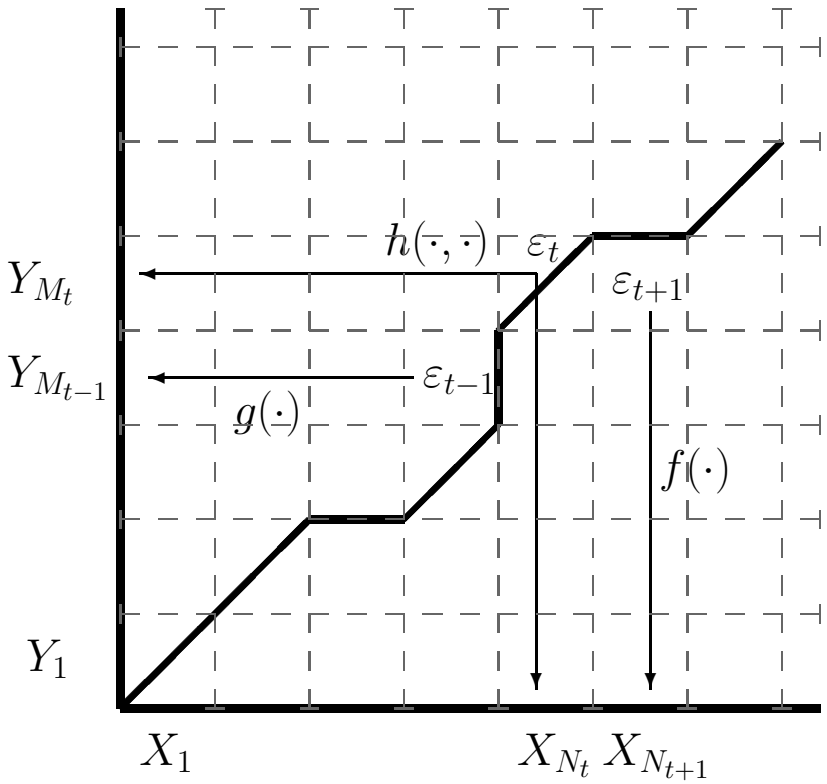


FIG. 1.6 – *Schéma de l'émission des observations dans le modèle pair Markov caché.*

Conditionnellement à la chaîne de Markov cachée toutes les variables émises sont indépendantes. Ainsi on peut écrire

$$\mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|\varepsilon_{1:t}) = \prod_{s=1}^{t} f(X_{N_s})^{\mathbb{1}\{\varepsilon_s=(1,0)\}} g(Y_{M_s})^{\mathbb{1}\{\varepsilon_s=(0,1)\}} h(X_{N_s}, Y_{M_s})^{\mathbb{1}\{\varepsilon_s=(1,1)\}},$$

ou $\theta = (\pi, f, g, h)$ est le paramètre du modèle.
Cette définition du pair-HMM contient, par exemple, le modèle FID de Metzler [55]. Ce

n'est pas le cas pour le modèle TKF91 car la chaîne de Markov sous ce modèle n'est pas stationnaire. Ceci dit, on ne peut pas espérer étudier des propriétés asymptotiques sous le modèle TKF91 car la longueur des séquences observées dépend de la valeur des paramètres (voir le Chapitre 3 pour plus de détails).

La définition de la vraisemblance des observations dans le modèle pair-HMM est ambiguë. On peut en effet considérer comme vraisemblance la quantité calculée par les algorithmes pair-HMM, c'est à dire la somme sur tous les alignements des deux séquences de la probabilité d'observer les deux séquences et un alignement. On appellera cette quantité Q_θ. On peut aussi s'intéresser à la loi jointe des observations émises par le processus caché jusqu'à une longueur déterminée, qu'on notera $\mathbb{P}_\theta$. Les résultats qu'on va obtenir sont valables pour les deux définitions de la vraisemblance.

On se base sur la méthode classique de Wald [78] pour prouver la consistance des estimateurs bayésien et par maximum de vraisemblance dans notre modèle. Celle-ci comporte trois points : prouver la convergence de la log-vraisemblance renormalisée vers une fonction de contraste limite, montrer que ce contraste limite est maximum uniquement pour la vraie valeur du paramètre et enfin établir un critère d'uniformité par rapport au paramètre dans la convergence de ce contraste limite.

Pour le premier point on utilise une version pour des processus sur-additifs du théorème ergodique sous-additif de Kingman [43]. En effet la log-vraisemblance sur notre modèle est un processus sur-additif.

Pour le deuxième point, c'est à dire montrer que le contraste limite est maximum uniquement pour la vraie valeur du paramètre, on obtient des résultats partiels. En effet on arrive à identifier les lois d'émission f et g si elles sont les marginales de h. De même on est capables de distinguer des paramètres qui donnent des *directions principales* de l'alignement différentes (la direction principale étant la droite définie par (0,0) et $\mathbb{E}_\theta[\varepsilon_1]$). Cependant il reste des cas intéressants à couvrir. Notamment lorsqu'on impose que le modèle d'alignement soit réversible dans le temps, et donc que les probabilités d'insertion et délétion soient égales (ce qui implique que la direction principale de l'alignement est la droite (t, t) pour toute valeur du paramètre).

Finalement on s'intéresse à la consistance des estimateurs fondés sur la quantité Q_θ qui est celle qui sert de base à l'estimation dans les algorithmes pair-HMM. Dans les schémas de paramétrisation pour lesquels le contraste limite est maximum uniquement pour la vraie valeur du paramètre, la consistance de l'estimateur du maximum de vraisemblance s'obtient immédiatement avec des arguments classiques sur les M-estimateurs. Pour l'estimateur bayésien, cela n'est pas une conséquence directe puisque notre vraisemblance Q_θ n'est pas la loi jointe des observations. Cependant la preuve de la consistance suit les

idées classiques de la théorie bayésienne.

1.2.3 Estimation paramétrique dans les modèles d'alignement multiple issus du modèle TKF91

Dans le troisième chapitre de cette thèse on s'intéresse à l'étude des propriétés statistiques des estimateurs bayésien et par maximum de vraisemblance dans les modèles d'alignement multiple issus du modèle TKF91.

Dans un premier temps, dans l'idée de généraliser les résultats obtenus pour les pair-HMMs, on s'était intéressé aux HMMs multiples. Cependant cette modélisation des alignements multiples repose sur des artifices (tels que l'introduction d'une *mémoire* dans les états cachés pour que le processus reste markovien, ou l'imposition d'un ordre pour l'apparition des insertions) qui rendent le modèle peu compréhensible. En fait ce qui est important dans le problème de l'alignement multiple de séquences c'est de retrouver les positions homologues. Entre deux positions homologues il va y avoir des insertions, mais l'ordre dans lequel elles se produisent n'a aucune importance. Ainsi il est naturel de s'intéresser à la structure d'homologie (spécification des positions homologues) plutôt qu'à l'alignement en soi.

On se place dans le cas d'un arbre en étoile avec k séquences. On définit la structure d'homologie comme une suite de variables i.i.d. $\{\varepsilon_n\}_{n\geq 1}$ à valeurs dans

$$\mathcal{E} = \{(\delta_{1:k}, a_{1:k}) \,|\, \delta_i \in \{0,1\}, a_i \in \mathbb{N}\ i = 1, \ldots, k\}\,.$$

Le vecteur $\delta_{1:k} = \delta_1, \ldots, \delta_k$ correspond à un caractère dans la séquence ancestrale qui est conservé (1) ou non (0) dans chacune des séquences observées. Le vecteur $a_{1:k}$ correspond au nombre d'insertions (qui peut être zéro) sur chaque séquence entre deux positions homologues. L'indépendance et l'égalité en distribution des $\{\varepsilon_i\}$ découlent de ces deux mêmes propriétés pour l'évolution des *liens* dans le modèle TKF91. La loi des $\{\varepsilon_i\}$ s'écrit facilement en fonction de la loi du nombre de descendants d'un *lien mortel*.

Pour l'étude des propriétés asymptotiques des estimateurs dans ce modèle, on se doit évidemment de disposer de séquences dont la longueur tend vers l'infini, ce qui n'est pas possible sous les hypothèses du modèle TKF91. En effet, dans le modèle TKF91 la longueur des séquences dépend des valeurs des paramètres. Ainsi, si la longueur des séquences tend vers l'infini, le rapport λ/μ doit tendre vers 1. Pour pouvoir faire une étude asymptotique on doit donc se placer dans le cas limite $\lambda = \mu$. Ceci implique que les séquences

observées doivent êtres considérées comme des morceaux de séquences plus longues extraites entre des positions homologues connues.

Sous ces hypothèses, et avec le même schéma que dans le Chapitre 2, on montre la convergence de la log-vraisemblance renormalisée vers une fonction de contraste limite et que ce contraste limite est maximum pour la vraie valeur du paramètre (mais on ne sait pas montrer l'unicité). Finalement on présente des simulations dans lesquelles le contraste limite semble avoir un maximum unique à la vraie valeur du paramètre.

Chapitre 2

Pairwise alignment with an evolution model allowing rate heterogeneity

Abstract

We present a stochastic sequence evolution model to obtain alignments and estimate mutation rates between two homologous sequences. The model allows two possible evolutionary behaviors along a DNA sequence in order to determine conserved regions and take its heterogeneity into account. In our model the sequence is divided into slow and fast evolution regions. The boundaries between these sections are not known and must be estimated. This model induces a pair hidden Markov structure at the level of alignments thus making efficient statistical alignment algorithms possible. We propose two complementary estimation methods, namely a Gibbs sampler for Bayesian estimation and a stochastic version of the EM algorithm for maximum likelihood estimation. Both algorithms involve the simulation of alignments. We propose a partial alignment sampler computationally less expensive than the typical whole alignment sampler. We show the convergence of the two estimation algorithms when used with this partial sampler. Our algorithms provide consistent estimates for the mutation rates and plausible alignments and sequence segmentations on both simulated and real data.

Contents

Ce chapitre correspond, avec plus de details, à un article en collaboration avec Dirk Metzler et Jean-Louis Plouhinec soumis à *IEEE Transactions on Computational Biology and Bioinformatics [4].*

2.1 Introduction

Models for the evolution of biological sequences like DNA, RNA and proteins play an important role in modern methods of sequence analysis, as for example Bayesian or maximum-likelihood based phylogeny reconstruction (cf. [19], [68], [22]). Since their initial introduction by Jukes and Cantor [40] and Dayhoff, Schwartz and Orcutt [16], models for the process of amino acid or nucleotide substitutions have been continuously refined. In order to provide a sound basis of statistical reasoning for sequence alignment algorithms, the availability of explicit models for the process of insertions and deletions (indels) is important. The first indel models for sequence alignment were the TKF91 and TKF92 models introduced by Thorne, Kishino and Felsenstein ([76, 77]). In TKF91 only insertions and deletions of single positions are permited, which is not appropriate for many data sets. The TKF92 model allows insertions and deletions of sequence fragments of geometric length. The parameters of this model are the rates of insertions and deletions and the mean length of the fragments. For given values of these parameters and the substitution rates it is possible to compute the likelihood of the set of parameter values as well as the most probable alignment of two sequences. In [56] and [55] a Markov chain Monte Carlo (MCMC) method for the joint sampling of alignments and mutation parameter values based on the TKF91 model and the fragment-insertion-deletion (FID) model similar to TKF92 model is suggested. The efficiency of the algorithms in all these methods takes advantage of a hidden Markov structure in the models TKF91, TKF92 and FID. For this structure it is crucial that fragments that have once been inserted can not be split by later insertions or deletions. Miklós *et al.* [60] discuss more costly statistical alignment algorithms for a model of insertions and deletions without this fixed-fragments assumption. Simulation studies in [55] provide evidence that parameter estimations based on models assuming fixed-fragments are quite robust against violations of this assumption.

In the previous indel models, it is assumed that the indel rates as well as the typical indel lengths are constant along the sequence. Nevertheless, while we still lack an in-depth understanding of the functional constraints shaping non-coding genomic sequences, it is clear that the evolutionary process is heterogeneous along the sequence : most genomic regions evolve merely neutrally [12, 54], while some are conserved between distant species and evolve under purifying selection [8, 69]. Some of these conserved regions have been shown to consist in compact arrays of transcription factor binding sites termed cis-regulatory modules (CRM), [71]. These CRMs consist in conserved blocks of nucleotides recognized by transcription factors separated by more variable regions allowing indel events.

In line with the current understanding of the evolution of genomic non-coding regions, we propose an indel model of sequence evolution with two different rates of evolution,

namely *slow* evolution rates and *fast* evolution rates. We allow two different speeds of evolution along the sequence. Given two unaligned sequences, the model output is not only an alignment and estimates of the mutation parameters, but also a segmentation of the alignment into slow (conserved) and fast sections. The boundaries between these sections are not a priori known and must therefore also be estimated from the data. This is done by turning the evolution process into a pair Hidden Markov model at the level of sequences where the hidden process is the alignment and the segmentation. The substitution model to go with this indel model allows two different substitution rates, one for each kind of section.

For the estimation of evolution parameters, alignments and segmentations of the sequences we propose two complementary approaches. On the one hand, we follow a maximum likelihood approach that allows us to retrieve the most probable evolution parameter. From the estimated value we can give a reliability measure of any alignment and segmentation. Since the direct maximization of the likelihood over all possible alignments and segmentations is computationally very expensive, and since the EM algorithm, [18], (which is often used to maximize the likelihood in such cases) does not reduce significantly the complexity of this particular problem, we propose to use some stochastic version of EM, such as the SAEM algorithm [17]. The principle of SAEM is to simulate an alignment at each iteration and to maximize the complete likelihood of the observed sequences given that alignment. This algorithm has been proved to have the same convergence properties as the EM algorithm itself.

On the other hand, in a Bayesian framework, we will provide the posterior distribution of alignments, segmentations and evolution rates given the observed sequences. This will be done via the Gibbs sampling algorithm (see [67]). This algorithm consists of simulating reiteratively alignments and segmentations given a parameter value and parameters given an alignment-segmentation of the two sequences. This procedure has been proven to provide values following the joint posterior distribution given the observed sequences.

We prove that the convergence of both Gibbs sampling and SAEM is still guaranteed even if we do not simulate a whole alignment at each iteration but just a part of it, which significantly reduces the complexity of the algorithms.

The results obtained on simulated data confirm the efficiency of both estimation methods. Moreover, their application to the analysis of five vertebrate sequences of the Otx2 locus (1.9-2 kb) provides coherent segmentations into *fast* and *slow* evolution regions.

2.2 The model

The introduced evolution model allows two possible evolution behaviors along a DNA sequence in order to consider its heterogeneity and to determine conserved regions. The sequence is assumed to be divided into slow and fast evolution regions. Slow regions are conserved along the time, i. e. no insertion or deletion can be produced. However, we do allow for nucleotide substitutions to take place in these regions. In fast regions any kind of evolution events (insertions, deletions or substitutions) may occur. To highlight the fact that slow regions stand for almost invariable parts of the sequence, substitutions in slow regions should be much less probable than in fast ones. Then, substitution rates will be $\alpha_1 > 0$ and $\alpha_2 > 0$ respectively, with $\alpha_1 << \alpha_2$. The substitution model to go with the proposed insertion and deletion process should be reversible in time, given the alignment. The indel model can also be transferred to protein sequences and combined with the PAM substitution model [16].

2.2.1 The Insertion and Deletion Process

Our model is a fragment insertion and deletion model, i. e. insertions and deletions of more than one nucleotide at a time are possible. In our model, a DNA sequence is split into a geometrically distributed number ≥ 1 of *stretches*. Each stretch starts with one slow fragment, followed by a finite number ≥ 0 of fast fragments. Each fragment consist of a finite number ≥ 1 of *positions*, with one exception : The slow fragment of the very first stretch is of length 0.

The parameters of this process will be $\gamma_1 \geq 1$ and $\gamma_2 \geq 1$ the expectations of slow and fast fragments lengths, $\lambda > 0$ the insertion rate in fast regions, and $\mu > 0$ the deletion rate in fast regions. We will take $\lambda < \mu$ to avoid sequence lengths increasing to infinity.

We suppose the length of fragments to be geometrically distributed with expectation γ_1 for the slow ones and γ_2 for the fast ones, i. e. the probability that the length of a fragment equals k is $(1 - 1/\gamma_1)^{k-1} 1/\gamma_1$ and $(1 - 1/\gamma_2)^{k-1} 1/\gamma_2$ respectively.

The fast fragments of a stretch evolve under the TKF92 model, and the corresponding slow fragment acts as their immortal link in the sense of [76, 77]. Thus, in the stationary distribution the number of fragments in each stretch is geometrically distributed. This results in the following picture of a DNA sequence :

where B stands for a nucleotide base. Of course, when analyzing data, the subdivision of the sequence into stretches and fragments is not observable.

Each fragment produces new fragments to its right at rate λ. The new fragment is of type "fast", independently of the type of its ancestor, and its length is geometrically distributed with expectation γ_2. Every fast fragment is deleted at rate μ. Under these assumptions and with a deletion rate that exceeds the insertion rate, the stationary distribution of the number of fast fragments in a stretch is geometric on $\{0, 1, 2, \ldots\}$ with expectation $\lambda/(\mu - \lambda)$ (see [76]). Consequently, the number of fragments in a stretch is geometric on $\{1, 2, \ldots\}$ with expectation $\mu/(\mu - \lambda)$. Slow fragments cannot be deleted.

2.2.2 The Markov property of the alignment

Let us consider the bare alignment (the alignment without specification of nucleotides) as a sequence of the following four states $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{S}}$, $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{F}}$, $\overset{-}{\mathtt{B}_\mathrm{F}}$ and $\overset{\mathtt{B}}{-_\mathrm{F}}$, where S stands for sites in slow regions and F for sites in fast regions. States $\mathtt{B}_\mathrm{S}$ and $\overset{\mathtt{B}}{-_\mathrm{S}}$ never appear because we do not allow insertions or deletions in slow regions. We are going to see that the bare alignment is a Markov chain on these states.

Since this property is well-known for the TKF92 model (cf. [77]), we can conclude that transitions between $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{F}}$, $\overset{-}{\mathtt{B}_\mathrm{F}}$ and $\overset{\mathtt{B}}{-_\mathrm{F}}$ are Markovian. We only have to consider transitions going from or into $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{S}}$.

The transition from $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{S}}$ to itself is markovian because the length of slow fragments is geometrically distributed, i. e. the probability that the fragment ends at a particular site is independent of the number of previous sites in the fragment.

Transitions from $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{S}}$ to any state in a fast region do not depend on the precedent states because slow and fast regions behave independently, it only matters that we are at the end of the slow fragment. In fact, the probability of such a transition is simply the product of the probability that a slow fragment ends and the probability that a fast fragment under the TKF92 model starts with the respective state in the fast region.

In the same way, transitions from any state in a fast region to $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{S}}$ do not depend on the precedent states, the only remarkable events being the end of the fragment and the end of the fast region. If we are at the end of the fragment but not at the end of the fast region, we start a new fragment in the same fast region. If we are at the end of the fast region, we start a new slow region. As the length of fast fragments is supposed to be geometric, and the number of fragments in a fast region is also geometric (stationary distribution of the number of normal links in a sequence under the TKF92 model, see [76, 77]) the end of a fast fragment and the end of a fast region in a particular position do not depend on the precedent positions in the sequence. Thus, any transition to the slow state is also Markovian.

The stationary distribution of the alignment length will be a mixture of geometric distribution depending on the length of the initial sequence in each case. However, we will

not consider it here because as in [55], for any given sequence data we will assume that the sequences were cut out of very much longer sequences, which is a realistic biological assumption. This is also the reason why there is no *End* state in our model.

Let us denote $\{\varepsilon_i\}_{i\geq 1}$ the Markov chain of the alignment, whose space state is $\mathcal{E} = \{\overset{\texttt{B}}{\texttt{B}_\texttt{S}}, \overset{\texttt{B}}{\texttt{B}_\texttt{F}}, \overset{\texttt{-}}{\texttt{B}_\texttt{F}}, \overset{\texttt{B}}{\texttt{-}_\texttt{F}}\}$. The transition probabilities between the three *fast* states are those of the TKF92 model. We recall (see [76]) the distribution $p_n^H(t)$ of the number of descendants from a surviving fast fragment, including the fragment itself, after a time t, the distribution $p_n^N(t)$ of the number of descendants from a deleted fast fragment after a time t, and the distribution $p_n^I(t)$ of the number of descendants from a slow fragment, including itself, after a time t, always under the assumptions of the birth and death process :

$$\begin{aligned} p_n^H(t) &= e^{-\mu t}[1-\lambda\beta(t)][\lambda\beta(t)]^{n-1} \; n>0 \\ p_n^N(t) &= [1-e^{-\mu t}-\mu\beta(t)][1-\lambda\beta(t)][\lambda\beta(t)]^{n-1} \; n>1; \quad p_0^N(t)=\mu\beta(t) \\ p_n^I(t) &= [1-\lambda\beta(t)][\lambda\beta(t)]^{n-1} \; n>0 \end{aligned} \tag{2.1}$$

where $\beta(t) = \frac{1-e^{(\lambda-\mu)t}}{\mu-\lambda e^{(\lambda-\mu)t}}$.

Because of time scaling we can set the time distance between the two sequences to 1 and so t not appear in the following. From (2.1) we have the probabilities of all the transitions inside a stretch of fast fragments. For instance, the probability of getting $\overset{\texttt{-}}{\texttt{B}_\texttt{F}}$ from $\overset{\texttt{B}}{\texttt{-}_\texttt{F}}$ is the probability of being at the end of a fast fragment, multiplied with the probability that a fast fragment, given that it dies, has at least one surviving descendent, so we obtain

$$P_{\lambda,\mu,\gamma_1,\gamma_2}(\varepsilon_{n+1} = \overset{\texttt{-}}{\texttt{B}_\texttt{F}} \,|\varepsilon_n = \overset{\texttt{B}}{\texttt{-}_\texttt{F}}) = \frac{\sum_{n=1}^{\infty} p_n^N(t)}{\gamma_2\cdot(1-e^{-\mu})} = \frac{1-e^{-\mu}-\mu\beta}{\gamma_2\cdot(1-e^{-\mu})}.$$

We recall that the distribution of the number of fast fragments in a stretch is geometric on $\{0,1,2,\ldots\}$, so it is possible to have stretches with no fast fragments.

From the previous considerations we can now compute the transition probabilities involving slow fragments. Remaining in state $\overset{\texttt{B}}{\texttt{B}_\texttt{S}}$ can occur when we stay inside a slow fragment or when we pass from one slow fragment into another one through an empty stretch of fragments. So $P_{\lambda,\mu,\gamma_1,\gamma_2}(\varepsilon_{n+1} = \overset{\texttt{B}}{\texttt{B}_\texttt{S}} \,|\varepsilon_n = \overset{\texttt{B}}{\texttt{B}_\texttt{S}})$ is the probability that a slow fragment does not finish, $1-\frac{1}{\gamma_1}$, or that it finishes, $\frac{1}{\gamma_1}$, but without descendants, $1-\lambda\beta$, and inside a stretch with no fast fragments, $1-\frac{\lambda}{\mu}$.
Transition probabilities from any fast state to a slow fragment are the probabilities of being at the end of the last fast fragment of a stretch at this particular state. For instance, the probability of going from a match in a fast fragment to a slow fragment is the probability of being at the last fast fragment of the stretch, $1-\frac{\lambda}{\mu}$, at the end of this fragment, $\frac{1}{\gamma_2}$, and without any descendants, $1-\lambda\beta$. The probability of getting $\overset{\texttt{B}}{\texttt{B}_\texttt{S}}$ from $\overset{\texttt{B}}{\texttt{-}_\texttt{F}}$ is again the

probability of being at the last fast fragment of the stretch, $1-\frac{\lambda}{\mu}$, at the end of this fragment, $\frac{1}{\gamma_2}$, and now the probability that a fragment which has died has no survivors, $\frac{\mu\beta}{1-e^{-\mu}}$.
Finally, the probability of getting any fast state from $\mathtt{B}_{\mathtt{S}}$ is the probability that a slow fragment ends, $\frac{1}{\gamma_1}$, multiplied with the probability of starting a fast fragment with the given fast state. If this fast state is, for instance, $\bar{\mathtt{B}}_{\mathtt{F}}$, this last one would be $\lambda\beta$, the probability that a slow fragment has at least one survivor.

The whole transition matrix is the following :

$$\Big(\pi_{\lambda,\mu,\gamma_1,\gamma_2}(i,j) = P_{\lambda,\mu,\gamma_1,\gamma_2}(\varepsilon_{n+1}=j|\varepsilon_n=i)\Big)_{i,j\in\mathcal{E}} =$$

$$\begin{array}{c} \\ \mathtt{B}_{\mathtt{S}} \\ \mathtt{B}_{\mathtt{F}} \\ \bar{\mathtt{B}}_{\mathtt{F}} \\ \mathtt{B}_{\mathtt{S}} \\ \mathtt{B}_{\mathtt{F}} \\ \bar{\mathtt{B}}_{\mathtt{F}} \end{array}
\begin{array}{c}
\begin{array}{cccc} \mathtt{B}_{\mathtt{S}} & \mathtt{B}_{\mathtt{F}} & \bar{\mathtt{B}}_{\mathtt{F}} & \mathtt{B}_{\mathtt{S}} \end{array} \\
\left(\begin{array}{cccc}
\frac{1}{\gamma_1}(1-\lambda\beta)(1-\frac{\lambda}{\mu})+1-\frac{1}{\gamma_1} & \frac{1}{\gamma_1}(1-\lambda\beta)\frac{\lambda}{\mu}e^{-\mu} & \frac{1}{\gamma_1}(1-\lambda\beta)\frac{\lambda}{\mu}(1-e^{-\mu}) & \frac{1}{\gamma_1}\lambda\beta \\
\frac{1}{\gamma_2}(1-\lambda\beta)(1-\frac{\lambda}{\mu}) & \frac{1}{\gamma_2}(1-\lambda\beta)\frac{\lambda}{\mu}e^{-\mu}+1-\frac{1}{\gamma_2} & \frac{1}{\gamma_2}(1-\lambda\beta)\frac{\lambda}{\mu}(1-e^{-\mu}) & \frac{1}{\gamma_2}\lambda\beta \\
\frac{1}{\gamma_2}\frac{\mu-\lambda}{1-e^{-\mu}}\beta & \frac{1}{\gamma_2}\lambda\beta\frac{e^{-\mu}}{1-e^{-\mu}} & \frac{1}{\gamma_2}\lambda\beta+1-\frac{1}{\gamma_2} & \frac{1}{\gamma_2}\frac{1-e^{-\mu}-\mu\beta}{1-e^{-\mu}} \\
\frac{1}{\gamma_2}(1-\lambda\beta)(1-\frac{\lambda}{\mu}) & \frac{1}{\gamma_2}(1-\lambda\beta)\frac{\lambda}{\mu}e^{-\mu} & \frac{1}{\gamma_2}(1-\lambda\beta)\frac{\lambda}{\mu}(1-e^{-\mu}) & \frac{1}{\gamma_2}\lambda\beta+1-\frac{1}{\gamma_2}
\end{array}\right)
\end{array}
\tag{2.2}$$

2.2.3 Reversibility of the homology structure

This insertion-deletion model is time reversible on the homology structure (see [55]). We do not have the reversibility on the alignment because of the TKF convention (births always happen to the right of a link). However, the homology structure is sufficient because it gives us the number of inserted and deleted nucleotides between two homologous sites.

In our model insertions and deletions between homologous sites only take place in fast regions, which behave under the TKF92 model. As the latter fulfils the reversibility of the homology structure, the property translates immediately to our model.

2.3 Algorithms

We present two approaches to analyze DNA sequences with our model. The first one aims at a point estimation of the evolution parameters, which is achieved using a maximum

likelihood approach. We propose to use a stochastic version of the EM algorithm, namely the SAEM algorithm, [17], which is an efficient method to compute it. As the main point of our model is to provide alignments and segmentations into slow and fast evolution regions of the sequences, one should consider to give some optimal alignment (in the following *alignment* will denote the whole Markov chain, that is the alignment and the segmentation) for the estimated parameters. However, it is known that optimal alignments (as the most probable *a posteriori* alignment given by the Viterbi algorithm for instance, see [19]) look different from typical ones and giving a unique alignment would be very arbitrary. For this reason we propose to study, for the ML estimation, the probability distribution of states over each pair of nucleotides from the observed sequences, which give us a *reliability measure* for any alignment of the two sequences.

The second approach should be considered in a Bayesian framework. Its aim is to provide a sample of plausible alignments and evolution rates from the observed sequences, rather than a single alignment or parameter value, (see [56] and [55]). This is done with an MCMC strategy, sampling alignments from parameters and vice versa.

2.3.1 The likelihood

Let us denote $\theta = (\lambda, \mu, \gamma_1, \gamma_2, \alpha_1, \alpha_2)$ the vector of evolution parameters describing our model. Let $h(\cdot)$ be the emission probability of a nucleotide (in practice it will simply be the equilibrium probability of each nucleotide and it will not depend on θ) and for a substitution rate $\alpha > 0$ let $g_\alpha(\cdot,\cdot)$ be the emission probability of a pair of nucleotides, that is the chosen substitution function. Any substitution function can be combined with our model, with the only condition of being symmetric to maintain the time reversibility.

The model described in Section 2 provides a pair-HMM structure at the level of alignments and observed sequences. This means that the hidden alignment $\{\varepsilon_i\}_{i\geq 1}$ is a Markov chain and conditionally to each ε_i the emission of the corresponding nucleotide or pair of nucleotides on the observed sequences is independent of the rest of the positions on the alignment and of the rest of the observed nucleotides. The emissions are produced in the following way. If $\varepsilon_i = \overset{\mathtt{B}}{\underset{}{\mathtt{-}_\mathtt{F}}}$ then a nucleotide is emitted on the first sequence according to h; if $\varepsilon_i = \overset{\mathtt{-}}{\mathtt{B}_\mathtt{F}}$ then a nucleotide is emitted on the second sequence according to h; if $\varepsilon_i = \overset{\mathtt{B}}{\mathtt{B}_\mathtt{S}}$ a pair of nucleotides (one on each sequence) is emitted according to g_{α_1}; finally, if $\varepsilon_i = \overset{\mathtt{B}}{\mathtt{B}_\mathtt{F}}$ a pair of nucleotides is emitted according to g_{α_2}.

Let $\mathcal{E}(n,m)$ be the set of all possible alignments of sequences with n and m nucleotides in length. Let e be an alignment in $\mathcal{E}(n,m)$. For each index i such that e_i emits a nucleotide on the first (resp. the second) sequence, let $n_i = n_i(e)$ (resp. $m_i = m_i(e)$) be the corresponding position on that sequence. For instance, for the alignment $e = \left(\overset{\mathtt{B}}{\mathtt{B}_\mathtt{S}}, \overset{\mathtt{-}}{\mathtt{B}_\mathtt{F}}, \overset{\mathtt{-}}{\mathtt{B}_\mathtt{F}}, \overset{\mathtt{B}}{\mathtt{-}_\mathtt{F}}, \overset{\mathtt{B}}{\mathtt{B}_\mathtt{F}}\right)$, we have $n_1 = 1, n_4 = 2, n_5 = 3$ and $m_1 = 1, m_2 = 2, m_3 = 3, m_5 = 4$.

The conditional distribution of observed sequences $x_{1:n}$ and $y_{1:m}$ given an alignment thus writes

$$\mathbb{P}_\theta(x_{1:n}, y_{1:m}|\varepsilon_{1:|e|} = e) = \prod_{i=1}^{|e|} h(x_{n_i})^{\mathbb{1}\{e_i=\overset{\texttt{B}}{\texttt{-}_\texttt{F}}\}} h(y_{m_i})^{\mathbb{1}\{e_i=\overset{\texttt{-}}{\texttt{B}_\texttt{F}}\}} g_{\alpha_1}(x_{n_i}, y_{m_i})^{\mathbb{1}\{e_i=\overset{\texttt{B}}{\texttt{B}_\texttt{S}}\}} g_{\alpha_2}(x_{n_i}, y_{m_i})^{\mathbb{1}\{e_i=\overset{\texttt{B}}{\texttt{B}_\texttt{F}}\}} \quad (2.3)$$

where $\mathbb{1}\{\cdot\}$ stands for the indicator function. Moreover, the complete distribution (or complete likelihood) of the observed sequences and an alignment is given by

$$\mathbb{P}_\theta(\varepsilon_{1:|e|} = e, x_{1:n}, y_{1:m}) = \left\{ \mu_\theta(e_1) \prod_{i=2}^{|e|} \pi_\theta(e_{i-1}, e_i) \right\} \mathbb{P}_\theta(x_{1:n}, y_{1:m}|\varepsilon_{1:|e|} = e) \quad (2.4)$$

where $\mu_\theta(\cdot)$ is the stationary distribution of the Markov chain. This complete likelihood can be written in the classical exponential parametrisation for discrete Markov Chains :

$$\mathbb{P}_\theta(\varepsilon_{1:|e|} = e, x_{1:n}, y_{1:m}) = \exp\{-\Psi(\theta) + \langle \tilde{S}(e, x_{1:n}, y_{1:m}), \Phi(\theta) \rangle\}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product. This is the expression of the complete likelihood that we will use in the following. In our model $\Phi(\theta)$ is the vector whose components are the logarithm of all initial, transition and emission probabilities and the components of $\tilde{S}(e, x_{1:n}, y_{1:m})$ are the frequencies of these events in the given alignment e. There are no terms in the complete likelihood involving only the parameter values so $\Psi(\theta)$ will not appear in our case.

Finally, the likelihood of the observed sequences $x_{1:n}$ and $y_{1:m}$ is just the sum over all possible hidden alignments of the complete likelihood, namely

$$L_\theta(x_{1:n}, y_{1:m}) = \sum_{e \in \mathcal{E}(n,m)} \mathbb{P}_\theta(\varepsilon_{1:|e|} = e, x_{1:n}, y_{1:m}). \quad (2.5)$$

2.3.2 ML-estimation of parameters

The ML-estimation of θ for observed sequences $x_{1:n}$ and $y_{1:m}$ is

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L_\theta(x_{1:n}, y_{1:m}). \quad (2.6)$$

This likelihood can be computed recursively by the Forward algorithm for pair-HMMs (see [19]) and then maximized numerically as in the original TKF91 paper [76]. However, this is quite expensive for relatively long sequences because the maximization algorithm would need several evaluations of the likelihood and each one of them requires a number of iterations similar to the square of one of the sequences length.

Therefore, one should use some version of the EM algorithm, [18], which maximizes the likelihood in missing data models (indeed, we observe the sequences but not the alignment). The principle of EM is to maximize at each iteration the conditional expectation of the complete likelihood (the likelihood of the observations and a realization of the missing data) given a value of θ, which is shown to be equivalent to the maximization of the likelihood. However, this expectation needs again to be computed over all possible alignments and since in a pair-HMM the length of the hidden process (the alignment) is unknown, the maximization of this quantity is not explicit and we would have to use a numerical procedure as complex as the one for the maximization of the likelihood.

For this reason we propose to use some stochastic version of the EM algorithm to maximize the likelihood. This allows us to replace the computation of the conditional expectation by the much simpler of the complete likelihood given an alignment. Indeed, at each iteration we will sample an alignment from $\mathbb{P}_\theta(\cdot|\, x_{1:n}, y_{1:m})$, the conditional distribution of alignments given the observed sequences. This is done by the Forward algorithm with backwards sampling. It consists on computing the probability $P(i,j,k)$ of aligning $x_{1:i}$ to $y_{1:j}$ with final state k, for each $i = 1,\dots,n$, $j = 1,\dots,m$ and $k \in \mathcal{E}$ (Forward algorithm) and then tracing back through the matrix P (see [19] for details). We then maximize on θ the complete likelihood given that alignment.

There are many stochastic versions of the EM algorithm. We have chosen to use the SAEM algorithm (stochastic approximation version of EM, proposed by Delyon, Lavielle and Moulines [17]) which has good convergence properties. It is defined as follows.

SAEM algorithm :

---

1. Simulation of the alignment e^{k+1} from $\mathbb{P}_{\theta^k}(\cdot|\, x_{1:n}, y_{1:m})$

2. $s^{k+1} = s^k + \tau_k(\tilde{S}(e^{k+1}, x_{1:n}, y_{1:m}) - s^k)$

3. $\theta^{k+1} = \arg\max_\theta \langle s^{k+1}, \Phi(\theta)\rangle$

---

where τ_k is a sequence decreasing to zero (for instance $\tau_k = 1$ if $k \leq k_0$ and $\tau_k = \frac{1}{k-k_0}$ if $k > k_0$ for some number of iterations k_0) and $\tilde{S}(\cdot,\cdot,\cdot)$ and $\Phi(\cdot)$ are the ones described in the precedent section. It has been shown by Delyon *et al.* that under mild regularity conditions (satisfied by our model) the sequence (θ^k) converges almost surely to a local maximum of the likelihood L_θ (cf. [17]).

Like in the EM algorithm, the maximization step in the SAEM algorithm need to be performed numerically, but now the evaluation of the objective function is straightforward. In this case (maximization of the complete likelihood given a realization of the hidden Markov chain), one could think of an explicit maximization of the indel parameters

$(\lambda, \mu, \gamma_1, \gamma_2)$ by simply considering the empiric estimators of the transition probabilities for the given alignment. However, this is not possible since the transition matrix of our model (2.2) is not a general stochastic matrix and also linear restrictions on the parameters need to be considered.

2.3.3 MCMC sampling of parameters and alignments

On the aim of producing a distribution of parameter values and alignments one can consider non-informative priors on the parameters and apply an MCMC strategy to simulate the joint posterior law of alignments and parameter values given the observed sequences.

We will apply the idea of the Gibbs sampling (see [67]) by sampling reiteratively parameters and alignments from their posterior marginal distributions. After a large enough number of iterations the sampled parameters and alignments are distributed according to the joint posterior law.

As in the first step of SAEM, we sample an alignment from the distribution of alignments given the observed sequences and a parameter value θ, via the Forward algorithm with backwards sampling.

Given a prior distribution $\pi(\cdot)$ on Θ, the space of parameter values, the posterior distribution conditioned to observed sequences $x_{1:n}$ and $y_{1:m}$ and alignment e is

$$\pi(\theta \,|\, e, x_{1:n}, y_{1:m}) = \frac{\mathbb{P}(\varepsilon_{1:|e|} = e, x_{1:n}, y_{1:m} \,|\, \theta)\pi(\theta)}{\int_\Theta \mathbb{P}(\varepsilon_{1:|e|} = e, x_{1:n}, y_{1:m} \,|\, \theta')\pi(\theta')d\theta'}$$

where $\mathbb{P}(\varepsilon_{1:|e|} = e, x_{1:n}, y_{1:m} \,|\, \theta)$ is just $\mathbb{P}_\theta(\varepsilon_{1:|e|} = e, x_{1:n}, y_{1:m})$ in Bayesian notation. As we can not compute this posterior density directly we use the Metropolis-Hastings sampling (see [67]) to produce values from it. The whole sampling algorithm is defined as follows.

Gibbs sampling algorithm with Metropolis-Hastings step :

---

1. Alignment sampling : Simulation of e^{k+1} from $\mathbb{P}_{\theta^k}(\cdot \,|\, x_{1:n}, y_{1:m})$

2. Parameter sampling : Simulation of θ^{k+1} from $\pi(\cdot \,|\, e^{k+1}, x_{1:n}, y_{1:m})$ via a Metropolis-Hastings step :

2.1 $\tilde{\theta} \sim q(\cdot|\theta^k)$ the proposal law

2.2 $\theta^{k+1} = \begin{cases} \tilde{\theta} & \text{with probability } \rho \\ \theta^k & \text{with probability } 1-\rho \end{cases}$; $\rho = \min\left\{1, \frac{\mathbb{P}(e^{k+1}, x_{1:n}, y_{1:m} \,|\, \tilde{\theta})\pi(\tilde{\theta})}{\mathbb{P}(e^{k+1}, x_{1:n}, y_{1:m} \,|\, \theta^k)\pi(\theta^k)} \frac{q(\theta^k|\tilde{\theta})}{q(\tilde{\theta}|\theta^k)}\right\}$

This kind of algorithm is known as *hybrid version of Gibbs sampling*. Note that at point

2 of each iteration the Metropolis-Hastings step is performed only once : we do not need to accurately approximate the marginal $\pi(\cdot \mid e, x_{1:n}, y_{1:m})$ but just to provide a simulation from the proposal law. The resulting hybrid algorithm is valid, that is, its stationnary distribution is the joint law of parameters and alignments, as soon as the Metropolis-Hasting algorithm is valid itself (for a given alignment e its stationnary distribution is the marginal $\pi(\cdot \mid e, x_{1:n}, y_{1:m})$). This condition is fulfilled if, for instance, we chose the proposal law q to be positive over the support of $\pi(\cdot \mid e, x_{1:n}, y_{1:m})$ (see [67] for the details).

2.3.4 Alignment sampling

In both algorithms SAEM and Gibbs sampling we have to simulate alignments from their conditional distribution given the observed sequences and a value of the evolution parameters. As already said, this sampling step is accomplished via the Forward algorithm with backwards sampling. This Forward algorithm needs a number of iterations proportional to the product of the sequences lengths. This makes alignment sampling step very expensive in terms of computing time for long sequences. Therefore, we will use for the alignment sampling the method proposed by Metzler in [55], which is far less computer time demanding. We will show that when using it within SAEM and Gibbs sampling the convergence of both algorithms is still guaranteed.

The principle of our sampling strategy is not to resample the whole alignment but just a little piece of it at each iteration of the algorithms. However, to guarantee that the iteration of these *partial resamplings* is equivalent to the resampling of the whole alignment, the choice of the subalignment to resample is to be done as described below.

Let $n_0 \in \{1, \dots, n\}$ be fixed. Let ε^k denote the random alignment in iteration k of the algorithm. In iteration k, instead of resampling the whole alignment, we will choose a subalignment of ε^{k-1} containing n_0 nucleotides of the first sequence, x, and we will resample it from the law of alignments given the observed sequences. This procedure produce a Markov chain (ε^k) on the set of all alignments of two observed sequences. We want to show that the stationary law of this chain is exactly the distribution of alignments given the observed sequences.

For any alignment e we will note $e_{(i)_x}$, $i = 1, \dots, n$ the position on the alignment corresponding to position i on the first sequence. For $l = 1, \dots, n - n_0 + 1$ let $e[l]$ be the subalignment $e_{(l-1)_x+1:(l+n_0)_x-1}$ and $y_e[l]$ the subsequence of y aligned to $x[l] = x_{l:l+n_0-1}$ by $e[l]$. We will also note $e_{[l]_-} = e_{(l-1)_x}$ and $e_{[l]_+} = e_{(l+n_0)_x}$. The same notations will apply to the random alignments. Let us show an example. If $n_0 = 4$,

$$x = AACGCTC \quad y = ATTGTT \quad e = \begin{array}{cccccccccc} A & A & | & - & - & C & G & C & T & - & | & C \\ A_S & -_F & | & T_F & T_F & -_F & G_S & -_F & T_S & T_F & | & -_F \end{array},$$

and we choose $l = 3$, then

$$x[l] = CGCT \qquad e[l] = \begin{array}{cccccccc} - & - & C & G & C & T & - \\ T_F & T_F & -_F & G_S & -_F & T_S & T_F \end{array} \qquad \text{and} \qquad y_e[l] = TTGTT.$$

For $l = 1, \ldots, n - n_0 + 1$ let

$$\mathcal{R}_l(e^k) = \left\{ e^{k-1} \text{ alig. of } x_{1:n}, y_{1:m} \quad | \quad e^{k-1}_{1:[l]_-} = e^k_{1:[l]_-} \text{ and } e^{k-1}_{[l]_+:|e^{k-1}|} = e^k_{[l]_+:|e^k|} \right\}$$

be the set of all alignments *related at position l to e^k* and

$$\mathcal{R}(e^k) = \bigcup_l \mathcal{R}_l(e^k).$$

Now, we can write the alignment resampling procedure as follows. At iteration k of the algorithms we will generate the new alignment e^k from the transition probability

$$\Pi_\theta \left(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}, \varepsilon^k_{1:|e^k|} = e^k \;\middle|\; x_{1:n}, y_{1:m} \right) = \sum_{l; e^{k-1} \in \mathcal{R}_l(e^k)} p_l \times \mathbb{P}_\theta \left(\varepsilon^k[l] = e^k[l] \;\middle|\; x[l], y_{e^{k-1}}[l], e^{k-1}_{[l]_-}, e^{k-1}_{[l]_+} \right)$$

if $e^{k-1} \in \mathcal{R}(e^k)$ and 0 elsewhere, where $p_l = \frac{1}{n-n_0+1}$. This means, given $\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}$ we choose l uniformly from $\{0, 1, \ldots, n+n_0-1\}$, we set $e^k_{1:[l]_-} = e^{k-1}_{1:[l]_-}$ and $e^k_{[l]_+:|e^{k-1}|} = e^{k-1}_{[l]_+:|e^k|}$ and we sample $e^k[l]$ from $\mathbb{P}_\theta(\cdot|x[l], y_{e^{k-1}}[l])$. We recall that for $\theta \in \Theta$, $\mathbb{P}_\theta(\cdot|x_{1:n}, y_{1:m})$ is the conditional distribution of alignments given the observed sequences. Here, as we do not compute the probability of the whole alignment, the probability of the new subalignment is also conditioned by its neighbouring positions (we recall that the alignment is a Markov chain). To make the formulae easier to read, we will write $e^{k-1}_{[l]_-}, e^{k-1}_{[l]_+}$ instead of $\varepsilon^{k-1}_{[l]_-} = e^{k-1}_{[l]_-}, \varepsilon^{k-1}_{[l]_+} = e^{k-1}_{[l]_+}$.

Remark 1 *If $e^{k-1} \in \mathcal{R}_l(e^k)$ then $y_{e^{k-1}}[l] = y_{e^k}[l]$.*
This fact is necessary in the proof of Proposition 1. This is why, given $x[l]$, it is important to choose the subalignment to resample in the way described before. Let us illustrate this point.
Suppose that the alignment e^k at iteration k of our algorithm is the alignment e given in the precedent example. Suppose that we had decided to chose the subalignment to resample to be $e_{(l)_x:(l+n_0-1)_x}$ instead of $e_{(l-1)_x+1:(l+n_0)_x-1}$, that is to take the subalignment as exactly corresponding to the chosen subsequence of x instead of extending it to the contiguous insertion positions. If $n_0 = 4$ and $l = 3$, then e^k could have been sampled, for instance, from any of the two alignments

$$e^{k-1,1} = \begin{array}{cccccccccc} A & A & | & C & - & G & C & - & T & | & C \\ A_S & -_F & | & T_F & T_F & G_S & -_F & T_S & T_F & | & -_F \end{array}$$

$$e^{k-1,2} = \begin{array}{cccc|cccc|c} A & A & - & - & C & G & C & T & C \\ A_S & -_F & T_F & T_F & -_F & G_S & T_S & T_F & -_F \end{array}$$

that is, we could have different subsequences of y in the resampling region. In our set up, for any e^{k-1} leading to e^k by resampling at position l, we always have the same subsequences of x and y in the resampling region. This implies that the sum over all $e^{k-1} \in \mathcal{R}_l(e^k)$ is just the sum over all possible alignments of these two subsequences.

The transition matrix Π_θ defines a discrete Markov chain $\{\varepsilon^k\}_{k\geq 1}$ on the set of all alignments of sequences n and m nucleotides in length, $\mathcal{E}(n,m)$. When we use this distribution to sample alignments, our algorithms based on SAEM and Gibbs sampling write :

Our SAEM algorithm :

---

1. Simulation of e^{k+1} from $\Pi_{\theta^k}\left(e^k, \cdot \,\middle|\, x_{1:n}, y_{1:m}\right)$

2. $s^{k+1} = s^k + \tau_k(\tilde{S}(e^{k+1}, x_{1:n}, y_{1:m}) - s^k)$

3. $\theta^{k+1} = \arg\max_\theta \langle s^{k+1}, \Phi(\theta)\rangle$

Our Gibbs sampling algorithm :

---

1. Simulation of e^{k+1} from $\Pi_{\theta^k}\left(e^k, \cdot \,\middle|\, x_{1:n}, y_{1:m}\right)$

2. Simulation of θ^{k+1} from $\pi(\cdot \,|\, e^{k+1}, x_{1:n}, y_{1:m})$ via a Metropolis-Hastings step :

2.1 $\tilde{\theta} \sim q(\cdot|\theta^k)$ the proposal law

2.2 $\theta^{k+1} = \begin{cases} \tilde{\theta} & \text{with probability } \rho \\ \theta^k & \text{with probability } 1-\rho \end{cases}$; $\rho = \min\left\{1, \frac{\mathbb{P}(e^{k+1}, x_{1:n}, y_{1:m} \,|\, \tilde{\theta})\pi(\tilde{\theta})}{\mathbb{P}(e^{k+1}, x_{1:n}, y_{1:m} \,|\, \theta^k)\pi(\theta^k)} \frac{q(\theta^k|\tilde{\theta})}{q(\tilde{\theta}|\theta^k)}\right\}$.

We will now show the validity of both algorithms, i. e. that the values generated by the SAEM procedure converge towards a local maximum of the likelihood and that the Gibbs sampling procedure has as stationary law the joint posterior distribution of alignments and parameter values given the observed sequences.
For the first point, Kuhn et Lavielle [46], have established the convergence of SAEM when in the generation step of the missing data we replace the conditional law given the observed data by a transition probability that generates an uniformly ergodic chain with invariant probability this conditional distribution.
For the second point, namely the Gibbs sampling, the substitution of the generation from $\mathbb{P}_\theta(\cdot|x_{1:n}, y_{1:m})$ by a step of simulation from the transition kernel Π_θ is a new hybrid Gibbs

sampler. As we have already seen for the introduction of Metropolis-Hastings steps, this algorithm is valid as soon as the transition kernel Π_θ has $\mathbb{P}_\theta(\cdot|x_{1:n}, y_{1:m})$ as stationary law.

Then, the following result establishes the validity of the proposed algorithms. Note that in both cases we only simulate *one* alignment at each iteration of the algorithms (we do not need to approximate $\mathbb{P}_\theta(\cdot|x_{1:n}, y_{1:m})$ by generating long chains).

Proposition 1 *The Markov chain $\{\varepsilon^k\}_{k\geq 1}$ is uniformly ergodic and its invariant distribution is $\mathbb{P}_\theta(\cdot|x_{1:n}, y_{1:m})$.*

Proof.
Even if not all transitions between alignments are allowed, we can reach any alignment from any other in a finite number of steps, i. e. the Markov chain $\{\varepsilon^k\}_{k\geq 1}$ is irreducible. Indeed, the maximum number of iterations required for getting an alignment from any other is $[n/n_0] + 1$, where $[\cdot]$ stands for the integer part. This is true because the way in which we choose subalignments allows *shifting* sequence y through sequence x and so after at most $[n/n_0]+1$ steps we can have any y_j aligned to any x_i. Since $\{\varepsilon^k\}_{k\geq 1}$ is a finite state Markov chain, being irreducible implies that it is also uniformly ergodic. So we only have to show that $\mathbb{P}_\theta(\cdot|x_{1:n}, y_{1:m})$ is invariant for the transition probability $\Pi_\theta\left(\cdot, \cdot \,\middle|\, x_{1:n}, y_{1:m}\right)$. We have

$$\sum_{e^{k-1}\in\mathcal{E}(n,m)} \mathbb{P}_\theta(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}|x_{1:n}, y_{1:m}) \times \Pi_\theta\left(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}, \varepsilon^{k}_{1:|e^{k}|} = e^{k} \,\middle|\, x_{1:n}, y_{1:m}\right) =$$

$$\sum_{e^{k-1}\in\mathcal{R}(e^k)} \mathbb{P}_\theta(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}|x_{1:n}, y_{1:m}) \sum_{l; e^{k-1}\in\mathcal{R}_l(e^k)} p_l \times \mathbb{P}_\theta\left(\varepsilon^k[l] = e^k[l] \,\middle|\, x[l], y_{e^{k-1}}[l], e^{k-1}_{[l]_-}, e^{k-1}_{[l]_+}\right) =_{(a)}$$

$$\sum_{l=1}^{n-n_0+1} p_l \times \mathbb{P}_\theta\left(\varepsilon^k[l] = e^k[l] \,\middle|\, x[l], y_{e^k}[l], e^{k}_{[l]_-}, e^{k}_{[l]_+}\right) \sum_{e^{k-1}\in\mathcal{R}_l(e^k)} \mathbb{P}_\theta(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}|x_{1:n}, y_{1:m})$$

where (a) comes from Remark 1. Now, from the pair-HMM structure, $\{\varepsilon^k_i\}_{i\geq 1}$ is again a Markov chain and the observed sequences are conditionally independent of it, and since for $e^{k-1}\in\mathcal{R}_l(e^k)$, e^{k-1} and e^k are equal out of $[l]$, we can rearrange the terms in $\mathbb{P}_\theta(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}|x_{1:n}, y_{1:m})$ and $\mathbb{P}_\theta(\varepsilon^k[l] = e^k[l] \,|\, x[l], y_{e^k}[l], e^k_{[l]_-}, e^k_{[l]_+})$ to write

$$\sum_{e^{k-1}\in\mathcal{E}(n,m)} \mathbb{P}_\theta(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}|x_{1:n}, y_{1:m}) \times \Pi_\theta\left(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}, \varepsilon^{k}_{1:|e^{k}|} = e^{k} \,\middle|\, x_{1:n}, y_{1:m}\right) =$$

$$\sum_{l=1}^{n-n_0+1} p_l \times \mathbb{P}_\theta\left(\varepsilon^{k}_{1:|e^{k}|} = e^{k} \,\middle|\, x_{1:n}, y_{1:m}\right) \sum_{e^{k-1}\in\mathcal{R}_l(e^k)} \mathbb{P}_\theta(\varepsilon^{k-1}[l] = e^{k-1}[l] |\, x[l], y_{e^k}[l], e^k_{[l]_-}, e^k_{[l]_+}).$$

The last sum is taken over all possible alignments of subsequences $x[l]$ and $y_{e^k}[l]$ given the two neighbouring positions and thus takes the value 1. Since $\sum_{l=1}^{n-n_0+1} p_l = 1$, we have

$$\sum_{e^{k-1}\in\mathcal{E}(n,m)} \mathbb{P}_\theta(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}|x_{1:n}, y_{1:m}) \times \Pi_\theta\left(\varepsilon^{k-1}_{1:|e^{k-1}|} = e^{k-1}, \varepsilon^{k}_{1:|e^{k}|} = e^{k} \,\middle|\, x_{1:n}, y_{1:m}\right) = \mathbb{P}_\theta\left(\varepsilon^{k}_{1:|e^{k}|} = e^{k} \,\middle|\, x_{1:n}, y_{1:m}\right)$$

which concludes the proof. □

2.4 Applications

We show in this section the performance of the algorithms described in Section 3.4 via some examples on simulated and real data.

For the sake of simplicity we use the Felsentein-81 substitution model (cf. [22]) with

$$\begin{aligned} g_\alpha(i,j) &= \begin{cases} \pi_i(e^{-\alpha} + \pi_i(1-e^{-\alpha})) & \text{for } i = j \\ \pi_i(1-e^{-\alpha})\pi_j & \text{for } i \neq j \end{cases} \qquad i,j \in \{A,C,G,T\} \\ h(i) &= \pi_i \end{aligned}$$

where π_i are the equilibrium probabilities of the four nucleotides. They will be replaced by their frequencies in the sequences (they are not re-estimated).
We recall that the insertion and deletion process described in this paper can be combined with any other time reversible substitution process.

Since insertion-deletion events are much less frequent than point substitutions, we will also assume that $\mu < \alpha_1$.

For the Bayesian procedure we have to consider a prior distribution for θ. We will take exponentials of mean 1 on λ, μ, α_1 and α_2 conditioned to $\lambda < \mu < \alpha_1 < \alpha_2$. This means that we sample λ from an exponential of mean 1 and then we sample μ from the density $e^{\mu-\lambda}$ for $\mu > \lambda$, α_1 from the density $e^{\alpha_1-\mu}$ for $\alpha_1 > \mu$ and finally α_2 from the density $e^{\alpha_2-\alpha_1}$ for $\alpha_2 > \alpha_1$. We will also use an exponential prior of mean 1 for $\gamma_1 - 1$ and $\gamma_2 - 1$. The proposal distribution for the Metropolis-Hastings step will be the prior distribution centered at the current value of the parameter.

We use Powell's optimization method, [65], for the maximization of the complete likelihood on $\lambda, \mu, \gamma_1, \gamma_2$ at each iteration of SAEM. Maximization on the substitution rates is done by solving

$$\sum_{i=1, j\neq i}^{4} n(i,j,k) = \sum_{i=1}^{4} \frac{n(i,i,k)(1-e^{-\alpha_k})}{e^{-\alpha_k} + \frac{\pi_i}{1-\pi_i}} \quad , \quad k = 1,2 \tag{2.7}$$

where $n(i, j, k)$ is the frequency of the pair of nucleotides (i, j) in an aligned pair of positions of type k for the given alignment.

2.4.1 Application to simulated data

The whole simulation procedure is the following. We choose the true value of the parameter, $\theta^\star$, and the length of the alignment, t. Then we simulate R alignment Markov chains of length t and R pairs of DNA sequences of random lengths n and m from the alignments. For each simulation we set the number of iterations to a value K and the number of iterations at the *burn-in* phase to a value k_0. For the choice of n_0, the length of the subalignment (based on the first sequence) to resample at each iteration, we have to take into account the algorithm's complexities and their speeds of convergence. Indeed, the convergence is guaranteed for any value of n_0 as we have seen in Section 3.4, however, the larger n_0 is, the more the alignments will change and the faster (in number of iterations) the algorithms will converge. On the other hand, as the runtime for realignment depends quadratically on n_0, the larger n_0 is the slower (in runtime per iteration) become the algorithms.

The initial parameter value θ_0 is chosen randomly from the prior distribution in both Gibbs sampling and SAEM. The initial alignment is sampled from the posterior distribution of alignments given the sequences and θ_0.

For the rth simulation we obtain an estimation $\hat{\theta}(r)$ of the parameters. For SAEM $\hat{\theta}(r) = \theta^K(r)$ and for the Gibbs sampling $\hat{\theta}(r) = \frac{1}{K-k_0}\sum_{k=k_0+1}^{K}\theta^k(r)$. After R simulations we get the mean and the variance of all estimates

$$\hat{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}(r), \quad \sigma^2 = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}(r) - \hat{\theta}\right)^2$$

and the mean square relative error

$$MSRE = \frac{1}{R}\sum_{r=1}^{R}\left(1 - \frac{\hat{\theta}(r)}{\theta^\star}\right)^2.$$

We now show the results obtained from two sets of simulated data. In the first one we set $\lambda^\star = 0.01$, $\mu^\star = 0.025$, $\gamma_1^\star = 10$, $\gamma_2^\star = 3$, $\alpha_1^\star = 0.05$ and $\alpha_2^\star = 0.5$, and all nucleotide frequencies are equal to 0.25 so that equation (2.7) has an explicit solution. For the second set we take $\lambda^\star = 0.015$, $\mu^\star = 0.02$, $\gamma_1^\star = 8$, $\gamma_2^\star = 2.5$, $\alpha_1^\star = 0.04$ and $\alpha_2^\star = 0.4$, and different nucleotide frequencies ($\pi_A = 0.2$, $\pi_C = 0.1$,$\pi_G = 0.15$ and $\pi_C = 0.55$). In this case (2.7) has no explicit solution and therefore the maximization of substitution rates in SAEM has

to be done numerically. We sample 400 alignments of length $t = 15000$ from each one of the parameter sets and we estimate $\theta^\star$ with both Gibbs sampling and SAEM. Table 1 and Table 2 show the parameter estimation results for these simulations. The performances of the two algorithms seem to be very similar given that Gibbs sampling needs a lot more iterations than SAEM. Indeed, we can see in Figures 2.1and 2.2 the convergence of the estimations for the two algorithms in a single simulation (remark that we are using a logarithmic scale for the x-axis to highlight the very fast convergence of some parameters ; therefore the parameters reaching their true values by the middle of the total number of runs would apparently seem to converge just at the stop of the algorithm, which is not the case).

		Gibbs sampling $K = 100000$ $k_0 = 30000$ $n_0 = 30$			SAEM $K = 15000$ $k_0 = 10000$ $n_0 = 30$		
	$\theta^\star$	$\hat{\theta}$	σ^2	$MSRE$	$\hat{\theta}$	σ^2	$MSRE$
λ	0.01	0.0111	1.0307e-05	0.1155	0.0104	1.0295e-05	0.1039
μ	0.025	0.0303	7.0243e-05	0.1564	0.0256	7.8476e-05	0.1258
γ_1	10	11.4729	15.1604	0.1729	11.5506	20.8702	0.2322
γ_2	3	3.1441	0.1462	0.0185	3.1336	0.5143	0.0590
α_1	0.05	0.0537	5.2662e-05	0.0264	0.0498	8.0462e-05	0.0321
α_2	0.5	0.5379	0.0078	0.0367	0.4834	0.0098	0.0403

TAB. 2.1 – *Estimations of θ for the first set of simulations.*

Figure 2.3 and 2.4 show the distributions of parameter estimations with both algorithms for the two sets of simulated data. We can appreciate that the estimations are concentrated around $\theta^\star$. Finally we can see that there is no remarkable difference in the performance of SAEM on the different sets of simulations, even if the second one requires an additional numerical algorithm to solve (2.7) at each iteration.

Let us just stress that the lengths of the sequences are an important factor on the quality of the estimation. Indeed, the likelihood in this model is very flat on the fragment length expectations γ_1 and γ_2 and simulations with shorter sequences have given results that are not as good as the ones shown here, especially on the estimation of these parameters. We recall that from a theoretical point of view, none of the ML or the Bayesian estimation under a pair-HMM have been totally validated (see Chapter 2).

At the level of alignments (and their associated segmentation) we can observe that the posterior distribution of alignments generated by the Gibbs sampling concentrates around the true alignment. This is shown in Figure 2.5 for a pair of simulated sequences from the

		Gibbs sampling $K = 100000$ $k_0 = 30000$ $n_0 = 30$			SAEM $K = 15000$ $k_0 = 10000$ $n_0 = 30$		
	$\theta^\star$	$\hat{\theta}$	σ^2	$MSRE$	$\hat{\theta}$	σ^2	$MSRE$
λ	0.015	0.0173	1.1213e-05	0.0726	0.0161	4.4007e-06	0.0246
μ	0.02	0.0255	3.6774e-05	0.0597	0.1656	9.2987e-06	0.0286
γ_1	8	8.7562	3.1942	0.0585	8.9133	2.7208	0.0554
γ_2	2.5	2.5931	0.0583	0.0107	2.6257	0.0519	0.0108
α_1	0.04	0.0562	2.3519e-04	0.3094	0.0444	1.4964e-04	0.1056
α_2	0.4	0.4459	0.0025	0.0286	0.4009	9.4219e-04	0.0059

TAB. 2.2 – *Estimations of θ for the second set of simulations.*

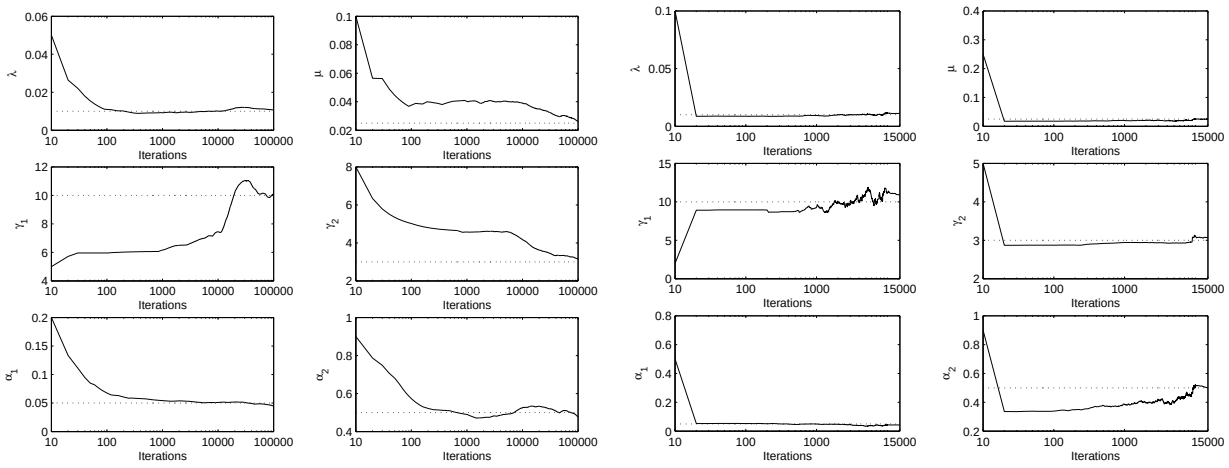


FIG. 2.1 – *Evolution over iterations of the Gibbs approximation of posterior means (left) and the SAEM estimation (right) in a single simulation from the first set of parameters. A logarithmic scale is used for the x-axis. The real values of the parameters are displayed in dotted line.*

second set of simulations ($\lambda^\star = 0.015, \mu^\star = 0.02, \gamma_1^\star = 8, \gamma_2^\star = 2.5, \alpha_1^\star = 0.04, \alpha_2^\star = 0.4$). This kind of representation of sampled alignments allow us to visualize the behavior of different possible alignments and if we are interested in keeping only one or a few alignments it can be useful for the choice of these alignments. Indeed, portions of alignment where all sampled alignments seem to coincide correspond to high probability alignment regions, so we will keep these regions in our choosen alignment. On the other hand, portions of alignment where more variability exists will need to be examined more carefully (see Figure2.6). For each sampled alignment proposed by the Gibbs sampling algorithm we can give a measure of reliability that can help us to compare several alignments. This

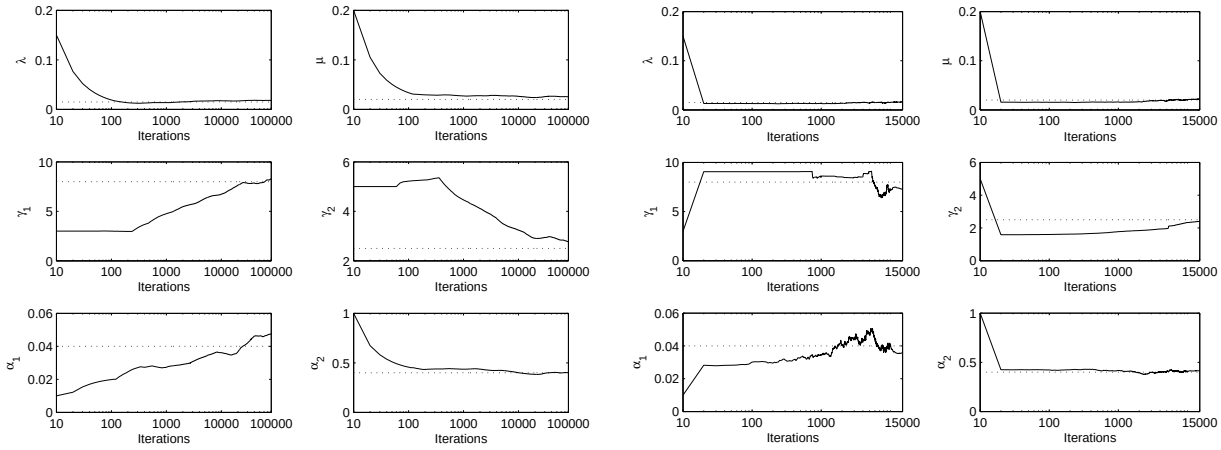


FIG. 2.2 – *Evolution over iterations of the Gibbs approximation of posterior means (left) and the SAEM estimation (right) in a single simulation from the second set of parameters. A logarithmic scale is used for the x-axis. The real values of the parameters are displayed in dotted line.*

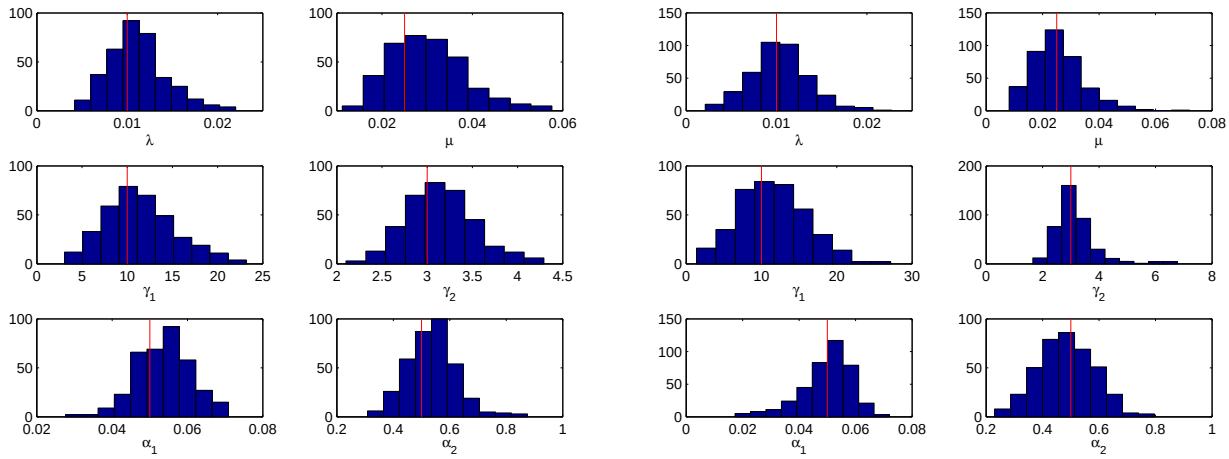


FIG. 2.3 – *Distribution of 400 estimations produced by Gibbs sampling (left) and SAEM (right) for the first set of simulations* ($\lambda^\star = 0.01, \mu^\star = 0.025, \gamma_1^\star = 10, \gamma_2^\star = 3, \alpha_1^\star = 0.05, \alpha_2^\star = 0.5$*).*

is done as follows : for a given value of the parameter θ (in general we will use the ML estimation) and for any pair of nucleotides x_i and y_j of the first and second sequence respectively, the Forward algorithm give us for all $k \in \{\mathtt{B}_{\mathrm{S}}, \mathtt{B}_{\mathrm{F}}, \bar{\mathtt{B}}_{\mathrm{F}}, \underline{\mathtt{B}}_{\mathrm{F}}\}$, the total probability $\mathbb{P}_\theta(x_{1:i}, y_{1:j}, (i,j) \to k) = \sum_{e \in \mathcal{E}(i,j), e_{|e|}=k} \mathbb{P}_\theta(x_{1:i}, y_{1:j}, \varepsilon_{1:|e|} = e)$ of all alignments up to x_i and y_j whose last state is k. The Backward calculation for pairHMMs (see [19]) give us

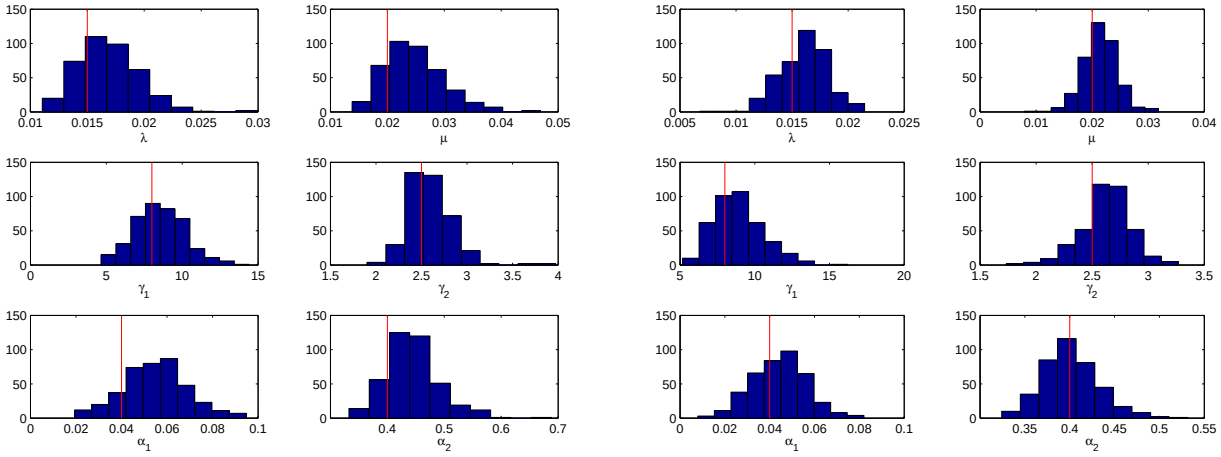


FIG. 2.4 – *Distribution of 400 estimations produced by Gibbs sampling (left) and SAEM (right) for the second set of simulations* ($\lambda^\star = 0.015, \mu^\star = 0.02, \gamma_1^\star = 8, \gamma_2^\star = 2.5, \alpha_1^\star = 0.04, \alpha_2^\star = 0.4$*).*

$\mathbb{P}_\theta(x_{i+1:n}, y_{j+1:m}|(i,j)\rightarrow k) =$
$\sum_{e\in\mathcal{E}(n-i,m-j)} \mathbb{P}_\theta(x_{i+1:n}, y_{j+1:m}, \varepsilon_{2:|e|+1} = e|\varepsilon_1 = k)$, the joint probablity of subsequences $x_{i+1:n}$ and $y_{j+1:m}$ given that nucleotides x_i and y_j are aligned in state k. Then, the joint probability of sequences x and y and nucleotides x_i and y_j being aligned in state k is $\mathbb{P}_\theta(x_{1:n}, y_{1:m}, (i,j)\rightarrow k) = \mathbb{P}_\theta(x_{1:i}, y_{1:j}, (i,j)\rightarrow k) \times \mathbb{P}_\theta(x_{i+1:n}, y_{j+1:m}|(i,j)\rightarrow k)$. The posterior probability of nucleotides x_i and y_j being aligned in state k is just $\frac{\mathbb{P}_\theta((i,j)\rightarrow k|x_{1:n},y_{1:m})}{\sum_l \mathbb{P}_\theta((i,j)\rightarrow l|x_{1:n},y_{1:m})} = \frac{\mathbb{P}_\theta(x_{1:n},y_{1:m},(i,j)\rightarrow k)}{\sum_l \mathbb{P}_\theta(x_{1:n},y_{1:m},(i,j)\rightarrow l)}$. Let us explain what "a pair of nucleotides x_i and y_j being aligned in state k" means when k is not a match state. For instance, if $k = \bar{\mathtt{B}}_{\mathtt{F}}$, a pair of nucleotides x_i and y_j is aligned in state k if the alignment until the current alignment position is an alignment of subsequences $x_{1:i-1}$ and $y_{1:j-1}$ and nucleotide y_j is inserted in the current alignment position.

So, for a pair of sequences and a given alignment we can compute this posterior distribution of states on each alignment position. That is what is done in Figure 2.7 for the sequences used in Figures 2.5 and 2.6 and the true alignment from which this sequences where generated. We used the ML estimation of parameters to compute the posterior distribution of states. We can observe that, as expected, for most positions in the true alignment the true state and our 'state prediction' (the state with the highest probability) coincides.

In practice we do not know the true alignment but, as we have already said, we can use this procedure with any given alignment namely with those given by the Gibbs sampling

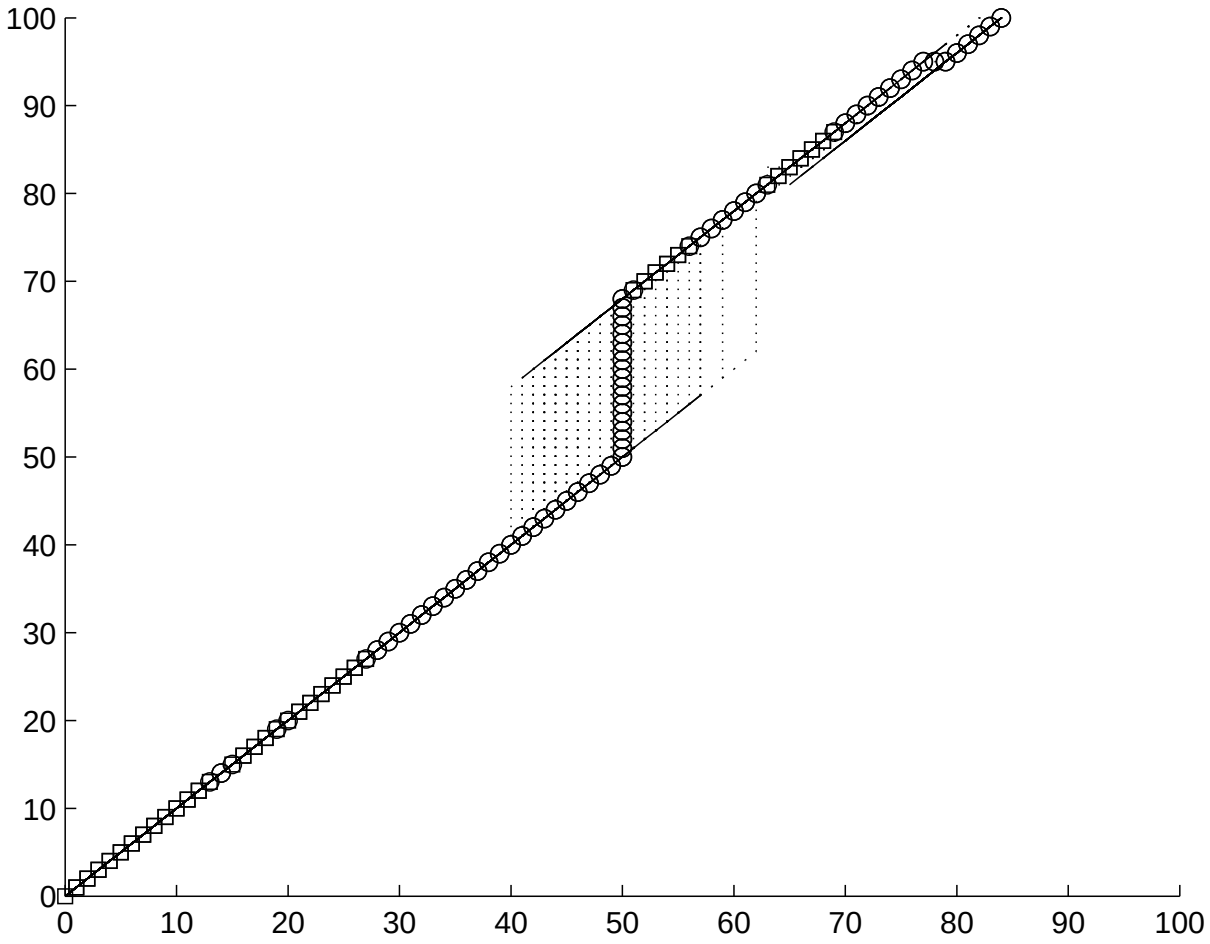


FIG. 2.5 – *True alignment (circles and squares) of a simulation from the second set* ($\lambda^\star = 0.015, \mu^\star = 0.02, \gamma_1^\star = 8, \gamma_2^\star = 2.5, \alpha_1^\star = 0.04, \alpha_2^\star = 0.4$) *versus the alignments sampled from their posterior distribution given the simulated sequences after 100000 iterations of Gibbs sampling (dotted and continuous line). We show the alignment over the first 100 positions on each sequence. In the true alignment squares represent the slow states and circles represent fast states. In the sampled alignments slow states are set in continuous line whereas fast states are set in dotted line.*

algorithm. A reliability measure for the chosen alignment is then obtained by considering only, at each position of the alignment, the probability of the state *predicted* at this position.

If we are interested in a segmentation of one of the sequences (the first one for instance)

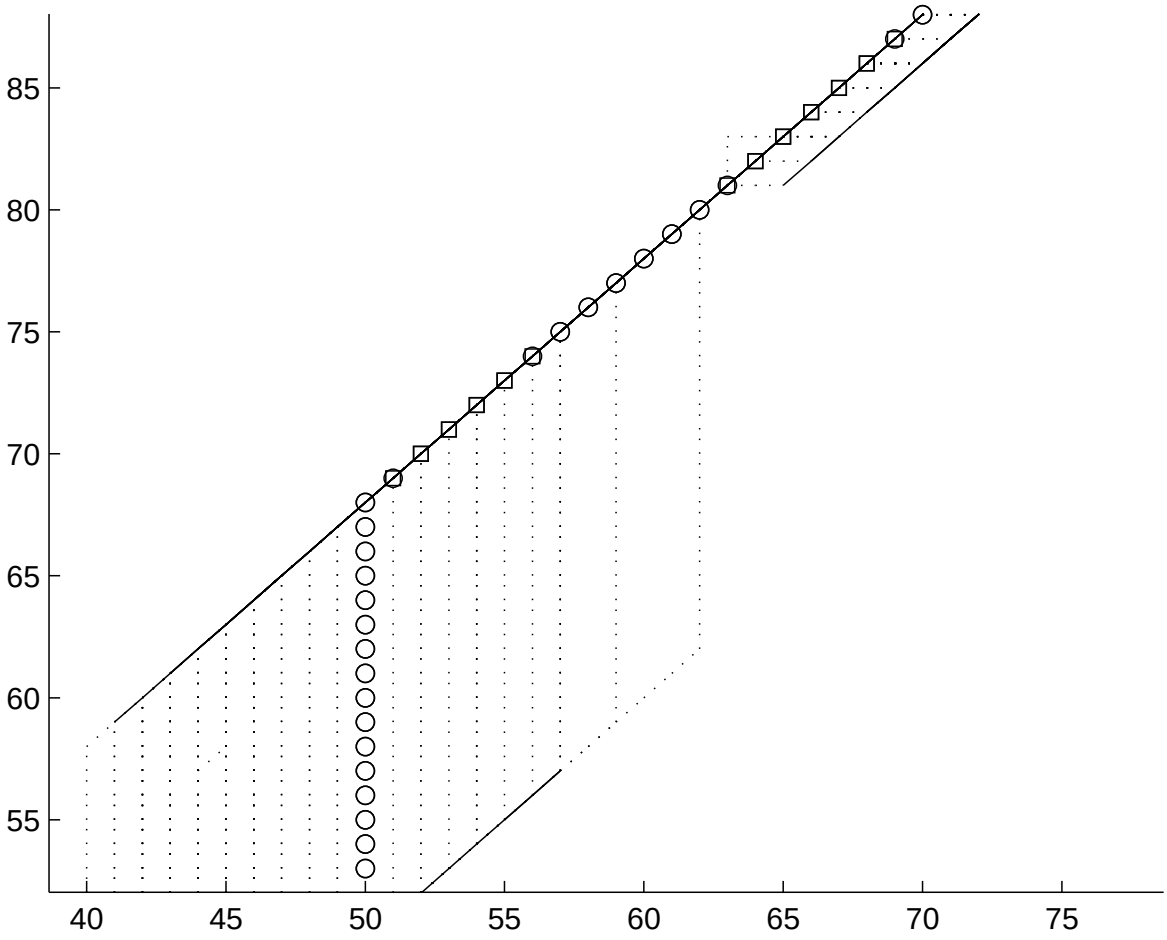
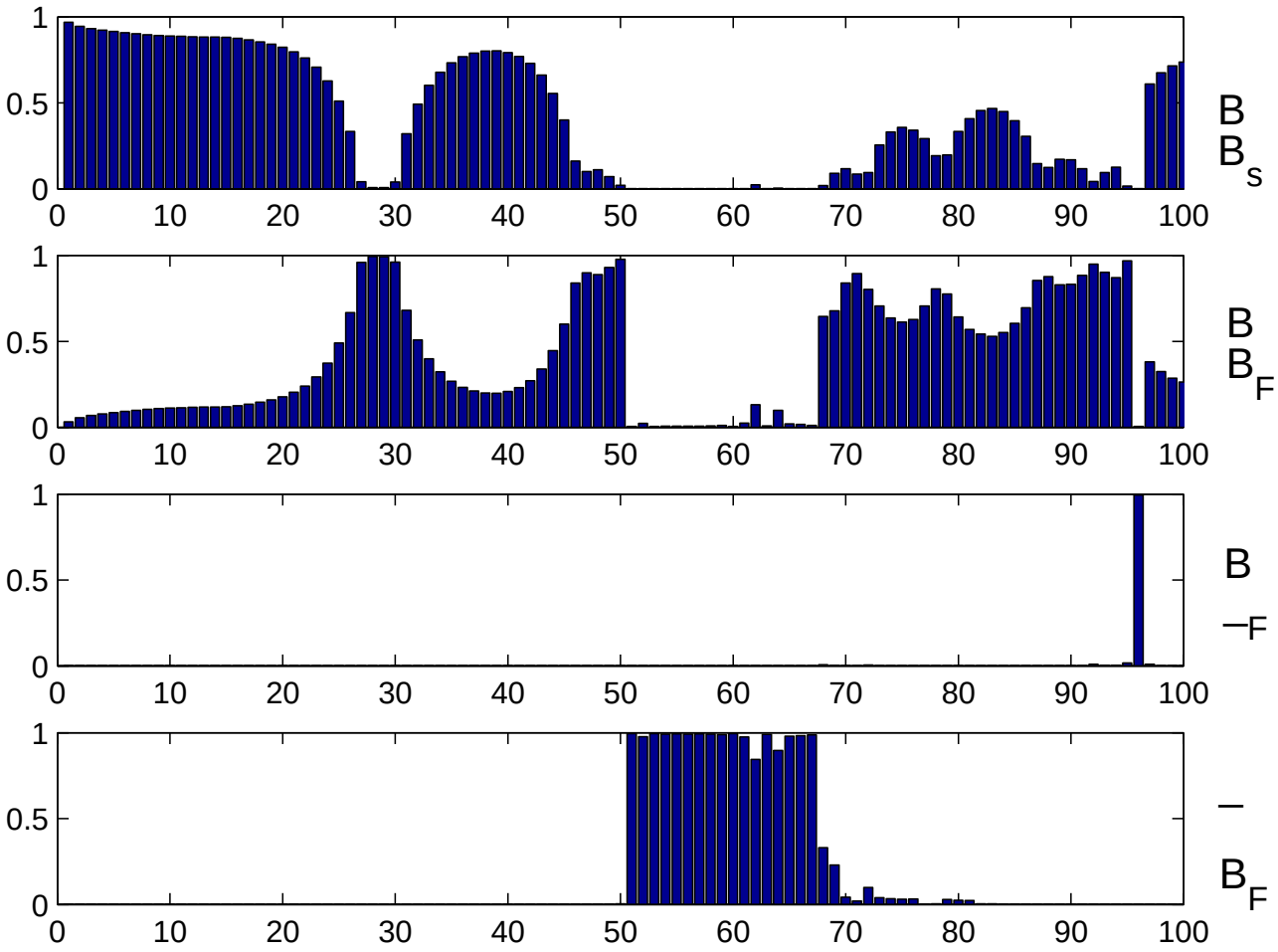


FIG. 2.6 – *True alignment (circles and squares) versus sampled alignments (dotted and continuous line) from Gibbs sampling. Zoom from Figure 2.5. In the true alignment squares represent the slow states and circles represent fast states. In the sampled alignments slow states are set in continuous line whereas fast states are set in dotted line.*

into conserved, less conserved and deleted positions this can be obtained by choosing a consensus alignment (from those given by the Gibbs sampling) and by computing the posterior distribution of states over the positions of this alignment, but now normalizing at each position over three of the states, namely $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{S}}$, $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{F}}$ and $\overset{\mathtt{B}}{\__\mathrm{F}}$. If there is no such a consensus alignment, another posibility to give a segmentation of the sequence is just to give for each position of the sequence the empirical distribution of states $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{S}}$, $\overset{\mathtt{B}}{\mathtt{B}_\mathrm{F}}$ and $\overset{\mathtt{B}}{\__\mathrm{F}}$ on the alignements sampled by the Gibbs sampling algorithm, that is to compute for each




```
1 TTTGGGATGTTATTAAATTGTGGGTCAAACAAGTTGATAGACCATGCGTC------------------TTTGAGGAATGCTTTATTCCACTGTGTGATGT 100
1 TTTGGGATGTTATTAAATTGTGGGTCAGGTAAGTTGATAGACCATGGGTGTGGGAGGAGTGCTATTTGTTTTAGGAATCCTTTATTCTACTTTGA--TGT 100
1 SSSSSSSSSSSSSFFSSSSFSSSSSSSFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFSSSSSFFFFFFFSSSSSSFFFFFFFFFFFFF 100
```

FIG. 2.7 – *Probability distribution of states for the ML estimation* ($\hat{\lambda}_{ML} = 0.016, \hat{\mu}_{ML} = 0.021, \hat{\gamma}_{1ML} = 7.138, \hat{\gamma}_{2ML} = 2.516, \hat{\alpha}_{1ML} = 0.031, \hat{\alpha}_{2ML} = 0.387$ *obtained after 15000 iterations of the SAEM algorithm) over the true simulated alignment of a simulation from the second set* ($\lambda^\star = 0.015, \mu^\star = 0.02, \gamma_1^\star = 8, \gamma_2^\star = 2.5, \alpha_1^\star = 0.04, \alpha_2^\star = 0.4$*). Zoom over the first 100 positions of the alignment.*

position of the sequence the number of alignments in wich it appears in each one of the states $\overset{B}{B_S}$, $\overset{B}{B_F}$ and $\overset{B}{-_F}$.

2.4.2 Application to real data

In order to test if our model is capable of correctly distinguishing regions under purifying selection from neutrally evolving sites, we tested our algorithm against a data set of conserved homologous non-coding sequences from the locus of the vertebrate Otx2 homeodomain gene, which is located in a gene desert and subject to a complex regulation during embryonic development (see [47, 48]). The data set contains five sequences from different vertebrate species (human(Hs), dog (Cf), mouse (Mm), marsupial (Md) and chicken (Gg)), homologous to a conserved 1.9kb[1] sequence located 83kb upstream of the initiation codon[2] of the human Otx2 gene. The aim is to use the algorithms presented in Section 3 to provide pairwise alignments and parameter estimations, and especially to give distributions of the degree of conservation along the sequences, which is the main innovation of this paper. We have considered four pairs of sequences, the resulting ones from coupling the human sequence with any of the four other sequences. The following has been done for any pair of sequences. In a first run, as the sequences are not long enough to allow reliable estimations of γ_1 and γ_2, we have fixed their values to 5, a reasonable fragment size. Then we ran the Gibbs sampling algorithm (100000 iterations), generating a distribution of parameter values and plausible alignments. We realized that all the sampled alignments (after the burn-in phase) were very similar on the positions of matches, insertions and deletions, but were not on the distribution of fast and slow evolution behaviors. Then we chose a consensus alignment without specifying fast and slow positions.

In a second run we applied the SAEM algorithm (15000 iterations) to compute the ML estimation of evolution parameters. Finally, for the chosen alignment and the ML estimation we have computed the probability distribution of states $\overset{\mathtt{B}}{\mathtt{B}_\mathtt{S}}$, $\overset{\mathtt{B}}{\mathtt{B}_\mathtt{F}}$, and $\overset{\mathtt{B}}{\mathtt{-}_\mathtt{F}}$ at every position (this has been done as explained in 3.4 for simulated data, except that we normalize here over only three of the states). We do not consider state $\overline{\mathtt{B}_\mathtt{F}}$ because we want to use the human sequence as the reference. For each nucleotide on it we want to give the probability of being in a fast position, in a slow position or in a deletion state. In this way we would provide a segmentation on evolution behaviors of the human sequence by marking as *conserved* the sites with high probability of being in state $\overset{\mathtt{B}}{\mathtt{B}_\mathtt{S}}$, as *variable* the sites with high probability of being in state $\overset{\mathtt{B}}{\mathtt{B}_\mathtt{F}}$ and as *deleted* the sites with high probability of being in state $\overset{\mathtt{B}}{\mathtt{-}_\mathtt{F}}$. The distributions obtained for all the alignments show a very conserved region at the beginning of the sequence (0-1000 b) and a more variable region after that. Let us focus on this last one, which is more informative about the differences on the degree of conservation on the human sequence in the different alignments.

[1] Kb stands for kilo base and is the common unit for DNA sequences length.

[2] The initiation codon indicates the begining of the coding region of a gene.

The results provided by the four alignments (Figure 2.8) are coherent with the evolution distance between sequences, i. e. the greater the evolution distance the smaller the degree of conservation.

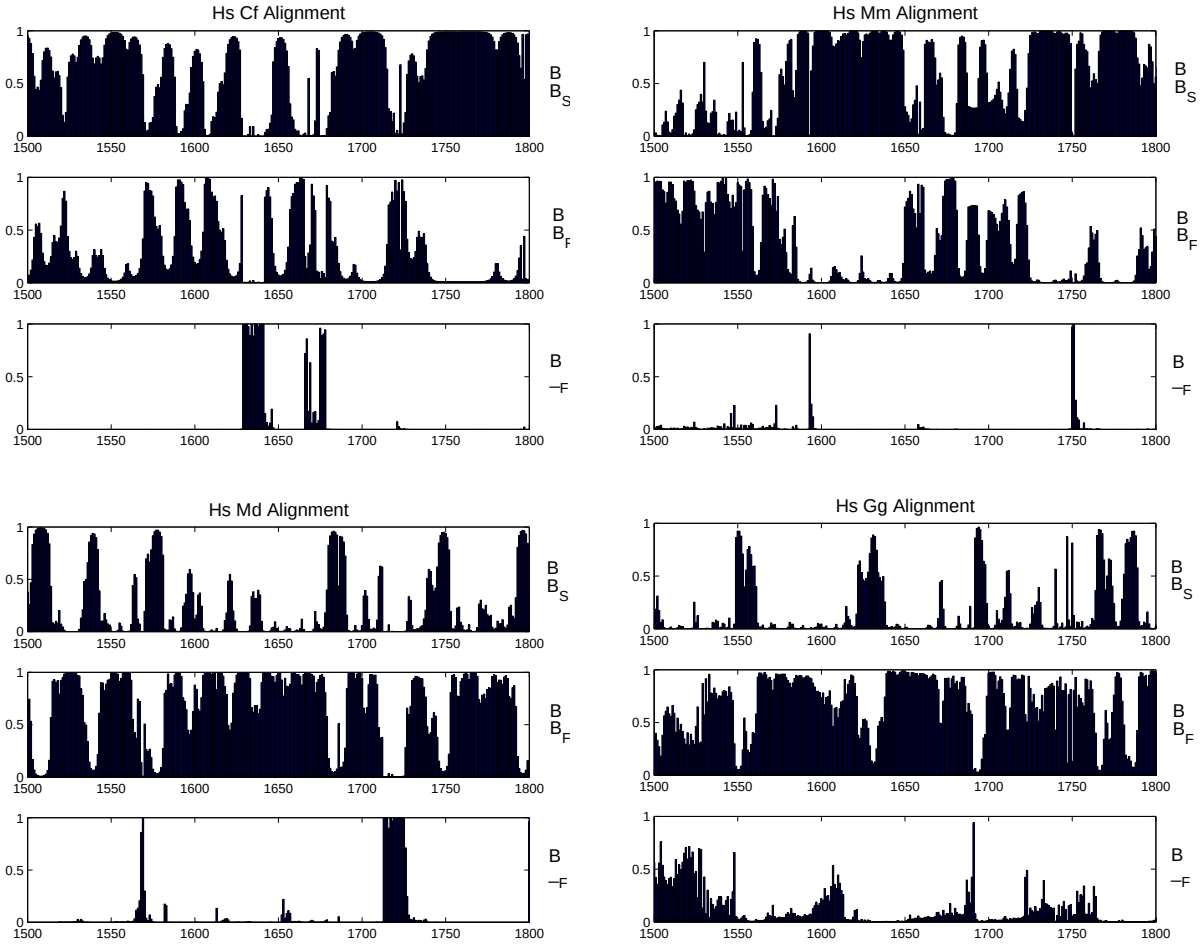


FIG. 2.8 – *Distribution of states* $\overset{B}{B_S}$, $\overset{B}{B_F}$ *and* $\underline{B}_F$ *over the human sequence of Otx2 for the alignments Hs_Cf, Hs_Mm, Hs_Md and Hs_Gg.*

2.5 Discussion

The fragment insertion and deletion process described in this paper makes it possible to consider rate heterogeneity along a DNA sequences. This process induces a pair-HMM structure at the level of pairwise alignments which give us the frame for the statistical estimation of alignments and mutation rates and for the segmentation of the sequences into *conserved* and *variable* regions.

The two approaches of estimation that we propose, namely a Bayesian approach implemented via the Gibbs sampling algorithm and an ML approach implemented via the SAEM algorithm, provide complementary information on the alignments and the evolution parameters and enables us to give a distribution of the degree of conservation along a sequence.

The promising results obtained with simulated data show the quality of both procedures of estimation. Analysis on real data give us coherent results with the previous knowledge we had about the analyzed sequences.

Two main extensions of our work can be considered. The evolution model could be easily generalized to allow more than two evolution regions by just adding states to the Markov chain. However, the practical issues of this new set up (computing time, numerical algorithms for maximization) should be studied carefully. Also, a more realistic analysis of sequences conservations would be made via multiple alignment. We are planning to generalize our model to the alignment of more than two sequences.

Chapitre 3

Parameter Estimation in Pair Hidden Markov Models

Abstract

In this chapter we deal with parameter estimation in pair hidden Markov models (pair-HMMs). We first provide a rigorous formalism for these models and discuss possible definitions of likelihoods. The model is biologically motivated and therefore naturally leads to restrictions on the parameter space. Existence of two different Information divergence rates is established and a divergence property is shown under additional assumptions. This yields consistency for the parameter in parametrization schemes for which the divergence property holds. Simulations illustrate different cases which are not covered by our results.

Contents

Ce chapitre correspond à un article en collaboration avec Elisabeth Gassiat et Cathérine Matias paru dans *Scandinavian Journal of Statistics [3].*

3.1 Introduction

3.1.1 Background

Sequence alignment has become one of the most powerful tools in bioinformatics. Biological sequences are aligned for instance (and among many other examples) to infer gene functions, to construct or use protein databases or to construct phylogenetic trees. Concerning this last topic, current methods first align the sequences and then infer the phylogeny given this fixed alignment. This approach contains a major flaw since the two problems are largely intertwined. Indeed, the alignment problem consists in retrieving the places, in the observed sequences, where substitution/deletion/insertion events have occurred, due to the evolution process. In the pair alignment problem, the observations consist in a couple of sequences $X_{1:n} = X_1 \dots X_n$ and $Y_{1:m} = Y_1 \dots Y_m$ with values on a finite state alphabet $\mathcal{A}$ ($\mathcal{A} = \{A, C, G, T\}$ for DNA sequences). It is assumed that the sequences share a common ancestor. According to biological evolution, the sequence of the ancestor evolves and letters in each site may change (substitution event), or be deleted (deletion event), or new letters may be inserted in the sequence (insertion event). This process finally leads to the two different observed sequences. A most convenient way of displaying alignments is a graphical representation as a path through a rectangular grid (see Figure 3.1). A diagonal move corresponds to a match between the two sequences, whereas horizontal and vertical moves correspond to insertion-deletion events. This path consists of steps ε_t, $t = 1, \dots, l$, where ε_t represents either a match ($\varepsilon_t = (1, 1)$) or an insertion-deletion event ($\varepsilon_t = (1, 0)$ or $(0, 1)$). The length of the alignment is l, and satisfies

$$n \vee m \leq l \leq n + m. \tag{3.1}$$

Here $n \vee m$ denotes the maximum value between n and m. The multiple alignment problem is the same, except that one has to retrieve the places where substitution/deletion/insertion events have occurred on the basis of a set of (more than two) sequences.

Aligning two sequences relies on the choice of a score optimization scheme (for instance, the Needleman-Wunsch algorithm [62]) and therefore the obtained alignments depend on the score parameters. Choosing these score parameters in the most objective way appears as a crucial issue. Because evolution is the force that promotes divergence between biological sequences, it is desirable to consider biological alignment in the context of evolution. Now, given an evolution model, optimal choices of the score parameters depend on the underlying unknown mutation rates and thus on the phylogeny to be inferred after the alignment. The existence of such a vicious circle explains the emergence of probabilistic models where optimal alignment and evolution parameters estimation are achieved at the same time.

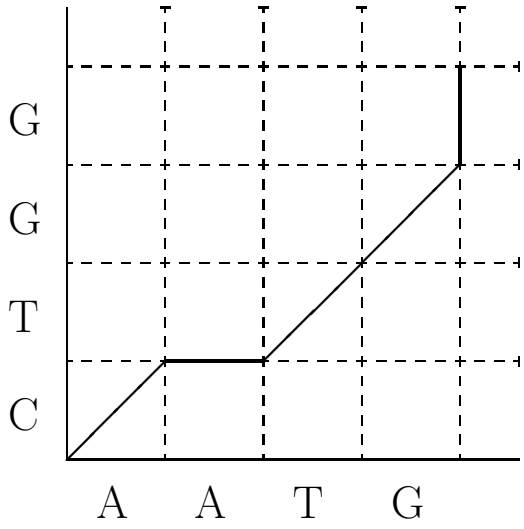


FIG. 3.1 – *Graphical representation of an alignment between two sequences* $X = AATG$ *and* $Y = CTGG$*. The displayed alignment is* $\begin{smallmatrix} A & A & TG & - \\ C & - & TG & G \end{smallmatrix}$.

Relying on a pioneering work by Bishop & Thompson [9], Thorne, Kishino & Felsenstein [76] were the first to provide a maximum likelihood approach to the alignment of a pair of DNA sequences based on a rigorous model of sequence evolution (referred to as the TKF model). This model has become quite classical nowadays. In this setup, each site is independently hit by a substitution or deleted, and insertions occur between two sites or at both ends of the sequence. Each one of those events occurs at a specific rate. When a substitution or an insertion occurs, a new nucleotide is drawn randomly according to some probability distribution on the state space $\{A, C, G, T\}$. One of the advantages of the TKF model lies in its exact correspondence with a model containing a hidden Markov structure, ensuring the existence of powerful algorithmic tools based on dynamic programming methods. More precisely, the TKF evolution model falls within the concept of a pair hidden Markov model (pair-HMM), as first formally described in Durbin *et al.* [19].

Observations in a pair-HMM are formed by a couple of sequences (the ones to be aligned) and the model assumes that the hidden (i.e. non observed) alignment sequence $\{\varepsilon_t\}_t$ is a Markov chain that determines the probability distribution of the observations. Since the seminal paper by Thorne *et al.* [76], an abundant literature aroused in which

parameter estimation occurs in a pair-HMM. Thorne, Kishino & Felsenstein [77] slightly improved their original model to take into account insertion and deletion of entire fragments (and not only single nucleotides). The TKF model approaches have been further developed, for instance in Hein *et al.* [31], Metzler [55], Knudsen & Miyamoto [44] and Miklos *et al.* [60]. Let us also mention that pair-HMMs were recently combined with classical hidden Markov models (HMMs) for *ab initio* prediction of genes (Meyer & Durbin [58] ; Pachter *et al.* [63] ; Hobolth & Jensen [34]).

The main difference between pair-HMMs and classical HMMs lies in the observation of a *pair* of sequences instead of a *single* one. From a practical point of view, the two above models are not very different and classical algorithms such as forward or Viterbi algorithms are still valid and efficient in the pair-HMM context (we refer to Durbin *et al.*, [19] for a complete description of those techniques). Forward algorithm allows to compute the likelihood of the two observed sequences and thus, by means of a maximization technique, to approximate the maximum likelihood estimator (MLE) of the parameters. Numerical maximization approaches are commonly used (see for instance [76]) but statistical approaches using the Expectation-Maximization (EM) algorithm and its variants (Stochastic EM, Stochastic Approximation EM) have recently been explored (Holmes [35] ; Arribas-Gil *et al.* [4] (Chapter 1 of this thesis)). Viterbi algorithm is designed to reconstruct the most probable hidden path, thus giving the alignment. From a Bayesian point of view, it is also interesting to provide a posterior distribution for parameters and alignments. This can be done with MCMC procedures needing again the use of the forward algorithm (Metzler [55] ; Arribas-Gil *et al.* [4] (Chapter 1 of this thesis)).

Nonetheless, from a theoretical point of view, pair-HMMs and classical HMMs are completely different. In particular, up to our knowledge, there is no theoretical proofs that the maximum likelihood procedure nor the Bayesian estimation give consistent estimators of the pair-HMM parameters (though it is the case for instance for regular HMMs with finite state space, (see Baum & Petrie [6] concerning MLE consistency ; see also Caliebe & Rösler [11], for the convergence of the maximum a posteriori hidden path).

This paper is thus concerned with statistical properties of parameter estimation procedures in pair-HMMs.

3.1.2 Roadmap

In Section 2, the pair-HMM is described, together with some properties of the distribution of observed sequences. Then we state possible likelihood functions, to be compared

with the criterion that is optimized in pair-HMM algorithms. We then interpret this last one as a likelihood function.
To investigate consistency of estimators obtained by maximization, one has to understand the asymptotic behaviour of the criteria. We adopt the Information Theory terminology and call *Information divergence rates* the difference between the limiting values of the log-likelihoods at the (unknown) true parameter value and at another parameter value. Indeed, the general model described below may be interpreted as a channel transmitting the input $X_{1:n}$ with possible errors, insertions or deletions, leading to the output $Y_{1:m}$ (see for instance Davey & MacKay [15] ; Levenshtein [50], on the topic of error correcting codes and also Csiszár & Körner [14], and Cover & Thomas [13], for a general introduction to Information Theory). In this setting, *Information divergence rates* have a precise meaning (in terms of coding or transmission qualities). In a statistical setting such as ours, they are interpreted as divergences that should have a unique minimum at the true parameter value (divergence property). Section 3 is devoted to the existence and properties of such limit functions (see Theorems 1 and 2).
Section 4, then, gives the statistical consequences in terms of consistent estimation of the parameters obtained via MLE or Bayesian estimation using pair-HMM algorithms (see Theorems 3, 4). According to these results, consistency holds for the parameter in parametrization schemes for which the divergence property holds for the associated Information divergence rate.
In Section 3.5, we present several simulation results to investigate situations in which the divergence property is not established. We illustrate the consistency results in cases where Theorem 3 applies, as may be seen on numerical computations of information divergence rates. We also compare the limiting values of different criteria and give some interpretations. The paper ends with a discussion on this work.

3.2 The pair hidden Markov model

3.2.1 Model description

We now describe in details the pair-HMM. Consider a stationary ergodic Markov chain $\{\varepsilon_t\}_{t\geq 1}$ on the state space $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$, with transition matrix π and stationary distribution $\mu = (p,q,r)$. This chain generates a random walk $\{Z_t\}_{t\geq 0}$ with values in the two-dimensional integer lattice $\mathbb{N}\times\mathbb{N}$, by letting $Z_0 = (0,0)$ and $Z_t = \sum_{1\leq s\leq t}\varepsilon_s$. The coordinate random variables corresponding to Z_t at time t are denoted by (N_t, M_t) (*i.e.* $Z_t = (N_t, M_t)$). We shall either use the notation $\pi(\varepsilon_s, \varepsilon_{s+1})$ to denote the transitions probabilities of the matrix π, or explicit symbols like π_{HV} indicating a

transition from state $H = (1, 0)$ to state $V = (0, 1)$ (H stands for *horizontal* move, V for *vertical* move and $D = (1, 1)$ for *diagonal* move).

Conditional on the hidden random walk, the observations are drawn according to the following scheme. At time t, if $\varepsilon_t = (1, 0)$ then a random variable X is drawn (emitted) according to some probability distribution f on $\mathcal{A}$, if $\varepsilon_t = (0, 1)$ then a random variable Y is drawn (emitted) according to some probability distribution g on $\mathcal{A}$ and finally, if $\varepsilon_t = (1, 1)$ then a couple of random variables (X, Y) is drawn (emitted) according to some probability distribution h on $\mathcal{A} \times \mathcal{A}$. Conditionally to the hidden Markov chain $\{\varepsilon_t\}_{t\geq 1}$, all emitted random variables are independent. This model is described by the parameter $\theta = (\pi, f, g, h) \in \Theta$. The conditional distribution of the observations thus writes

$$\begin{aligned}\mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|\varepsilon_{1:t}, \{\varepsilon_s\}_{s>t}, \{X_i, Y_j\}_{i\neq N_s, j\neq M_s, 0\leq s\leq t}) &= \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|\varepsilon_{1:t}) \\ &= \prod_{s=1}^{t} f(X_{N_s})^{\mathbb{1}\{\varepsilon_s=(1,0)\}} g(Y_{M_s})^{\mathbb{1}\{\varepsilon_s=(0,1)\}} h(X_{N_s}, Y_{M_s})^{\mathbb{1}\{\varepsilon_s=(1,1)\}}, \quad (3.2)\end{aligned}$$

where $\mathbb{1}\{\cdot\}$ stands for the indicator function. Moreover, the complete distribution $\mathbb{P}_\theta$ is given by

$$\mathbb{P}_\theta(\varepsilon_{1:t}, X_{1:N_t}, Y_{1:M_t}) = \mu(\varepsilon_1)\Big\{\prod_{s=2}^{t} \pi(\varepsilon_{s-1}, \varepsilon_s)\Big\}\mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|\varepsilon_{1:t}).$$

Here we denote by $\mathbb{P}_\theta$ (and $\mathbb{E}_\theta$) the induced probability distribution (and corresponding expectation) on $\mathcal{E}^{\mathbb{N}} \times \mathcal{A}^{\mathbb{N}} \times \mathcal{A}^{\mathbb{N}}$ and θ_0 the true parameter corresponding to the distribution of the observations (we shall abbreviate to $\mathbb{P}_0$ and $\mathbb{E}_0$ the probability distribution and expectation under parameter θ_0). Note that a necessary condition for identifiability of the parameter θ is that the occurrence probability of two aligned letters differs from the product probabilities of these letters. That is :

Assumption 1

$$\exists x, y \in \mathcal{A}, \text{ such that } h(x, y) \neq f(x)g(y).$$

Indeed, if $h = fg$, then (3.2) gives

$$\mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|\varepsilon_{1:t}) = \Big\{\prod_{i=1}^{N_t} f(X_i)\Big\}\Big\{\prod_{j=1}^{M_t} g(Y_j)\Big\} = \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}).$$

Thus, in this case, the distribution of the observations is independent from the hidden process and the parameter π cannot be identified. In the following, we shall always work under Assumption 1.

3.2.2 Observations and likelihoods

Statisticians define log-likelihoods to be functions of the parameter, that are equal to the logarithm of the probability of the observations. Here, to state what log-likelihoods are, one has to decide what do the observed sequences $(X_{1:n}, Y_{1:m})$ represent. Indeed, one may interpret it in at least two different ways :

(a) It is the observation of emitted sequences until some time t, so that the log-likelihood should be $\log \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t})$. Here, the probability is that one of the observed sequences *and* a point of the hidden process $Z_t = (N_t, M_t)$;

(b) Each observed sequence is one of the emitted sequences $X_{1:N_t}$ for some t and $Y_{1:M_s}$ for some s, knowing nothing on the hidden process (that is whether $t = s$, or $t > s$, or $t < s$), so that the log-likelihood should be $\log \mathbb{P}_\theta(X_{1:n}, Y_{1:m})$. Here, the probability is the marginal distribution of the sequences.

It should be now noted that none of those quantities is the one computed by pair-HMM algorithms. We will come back to this fact later (see (3.4) below). Note also that we imposed the true underlying alignment to pass through the fixed point $(0, 0)$ (namely, we assumed $Z_0 = (0, 0)$) which is not the more general setup (and may introduce a bias in practical applications). However, we restrict our attention to this particular setup.

First, we introduce some notations to make the previous quantities more precise. Let us consider the set $\mathcal{E}_\infty$ of all the possible trajectories of the hidden path and the set $\mathcal{E}_{n,m}$ of trajectories passing through the point (n, m) :

$$\begin{aligned}\mathcal{E}_\infty &= \{(0,1);(1,0);(1,1)\}^{\mathbb{N}} = \{e = (e_1, e_2, \ldots)\} = \mathcal{E}^{\mathbb{N}},\\ \mathcal{E}_{n,m} &= \{e \in \{(0,1);(1,0);(1,1)\}^l; n \vee m \leq l \leq n+m; \sum_{i=1}^{l} e_i = (n,m)\}.\end{aligned}$$

The length of any trajectory $e \in \mathcal{E}_{n,m}$ is denoted by $|e|$. Then, we have the following equations

$$\mathbb{P}_\theta(X_{1:n}, Y_{1:m}) = \sum_{e \in \mathcal{E}_\infty} \mathbb{P}_\theta(\varepsilon_{1:\infty} = e_{1:\infty}, X_{1:n}, Y_{1:m}), \tag{3.3}$$

$$\mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}) = \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}, Z_t) = \sum_{e \in \mathcal{E}_{N_t,M_t}; |e|=t} \mathbb{P}_\theta(\varepsilon_{1:t} = e_{1:t}, X_{1:N_t}, Y_{1:M_t}).$$

As Equation (3.3) shows, if one uses the marginal distributions as likelihood, it means that when observing two sequences $X_{1:n}$ and $Y_{1:m}$, it is not assumed that the hidden process passes through the observed point (n, m). This results in an alignment with not necessarily bounded length (see Figure 3.2). We shall now detail Equation (3.3) according to possible alignments. Among all the trajectories in $\mathcal{E}_\infty$, we shall distinguish the ones in $\mathcal{E}_{n,m}$ and the ones belonging to some set $\mathcal{E}_{n,p}$ (with $p > m$) or $\mathcal{E}_{p,m}$ (with $p > n$). Those

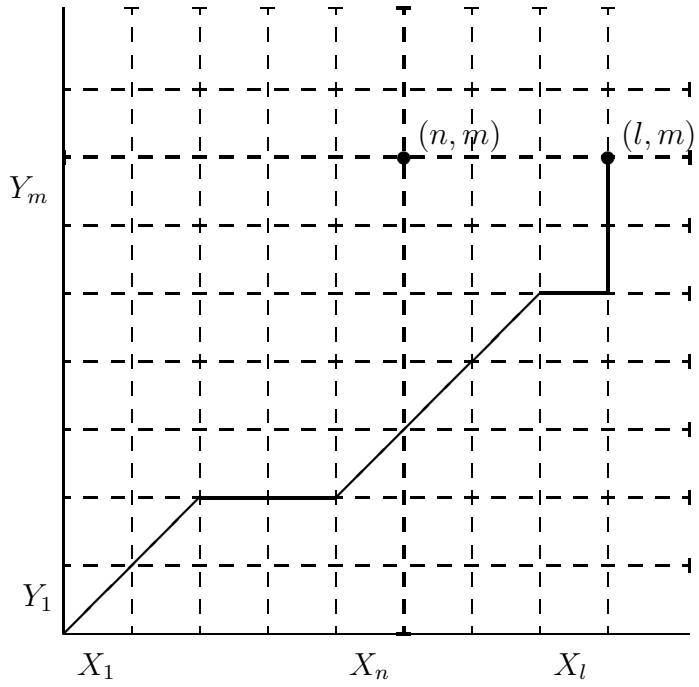


FIG. 3.2 – *Graphical representation of an alignment of sequences $X_{1:n}$ and $Y_{1:m}$ not passing through the point (n,m).*

last ones need to be constrained in order to avoid multiple counting. Let us denote by $\mathcal{E}_{n,m}^{-H}$ (resp. $\mathcal{E}_{n,m}^{-V}$) the restriction of the set $\mathcal{E}_{n,m}$ to trajectories not ending with an horizontal (resp. vertical) part. More precisely,

$$\mathcal{E}_{n,m}^{-H} = \{e = (e_1, \ldots, e_{|e|}) \in \mathcal{E}_{n,m}; e_{|e|} \neq (1,0)\}, \quad \mathcal{E}_{n,m}^{-V} = \{e \in \mathcal{E}_{n,m}; e_{|e|} \neq (0,1)\}.$$

These notations allow to express the marginal distribution $\mathbb{P}_\theta(X_{1:n}, Y_{1:m})$ as a sum over three different path types.

$$\begin{aligned}\mathbb{P}_\theta(X_{1:n}, Y_{1:m}) = &\sum_{e \in \mathcal{E}_{n,m}} \mathbb{P}_\theta(\varepsilon_{1:|e|} = e, X_{1:n}, Y_{1:m}) \\ &+ \sum_{l>n} \sum_{e \in \mathcal{E}_{l,m}^{-H}} \sum_{x_{n+1:l}} \mathbb{P}_\theta(\varepsilon_{1:|e|} = e, X_{1:n}, X_{n+1:l} = x_{n+1:l}, Y_{1:m}) \\ &+ \sum_{l>m} \sum_{e \in \mathcal{E}_{n,l}^{-V}} \sum_{y_{m+1:l}} \mathbb{P}_\theta(\varepsilon_{1:|e|} = e, X_{1:n}, Y_{1:m}, Y_{m+1:l} = y_{m+1:l}).\end{aligned}$$

This form may not be used for the computation of the marginal distribution $\mathbb{P}_\theta(X_{1:n}, Y_{1:m})$.

We now give some recursion formulas that could lead to practical implementations of this last quantity. For any state $e \in \mathcal{E}$, define $\mathbb{P}_\theta^e$ as the distribution induced by $\mathbb{P}_\theta$ conditional on $\varepsilon_1 = e$. Let us also denote by h_X (resp. h_Y) the marginal with respect to the first (resp. second) coordinate of the distribution h.

Lemma 1 *For any $n \geq 1, m \geq 1$,*

$$\mathbb{P}_\theta(X_{1:n}, Y_{1:m}) = p\,\mathbb{P}_\theta^H(X_{1:n}, Y_{1:m}) + q\,\mathbb{P}_\theta^V(X_{1:n}, Y_{1:m}) + r\,\mathbb{P}_\theta^D(X_{1:n}, Y_{1:m}),$$

with the following recursions

$$\begin{aligned}
\mathbb{P}_\theta^H(X_{1:n}, Y_{1:m}) &= f(X_1)\{\pi_{HH}\mathbb{P}_\theta^H(X_{2:n}, Y_{1:m}) + \pi_{HV}\mathbb{P}_\theta^V(X_{2:n}, Y_{1:m}) + \pi_{HD}\mathbb{P}_\theta^D(X_{2:n}, Y_{1:m})\} \\
\mathbb{P}_\theta^V(X_{1:n}, Y_{1:m}) &= g(Y_1)\{\pi_{VH}\mathbb{P}_\theta^H(X_{1:n}, Y_{2:m}) + \pi_{VV}\mathbb{P}_\theta^V(X_{1:n}, Y_{2:m}) + \pi_{VD}\mathbb{P}_\theta^D(X_{1:n}, Y_{2:m})\} \\
\mathbb{P}_\theta^D(X_{1:n}, Y_{1:m}) &= h(X_1, Y_1)\{\pi_{DH}\mathbb{P}_\theta^H(X_{2:n}, Y_{2:m}) + \pi_{DV}\mathbb{P}_\theta^V(X_{2:n}, Y_{2:m}) + \pi_{DD}\mathbb{P}_\theta^D(X_{2:n}, Y_{2:m})\}
\end{aligned}$$

and initializations :

$$\begin{aligned}
\mathbb{P}_\theta^H(X_1) = f(X_1), \quad \mathbb{P}_\theta^V(Y_1) = g(Y_1), \quad \mathbb{P}_\theta^D(X_1, Y_1) &= h(X_1, Y_1), \\
\mathbb{P}_\theta^V(X_{1:n}) &= (1-\pi_{VV})^{-1}\{\pi_{VH}\, f(X_1)\mathbb{P}_\theta^H(X_{2:n}) + \pi_{VD}\, h_X(X_1)\mathbb{P}_\theta^D(X_{2:n})\}, \\
\mathbb{P}_\theta^H(Y_{1:m}) &= (1-\pi_{HH})^{-1}\{\pi_{HV}\, g(Y_1)\mathbb{P}_\theta^V(Y_{2:m}) + \pi_{HD}\, h_Y(Y_1)\mathbb{P}_\theta^D(Y_{2:m})\}.
\end{aligned}$$

Proof of Lemma 1 is trivial and therefore omitted.

Interpretation (a) leads to define the log-likelihood $\ell_t(\theta)$ as

$$\ell_t(\theta) = \log \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}), \quad t \geq 1.$$

But since the underlying process $\{Z_t\}_{t\geq 0}$ is not observed, the quantity $\ell_t(\theta)$ is not a function of the observations alone. More precisely, the *time* t at which observation is made is not observed itself. Though, if one decides to use interpretation (a), namely that $(X_{1:n}, Y_{1:m})$ corresponds to the observation of the emitted sequences at a point of the hidden process $Z_t = (N_t, M_t)$ and some *unknown* time t, one does not use $\ell_t(\theta)$ as a log-likelihood, but rather

$$w_t(\theta) = \log Q_\theta(X_{1:N_t}, Y_{1:M_t}), \quad t \geq 1$$

where for any integers n and m

$$Q_\theta(X_{1:n}, Y_{1:m}) = \mathbb{P}_\theta(\exists s \geq 1, Z_s = (n, m); X_{1:n}, Y_{1:m}).$$

In other words, Q_θ is the probability of the observed sequences under the assumption that the underlying process $\{\varepsilon_t\}_{t\geq 1}$ passes through the point (n, m). But the length of the hidden trajectory remains unknown when computing Q_θ. This gives the formula :

$$Q_\theta(X_{1:n}, Y_{1:m}) = \sum_{e\in\mathcal{E}_{n,m}} \mathbb{P}_\theta(\varepsilon_{1:|e|} = e, X_{1:n}, Y_{1:m}). \tag{3.4}$$

Let us stress that we have

$$w_t(\theta) = \log \mathbb{P}_\theta(\exists s \geq 1, Z_s = (N_t, M_t); X_{1:N_t}, Y_{1:M_t}), \quad t \geq 1,$$

meaning that the length of the trajectory is not necessarily t, but is in fact unknown.

Q_θ is the quantity that is computed by forward algorithm (see [19]) and which is used as likelihood in biological applications. It is computed via recursive equations similar to those of Lemma 1. In practice, paths with highest scores according to the the Needleman-Wunsch scoring scheme exactly correspond to highest probability paths in a pair-HMM, with a corresponding choice of the parameters ([19]). Thus, the quantity Q_θ is used for finding the best alignment between two sequences. Moreover, as we explained it in the introduction, the idea of maximizing this quantity with respect to the parameter θ has now widely spread among practitioners (Thorne *et al.* [76, 77] ; Hein *et al.* [31] ; Metzler [55] ; Knudsen & Miyamoto [44] ; Miklos *et al.* [60]). The goal is to obtain an objective choice of the parameters appearing in the scoring scheme, taking evolution into account. Thus, asymptotic properties of criterion Q_θ and consequences on asymptotic properties of the estimator derived from Q_θ are of primarily interest.
According to the relation (3.1), asymptotic results for $t \to \infty$ will imply equivalent ones for $n, m \to \infty$. In other words, consistency results obtained when $t \to \infty$ can be interpreted as valid for long enough observed sequences, even if one does not know t.

3.2.3 Biologically motivated restrictions

Evolution models are commonly chosen time reversible, in the limit of infinitely long sequences. The reversibility property implies that the joint probability of sequence X and an ancestor sequence U is not influenced by the fact that X is a descendant of sequence U : this joint probability would be the same if X were an ancestor of U or if both were descendants of a third sequence. Note that this assumption does not apply on the level of alignments. Indeed, for single alignments, one may have $\mathbb{P}_\theta(\varepsilon = e, X, Y) \neq \mathbb{P}_\theta(\varepsilon = e', Y, X)$, where e and e' are equal on diagonal steps and have switched insertions and deletions (namely, corresponding paths are symmetric around the axis $x = y$). In fact, it is the probability of a whole given set of evolution events (namely mutations, insertions or

deletions occurring in the evolution process), which is a sum over different alignments e (all representing this same set of evolution events) of probabilities $\mathbb{P}_\theta(\varepsilon = e, X, Y)$, which is conserved if we interchange the two observed sequences. More precisely, we always have $\sum_{e\in\mathcal{E}_1} \mathbb{P}_\theta(\varepsilon = e, X, Y) = \sum_{e\in\mathcal{E}_2} \mathbb{P}_\theta(\varepsilon = e, Y, X)$ where $\mathcal{E}_1$ and $\mathcal{E}_2$ are alignments subsets representing the same set of evolution events.
Evolution models rely on two separate processes : the insertion-deletion (indel) and the substitution process and both are supposed to be time reversible. As a consequence of time reversibility of indel process, the stationary probability of appearance of an insertion or of a deletion is the same, meaning that $p = q$. We thus introduce the following assumption on the stationary distribution of the hidden Markov chain :

Assumption 2 $p = q$.

Time reversibility assumption on the substitution process implies equality between the marginals of h and individual distributions of the letters, namely $h_X = f$ and $h_Y = g$. We thus also introduce the following assumption on the emission distributions :

Assumption 3 $h_X = f$ *and* $h_Y = g$.

This last assumption has an interesting consequence on the distribution of only one sequence :

Lemma 2 *Under Assumption 3, for any integers n and m, any $x_{1:n}$ and any $y_{1:m}$*

$$\mathbb{P}_\theta\left(Z_t = (n,m), X_{1:n} = x_{1:n}\right) = \mathbb{P}_\theta\left(Z_t = (n,m)\right) f^{\otimes n}(x_{1:n}),$$

$$\mathbb{P}_\theta\left(Z_t = (n,m), Y_{1:m} = y_{1:m}\right) = \mathbb{P}_\theta\left(Z_t = (n,m)\right) g^{\otimes m}(y_{1:m}).$$

Here, $f^{\otimes n}(x_{1:n}) \triangleq f(x_1)\dots f(x_n)$.

Proof. One has

$$\begin{aligned}
\mathbb{P}_\theta\left(Z_t = (n,m), X_{1:n} = x_{1:n}\right) &= \sum_{y_{1:m}} \mathbb{P}_\theta\left(Z_t = (n,m), X_{1:n} = x_{1:n}, Y_{1:m} = y_{1:m}\right) \\
&= \sum_{e\in\mathcal{E}_{n,m},|e|=t} \sum_{y_{1:m}} \mathbb{P}_\theta\left(\varepsilon_{1:t} = e, X_{1:n} = x_{1:n}, Y_{1:m} = y_{1:m}\right) \\
&= \sum_{e\in\mathcal{E}_{n,m},|e|=t} \mathbb{P}_\theta\left(\varepsilon_{1:t} = e\right) \sum_{y_{1:m}} \mathbb{P}_\theta\left(X_{1:n} = x_{1:n}, Y_{1:m} = y_{1:m} | \varepsilon_{1:t} = e\right),
\end{aligned}$$

so that use of equation (3.2) and Assumption 3 gives the first assertion of the Lemma. Proof of the second assertion is similar. □

3.3 Information divergence rates

3.3.1 Definition of Information divergence rates

In this section, we investigate the asymptotic properties of the *log-likelihoods* $\ell_t(\theta)$ and $w_t(\theta)$ when properly normalized. We first prove that limiting functions exist. We shall need the following parameter sets Θ_δ, $\delta > 0$ and $\Theta_0 = \cap_{\delta>0}\Theta_\delta$:

$$\begin{aligned}\Theta_\delta &= \{\theta \in \Theta \,|\, \pi(i,j) \geq \delta,\ f(x) \geq \delta,\ g(y) \geq \delta,\ h(x,y) \geq \delta, \forall i,j \in \mathcal{E},\ \forall x,y \in \mathcal{A}\}\,,\\ \Theta_0 &= \{\theta \in \Theta \,|\, \pi(i,j) > 0,\ f(x) > 0,\ g(y) > 0,\ h(x,y) > 0,\ \forall i,j \in \mathcal{E},\ \forall x,y \in \mathcal{A}\}\,.\end{aligned}$$

We shall always assume that $\theta_0 \in \Theta_0$.

Theorem 1 *The following holds for any $\theta \in \Theta_0$:*

i) $t^{-1}\ell_t(\theta)$ converges $\mathbb{P}_0$-almost surely and in $\mathbb{L}_1$, as t tends to infinity to

$$\ell(\theta) = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}_0\left(\log \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t})\right) = \sup_t \frac{1}{t}\mathbb{E}_0\left(\log \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t})\right).$$

ii) $t^{-1}w_t(\theta)$ converges $\mathbb{P}_0$-almost surely and in $\mathbb{L}_1$, as t tends to infinity to

$$w(\theta) = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}_0\left(\log Q_\theta(X_{1:N_t}, Y_{1:M_t})\right) = \sup_t \frac{1}{t}\mathbb{E}_0\left(\log Q_\theta(X_{1:N_t}, Y_{1:M_t})\right).$$

We then define Information divergence rates :

Definition 1 $\forall \theta \in \Theta_0,\ D(\theta|\theta_0) = w(\theta_0) - w(\theta) \quad \textit{and} \quad D^*(\theta|\theta_0) = \ell(\theta_0) - \ell(\theta).$

Note that D^* is what is usually called the Information divergence rate in Information Theory : it is the limit of the normalized Kullback-Leibler divergence between the distributions of the observations at the true parameter value and another parameter value. However, we also call D an Information divergence rate since Q_θ may be interpreted as a likelihood.

Proof of Theorem 1. This proof follows the lines of Leroux ([49], Theorem 2). We shall use the following version of the sub-additive ergodic Theorem due to Kingman [43] to prove point *i)*. A similar proof may be written for *ii)* and is left to the reader.
Let $(W_{s,t})_{0\leq s<t}$ be a sequence of random variables such that

1. For all $s < t$, $W_{0,t} \geq W_{0,s} + W_{s,t}$,
2. For all $k > 0$, the joint distributions of $(W_{s+k,t+k})_{0\leq s<t}$ are the same as those of $(W_{s,t})_{0\leq s<t}$,
3. $\mathbb{E}_0(W_{0,1}) > -\infty$.

Then $\lim_{t\to\infty} t^{-1}W_{0,t}$ exists almost surely. If moreover the sequences $(W_{s+k,t+k})_{k>0}$ are ergodic, then the limit is almost surely deterministic and equals $\sup_t t^{-1}\mathbb{E}_0(W_{0,t})$. If moreover $\mathbb{E}_0(W_{0,t}) \leq At$, for some constant $A \geq 0$ and all t, then the convergence holds in $\mathbb{L}_1$.

We apply this theorem to the auxiliary process

$$W_{s,t} = \max_{e\in\mathcal{E}} \log \mathbb{P}_\theta(X_{N_s+1:N_t}, Y_{M_s+1:M_t}|\varepsilon_{s+1} = e) + \log(\delta_\theta), \quad 0 \leq s < t,$$

where $\delta_\theta = \min_{e,e'\in\mathcal{E}} \pi(e,e') > 0$. We are interested in the behaviour of

$$U_{s,t} = \log \mathbb{P}_\theta(X_{N_s+1:N_t}, Y_{M_s+1:M_t}), \quad 0 \leq s < t.$$

Since we have $\exp(U_{s,t}) = \sum_{e\in\mathcal{E}} \mathbb{P}_\theta(\varepsilon_{s+1} = e)\mathbb{P}_\theta(X_{N_s+1:N_t}, Y_{M_s+1:M_t}|\varepsilon_{s+1} = e)$ leading to $\exp(W_{s,t} - \log\delta_\theta) \min_{e\in\mathcal{E}} \mathbb{P}_\theta(\varepsilon_1 = e) \leq \exp(U_{s,t}) \leq \exp(W_{s,t} - \log\delta_\theta)$, we can conclude that the desired results on $\lim_{t\to\infty} t^{-1}U_{0,t}$ and $\lim_{t\to\infty} t^{-1}\mathbb{E}_0(U_{0,t})$ follow from corresponding ones on the process W.

Note that since $Z_0 = (0,0)$ is deterministic, we have $W_{0,t} = \max_{e\in\mathcal{E}} \log \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|$ $\varepsilon_1 = e) + \log\delta_\theta$. Super-additivity (namely point 1.) follows since for any $0 \leq s < t$,

$$\begin{aligned}
\mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|\varepsilon_1 = e_1) &= \sum_{\substack{e\in\mathcal{E}_{N_t,M_t} \\ |e|=t}} \mathbb{P}_\theta(\varepsilon_{2:t} = e_{2:t}, X_{1:N_t}, Y_{1:M_t}|\varepsilon_1 = e_1) \\
&\geq \sum_{\substack{e^1\in\mathcal{E}_{N_s,M_s} \\ |e^1|=s}} \sum_{\substack{e^2\in\mathcal{E}_{N_t-N_s,M_t-M_s} \\ |e^2|=t-s}} \mathbb{P}_\theta(\varepsilon_{2:s} = e^1_{2:s}, \varepsilon_{s+1:t} = e^2, X_{1:N_t}, Y_{1:M_t}|\varepsilon_1 = e_1) \\
&= \sum_{\substack{e^1\in\mathcal{E}_{N_s,M_s} \\ |e^1|=s}} \sum_{\substack{e^2\in\mathcal{E}_{N_t-N_s,M_t-M_s} \\ |e^2|=t-s}} \mathbb{P}_\theta(\varepsilon_{s+2:t} = e^2_{2:t-s}, X_{N_s+1:N_t}, Y_{M_s+1:M_t}|\varepsilon_{s+1} = e^2_1) \\
&\qquad \times \pi(e_s, e_{s+1})\mathbb{P}_\theta(\varepsilon_{2:s} = e^1_{2:s}, X_{1:N_s}, Y_{1:M_s}|\varepsilon_1 = e_1) \\
&= \sum_{e_s,e_{s+1}\in\mathcal{E}} \mathbb{P}_\theta(X_{N_s+1:N_t}, Y_{M_s+1:M_t}|\varepsilon_{s+1} = e_{s+1})\pi(e_s, e_{s+1})\mathbb{P}_\theta(\varepsilon_s = e_s, X_{1:N_s}, Y_{1:M_s}|\varepsilon_1 = e_1) \\
&\geq \{\max_{e'\in\mathcal{E}} \mathbb{P}_\theta(X_{N_s+1:N_t}, Y_{M_s+1:M_t}|\varepsilon_{s+1} = e')\}\{\min_{e,e'} \pi(e,e')\}\mathbb{P}_\theta(X_{1:N_s}, Y_{1:M_s}|\varepsilon_1 = e_1)\ ,
\end{aligned}$$

so that we get $W_{0,t} \geq W_{0,s} + W_{s,t}$, for any $0 \leq s < t$.

To understand the distribution of $(W_{s,t})_{0\leq s<t}$, note that $W_{s,t}$ only depends on trajectories of the random walk going from the point (N_s, M_s) to the point (N_t, M_t) with length $t-s$. Since the process $(\varepsilon_t)_{t\in\mathbb{N}}$ is stationary, one gets that the distribution of $(W_{s,t})$ is the same as that of $(W_{s+k,t+k})$ for any k, so that point 2. holds.

Point 3. comes from :

$$\mathbb{E}_0(W_{0,1}) - \log\delta_\theta = \mathbb{E}_0 \max\{\log f(X_1); \log g(Y_1); \log h(X_1, Y_1)\} > -\infty,$$

$\mathbb{P}_0$-almost surely, since $\theta \in \Theta_0$. Let us fix $0 \leq s < t$. The proof that $W^{s,t} = (W_{s+k,t+k})_{k>0}$ is ergodic is the same as that of Leroux ([49], Lemma 1). Let T be the shift operator, so that if $u = (u_k)_{k\geq 0}$, the sequence Tu is defined by $(Tu)_k = (u)_{k+1}$ for any $k \geq 0$. Let B be an event which is T-invariant. We need to prove that $\mathbb{P}_0(W^{s,t} \in B)$ equals 0 or 1. For any integer n, there exists a cylinder set B_n, depending only on the coordinates u_k with $-m_n \leq k \leq m_n$ for some sub-sequence m_n, such that $\mathbb{P}_0(W^{s,t} \in B \Delta B_{m_n}) \leq 1/2^n$. Here, Δ denotes the symmetric difference between sets. Since $W^{s,t}$ is stationary and B is T-invariant :

$$\mathbb{P}_0\left(W^{s,t} \in B\Delta B_{m_n}\right) = \mathbb{P}_0\left(T^{2m_n}W^{s,t} \in B\Delta B_{m_n}\right) = \mathbb{P}_0\left(W^{s,t} \in B\Delta T^{-2m_n}B_{m_n}\right).$$

Let $\tilde{B} = \cap_{n\geq 1} \cup_{j\geq n} T^{-2m_j}B_{m_j}$. Borel-Cantelli's Lemma leads to $\mathbb{P}_0(W^{s,t} \in B\Delta\tilde{B}) = 0$, so that $\mathbb{P}_0(W^{s,t} \in B) = \mathbb{P}_0(W^{s,t} \in \tilde{B}) = \mathbb{P}_0(W^{s,t} \in B \cap \tilde{B})$. Now, conditional on $(\varepsilon_t)_{t\in\mathbb{N}}$, the random variables $(W_{s+k,t+k})_{k>0}$ are strongly mixing, since $W_{s+k,t+k}$ only depends on a finite number of other $W_{s+l,t+l}$, $l > 0$. Then the $0-1$ law implies (see [73]) that for any fixed sequence e with values in $\mathcal{E}_\infty$, the probability $\mathbb{P}_0(W^{s,t} \in \tilde{B}|(\varepsilon_t)_t = e)$ equals 0 or 1, so that

$$\mathbb{P}_0\left(W^{s,t} \in \tilde{B}\right) = P\left((\varepsilon_t)_t \in C\right)$$

where C is the set of sequences e such that $P(W^{s,t} \in \tilde{B}|(\varepsilon_t)_t = e) = 1$. But it is easy to see that C is T-invariant. Indeed, if $e \in C$ then, since $W^{s,t}$ is stationary and $\tilde{B}$ invariant,

$$1 = \mathbb{P}_0(W^{s,t} \in \tilde{B}|(\varepsilon_t)_t = e) = \mathbb{P}_0(TW^{s,t} \in \tilde{B}|(\varepsilon_t)_t = Te) = \mathbb{P}_0(W^{s,t} \in \tilde{B}|(\varepsilon_t)_t = Te)$$

so that $Te \in C$. Now, since a stationary irreducible Markov chain is ergodic, $\mathbb{P}_0\left((\varepsilon_t)_t \in C\right)$ equals 0 or 1. This concludes the proof of ergodicity of the sequence $W^{s,t}$.
To end with, note that for any $t \geq 0$, the random variable $W_{0,t}$ is non positive, ensuring the convergence of $\{t^{-1}W_{0,t}\}$ in $\mathbb{L}_1$. □

3.3.2 Divergence properties of Information divergence rates

Information divergence rates should be non negative : this is proved below. They also should be positive for parameters that are different than the true one : we only prove it for some subsets of the parameter set. We thus define Θ_{exp} as the subset of Θ_0 such that the expectations of ε_1 under θ and under θ_0 are not aligned with $(0,0)$:

$$\Theta_{exp} = \left\{\theta \in \Theta_0 \ : \ \forall\lambda > 0, \mathbb{E}_\theta(\varepsilon_1) \neq \lambda\mathbb{E}_0(\varepsilon_1)\right\}.$$

Θ_{marg} is the subset of Θ_0 such that Assumption 3 holds :

$$\Theta_{marg} = \left\{\theta \in \Theta_0 \ : \ h_X = f, \ h_Y = g\right\}.$$

Theorem 2 *Information divergence rates satisfy :*

- *For all* $\theta \in \Theta_0$, $D(\theta|\theta_0) \geq 0$ *and* $D^*(\theta|\theta_0) \geq 0$.
- *For any* $\theta \in \Theta_{exp}$, $\theta \neq \theta_0$, *we have* $D(\theta|\theta_0) > 0$ *and* $D^*(\theta|\theta_0) > 0$.
- *If* θ_0 *and* θ *are in* Θ_{marg}, $D(\theta|\theta_0) > 0$ *and* $D^*(\theta|\theta_0) > 0$ *as soon as* $f \neq f_0$ *or* $g \neq g_0$.

Notice that in case Assumption 2 holds, the expectations of ε_1 under θ and under θ_0 are aligned with $(0,0)$. In this case, we were not able to prove that $h \neq h_0$ implies positivity of information divergence rates.

Proof. Since for all t,

$$\mathbb{E}_0\left(\log \mathbb{P}_0(X_{1:N_t}, Y_{1:M_t})\right) - \mathbb{E}_0\left(\log \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t})\right)$$

is a Kullback-Leibler divergence, it is non negative, and the limit $D^*(\theta|\theta_0)$ is also non negative.

Let us prove that $D(\theta|\theta_0)$ is also non negative. To compute the value of the expectation $\mathbb{E}_0[w_t(\theta)]$, introduce the set A_t of all possible values of Z_t :

$$A_t = \left\{(n,m) \in \mathbb{N}^2 \ : \ n \vee m \leq t \leq n+m\right\}.$$

Then,

$$\mathbb{E}_0[w_t(\theta)] = \sum_{(n,m)\in A_t} \sum_{x_{1:n}, y_{1:m}} \mathbb{P}_0(Z_t = (n,m), X_{1:n} = x_{1:n}, Y_{1:m} = y_{1:m}) \log Q_\theta(x_{1:n}, y_{1:m}).$$

Now, by definition,

$$D\left(\theta|\theta_0\right) = \lim_{t\to+\infty} \frac{1}{t}\mathbb{E}_0\left(\log \frac{Q_{\theta_0}(X_{1:N_t}, Y_{1:M_t})}{Q_\theta(X_{1:N_t}, Y_{1:M_t})}\right).$$

By using Jensen's inequality,

$$\mathbb{E}_0\left(\log \frac{Q_\theta(X_{1:N_t}, Y_{1:M_t})}{Q_{\theta_0}(X_{1:N_t}, Y_{1:M_t})}\right) \leq \log \mathbb{E}_0\left(\frac{Q_\theta(X_{1:N_t}, Y_{1:M_t})}{Q_{\theta_0}(X_{1:N_t}, Y_{1:M_t})}\right).$$

But

$$\begin{aligned}
&\mathbb{E}_0\left(\frac{Q_\theta(X_{1:N_t}, Y_{1:M_t})}{Q_{\theta_0}(X_{1:N_t}, Y_{1:M_t})}\right) \\
&\quad = \sum_{(n,m)\in A_t} \sum_{x_{1:n}, y_{1:m}} \mathbb{P}_0(Z_t = (n,m), X_{1:n} = x_{1:n}, Y_{1:m} = y_{1:m}) \times \frac{Q_\theta(x_{1:n}, y_{1:m})}{Q_{\theta_0}(x_{1:n}, y_{1:m})} \\
&\quad \overset{(a)}{\leq} \sum_{(n,m)\in A_t} \sum_{x_{1:n}, y_{1:m}} \mathbb{P}_\theta(\exists s \geq 1, Z_s = (n,m), X_{1:n} = x_{1:n}, Y_{1:m} = y_{1:m}) \\
&\quad = \sum_{(n,m)\in A_t} \mathbb{P}_\theta(\exists s \geq 1, Z_s = (n,m)),
\end{aligned}$$

where (a) comes from expression (3.4). Finally,

$$\lim_{t\to+\infty}\frac{1}{t}\left(w_t(\theta)-w_t(\theta_0)\right)\leq\liminf_{t\to+\infty}\frac{1}{t}\log\left[\sum_{(n,m)\in A_t}\mathbb{P}_\theta\left(\exists s\geq 1, Z_s=(n,m)\right)\right].$$

But the cardinality of A_t is at most t^2, so that

$$\lim_{t\to+\infty}\frac{1}{t}\left(w_t(\theta)-w_t(\theta_0)\right)\leq\liminf_{t\to+\infty}\frac{1}{t}\log t^2=0,$$

and

$$\forall\theta\in\Theta_0,\ D(\theta|\theta_0)\geq 0.$$

Since $\theta\in\Theta_0$, there exists δ_θ such that $\theta\in\Theta_{\delta_\theta}$. By using (3.4), one gets the lower bound

$$Q_\theta(x_{1:n},y_{1:m})\geq\delta_\theta^{n+m}\inf_{e\in\mathcal{E}_{n,m}}\left[\mathbb{P}_\theta\left(\varepsilon_{1:|e|}=e\right)\right].$$

Since trajectories e in $\mathcal{E}_{n,m}$ have length at most $n+m$,

$$\inf_{e\in\mathcal{E}_{n,m}}\left[\mathbb{P}_\theta\left(\varepsilon_{1:|e|}=e\right)\right]\geq\delta_\theta^{n+m}.$$

Note also that if (n,m) belongs to A_t then we have $n+m\leq 2t$ and $n\vee m\geq t/2$. Thus, uniformly with respect to $(n,m)\in A_t$ and to $x_{1:n}$ and $y_{1:m}$,

$$4t\log\delta_\theta\leq\log Q_\theta(x_{1:n},y_{1:m})\leq 0. \tag{3.5}$$

Moreover, with

$$\rho_\theta=\|f\|_\infty\vee\|g\|_\infty\vee\|h\|_\infty\leq 1-\delta_\theta<1$$

one has for any integers n, m, any $x_{1:n}$ and $y_{1:m}$

$$Q_\theta(x_{1:n},y_{1:m})\leq\rho_\theta^{n\vee m}.$$

In this case, for all t, and uniformly with respect to $(n,m)\in A_t$ and to $x_{1:n}$ and $y_{1:m}$,

$$\log Q_\theta(x_{1:n},y_{1:m})\leq\frac{t}{2}\log(1-\delta_\theta). \tag{3.6}$$

Inequalities (3.5) and (3.6) allow to conclude that for some positive numbers c_θ and C_θ,

$$-C_{\theta_0}\leq w(\theta_0)\leq -c_{\theta_0}\quad\text{and}\quad -C_\theta\leq w(\theta)\leq -c_\theta.$$

Then, as soon as B_t is a set such that

$$\lim_{t\to+\infty}\mathbb{P}_0\left(Z_t\notin B_t\right)=0, \tag{3.7}$$

we have

$$D\left(\theta|\theta_0\right) = \lim_{t\to+\infty} \frac{1}{t}\mathbb{E}_0\left[\left(\log\frac{Q_{\theta_0}(X_{1:N_t},Y_{1:M_t})}{Q_{\theta}(X_{1:N_t},Y_{1:M_t})}\right)\mathbb{1}\{Z_t\in B_t\}\right].$$

Now, using Jensen's inequality,

$$\mathbb{E}_0\left[\left(\log\frac{Q_{\theta}(X_{1:N_t},Y_{1:M_t})}{Q_{\theta_0}(X_{1:N_t},Y_{1:M_t})}\right)\mathbb{1}\{Z_t\in B_t\}\right] \leq \mathbb{P}_0\left(Z_t\in B_t\right)\log\mathbb{E}_0\left(\frac{Q_{\theta}(X_{1:N_t},Y_{1:M_t})}{Q_{\theta_0}(X_{1:N_t},Y_{1:M_t})}\Big|Z_t\in B_t\right).$$

But as previously seen,

$$\begin{aligned}
&\mathbb{E}_0\left(\frac{Q_{\theta}(X_{1:N_t},Y_{1:M_t})}{Q_{\theta_0}(X_{1:N_t},Y_{1:M_t})}|Z_t\in B_t\right)\\
=&\sum_{(n,m)\in B_t}\sum_{x_{1:n},y_{1:m}}\frac{\mathbb{P}_0(Z_t=(n,m),X_{1:n}=x_{1:n},Y_{1:m}=y_{1:m})}{\mathbb{P}_0(Z_t\in B_t)}\times\frac{Q_\theta(x_{1:n},y_{1:m})}{Q_{\theta_0}(x_{1:n},y_{1:m})}\\
\leq&\sum_{(n,m)\in B_t}\frac{\mathbb{P}_\theta(\exists s\geq 1, Z_s=(n,m))}{\mathbb{P}_0(Z_t\in B_t)}.
\end{aligned}$$

Finally,

$$-D\left(\theta|\theta_0\right)\leq\lim_{t\to+\infty}\frac{1}{t}\log\mathbb{P}_\theta(\exists s\geq 1, Z_s\in B_t). \tag{3.8}$$

Let us now consider the case where the expectations of ε_1 under parameters θ and θ_0 are not aligned with $(0,0)$, that is $\theta\in\Theta_{exp}$. We have

$$\eta=\inf_{\lambda\in\mathbb{R}}\|\mathbb{E}_\theta(\varepsilon_1)-\lambda\mathbb{E}_0(\varepsilon_1)\|>0,$$

where $\|\cdot\|$ denotes the euclidean norm. Define

$$B_t=\left\{(n,m)\in A_t\ :\ \left\|\frac{(n,m)}{t}-\mathbb{E}_0(\varepsilon_1)\right\|\leq\frac{\eta}{4}\right\}.$$

Then, (3.7) holds. Any trajectory e ending at point (n,m) has length at least $n\vee m$ which is at least $t/2$ when $(n,m)\in B_t$. Thus for such (n,m) :

$$\begin{aligned}
\mathbb{P}_\theta\left(\exists s\geq 1, Z_s=(n,m)\right)&\leq\mathbb{P}_\theta\left(\exists s\geq\frac{t}{2},\inf_{\lambda\in\mathbb{R}}\left\|\frac{Z_s}{s}-\lambda\mathbb{E}_0(\varepsilon_1)\right\|\leq\frac{t}{s}\left\|\frac{Z_s}{t}-\mathbb{E}_0(\varepsilon_1)\right\|\right)\\
&\leq\mathbb{P}_\theta\left(\exists s\geq\frac{t}{2},\inf_{\lambda\in\mathbb{R}}\left\|\frac{Z_s}{s}-\lambda\mathbb{E}_0(\varepsilon_1)\right\|\leq\frac{\eta}{2}\right)\leq\mathbb{P}_\theta\left(\exists s\geq\frac{t}{2},\left\|\frac{Z_s}{s}-\mathbb{E}_\theta(\varepsilon_1)\right\|\geq\frac{\eta}{2}\right).
\end{aligned}$$

Now, using easy Cramer-Chernoff bounds, since π is irreducible, one has that there exists a positive $c(\eta)$ and some $s_0>0$ such that as soon as $s\geq s_0$,

$$\mathbb{P}_\theta\left(\left\|\frac{Z_s}{s}-\mathbb{E}_\theta(\varepsilon_1)\right\|\geq\frac{\eta}{2}\right)\leq\exp\left(-sc(\eta)\right),$$

and by summing over s, there also exists a positive C such that for large enough t,

$$\mathbb{P}_\theta\left(\exists s \geq \frac{t}{2} \ : \ \left\|\frac{Z_s}{s} - \mathbb{E}_\theta(\varepsilon_1)\right\| \geq \frac{\eta}{2}\right) \leq C\exp\left(-tc(\eta)/2\right).$$

Thus, using (3.8), one obtains that for $\theta \in \Theta_{exp}$:

$$D(\theta|\theta_0) \geq \frac{c(\eta)}{2} > 0.$$

Let us now consider the case where θ_0 and θ are in Θ_{marg}. Then, using Jensen's Inequality and definition (3.4),

$$\begin{aligned}
&\mathbb{E}_0\left(\log\frac{Q_\theta(X_{1:N_t}, Y_{1:M_t})}{Q_{\theta_0}(X_{1:N_t}, Y_{1:M_t})}\right)\\
= &\sum_{(n,m)\in A_t}\sum_{x_{1:n}}\sum_{y_{1:m}} \mathbb{P}_0(Z_t=(n,m), X_{1:n}=x_{1:n}, Y_{1:m}=y_{1:m})\log\frac{Q_\theta(x_{1:n},y_{1:m})}{Q_{\theta_0}(x_{1:n},y_{1:m})}\\
\leq &\sum_{(n,m)\in A_t}\sum_{x_{1:n}} \mathbb{P}_0(Z_t=(n,m), X_{1:n}=x_{1:n})\\
&\log\left(\sum_{y_{1:m}}\frac{\mathbb{P}_0(Z_t=(n,m), X_{1:n}=x_{1:n}, Y_{1:m}=y_{1:m})Q_\theta(x_{1:n},y_{1:m})}{\mathbb{P}_0(Z_t=(n,m), X_{1:n}=x_{1:n})Q_{\theta_0}(x_{1:n},y_{1:m})}\right)\\
\leq &\sum_{(n,m)\in A_t}\sum_{x_{1:n}} \mathbb{P}_0(Z_t=(n,m))f_0^{\otimes n}(x_{1:n})\log\left(\frac{\mathbb{P}_\theta(\exists s\geq 1, Z_s=(n,m))f^{\otimes n}(x_{1:n})}{\mathbb{P}_0(Z_t=(n,m))f_0^{\otimes n}(x_{1:n})}\right),
\end{aligned}$$

where the last inequality comes from Lemma 2 and the fact that $\mathbb{P}_0(Z_t=(n,m), X_{1:n}=x_{1:n}, Y_{1:m}=y_{1:m}) \leq Q_{\theta_0}(x_{1:n},y_{1:m})$.
Thus, since $t^{-1}N_t$ tends to $(1-p)$, $\mathbb{P}_0$-a.s. as t tends to infinity, and $(1-p)>0$ since $\theta\in\Theta_0$, we have

$$\begin{aligned}
-D(\theta|\theta_0) \leq \limsup_{t\to+\infty}\frac{1}{t}\sum_{(n,m)\in A_t, n\geq\frac{(1-p)}{2}t}&\mathbb{P}_0(Z_t=(n,m))\Big\{\log\frac{\mathbb{P}_\theta(\exists s\geq 1, Z_s=(n,m))}{\mathbb{P}_0(Z_t=(n,m))}\\
&+\frac{(1-p)}{2}t\sum_x f_0(x)\log\frac{f(x)}{f_0(x)}\Big\} \leq \frac{(1-p)}{2}\sum_x f_0(x)\log\frac{f(x)}{f_0(x)} < 0,
\end{aligned}$$

as soon as $f\neq f_0$. A similar proof applies if $g\neq g_0$.
Proofs of divergence properties for D^* follow the same lines. □

3.3.3 Continuity properties

On Θ_δ, the log-likelihoods are uniformly equicontinuous, with a modulus of continuity that does not depend on trajectories, as appears in the proof of the following Lemma.

Lemma 3 *The families of functions* $\{t^{-1}w_t(\theta)\}_{t\geq 1}$ *and* $\{t^{-1}\ell_t(\theta)\}_{t\geq 1}$ *are uniformly equicontinuous on* Θ_δ.

A consequence of this Lemma and the compactness of Θ_δ is :

Corollary 1 *The following holds :*

i) $\{t^{-1}w_t(\theta)\}_t$ *(resp.* $\{t^{-1}\ell_t(\theta)\}_t$*) converges* $\mathbb{P}_0$*-almost surely to* $w(\theta)$ *(resp. to* $\ell(\theta)$*) uniformly on* Θ_δ *;*

ii) $\ell(\theta)$ *and* $w(\theta)$ *are uniformly continuous on* Θ_δ.

Proof of Lemma 3. Let $\alpha > 0$, and $\theta_1, \theta_2 \in \Theta_\delta$ such that $\|\theta_1 - \theta_2\|_\infty \leq \alpha$.
Let us denote μ_{θ_i}, π_{θ_i}, f_{θ_i}, g_{θ_i} and h_{θ_i} the parameters of the hidden Markov chain and of the emission distributions under θ_i, $i = 1, 2$.
For any $e \in \mathcal{E}_{N_t,M_t}$:

$$\begin{aligned}
&\frac{1}{t}\left|\log \mathbb{P}_{\theta_1}(\varepsilon_{1:|e|}=e, X_{1:N_t}, Y_{1:M_t}) - \log \mathbb{P}_{\theta_2}(\varepsilon_{1:|e|}=e, X_{1:N_t}, Y_{1:M_t})\right| \\
&\leq \frac{1}{t}\left|\log \mu_{\theta_1}(e_1) - \log \mu_{\theta_2}(e_1)\right| + \frac{1}{t}\sum_{k,l\in\mathcal{E}}\Big(\sum_{i=2}^{|e|}\mathbb{1}\{e_{i-1}=k, e_i=l\}\Big)\left|\log \pi_{\theta_1}(k,l) - \log \pi_{\theta_2}(k,l)\right| \\
&+ \frac{1}{t}\sum_{a\in\mathcal{A}}\Bigg\{\Big(\sum_{i=1}^{|e|}\mathbb{1}\{e_i=(1,0), X_{N_i}=a\}\Big)\left|\log f_{\theta_1}(a) - \log f_{\theta_2}(a)\right| \\
&\qquad\qquad + \Big(\sum_{i=1}^{|e|}\mathbb{1}\{e_i=(0,1), Y_{M_i}=a\}\Big)\left|\log\, g_{\theta_1}(a) - \log\, g_{\theta_2}(a)\right|\Bigg\} \\
&\quad + \frac{1}{t}\sum_{a,a'\in\mathcal{A}}\Big(\sum_{i=1}^{|e|}\mathbb{1}\{e_i=(1,1), X_{N_i}=a, Y_{M_i}=a'\}\Big)\left|\log h_{\theta_1}(a,a') - \log h_{\theta_2}(a,a')\right|.
\end{aligned}$$

In this sum, at most $2|e|$ terms are non null. Since all the components of θ_i, $i = 1, 2$ are bounded below by δ and $\|\theta_1 - \theta_2\|_\infty \leq \alpha$, we have :

$$\frac{1}{t}\left|\log \mathbb{P}_{\theta_1}(\varepsilon_{1:|e|}=e, X_{1:N_t}, Y_{1:M_t}) - \log \mathbb{P}_{\theta_2}(\varepsilon_{1:|e|}=e, X_{1:N_t}, Y_{1:M_t})\right| \leq \frac{2|e|}{t}\frac{\alpha}{\delta}.$$

But for any $e \in \mathcal{E}_{N_t,M_t}$, we have $|e| \leq 2t$, so that

$$\frac{1}{t}\left|\log \mathbb{P}_{\theta_1}(\varepsilon_{1:|e|}=e, X_{1:N_t}, Y_{1:M_t}) - \log \mathbb{P}_{\theta_2}(\varepsilon_{1:|e|}=e, X_{1:N_t}, Y_{1:M_t})\right| \leq \frac{4\alpha}{\delta},$$

as soon as $\|\theta_1 - \theta_2\|_\infty \leq \alpha$.
Now we get

$$Q_{\theta_1}(X_{1:N_t}, Y_{1:M_t}) = \sum_{e \in \mathcal{E}_{N_t,M_t}} \mathbb{P}_{\theta_1}(\varepsilon_{1:|e|} = e, X_{1:N_t}, Y_{1:M_t})$$
$$\leq \exp\left\{\frac{4\alpha}{\delta}t\right\} \sum_{e \in \mathcal{E}_{N_t,M_t}} \mathbb{P}_{\theta_2}(\varepsilon_{1:|e|} = e, X_{1:N_t}, Y_{1:M_t}) \leq \exp\left\{\frac{4\alpha}{\delta}t\right\} Q_{\theta_2}(X_{1:N_t}, Y_{1:M_t}),$$

and $t^{-1}\log Q_{\theta_1}(X_{1:N_t}, Y_{1:M_t}) \leq 4\alpha/\delta + t^{-1}\log Q_{\theta_2}(X_{1:N_t}, Y_{1:M_t})$. Since this is symmetric in θ_1 and θ_2, one obtains that for any $\theta_1, \theta_2 \in \Theta_\delta$ such that $\|\theta_1 - \theta_2\|_\infty \leq \alpha$,

$$|\frac{1}{t}w_t(\theta_1) - \frac{1}{t}w_t(\theta_2)| \leq \frac{4\alpha}{\delta}.$$

The same proof applies to $t^{-1}\ell_t$. □

3.4 Statistical properties of estimators

We now want to focus on a particular form of the pair-HMM, relying on a re-parametrization of the model. Indeed, the pair-HMM has been introduced to take into account evolutionary events. The corresponding evolutionary parameters are the ones of interest and practitioners aim at estimating those parameters rather than the full pair-HMM. Examples of such re-parametrization may be found for instance in [76, 77] (see also Section 5 of this paper). Let $\beta \mapsto \theta(\beta)$ be a continuous parametrization from some set B to Θ. For any $\delta > 0$, let $B_\delta = \theta^{-1}(\Theta_\delta)$. We assume that $\beta_0 = \theta^{-1}(\theta_0)$ in B_δ for some $\delta > 0$. Use of pair-HMM algorithms to estimate evolutionary parameters corresponds to the estimator

Definition 2 $\widehat{\beta}_t = \mathit{Argmax}_{\beta \in B_\delta} w_t(\theta(\beta))$.

Then,

Theorem 3 *If the set of maximizers of $w(\theta(\beta))$ over B_δ reduces to $\{\beta_0\}$, $\widehat{\beta}_t$ converges $\mathbb{P}_0$-almost surely to β_0.*

The proof of this theorem follows from Corollary 1 and usual arguments for M-estimators. The condition that the set of maximizers of $w(\theta(\beta))$ over B_δ reduces to $\{\beta_0\}$ corresponds to some identifiability condition and thus may not be avoided.

Another interesting approach to sequence alignment by pair-HMMs is to consider a non-informative prior distribution on the parameters to produce, via a MCMC procedure, the posterior distribution of the alignments and parameters given the observed sequences.

Using Q_θ as the likelihood of the observed sequences produces a posterior distribution as follows. Let ν be a prior probability measure on B_δ. MCMC algorithms approximate the random distribution $\nu_{|X_{1:N_t},Y_{1:M_t}}$ interpreted as the posterior measure given observations $X_{1:N_t}$ and $Y_{1:M_t}$:

$$\nu_{|X_{1:N_t},Y_{1:M_t}}(d\beta) = \frac{Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)}{\int_{B_\delta} Q_{\theta(\beta')}(X_{1:N_t},Y_{1:M_t})\nu(d\beta')}.$$

This leads to Bayesian consistent estimation of β_0 as in classical statistical models (see [37], for instance). Notice that since w_t is not the logarithm of a probability distribution on the observation space, these results are not direct consequences of classical ones. Though, the proof follows classical ideas of Bayesian theory.

Theorem 4 *If the set of maximizers of $w(\theta(\beta))$ over B_δ reduces to $\{\beta_0\}$, and if ν weights β_0, then the sequence of posterior measures $\nu_{|X_{1:N_t},Y_{1:M_t}}$ converges in distribution $\mathbb{P}_0$-almost surely to the Dirac mass at β_0.*

Proof. Let $m : B_\delta \to \mathbb{R}$ be any continuous, bounded function. For any $\epsilon > 0$, let α such that $|m(\beta) - m(\beta')| \leq \epsilon$ as soon as $\|\beta - \beta'\| \leq \alpha$. We have

$$\begin{aligned}\left|\int_{B_\delta} m(\beta)\nu_{|X_{1:N_t},Y_{1:M_t}}(d\beta) - m(\beta_0)\right| &\leq \int_{B_\delta} |m(\beta) - m(\beta_0)|\ \nu_{|X_{1:N_t},Y_{1:M_t}}(d\beta) \\ &\leq \epsilon + 2\|m\|_\infty \int_{\|\beta-\beta_0\|>\alpha} \nu_{|X_{1:N_t},Y_{1:M_t}}(d\beta) \\ &= \epsilon + 2\|m\|_\infty \frac{\int_{\|\beta-\beta_0\|>\alpha} \exp\left\{t\left(\frac{1}{t}w_t(\theta(\beta))\right)\right\}\nu(d\beta)}{\int_{B_\delta} \exp\left\{t\left(\frac{1}{t}w_t(\theta(\beta))\right)\right\}\nu(d\beta)}.\end{aligned}$$

Use of Corollary 1 and the fact that the set of maximizers of $w(\theta(\beta))$ over B_δ reduces to $\{\beta_0\}$ gives $\eta > 0$ and T such that for $t > T$ and $\|\beta - \beta_0\| > \alpha$, $t^{-1}w_t(\theta(\beta)) - t^{-1}w_t(\theta(\beta_0)) \leq -\eta$, and then there exists $\gamma > 0$ such that for $t > T$ and $\|\beta - \beta_0\| \leq \gamma$, $t^{-1}w_t(\theta(\beta)) - t^{-1}w_t(\theta(\beta_0)) \geq -\frac{\eta}{2}$. Then

$$\begin{aligned}&\frac{\int_{\|\beta-\beta_0\|>\alpha} \exp\left\{t\left(\frac{1}{t}w_t(\theta(\beta))\right)\right\}\nu(d\beta)}{\int_{B_\delta} \exp\left\{t\left(\frac{1}{t}w_t(\theta(\beta))\right)\right\}\nu(d\beta)} \\ &\leq \frac{\int_{\|\beta-\beta_0\|>\alpha} \exp\left\{t\left(\frac{1}{t}w_t(\theta(\beta)) - \frac{1}{t}w_t(\theta(\beta_0))\right)\right\}\nu(d\beta)}{\int_{\|\beta-\beta_0\|\leq\gamma} \exp\left\{t\left(\frac{1}{t}w_t(\theta(\beta)) - \frac{1}{t}w_t(\theta(\beta_0))\right)\right\}\nu(d\beta)} \leq \left(\exp\left\{-t\frac{\eta}{2}\right\}\right)\frac{\int_{\|\beta-\beta_0\|>\alpha}\nu(d\beta)}{\int_{\|\beta-\beta_0\|\leq\gamma}\nu(d\beta)}.\end{aligned}$$

Using that ν weights β_0 we finally obtain

$$\lim_{t\to\infty}\left|\int_{B_\delta} m(\beta)\ \nu_{|X_{1:N_t},Y_{1:M_t}}(d\beta) - m(\beta_0)\right| = 0 \quad \mathbb{P}_0 - a.s. \tag{3.9}$$

But it exists a countable collection of continuous and bounded functions that are determining for convergence in distribution and the union of the corresponding null sets in which (3.9) does not hold is still a null set. Then

$$\nu_{|X_{1:N_t},Y_{1:M_t}} \rightsquigarrow \delta_{\beta_0} \quad \mathbb{P}_0 - a.s.$$

□

3.5 Simulations

3.5.1 A simple model

For the whole simulation procedure we consider the following substitution model :

$$h(x,y) = \begin{cases} f(x)(1-e^{-\alpha})f(y) & \text{if } x \neq y \\ f(x)\{(1-e^{-\alpha})f(x) + e^{-\alpha}\} & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ is called the substitution rate and for every letter x, $f(x)$ equals the equilibrium probability of x. This equilibrium probability distribution is assumed to be known and will not be part of the parameter. Here, the emission distribution g equals f, and Assumption 3 holds. The unknown parameter is thus $\beta = (\pi, \alpha)$. This is a classical substitution model known as Felsenstein81 [21] for which the substitution rate is independent of the type of nucleotide being replaced and $1 - e^{-\alpha}$ represents the probability that a substitution occurs. We shall consider hidden Markov chains that satisfy Assumption 2, and will present :

- Simulations with i.i.d. $(\varepsilon_s)_s$ where probabilities of horizontal or vertical moves equal p_0 and probability of diagonal moves equals $r_0 = 1 - 2p_0$. Here, the parameter reduces to $\beta = (p, \alpha)$.
- Simulations with stationary Markov chains such that $p_0 = q_0$. The parameter dimension then reduces to 6 (including α).

Notice that none of these situations is covered by Theorem 2 : we do not know in those cases whether the information divergence rates are positive at a parameter value different from the true one.

In both cases, we get estimations of the parameters via MLE (taking Q_θ as the likelihood as it is done in practice), and in the i.i.d. case we compute and compare the functions w and ℓ.

3.5.2 Simulations with i.i.d. $(\varepsilon_s)_s$

We have simulated 200 alignments of length 15000 with substitution rate $\alpha_0 = 0.05$ and $p_0 = q_0 = 0.25$. We have set the equilibrium probability of every nucleotide to

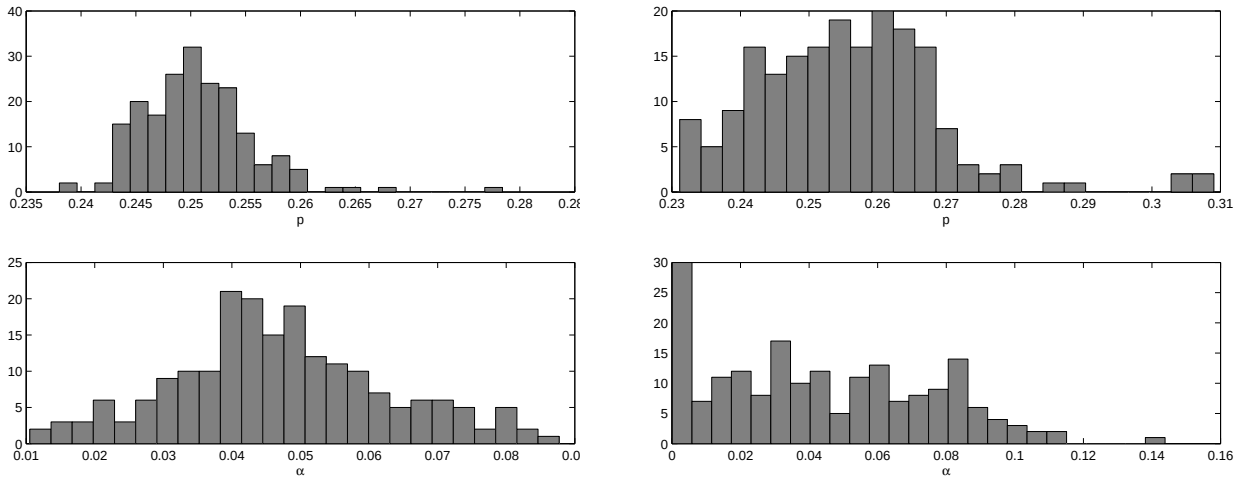


FIG. 3.3 – *Histograms of maximum likelihood estimations of parameters obtained with 200 simulations from the i.i.d. model. On the left : estimation of p given $\alpha = \alpha_0 = 0.05$ and estimation of α given $p = p_0 = 0.25$. On the right : joint estimation of p and α.*

0.25 (in this case the Felsenstein81 substitution model becomes the Jukes-Cantor model [40]). We show in Figure 3.3 histograms for the maximum likelihood estimations of both parameters. In a first part we keep α fixed at α_0 and estimate p and then we keep p fixed at p_0 and estimate α. That produces good estimations of the parameters even if α is a bit underestimated. However when estimating p and α simultaneously (second part) we obtain no satisfying results especially on α (see Figure 3.3).

That can be explained by looking at the graph of $w(\beta)$ and comparing it to $\ell(\beta)$ (Figure 3.4). We see that both w and ℓ are very flat with respect to α and as we deal with numerical precision errors, finding out the true maximum value becomes impossible. However, for $p = p_0$ if we look closely at the cuts of ℓ and w we appreciate that ℓ takes its maximum on α_0 and w near this point. As the maximisation problem complexity is reduced in this case we are able to find a quite good estimation for α. Concerning p, we see that both ℓ and w have a clear maximum near p_0, but again ℓ is less flat than w at this point. This is not surprising since ℓ really is the information divergence rate of the model.

3.5.3 Simulations with Markov chains satisfying Assumption 2

We have simulated 200 alignments of length 15000 with substitution rate $\alpha_0 = 0.05$ and the following transition matrix for $(\varepsilon_s)_s$

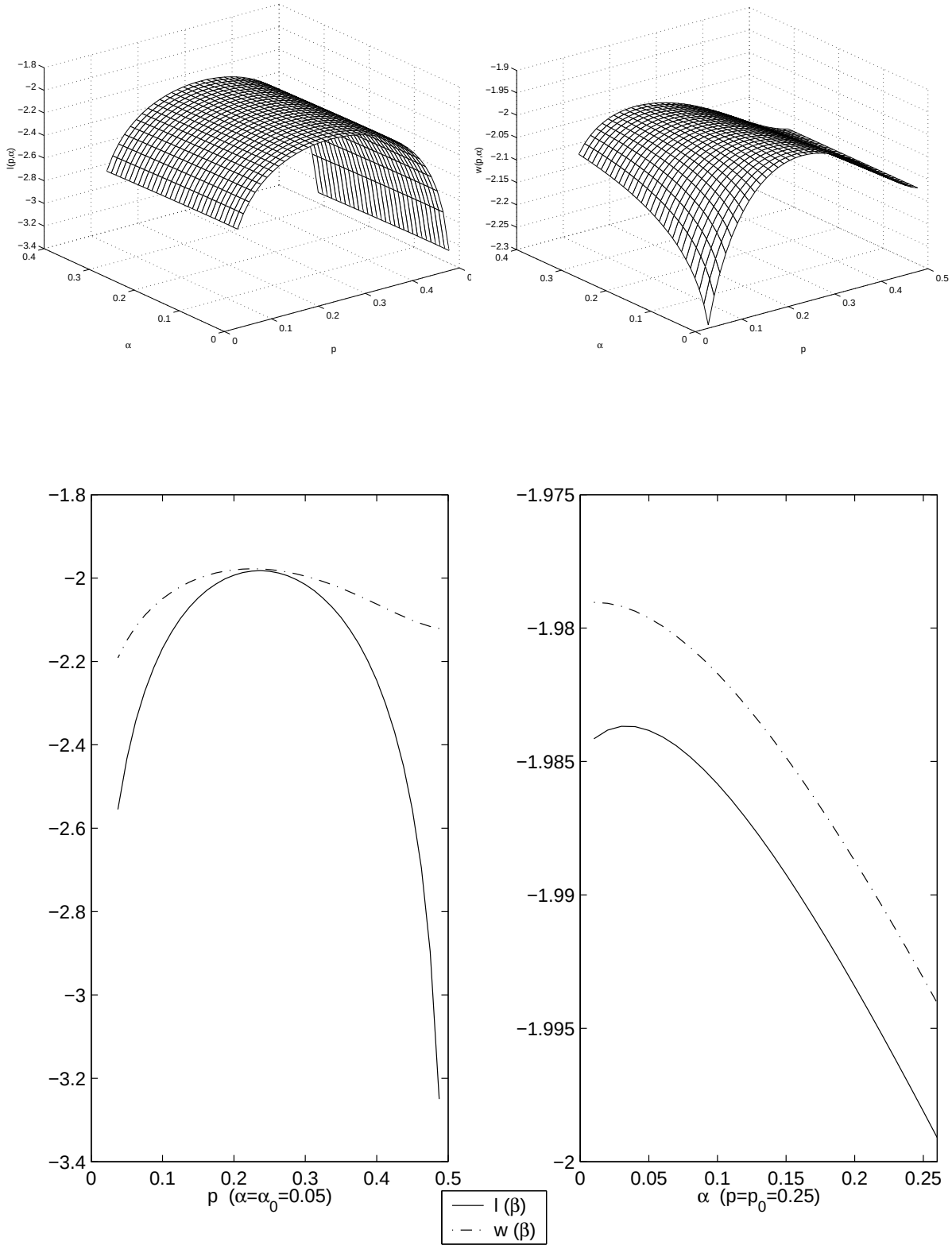


FIG. 3.4 – *On top : ℓ and w for the i.i.d. model ($p_0 = 0.25, \alpha_0 = 0.05$). On bottom : cuts of ℓ and w for $\alpha = \alpha_0$ fixed and for $p = p_0$ fixed.*

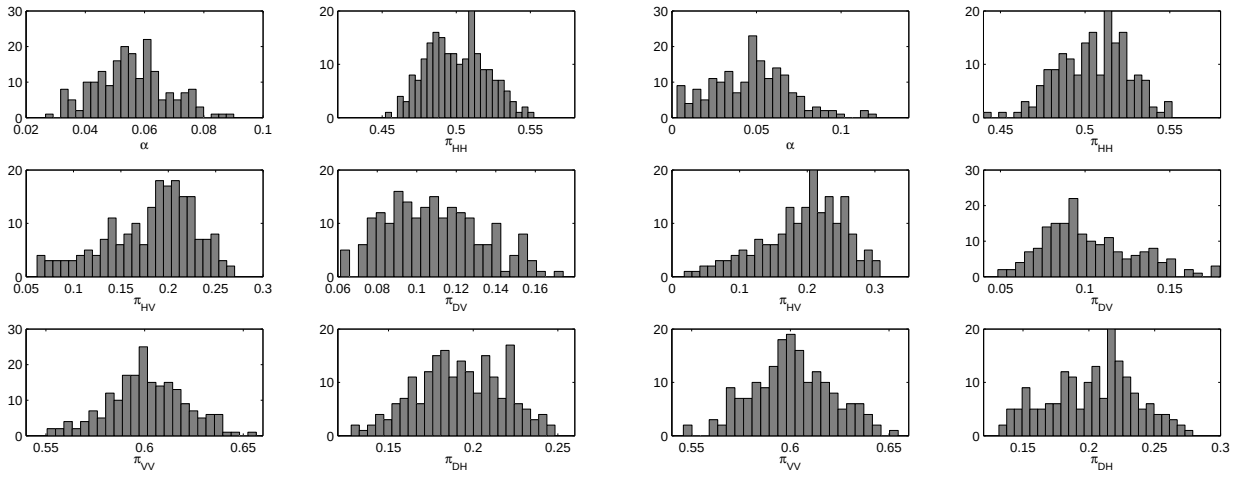


FIG. 3.5 – *Histograms of maximum likelihood estimations of parameters obtained with 200 simulations from the Markov chain model. On the left : estimation of the transition probabilities given $\alpha = \alpha_0$ and estimation of α given the true value of the transition probabilities. On the right : joint estimation of the transition probabilities and α.*

$$\begin{array}{c} \\ D \\ H \\ V \end{array} \begin{array}{c} \begin{array}{ccc} D & H & V \end{array} \\ \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.3 & 0.1 & 0.6 \end{pmatrix} \end{array}$$

with initial distribution $p_0 = q_0 = 0.25$. With this set up there are only five free parameters in the transition matrix. We have set as free parameters π_{HH}, π_{HV}, π_{DV}, π_{VV} and π_{DH}. The equilibrium probability of every nucleotide is again fixed to 0.25. We can observe in Figure 3.5 that the maximum likelihood estimators for these parameters and for α are close to their true values even when the estimation is done jointly.

3.6 Discussion

Our first contribution is to provide a rigorous probabilistic and statistical background to the study of pair hidden Markov models. This background is at the core of theoretical studies on these models and it is also a first step towards other biological models, such as those used in the context of multiple alignments. Our main results are given in Theorems 1 and 2, where we first prove convergence of normalized log-likelihoods and identify cases where a divergence property holds. Unfortunately, despite the positive results that we

obtain, it is not yet possible to validate pair-HMM algorithms in every situation. From a theoretical point of view, we were not able to prove that, under Assumption 2 (namely, $p = q$ and thus the parameter does not belong to Θ_{exp}), divergence property holds in case $h \neq h_0$. Consequences in terms of evolutionary parameters (in some particular re-parametrization schemes) remains a challenging issue. Simulation studies investigate situations in which Theorem 2 does not hold. Despite the fact that the i.i.d. framework works poorly, the Markov one seems to give satisfying results. These results are rather encouraging since the Markov case is the interesting one in biological applications.

Chapitre 4

Parameter Estimation in multiple alignment under the TKF91 evolution model

Abstract

In this chapter we deal with parameter estimation in a multiple alignment model derived from the TKF91 sequence evolution model. We first provide a rigorous formalism for the homology structure of k sequences related by a star tree. In the case of two sequences we compare this model to the pair-HMM. As in the pair-HMM we discuss possible definitions of likelihoods. Existence of two different Information divergence rates is established and a divergence property is shown under additional assumptions. This yields consistency for the parameter in parametrization schemes for which the divergence property holds. Simulations illustrate different cases which are not covered by our results.

Contents

4.1 Introduction

In the multiple alignment problem the observations consist in k ($k > 2$) sequences $X^1_{1:n_1}, ..., X^k_{1:n_k}$, where $X^i_{1:n_i} = X^i_1 \dots X^i_{n_i}$, with values in a finite alphabet $\mathcal{A}$ (for instance $\mathcal{A} = \{A, C, G, T\}$ for DNA sequences). It is assumed that the sequences are related by a phylogenetic tree, that is, a tree where the nodes represent the sequences and the edges represent the evolutionary relationships between them. The observed sequences are placed at the k leaves of the tree, whereas the inner nodes stand for ancestral (non-observable) sequences. The most ancestral sequence is placed at the root, $\mathcal{R}$, of the tree. The choice of the root assigns to each edge a direction (from the root to the leaves) and to each inner node its descendants nodes, but since the evolutionary process between the sequences is usually assumed to be time reversible, the placement of the root node is irrelevant (cf. [74]). A path from the root to a leaf represents the evolution through time and through a series of intermediate sequences of the ancestral sequence, leading to the corresponding observed sequence. The evolution on each edge (from its *parent* node to its *child* node) is described by some evolutionary proccess. We assume that the same evolutionary process works on every edge of the tree. A main hypothesis is that the evolutionary processes working on two edges with the same *parent* node are independent, i.e. a sequence evolves independently to each one of its descendants.

In this paper we consider that sequences evolve according to the TKF91 model of evolution [76]. Let us briefly recall how the TKF91 model works on pairwise alignments. This model is formulated in terms of *links* and associated letters. To each link is associated a letter that undergoes changes, independently of other letters, according to a reversible substitution process. The insertion and deletion process is described by a birth-death process on these links. Indeed, a link and its associated letter is deleted at the rate $\mu > 0$. While a link is present it gives rise to new links at the rate λ. A new link is placed immediately to the right of the link from which it originated, and the associated letter is chosen from the stationary distribution of the substitution process. At the very left of the sequence is a so-called immortal link that never dies and gives rise to new links at the rate λ. We need the death rate per link to exceed the birth rate per link to have a distribution of sequence lengths. Indeed, if $\lambda < \mu$ then the equilibrium distribution of length sequence is geometric with parameter λ/μ.

Let $p^H_n(t)$ be the probability that a normal link survives and has n descendants, including itself, after a time t. Let $p^N_n(t)$ be the probability that a normal link dies but leaves n descendants after a time t. Finally let $p^I_n(t)$ be the probability that an immortal link has n descendants, including itself, after a time t. Here H stands for homologous, N for

non-homologous and I for immortal. We have :

$$\begin{array}{lcll} p_n^H(t) & = & e^{-\mu t}[1-\lambda\beta(t)][\lambda\beta(t)]^{n-1} & for\ n>0 \\ p_n^N(t) & = & \mu\beta(t) & for\ n=0 \\ & = & [1-e^{-\mu t}-\mu\beta(t)][1-\lambda\beta(t)][\lambda\beta(t)]^{n-1} & for\ n>0 \\ p_n^I(t) & = & [1-\lambda\beta(t)][\lambda\beta(t)]^{n-1} & for\ n>0 \end{array} \tag{4.1}$$

where

$$\beta(t) = \frac{1-e^{(\lambda-\mu)t}}{\mu-\lambda e^{(\lambda-\mu)t}}.$$

Conceptually, $e^{-\mu t}$ is the probability of ancestral residue survival, $\lambda\beta(t)$ is the probability of more insertions given one or more existent descendants and $\gamma(t) := \frac{1-e^{-\mu t}-\mu\beta(t)}{1-e^{-\mu t}}$ is the probability of insertion given that the ancestral residue did not survive. See [76] for details.

If we want to investigate the asymptotic properties of parameter estimators we must consider observed sequences of growing lengths. However, this is not possible under the hypothesis of the TKF91 model. Indeed, the ancestral sequence length distribution depends on λ/μ, and so, for a given value of these parameters we can not make the ancestral sequence length to tend to infinity. As one would expect (and as we will show later) the lengths of the observed sequences are equivalent to the length of the root sequence, so under this setup we can not expect to observe infinitely long sequences.
Following the ideas in the FID model of Metzler ([55]) we will consider the case in which the TKF91 model can produce long sequences, that is, the case where $\lambda = \mu$. With this configuration finite length sequences are to be considered as cut out of very much longer sequences between known homologous positions. The length of the ancestral sequence is now considered to be non random.
We will note $q_n^H(t)$ and $q_n^N(t)$ the probability distributions of the number of descendants for a normal *link* under these assumptions. We do not need to consider the distribution for the immortal *link* anymore, since now all the positions on the observed sequences are descendants of normal *links*.
Since $\lim_{\mu\to\lambda}\beta(t) = \frac{t}{1+\lambda t}$ we get

$$\begin{array}{lcll} q_n^H(t) = \lim\limits_{\mu\to\lambda} p_n^H(t) & = & e^{-\lambda t}\dfrac{1}{1+\lambda t}\left(\dfrac{\lambda t}{1+\lambda t}\right)^{n-1} & for\ n>0 \\ q_n^N(t) = \lim\limits_{\mu\to\lambda} p_n^N(t) & = & \dfrac{\lambda t}{1+\lambda t} & for\ n=0 \\ & = & \left(\dfrac{1}{1+\lambda t}-e^{-\lambda t}\right)\dfrac{1}{1+\lambda t}\left(\dfrac{\lambda t}{1+\lambda t}\right)^{n-1} & for\ n>0 \end{array} \tag{4.2}$$

4.1.1 A star tree

Let us now consider a k-star phylogenetic tree, that is, a tree with a root, k leaves and no inner nodes. See Figure 4.1 for an example. We will note t_i, $i = 1, \ldots, k$, the branches lengths, that is the evolutionary time separating each sequence to the root. In this context, an alignment of the k sequences and the root consists in a composition of the k pairwise alignments of the root with any of the observed sequences. This is done as follows. Two characters X_j^i and X_h^l will be aligned in the same column if and only if they are homologous to the same character of the root sequence. So there is a column for each nucleotide at the root containing all its homologous positions on the leaves, and between two columns of this kind, there is one column for each inserted position on the leaves between the two corresponding nucleotide positions at the root. Insertions to the root sequence occur independently on each sequence and we assume that the probability of having two insertions on different sequences at the same time is 0. That is why insertion columns are composed by one nucleotide position in some of the sequences and gaps in all the others. We know

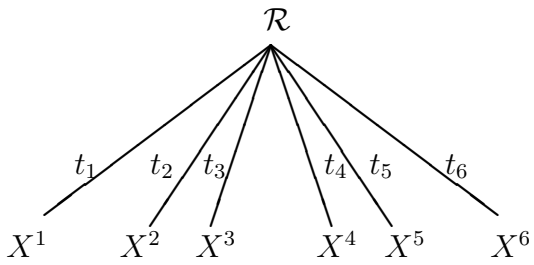


FIG. 4.1 – *A 6-star tree.*

that under the TKF91 model the pairwise alignment is a Markov chain on the state space $\{\binom{\mathtt{B}}{\mathtt{B}}, \binom{-}{\mathtt{B}}, \binom{\mathtt{B}}{-}\}$. However, when we apply it to multiple alignment we do not get a Markov chain on the set of all possible multiple alignment columns. In fact, Markov models for multiple alignment exist but states do not exactly correspond to alignment columns. Indeed, insertion states in these models describe not only an insertion on one sequence but also a kind of "memory" of what is happening in other sequences (see [30] for instance). This is because the Markov dependency for pairwise root-leaf alignments applies independently on each sequence due to the branch independence of the evolutionary process. So that, an alignment column describing an insertion on sequence i depends on the last column of the alignment describing any evolutionary event on sequence i, but there may be between them several alignment columns describing insertions on other sequences. So one could say that insertions to the root sequence break the Markov dependency between alignment columns. Also, the order of the insertions between two homologous positions is irrelevant,

the only important fact being which positions are homologous to which. Then, the interesting objet is not the alignment but the homology structure, essentially an alignment of homologous positions with specification of the number of insertions on each sequence between any two homologous positions.
The homology structure can be described in terms of the nucleotides at the root sequence. Indeed the homology structure is just the sequence of root positions in which we specify, for each ancestral residue, the sequences in which it has survived and all the insertions occurred at its right. We recall that the TKF91 evolution model is defined in terms of links, and evolution on each link is independent of evolution on other links (see [76]). That is why the homology structure can be described as a sequence of i.i.d. random variables. Moreover, due to the branch independence, in the original TKF91 model the probability of an homology structure $\mathcal{H}$ is computed as

$$P_{\lambda<\mu}(\mathcal{H}) = P_{\lambda<\mu}(L_0)\prod_{i=1}^{k} P_{\lambda<\mu}(\mathcal{H}^i|L_0)$$

where $\mathcal{H}^i$ is the homology structure restricted to the root and sequence i, L_0 is the length of the ancestral sequence and $P_{\lambda<\mu}(\mathcal{H}^i, L_0) = P_{\lambda<\mu}(\mathcal{H}^i|L_0)P_{\lambda<\mu}(L_0)$ is just the probability of the homology structure of two sequences, the first one having length L_0, under the TKF91 model. Indeed, when computing the probability of a single pairwise alignment (and the probability of an homology structure is obtained as the sum of the probabilities of several alignments) under the TKF91 as the product of the transition probabilities given in (1.5) there is a factor $\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{L_0}$ corresponding to the length of the first sequence (cf. (1.11) ; see [74] for more details).
In our new set up, as we are conditioning on the ancestral sequence being a cut out of a much longer sequence, we do not take lengths into account and we get

$$P_{\lambda=\mu}(\mathcal{H}) = \prod_{i=1}^{k} P_{\lambda=\mu}(\mathcal{H}^i) \tag{4.3}$$

where in fact $P_{\lambda=\mu}(\mathcal{H}^i) = P_{\lambda<\mu}(\mathcal{H}^i|L_0) = \frac{P_{\lambda<\mu}(\mathcal{H}^i,L_0)}{\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{L_0}}$.

4.2 The homology structure model

4.2.1 Model description

Consider a k-star phylogenetic tree $\mathcal{T}$ with branch lengths $t_1, \ldots, t_k$. The homology structure of the sequences related by $\mathcal{T}$ is a sequence of independent and identically

distributed random variables $\{\varepsilon_n\}_{n\geq 1}$ on

$$\mathcal{E}^k = \left\{(e(1), e(2)) = (\delta^{1:k}, a^{1:k}) \,|\, \delta^i \in \{0,1\},\, a^i \in \mathbb{N}\ i = 1, \ldots, k\right\}.$$

The variable ε_n represents the fate of the nth ancestral sequence character. The first column of ε_n correspond to the homologous positions from this character. If it is conserved in sequence i, $i = 1, \ldots, k$ then $\varepsilon_n^i(1) = 1$, else $\varepsilon_n^i(1) = 0$. It is possible for an ancestral character to have been deleted in all the observed sequences ($\varepsilon_n(1) = 0_{\mathbb{N}^k}$). The second column of ε_n represents the number of insertions on the observed sequences between the nth and the $(n+1)$th ancestral sequence characters. It is possible to have none insertions in any of the observed sequences between two homologous positions. From (4.3), the law of ε_n is given by

$$\mathbb{P}_\lambda\left(\varepsilon_n = (\delta^{1:k}, a^{1:k})\right) = \prod_{i=1}^{k} \left(q^H_{a^i+1}(t_i)\right)^{\mathbb{1}\{\delta^i=1\}} \left(q^N_{a^i}(t_i)\right)^{\mathbb{1}\{\delta^i=0\}}, \quad (\delta^{1:k}, a^{1:k}) \in \mathcal{E}^k,\ n \geq 1. \tag{4.4}$$

The process $\{\varepsilon_n\}_{n\geq 1}$ generates a random walk $\{Z_n\}_{n\geq 1}$ with values on $\mathbb{N}^k$ by letting $Z_0 = 0_{\mathbb{N}^k}$ and $Z_n = \sum_{1\leq i\leq n}\{\varepsilon_i(1) + \varepsilon_i(2)\}$ for $n \geq 1$. The coordinate random variables corresponding to Z_n at position n are denoted by $(Z_n^1, \ldots, Z_n^k)$ (*i.e.* $Z_n = (Z_n^1, \ldots, Z_n^k)$) and represent the length of each observed sequence up to position n on the ancestral sequence.

Let us now describe the emission of the observed sequences. For $n \geq 1$, if $\varepsilon_n = (\delta^{1:k}, a^{1:k})$ then a vector of $r = \sum \delta^i$ r.v. is emitted according to some probability distribution h_J, $J = \{i|\delta^i = 1\}$, on $\mathcal{A}^r$ and $\sum_{i=1}^k a^i$ r.v. $\{X^i_{1:a^i}, a^i \geq 1\}$, $i = 1, \ldots, k$ are drawn (emitted) according to the following scheme : $\{X^i_j\}_{j=1,a^i}^{i=1,k}$ are independent and identically distributed from some probability distribution f on $\mathcal{A}$. Conditionally to the process $\{\varepsilon_n\}_{n\geq 1}$, the random variables emitted at different instants are independent. This model is described by the parameter $\theta = (\lambda, \{h_J\}_{J\subseteq K}, f) \in \Theta$. We use K to denote the set $\{1, \ldots, k\}$. We do not consider the branch lengths as a component of the parameter and assume they are known.

Remark 2 *In practice the substitution process is described by a continuous time Markov chain defined on $\mathcal{A}$ and depending on the branch lengths. Let us note ν the stationary law of this process and $P_t(\cdot,\cdot)$ the transition probability matrix for a transition time $t > 0$. In this case we have :*

$$h_{\{i_1,\ldots,i_l\}}(x^{i_1}, \ldots, x^{i_l}) = \sum_{R\in\mathcal{A}} \nu(R) \prod_{j=1}^{l} P_{t_{i_j}}(R, x^{i_j})$$

and

$$f(\cdot) = \nu(\cdot).$$

That is, h_J does not only depends on the cardinal of J, but also on its elements via the branch lengths $\{t_i\}_{i=1,\dots,k}$.

The conditional distribution of the observations given an homology structure $e_{1:n} = (e_j)_{1\le j\le n} = ((\delta_j^{1:k}, a_j^{1:k}))_{1\le j\le n}$, writes

$$\begin{aligned}
&\mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}|\varepsilon_{1:n} = e_{1:n}, \{\varepsilon_m\}_{m>n}, \{X_{n_i}\}_{i\in K, n_i > Z_n^i}) = \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}|\varepsilon_{1:n} = e_{1:n}) \\
= &\prod_{j=1}^{n} \mathbb{P}_\theta(\mathbb{X}_{Z_{j-1}+1_k:Z_j}|\varepsilon_j = e_j) \\
= &\prod_{j=1}^{n} \left\{ h_{\{i|\delta_j^i=1\}}\big(\{X^i_{Z^i_{j-1}+1}\}_{i|\delta^i_j=1}\big) \prod_{i=1}^{k}\prod_{s=1}^{a_j^i} f\big(X^i_{Z^i_{j-1}+\delta^i_j+s}\big) \right\} \qquad (4.5)
\end{aligned}$$

where 1_k stands for the k-dimensional vector with all components equal to 1 and $\mathbb{X}_{1_k:Z_n} = (X^1_{1:Z^1_n}, \dots, X^k_{1:Z^k_n})$. This notation can be confusing since it is possible to have $Z^i_{j-1} + 1_k > Z^i_j$ for some $i \in K$ and for some $j \ge 1$. However when writing $\mathbb{X}_{Z_{j-1}+1_k:Z_j}$ we will only be considering the variables corresponding to those sequences $i \in K$ for which $Z^i_{j-1}+1_k \le Z^i_j$. The complete distribution $\mathbb{P}_\theta$ is given by

$$\mathbb{P}_\theta(\varepsilon_{1:n} = e_{1:n}, \mathbb{X}_{1_k:Z_n}) = \mathbb{P}_\theta(\varepsilon_{1:n} = e_{1:n}, \mathbb{X}_{1_k:Z_n}) = \left\{ \prod_{j=1}^{n} \mathbb{P}_\theta(\varepsilon_j = e_j) \right\} \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}|\varepsilon_{1:n} = e_{1:n})$$

where $\mathbb{P}_\theta(\varepsilon_j = e_j)$ is given in (4.4). Here we denote by $\mathbb{P}_\theta$ (and $\mathbb{E}_\theta$) the induced probability distribution (and corresponding expectation) on $\mathcal{E}^{\mathbb{N}} \times (\mathcal{A}^{\mathbb{N}})^k$ and θ_0 the true parameter corresponding to the distribution of the observations (we shall abbreviate to $\mathbb{P}_0$ and $\mathbb{E}_0$ the probability distribution and expectation under parameter θ_0).

4.2.2 Observations and likelihoods

As in the pair-HMM case (Chapter 2) there are different interpretations of what the observations represent on this model, and thus different definitions for the log-likelihood of the observed sequences $(X^1_{1:n_1}, \dots, X^k_{1:n_k})$. However, the difference with the pair-HMM is that in our model we suppose that the observed sequences are cut out of very much longer sequences between known homologous positions. This implies that any interpretation of what observations represent must assume that the underlying process $\{\varepsilon_n\}_{n\ge1}$ passes through the point $(n_1, \dots, n_k)$.

One may consider that what we observe are sequences that have evolved from an ancestral sequence of length n so that the likelihood should be $\mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}) = \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}, Z_n)$. This term is computed by summing, over all possible homology structures from an ancestral sequence of length n, the probability of observing the sequences and a homology structure.
Let us define $\mathcal{E}_{n_1,\ldots,n_k}$ the set of all possible homology structures of k sequences of lengths $n_1, \ldots, n_k$:

$$\mathcal{E}_{n_1,\ldots,n_k} = \{e \in (\mathcal{E}^k)^n;\ n \in \mathbb{N},\ \sum_{j=1}^{n} \{e_j(1) + e_j(2)\} = (n_1, \ldots, n_k)\}. \tag{4.6}$$

For any homology structure $e \in \mathcal{E}_{n_1,\ldots,n_k}$, if $e \in (\mathcal{E}^k)^n$, then n is the length of the path e and is denoted by $|e|$ ($|e|$ stands for the length of the ancestral sequence). Then we have

$$\mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}) = \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}, Z_n) = \sum_{e \in \mathcal{E}_{Z_n}; |e|=n} \mathbb{P}_\theta(\varepsilon_{1:n} = e, \mathbb{X}_{1_k:Z_n}).$$

Then, we would define the log-likelihood $\ell_n(\theta)$ as

$$\ell_n(\theta) = \log \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}), \quad n \geq 1. \tag{4.7}$$

But since the underlying process $\{Z_n\}_{n\geq 1}$ is not observed, the quantity $\ell_n(\theta)$ is not a measurable function of the observations. More precisely, the length n at which the observation is made is not observed itself. Though, if one decides that $(X^1_{1:n_1}, \ldots, X^k_{1:n_k})$ corresponds to the observation of the emitted sequences at a point of the hidden process $Z_n = (Z^i_n)_{i=1,\ldots,k}$ and some *unknown* "ancestral length" n, one does not use $\ell_n(\theta)$ as a log-likelihood, but rather

$$w_n(\theta) = \log Q_\theta(\mathbb{X}_{1_k:Z_n}), \quad n \geq 1 \tag{4.8}$$

where for any integers $n_i, i = 1, \ldots, k$

$$Q_\theta(X^1_{1:n_1}, \ldots, X^k_{1:n_k}) = \mathbb{P}_\theta(\exists m \geq 1, Z_m = (n_1, \ldots, n_k); X^1_{1:n_1}, \ldots, X^k_{1:n_k}). \tag{4.9}$$

In other words, Q_θ is the probability of the observed sequences under the assumption that the underlying process $\{\varepsilon_n\}_{n\geq 1}$ passes through the point $(n_1, \ldots, n_k)$. But the length of the ancestral sequence remains unknown when computing Q_θ. This gives the formula :

$$Q_\theta(X^1_{1:n_1}, \ldots, X^k_{1:n_k}) = \sum_{e \in \mathcal{E}_{n_1,\ldots,n_k}} \mathbb{P}_\theta(\varepsilon_{1:n} = e, X^1_{1:n_1}, \ldots, X^k_{1:n_k}). \tag{4.10}$$

Let us stress that we have

$$w_n(\theta) = \log \mathbb{P}_\theta(\exists m \geq 1, Z_m = (Z^i_n)_{i=1,\ldots,k}; X^1_{1:Z^1_n}, \ldots, X^k_{1:Z^k_n}), \quad n \geq 1,$$

meaning that the length of the ancestral sequence is not necessarily n, but is in fact unknown.

Q_θ is the quantity that is computed by the multiple alignment algorithms (see for instance [36, 53, 72]) and which is used as likelihood in biological applications. The more extended application is to use this quantity to co-estimate alignments and evolution parameters in a Bayesian framework (cf. [25, 52]). Thus, asymptotic properties of the criterion Q_θ and consequences on asymptotic properties of the estimators derived from Q_θ are of primarily interest.
We will look for asymptotic results for $n \to \infty$. We need to establish some kind of relationship between n and $n_1, \ldots, n_k$, to derive asymptotic results for $n_i \to \infty$. From our definition of the homology structure it is clear that it does not exist a deterministic relationship between the length of the ancestral sequence and the lengths of the observed sequences. However, a natural assumption is that very big insertions and deletions occur rarely and thus the length of the root sequence should be equivalent to the lengths of the observed sequences. In fact we have the following result.

Lemma 4 *For any $\theta \in \Theta$ $Z_n^i \sim n$, $i = 1, \ldots, k$, $\mathbb{P}_\theta$-almost surely.*

Proof. For all $i = 1, \ldots, k$ and for all $n \geq 1$ we have that

$$Z_n^i = \sum_{j=1}^{n} (\varepsilon_j^i(1) + \varepsilon_j^i(2))$$

where $\{\varepsilon_j^i\}_{j \geq 1}$ are i.i.d. Moreover, for any $\theta \in \Theta$

$$\begin{aligned}
& \mathbb{E}_\theta \left[\varepsilon_j^i(1) + \varepsilon_j^i(2)\right] \\
= & \sum_{m \geq 1} m \left\{ \mathbb{P}_\theta(\varepsilon_j^i(1) + \varepsilon_j^i(2) = m, \varepsilon_j^i(1) = 0) + \mathbb{P}_\theta(\varepsilon_j^i(1) + \varepsilon_j^i(2) = m, \varepsilon_j^i(1) = 1) \right\} \\
= & \sum_{m \geq 1} m \left\{ q_m^N(t_i) + q_m^H(t_i) \right\} \\
= & \sum_{m \geq 1} m \left\{ \left(\frac{1}{1+\lambda t_i} - e^{-\lambda t_i} \right) \frac{1}{1+\lambda t_i} \left(\frac{\lambda t_i}{1+\lambda t_i} \right)^{m-1} + e^{-\lambda t_i} \frac{1}{1+\lambda t_i} \left(\frac{\lambda t_i}{1+\lambda t_i} \right)^{m-1} \right\} = 1.
\end{aligned}$$

Now the result holds from the strong law of large numbers. □

According to this lemma, asymptotic results for $n \to \infty$ will imply equivalent ones for $n_i \to \infty$, $i = 1, \ldots, k$.

4.2.3 The case of two sequences

When $k = 2$ we are in the case of pairwise alignment. Indeed, two sequences evolving from a common unknown ancestor under the TKF91 can be considered as being one of them the ancestor of the other one (see the Appendix for details). It can be interesting to compare in this case the homology structure model to the pair-HMM.

First of all, let us remark that the likelihood (Q_θ) of two sequences $x_{1:n}$ and $y_{1:m}$ is the same under the two models, when considering for the pair-HMM the following transition matrix :

$$\begin{array}{c} \\ D \\ H \\ V \end{array} \begin{array}{c} \begin{array}{ccc} D & H & V \end{array} \\ \begin{pmatrix} \frac{\alpha(t)}{(1+\lambda t)} & \frac{1-\alpha(t)}{(1+\lambda t)} & \frac{\lambda t}{(1+\lambda t)} \\ (1-\kappa(t))\alpha(t) & (1-\kappa(t))(1-\alpha(t)) & \kappa(t) \\ \frac{\alpha(t)}{(1+\lambda t)} & \frac{1-\alpha(t)}{(1+\lambda t)} & \frac{\lambda t}{(1+\lambda t)} \end{pmatrix} \end{array} \tag{4.11}$$

where D, H and V stand for diagonal, horizontal and vertical mouvements respectively and $\alpha(t) = e^{-\lambda t}$, $\kappa(t) = 1 - \frac{\lambda t}{(1+\lambda t)(1-\alpha(t))}$. Indeed the probability of an homology structure is just the sum of the probabilities of all possible alignments leading to that homology structure. Then the sum over all possible alignments and all possible homology structures of two sequences is equivalent.

Finally, note that for the transition matrix in (4.11) the stationnary probabilities of insertions and deletions are the same, that is $p = q$ with the notations of Chapter 2. That means that we are in the case where the *main direction* of the alignment is always the straight line (n, n) for every value of the parameter.
This is also the case when $k > 2$, if we imagine the homology structure as a path in a $\mathbb{N}^k$-grid. Indeed Lemma 4 shows that the length of every observed sequence is equivalent to the length of the ancestral sequence.

4.3 Information divergence rates

4.3.1 Definition of Information divergence rates

In this section we prove the convergence of the normalized *log-likelihoods* $\ell_n(\theta)$ and $\omega_n(\theta)$. Let us note

$$\Theta_0 \quad = \quad \left\{\theta \in \Theta \mid \lambda > 0,\ h_J(x^{1:|J|}) > 0,\ \forall x^{1:|J|} \in \mathcal{A}^{|J|},\ \forall J \subseteq K,\ f(y) > 0,\ \forall y \in \mathcal{A}\right\}.$$

We shall always assume that $\theta_0 \in \Theta_0$.

Theorem 5 *The following holds for any $\theta \in \Theta_0$:*

i) $n^{-1}\ell_n(\theta)$ converges $\mathbb{P}_0$-almost surely and in $\mathbb{L}_1$, as n tends to infinity to

$$\ell(\theta) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_0\left(\log \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n})\right) = \sup_n \frac{1}{n}\mathbb{E}_0\left(\log \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n})\right).$$

ii) $n^{-1}w_n(\theta)$ converges $\mathbb{P}_0$-almost surely and in $\mathbb{L}_1$, as n tends to infinity to

$$w(\theta) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_0\left(\log Q_\theta(\mathbb{X}_{1_k:Z_n})\right) = \sup_n \frac{1}{n}\mathbb{E}_0\left(\log Q_\theta(\mathbb{X}_{1_k:Z_n})\right).$$

Using the terminology of Chapter 2 we then define Information divergence rates :

Definition 3 $\forall \theta \in \Theta_0,\ D(\theta|\theta_0) = w(\theta_0) - w(\theta) \quad \textit{and} \quad D^*(\theta|\theta_0) = \ell(\theta_0) - \ell(\theta).$

We recall that D^* is what is usually called the Information divergence rate in Information Theory : it is the limit of the normalized Kullback-Leibler divergence between the distributions of the observations at the true parameter value and another parameter value. However, we also call D an Information divergence rate since Q_θ may be interpreted as a likelihood.

Proof of Theorem 5. This proof is similar to the proof of Theorem 1 in Chapter 2. We shall use the following version of the sub-additive ergodic Theorem due to Kingman (1968) to prove point *i)*. A similar proof may be written for *ii)* and is left to the reader. Let $(W_{s,t})_{0\le s<t}$ be a sequence of random variables such that

1. For all $m < n$, $W_{0,n} \ge W_{0,m} + W_{m,n}$,
2. For all $l > 0$, the joint distributions of $(W_{m+l,n+l})_{0\le m<n}$ are the same as those of $(W_{m,n})_{0\le m<n}$,
3. $\mathbb{E}_0(W_{0,1}) > -\infty$.

Then $\lim_{n\to\infty} n^{-1}W_{0,n}$ exists almost surely. If moreover the sequences $(W_{m+l,n+l})_{l>0}$ are ergodic, then the limit is almost surely deterministic and equals $\sup_n n^{-1}\mathbb{E}_0(W_{0,n})$. If moreover $\mathbb{E}_0(W_{0,n}) \le An$, for some constant $A \ge 0$ and all n, then the convergence holds in $\mathbb{L}_1$.
We apply this theorem to the process

$$W_{m,n} = \log \mathbb{P}_\theta(\mathbb{X}_{Z_m+1_k:Z_n}), \quad 0 \le m < n.$$

Note that since $Z_0 = (0,0)$ is deterministic, we have $W_{0,n} = \log \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n})$. Super-additivity (namely point 1.) follows since for any $0 \leq m < n$,

$$\begin{aligned}
\mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n}) &= \sum_{\substack{e\in\mathcal{E}_{Z_n}\\|e|=n}} \mathbb{P}_\theta(\varepsilon_{1:n} = e_{1:n},\, \mathbb{X}_{1_k:Z_n}) \\
&\geq \sum_{\substack{e\in\mathcal{E}_{Z_m}\\|e|=m}} \sum_{\substack{e'\in\mathcal{E}_{Z_n-Z_m}\\|e'|=n-m}} \mathbb{P}_\theta(\varepsilon_{1:m} = e_{1:m},\, \varepsilon_{m+1:n} = e'_{1:n-m},\, \mathbb{X}_{1_k:Z_n}) \\
\geq \sum_{\substack{e\in\mathcal{E}_{Z_m}\\|e|=m}} \sum_{\substack{e'\in\mathcal{E}_{Z_n-Z_m}\\|e'|=n-m}} &\mathbb{P}_\theta(\varepsilon_{m+1:n} = e'_{1:n-m},\, \mathbb{X}_{Z_m+1_k:Z_n}) \times \mathbb{P}_\theta(\varepsilon_{1:m} = e_{1:m},\, \mathbb{X}_{1_k:Z_m}) \\
&= \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_m}) \times \mathbb{P}_\theta(\mathbb{X}_{Z_m+1_k:Z_n})
\end{aligned}$$

so that we get $W_{0,n} \geq W_{0,m} + W_{m,n}$, for any $0 \leq m < n$.
To understand the distribution of $(W_{m,n})_{0\leq m<n}$, note that $W_{m,n}$ only depends on trajectories of the random walk going from the point $(Z_m^1, \ldots, Z_m^k)$ to the point $(Z_n^1, \ldots, Z_n^k)$ with length $n-m$. Since the variables $(\varepsilon_n)_{n\geq 1}$ are i.i.d., one gets that the distribution of $(W_{m,n})$ is the same as that of $(W_{m+l,n+l})$ for any l, so that point 2. holds.
Point 3. comes from :

$$\begin{aligned}
\mathbb{P}_\theta(\mathbb{X}_{1_k:Z_1}) &= \sum_{\substack{e\in\mathcal{E}_{Z_1}\\|e|=1}} \mathbb{P}_\theta(\varepsilon_1 = e)\, \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_1} | \varepsilon_1 = e) \\
&= \sum_{\substack{e\in\mathcal{E}_{Z_1}\\|e|=1}} \mathbb{P}_\theta(\varepsilon_1 = e) \left\{ h_{\{i|\delta_1^i=1\}}\left(\{X_1^i\}_{i|\delta_1^i=1}\right) \prod_{i=1}^{k} \prod_{s=1}^{a_1^i} f\left(X^i_{\delta_1^i+s}\right) \right\} > 0
\end{aligned}$$

$\mathbb{P}_0$-almost surely, since $\theta \in \Theta_0$, provided that $Z_1^i \geq 1$ for some $i \in K$. So $\mathbb{E}_0(W_{0,1}) = \mathbb{E}_0 \log \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_1}) > -\infty$.
Let us fix $0 \leq m < n$. The proof that $W^{s,t} = (W_{m+l,n+l})_{l>0}$ is ergodic is the same as that of Leroux (1992, Lemma 1). Let T be the shift operator, so that if $u = (u_l)_{l\geq 0}$, the sequence Tu is defined by $(Tu)_l = (u)_{l+1}$ for any $l \geq 0$. Let B be an event which is T-invariant. We need to prove that $\mathbb{P}_0(W^{m,n} \in B)$ equals 0 or 1. For any integer i, there exists a cylinder set B_i, depending only on the coordinates u_l with $-j_i \leq l \leq j_i$ for some sub-sequence j_i, such that $\mathbb{P}_0(W^{m,n} \in B \Delta B_{j_i}) \leq 1/2^i$. Here, Δ denotes the symmetric difference between sets. Since $W^{m,n}$ is stationary and B is T-invariant :

$$\mathbb{P}_0\left(W^{m,n} \in B\Delta B_{j_i}\right) = \mathbb{P}_0\left(T^{2j_i} W^{m,n} \in B\Delta B_{j_i}\right) = \mathbb{P}_0\left(W^{m,n} \in B\Delta T^{-2j_i} B_{j_i}\right).$$

Let $\tilde{B} = \cap_{i\geq 1} \cup_{h\geq i} T^{-2j_h} B_{j_h}$. Borel-Cantelli's Lemma leads to $\mathbb{P}_0(W^{m,n} \in B\Delta\tilde{B}) = 0$, so that $\mathbb{P}_0(W^{m,n} \in B) = \mathbb{P}_0(W^{m,n} \in \tilde{B}) = \mathbb{P}_0(W^{m,n} \in B \cap \tilde{B})$. Now, conditional

on $(\varepsilon_n)_{n\in\mathbb{N}}$, the random variables $(W_{m+l,n+l})_{l>0}$ are strongly mixing. Indeed $W_{m+l,n+l} = \sum_{j=m+l}^{n+l} \log \mathbb{P}_\theta(\mathbb{X}_{Z_j+1_k:Z_{j+1}})$ so it depends only on $(W_{m+k,n+k})_{k=max(1,m+l-n+1),\ldots,n+l-m+1}$. Then the $0-1$ law implies (see [73]) that for any fixed sequence e with values in $(\mathcal{E}^k)^{\mathbb{N}}$, the probability $\mathbb{P}_0(W^{m,n} \in \tilde{B} | (\varepsilon_n)_n = e)$ equals 0 or 1, so that

$$\mathbb{P}_0\left(W^{m,n} \in \tilde{B}\right) = P\left((\varepsilon_n)_n \in C\right)$$

where C is the set of sequences e such that $P(W^{m,n} \in \tilde{B} | (\varepsilon_n)_n = e) = 1$. But it is easy to see that C is T-invariant. Indeed, if $e \in C$ then, since $W^{m,n}$ is stationary and $\tilde{B}$ invariant,

$$1 = \mathbb{P}_0(W^{m,n} \in \tilde{B} | (\varepsilon_n)_n = e) = \mathbb{P}_0(TW^{m,n} \in \tilde{B} | (\varepsilon_n)_n = Te) = \mathbb{P}_0(W^{m,n} \in \tilde{B} | (\varepsilon_n)_n = Te)$$

so that $Te \in C$. Now, since $(\varepsilon_n)_{n\geq 1}$ is ergodic, $\mathbb{P}_0\left((\varepsilon_n)_n \in C\right)$ equals 0 or 1. This concludes the proof of ergodicity of the sequence $W^{m,n}$.
To end with, note that for any $n \geq 0$, the random variable $W_{0,n}$ is non positive, ensuring the convergence of $\{n^{-1}W_{0,n}\}$ in $\mathbb{L}_1$. □

4.3.2 Divergence properties of Information divergence rates

Information divergence rates should be non negative : this is proved below. They also should be positive for parameters that are different than the true one : we only prove it in a particular subset of the parameter set. Let us define the set

$$\Theta_{marg} = \left\{\theta \in \Theta_0 \ : \ h^i_J = f, \forall J \subseteq K, \forall i \in J\right\}.$$

where h^i_J denotes the i-th marginal of h_J.

Theorem 6 *Information divergence rates satisfy :*
- *For all $\theta \in \Theta_0$, $D(\theta|\theta_0) \geq 0$ and $D^*(\theta|\theta_0) \geq 0$.*
- *If θ_0 and θ are in Θ_{marg}, $D(\theta|\theta_0) > 0$ and $D^*(\theta|\theta_0) > 0$ as soon as $f \neq f_0$.*

Note that since in the homology structure model the *main direction* is always the diagonal for every value of the parameter, then this theorem is equivalent to Theorem 2 of Chapter 2 for pair-HMMs.

Proof. Since for all n,

$$\mathbb{E}_0\left(\log \mathbb{P}_0(\mathbb{X}_{1_k:Z_n})\right) - \mathbb{E}_0\left(\log \mathbb{P}_\theta(\mathbb{X}_{1_k:Z_n})\right)$$

is a Kullback-Leibler divergence, it is non negative, and the limit $D^*(\theta|\theta_0)$ is also non negative.

Let us prove that $D(\theta|\theta_0)$ is also non negative. To compute the value of the expectation $\mathbb{E}_0[w_n(\theta)]$, note that the set of all possible values of Z_n is $\mathbb{N}^k$. Then,

$$\mathbb{E}_0[w_n(\theta)] = \sum_{(n_1,\dots,n_k)\in\mathbb{N}^k} \sum_{(x^i_{1:n_i})_{i=1,\dots,k}} \mathbb{P}_0\big(Z_n = (n_1,\dots,n_k), X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}\big) \times \log Q_\theta(x^1_{1:n_1},\dots,x^k_{1:n_k}).$$

Now, by definition,

$$D\left(\theta|\theta_0\right) = \lim_{n\to+\infty} \frac{1}{n}\mathbb{E}_0\left(\log\frac{Q_{\theta_0}(\mathbb{X}_{1_k:Z_n})}{Q_\theta(\mathbb{X}_{1_k:Z_n})}\right).$$

By using Jensen's inequality,

$$\mathbb{E}_0\left(\log\frac{Q_\theta(\mathbb{X}_{1_k:Z_n})}{Q_{\theta_0}(\mathbb{X}_{1_k:Z_n})}\right) \le \log\mathbb{E}_0\left(\frac{Q_\theta(\mathbb{X}_{1_k:Z_n})}{Q_{\theta_0}(\mathbb{X}_{1_k:Z_n})}\right) = \log\mathbb{E}_0\left[\mathbb{E}_0\left(\frac{Q_\theta(\mathbb{X}_{1_k:Z_n})}{Q_{\theta_0}(\mathbb{X}_{1_k:Z_n})}\right)\big|Z_n\right].$$

Now, for all $(n_1,\dots,n_k)\in\mathbb{N}^k$

$$\begin{aligned}
&\mathbb{E}_0\left(\frac{Q_\theta(\mathbb{X}_{1_k:Z_n})}{Q_{\theta_0}(\mathbb{X}_{1_k:Z_n})}\big|Z_n = (n_1,\dots,n_k)\right)\\
&= \sum_{(x^i_{1:n_i})_{i=1,\dots,k}} \mathbb{P}_0\big(Z_n = (n_1,\dots,n_k), X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}\big)\frac{Q_\theta(x^1_{1:n_1},\dots,x^k_{1:n_k})}{Q_{\theta_0}(x^1_{1:n_1},\dots,x^k_{1:n_k})}\\
&\overset{(a)}{\le} \sum_{(x^i_{1:n_i})_{i=1,\dots,k}} \mathbb{P}_\theta\big(\exists m\ge 1, Z_m = (n_1,\dots,n_k),\ X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}\big)\\
&= \mathbb{P}_\theta\big(\exists m\ge 1, Z_m = (n_1,\dots,n_k)\big) \le 1
\end{aligned}$$

where (a) comes from expression (4.10). Thus, $\mathbb{E}_0\left[\mathbb{E}_0\left(\frac{Q_\theta(\mathbb{X}_{1_k:Z_n})}{Q_{\theta_0}(\mathbb{X}_{1_k:Z_n})}\right)|Z_n\right] \le 1$, and

$$\lim_{n\to+\infty}\frac{1}{n}\left(w_n(\theta) - w_n(\theta_0)\right) \le \liminf_{n\to+\infty}\frac{1}{n}\log\mathbb{E}_0\left[\mathbb{E}_0\left(\frac{Q_\theta(\mathbb{X}_{1_k:Z_n})}{Q_{\theta_0}(\mathbb{X}_{1_k:Z_n})}\big|Z_n\right)\right] \le 0.$$

So finally

$$\forall\theta\in\Theta_0,\ D(\theta|\theta_0)\ge 0.$$

Let us now consider the case where θ_0 and θ are in Θ_{marg}. Let us remark thar for any $\theta \in \Theta_{marg}$ we have

$$\begin{aligned}
&\mathbb{P}_\theta\big(Z_n = (n_1,\dots,n_k),\, X^1_{1:n_1} = x^1_{1:n_1}\big) \\
&\quad= \sum_{(x^i_{1:n_i})_{i=2,\dots,k}} \mathbb{P}_\theta\big(Z_n = (n_1,\dots,n_k),\, X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}\big) \\
&\quad= \sum_{e\in\mathcal{E}_{n_1,\dots,n_k};|e|=n} \sum_{(x^i_{1:n_i})_{i=2,\dots,k}} \mathbb{P}_\theta(\varepsilon_{1:n} = e,\, X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}) \\
&= \sum_{e\in\mathcal{E}_{n_1,\dots,n_k};|e|=n} \sum_{(x^i_{1:n_i})_{i=2,\dots,k}} \mathbb{P}_\theta(\varepsilon_{1:n} = e)\,\mathbb{P}_\theta(X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}|\varepsilon_{1:n} = e) \\
&\qquad= \mathbb{P}_\theta\big(Z_n = (n_1,\dots,n_k)\big) f^{\otimes n_1}(x^1_{1:n_1}) \quad (4.12)
\end{aligned}$$

where the last equality comes from 4.5. In the same way, for any $\theta \in \Theta_{marg}$ we have $\mathbb{P}_\theta(\exists m \le 1, Z_m = (n_1,\dots,n_k), X^1_{1:n_1} = x^1_{1:n_1}) = \mathbb{P}_\theta(\exists m \le 1, Z_m = (n_1,\dots,n_k)) f^{\otimes n_1}(x^1_{1:n_1})$. This is also true for any other sequence $X^i_{1:n_i}$, $i = 1\dots,k$.
Then, using Jensen's Inequality and definition (4.10),

$$\begin{aligned}
&\mathbb{E}_0\left(\log\frac{Q_\theta(\mathbb{X}_{1:Z_n})}{Q_{\theta_0}(\mathbb{X}_{1:Z_n})}\right) \\
&= \sum_{(n_1\dots,n_k)\in\mathbb{N}^k} \sum_{(x^i_{1:n_i})_{i=1,\dots,k}} \mathbb{P}_0\big(Z_n = (n_1\dots,n_k),\, X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}\big) \\
&\times\ \log\frac{Q_\theta(x^1_{1:n_1},\dots,x^k_{1:n_k})}{Q_{\theta_0}(x^1_{1:n_1},\dots,x^k_{1:n_k})} \quad \le \sum_{(n_1\dots,n_k)\in\mathbb{N}^k} \sum_{x^1_{1:n_1}} \mathbb{P}_0\big(Z_n = (n_1,\dots,n_k),\, X^1_{1:n_1} = x^1_{1:n_1}\big) \\
&\times\ \log\left(\sum_{(x^i_{1:n_i})_{i=2}^k} \frac{\mathbb{P}_0\big(Z_n = (n_1\dots,n_k),\, X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}\big) Q_\theta(x^1_{1:n_1},\dots,x^k_{1:n_k})}{\mathbb{P}_0\big(Z_n = (n_1\dots,n_k),\, X^1_{1:n_1} = x^1_{1:n_1}\big) Q_{\theta_0}(x^1_{1:n_1},\dots,x^k_{1:n_k})}\right) \\
&\le \sum_{(n_1\dots,n_k)\in\mathbb{N}^k} \sum_{x^1_{1:n_1}} \mathbb{P}_0\big(Z_n = (n_1,\dots,n_k)\big) f_0^{\otimes n_1}(x^1_{1:n}) \\
&\qquad\qquad \times\log\left(\frac{\mathbb{P}_\theta\big(\exists m \ge 1, Z_m = (n_1,\dots,n_k)\big) f^{\otimes n_1}(x^1_{1:n_1})}{\mathbb{P}_0\big(Z_n = (n_1,\dots,n_k)\big) f_0^{\otimes n_1}(x^1_{1:n_1})}\right),
\end{aligned}$$

where the last inequality comes from (4.12) and the fact that $\mathbb{P}_0\big(Z_n = (n_1,\dots,n_k),\, X^1_{1:n_1} = x^1_{1:n_1},\dots,X^k_{1:n_k} = x^k_{1:n_k}\big) \le Q_{\theta_0}(x^1_{1:n_1},\dots,x^k_{1:n_k})$.

Thus, we have

$$\begin{aligned}
& -D(\theta|\theta_0) \\
\leq\ & \limsup_{n\to+\infty} \frac{1}{n} \sum_{(n_1\ldots,n_k)\in\mathbb{N}^k} \mathbb{P}_0(Z_n=(n_1,\ldots,n_k))\Big\{ \log \frac{\mathbb{P}_\theta\big(\exists m\geq 1, Z_m=(n_1,\ldots,n_k)\big)}{\mathbb{P}_0\big(Z_n=(n_1,\ldots,n_k)\big)} \\
& \qquad\qquad +n_1 \sum_x f_0(x)\log\frac{f(x)}{f_0(x)}\Big\} \\
\leq\ & \limsup_{n\to+\infty} \frac{1}{n}\Big\{ \log \sum_{(n_1\ldots,n_k)\in\mathbb{N}^k} \mathbb{P}_\theta\big(\exists m\geq 1, Z_m=(n_1,\ldots,n_k)\big) \\
& \qquad\qquad + \sum_{(n_1\ldots,n_k)\in\mathbb{N}^k} \mathbb{P}_0\big(Z_n=(n_1,\ldots,n_k)\big) n_1 \sum_x f_0(x)\log\frac{f(x)}{f_0(x)}\Big\} \\
\leq\ & \limsup_{n\to+\infty} \frac{1}{n}\Big\{ \mathbb{E}_0[Z_n^1] \sum_x f_0(x)\log\frac{f(x)}{f_0(x)}\Big\} = \limsup_{n\to+\infty} \frac{1}{n}\Big\{ n\sum_x f_0(x)\log\frac{f(x)}{f_0(x)}\Big\} < 0,
\end{aligned}$$

as soon as $f \neq f_0$.
The proof for D^* follow the same lines. □

It would be interesting to prove the uniqueness of the maximum of the functions $\ell(\theta)$ and $w(\theta)$ at the true value of the parameter θ_0. If that was true, the consistency of maximum likelihood and bayesian estimators would be obtained with classical arguments (see Chapter 2). We propose to investigate the behavior of functions $\ell(\theta)$ and $w(\theta)$ via some simulations.

4.4 Simulations

We have considered for the simulations a 3-star phylogenetic tree, the most simple non trivial example of multiple alignment. The branches lengths, or evolutionnary distance from the ancestral sequence to the observed sequences, are set to 1 in all branches. Let us recall that this distance is not the real time of evolution between sequences but a measure given in terms of the number of expected evolutionary events per site. Indeed, under our indel evolutionnary model λt is the expected number of indels per site between two sequences at distance t.
We have used a multiple-HMM (see [30] for instance) scheme to simulate the sequences. Indeed in practice it is easier to simulate from a finite state Markov chain than from our i.i.d. variables on $\mathbb{N}^3$. The number of states for the Markov chain for three sequences is 15 (2^4-1). The simulated sequences have been used to compute the quantities $\ell(\theta)$ and $w(\theta)$.

The log-likelihood $\omega_n(\theta)$ has been computed with the Forward algorithm for multiple-HMM (cf. [19]). Note that this algorithm computes the log-likelihood by summing over all possible alignments of the three sequences. However, since a homology structure is just a set of alignments, this is equivalent to sum over all possible homology structures, and the final result is exactly $\omega_n(\theta)$. The time complexity for a non-improved version of this algorithm is $O(15^2 n_1 n_2 n_3)$, where n_1, n_2 and n_3 are the lengths of the observed sequences. Computation of $\ell_n(\theta)$ is done with a modified version of the Forward algorithm that takes into account the length of the ancestral sequence. The time complexity grows now to $O(15\, n\, n_1 n_2 n_3)$. This is the reason for having limited the simulations to 3 sequences.
The substitution model chosen for the simulations is described below.

4.4.1 The substitution model

For the whole simulation procedure we consider the following pairwise markovian substitution model :

$$p_t(x,y) = \begin{cases} (1-e^{-\alpha t})\nu(y) & \text{if } x \neq y \\ \{(1-e^{-\alpha t})\nu(x) + e^{-\alpha t}\} & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ is called substitution rate, t is the evolutionary distance, and for every letter x, $\nu(x)$ equals the equilibrium probability of x. This model is known as the Felsenstein81 substitution model [21]. We will take $f(\cdot) = \nu(\cdot)$. We define the emission function h as

$$h_J((x_i)_{i\in J}) = \sum_{R\in\mathcal{A}} \nu(R) \prod_{i\in J} p_{t_i}(R, x_i)$$

for all $J \subseteq \{1,2,3\}$.
The equilibrium probability distribution $\nu(\cdot)$ is assumed to be known and will not be part of the parameter. Then we have $f(\cdot) = f_0(\cdot)$. We will set it to $\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$ for the whole simulation procedure. The unknown parameter is $\theta = (\lambda, \alpha)$.

4.4.2 Simulation results

We have computed the functions $\ell(\theta)$ and $w(\theta)$ for two different values of θ_0 :
- $\lambda_0 = 0.02$, $\alpha_0 = 0.1$ and
- $\lambda_0 = 0.01$, $\alpha_0 = 0.08$.

The substitution rate is much bigger than the insertion-deletion rate and both are quite small, as expected by biologists.
The graphs of $\ell(\theta)$ and $w(\theta)$ for these parameterizations are shown in Figures 4.2 and 4.3. For the fisrt parametrization we can see that $w(\theta)$ seems to take its maximum at

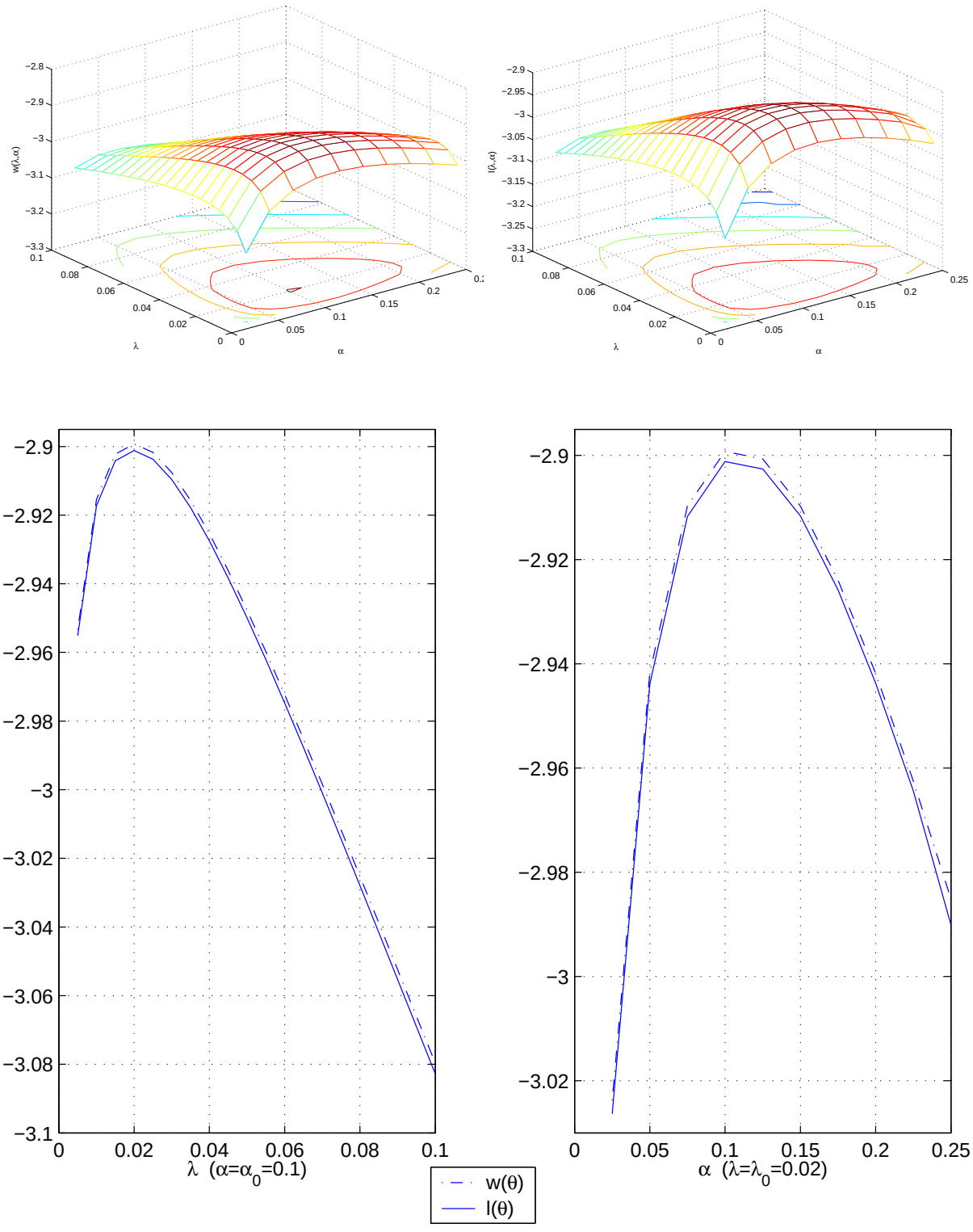


FIG. 4.2 – *On top : w and ℓ for parametrization ($\lambda_0 = 0.02, \alpha_0 = 0.1$). On bottom : cuts of ℓ and w for $\alpha = \alpha_0$ fixed and for $\lambda = \lambda_0$ fixed.*

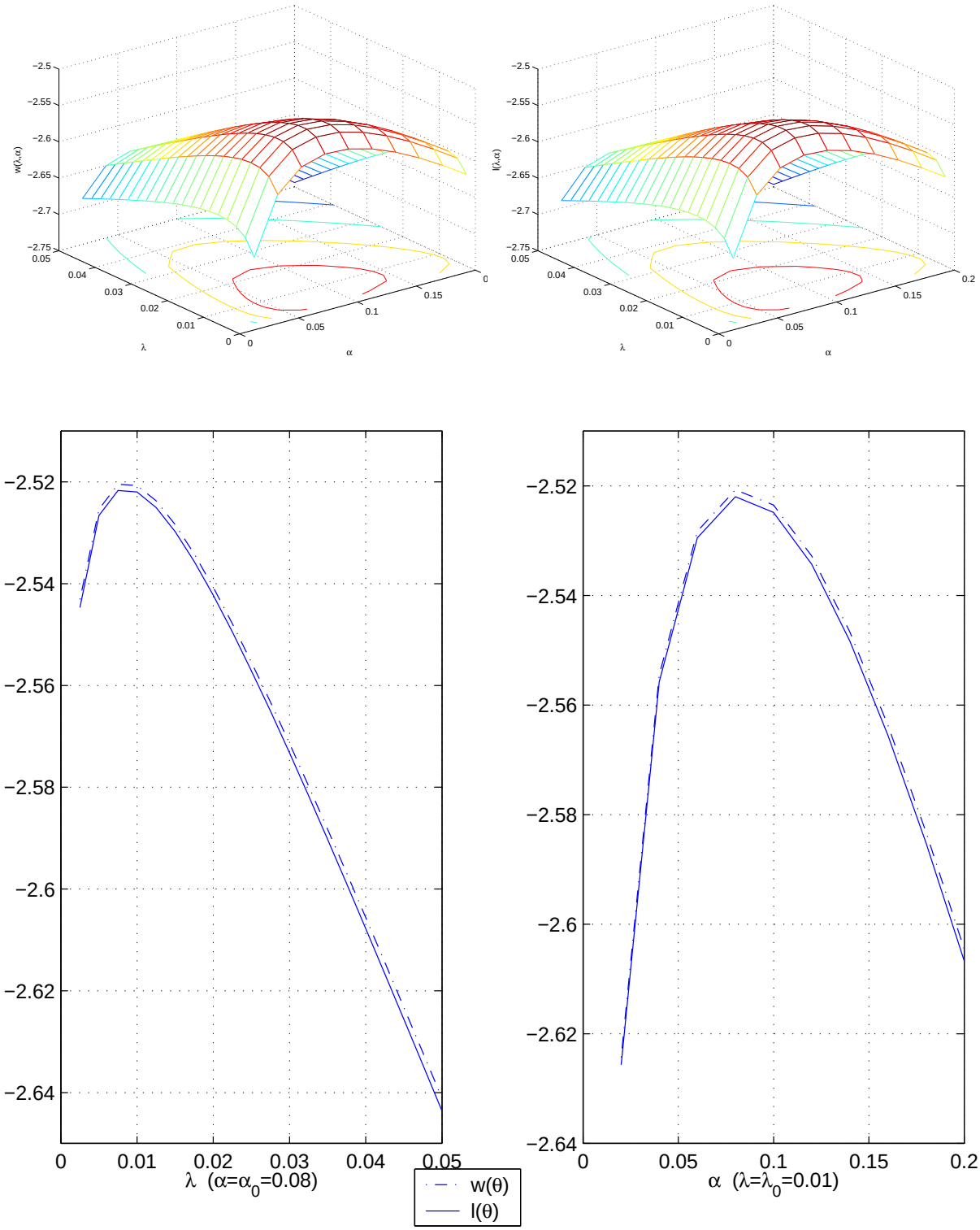


FIG. 4.3 – *On top : w and ℓ for parametrization ($\lambda_0 = 0.01, \alpha_0 = 0.08$). On bottom : cuts of ℓ and w for $\alpha = \alpha_0$ fixed and for $\lambda = \lambda_0$ fixed.*

(λ_0, α_0) (Figure 4.2, top left). For $\ell(\theta)$ this is not so evident. Neither for any of the two functions for the second parametrization. However, when looking at the cuts of w and $\ell(\theta)$ for $\alpha = \alpha_0$ and $\lambda = \lambda_0$ we appreciate that in both parametrizations both seem to take their maximums near λ_0 and α_0 respectively. We remark that in the two examples, the functions $\ell(\theta)$ and $w(\theta)$ are very close to each other.

4.5 Discussion

Our first contribution is to provide a probabilistic and statistical background to the study of multiple alignment of sequences related by a rigourous model of evolution. We define the homology structure of k sequences related by star shaped phylogenetic tree as a sequence of i.i.d. random variables whose distribution is determined by the evolution process. Our main results are given in Theorems 5 and 6, where we first prove convergence of normalized log-likelihoods and identify cases where a divergence property holds. Despite the positive results that we obtain, it is not yet possible to validate the estimation of evolution parameters under the homology structure model in every situation. However, the simulation studies that we present to investigate situations that are not covered by Theorem 6 provide encouraging results.
The extension of this work to the case of an arbitrary phylogenetic tree would be of main interest.

Annexe A

Note sur la reversibilité du modèle TKF91

Le modèle TKF91 [76] est reversible dans le temps, c'est à dire le résultat d'aligner deux séquences $x_{1:n}$ et $y_{1:m}$ est le même si on considère $x_{1:n}$ comme etant l'ancêtre de $y_{1:m}$ ou $y_{1:m}$ comme etant l'ancêtre $x_{1:n}$. Ceci a été montré par Metzler *et al.* [57].

Proposition 2 [Metzler *et al.* [57], Corollary 1]
Soit $TKF_{n,m;[0,t];\lambda,\mu}$ la distribution des structures d'homologie sous le modèle TKF91 de paramètres λ et μ, conditionnellement à l'observation d'une séquence de longueur n ayant evoloué en une séquence de longueur m après un temps t. Alors :
Pour tous $0 < \lambda < \mu$, les distributions $TKF_{n,m;[0,t];\lambda,\mu}$ et $TKF_{m,n;[0,t];\lambda,\mu}$ sont égales.

En effet la reversibilité n'existe pas au niveau de l'alignement mais au niveau de la structure d'homologie (voir Exemple 1). La structure d'homologie est la spécification des positions homologues (*matchs*) de l'alignement et du nombre d'insertions entre deux positions homologues. Elle peut être representée par une suite de vecteurs $(n_1, m_1), \ldots, (n_k, m_k)$ dont la première composante ($n_i \geq 0$) indique le nombre d'insertions dans la première séquence entre deux *matchs* et la deuxième composante ($m_i \geq 0$) indique le nombre d'insertions dans la deuxième séquence entre ces mêmes *matchs*. A une même structure d'homologie correspondent plusieurs alignements.

Exemple 1 *Soit la structure d'homologie $(2, 1)$, c'est à dire tous les alignements avec deux insertions dans la première séquence et une insertion dans la deuxième entre deux* matchs. *Ces alignements sont :*

$$(Al1)\ \begin{matrix} B & B & B & - & B \\ B & - & - & B & B \end{matrix} \qquad (Al2)\ \begin{matrix} B & B & - & B & B \\ B & - & B & - & B \end{matrix} \qquad et \quad (Al3)\ \begin{matrix} B & - & B & B & B \\ B & B & - & - & B \end{matrix}.$$

La probabilité de $(2, 1)$ est la somme des probabilités de ces trois alignements et est égale à la probabilité de la structure d'homologie $(1, 2)$, qui est la somme des probabilités des alignements

$$(Al4)\ \begin{matrix} B & B & - & - & B \\ B & - & B & B & B \end{matrix} \qquad (Al5)\ \begin{matrix} B & - & B & - & B \\ B & B & - & B & B \end{matrix} \qquad et \quad (Al6)\ \begin{matrix} B & - & - & B & B \\ B & B & B & - & B \end{matrix}.$$

En effet, si on note par M l'état match, par I l'état insertion, par D l'état délétion et par π_{ij} la probabilité de transition de l'état i à l'état j, à partir de (1.5) on a :

$$\begin{aligned}
\mathbb{P}_{TKF91,\lambda,\mu}((2,1)) &= \mathbb{P}_{TKF91,\lambda,\mu}((Al1)) + \mathbb{P}_{TKF91,\lambda,\mu}((Al2)) + \mathbb{P}_{TKF91,\lambda,\mu}((Al3)) \\
&\overset{(a)}{=} \pi_{MD}\pi_{DD}\pi_{DI}\pi_{IM} + \pi_{MD}\pi_{DI}\pi_{ID}\pi_{DM} + \pi_{MI}\pi_{ID}\pi_{DD}\pi_{DM} \\
&= (1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t))(1-\kappa(t))(\frac{\lambda}{\mu})(1-\alpha(t))\kappa(t)(1-\lambda\beta(t))(\frac{\lambda}{\mu})\alpha(t) \\
&+(1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t))\kappa(t)(1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t))(1-\kappa(t))(\frac{\lambda}{\mu})\alpha(t) \\
&+\lambda\beta(t)(1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t))(1-\kappa(t))(\frac{\lambda}{\mu})(1-\alpha(t))(1-\kappa(t))(\frac{\lambda}{\mu})\alpha(t) \\
&\overset{(b)}{=} \pi_{MD}\pi_{DI}\pi_{II}\pi_{IM} + \pi_{MI}\pi_{ID}\pi_{DI}\pi_{IM} + \pi_{MI}\pi_{II}\pi_{ID}\pi_{DM} \\
&\overset{(a)}{=} \mathbb{P}_{TKF91,\lambda,\mu}((Al4)) + \mathbb{P}_{TKF91,\lambda,\mu}((Al5)) + \mathbb{P}_{TKF91,\lambda,\mu}((Al6)) = \mathbb{P}_{TKF91,\lambda,\mu}((1,2))
\end{aligned}$$

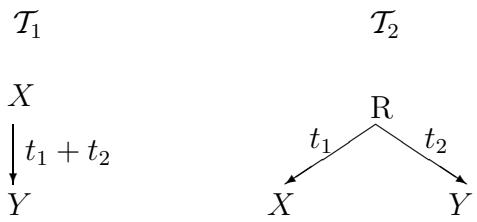
où (b) viens de $(1-\kappa(t))(\frac{\lambda}{\mu})(1-\alpha(t)) = \lambda\beta(t)$ et l'égalité est terme à terme, et (a) se verifie à une constante multiplicative près (la probabilité initiale de l'état match*).*
Si on regarde maintenant ce qui se passe au niveau des alignements, on voit que $(Al1)$ et l'alignement qu'on obtient si on inverse les deux séquences, c'est à dire $(Al6)$, n'ont pas la même probabilité. En effet

$$\begin{aligned}
&(1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t))(1-\kappa(t))(\frac{\lambda}{\mu})(1-\alpha(t))\kappa(t)(1-\lambda\beta(t))(\frac{\lambda}{\mu})\alpha(t) \\
&\quad = \lambda\beta(t)(1-\lambda\beta(t))(\frac{\lambda}{\mu})(1-\alpha(t))(1-\kappa(t))(\frac{\lambda}{\mu})(1-\alpha(t))(1-\kappa(t))(\frac{\lambda}{\mu})\alpha(t) \\
\iff \lambda\beta(t) = \kappa(t) &\iff \lambda\beta(t) = \frac{1-\alpha(t)-\mu\beta(t)}{1-\alpha(t)} \iff (\mu-\lambda)(1-e^{-\mu t}) = \mu(1-e^{(\lambda-\mu)t}) \\
&\iff \lambda = 0
\end{aligned}$$

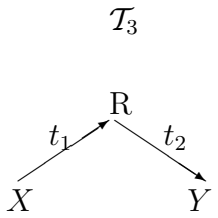
ce qui n'est pas possible sous le modèle TKF91.

La proprieté de reversibilité dans le temps de la structure d'homologie est aussi verifiée pour les modèles FID de Metzler [55] et TKF92 [77] (voir [57]). Pour le modèle FID, où la chaîne de Markov de l'alignement est stationnaire, ceci est équivalent à dire que la probabilité stationnaire d'insertion est égale à la probabilité stationnaire de délétion.

De la reversibilité dans le temps découle la proprieté qui nous permet de traiter deux séquences comme si l'une était l'ancêtre de l'autre, au lieu de devoir considérer tous les ancêtres communs possibles. Il s'agit alors de montrer que la distribution de la structure d'homologie sur les deux arbres suivants



est la même. A cause de la reversibilité dans le temps l'arbre $\mathcal{T}_2$ est équivalent à l'arbre



et donc intuitivement, en sommant sur toutes les possibles positions de la racine et toutes les possibles séquences ancestrâles on aurait le résultat. Celui-ci a été montré par Thatte [74] qui prouve que la distribution de la structure d'homologie de deux séquences reliées par l'arbre $\mathcal{T}_2$ et ayant evolué selon le modèle TKF91 est une fonction de $t_1 + t_2$. Ceci peut aussi être prouvé par un argument recursif sur le nombre d'insertions à partir des probabilités de transition données dans la Figure 1.5.

Comme une conséquence directe de cette proprieté et de l'hypothèse d'indépendance des branches d'un arbre par rapport au processus évolutif, Thate [74] montre que la distribution de la structure d'homologie des séquences (plus que deux) reliées par un arbre phylogénétique quelconque sous le modèle TKF91 est indépendante de la position de la racine dans l'arbre. Celle ci est une proprieté très importante dans la construction jointe d'alignements multiples et arbres phylogenetiques car elle nous permet d'aligner des séquences sur des arbres dont on ne connait pas la position de la racine (voir par exemple [52]).

Références

[1] L. Allison and C. Wallace. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimisation of multiple alignments. *J. Mol. Evol.*, 34 :418–430, 1994.

[2] S.F. Altschul, W. Gish, E. Miller, E.W. Meyers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3) :403–410, 1960.

[3] A. Arribas-Gil, E. Gassiat, and C. Matias. Parameter estimation in pair-hidden Markov models. *Scand. J. Statist.*, 33(4) :651–671, 2006.

[4] A. Arribas-Gil, D. Metzler, and J.-L. Plouhinec. Statistical alignment with a sequence evolution model allowing rate heterogeneity along the sequence. *Preprint.*, 2006.

[5] P. Baldi, Y. Chauvin, T Hunkapiller, and M.A. McClure. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA.*, 91 :1059–1063, 1994.

[6] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37 :1554–1563, 1966.

[7] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41 :164–171, 1970.

[8] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S Mattick, and D. Haussler. Ultraconserved elements in the human genome. *Science*, 304 :1321–1325, 2004.

[9] M.J. Bishop and E.A. Thompson. Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.*, 190 :159–165, 1986.

[10] J. Bérard, J.-B. Gouéré, and D. Piau. Solvable models of neighbor-dependent nucleotide substitution processes. *Preprint. http ://fr.arxiv.org/abs/math.PR/0510034*, 2005.

[11] A. Caliebe and U. Rösler. Convergence of the maximum a posteriori path estimator in hidden Markov models. *IEEE Trans. Inf. Theory*, 48(7) :1750–1758, 2002.

[12] F. Chiaromonte, R.J. Weber, K.M. Roskin, M. Diekhans, W.J. Kent, and D. Haussler. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb Symp Quant Biol.*, 68 :245–254, 2003.

[13] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley Series in Telecommunications, John Wiley & Sons, New York, USA, 1991.

[14] I. Csiszàr and J. Körner. *Information theory. Coding theorems for discrete memoryless systems.* Probability and Mathematical Statistics. Academic Press., New York-San Francisco-London, 1981.

[15] M.C. Davey and D.J.C. MacKay. Reliable communication over channels with insertions, deletions, and substitutions. *IEEE Trans. Inf. Theory*, 47(2) :687–698, 2001.

[16] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. *Atlas of protein structure. National Biomedical Research Fondation, Washington D.C.*, 5, Suppl. 3. :345–352, 1978.

[17] B. Delyon, M. Lavielle, and Moulines E. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Stat.*, 27 :94–128, 1999.

[18] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc., Ser. B*, 39 :1–38, 1977.

[19] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis : probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK, 1998.

[20] S.R. Eddy, G. Mitchison, and R. Durbin. Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.*, 2 :9–23, 1995.

[21] J. Felsenstein. Evolutionary trees from DNA sequences : a maximum likelihood approach. *J. Mol. Evol.*, 17(6) :368–376, 1981.

[22] J. Felsenstein. *Inferring Phylogenies*. Sunderland, MA : Sinauer Associates, 2004.

[23] J. Felsenstein and A. Churchill. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13 :93–104, 1996.

[24] W.M. Fitch and T.F. Smith. Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA*, 80 :1382–1386, 1983.

[25] R. Fleissner, D. Metzler, and A. von Haeseler. Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst. Biol.*, 54(4) :548–561, 2005.

[26] A. Gambin, J. Tiuryn, and J. Tyszkiewicz. Alignment with context dependent scoring function. *J.Comput. Biol.*, 13(1) :81–101, 2006.

[27] M. Gribskov, A.D. McLachlan, and D. Eisenberg. Profile analysis : detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, 84 :4355–4358, 1987.

[28] S. Grossmann and B. Yakir. Large Deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignments. *Bernoulli*, 10(5) :829–845, 2004.

[29] J. Hein. An algorithm for statistical alignment of sequences related by a binary tree. *Pac. Symp. Biocomp.*, pages 179–190, 2001.

[30] J. Hein, J.L. Jensen, and C.N.S. Pedersen. Recursions for statistical multiple alignment. *Proc. Natl. Acad. Sci. USA*, 100(25) :14960–14965, 2003.

[31] J. Hein, C. Wiuf, B. Knudsen, M.B. Moller, and G. Wibling. Statistical alignment : computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.*, 302 :265–279, 2000.

[32] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22) :10915–10919, 1992.

[33] D. Higgins, A. Bleasby, and R. Fuchs. CLUSTAL V : improved software for multiple sequence alignment. *CABIOS*, 8 :189–191, 1992.

[34] A. Hobolth and J.L. Jensen. Applications of hidden Markov models for characterization of homologous DNA sequences with a common gene. *J. Comput. Biol.*, 12(2) :186–203, 2005.

[35] I. Holmes. Using evolutionary Expectation Maximization to estimate indel rates. *Bioinformatics*, 21(10) :2294–2300, 2005.

[36] I. Holmes and W.J. Bruno. Evolutionary HMMs : a Bayesian approach to multiple alignment. *Bioinformatics*, 17 :803–820, 2001.

[37] I.A. Ibragimov and R.Z. Has'minskii. *Statistical Estimation. Asymptotic Theory.* Applications of Mathematics, Vol. 16. Springer-Verlag, New York - Heidelberg - Berlin, 1981.

[38] J.L. Jensen and A.-M. Pedersen. Probabilistic models for DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob.*, 32 :499–517, 2000.

[39] B.-H. Juang and L.R. Rabiner. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust. Speech Signal Process.*, 38(9) :1639–1641, 1990.

[40] T.H. Jukes and C.R. Cantor. *Evolution of protein molecules in Mammalian Protein Metabolism, pp. 21–132.* H.N. Munro, ed., Academic Press, New York, 1969.

[41] S. Karlin and S. F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA.*, 90 :5873–5877, 1993.

[42] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16(2) :111–120, 1980.

[43] J.F.C. Kingman. The ergodic theory of subadditive stochastic processes. *J. R. Stat. Soc., Ser. B*, 30 :499–510, 1968.

[44] B. Knudsen and M.M. Miyamoto. Sequence alignments and pair hidden Markov models using evolutionary history. *J. Mol. Biol.*, 333 :453–460, 2003.

[45] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology : Applications to protein modelling. *J. Mol. Biol.*, 235 :1501–1531, 1994.

[46] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM P&S*, 8 :115–131, 2004.

[47] D. Kurokawa, H. Kiyonari, R. Nakayama, C. Kimura-Yoshida, I. Matsuo, and Aizawa S. Regulation of Otx2 expression and its functions in mouse forebrain and midbrain. *Development*, 131(14) :3319–3331, 2004.

[48] D. Kurokawa, N. Takasaki, H. Kiyonari, R. Nakayama, C. Kimura-Yoshida, I. Matsuo, and Aizawa S. Regulation of Otx2 expression and its functions in mouse epiblast and anterior neuroectoderm. *Development*, 131(14) :3307–3317, 2004.

[49] B.G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.*, 40(1) :127–143, 1992.

[50] V.I. Levenshtein. Efficient reconstruction of sequences. *IEEE Trans. Inf. Theory*, 47(1) :2–22, 2001.

[51] D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227 :1435–1441, 1985.

[52] G. Lunter, I. Miklos, A. Drummond, J.L. Jensen, and J. Hein. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, pages 6–83, 2005.

[53] G. Lunter, I. Miklos, Y.S. Song, and J. Hein. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol*, 10(6) :869–889, 2003.

[54] G. Lunter, C.P. Ponting, and J. Hein. Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model. *PLoS Comput Biol.*, 13 :2(1) :e5, 2006.

[55] D. Metzler. Statistical alignment based on fragment insertion and deletion models. *Bioinformatics*, 19(4) :490–499, 2003.

[56] D. Metzler, R. Fleißner, A. Wakolbinger, and A. von Haeseler. Assessing variability by joint sampling of alignments and mutation rates. *J. Mol. Evol.*, 53(6) :660–669, 2001.

[57] D. Metzler, R. Fleißner, A. Wakolbinger, and A. von Haeseler. Stochastic insertion-deletion processes and statistical sequence alignment. *Interacting Stochastic Systems. Springer-Verlag Berlin Heidelberg.*, pages 247–267, 2005.

[58] I.M. Meyer and R. Durbin. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics*, 18(10) :1309–1318, 2002.

[59] I. Miklós. An improved algorithm for statistical alignment of sequences related by a star tree. *Bull. Math. Biol.*, 64 :771–779, 2002.

[60] I. Miklós, G. A. Lunter, and I. Holmes. A "Long Indel" Model For Evolutionary Sequence Alignment. *Mol. Biol. Evol.*, 21(3) :529–540, 2004.

[61] R. Mott and R. Tribe. Approximate statistics of gapped alignments. *J. Comput. Biol.*, 6(1) :91–112, 1999.

[62] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48 :443–453, 1970.

[63] L. Pachter, M. Alexandersson, and S. Cawley. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comput. Biol.*, 9(2) :389–399, 2002.

[64] W.R. Pearson. Rapid and sensitive sequence comparaison with FASTP and FASTA. *Methods in enzymology*, 183 :63–98, 1990.

[65] M.J.D. Powell. A fast algorithm for nonlinearly constrained optimization calculations. *Lecture Notes in Mathematics*, 630 :144–157, 1978.

[66] T.A. Reichert, D.N. Cohen, and A.K.C. Wong. An application of information theory to genetic mutations and the matching of polypeptide sequences. *J. Theor. Biol.*, 42 :245–261, 1973.

[67] C. Robert. *Méthodes de Monte Carlo par chaînes de Markov*. Economica, Paris, 1996.

[68] F. Ronquist and J. P. Huelsenbeck. Mrbayes 3 : Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19 :1572–1574, 2003.

[69] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, and Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8) :1034–1050, 2005.

[70] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147 :195–197, 1981.

[71] A. Stathopoulos and M. Levine. Genomic regulatory networks and animal development. *Dev Cell.*, 9(4) :449–462, 2005.

[72] M. Steel and J. Hein. Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. *Appl. Math. Let.*, 14 :679–684, 2001.

[73] L. Sucheston. On mixing and the Zero-One law. *J. Math. Anal. and Appl.*, 6 :447–456, 1963.

[74] B.D. Thatte. Invertibility of the TKF model of sequence evolution. *Mathematical Biosciences*, 200(1) :58–75, 2006.

[75] J. Thompson, D. Higgins, and T. Gibson. CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22 :4673–4680, 1994.

[76] J.L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33 :114–124, 1991.

[77] J.L. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality : an improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34 :3–16, 1992.

[78] A. Wald. Note on the consistency of the maximun likelihood estimate. *Ann. Math. Stat.*, 20 :595–601, 1949.

Résumé : Cette thèse est consacrée à l'estimation paramétrique dans certains modèles d'alignement de séquences biologiques. Ce sont des modèles construits à partir des considérations sur le processus d'évolution des séquences. Dans le cas de deux séquences, le processus d'évolution classique résulte dans un modèle d'alignement appelé pair-Hidden Markov Model (pair-HMM). Dans le pair-HMM les observations sont formées par le couple de séquences à aligner et l'alignement caché est une chaîne de Markov. D'un point de vue théorique nous donnons un cadre rigoureux pour ce modèle et étudions la consistance des estimateurs bayésien et par maximum de vraisemblance. D'un point de vue appliqué nous nous intéressons à la détection de motifs conservés dans les séquences à travers de l'alignement. Pour cela nous introduisons un processus d'évolution permettant différents comportements évolutifs à différents endroits de la séquence et pour lequel le modèle d'alignement est toujours un pair-HMM. Nous proposons des algorithmes d'estimation d'alignements et paramètres d'évolution adaptés à la complexité du modèle. Finalement, nous nous intéressons à l'alignement multiple (plus de deux séquences). Le processus d'évolution classique résulte dans ce cas dans un modèle d'alignement à variables cachées plus complexe et dans lequel il faut prendre en compte les relations phylogénétiques entre les séquences. Nous donnons le cadre théorique pour ce modèle et étudions, comme dans le cas de deux séquences, la propriété de consistance des estimateurs.
Mots-clés : Alignement de séquences, modèles d'évolution, pair-Hidden Markov Model, consistance, estimation par maximum de vraisemblance, estimation bayésienne, algorithme SAEM, algorithmes MCMC.

Abstract: This thesis is devoted to parameter estimation in models for biological sequence alignment. These are models constructed considering an evolution process on the sequences. In the case of two sequences evolving under the classical evolution process, the alignment model is called a pair-Hidden Markov Model (pair-HMM). Observations in a pair-HMM are formed by the couple of sequences to be aligned and the hidden alignment is a Markov chain. From a theoretical point of view, we provide a rigorous formalism for these models and study consistency of maximum likelihood and bayesian estimators. From the point of view of applications, we are interested in detection of conserved motifs in the sequences. To do this we present an evolution process that allows heterogeneity along the sequence. The alignment under this process still fits the pair-HMM. We propose efficient estimation algorithms for alignments and evolution parameters. Finally we are interested in multiple alignment (more than two sequences). The classical evolution process for the sequences provides a complex hidden variable model for the alignment in which the phylogenetic relationships between the sequences must be taken into account. We provide a theoretical framework for this model and study, as for the pairwise alignment, the consistency of estimators.
Keywords: Sequence alignment, evolution models, pair-Hidden Markov Model, consistency, maximum likelihood estimation, bayesian estimation, SAEM algorithm, MCMC algorithms.

Classification AMS : 62F10, 62F12, 62F15, 62M05, 62P10, 92D20