Parameter estimation in pair hidden Markov models.

Ana Arribas-Gil *, Elisabeth Gassiat *, Catherine Matias**

29th September 2005

* UMR CNRS 8628. Équipe Probabilités, Statistique et Modélisation, Université Paris-Sud, Bâtiment 425, Université de Paris-Sud, 91405 Orsay Cedex, France. E-mail: {ana.arribasgil, elisabeth.gassiat}@math.u-psud.fr
** UMR CNRS 8071. Laboratoire Statistique et Génome, Tour Evry 2, 523 pl. des Terrasses de l'Agora, 91 000 Evry, France. E-mail: matias@genopole.cnrs.fr

Abstract

This paper deals 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 being biologically motivated, some restrictions with respect to the full parameter space naturally occur. Existence of two different Information divergence rates is established and divergence property (namely positivity at values different from the true one) 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.

Key words and phrases: Pair-HMM, pair hidden Markov models, sequence alignment, score parameters estimation, TKF evolution model.

1 Introduction

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 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

$$n \vee m \leq l \leq n + m. \tag{1}$$

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.

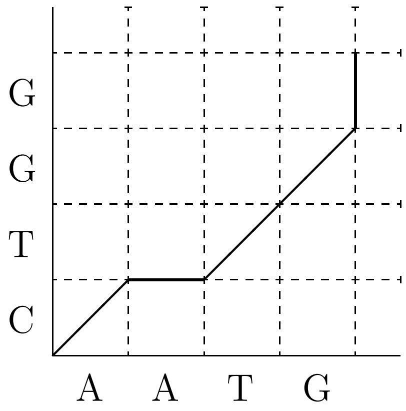

Figure 1: Graphical representation of an alignment between two sequences $X = AATG$ and $Y = CTGG$. The displayed alignment is $\begin{smallmatrix} A & A & TG & - \\ C & - & TG & G \end{smallmatrix}$.

Aligning two sequences relies on the choice of a score optimization scheme (for instance, the Needleman-Wunsch algorithm [Needleman and Wunsch 1970]) 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.

Relying on a pioneering work by Bishop and Thompson [1986], Thorne, Kishino and Felsenstein [1991] 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 fits into the concept of a pair hidden Markov model (pair-HMM), as first formally described in [Durbin, Eddy, Krogh, and Mitchison 1998].

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 [Thorne, Kishino, and Felsenstein 1991], an abundant literature aroused in which parameter estimation occurs in a pair-HMM. Thorne, Kishino and Felsenstein [1992] 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 in [Hein, Wiuf, Knudsen, Moller, and Wibling 2000; Metzler 2003; Knudsen and Miyamoto 2003; Miklos, Lunter, and Holmes 2004], for instance. Let us also mention that pair-HMMs were recently combined with classical hidden Markov models (HMMs) for *ab initio* prediction of genes [Meyer and Durbin 2002; Pachter, Alexandersson, and Cawley 2002; Hobolth and Jensen 2005].

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, Eddy, Krogh, and Mitchison 1998] 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 maximisation technique, to approximate the maximum likelihood estimator (MLE) of the parameters. Numerical maximisation approaches are commonly used ([Thorne, Kishino, and Felsenstein 1991]) but statistical approaches using the Expectation-Maximisation (EM) algorithm and its variants (Stochastic EM, Stochastic Approximation EM) have recently been explored [Holmes 2005; Arribas Gil, Metzler, and Plouhinec 2005]. 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 forward algorithm [Metzler 2003; Arribas Gil, Metzler, and Plouhinec 2005].

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 and Petrie 1966] concerning MLE consistency; see also [Caliebe and Rösler 2002] 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.

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 and MacKay 2001; Levenshtein 2001] on the topic of error correcting codes and also [Csiszár and Körner 1981; Cover and Thomas 1991] 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 a last section, 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. Unfortunately, despite the positive results that we obtain we are not yet in terms of completely validate pair-HMM algorithms in every situation.

2 The pair hidden Markov model

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

$$\begin{aligned}\mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|\varepsilon_{1:t}, \{\varepsilon_s\}_{s>t}, \{X_i, Y_j\}_{i\neq N_s, j\neq M_s, 0\leq s\leq t}) &= \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}|\varepsilon_{1:t}) \\ &= \prod_{s=1}^{t} f(X_{N_s})^{1\!\!1\{\varepsilon_s=(1,0)\}} g(Y_{M_s})^{1\!\!1\{\varepsilon_s=(0,1)\}} h(X_{N_s}, Y_{M_s})^{1\!\!1\{\varepsilon_s=(1,1)\}}, \quad (2)\end{aligned}$$

where $1\!\!1\{\cdot\}$ 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

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

Indeed, if $h = fg$, then (2) gives

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

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.

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 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. 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):

$$\begin{aligned} \mathcal{E}_\infty &= \{(0,1); (1,0); (1,1)\}^{\mathbb{N}} = \{e = (e_1, e_2, \dots)\} = \mathcal{E}^{\mathbb{N}}, && (3)\\ \mathcal{E}_{n,m} &= \{e \in \{(0,1); (1,0); (1,1)\}^l; n \vee m \le l \le n + m; \sum_{i=1}^{l} e_i = (n,m)\}. && (4) \end{aligned}$$

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

$$\mathbb{P}_\theta(X_{1:n}, Y_{1:m}) = \sum_{e \in \mathcal{E}_\infty} \mathbb{P}_\theta(\varepsilon_{1:\infty} = e_{1:\infty}, X_{1:n}, Y_{1:m}), \quad (5)$$

$$\mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}) = \mathbb{P}_\theta(X_{1:N_t}, Y_{1:M_t}, Z_t) = \sum_{e \in \mathcal{E}_{N_t, M_t}; |e| = t} \mathbb{P}_\theta(\varepsilon_{1:t} = e_{1:t}, X_{1:N_t}, Y_{1:M_t}). \quad (6)$$

As Equation (5) 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 2).

We shall now detail Equation (5) 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 last ones need to be constrained in order to avoid multiple counting. Let us denote by $\mathcal{E}^{-H}_{n,m}$ (resp. $\mathcal{E}^{-V}_{n,m}$) the restriction of the set $\mathcal{E}_{n,m}$ to trajectories not ending with an horizontal (resp. vertical) part. More precisely,

$$\begin{aligned} \mathcal{E}^{-H}_{n,m} &= \{e = (e_1, \dots, e_{|e|}) \in \mathcal{E}_{n,m};\ \text{If for some } s,\\ &\qquad e_{|e|} = e_{|e|-1} = \dots = e_{|e|-s+1} = (1,0) \text{ then } s = 0\},\\ \mathcal{E}^{-V}_{n,m} &= \{e = (e_1, \dots, e_{|e|}) \in \mathcal{E}_{n,m};\ \text{If for some } s,\\ &\qquad e_{|e|} = e_{|e|-1} = \dots = e_{|e|-s+1} = (0,1) \text{ then } s = 0\}. \end{aligned}$$

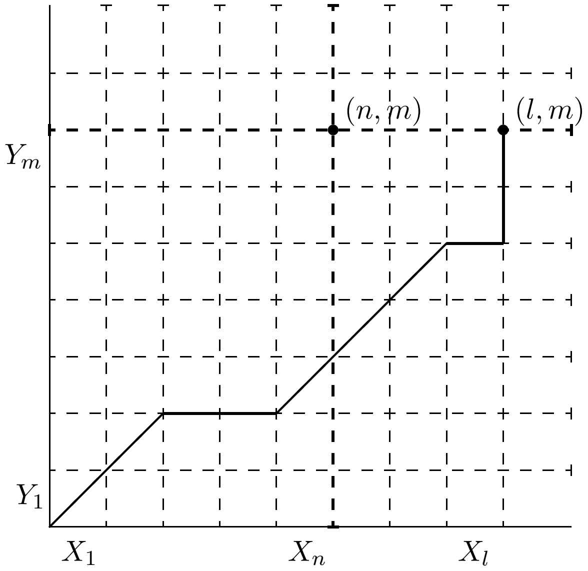

Figure 2: Graphical representation of an alignment of sequences $X_{1:n}$ and $Y_{1:m}$ not passing through the point (n, m).

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.

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

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 $\mathbb{P}^e_\theta$ 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 $n \geq 1, m \geq 1$,*

$$\mathbb{P}_\theta(X_{1:n}, Y_{1:m}) = p\,\mathbb{P}^{(1,0)}_\theta(X_{1:n}, Y_{1:m}) + q\,\mathbb{P}^{(0,1)}_\theta(X_{1:n}, Y_{1:m}) + r\,\mathbb{P}^{(1,1)}_\theta(X_{1:n}, Y_{1:m}), \qquad (7)$$

with the following recursions

$$\begin{aligned}
\mathbb{P}_\theta^{(1,0)}(X_{1:n},Y_{1:m}) &= f(X_1)\{\pi_{HH}\mathbb{P}_\theta^{(1,0)}(X_{2:n},Y_{1:m})+\pi_{HV}\mathbb{P}_\theta^{(0,1)}(X_{2:n},Y_{1:m}) \\
&\quad +\pi_{HD}\mathbb{P}_\theta^{(1,1)}(X_{2:n},Y_{1:m})\} \\
\mathbb{P}_\theta^{(0,1)}(X_{1:n},Y_{1:m}) &= g(Y_1)\{\pi_{VH}\mathbb{P}_\theta^{(1,0)}(X_{1:n},Y_{2:m})+\pi_{VV}\mathbb{P}_\theta^{(0,1)}(X_{1:n},Y_{2:m}) \\
&\quad +\pi_{VD}\mathbb{P}_\theta^{(1,1)}(X_{1:n},Y_{2:m})\} \\
\mathbb{P}_\theta^{(1,1)}(X_{1:n},Y_{1:m}) &= h(X_1,Y_1)\{\pi_{DH}\mathbb{P}_\theta^{(1,0)}(X_{2:n},Y_{2:m})+\pi_{DV}\mathbb{P}_\theta^{(0,1)}(X_{2:n},Y_{2:m}) \\
&\quad +\pi_{DD}\mathbb{P}_\theta^{(1,1)}(X_{2:n},Y_{2:m})\}
\end{aligned}$$

and initializations:

$$\begin{aligned}
\mathbb{P}_\theta^{(1,0)}(X_1) &= f(X_1), \quad \mathbb{P}_\theta^{(0,1)}(Y_1)=g(Y_1), \quad \mathbb{P}_\theta^{(1,1)}(X_1,Y_1)=h(X_1,Y_1), \\
\mathbb{P}_\theta^{(0,1)}(X_{1:n}) &= \frac{1}{1-\pi_{VV}}\{\pi_{VH}\, f(X_1)\mathbb{P}_\theta^{(1,0)}(X_{2:n})+\pi_{VD}\, h_X(X_1)\mathbb{P}_\theta^{(1,1)}(X_{2:n})\}, \\
\mathbb{P}_\theta^{(1,0)}(Y_{1:m}) &= \frac{1}{1-\pi_{HH}}\{\pi_{HV}\, g(Y_1)\mathbb{P}_\theta^{(0,1)}(Y_{2:m})+\pi_{HD}\, h_Y(Y_1)\mathbb{P}_\theta^{(1,1)}(Y_{2:m})\}.
\end{aligned}$$

Proof of Lemma 1 is trivial and therefore omitted.

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

$$\ell_t(\theta)=\log\mathbb{P}_\theta(X_{1:N_t},Y_{1:M_t}), \quad t\geq 1. \tag{8}$$

But since the underlying process $\{Z_t\}_{t\geq 0}$ is not observed, the quantity $\ell_t(\theta)$ is not a measurable function of the observations. 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

$$w_t(\theta)=\log Q_\theta(X_{1:N_t},Y_{1:M_t}), \quad t\geq 1 \tag{9}$$

where for any integers n and m

$$Q_\theta(X_{1:n},Y_{1:m})=\mathbb{P}_\theta(\exists s\geq 1, Z_s=(n,m);X_{1:n},Y_{1:m}). \tag{10}$$

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:

$$Q_\theta(X_{1:n},Y_{1:m})=\sum_{e\in\mathcal{E}_{n,m}}\mathbb{P}_\theta(\varepsilon_{1:|e|}=e,X_{1:n},Y_{1:m}). \tag{11}$$

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 [Durbin, Eddy, Krogh, and Mitchison 1998]) 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 ([Durbin, Eddy, Krogh, and Mitchison 1998]). 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, Kishino, and Felsenstein 1991; Thorne, Kishino, and Felsenstein 1992; Hein, Wiuf, Knudsen, Moller, and Wibling 2000; Metzler 2003; Knudsen and Miyamoto 2003; Miklos, Lunter, and Holmes 2004]). 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 (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.

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 (2) and Assumption 3 gives the first assertion of the Lemma. Proof of the second assertion is similar. ■

3 Information divergence rates

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_0 and Θ_δ, $\delta>0$:

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

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 ([1992], Theorem 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 $(W_{s,t})_{0\leq s<t}$ 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, 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 $0 \leq s < t$,

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

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:

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

$\mathbb{P}_0$-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 ([1992], 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:

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

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, so that the $0-1$ law implies (see [Sucheston 1963]) 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. ■

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:

$$A_t = \left\{(n,m) \in \mathbb{N}^2 \ : \ n \vee m \leq t \leq n+m\right\}. \tag{12}$$

Then,

$$\mathbb{E}_0[w_t(\theta)] = \sum_{(n,m)\in A_t} \sum_{x_{1:n}, y_{1:m}} \mathbb{P}_0(Z_t = (n,m), X_{1:n} = x_{1:n}, Y_{1:m} = y_{1:m}) \log Q_\theta(x_{1:n}, y_{1:m}). \tag{13}$$

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,

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

But

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

where (a) comes from expression (11). Finally,

$$\lim_{t\to+\infty}\frac{1}{t}\left(w_t(\theta)-w_t(\theta_0)\right) \le \liminf_{t\to+\infty}\frac{1}{t}\log\left[\sum_{(n,m)\in A_t}\mathbb{P}_\theta\left(\exists s\ge 1, Z_s=(n,m)\right)\right]. \tag{14}$$

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

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

and

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

Since $\theta\in\Theta_0$, there exists δ_θ such that $\theta\in\Theta_{\delta_\theta}$. By using (11), one gets the lower bound

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

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

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

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

$$4t\log\delta_\theta \le \log Q_\theta(x_{1:n}, y_{1:m}) \le 0. \tag{15}$$

Moreover, with

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

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 $y_{1:m}$,

$$\log Q_\theta(x_{1:n}, y_{1:m}) \leq \frac{t}{2} \log(1-\delta_\theta). \tag{16}$$

Inequalities (15) and (16) allow to conclude that

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

Then, as soon as B_t is a set such that

$$\lim_{t \to +\infty} \mathbb{P}_0\left(Z_t \notin B_t\right) = 0, \tag{17}$$

we have

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

Now, using Jensen's inequality,

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

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,

$$-D\left(\theta|\theta_0\right) \leq \lim_{t \to +\infty} \frac{1}{t} \log \mathbb{P}_\theta(\exists s \geq 1, Z_s \in B_t). \tag{18}$$

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, (17) 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 (18), 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 (11),

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

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 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 $e \in \mathcal{E}_{N_t,M_t}$:

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

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

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

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 $t^{-1}\ell_t$. ■

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 [Thorne, Kishino, and Felsenstein 1991; Thorne, Kishino, and Felsenstein 1992] (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

$$\widehat{\beta}_t = \underset{\beta \in B_\delta}{Argmax}\ w_t(\theta(\beta)).$$

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 and $\bar{\beta}$ a random vector distributed according to ν. 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 $Y_{1:M_t}$:

$$\frac{Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)}{\int_{B_\delta} Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)}. \tag{19}$$

This leads to Bayesian consistent estimation of β_0 as in classical statistical models (see [Ibragimov and Has'minskii 1981] 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

$$\left| \int_{B_\delta} m(\beta) \frac{Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})}{\int_{B_\delta} Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)} \nu(d\beta) - m(\beta_0) \right| \leq \frac{\int_{B_\delta} |m(\beta) - m(\beta_0)| Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)}{\int_{B_\delta} Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)}.$$

But

$$\frac{\int_{\|\beta-\beta_0\|\leq\alpha} |m(\beta) - m(\beta_0)| Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)}{\int_{B_\delta} Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)} \leq \epsilon$$

so that

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

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

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

Using that ν weights β_0 we finally obtain

$$\lim_{t\to\infty}\left|\int_{B_\delta} m(\beta)\frac{Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})}{\int_{B_\delta} Q_{\theta(\beta)}(X_{1:N_t},Y_{1:M_t})\nu(d\beta)}\nu(d\beta)-m(\beta_0)\right| = 0 \quad \mathbb{P}_0 - a.s. \tag{20}$$

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 (20) does not hold is still a null set. Then

$$\nu_{|X_{1:N_t},Y_{1:M_t}} \rightsquigarrow \delta_{\beta_0} \quad \mathbb{P}_0 - a.s. \tag{21}$$

■

5 Simulations

5.1 A simple model

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

$$h(x,y) = \begin{cases} f(x)(1-e^{-\alpha})f(y) & \text{if } x \neq y \\ f(x)\{(1-e^{-\alpha})f(x)+e^{-\alpha}\} & \text{otherwise,} \end{cases} \tag{22}$$

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 (used for instance in [Thorne, Kishino, and Felsenstein 1991]) where 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.

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 0.25. We show in Figure 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).

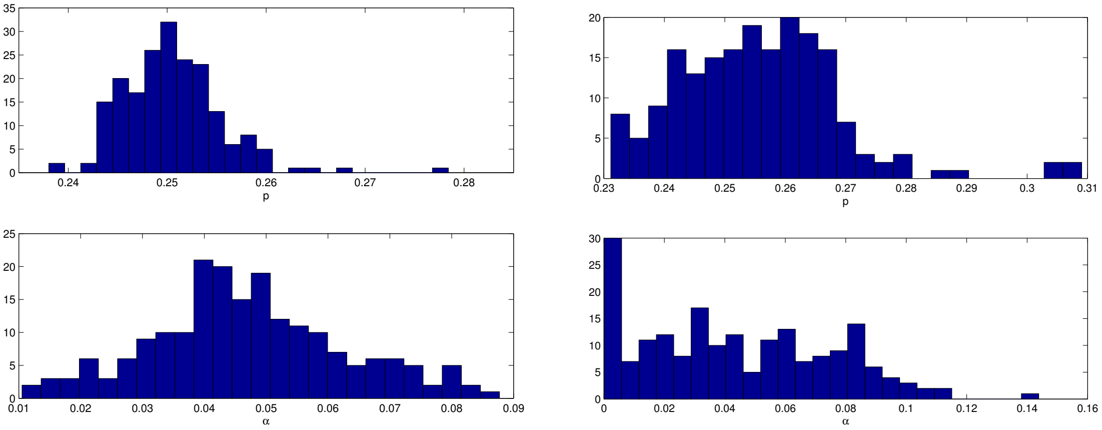

Figure 3: Histograms of maximum likelihood estimations of parameters obtained with 200 simulations from the i.i.d. model. On the left: estimation of p given $\alpha = \alpha_0$ and estimation of α given $p = p_0$. On the right: joint estimation of p and α.

That can be explained by looking at the graph of $w(\beta)$ and comparing it to $\ell(\beta)$ (Figure 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.

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 $(\varepsilon_s)_s$

$$\begin{array}{cc} & \begin{array}{ccc} D & H & V \end{array} \\ \begin{array}{c} D \\ H \\ V \end{array} & \left(\begin{array}{ccc} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.3 & 0.1 & 0.6 \end{array} \right) \end{array}$$

with initial distribution $p_0 = q_0 = 0.25$. 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 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. These results are rather encouraging since the Markov case is the interesting one in biological applications.

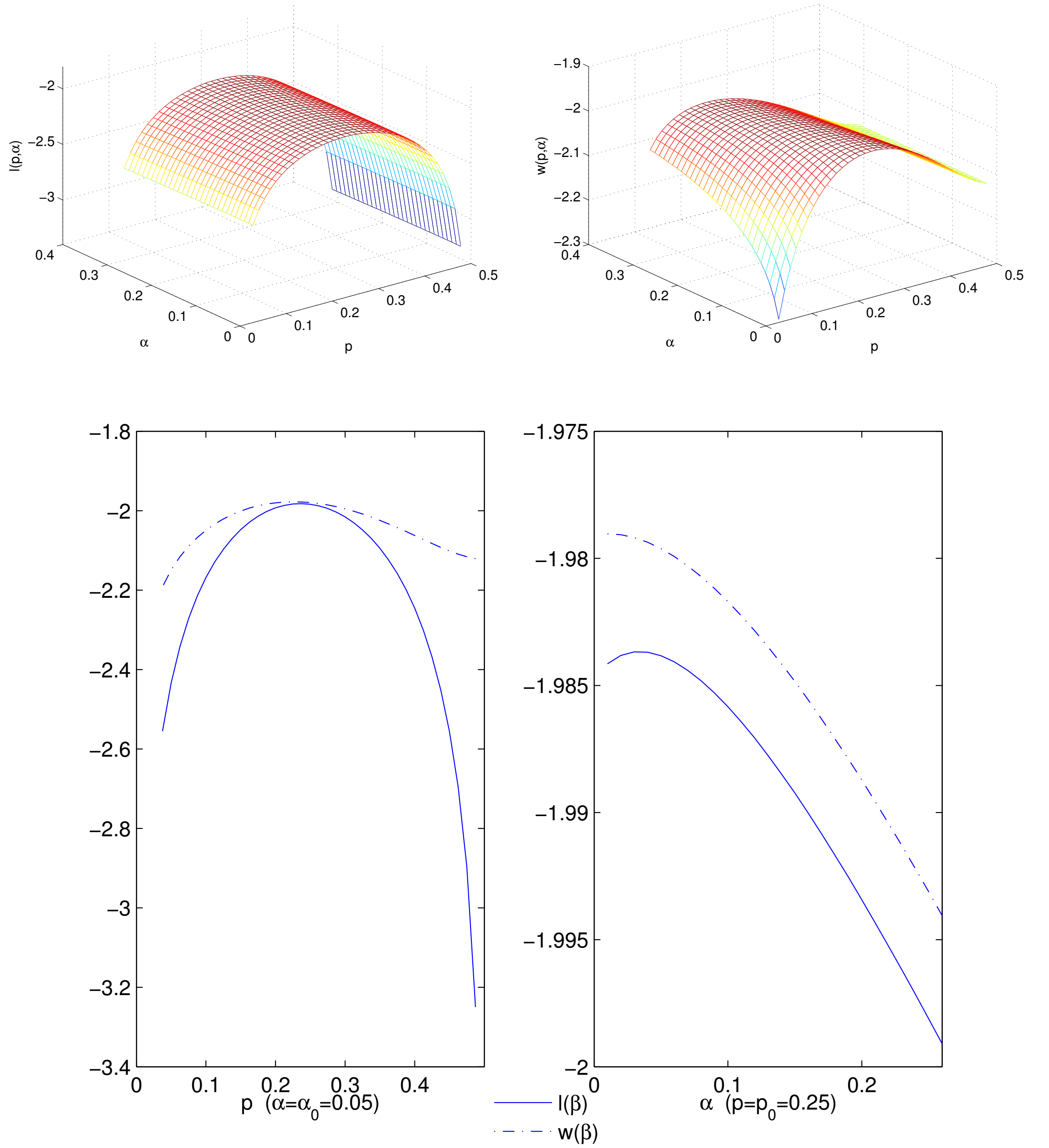

Figure 4: On top: ℓ and w for the i.i.d. model ($p_0 = 0.25, \alpha_0 = 0.05$). On bottom: cuts of ℓ and w for $\alpha = \alpha_0$ fixed and for $p = p_0$ fixed.

References

Arribas Gil, A., D. Metzler, and J.-L. Plouhinec (2005). A fragment insertion and deletion model allowing fast and slow fragments. Manuscript.

Baum, L. and T. Petrie (1966). Statistical inference for probabilistic functions of finite

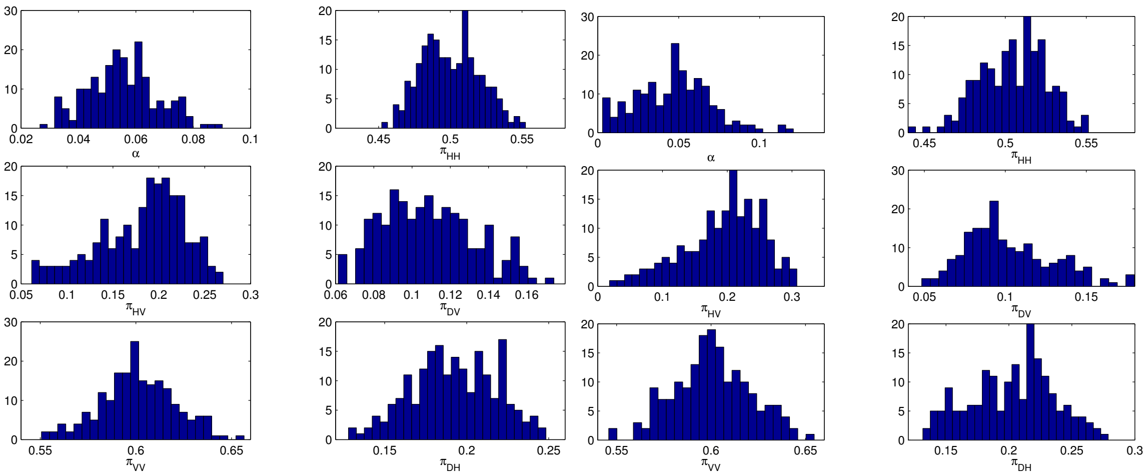

Figure 5: Histograms of maximum likelihood estimations of parameters obtained with 200 simulations from the Markov chain model. On the left: estimation of the transition probabilities given $\alpha = \alpha_0$ and estimation of α given the true value of the transition probabilities. On the right: joint estimation of the transition probabilities and α.

state Markov chains. *Ann. Math. Statist. 37*, pp. 1554–1563.

Bishop, M. and E. Thompson (1986). Maximum likelihood alignment of DNA sequences. *J. Mol. Biol. 190*, pp. 159–165.

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

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

Csiszár, I. and J. Körner (1981). *Information theory. Coding theorems for discrete memoryless systems.* New York-San Francisco-London: Probability and Mathematical Statistics. Academic Press.

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

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

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

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

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

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

Kingman, J. F. C. (1968). The ergodic theory of subadditive stochastic processes. *J. Roy. Statist. Soc. Ser. B 30*, pp. 499–510.

Knudsen, B. and M. Miyamoto (2003). Sequence alignments and pair hidden Markov models using evolutionary history. *J. Mol. Biol. 333*, pp. 453–460.

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

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

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

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

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

Needleman, S. and C. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol. 48*, pp. 443–453.

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

Sucheston, L. (1963). On mixing and the Zero-One law. *J. Math. Anal. and Appl. 6*, pp. 447–456.

Thorne, J., H. Kishino, and J. Felsenstein (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol. 33*, pp. 114–124.

Thorne, J., H. Kishino, and J. Felsenstein (1992). Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol. 34*, pp. 3–16.