Statistical alignment based on fragment insertion and deletion models

Dirk Metzler
Johann Wolfgang Goethe-Universität, Fachbereich Mathematik
Frankfurt am Main, Germany
www.math.uni-frankfurt.de/∼stoch/metzler
dmetzler@math.uni-frankfurt.de

October 7, 2002

Abstract

Motivation The topic of this paper is the estimation of alignments and mutation rates based on stochastic sequence-evolution models that allow insertions and deletions of subsequences ("fragments") and not just single bases. The model we propose is a variant of a model introduced by Thorne, Kishino, and Felsenstein (1992). The computational tractability of the model depends on certain restrictions in the insertion/deletion process; possible effects we discuss.

Results The process of fragment insertion and deletion in the sequence-evolution model induces a hidden Markov structure at the level of alignments and thus makes possible efficient statistical alignment algorithms. As an example we apply a sampling procedure to assess the variability in alignment and mutation parameter estimates for HVR1 sequences of human and orangutan, improving results of previous work. Simulation studies give evidence that estimation methods based on the proposed model also give satisfactory results when applied to data for which the restrictions in the insertion/deletion process do not hold.

Availability The source code of the software for sampling alignments and mutation rates for a pair of DNA sequences according to the fragment insertion and deletion model is freely available from
www.math.uni-frankfurt.de/∼stoch/software/mcmcsalut
under the terms of the GNU public license (GPL, 2000).

Contact dmetzler@math.uni-frankfurt.de

Key words: sequence alignment, Thorne Kishino Felsenstein model, statistical alignment, hidden Markov model, pair HMM, mutation parameter estimation, Markov chain Monte Carlo Method, hypervariable region, FID model

1 Introduction

To find a good alignment for a pair of DNA (or protein) sequences, one needs to have an idea of the mutation rates. This is true even when using score optimizing alignment algorithms (cf. Needleman, Wunsch 1970) that do not explicitly refer to a model of the insertion-deletion process, since prior opinions about mutation rates are reflected in the choice of the mismatch and gap penalties. On the other hand, mutation parameters are usually estimated from aligned sequences. Thus the choice of parameters for alignment algorithms can bias mutation parameter estimates (cf. Fleißner et al., 2000). Given a stochastic alignment model one can avoid this dilemma by estimating the mutation parameters from unaligned sequences. The likelihoods are then computed by summing the posterior probabilities of all possible alignments, using a dynamic programming approach (Thorne et al. 1991). In Metzler et al. (2001) alignments and mutation parameters are estimated simultaneously, and a Markov chain Monte Carlo sampling strategy is applied to assess the variability of such estimates. Like several other "statistical alignment" procedures (see for example Thorne et al., 1991, 1992, Hein et al., 2000, Hein, 2001, Holmes and Bruno, 2001) this procedure is based on a model proposed by Thorne, Kishino, and Felsenstein (1991), the TKF1 model, which generates a pair-HMM structure (Durbin et al. 1998) on the level of alignments. Thus it is compatible with some efficient algorithms for sequence alignment and mutation parameter estimation.

In the present paper we show how, for example, the sampling method of Metzler et al. (2001) can be adapted to a more general model to overcome the major drawback of the TKF1 model, namely, that it allows insertions and deletions of no more than a single nucleotide at a time. Since this seems unrealistic for many data sets, Thorne et al. (1992) have extended the TKF1 model to the TKF2 model, which describes the insertion and deletion process of longer fragments. Surprisingly, the TKF2 model has not become as popular as the TKF1 model. We can only speculate about the reasons. One of them might be the assumption of indivisible fragments: a fragment that has once been inserted can only be deleted as a whole, and no other fragments can be inserted in between it. We shall argue that one should not too much worry about these assumptions, which are necessary to obtain a pair-HMM structure on the alignments and thus make possible efficient computation. Computer simulations show that mutation parameter estimation procedures which are optimized for a model using these assumptions also work quite well when applied to data generated without the fragmentation restrictions.

Another drawback of the TKF2 model might be seen in the fact that it needs an extra parameter: in addition to the two parameters of the TKF1 model (insertion rate and deletion rate), TKF2 needs a third one, the average fragment length. In both models the deletion rate must be slightly higher than the insertion rate, but when applied to data, the differences between the estimates for the two parameters are negligible. Therefore we use a variant of the TKF2 model with one parameter for the mean fragment length and only one more parameter for the insertion and deletion (indel) rate (cf. section 2).

For simplicity we focus on DNA sequences, but the insertion-deletion-models and the methods are also applicable to protein sequences.

2 Model

Our model of fragment insertions and deletions (FID) has two parameters: $\gamma \geq 1$, the expected fragment length, and $\lambda \geq 0$, the indel rate per site. The fragment insertion and deletion process operates on sequences of "sites". The rightmost site in the ancestral sequence is marked as "fragment end". Each other site in the ancestral sequence is marked as "fragment end" independently with probability $1/\gamma$. In this way, the ancestral sequence is partitioned into *fragments*. In the course of evolution, each fragment end is selected at rate 2λ. When this happens a fair coin is tossed: With probability 1/2 the fragment is deleted. Otherwise, a new fragment is inserted to its right. In addition, new fragments are inserted with rate λ to the left of the first site. The length of the new fragment is geometrically distributed with expectation γ, i. e. the probability that the fragment length equals k is $(1-\gamma^{-1})^{k-1}\gamma^{-1}$ for $k \geq 1$. The rightmost site of the new fragment is marked as fragment end.

2.1 The FID-model and the TKF2-model

The FID-model is a slight modification of the TKF2-model (Thorne et al., 1992). Some of the resulting differences are only apparent: the same reality is modelled in slightly different ways. In the FID-model insertion and deletion rates are equal and thus we have no equilibrium distribution on the space of finitely long sequences. When sequence data are given, we assume that the sequences were cut out of very much longer sequences between known homologous positions. In the TKF models it is assumed that the ancestral sequence is taken from the stationary distribution of the process of the TKF model. Thus, its length is geometrically distributed. In the FID model it is only assumed that the base types of the ancestral sequence are in the equilibrium of the substitution process. The lengths of the given sequences are considered to be non-random. We will come back to this at the end of section 3.2.

2.2 Alignments and homology structures

Like Thorne et al. (1991, 1992) we consider insertions as happening to the right of positions rather than between positions and consider the inserted fragments as offspring of their left neighbors. So when notating alignments we apply the "TKF-convention": We write inserted fragments directly to the right of their "ancestors". For example, if in the sequence `ATTAGAA` the fragment `TT` is deleted and a fragment `GCC` is inserted later at its place (and the `G` in the first sequence is deleted and the last `A` substituted by a `C`), the resulting alignment must be denoted as $\begin{smallmatrix}\texttt{A\_\_\_TTAGAA}\\ \texttt{AGCC\_\_A\_AC}\end{smallmatrix}$ instead of $\begin{smallmatrix}\texttt{ATT\_\_\_AGAA}\\ \texttt{A\_\_GCCA\_AC}\end{smallmatrix}$.

This results in a pair-HMM-structure (from left to right) in the alignments and makes it possible to use some efficient algorithms, as we shall see in section 3.2.

Usually, the order of gaps between two homologous pairs of sites is not relevant. We can describe the homology structure by the numbers of unaligned positions in the sequences between each pair of homologous sites. Thus the homology structure of the alignment $\begin{smallmatrix}\texttt{A---TT|AG|A|A}\\ \texttt{AGCC--|A-|A|C}\end{smallmatrix}$ is given by the sequence of number pairs (2,3),(1,0),(0,0),(0,0) meaning that two unaligned sites in the ancestral sequence and three unaligned sites in the other sequence follow the first homologous pair and after the second homologous pair there is only one unaligned site in the ancestral sequence. The homology structure of the alignment $\begin{smallmatrix}\texttt{A---TT-|AG}\\ \texttt{AGCC--G|A-}\end{smallmatrix}$ is (2,4),(1,0). Note that the homology structures of sequence pairs which have evolved according to the FID-model are i. i. d. sequences of number pairs and the distribution on the number pairs can be computed efficiently using the pair-HMM structure (cf. section 3.2).

In the limit of (infinitely) long sequences, the FID evolution model is time reversible. Note that the TKF-convention breaks this reversibility on the level of alignments, but the homology structures are still time reversible: when we jump randomly into a long homology structure and consider the next k number pairs, then $(n_1, m_1), (n_2, m_2), \ldots, (n_k, m_k)$ is as probable as $(m_1, n_1), (m_2, n_2), \ldots, (m_k, n_k)$ in the FID model. Since the homology structure is the interesting object, we can always act as if one sequence was the ancestor of the other one, even if both sequences stem from some common ancestor.

2.3 Substitution models to go with FID

As long as the substitution model is not specified, the FID model only induces a probability distribution on the *bare* alignments, which ignore the base types. The bare alignments for the above examples are $\begin{smallmatrix}\texttt{B---BBBBBB}\\ \texttt{BBBB--B-BB}\end{smallmatrix}$ and $\begin{smallmatrix}\texttt{B---BB-BB}\\ \texttt{BBBB--BB-}\end{smallmatrix}$, where `B` stands for "base".

The substitution model should be Markovian and reversible in time, and all sites should evolve independently and with identical distributions. When a new site is inserted, its base type is drawn randomly from the equilibrium distribution of the substitution process. Given the homology structure, the substitution process is independent of the alignment. The base types (or amino acid residuals in the case of protein sequences) at all sites in the ancestral sequence are assumed to be drawn independently from the equilibrium distribution. Examples for substitution models fulfilling these requirements are the Jukes Cantor model, Kimura's models, and the model of Hasegawa, Kishino and Yano (cf. Hasegawa et al., 1985, Swofford et al., 1996, Zharkikh, 1994).

3 Algorithm

In Metzler et al. (2001) it was shown that the variability in mutation parameter estimates can be strongly underestimated if the estimated alignment is assumed to be the

true one. In order to assess the actual variability in the parameter estimates one has to take into account that the true alignment is unknown and probably different from the most plausible one. In addition, the use of "optimal" alignments can cause a bias in mutation rate estimates (cf. Fleißner et al., 2000). The algorithm described in Metzler et al. (2001) was based on the TKF1 model (Thorne et al., 1991), allowing only insertions and deletions of single nucleotides. Here we adapt the sampling algorithm to the FID model. This is possible since the FID model induces a pair-HMM structure on the alignments, as we will show in section 3.2. This means that one could generate a bare alignment according to the FID-induced distribution by building it from left to right according to certain Markov chain transition probabilities $P(x \to y)$ for $x, y \in \{ \binom{B}{B}, \binom{B}{\_}, \binom{\_}{B} \}$ instead of simulating the insertion deletion process. The aligned sequences are then "emitted" from the positions in the bare alignment according to the substitution model.

Let sequences s_1 and s_2 of lengths n and m be given.

At first we assume that the mutation parameters are also given. We can then sample alignments according to their posterior probability by the following procedure. For $0 \leq i \leq n$ and $0 \leq j \leq m$ let $f(i,j,x)$ be the probability that a bare alignment generated according to FID begins with an alignment of i positions which coincide with the first i positions in s_1 against j positions which coincide with j positions in s_2, and that the last bare alignment state in this partial alignment is x. Since we assume that the observed sequences s_1 and s_2 are cut out of longer sequences between homologous positions we set $f(0,0,x) = 1$ for $x = \binom{B}{B}$ and $f(0,0,x) = 0$ otherwise. The pair-HMM structure implies $f(i,j,\binom{B}{\_}) = \sum_{x \in \{\binom{B}{B}, \binom{B}{\_}, \binom{\_}{B}\}} f(i-1,j,x) \cdot P(x \to \binom{B}{\_}) \cdot e_i$ and $f(i,j,\binom{B}{B}) = \sum_{x \in \{\binom{B}{B}, \binom{B}{\_}, \binom{\_}{B}\}} f(i-1,j-1,x) \cdot P(x \to \binom{B}{B}) \cdot e_{ij}$, where e_i and e_{ij} depend on the substitution model: e_i is the emission probability of the base at the ith position in sequence s_1 and e_{ij} is a probability that a pair of homologous sites emits the bases at site i in s_1 and at site j in s_2. With the analogous formula for $f(i,j,\binom{\_}{B})$, we can efficiently compute all values for $f(.,.,.)$ iteratively while increasing i and j from $i = j = 0$ up to $(i,j) = (n,m)$. Since we assume that there are homologous sites to the right of our observed sequences, we set $f(n+1,m+1,\binom{B}{B}) = \sum_x f(n,m,x) P(x \to \binom{B}{B})$. For the computation of $f(i,j,x)$ we need the mutation parameter values, since $P(x \to y)$ depends on λ and γ and the emission probabilities e_i and e_{ij} depend on the substitution rates. Once we have $f(.,.,.)$ we can easily sample an alignment according to its conditional probability given the sequences. We generate it from right to left, starting with a triple (n,m,X), where the random $X \in \{ \binom{B}{B}, \binom{B}{\_}, \binom{\_}{B} \}$ is chosen to be x with probability $f(n,m,x)P(x \to \binom{B}{B}) / f(n+1,m+1,\binom{B}{B})$. We iterate this procedure: Given that the last chosen triple was (i,j,x) we draw the next one according to a probability distribution that reflects the relative contributions to $f(i,j,x)$ in the above sums. The alignment is complete when we draw a triple (i,j,x) with $i,j \leq 1$.

If the alignment is given and values of the mutation parameters are to be sampled, we can apply a Metropolis-Hastings strategy (cf. Gamerman, 1997): We start with initial estimates of the mutation rates. New values for them are proposed randomly

from geometric distributions with the old values as mean and accepted or rejected with a probability depending on the likelihood ratio and prior probabilities of the new and the old values, such that the iterations of this step form a Markov chain that converges to the posterior probability distribution of the parameters.

When – as usual – neither the alignment nor the parameters are given, we apply the idea of Gibbs sampling (cf. Gamerman 1997) to combine the two sampling schemes: We start with estimates of the mutation rates and sample an alignment. Then we apply a Metropolis-Hastings step to the mutation parameters. Then we randomly pick a part of the alignment of, say, about 30 bp and resample the alignment in this section, relying on the current parameter values. (We could also resample the alignment completely, but we would then have to calculate $O(nm)$ values of $f(,,)$ always when the parameter values change.) Thus we obtain a Markov chain on the tuples consisting of alignments and parameter values that converges to their joint posterior probability distribution.

3.1 Example: HVR-1 from human and orangutan

The FID model sheds new light on the HVR-1 example in Metzler et al. (2001) where we analyzed the human sequence ID 1244 (Anderson et al. 1981) and the orangutan sequence ID 389 (Xu et al. 1996) from the HVRBASE (Handt et al. 1998, see also `http://db.eva.mpg.de/Hvrbase`) on the base of the TKF1 model.

The sampling strategy is now based on the FID model and Felsenstein's substitution model "F84" (Swofford et al., 1996, Felsenstein, Churchill, 1996): the rate of transitions is increased compared to transversions by allowing transitions in addition to a general type of substitution that could change a nucleotide into any other base type (according to the base type probabilities). The rates of general substitutions and of transitions were sampled jointly with the indel process parameters λ and γ and the alignments according to their posterior probability using a Metropolis-Hastings-strategy. From the observed base frequencies we estimated the base type probabilities $\pi_{\mathtt{A}} = 0.3, \pi_{\mathtt{C}} = 0.35, \pi_{\mathtt{G}} = 0.15$ and $\pi_{\mathtt{T}} = 0.2$. Using the freedom of time scaling we assume that the time distance between the sequences is $t = 1$, such that the mutation rates are the expected numbers of mutations per site. We used exponential prior distributions on the general substitution rate, the transition rate and indel rate λ with expectation 1. This makes the probabilities that no mutation occurs at a given site uniform on the interval [0,1] for each of the three mutation types. We also used an exponential prior for $\gamma - 1$, the expected number of gap extensions. Since there was no such natural choice for its expectation as with the mutation rates we tried three different values: 0.5, 5 and 50. We call the latter choice a nearly flat prior since in this case the density of the prior is almost flat near relevant values for γ.

As initial alignment for the Markov chain Monte Carlo (MCMC) procedure we used the alignment given in the data base. The initial values of the mutation rates were 0.05 for the general substitution rate and 0.01 for λ and the transition rate. For the initial value of γ we chose the expectation value of its prior. After an initial run of 10000 steps ("burn in") we sampled 1000 alignments and corresponding mutation

parameters, performing 1000 steps between each two samples.

It is *not* our aim to come up with only *one* alignment and/or set of mutation parameter values which one should believe in. We rather want to assess the variability inherent in such estimates.

Figure 1 shows that in some positions more than 80% of the sampled alignments (using the most uninformative of the priors for γ) do not coincide with the most probable of the sampled alignments. (With other priors for γ we obtained slightly different most probable alignments but the regions of uncertainty were essentially the same.) For the alignment given in the data base the non-coincidence was even slightly higher (see Figure A on
`www.math.uni-frankfurt.de/~stoch/software/mcmcsalut/figures.html`).

Figure 2 shows the mutation parameters that were sampled simultaneously with the alignments. Obviously, the choice of the prior for γ has an effect on the distributions of the sampled values for λ and γ. This is no surprise because the data for estimating λ and γ are rare since there are only a few gaps in the probable alignments of the sequences. The effect influences mainly the length of the right tail of the distributions and is rather weak around the modes.

3.2 The Markov property of alignments in the FID-model

The alignment of two sequences that have evolved according to the FID model (or one of the TKF models) is a Markov chain on the states $\binom{-}{B}$ (insertion), $\binom{B}{-}$ (deletion) and $\binom{B}{B}$ (homologous sites). One way to understand this Markovian structure is to use ideas from the theory of Galton Watson processes (Harris, 1963, Geiger, 1996). In fact, the offspring population of a fragment in the FID model forms a Galton Watson process which is critical (i. e. the expected population size is constant in time) because of the equality of the rates of insertions and deletions. Given that the offspring population of a fragment is not extinct at some time $t \geq 0$, its size is geometrically distributed.

This can easily be seen: if there are any survivors, take the leftmost of them, where the left-right scheme follows the TKF-alignment notation convention. Consider the branch b_1 from the ancestor to this survivor in the phylogenetic tree of the offspring population (the dashed lines in the left of Figure 3). Nothing survives to the left of b_1, which has length t. It produces offspring to the right (the grey area in Figure 3). If one of them survives, consider again the leftmost of them and call the branch from the root to it b_2 (the dashed lines in the middle of Figure 3). Given that this b_2 exists, it has length t again and starts together with b_1 at time 0. Therefore we are again in the same situation as before. Iterating this argument we see that if X is the size of the offspring population, then the distribution of $X - n$, given $X \geq n$ for some $n \geq 1$, does not depend on n. This "lack of memory" characterizes the geometric distribution.

We still have to answer two questions: what is the parameter of the geometric distribution and what is the probability that $X > 0$? Both are easy to answer because the expected generation size of a critical Galton Watson process is a constant over time: $\mathbb{E}X = 1$. Given that the leftmost survivor exists (for instance if the ancestor

$P(x \to y)$	$y = \begin{smallmatrix}\mathtt{B}\\ \mathtt{B}\end{smallmatrix}$	$y = \begin{smallmatrix}\bar{}\\ \mathtt{B}\end{smallmatrix}$	$y = \begin{smallmatrix}\mathtt{B}\\ \_\end{smallmatrix}$
$x = \begin{smallmatrix}\mathtt{B}\\ \mathtt{B}\end{smallmatrix}$	$1 - \frac{1+\lambda-e^{-\lambda}}{\gamma(1+\lambda)}$	$\frac{\lambda}{\gamma(1+\lambda)}$	$\frac{1-e^{-\lambda}}{\gamma(1+\lambda)}$
$x = \begin{smallmatrix}\bar{}\\ \mathtt{B}\end{smallmatrix}$	$\frac{e^{-\lambda}}{\gamma(1+\lambda)}$	$\frac{\gamma(1+\lambda)-1}{\gamma(1+\lambda)}$	$\frac{1-e^{-\lambda}}{\gamma(1+\lambda)}$
$x = \begin{smallmatrix}\mathtt{B}\\ \_\end{smallmatrix}$	$\frac{\lambda e^{-\lambda}}{\gamma(1-e^{-\lambda})(1+\lambda)}$	$\frac{1-e^{-\lambda}(1+\lambda)}{\gamma(1-e^{-\lambda})(1+\lambda)}$	$\frac{\gamma(1+\lambda)-1}{\gamma(1+\lambda)}$

Table 1: The transition probabilities for alignment states from left to right in the FID-model (for $t = 1$).

itself is not deleted) the expected number of birth events that happen on its branch is λt. Each of the born individuals has an offspring population of expected size 1 at any later time. This implies $\mathbb{E}(X \mid X > 0) = 1 + \lambda t$. Thus the parameter of the geometric distribution is $1/(\lambda t + 1)$. From

$$\begin{aligned} 1 &= \mathbb{E}X = \Pr(X > 0) \cdot \mathbb{E}(X|X > 0) + \Pr(X = 0) \cdot \mathbb{E}(X|X = 0) \\ &= \Pr(X > 0) \cdot (1 + \lambda t) + \Pr(X = 0) \cdot 0 \end{aligned}$$

we get $\Pr(X > 0) = 1/(\lambda t + 1)$. Since each fragment has geometrically many sites with expectation γ and since the sum of geometrically many i. i. d. geometrically distributed random variables is also geometrically distributed, we get the following result. Let Y be the number of sites in the offspring population of some fragment at time t, then given $Y > 0$, Y is geometrically distributed with $\mathbb{E}Y = (\lambda t + 1) \cdot \gamma$. Since the offspring populations of different fragments are independent of each other and since the geometric distribution is memoryless, we obtain the Markov property of the alignments in the FID model.

The computation of the transition probabilities between the states $\begin{smallmatrix}\bar{}\\ \mathtt{B}\end{smallmatrix}$, $\begin{smallmatrix}\mathtt{B}\\ \_\end{smallmatrix}$ and $\begin{smallmatrix}\mathtt{B}\\ \mathtt{B}\end{smallmatrix}$ is not very difficult, as the following examples show. (As in section 3.1 we set $t = 1$). If we are in $\begin{smallmatrix}\mathtt{B}\\ \mathtt{B}\end{smallmatrix}$, the probability that the next state is $\begin{smallmatrix}\bar{}\\ \mathtt{B}\end{smallmatrix}$ is the product of the probability that the current fragment ends and the probability that the fragment has another surviving offspring: $P(\begin{smallmatrix}\mathtt{B}\\ \mathtt{B}\end{smallmatrix} \to \begin{smallmatrix}\bar{}\\ \mathtt{B}\end{smallmatrix}) = \gamma^{-1} \cdot (1 - (\lambda + 1)^{-1})$.

$P(\begin{smallmatrix}\mathtt{B}\\ \mathtt{B}\end{smallmatrix} \to \begin{smallmatrix}\mathtt{B}\\ \_\end{smallmatrix})$ is the product of the probability for the end of the fragment, the probability that it has no further offspring and the probability of the death of the next fragment before time $t = 1$. The result is $\gamma^{-1} \cdot (\lambda + 1)^{-1} \cdot (1 - \exp(-\lambda))$.

Let p be the probability for a fragment that dies before $t = 1$ to have an offspring at $t = 1$. Then we have $P(\begin{smallmatrix}\mathtt{B}\\ \_\end{smallmatrix} \to \begin{smallmatrix}\bar{}\\ \mathtt{B}\end{smallmatrix}) = \gamma^{-1} \cdot p$. Together with $1/(\lambda + 1) = \Pr(X > 0) = e^{-\lambda} + (1 - e^{-\lambda}) \cdot p$ this implies $P(\begin{smallmatrix}\mathtt{B}\\ \_\end{smallmatrix} \to \begin{smallmatrix}\bar{}\\ \mathtt{B}\end{smallmatrix}) = \frac{1-e^{-\lambda}(1+\lambda)}{\gamma(1-e^{-\lambda})(1+\lambda)}$.

All transition probabilities are given in table 1. Note that these probabilities are the limits of those in the TKF2 model when the deletion rate converges to the insertion rate λ.

As already mentioned in section 2.1, we assume that the sequences are taken from a region between known positions. Therefore we let the Markov chain start in a $\begin{smallmatrix}\mathtt{B}\\ \mathtt{B}\end{smallmatrix}$ state (that does not emit base types to the sequences). Unlike the TKF models, the

FID model provides no END state for the Markov chain. Instead, this is covered by conditioning on the data: Given sequences of length n and m, the Markov chain is conditioned on emitting the base types *and* to jump to $\begin{smallmatrix}\mathtt{B}\\\mathtt{B}\end{smallmatrix}$ after the total number of jumps to $\begin{smallmatrix}\mathtt{B}\\\mathtt{B}\end{smallmatrix}$ and $\begin{smallmatrix}\mathtt{B}\\\_\end{smallmatrix}$ was n and the total number of jumps to $\begin{smallmatrix}\mathtt{B}\\\mathtt{B}\end{smallmatrix}$ and $\begin{smallmatrix}\_\\\mathtt{B}\end{smallmatrix}$ was m. This differs slightly from the original pair-HMM concept but is still compatible with the dynamic programming algorithms, as we saw before.

4 Robustness

To achieve computational tractability the FID model makes an unnatural assumption: when a fragment has been inserted into a sequence, it can only be deleted as a whole and new fragments can only be inserted left or right of the first fragment but not into it. This assumption might be unrealistic in principle; its effect in practice is unclear. We therefore explore how well mutation-parameter estimators based on the FID model work for data generated without these restrictions.

4.1 A more general insertion deletion (GID) model

Set $\mu := \lambda/\gamma$. The GID model is like FID but without fixed fragmentation. Each site is selected with rate 2μ, then a coin is tossed to decide if randomly many sites are deleted, ending with the selected one, or inserted to the right of the selected one. The number of inserted or deleted sites is geometrically distributed with expectation γ.

The GID model is time reversible in the same sense as the FID model. As in the case of the FID model, the TKF-convention breaks this reversibility on the level of alignments but not on the level of homology structures. Note that the homology structure is also an i.i.d. sequence of number pairs when the sequences evolved according to the GID model, but it seems hard to compute the probability of each number pair in this case.

To simulate a pair of sequences (s_1, s_2) of a desired approximate length n, we first generate a sequence $\tilde{s_1}$ of length $n + 2k$ (for sufficiently large k) according to the equilibrium of the substitution process and let it evolve to a sequence $\tilde{s_2}$. From site $n + 1$ of sequence $\tilde{s_1}$ we move left and from position $n + k$ we move right until we come to positions that are homologous to positions in $\tilde{s_2}$. The subsequences between these homologous pairs are s_1 and s_2.

4.2 Robustness of the ML estimator for the mutation parameters

Suppose we want to estimate mutation parameters (λ and γ and/or substitution rates) from a pair of unaligned sequences. If the sequences were generated according to the FID-model, we could use the maximum likelihood (ML) estimator as in Thorne et al. (1991,1992). How good is the estimator if the sequences were actually generated by a GID instead of a FID process? As a first step, we checked the robustness of the

FID-ML-estimator for λ and γ based on the homology pattern instead of the sequence pairs. Some of the results are shown in Figures 4, 5, and 6. Again we used the time scaling $t = 1$ such that mutation rates are the expected numbers of mutations per site.

The FID-ML-estimator works for GID-generated homology structures almost as well as for those that were generated according to FID except for very high values for λ and γ ($\lambda = 0.5$, $\gamma = 3$, cf. Figure 6). But keep in mind that we usually do not observe homology structures but only sequences. For amounts of mutations like those mentioned above, most information is lost on the way from the homology structures to the sequences. Thus the parameter values where FID and GID lead to substantially different estimates could be irrelevant because they cannot reasonably be estimated from sequence data anyway. To illustrate this we applied the FID-ML-estimator to sequence pairs that were generated with the same parameters $\lambda = 0.5$ and $\gamma = 3$ as in Figure 6 and a substitution rate of $s = \lambda = 0.5$. For unaligned positions we drew the base uniformly from $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and for homologous positions we chose the base of the ancestral sequence uniformly and let it evolve according to the Jukes Cantor model: Substitutions occur with rate s, and then the new base is drawn uniformly from $\{\mathtt{A,C,G,T}\}$. The left side of Figure 7 shows the result of an experiment where 100 pairs of sequences of length $\sim$1000 were generated for the FID model and the FID-ML-estimator for λ and γ was applied to the unaligned sequence pairs. The same was done in the right side of Figure 7 with sequences generated according to the GID model. (For optimization we used Nelder and Mead's simplex algorithm as suggested in Thorne et al. 1991, see also Press et al. 1988.) We see that if the true parameters (s, λ, γ) are as high as $(0.5, 0.1, 3)$, the FID-ML-estimator is so variable that it does not become much worse when it is applied to data generated by GID instead of FID.

4.3 Robustness of the sampling algorithm

Now we check how much difference it makes when the FID sampling algorithm is applied to data generated by GID. We generated 30 pairs of sequences of length 1000 according to the FID model with Jukes Cantor substitution dynamics with parameters values $(s, \lambda, \gamma) = (0.5, 0.1, 3)$ and sampled for each sequence pair 50 alignment-mutation parameter combinations. Then we did the same again with sequences generated according to GID. For s and λ we used exponential priors with expectation 1.0 and for γ an exponential prior with expectation 100.0. The results for the GID data were highly similar to those for the FID data (see `www.math.uni-frankfurt.de/~stoch/software/mcmcsalut/figures.html`). Also for simulations with other parameter values ($(0.1, 0.1, 3)$ and $(0.5, 0.5, 3)$), there was a high similarity between the results in the FID and the GID cases (data not shown).

From these observations we cannot conclude that the sampling method is robust as an estimator for the posterior distribution of the parameters. To show that we would need to compare the results from the FID data to posterior distributions in the GID model, which we cannot efficiently compute. What we can say is the FID sampling

method was not completely thrown off the scent by the GID sequence pairs. This was to be expected since the posterior distributions in the GID model are probably not very different from those in the FID model, from which we sampled.

5 Implementation

The software that was used for simultaneous sampling of alignments and mutation parameters for a pair of DNA sequences is freely available from
`www.math.uni-frankfurt.de/~stoch/software/mcmcsalut/index.html`
under the terms of the GNU Public License (GPL, 2000). The program is written in C++. Optional user interfaces for comfortable parameter input and graphical representation of results are implemented in Tcl/Tk. The program allows a large variety of base substitution models to be combined with the FID model, including, for instance, the option of resampling of the ratio of transversion and transition rates.

6 Discussion

The FID model makes possible efficient algorithms for statistical alignment and mutation rate estimation for a pair of DNA or protein sequences. The results seem more realistic than those given in Metzler et al. (2001), which are based on the TKF1 model. The fixed fragmentation structure in the FID model (as in the TKF2 model, cf. Thorne et al. 1992) might be unrealistic. If we drop this assumption, the FID model changes into the GID model. Since this model lacks the pair-HMM structure, we cannot expect to find efficient exact algorithms. However, our computer simulation studies encourage us to hope that mutation-parameter estimation and sampling procedures that were optimized for the FID model also give good results if the data stem from the GID model. A possible explanation is that differences between the models only matter when a newly inserted fragment is hit by another insertion or deletion. This probably occurs rarely, except when the number of mutations between the sequences is very high. In the latter case, however, any estimate is quite rough.

When the probable alignments of the given sequences contain only a few gaps, the fragment length parameter γ is difficult to estimate. In section 3.1 the posterior distribution of γ depends on the prior. In cases like this it might be reasonable to estimate γ (or at least an informative prior for γ) from similar sequence data, if available, since the parameter γ probably does not depend on the evolutionary distance and is thought to be constant e. g. among the HVR-1 sequences.

Often more than two sequences are to be aligned. If the aim is the estimation of a phylogeny one is in the same dilemma as in the case of mutation parameter estimation: one could estimate the phylogeny if the sequences were aligned and for the alignment of the sequences it would be helpful to know the phylogeny. Here too, a way out might be to estimate multiple alignments and phylogenies simultaneously. For recent

work in this direction see Mitchison (1999), Holmes and Bruno (2001), and Fleißner et al. (2002). These methods too can probably be brought a step closer to reality with the help of the FID model.

Acknowledgments

I would like to thank Anton Wakolbinger, Brooks Ferebee, Roland Fleißner and Jeff Thorne for stimulating discussions, the referees for helpful suggestions, and the German Science Foundation DFG for financial support.

References

Anderson,S., Bankier,A.T., Barrell,B.G. de Bruijn,M.H.L., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., Schreier,P.H., Smith,A.J.H., Staden,R., Young,I.G. (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457-465.

Durbin,R., Eddy,S., Krogh,A., Mitchison,G. (1998) *Biological sequence analysis*. Cambridge University Press

Felsenstein,J., Churchill,G.A. (1996) A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution. *Mol. Biol. Evol.* **13.1**, 93-104.

Fleißner,R., Metzler,D., von Haeseler,A. (2000) Can one estimate distances from pairwise sequence alignments? In Bornberg-Bauer,E., Rost,U., Stoye,J., Vingron,M. (eds) *GCB 2000, Proceedings of the German Conference on Bioinformatics, October 5-7, 2000, Heidelberg*. Logos Verlag, Berlin, pp. 89-95

Fleißner,R., Metzler,D., von Haeseler,A. (2002) Simultaneous Probabilistic Multiple Alignment and Phylogeny Reconstruction. (in prep.)

Gamerman,D. (1997) *Markov Chain Monte Carlo*. Chapman & Hall, London

Geiger,J. (1996) Size-biased and conditioned random splitting trees.
Stoch. Proc. Appl., **65**, 187-207.

GPL, (2000) The GNU Public Licence.
Available in full from `www.fsf.org/copyleft/gpl.html`

Handt,O., Meyer,S., von Haeseler,A. (1998) Compilation of human mtDNA control region sequences. *Nucleic Acids Research* **26**, 126-129.

Harris,T. (1963) *The Theory of Branching Processes*. Springer, Berlin

Hasegawa,M., Kishino,H., Yano,T. (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160-174.

Hein,J. (2001) An algorithm for statistical alignment of sequences related by a binary tree. In Altman,R.B., Dunker,A.K., Hunter,L., Lauderdale,K., and Klein,T.E. (eds) *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 179-190.

Hein,J., Wiuf,C., Knudsen,B., Møller,M.B., Wibling,G. (2000) Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit. *J. Mol. Biol.*, **302**, 265-279.

Holmes,I., Bruno,J., (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17.9**, 803-820.

Metzler,D., Fleißner,R., Wakolbinger,A., von Haeseler,A. (2001) Assessing Variability by Joint Sampling of Alignments and Mutation Rates. *J. Mol. Evol.*, **53**, 660-669.

Mitchison,G.J. (1999) A probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol*, **49**, 11-22.

Needleman,S.B., Wunsch,C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443-453.

Press,W.H., Flannery,B.P., Teukolsky,S.A., Vetterling,W.T. (1988) *Numerical Recipes in C*. Cambridge University Press, Cambridge

Swofford,D.L., Olsen,G.J., Waddell,P.J., Hillis,D. (1996) Phylogenetic inference. in: *Molecular Systematics* (eds: Hillis,D.M., Moritz,C., Mable,B.K.) Sinauer & Associates, Sunderland, MA.

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

Thorne,J.L., Kishino,H., Felsenstein,J. (1992) Inching towards reality: an improved likelihood model for sequence evolution. *J. Mol. Evol.*, **34**, 3-16.

Xu,X., Gullberg,A., Arnason,U. (1996) The Complete Mitochondrial DNA (mtDNA) of the Donkey and mtDNA Comparisons Among Four Closely Related Mammalian Species-Pairs. *J. Mol. Evol.*, **43.5**, 431-437.

Zharkikh,A. (1994) Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.*,**39**, 315-329.

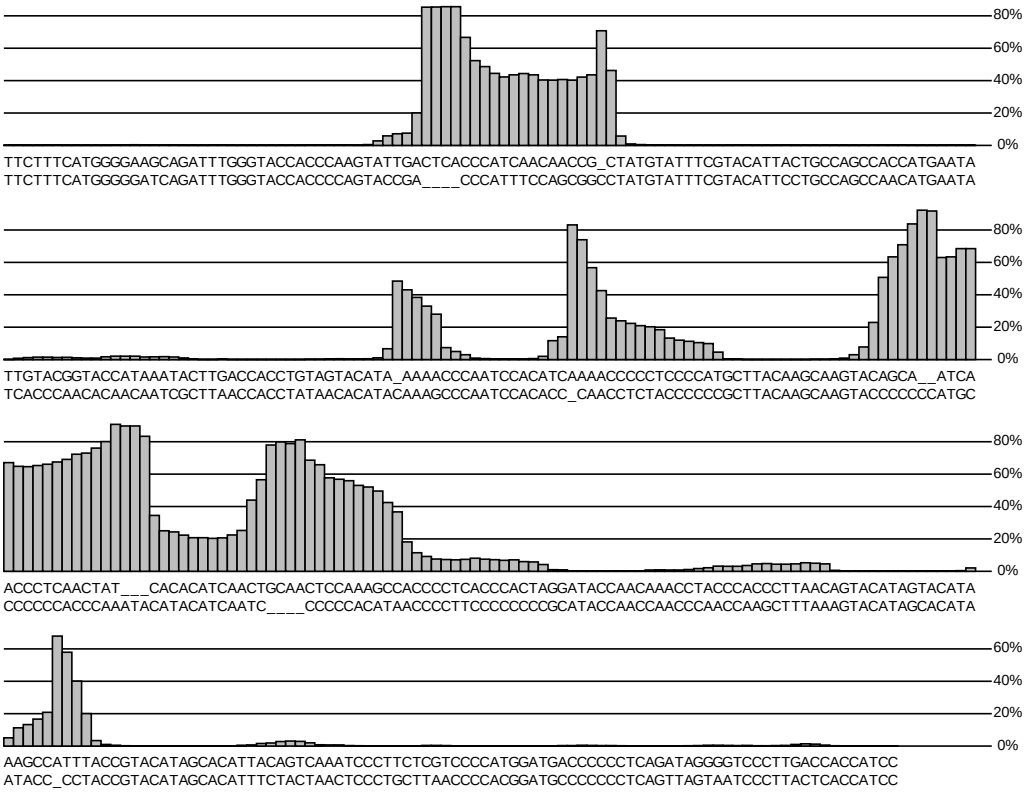

Figure 1: The most probable of 1000 sampled alignments (with almost flat prior) of a human (top sequence) and an orangutan HVR-1 sequence (bottom sequence) and the percentage of sampled alignments that differ from it in each position.

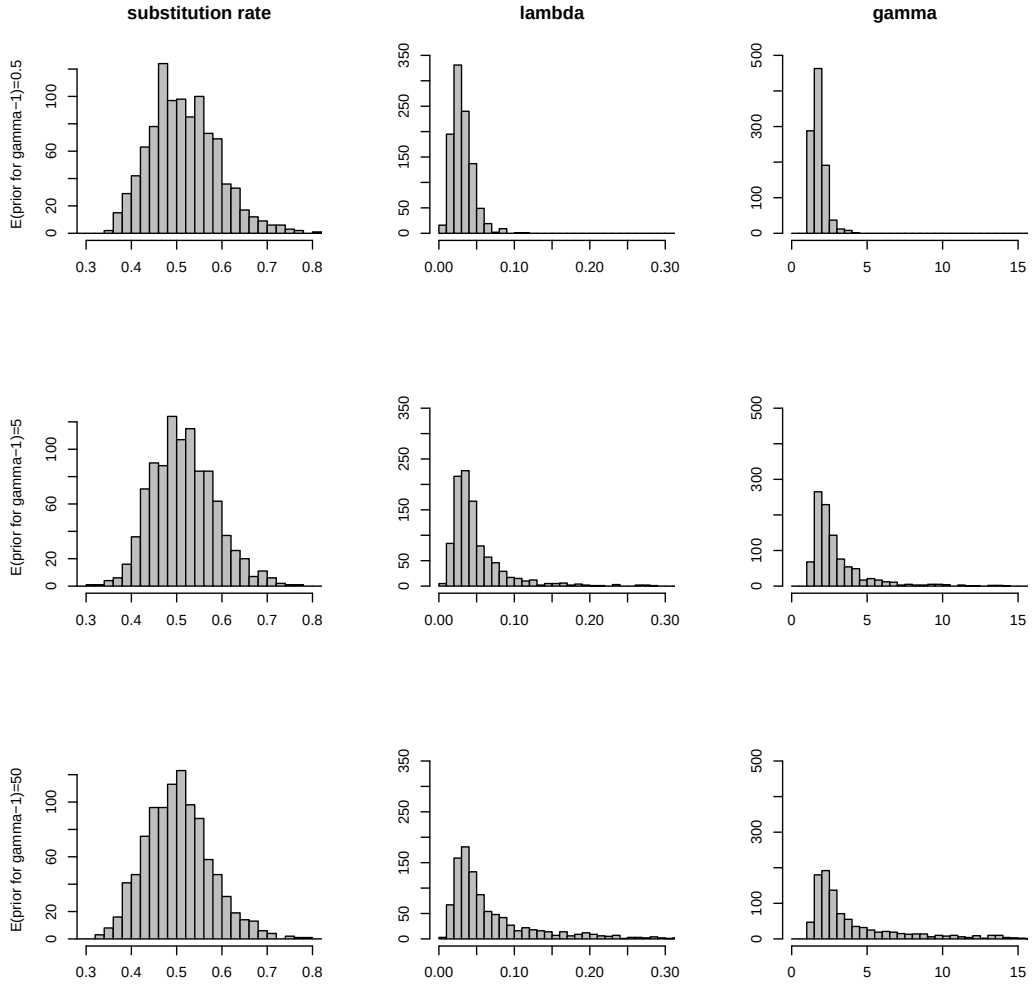

Figure 2: Each row shows the distributions of 1000 substitution rates (general substitution rate + transition rate), indel rate λ, and mean fragment length γ between HVR-1 of human and orangutan, sampled together with possible alignments of sequences according to their joint posterior distribution. For the exponential prior on $\gamma - 1$ we set the expectation to 0.5 (top row), 5 (middle), and 50 (bottom row). (In the the bottom row we cut the tails of the distributions of the sampled values for λ and γ in order to have the same scales as in the first two lines. In fact, 2.3% of the values for λ were greater than 0.3 and 2.9% of the values for γ were greater than 15 when the expectation of the prior for $\gamma - 1$ was set to 50.)

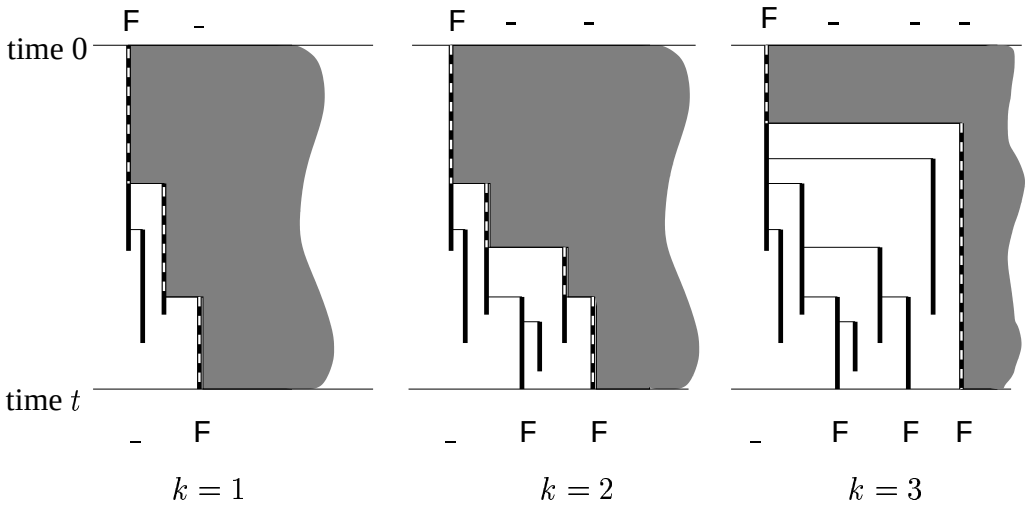

Figure 3: The process of insertions and deletions to the right of a fragment in the ancestral sequence (time 0) is a Galton Watson process ("F" stands for "fragment" and "_" for a series of gaps). Given that the ancestry of the $k = 1, 2, 3, \ldots$ leftmost survivors of a Galton Watson process at time t, the rest of the tree branches freely out of a branch from 0 to t (dashed line). This implies: Given that the number of survivors is not 0, it is geometrically distributed.

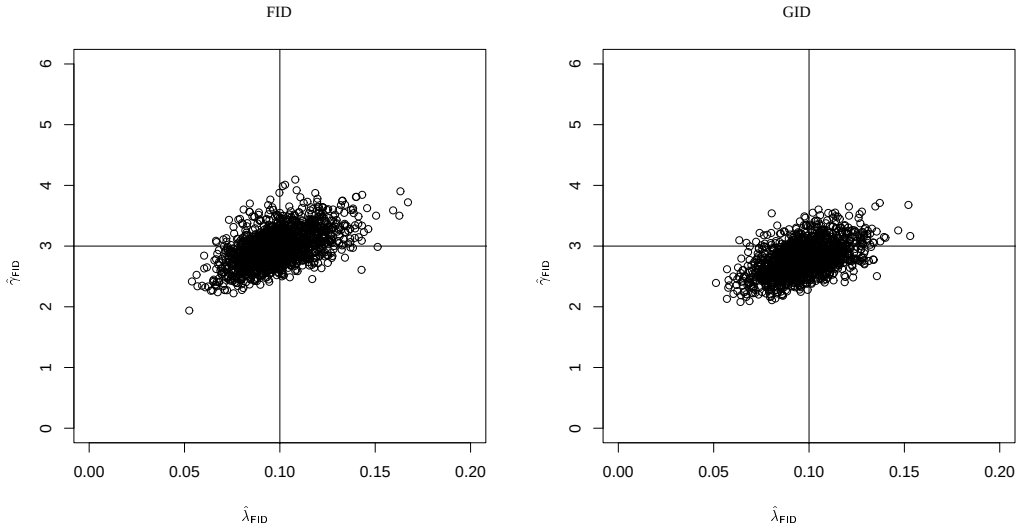

Figure 4: 1500 homology patterns with ancestral sequence length ~ 1000 were generated according to the FID (left) and the GID (right) model with $\lambda = 0.1$ and $\gamma = 3$. The point clouds show the ML-estimators in the FID model for (λ, γ) based on the homology patterns.

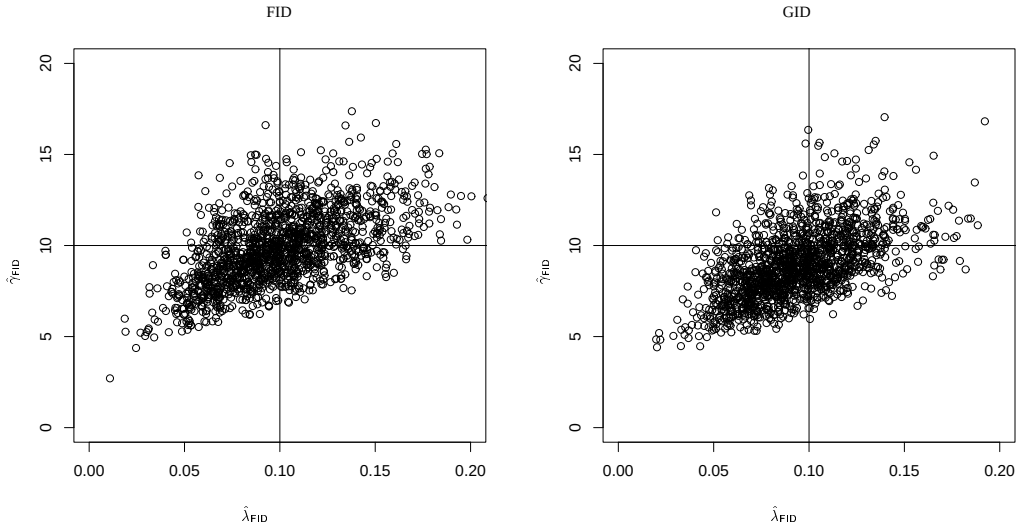

Figure 5: 1500 homology patterns with ancestral sequence length ~ 1000 were generated according to the FID (left) and the GID (right) model with $\lambda = 0.1$ and $\gamma = 10$. The point clouds show the ML-estimators in the FID model for (λ, γ) based on the homology patterns.

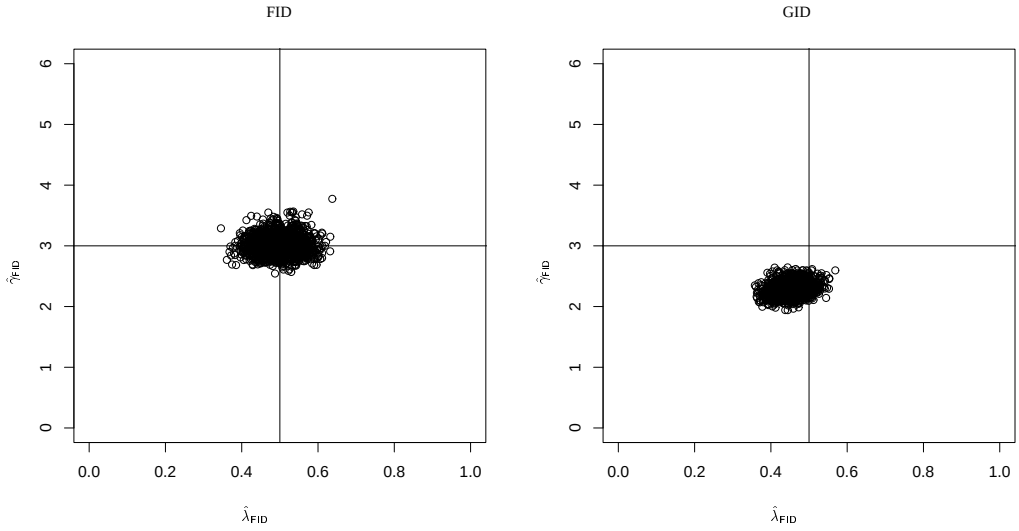

Figure 6: 1500 homology patterns with ancestral sequence length ≈ 1000 were generated according to the FID (left) and the GID (right) model with $\lambda = 0.5$ and $\gamma = 3$. The point clouds show the ML-estimators in the FID model for (λ, γ) based on the homology patterns.

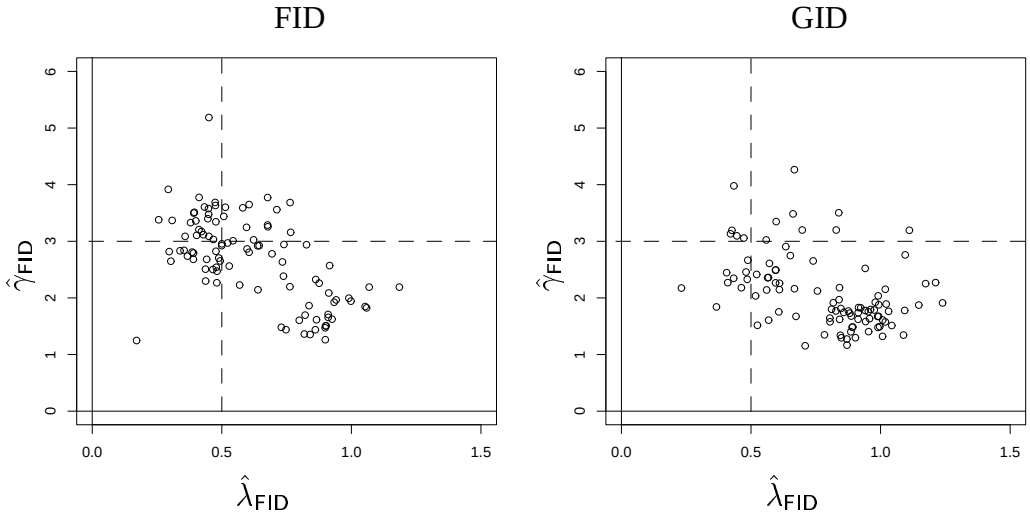

Figure 7: 100 sequence pairs of length ≈ 1000 were generated according to the FID (left) and the GID (right) model with substitution rate $s = 0.5$, $\lambda = 0.5$ and $\gamma = 3$. The point clouds show the ML-estimators in the FID model for (λ, γ) based on the sequence pairs.