

Assessing variability by joint sampling of alignments and mutation rates

Dirk Metzler*
Roland Fleißner**
Anton Wakolbinger*
Arndt von Haeseler**

corresponding author:

Dirk Metzler
telephone: +49 69 798 23512,
fax: +49 69 798 28444
e-mail: dmetzler@math.uni-frankfurt.de

*Johann Wolfgang Goethe-Universität, Fachbereich Mathematik, D-60054 Frankfurt am Main, Germany

**Max-Planck-Institut für evolutionäre Anthropologie, Inselstraße 22, D-04103 Leipzig, Germany

Abstract

When two sequences are aligned with a single set of alignment parameters, or when mutation parameters are estimated on the basis of a single “optimal” sequence alignment, the variability of both the alignment and the estimated parameters can be seriously underestimated. To obtain a more realistic impression of the actual uncertainty, we propose sampling sequence alignments and mutation parameters simultaneously from their joint posterior distribution given the two original sequences. We illustrate our method with human and orangutan sequences from the hyper variable region I and with gene-pseudogene pairs.

Keywords: sequence alignment, Markov chain Monte Carlo method, Thorne-Kishino-Felsenstein model, mutation parameter estimation, statistical alignment, hyper variable region, pseudogenes, Hidden Markov model.

Introduction

When two sequences are aligned using a score optimization like the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), the alignment obtained depends on the parameters used, and the optimal choices of these parameters depend on the unknown mutation rates. Similarly, when mutation rates are estimated from an alignment, the uncertainty in the alignment adds uncertainty to the parameter estimates. Treating either of these problems in isolation can thus result in an underestimate of the variability involved.

One way to break this vicious circle and achieve a more accurate assessment of the true uncertainty is to sample simultaneously alignments and mutation rates from an assumed joint posterior distribution. For this we need a model of sequence evolution, like that of Thorne, Kishino & Felsenstein (1991), which provides us with such a joint posterior distribution. In fact, the Markov chain Monte Carlo (MCMC) method (cf. Gamerman 1997) which we will employ for the joint sampling of alignments and mutation parameters does not use any specific properties of the Thorne-Kishino-Felsenstein (TKF) model. The only essential points are that the likelihoods of alignments as well as those of sequence pairs can be computed efficiently, and that the sampling of the alignments can be done in a tractable way. Both is guaranteed if the (unobserved) random alignments together with the (observed) sequences carry a Hidden Markov Structure when read from left to right – and indeed this is the case for the TKF model.

We think of a stochastic evolution dynamics depending on parameters like substitution, insertion and deletion rates which we collect into a parameter vector θ . This evolution dynamics takes an initial sequence into a final one by deleting some of the nucleotides, inserting others and changing some of the nucleotides by point mutations. In this way it produces an alignment a , say

$$\begin{array}{l} \text{ACGT_C} \\ \text{CC_TAC} \end{array} \quad (*)$$

The alignment, however, cannot be observed. What *is* observed are the sequences $s_1 = \text{ACGTC}$ and $s_2 = \text{CCTAC}$. Also, the model parameters θ are in general unknown; we are going to assume a Bayesian framework where we put some prior probability π on the parameters. For given θ , we denote the probability of an alignment a by $p_\theta(a)$.

For given sequences s_1, s_2 we would like to answer questions as follows:

1. What is the posterior distribution of the model parameters?

2. Which parts of an alignment are relatively certain and which ones are more questionable?
3. What do typical alignments look like?
4. How do the typical alignments depend on θ ?

We propose to attack these questions through a *joint* sampling of the alignments a and the model parameters θ with weights proportional to $p_\theta(a)\pi(d\theta)$, where a runs through the alignments compatible with s_1, s_2 . Using the idea of Gibbs sampling, we achieve this by an MCMC method in which alternately the alignment is sampled given the model parameters, and vice versa.

A well known and most convenient way of displaying alignments is their graphical representation as a path through a rectangular grid. In this way, the alignment (*) is represented by the path highlighted in Figure 1.

The TKF model for sequence evolution

We use a model of sequence evolution that was first described and applied to sequence alignment problems in Thorne et al. (1991). We refer to it as the TKF model. This section gives a short informal overview of the TKF model in the context of sequence alignment. The model contains three parameters: the substitution rate s , the insertion rate λ , and the deletion rate μ . Each site is independently deleted with rate μ , and hit by a substitution with rate s . Insertions occur between any two sites or at the ends of the sequence at rate λ . When a substitution or an insertion occurs, the new base is drawn randomly from $\{A, G, C, T\}$ according to a probability distribution $(\pi_A, \pi_G, \pi_C, \pi_T)$.

Given a sequence of length n , there are n sites which are candidates for deletion and $n + 1$ positions where a nucleotide can be inserted. Therefore, the length of the sequence grows with net rate $\lambda(n + 1) - \mu n$. In order to avoid a drift towards longer and longer sequences it is assumed that $\mu > \lambda$, which then ensures the existence of a stationary distribution of the sequence evolution process. The length of a random sequence at equilibrium is geometrically distributed with expectation $\frac{\lambda}{\mu - \lambda}$ and the bases are independently distributed according to $(\pi_A, \pi_G, \pi_C, \pi_T)$. This time stationary process is *reversible*: the probability of starting with an ancestral sequence s_1 and arriving after a time t at an offspring sequence s_2 is unaltered when s_1 and s_2 are interchanged.

In general, certain nucleotides of the ancestral sequence will be conserved in the offspring sequence; other nucleotides will appear only in the ancestral or only in the offspring sequence. Assume for instance that the ancestral sequence is ACGTC; it may happen that A is substituted by a C, the G is deleted and an A is inserted between the T and the C. Then, with the ancestral sequence on top, this results in the alignment

$$\begin{array}{l} \text{ACGT_C} \\ \text{CC_TAC} \end{array} \quad (*)$$

Other sequence histories lead to ambiguities in the alignment which must be resolved by convention. For instance, starting with ACG, suppose C is deleted and then T is inserted between A and G. There are two possibilities for the resulting alignment:

$$\text{alignment 1: } \begin{array}{l} \text{AC_G} \\ \text{A_TG} \end{array} \quad \text{or alignment 2: } \begin{array}{l} \text{A_CG} \\ \text{AT_G} \end{array}$$

Thorne et al. (1991) make the convention that insertions happen to the right of a nucleotide rather than between nucleotides. Thus in the example, T must be the offspring nucleotide of A and alignment 2 is the appropriate one. This would also be the case if T had been inserted between the A and the C before the deletion of the C, whereas alignment 1 would correspond to a history in which T was inserted between the C and the G and the C was deleted after that. This makes alignment 2 more probable than alignment 1. If the bottom sequence would be the ancestral one, it would be the other way round: alignment 1 would be more probable than alignment 2. Thus the alignment convention destroys the time reversibility. But this affects only the order of gaps in the two sequences and not the site homology, which is the primary objective of alignment problems. Therefore, if two sequences are given which have descended from some common ancestor, we may assume that the second sequence has evolved from the first.

Now suppose that for given parameters s , λ , and μ , a sequence is drawn from the equilibrium distribution and evolves for a time span t according to the TKF-model. How can we then compute the probability, that, say, the alignment (*) is the true one? The first observation is that we can separate the process into two independent components: the insertion/deletion process and the process of substitutions. First we calculate the probability of the *bare alignment*

$$\begin{array}{l} \text{BBBB_B} \\ \text{BB_BBB,} \end{array} \quad (**)$$

ignoring the base types (B just stands for “Base”). Then we compute the conditional probability for the alignment (*) given the bare alignment (**), which is easy, because all $\overset{B}{B}$, $\overset{B}{-}$, and $\overset{-}{B}$ take their base types independently of each other. For the pair $\overset{A}{C}$ and time t , this is the product of the probability of A, the probability that at least one substitution occurs in a time span t , and the probability that the new base is C: $\pi_A \cdot (1 - e^{-st}) \cdot \pi_C$. For the pair $\overset{C}{C}$ we have two possibilities: no substitution or one or more substitutions with the last one being C: $\pi_C \cdot [e^{-st} + (1 - e^{-st}) \cdot \pi_C]$. The probabilities for $\overset{G}{-}$ (given $\overset{B}{-}$) or $\overset{-}{A}$ (given $\overset{-}{B}$) are π_G and π_A . In the sequel we rescale the time so that the time elapsed between the sequences is $t = 1$.

The computation of the probability of the bare alignment (**) is slightly more difficult, because the positions are not independent. But as a consequence of the TKF alignment convention, the bare alignment is generated by a Markov chain on the states Start , $\overset{B}{B}$, $\overset{B}{-}$, $\overset{-}{B}$ and End . Therefore, we can compute the probability of a bare alignment by stepping through it from left to right and iteratively multiplying the result with the transition probability from the preceding state to the current one. For given states $x, y \in \{\text{Start}, \overset{B}{B}, \overset{B}{-}, \overset{-}{B}, \text{End}\}$, we denote the probability for the next state being y , given that the current state is x , by $P_{\lambda, \mu}(x \rightarrow y)$. For the calculation of the transition probabilities (see table 1) we refer to Thorne et al. (1991).

Likelihood computation

The TKF model fits into the concept of a *pair hidden Markov model (pair-HMM)* as described in Durbin et al. (1998). The usual HMM is a Markov chain which is hidden from an observer in the sense that he can only observe a sequence of “emissions” which depend on the current states. In the *pair-HMM* setting the observer sees a pair of sequences instead of a single sequence. In our case, the Markov chain is the bare alignment (corresponding to the path in the graphical representation) and the emissions are the DNA sequences. Durbin et al. (1998) describe various pair-HMMs that can be used for sequence alignment. In general, these models are constructed rather heuristically and are not exactly compatible with any stochastic sequence evolution model, as it is the case for the pair-HMM that arises from the TKF model.

Many algorithms for HMMs can be adapted to pair-HMMs. One of them is the *forward algorithm* for the calculation of the likelihood of possible values for the transition probabilities of the Markov chain, given the emitted sequence. In our

case, this likelihood is the probability $p_{s,\lambda,\mu}(s_1, s_2)$ that the observed sequences s_1 and s_2 are emitted from a TKF pair-HMM with parameters s , λ , and μ (here we think of the time span t for the evolution of one sequence into the other to be scaled to unit time.)

For the computation of this likelihood it is necessary to sum over all possible alignments of the sequences:

$$p_{s,\lambda,\mu}(s_1, s_2) = \sum_{\text{bare alignment } w} p_{\lambda,\mu}(w) \cdot p_s(s_1, s_2 | w) \quad (1)$$

$p_{\lambda,\mu}(w)$ is the probability of the bare alignment w for given insertion and deletion rates λ and μ , and $p_s(s_1, s_2 | w)$ is the probability of the sequences, given the bare alignment w and the substitution rate s .

The forward algorithm uses a dynamic programming approach to calculate the above sum. The main idea is the following (for details see Thorne et al. (1991) and Durbin et al. (1998)): Let the evolution parameters s , λ , and μ and a pair of sequences s_1 and s_2 of lengths n and m be fixed. For $0 \leq i \leq n$ and $0 \leq j \leq m$ let $f(i, j, \underline{B})$ be the probability that the following three events occur in a run of a TKF pair-HMM with the given parameters :

- The first i bases of the emitted sequence 1 coincide with the first i bases of s_1 .
- The first j bases of the emitted sequence 2 coincide with the first j bases of s_2 .
- In the bare alignment given by the hidden state sequence the i -th site of the first sequence is homologous to the j -th site of the second sequence.

$f(i, j, \underline{B})$ is defined similarly, except that site i is not homologous to site j but is aligned with a gap between sites j and $j + 1$ of sequence 2. Note that because of the Markov property of the bare alignment, one can easily compute $f(i, j, \underline{B})$ if the values for $f(i - 1, j - 1, \underline{B})$, $f(i - 1, j, \underline{B})$, and $f(i, j - 1, \underline{B})$ are known. For example, if there is a C at site i in the first given sequence and a G at site j in the second sequence (for $i, j > 1$), it is easy to see that the following equations hold:

$$f(i, j, \underline{B}) = \left[f(i - 1, j, \underline{B}) \cdot P\left(\frac{B}{B} \rightarrow \underline{B}\right) + f(i - 1, j, \underline{B}) \cdot P\left(\underline{B} \rightarrow \underline{B}\right) + f(i - 1, j, \overline{B}) \cdot P\left(\overline{B} \rightarrow \underline{B}\right) \right] \cdot \pi_C \quad (2)$$

$$\begin{aligned}
f(i, j, \frac{B}{B}) &= \left[f(i-1, j-1, \frac{B}{B}) \cdot P(\frac{B}{B} \rightarrow \frac{B}{B}) \right. \\
&\quad + f(i-1, j-1, \frac{B}{-}) \cdot P(\frac{B}{-} \rightarrow \frac{B}{B}) \\
&\quad \left. + f(i-1, j-1, \frac{-}{B}) \cdot P(\frac{-}{B} \rightarrow \frac{B}{B}) \right] \cdot \pi_C \cdot (1 - e^{-st}) \cdot \pi_G
\end{aligned} \tag{3}$$

Therefore the $n \cdot m \cdot 3$ values of the function $f(., ., .)$ can be computed efficiently by calculating $f(i, j, .)$ iteratively while increasing i and j from $i = j = 0$ up to $(i, j) = (n, m)$. The desired likelihood is then:

$$\begin{aligned}
p_{s, \lambda, \mu}(s_1, s_2) &= f(n, m, \frac{B}{B}) \cdot P(\frac{B}{B} \rightarrow \text{End}) \\
&\quad + f(n, m, \frac{B}{-}) \cdot P(\frac{B}{-} \rightarrow \text{End}) \\
&\quad + f(n, m, \frac{-}{B}) \cdot P(\frac{-}{B} \rightarrow \text{End})
\end{aligned} \tag{4}$$

Thorne et al. (1991) suggest using classical optimization algorithms in combination with the forward algorithm to search for the triple (s, λ, μ) that maximizes $p_{s, \lambda, \mu}(s_1, s_2)$ for given s_1 and s_2 . Recently, Hein et al. (2000) have improved the efficiency of the algorithm. Note that the maximum likelihood estimator, which is obtained by this procedure, does not rely on a single alignment: it takes every possible alignment of the given sequences into account.

The sampling method

We are now going to describe our method of joint sampling of alignments and parameters in more detail. In the framework of the TKF model, we have $\theta = (s, \lambda, \mu)$. (Again, we take $t = 1$ for simplicity.) As a prior for θ we propose $\pi(d\theta) = e^{-s} ds e^{-\lambda} d\lambda e^{-\mu} d\mu$; this makes the probability that no substitution occurs at a given surviving site uniformly distributed on $[0, 1]$. We write $p(d\theta, a)$ for $p_\theta(a)\pi(d\theta)$. Our task is to sample (θ, a) according to $p(d\theta, a | s_1, s_2)$.

If θ is fixed and we want to sample an alignment according to $p_\theta(a | s_1, s_2)$, then we can apply a classical HMM backward sampling algorithm (cf. Durbin et al. (1998), pp. 89-99), using the function f defined above. With f we can easily compute the probability distribution for the last alignment state before End given the sequences and the mutation parameters. For $x \in \{\frac{B}{B}, \frac{B}{-}, \frac{-}{B}\}$ the probability that the last state is x , is $c \cdot f(n, m, x) \cdot P(x \rightarrow \text{End})$, where c is a normalizing constant such that the three values for the tree states add up to 1. Therefore we can easily draw the last state at random according to this probability distribution. Assume

for example we drew a $\underline{\text{B}}$ for the last state. In the graphical representation of the alignment, this would mean that the last edge is a vertical one from $(n-1, m)$ to (n, m) . For all states $y \in \{\overset{\text{B}}{\text{B}}, \overset{\text{B}}{_}, \bar{\text{B}}\}$, given that the last state is $\underline{\text{B}}$, the probability that the preceding state equals y is $c' \cdot f(n-1, m, y) \cdot P(y \rightarrow \underline{\text{B}})$, where c' is a normalizing constant again. We draw the preceding state according to this distribution and continue in the same manner until we end up at the first state, i. e. until we have drawn an edge in the alignment graph that starts at vertex $(0, 0)$.

On the other hand, if an alignment is given, we can use a Metropolis-Hastings approach (cf. Gamerman (1997)) for sampling the parameter θ according to $p(d\theta | a)$. Starting with some parameter vector θ_0 we generate a Markov chain on the parameter space in the following manner: If the current state in step i is $\theta_i = (s_i, \lambda_i, \mu_i)$, then generate independent random numbers \tilde{s} , $\tilde{\lambda}$, and $\tilde{\mu}$, exponentially distributed with means s_i , λ_i , and μ_i . Consider $\tilde{\theta} = (\tilde{s}, \tilde{\lambda}, \tilde{\mu})$ as a proposal for the next state $(s_{i+1}, \lambda_{i+1}, \mu_{i+1})$. Then make a random decision: Accept it with probability

$$\min \left\{ 1, \frac{p(d\tilde{\theta} | a) \cdot s_i \cdot \mu_i \cdot \lambda_i}{p(d\theta_i | a) \cdot \tilde{s} \cdot \tilde{\mu} \cdot \tilde{\lambda}} \cdot \exp \left(\frac{\tilde{s}}{s_i} - \frac{s_i}{\tilde{s}} + \frac{\tilde{\lambda}}{\lambda_i} - \frac{\lambda_i}{\tilde{\lambda}} + \frac{\tilde{\mu}}{\mu_i} - \frac{\mu_i}{\tilde{\mu}} \right) \right\} \quad (5)$$

or set $\theta_{i+1} = \theta_i$ if you do not accept¹. As can be checked in a straightforward way, the posterior distribution for (s, λ, μ) is a (reversible) equilibrium for the process described above. By irreducibility, the process converges to this equilibrium distribution and can thus be used (at least approximatively) for sampling mutation parameters for given alignments. Of course, if one samples from one realization of the process, the results are not independent. However, the dependencies become small if one chooses the intervals between two samplings sufficiently large. Especially before the first sampling one should allow enough steps for the so called “burn-in”, because the initial state could have been a bad guess in a region of low probability.

Now that we have an alignment sampling strategy for given mutation parameters and a mutation parameter sampling strategy for given alignments, we can combine them using the idea of Gibbs sampling (cf. Gamerman (1997)) and obtain a method for sampling mutation parameters and alignments simultaneously. We construct a Markov chain on the space of mutation parameters and alignments as follows: If (θ, a) is the current state, then sample a new alignment a' for the

¹In our program we use a slightly different proposal chain, where the quotient $\lambda/(\mu - \lambda)$ is exponentially distributed (with the old value as expectation) instead of μ .

parameter triple θ , and afterwards perform a Metropolis-Hastings run to obtain a new parameter triple θ' given the alignment a' .

The disadvantage of this method is that we always have to sample alignments for new mutation parameters. For doing so, we have to compute the $3 \cdot n \cdot m$ values for the function f with the new parameters each time, which is very time consuming in general. Therefore, instead of sampling the whole alignment in each step, we realign only a part of it, about 30 nucleotides – then we have to compute only about 2700 values for f in each step. We use a part of the saved runtime to increase the number of steps between the samplings to compensate the dependencies in the alignments that arise from this strategy. Note that the posterior probability distribution $p(d\theta, a | s_1, s_2)$ is still a reversible equilibrium for the Markov chain on $\{(\theta, a)\}$ which we obtain by the algorithm described above. Therefore we can use this process for simultaneous Markov chain Monte Carlo (MCMC) sampling of alignments and mutation parameters from their common posterior distribution.

As in most applications of MCMC methods, appropriate algorithm parameters as the number of steps in the sampling intervals, the duration of burn-in, and the length of alignment resampling ranges can only be found with experience gained from careful analysis of the sampling results.

An example: HVR-1 from human and orangutan

As an illustration of our approach, we consider the alignment of the human sequence ID 1244 (Anderson et al. (1981)) and the orangutan (*Pongo pygmaeus pygmaeus*) sequence ID 389 (Xu et al. (1996)) from an updated version (<http://db.eva.mpg.de/Hvrbase>) of the hvrbase (Handt et al. (1998)). We only used the fragment that is known for both sequences.

When speaking about mutation parameters, we think of a time scaling which makes the time distance between the two sequences one unit. Thus, our parameters s , λ and μ are the expected numbers of substitutions, insertions and deletions per site.

Alignment reliability

After an initial run of 1000 steps (“burn in”) we sampled 1000 alignments and corresponding mutation parameters, performing 100 steps between each two samplings. In each sampling step, a (randomly chosen) piece of a length of 30 bp was

resampled, as explained above. Then we counted for each position of the alignment given in the data base how many of the sampled alignments differ from the data base alignment.

Figure 2 shows that in some regions the alignment given in the data base differs from more than 80% of the sampled alignments (for instance, the region \mathcal{R} corresponding to the segment TCACCCATCAACAACCG in the middle of the top line of Figure 2). In this region, even the most probable alignment of the 1000 alignments sampled shows a quite similar pattern of non-coincidence with the other sampled alignments (cf. Figure 3).

In Figure 4 we compare the graphical representation of the data base alignment with the sampled alignments around region \mathcal{R} . We see that many different alignments are possible.

Estimation of mutation parameters

Figure 5 shows that the range of possible values for the mutation parameters becomes much larger if we take into account that the true alignment is actually unknown. The left side of Figure 5 shows 1000 values for (s, λ) sampled from the posterior distribution given the sequences and *given that the data-base alignment is the true one*. Under this assumption the values for s and λ with the greatest posterior probability found in the sample are 0.48 and 0.0085. The deletion rate μ typically differs little from the insertion rate λ , so we concentrate on (s, λ) . The right side of Figure 5 shows the (s, λ) -part of 1000 alignment/parameter pairs sampled from their posterior distribution with our method. The most probable of these was associated with similar values for s (0.45) and λ (0.01) as in the case where the data-base alignment was assumed to be the true one, but the credibility range is much larger at the right side.

It may also be of interest to compare Figure 5 with the maximum likelihood (ML) estimator as suggested by Thorne et al. (1991). Maximizing the likelihood function of (s, λ) based on the given sequences (taking every alignment into account) via Nelder and Mead's simplex method (cf. Press et al. 1988) we obtained $(\hat{s}_{ML}, \hat{\lambda}_{ML}) = (0.42, 0.019)$. Note that $(\hat{s}_{ML}, \hat{\lambda}_{ML})$ lies at the edge of the left cloud but well within the right cloud in Figure 5. As also suggested in Thorne et al. (1991) we approximated the covariance matrix of the ML estimator by the negative inverse of its Hessian and arrived at $\text{sd}(\hat{s}_{ML}) \approx 0.057$, $\text{sd}(\hat{\lambda}_{ML}) \approx 0.01$ and $\text{cor}(\hat{s}_{ML}, \hat{\lambda}_{ML}) \approx -0.47$. These values are similar to the standard deviations and correlation in the right cloud of Figure 5 ($\text{sd}(s_{\text{sample}}) = 0.061$, $\text{sd}(\lambda_{\text{sample}}) = 0.013$, $\text{cor}(s_{\text{sample}}, \lambda_{\text{sample}}) = -0.54$).

A second example: Pseudogene/gene pairs

In order to compare the substitution and insertion rates obtained with our method with estimates of substitution rates and gap frequencies based on usual sequence alignments we applied it to the pseudogene gene pairs analyzed by Gu and Li (1995). Pseudogenes seem to be appropriate for testing our approach as their evolution is not constrained by functional necessity. Since they stem from functional sequences they should not contain repeats of short sequence motifs which could distort the insertion-deletion-process locally. Their functional homologs on the other hand should be practically devoid of insertions and deletions. Although the main focus of Gu and Li (1995) is on the distribution of gap lengths, they also provide estimates of the number of differences per position. Lacking access to the alignments on which Gu and Li based their study, we retrieved the unaligned sequences and aligned the coding regions of the functional genes to the presumably homologous part of the corresponding pseudogene. For 20 sequence pairs the numbers and lengths of gaps given by Gu and Li were not compatible with the lengths of the available sequences; these were not analyzed further. The remaining 58 of the 78 pseudogenes treated are given in table 2. For each sequence pair the sampling procedure was started using $4/3$ times the number of differences per position from Gu and Li (1995) as estimator of the substitution rate. (The factor $4/3$ takes into account that a nucleotide can be replaced by a nucleotide of the same type in the TKF model.) As an initial simple estimator of the insertion rate we used the number of gaps divided by twice the alignment length: $\frac{g}{L_x + L_y + g}$, where L_x and L_y are the sequence lengths and g is the number of gaps in Gu and Li (1995). With these values a first alignment was sampled. Then, after 10,000 burn-in runs, every thousandth of 100,000 runs was sampled. The range of resampling of the alignment in each of the steps was 30 nucleotides.

Figure 6 compares the most probable sampled substitution rates with Gu and Li's values. In most cases the estimates given by Gu and Li lie well within the central 95% of the sampled values. This is even true for large substitution rates. Most of our most probable substitution rates are bigger than Gu and Li's estimates. This may indicate that the score-optimization of alignments tends to make the aligned sequences more similar than they actually are (cf. Fleißner et al. (2000)).

For the sequence pairs which are labeled A-F in table 2 the 97.5% quantiles of the sampled substitution rates are by far smaller than the estimates given by Gu and Li. The gap frequencies as well as the sizes of the gaps in the respective sampled alignments, however, were always close to the values given by Gu and Li (1995) (data not shown). Thus, any reasonable scoring system (including the

human eye) would prefer our sampled alignments to Gu and Li's alignments. This calls for an explanation like a typing error in their paper or the use of a different sequence for the functional gene.

Table 3 shows examples of sampled alignments. The small variability seen in the first two frames is typical. The chaotic behavior in the last frame is probably due to a single long gap. This is of course not provided for in the assumptions of the TKF model.

Discussion and Outlook

The method presented here makes it possible to assess the joint variability of alignment and parameter estimates. Using the program one can also study the interaction of the two by clicking on individual parameter estimates and comparing the corresponding alignment. As shown in Figure 2 and Table 3, the program also reveals which parts of the sequences are difficult to align.

Various minor refinements of the method are easily implemented. For instance, different rates for transitions and transversions can be accommodated, as indeed could individual rates for each pair of bases, though this could lead to parameter over-fitting problems.

Our model's gravest defect is unfortunately not so easily remedied. The TKF model considers only insertions and deletions of single nucleotides. This assumption runs counter to a growing body of practical experience and it can lead to implausible alignments, as the last example in table 3 shows (see also Saitou and Ueda (1994)). The challenge is to find biologically more plausible models for the insertion-deletion process which still preserve the hidden Markov structure essential for computational feasibility. Thorne, Kishino and Felsenstein (1992) suggest a generalization of the TKF model that allows insertions and deletions of longer fragments, but they are forced to require that inserted fragments can only be deleted as a whole. Another approach, we are investigating is the approximation of biologically reasonable models by hidden Markov models.

In many situations, the phylogeny is not known but is to be estimated from the multiple alignment. On the other hand, the multiple alignments are usually based on a phylogeny. Thus we are in the same dilemma as in the case of two sequences with the relation between alignments and mutation parameters. Some suggestions have been made to overcome these difficulties in the case of multiple alignments, see for example Thorne and Kishino (1992), Vingron and von Haeseler (1997) and the literature cited therein. Mitchison (1999) suggests an algorithm, for si-

multaneous MCMC sampling of multiple alignments and phylogenies. However, it is not clear if the HMM he uses is compatible with a plausible sequence evolution model. We hope to find a biologically plausible model of sequence evolution, which can also be used in this context.

Acknowledgments

We thank Brooks Ferebee, Sonja Meyer, Steffen Grossmann, Jeff Thorne, Naruya Saitou, and Hirohisa Kishino for stimulating discussions and a referee for very helpful suggestions. We acknowledge the financial support of the Deutsche Forschungsgemeinschaft and the Max-Planck-Gesellschaft. Part of this work was done while Dirk Metzler was working at the MPI für evolutionäre Anthropologie.

Our programs are freely available from:

<http://www.math.uni-frankfurt.de/~stoch/software/mcmcalgn>

References

Anderson S, Bankier AT, Barrell BG de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (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

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

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

Gu X, Li WH (1995) The size distribution of insertions and deletions in human and rodent pseudogenes suggest the logarithmic gap penalty for sequence alignment. *J. Mol. Evol* 40:464-473

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

- Hein J, Wiuf C, Knudsen B, Møller MB, Wibling G (2000) Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit. *J. Mol. Biol.* 302:265-279
- Mitchison GJ (1999) A probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol.* 49:11-22
- Needleman SB, Wunsch CD (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 WH, Flannery BP, Teukolsky SA, Vetterling WT (1988) *Numerical Recipes in C*. Cambridge University Press, Cambridge
- Saitou N, Ueda S (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Mol. Biol. Evol.* 11(3):504-512
- Thorne JL, Kishino H (1992) Freeing phylogenies from artifacts of alignment. *Mol. Biol. Evol.* 9:1148-1162
- Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* 33:114-124
- Thorne JL, Kishino H, Felsenstein J (1992) Inching towards reality: an improved likelihood model for sequence evolution. *Methods in Enzymology* 34:3-16
- Vingron M, von Haeseler A (1997) Towards integration of multiple alignment and phylogenetic tree construction. *J. Comp. Biol.* 4:23-34
- 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

	$y = \frac{B}{B}$	$y = \frac{B}{-}$	$y = \bar{B}$	$y = \text{End}$
$x = \text{Start}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} e^{-\mu}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} (1 - e^{-\mu})$	$\lambda\beta$	$(1 - \frac{\lambda}{\mu})(1 - \lambda\beta)$
$x = \frac{B}{B}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} e^{-\mu}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} (1 - e^{-\mu})$	$\lambda\beta$	$(1 - \frac{\lambda}{\mu})(1 - \lambda\beta)$
$x = \frac{B}{-}$	$\lambda\beta \frac{e^{-\mu}}{1 - e^{-\mu}}$	$\lambda\beta$	$\frac{1 - e^{-\mu} - \mu\beta}{1 - e^{-\mu}}$	$\frac{(\mu - \lambda)\beta}{1 - e^{-\mu}}$
$x = \bar{B}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} e^{-\mu}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} (1 - e^{-\mu})$	$\lambda\beta$	$(1 - \frac{\lambda}{\mu})(1 - \lambda\beta)$

Table 1: The transition probabilities $P_{\lambda, \mu}(x \rightarrow y)$ in the TKF model with $\beta := \frac{1 - e^{\lambda - \mu}}{\mu - \lambda e^{\lambda - \mu}}$

Pseudogene	Acc.	Pseudogene	Acc.	Pseudogene	Acc.	
1. β actin ψ 1	V00479	24. Cytochrome c ψ C	M22880	49. Casein kinase II- α ψ	X64692	
2. β actin ψ 2	V00481	25. Cytochrome c ψ E	M22886	50. Keratin 19 ψ	M33101	
3. β actin ψ	M55014	26. Cytochrome c ψ F	M22889	53. HSC-70 ψ	Y00481	
4. γ actin ψ	M55082	27. Cytochrome c ψ H	M22891	55. Ferredoxin ψ -A	M34787	
6. γ actin ψ 1	X04224	28. Cytochrome c ψ I	M22892	56. Ferredoxin ψ -B	M34789	
7. Aldolase reductase ψ	M84454	29. α enolase ψ	X15277	57. Ferritin H ψ	J04755	
8. Cyclophilin ψ 133	X52856	30. ARS ψ 3	K01846	58. Tubulin- β ψ 67	M38484	E
9. Cyclophilin ψ 167	X52858	31. ARS ψ 1	K01845	60. Tubulin- β ψ 14P	K00840	D
10. Cyclophilin ψ 18	X52855	32. Aldolase B ψ	M21191	62. Tubulin- β ψ 21P	K00841	B
11. Cyclophilin ψ 192	X52857	33. Na/K-ATPase- β ψ	M25159	63. Tubulin- β ψ 46P	J00317	C
12. Cyclophilin ψ 29	X52853	38. D2-type cyclin ψ	M91003	64. Tubulin- β ψ 7P	K00842	F
13. Cyclophilin ψ 39	X52852	40. LDH-A ψ	X02153	66. TPI ψ 5A	K03224	
14. Cyclophilin ψ 43	X52854	41. LDH-B ψ	M60601	67. TPI ψ 19A	K03225	
15. Cytochrome c ψ 1	D00266	42. Lipocortin 2 ψ A	M62895	68. TPI ψ 13C	K03223	
16. Cytochrome c ψ 2	D00267	43. Lipocortin 2 ψ B	M62896	70. Prothymosin- α ψ D	J04800	
17. Cytochrome c ψ 3	D00268	44. Lipocortin 2 ψ C	M62898	71. Prothymosin- α ψ F	J04801	A
18. Cytochrome c ψ A	M22878	45. Metallothionein I ψ	M13073	72. Prothymosin- α ψ G	J04802	
20. Cytochrome c ψ G	M22890	46. Metallothionein II ψ	M13074	78. Adenylate kinase 3 ψ	X60674	
21. Cytochrome c ψ J	M22900	47. PGK ψ X	K03201			
22. Cytochrome c ψ K	M22893	48. PGK ψ A	K03019			

Table 2: The analyzed pseudogenes. The marks refer to the labels in Figure 6. The numbering is the same as in Table 1 of Gu and Li (1995).

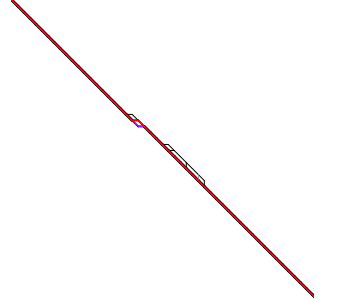
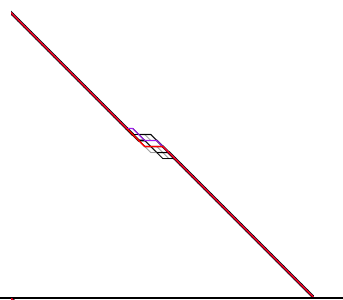
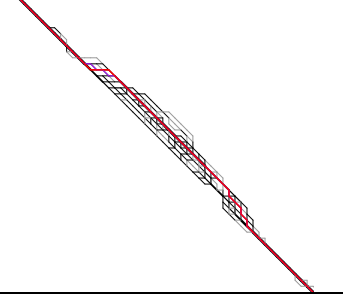
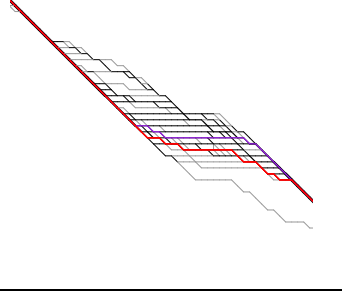
Paths	Most probable alignment
	<pre> CCAGTTGCGGAAGAAGAGGCA_CAGTCCAAAACAATAAGA CCAGTTGCGGAAGAAGAGGCAACAGTTCCAACAATAAGA TCACTGTAGT TCACTGTAGT </pre>
	<pre> CGCCGATAGGATGCAGAAG__ATCACCACCCTGGCGCCC TGCCGACAGGATGCAGAAGGAGATCACTGCCCTGGCACCC AGCACAAAT AGCACAAAT </pre>
	<pre> GAGAAAGGCA__AGATTTTGTTCAAAAGGTTGCCGCC GAGAAAGCAAGAAGATTTTATTATGAAGTGT_TC_C_C AGTGCCACACCATGGAAAAGGGAA AGTGCCACACCCTGAAAAGGGAG </pre>
	<pre> CTATATCCAGCAAGACACTAAGG__G_T_____GC_ TTATATCCAGCAAGACACTAAGGGCGACTACCAGAAAGCG _TG_T__ACC CTGCTGTACC </pre>

Table 3: Some typical samples of alignment paths together with the most probable of the sampled alignments. The paths and alignments shown are cutouts from the sampled alignments of LDH ψ , β actin $\psi 1$, cytochrome c ψG and lipocortin 2 ψB .

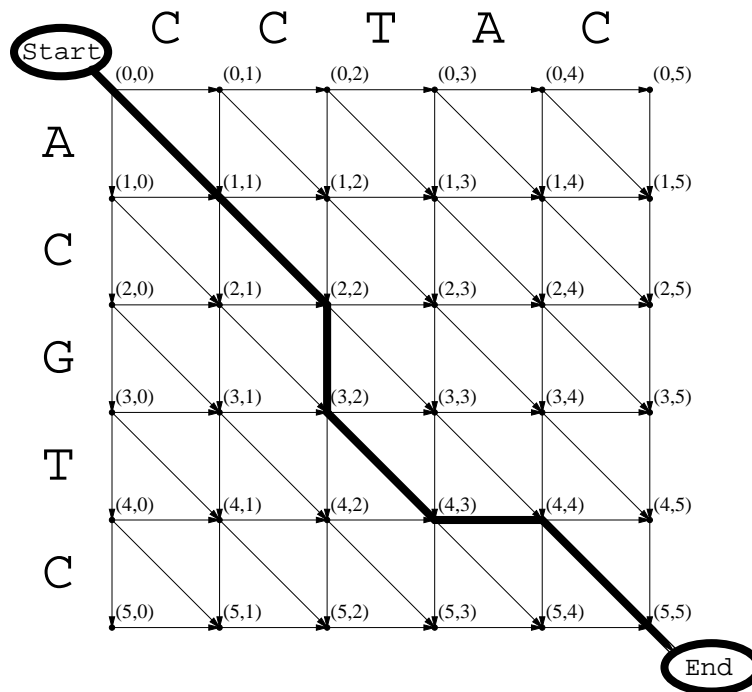


Figure 1: Graphical representation of the alignment `ACGT_C`
`CC_TAC`.

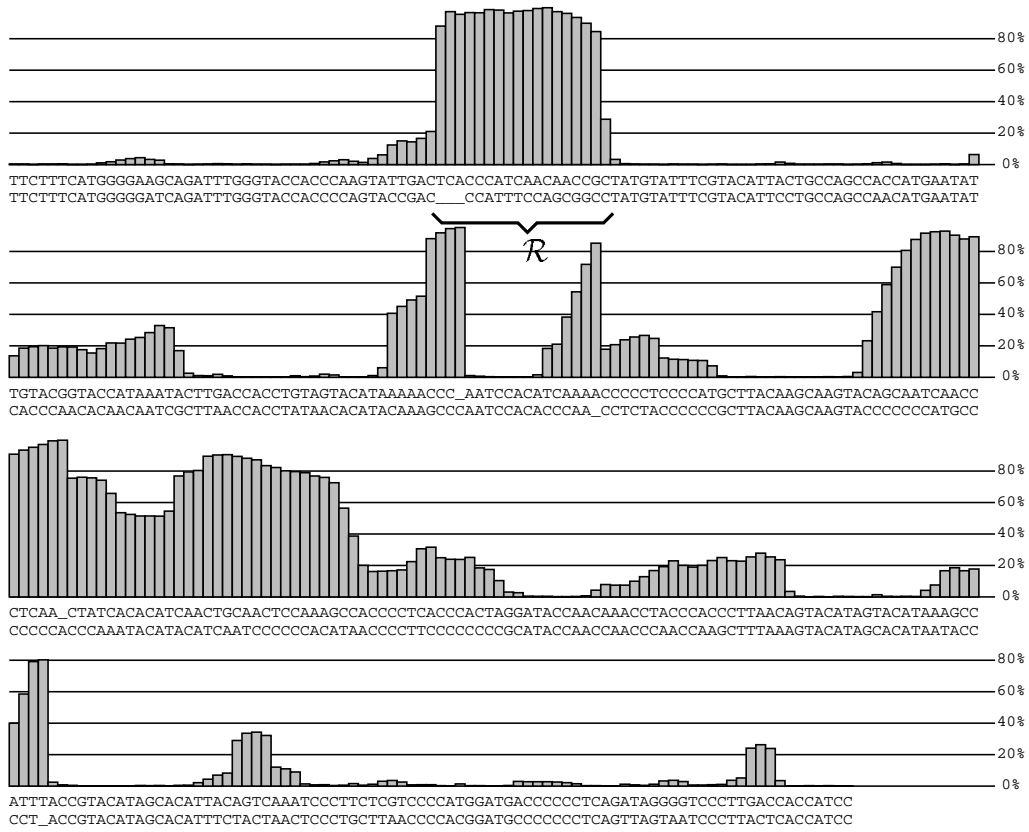


Figure 2: The data base alignment of a human (top sequence) and an orangutan HVR-1 sequence (bottom sequence) and the percentages of sampled alignments that differ from it in each position.

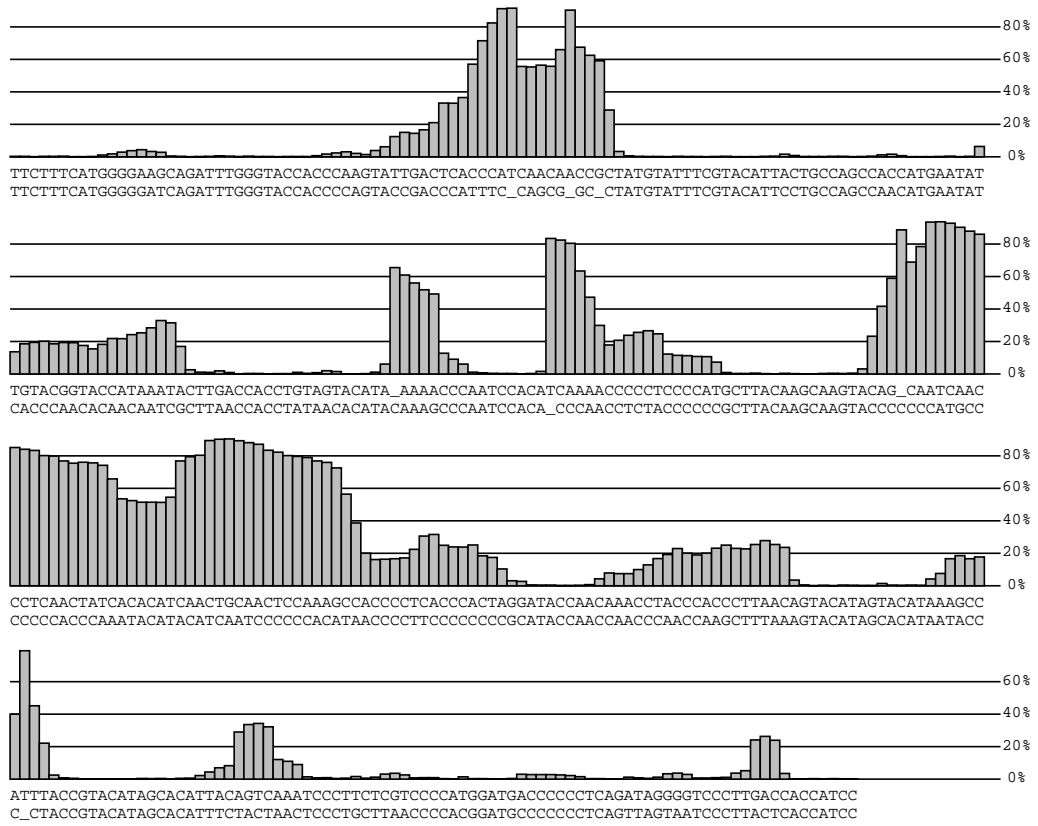


Figure 3: The most probable of the sampled alignments of a human and an orangutan HVR-1 sequence and the percentages of sampled alignments that differ from it in each position.

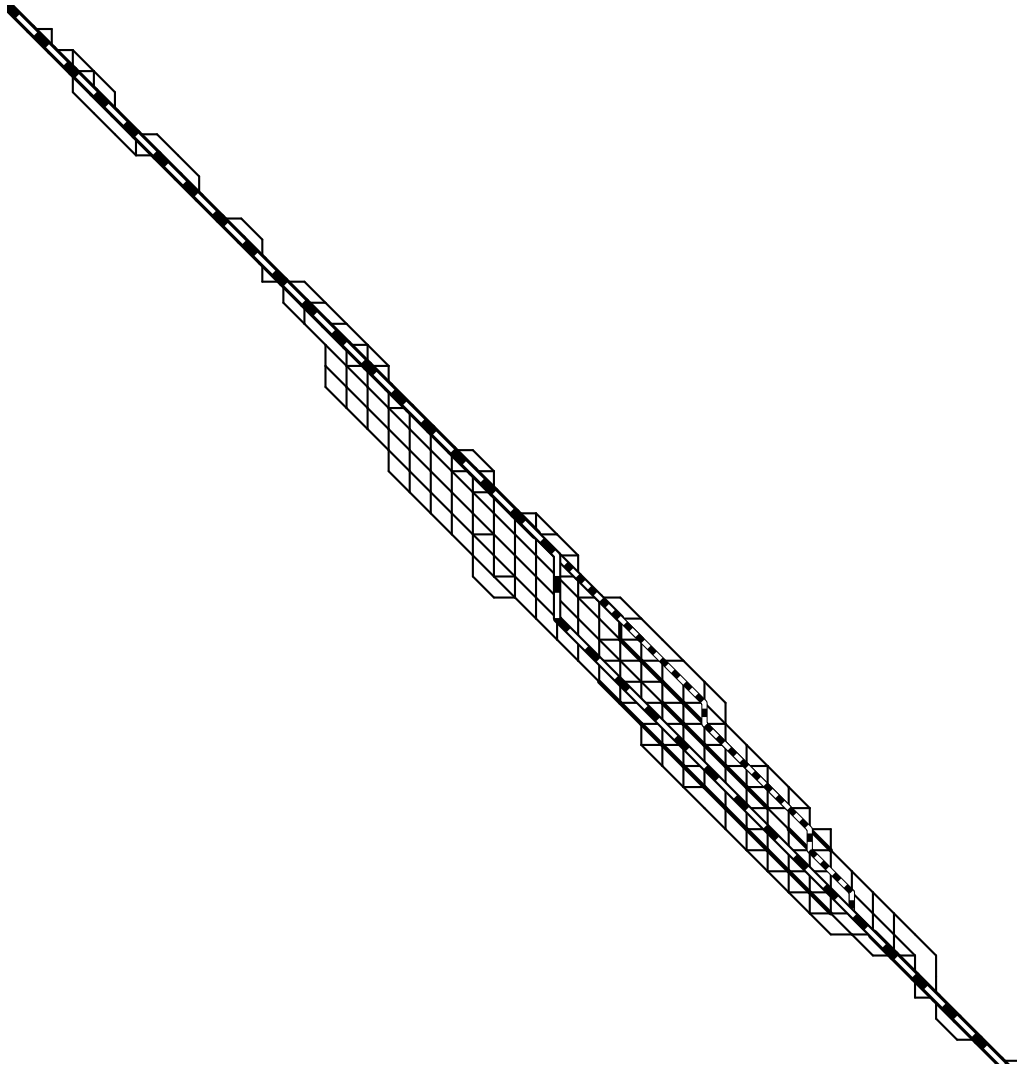


Figure 4: Alignment paths for the region \mathcal{R} (cf. Figure 2) between position 19 to 69 in the considered fragment of the human sequence and position 19 to 66 of the orangutan sequence: The alignment from the data base (dashed) and the most probable of the sampled alignments (dotted). The thickness of the black paths displays the frequency of the corresponding pairings in the sampled alignments.

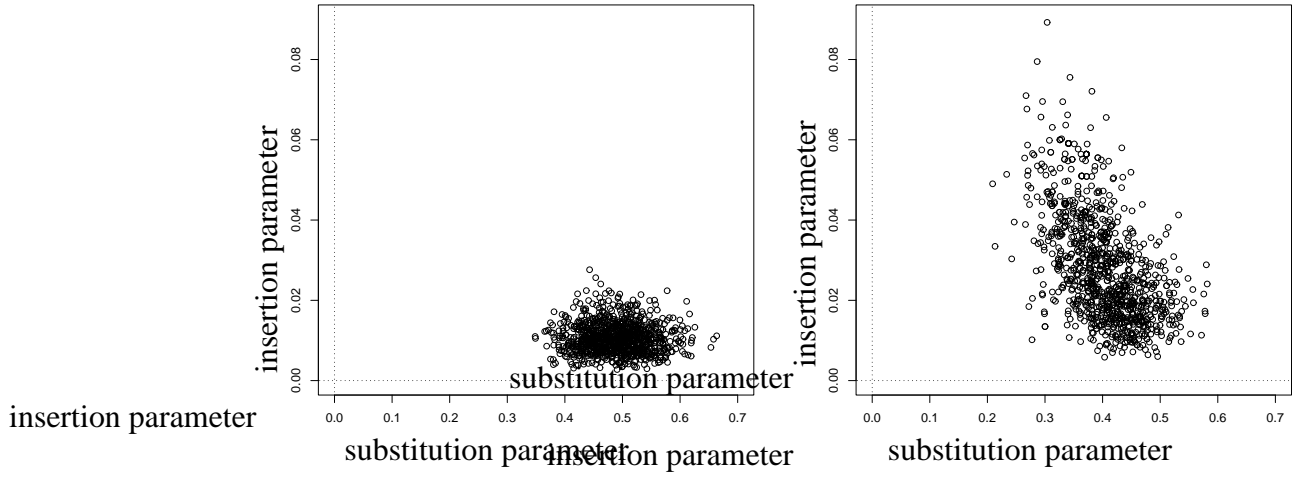


Figure 5: 1000 samples of substitution and insertion parameters between the HVR-1 sequences of human and orangutan according to their posterior probability given the data-base alignment (left) and the substitution and insertion parameters that were sampled together with the alignments according to the joint posterior probability of alignments and mutation parameters (right).

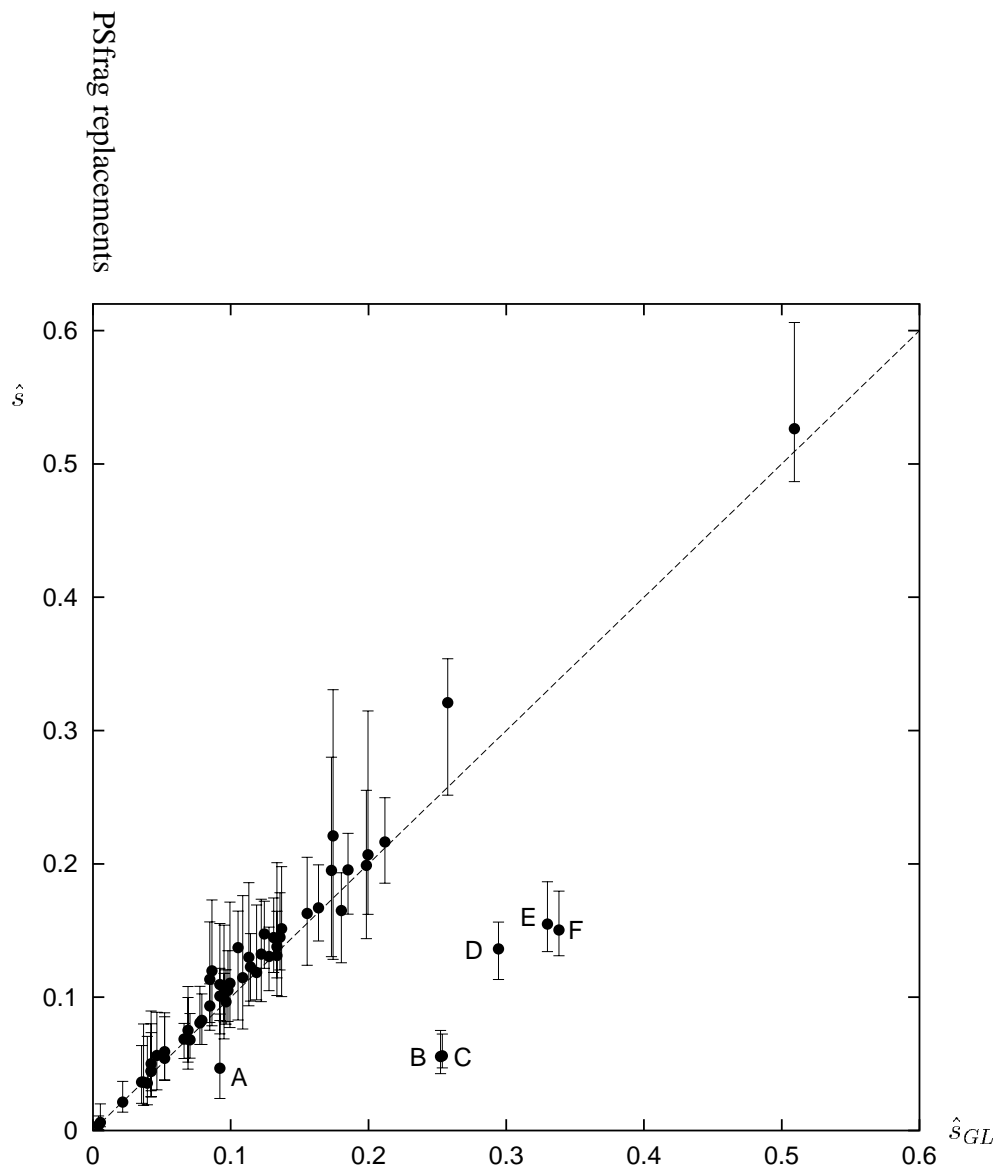


Figure 6: The most probable of the sampled substitution rates for each pseudogene gene pair (\hat{s}) plotted against the values estimated by Gu and Li (\hat{s}_{GL}). The latter was evaluated as $-\ln\left(1 - \frac{4}{3}p\right)$ where p is the number of differences per position given in Gu and Li (1995). The dotted line is the identity, the plotted intervals indicate the 2.5% and the 97.5% quantiles.