

Reconstruction of ancestral nucleotide sequences and estimation of substitution frequencies in a star phylogeny

Peter F. Arndt

Max Planck Institute for Molecular Genetics, Ihnestr. 63, 14195 Berlin, Germany

Received 13 June 2006; received in revised form 15 November 2006; accepted 15 November 2006

Available online 14 December 2006

Received by M. Batzer

Abstract

Maximum likelihood phylogeny reconstruction methods are widely used in uncovering and assessing the evolutionary history and relationships of natural systems. However, several simplifying assumptions commonly made in this analysis limit the explanatory power of the results obtained. We present an algorithm that performs the phylogenetic analysis without making the common assumptions for sequence data from at least three leaf nodes in a star phylogeny. In particular, the underlying nucleotide substitution model does not have to be reversible and may include neighborhood-dependent processes like the CpG methylation deamination process (CpG-effect). The base composition of the sequences at the external nodes and the one of the ancestral sequence may be different from each other and they do not have to be stationary state distributions of the corresponding substitution model. The algorithm is able to reconstruct the ancestral base composition and accurately estimate substitution frequencies in the branches of the star phylogeny. Extensive tests on simulated data validate the very favorable performance of the algorithm. As an application we present the analysis of aligned genomic sequences from human, mouse, and dog. Different substitution pattern can be observed in the three lineages.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Phylogeny; Ancient DNA reconstruction; Nucleotide substitution; CpG-effect

1. Introduction

As already noted by Zuckerkandl and Pauling about 40 years ago, living natural systems preserve inscribed into their genomes the largest amount of their own evolutionary history (Zuckerkandl and Pauling, 1965). However, it is still a challenging task to accurately reconstruct this evolutionary history from present-day sequence data. A very fruitful approach to this problem is the comparison of genomic sequences from different species. Nowadays this type of analysis can actually be performed on a genome-wide scale since recent genome sequencing projects provided us with a plethora of completely or nearly completely sequenced genomes over a wide range of species.

Parallel to the progress in sequencing, several important developments in pairwise and multiple sequence alignment, as well as phylogeny reconstruction have been achieved. One of the earliest and commonly used concepts in phylogeny reconstruction is parsimony. In this framework one tries to reconstruct ancestral sequences and explain observed sequence data with the least amount of substitutional changes necessary. While this concept is very simple and powerful for closely related sequences, it fails as divergence among sequences increases. Related problems have already been pointed out by other authors (Collins et al., 1994; Eyre-Walker, 1998; Perna and Kocher, 1995). One can avoid such problems using maximum likelihood phylogeny reconstruction, which was pioneered by Felsenstein (1981). In this framework one chooses a phylogeny that maximizes the likelihood of sequence data given a stochastic model of nucleotide substitutions. Several simple models for the substitutional processes have been discussed already, see Lio and Goldman (1998) for a review. More complicated substitution models take also rate variations

Abbreviations: MCML, Monte Carlo Maximum Likelihood; EM, Expectation Maximization; GC, guanine and cytosine.

E-mail address: arndt@molgen.mpg.de.

(Goldman and Yang, 1994; Nei and Gojobori, 1986; Uzzell and Corbin, 1971; Yang, 1994a,b) and neighbor dependencies (Arndt et al., 2003a,b; Lunter and Hein, 2004; Siepel and Haussler, 2004b; Whelan and Goldman, 2004) or the codon structure of exonic sequences (Pedersen and Jensen, 2001) into account and incorporate those into the framework of maximum likelihood phylogeny reconstruction (Siepel and Haussler, 2004a; Yang et al., 1994) and the interference of ancestral states (Krishnan et al., 2004). Besides these methods, Bayesian methods using Markov chain Monte Carlo (MCMC) methodology have been introduced (Huelsenbeck and Ronquist, 2001; Huelsenbeck et al., 2001; Bollback, 2002; Holder and Lewis, 2003; Hwang and Green, 2004) and are potentially very useful alternatives to maximum likelihood methods for the interference of ancestral states (Huelsenbeck and Bollback, 2001).

However most of these approaches to phylogeny reconstruction make at least one of the following assumptions: (i) the substitution model is time-reversible and the same in all branches of a given tree (only the branch length might vary from one branch to another, not all substitution processes are considered independently); (ii) the genomes under considerations are in the stationary state with respect to this model; and (iii) neighbor-dependent nucleotide substitutions can be neglected. These assumption are necessary to efficiently compute the likelihood for a given substitution model and tree topology (Felsenstein, 1981). However, all these simplifying assumptions are not necessarily granted. For vertebrates, they are actually violated to various degrees in the light of new sequence data and knowledge about the evolutionary processes. For instance, we know that the neighbor-dependent and irreversible CpG methylation deamination process (CpG → CpA/TpG) is the predominant nucleotide substitution process in vertebrates (Arndt et al., 2003a,b; Coulondre et al., 1978). Furthermore, the base composition is by far not constant and stationary for vertebrate species. The analysis of substitution pattern in repetitive elements, which can be regarded as genomic fossils, showed that there is strong AT-bias for substitutions in the human genome. The human genome, which today has a GC-content of about 42%, was more GC-rich in the past, and will lose more GC in the future until it reaches a stationary state at about 35% GC (Arndt et al., 2003a,b). Similar effects have also been found in other mammalian lineages (Duret et al., 2002).

We introduce a new method which does not rely on the above mentioned assumptions as long as one has sequence data available for at least three nodes in a star phylogeny, see Fig. 1 (b,c). This could be sequence data from 3 species (e.g. human,

mouse, and dog, which are believed to have diverged all from a common ancestor at the time of the mammalian radiation) or sequence data from copies of repetitive elements, which have deposited into a genome at about the same time. Specifically, we will show how to reconstruct ancestral sequences using a maximum likelihood approach which allows for different substitution models along the branches from the common ancestor to the leaf nodes. The sequences do not have to be in the stationary state and the underlying models may also be irreversible and include neighbor-dependent substitutions processes, like the CpG methylation deamination process. The incorporation of this process is especially important for the analysis of vertebrate sequences. The star phylogeny itself is already well established for a number of mammalian and other species. Using our approach we are able to reveal information about the differences in the evolutionary processes for different species, which is not possible using simpler existing approaches. Since we do not assume from the beginning that genomic sequences are at equilibrium, we are further able to show that this assumption is in fact not justified. The base distribution of genomic sequences is still evolving; in particular mammalian genomes are currently losing G and C nucleotides.

In the Method section, we will describe our new approach. First, we will introduce a neighbor-dependent nucleotide substitution model and show how the relevant transition probabilities can be computed. In a second part, we then state the likelihood function for given sequences and present a method to efficiently maximize this likelihood for substitutional processes with or without neighbor dependence. All relevant information to implement the presented algorithm is given. We also provide a public web server at <http://star.molgen.mpg.de>, where sequence data can be uploaded and analyzed as presented here. After presenting the method we describe extensive tests to validation of the method using synthetic sequence data. We further apply our method to genomic sequences from human, mouse, and dog and show first results on the ancestral base composition of the last common ancestor and on differences of the substitutional patterns and the strength of the CpG-effect in the 3 lineages.

2. Method

2.1. Nucleotide substitution model

In total there are 12 distinct neighbor-independent substitution processes of a single nucleotide by another. The rates of all these processes, $\alpha \rightarrow \beta$, will be denoted $r_{\alpha\beta}$, where Greek letters

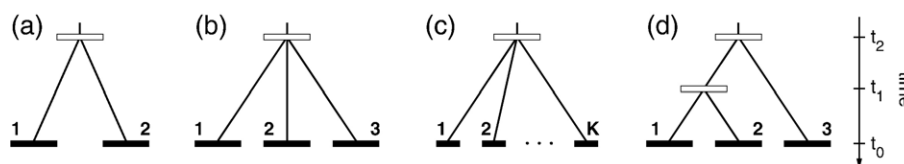


Fig. 1. Different phylogenies with two or more leaf nodes: the star phylogeny with 2 (a), 3 (b), and K leaves (c), and the tree topology with three leaves (d). Filled rectangles depict present-day sequences ($t=t_0$), open rectangles depict ancestral sequence data at the respective times of speciation. All phylogenies are rooted at the top.

represent nucleotides A, C, G, or T. These rates measure the number of substitutions per base pair (bp) and per time, in a sufficiently small time interval.

In addition to these 12 processes, we also want to consider neighbor-dependent processes of the kind $\kappa\lambda \rightarrow \kappa\sigma$ and $\kappa\lambda \rightarrow \sigma\lambda$ where either the right or left base of a dinucleotide changes, respectively. There might be several of those processes present in our model, their rates will be denoted by $r_{\kappa\lambda\kappa\sigma}$ or $r_{\kappa\lambda\sigma\lambda}$. The inclusion of such processes is motivated by the fact that the predominant substitution process in vertebrates is the substitution of cytosine in CpG resulting in TpG or CpA. Its rate is about 40 times higher than that of a transversion (Arndt et al., 2003a,b).

The model is parameterized by the substitution rates and the length of the time span, dt , the respective substitution processes act upon the sequence, which for our purposes and in the case of a star phylogeny as shown in Fig. 1(c) is $T = t_0 - t_2$. However, we have the freedom to rescale time and measure it in units of T . In this case, the time span is $dt=1$ and with this choice the substitution rates are equal to the substitution frequencies giving the number of nucleotide substitutions per bp. If we exclude neighbor-dependent processes, the model is parameterized by 12 substitution frequencies. For each additional neighbor-dependent process, we gain one additional parameter. The set of frequencies in the k th branch will be denoted by $\{r^k\}$. All frequencies are independent from each other.

In order to facilitate the subsequent maximum likelihood analysis we need to compute the transition probability, $P_{\{r\}}(\beta_1\beta_2\beta_3|\alpha_1\alpha_2\alpha_3)$, that a sequence of three bases $\alpha_1\alpha_2\alpha_3$ changes into the sequence $\beta_1\beta_2\beta_3$ for given substitution frequencies $\{r\}$. This probability can easily be calculated by numerically solving the time evolution of the probability to find three bases $p(\alpha\beta\gamma; t)$ at time t , which is given by a Master equation and can be written as the following set of 64 differential equations (Arndt and Hwa, 2005):

$$\begin{aligned} \frac{\partial}{\partial t} p(\alpha\beta\gamma; t) &= \sum_{\epsilon} [r_{\epsilon\alpha} p(\epsilon\beta\gamma; t) + r_{\epsilon\beta} p(\alpha\epsilon\gamma; t) + r_{\epsilon\gamma} p(\alpha\beta\epsilon; t)] \\ &+ \sum_{\epsilon, \epsilon'} r_{\epsilon\epsilon'\alpha\beta} p(\epsilon\epsilon'\gamma; t) + \sum_{\epsilon, \epsilon'} r_{\epsilon\epsilon'\beta\gamma} p(\alpha\epsilon\epsilon'; t), \end{aligned} \quad (1)$$

where ϵ and ϵ' are summed over all nucleotides. The rate parameters with the equal initial and final state, $r_{\alpha\alpha}$ and $r_{\alpha\beta\alpha\beta}$ are defined by

$$r_{\alpha\alpha} = - \sum_{\epsilon \neq \alpha} r_{\alpha\epsilon}, \quad r_{\alpha\beta\alpha\beta} = - \sum_{(\epsilon\epsilon') \neq (\alpha\beta)} r_{\alpha\beta\epsilon\epsilon'}, \quad (2)$$

where rates of neighbor-dependent substitution processes not included into the model are taken to be zero. To describe the evolution of three nucleotides $\alpha_1\alpha_2\alpha_3$, these differential equations have to be solved for initial conditions of the form

$$p(\alpha\beta\gamma; t=0) = \begin{cases} 1 & \text{if } (\alpha\beta\gamma) = (\alpha_1\alpha_2\alpha_3) \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

This can easily be done by numerical integration. The solutions will then yield the transition probabilities of three bases $\alpha_1\alpha_2\alpha_3$ being substituted by $\beta_1\beta_2\beta_3$:

$$P_{\{r\}}(\beta_1\beta_2\beta_3|\alpha_1\alpha_2\alpha_3) = p(\beta_1\beta_2\beta_3; t=1). \quad (4)$$

Please note that we can also state the time evolution of just one site (excluding any neighbor-dependent process), which is encoded by just 4 differential equations and yields us the transition probabilities $P_{\{r\}}(\beta|\alpha)$.

2.2. Maximum likelihood estimation

Let us consider K daughter sequences $\vec{\beta}^k = (\beta_1^k, \dots, \beta_N^k)$ of length N on the leaf nodes $k=1, \dots, K$ in Fig. 1(c). These sequences are aligned, i.e. orthologous sites have the same positional index $i=1, \dots, N$. If gaps are present in the alignment we exclude those sites from the further analysis.

The log likelihood that these K sequences are observed for a given substitution model is

$$\log L = \log \sum_{\vec{\alpha}} P(\vec{\alpha}) \prod_{k=1}^K P_{\{r^k\}}(\vec{\beta}^k|\vec{\alpha}), \quad (5)$$

where $P(\vec{\alpha})$ is the probability distribution of ancestral sequences and $P_{\{r^k\}}(\vec{\beta}^k|\vec{\alpha})$ are the transition probabilities from $\vec{\alpha}$ to $\vec{\beta}^k$ under given substitution models parameterized by the substitution frequencies $\{r^k\}$. Basically, the ancestral sequences are distributed following $P(\vec{\alpha})$ (which yet has to be specified through our analysis). For all ancestors $\vec{\alpha}$, the probability to observe the K daughter sequences is given by the product of probabilities that the ancestor $\vec{\alpha}$ evolved into daughter $\vec{\beta}^k$ under the substitution model $\{r^k\}$. An equivalent expression for the likelihood is already given by Felsenstein (1981). The likelihood in Eq. (5) has to be maximized to find the ancestral sequence and the frequencies for the substitution models along the K branches of the star phylogeny.

2.2.1. Neighbor-independent substitutions

Let us first discuss a special case. For independently and identically evolving nucleotides and excluding neighbor-dependent processes the above quantities factorize and the log likelihood is given by

$$\begin{aligned} \log L &= \sum_{i=1}^N \log \sum_{\alpha} P(\alpha) \prod_{k=1}^K P_{\{r^k\}}(\beta_i^k|\alpha) \\ &= \sum_{\gamma^1, \dots, \gamma^K} A(\gamma^1 \dots \gamma^K) \log \sum_{\alpha} P(\alpha) \prod_{k=1}^K P_{\{r^k\}}(\gamma^k|\alpha), \end{aligned} \quad (6)$$

where $P(\alpha)$ is the ancestral nucleotide distribution and the single-site transition probabilities are defined in the previous subsection. As shown in the second line, this likelihood depends on numbers $A(\gamma^1 \dots \gamma^K)$ that are defined as counts of 'site patterns' $\gamma^1 \dots \gamma^K$ at the K leaf nodes (there are in total 4^K such

numbers A). The likelihood has to be maximized over the $K \times 12$ free substitution frequencies and the ancestral nucleotide frequencies $P(\alpha)$ (3 additional free parameters). This maximization can easily be achieved using Powell's algorithm (Box, 1966; Press et al., 1992). This way we establish a map from the 4^K counts $A(\gamma^1 \dots \gamma^K)$ to the $12 \times K + 3$ free frequency parameters after maximization of the likelihood. At this point of the reasoning, it becomes apparent why the alignment of only two daughter sequences at the leaf nodes (cf. Fig. 1(a)) would not suffice to estimate 2 sets of substitution frequencies and the ancestral nucleotide distribution. In this case, there are only $4^2 = 16$ numbers $A(\gamma^1 \gamma^2)$ which is less than $2 \times 12 + 3 = 27$, the number of free parameters. Hence, we see that at least three species are needed to uniquely fix all substitution frequencies and the ancestral base distribution (Barry and Hartigan, 1987; Chang, 1996). Note that in some situations the number of substitutional parameters can be reduced by prior knowledge about the substitutional process, e.g. if the process is reverse-complement symmetric. However, this already introduces additional assumptions into the substitution model, which we wanted to avoid.

Once the maximum likelihood frequencies $\{r^k\}$ and the ancestral nucleotide distribution $P(\alpha)$ is found, the likelihood that a nucleotide α is the ancestral state for given site patterns $\gamma^1 \dots \gamma^K$ at the leaf nodes is $P(\alpha) \prod_{k=1}^K P_{\{r^k\}}(\gamma^k | \alpha)$. This can be subsequently used to reconstruct ancestral states at each position.

2.3. Monte Carlo maximum likelihood method

In the section above we maximized the likelihood by adjusting the frequency parameters $\{r^k\}$ and the ancestral nucleotide distribution. When the model does not include neighbor-dependent substitutions, this maximization is very fast. For models including neighbor-dependent processes, we propose a mixed Monte Carlo Maximum Likelihood (MCML) approach, which combines elements of the two methods in a very efficient way. In an iterative fashion we will first (a) estimate substitution frequencies for a given ancestral sequence (using a maximum likelihood approach) and then (b) get a new estimate for the ancestral sequence for given substitution frequencies (using a Monte Carlo approach). This iteration is initialized using the consensus of the K daughter sequences¹ as the ancestral sequence.

The procedure in step (a) to estimate substitution frequencies (including neighbor-dependent processes) from a set of ancestral and daughter sequences has already been described in the literature (Arndt et al., 2003a,b; Arndt and Hwa, 2005). This procedure takes an ancestral and aligned daughter sequence and finds the substitution frequencies for models as we have defined them in the last section. This is done using a maximum likelihood approach, which accounts for multiple and back substitutions at the same site, and estimates very accurately the substitution frequencies. This approach also admits to include

additional neighbor-dependent processes besides the CpG-effect. The statistical relevance of the inclusion of such additional processes can be assessed using a likelihood ratio test (Arndt and Hwa, 2005). Including only the relevant processes keeps the number of free parameters small and increases the fidelity of their estimates. The estimation of substitution frequencies is done independently in the various branches of the star phylogeny by comparison of the ancestor sequence with the daughter sequences yielding K sets of frequencies $\{r^k\}$.

In the second step (b) we generate a guess for an ancestral sequence. To do this we make use of a Monte Carlo procedure. Sequentially, we consider each site $i = 1, \dots, N$ and propose to substitute the nucleotide α_i by another nucleotide α'_i . The newly proposed nucleotide is accepted with the probability $p(\alpha_i \rightarrow \alpha'_i)$. To calculate this probability we observe that the likelihood in Eq. (5) can be approximated by

$$\log L \propto \sum_i \log L_i(\alpha_{i-1} \alpha_i \alpha_{i+1}, \{r\}), \quad (7)$$

where the local likelihood L_i is given by

$$L_i(\alpha_{i-1} \alpha_i \alpha_{i+1}, \{r\}) = P(\alpha_{i-1} \alpha_i \alpha_{i+1}) \prod_{k=1}^K P_{\{r^k\}}(\beta_{i-1}^k \beta_i^k \beta_{i+1}^k | \alpha_{i-1} \alpha_i \alpha_{i+1}). \quad (8)$$

The transition probabilities are defined in Eq. (4) with substitution frequencies taken from the estimates in step (a) and the distribution of 3-mers is simply estimated from the ancestral sequence before the Monte Carlo procedure. A substitution $\alpha_i \rightarrow \alpha'_i$ is always accepted if the likelihood increases, i.e. if the likelihood ratio

$$\lambda = L_i(\alpha_{i-1} \alpha'_i \alpha_{i+1}, \{r\}) / L_i(\alpha_{i-1} \alpha_i \alpha_{i+1}, \{r\}) \quad (9)$$

is larger than 1. If this ratio is smaller than one the substitution is accepted with probability λ :

$$p(\alpha_i \rightarrow \alpha'_i) = \begin{cases} 1 & \text{if } \lambda \geq 1 \\ \lambda & \text{if } \lambda < 1. \end{cases} \quad (10)$$

By the virtue of the Monte Carlo step, we allow that ancestral sites might not be in their most likely ancestral state. This is done by intention since such situations can actually be observed in a comparison of sufficiently long daughter sequences with their ancestral sequence $\vec{\alpha}$. For instance, there is always a finite probability that a nucleotide α is substituted by the same nucleotide β in all daughter sequences. The Monte Carlo step introduces such configurations into the ancestral sequence in as much as they are expected to occur with regard to the substitution model. This is crucial for the accurate estimation of substitution frequencies and ancestral single- and dinucleotide frequencies. Note that while the number of those sites that are not in their most likely state is given by the substitution models, their positions are not uniquely defined. Therefore, the ancestral sequence is one representative out of the set of sequences that maximize the likelihood.

This two-step iteration is performed several times until convergence of all the substitution frequencies and the 3-mer

¹ Initializing with a random sequence prolongs but not prevents the convergence of the algorithm to the maximum.

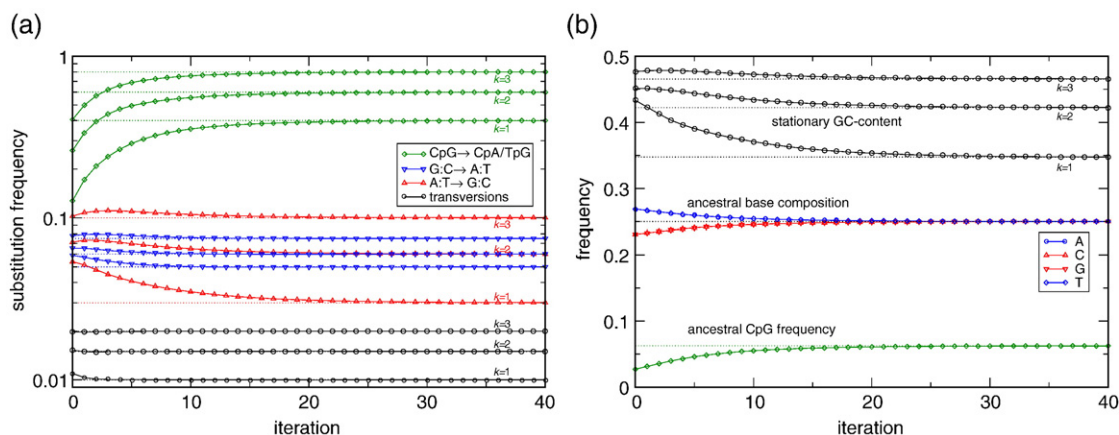


Fig. 2. Convergence of the estimates for various quantities to the underlying values (dotted lines, see also Table 1) used to generate the sequences. (a) Substitution frequencies in the $K=3$ branches (the y -axis is chosen logarithmically to present data on various magnitudes), (b) ancestral base composition, ancestral CpG frequency, and estimates for the stationary GC-content for the substitution pattern in the 3 branches.

distribution is established. Supplied with $K=3$ daughter sequences $\vec{\beta}^k$, the MCML algorithm converges after about 40 iterations corresponding to about 2 h CPU time (see Fig. 2). Both, the substitution frequencies in all branches, as well as the single and di-nucleotide distribution of the ancestral sequence

are very accurately recovered. These results are summarized in Table 1.

We provide a public web server at <http://star.molgen.mpg.de> where three aligned nucleotide sequences can be uploaded and analyzed using a substitution model specified by the user.

Table 1
Nucleotide and substitution frequencies for synthetic sequence data

	(a) General model	Reconstruction using:			
		(b) MCML	(c) Consensus	(d) No CpG-effect	(e) Real ancestor $\vec{\alpha}^a$
<i>Branch 1</i>					
r^1 (transversion) ^b	0.01000	0.01000±0.00008	0.01096±0.00007	0.01153±0.00007	0.01000±0.00007
r^1 (A:T→G:C)	0.03000	0.03002±0.00031	0.05386±0.00031	0.05441±0.00032	0.02998±0.00026
r^1 (G:C→A:T)	0.05000	0.05001±0.00047	0.05893±0.00038	0.07541±0.00037	0.04999±0.00035
r^1 (CpG→CpA/TpG)	0.40000	0.40039±0.00366	0.12725±0.00216	–	0.40012±0.00270
stat. GC-content	0.34742	0.34746±0.00258	0.43335±0.00143	0.43131±0.00160	0.34733±0.00168
<i>Branch 2</i>					
r^2 (transversion)	0.01500	0.01500±0.00009	0.01523±0.00008	0.01591±0.00009	0.01500±0.00008
r^2 (A:T→G:C)	0.06000	0.06003±0.00051	0.07118±0.00037	0.07082±0.00036	0.06003±0.00037
r^2 (G:C→A:T)	0.06000	0.05997±0.00056	0.06524±0.00045	0.09447±0.00041	0.05999±0.00045
r^2 (CpG→CpA/TpG)	0.60000	0.59964±0.00394	0.26060±0.00276	–	0.59951±0.00360
Stat. GC-content	0.42222	0.42238±0.00186	0.45150±0.00120	0.44003±0.00122	0.42232±0.00140
<i>Branch 3</i>					
r^3 (transversion)	0.02000	0.02000±0.00010	0.01984±0.00009	0.02062±0.00010	0.02000±0.00009
r^3 (A:T→G:C)	0.10000	0.10004±0.00059	0.10246±0.00048	0.10047±0.00046	0.10002±0.00046
r^3 (G:C→A:T)	0.07500	0.07488±0.00068	0.07797±0.00060	0.11916±0.00057	0.07491±0.00060
r^3 (CpG→CpA/TpG)	0.80000	0.80081±0.00681	0.40560±0.00454	–	0.80063±0.00594
Stat. GC-content	0.46561	0.46587±0.00149	0.47683±0.00119	0.46418±0.00123	0.46580±0.00127
<i>Ancestor</i>					
P (A)	0.250	0.250±0.000	0.269±0.000	0.269±0.000	–
P (C)	0.250	0.250±0.000	0.231±0.000	0.231±0.000	–
P (G)	0.250	0.250±0.000	0.231±0.000	0.231±0.000	–
P (T)	0.250	0.250±0.000	0.269±0.000	0.268±0.000	–
P (CpG)/ $(P$ (C) P (G))	1.000	1.000±0.005	0.510±0.003	–	–

Column (a): underlying frequencies which have been used to generate the ancestral and daughter sequences. Column (b): reconstruction of the frequencies in (a) using the MCML approach. Columns (c–e) report results on more simple approaches as discussed in the text. The sequence length is 1 Mbp and we give mean values and standard deviations from 100 independent runs.

^a No ancestral sequence is reconstructed.

^b For simplicity we report only the mean transversion frequency.

As output we provide the sets of substitution frequencies, the ancestral single- and di-nucleotide frequencies along with the ancestral sequence taken from the last iteration. A small example of a reconstructed ancestral sequence is shown in Fig. 3.

Note that the proposed algorithm actually falls into the class of stochastic Expectation Maximization (EM) algorithms, where (a) represents the M-step and (b) the E-step (McLachlan and Krishnan, 1997). While for a general EM algorithm one would require to take the expectation over all possible ancestral sequences (or a sample of those for a Monte Carlo EM algorithm), we rely here on only one representative ancestral sequence. This is possible since the average over all positions along the sequence offer an implicit equivalent of the expectation. If only little amounts of sequence data is available a sampling over different realization of ancestral sequences can easily be incorporated into the MCML approach.

2.4. Expected deviations of estimated frequencies

The errors introduced when reconstructing the ancestral sequence and estimating the substitution frequencies come from two sources. First, the assumed ancestral sequence can deviate from the real ancestor which will also have an effect on the substitution frequencies. Further the estimation of the substitution frequencies (for a given ancestral sequence) is not exact. The latter errors depend on the sequence length and are proportional to $1/\sqrt{N}$ if N is the sequence length (Arndt and Hwa, 2005).

In Fig. 4, we present the standard deviation of the substitution frequencies for different sequence length for the MCML method by closed symbols. The standard deviation again decreases proportional to $1/\sqrt{N}$. In the same figure, we also show (open symbols) the standard deviations for estimates of the substitution frequencies in case we would have knowledge of the real ancestral sequence $\vec{\alpha}$ and would not have to reconstruct it. In the latter analysis, one source of errors (due to the reconstruction of the ancestral sequence) is eliminated. However, both standard deviations are of the same order. This can also be observed comparing columns (b) and (e) in Table 1. In the same table, in column (e) we report estimates of the substitution frequencies taking the real ancestral sequence $\vec{\alpha}$ to estimate the substitution frequencies. The standard

```

ancestor  GCCAGCGGTTCCGGACGCCTCCACTGATGC
branch 1  .....T....T....T.T.C.....C.
branch 2  .....C...AA.T..T.....T
branch 3  A.....A....T....AT.....G.TT

```

Fig. 3. Reconstruction of the ancestral sequence. Shown is a small part of the three synthetically generated daughter sequences (see the text and Table 1 for parameters) together with the MCML reconstructed ancestral sequence, which coincides with the used real ancestral sequence. A dot denotes nucleotides identical to the ancestral one. Note that also the highly mutable CpG site is correctly reconstructed.

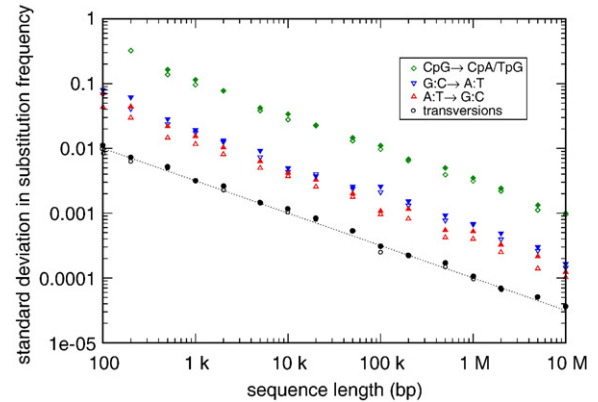


Fig. 4. The standard deviations of substitution frequency estimates (in the first branch, $k=1$) for various sequence lengths N and substitution frequencies as given in Table 1. Closed symbols represent data from 50 runs of the MCML algorithm, open symbols from equally many comparisons of the real ancestral sequences with their daughter sequences. The line represents the power law $0.1/\sqrt{N}$.

deviations in this type of analysis can be estimated by bootstrap (Arndt and Hwa, 2005). Consequently, good estimates for the standard deviations of the MCML analysis can also be generated by bootstrapping the frequency estimation using the ancestral sequence after the last iteration.

3. Results

3.1. Synthetic data

To test and validate the MCML algorithm we first applied it to synthetically generated sequence data. We first generated a 1-Mbp-long ancestral sequence $\vec{\alpha}$ with uniform base composition. This ancestral sequence was subsequently taken to generate $K=3$ daughter sequences by evolving them with respect to 3 previously fixed substitution models. These models included the neighbor-dependent CpG methylation deamination process as well as all other single-nucleotide substitutions. The relative substitution frequencies are chosen to mimic the substitution patterns as they might be observed in mammals after the mammalian radiation. The underlying frequencies are given Table 1, column (a). We have chosen a reverse-complement symmetric substitution models, which would describe the evolution of neutral DNA. While the four transversions occur with the same frequency in one branch, they occur, when compared to the first branch, 1.5 and 2 times more often in the second and third. Further, the neighbor-independent transitions introduce an AT-bias in branch 1 and a CG-bias in branch 3. On top of these processes we introduce the CpG-effect, which is taken to be predominant in all 3 branches. Note that the 3 substitution models are not related among each other through a simple rescaling of the frequencies. The three generated daughter sequences $\vec{\beta}^k$ show about 9%, 12%, and 16% divergence from the real ancestor $\vec{\alpha}$. The stationary GC-content for the three substitution patterns is 34.7%, 42.2%, and 46.6%, respectively. Neither the ancestral sequence (50% GC) nor the

three daughter sequences (45–46% GC) are in the stationary state with respect to the substitution models.

We then supplied only the 3 daughter sequences $\vec{\beta}^k$ to our MCML algorithm and iteratively generated ancestral sequences and estimated substitution frequencies. After initializing the ancestral sequence with the consensus sequence of the three daughter sequences (iteration 0), the algorithm converges after about 40 iterations. Both, the substitution frequencies in all three branches, as well as the single- and di-nucleotide distribution of the ancestral sequence are very accurately recovered. In Table 1 we summarize these results in the second column (b). We also list standard deviations of all measured quantities estimated from 100 independent tests involving newly created ancestral and daughter sequences. The deviations from the underlying frequencies and the standard deviations are always well below 1% for all substitution frequencies and base distributions in column (b). Subsequently, also the estimations for the stationary GC-content come out very good.

In Table 1, we also present the results of two other methods to estimate substitution frequencies and ancestral base compositions. The first one (in column (c)) takes the consensus sequence as the ancestor and compares this sequence to the three daughter sequences $\vec{\beta}^k$ to estimate substitution frequencies using the maximum likelihood approach from (Arndt and Hwa, 2005) to estimate the substitution frequencies. As can be seen the substitution frequencies show substantial deviations from the underlying frequencies (in column (a)). Up to 2-fold deviations are observed especially for the ancestral CpG frequency and the neighbor-dependent substitution frequencies. As a consequence, the values of the stationary GC-content for those substitution patterns deviate up to 10% (total).

In column (d) of Table 1, we present results which neglect neighbor dependencies, i.e. the CpG-effect, during the analysis of the sequences. We used the maximum likelihood approach with the likelihood given in Eq. (6). The results also deviate from the underlying frequencies, which have been used to generate the daughter sequences. However, we stress that those deviations are just a result of the exclusion of neighbor-dependent processes during the analysis. If such processes had not have been used to generate the three daughter sequences, the reconstructed frequencies would deviate again less than 1% from the underlying frequencies (data not shown). We still included this column to demonstrate that it is potentially very dangerous to neglect neighbor-dependent substitutions like the CpG-effect during the analysis. All measured quantities (substitution frequencies and ancestral base composition) as well as the computed stationary GC-content deviate substantially from the real values. If in doubt whether neighbor dependencies should be included, one may try to include different neighbor-dependent processes and assess their importance to be included into the analysis using a likelihood ratio test (Arndt and Hwa, 2005).

We extensively performed other tests allowing for different nucleotide compositions and non-trivial di-nucleotide distributions in the ancestral sequence or taking a piece of human genomic sequence as the ancestor. We further probed other sets of substitution frequencies in the 3 branches of the phylogeny.

We checked situations where the 3 sets of substitution frequencies were all the same and situations where the 3×14 frequencies were all different. All these tests clearly demonstrate the validity of our method and the results show basically the same behavior with respect to the deviations from the generating frequencies as presented in Table 1. We are ready to apply the MCML method to analyze genomic data.

3.2. Analysis of human–mouse–dog alignments

During the time of the mammalian radiation (about 90 Myr ago), several mammalian species diverged from one common ancestor (Hedges et al., 1996; Easteal, 1999). Today we have substantial amounts of sequence data for three species, which respect the star phylogeny: *Homo sapiens*, *Mus musculus*, and *Canis familiaris*. We analyzed multiple sequence alignments (multi alignments) for these three species from the UCSC website (genome.ucsc.edu). A total of about 950 Mbp have been aligned. In our analysis we include only those sequence segments that are at least 100 bp long and have less than 10% positions involving a gap in one of the three species. Further, we remove from the aligned sequences those segments that overlap with coding segments (exons) or repetitive sequences according to the annotation of human genome taken from www.ensembl.org. We are left with 37 Mbp of human–mouse–dog alignments in 125 000 sequence segments. The segments were concatenated and feed to the MCML algorithm. The algorithm converges after about 40 iterations and the resulting estimates for the ancestral base composition and the three sets of substitution frequencies are summarized in Table 2. We also report the GC-content and CpG odds ratio (i.e. the CpG dinucleotide frequency normalized for the base composition) of the three daughter sequences, as well as the average scaling factor of the neighbor-independent nucleotide substitution frequencies relative to the human lineage and the stationary GC-content computed from the three sets of frequencies.

The results show that the GC-content in the intergenic and non-repetitive sequences of all three species is different. The

Table 2
Estimated substitution frequencies and ancestral base composition from human–mouse–dog alignments

	Ancestor	Human	Mouse	Dog
GC-content	0.396	0.380	0.399	0.389
CpG odds ratio	0.349	0.201	0.235	0.232
$r(A:T \rightarrow C:G)$		0.01679	0.05340	0.02441
$r(A:T \rightarrow T:A)$		0.01298	0.04048	0.01622
$r(C:G \rightarrow G:C)$		0.01833	0.04674	0.02215
$r(C:G \rightarrow A:T)$		0.02054	0.05017	0.02212
$r(A:T \rightarrow G:C)$		0.04444	0.13827	0.06521
$r(C:G \rightarrow T:A)$		0.08469	0.17721	0.09817
$r(CpG \rightarrow CpA/TpG)$		0.64882	0.85770	0.63033
Average scaling factor ^a		1	2.7	1.3
Stationary GC		0.32229	0.39937	0.37086

The MCML approach was used to analyze about 37 Mbp of genome-wide intergenic, non-repetitive sequences. The uncertainties in the estimates are below 1% and omitted from the table.

^a For substitution frequencies, excluding $r(CpG \rightarrow CpA/TpG)$.

aligned sequences from human, mouse, and dog have 38%, 40%, and 39% GC, respectively. The ancestral GC content is estimated to be about 39.6%. Strikingly, the ancestral sequence has much more CpG dinucleotides, since the CpG odds ratio is estimated to be 0.34, while for the present-day sequences this ratio is about 0.22 for the three species. Many ancestral CpGs have been lost through the CpG-effect, which is predominant in all three branches. Compared to transversions the CpG methylation deamination frequency is about 40 times higher for human (Arndt et al., 2003a,b estimated the same value based on a study of repetitive elements), 20 times higher for mouse, and 30 times higher for dog. We can further read off that, when excluding the CpG-effect, the mouse lineage accumulated about 2.7 times more substitutions as the human lineage. The CpG methylation deamination frequency is only 1.3 times higher in mouse. Previous analysis on ancestral repeats in the human and mouse lineage revealed a factor of about two but did not distinguish different substitutional processes (Waterston et al., 2002). The dog lineage accumulated about 1.3 times more transition and transversion than the human lineage, while the CpG-effect is of the same magnitude in both lineages.

In the last row of Table 2, we computed the stationary GC-content for the substitution pattern in the three lineages. They are 32%, 40%, and 37% for human, mouse, and dog, respectively. Note that the GC-content of the mouse genome seems to be in equilibrium, i.e. the substitutions from A or T to C or G seem to be balanced with their reverse processes. This probably explains why at the much larger substitution frequency the mouse GC-content is still so high. But we want to caution the reader since we performed a genome-wide analysis of the three genomes. As for the human genome we expect large-scale regional fluctuations of the substitutional patterns along all chromosomes (Arndt et al., 2005). Therefore, a more careful analysis of different regions in the mouse genome has to be performed before coming to the conclusion that an equilibrium has been reached. Still, distinct differences in the substitutional patterns on the genome-wide scale can already be observed. In summary, much more information about different substitution frequencies along the human, mouse, and dog lineage and several properties of the ancestral sequence have been revealed using the MCML approach.

4. Discussion

We present an algorithm to maximize the likelihood to observe three or more nucleotide sequences in a star phylogeny. In our analysis we do not assume that the nucleotide substitution process is time reversible, stationary, and neighbor-independent, assumptions which are generally thought to be granted in phylogenetic analysis. Our approach is conceptually and numerically more intricate. However, without relying on the above assumptions we are actually able to show that they are in fact violated for mammalian species. In particular we can show that the GC-content human genome has decreased in time. This is possible since we reconstruct substitutional patterns as well as the ancestral sequence at the same time. We can further state that neither our genome today

nor that of our ancestors with all mammals is in equilibrium. In contrary, continuing the trend of the last 90 Myr the human genome will lose more G and C nucleotides until a new equilibrium is reached. The same is true for the dog genome whose substitution patterns can be, due to our approach, analyzed independently from the human ones.

Although designed for the star phylogeny, the algorithm still works and gives reliable results if the phylogeny considered is nearly star-like, i.e. if the difference in speciation times $t_1 - t_2$ is small compared to the time along the branches t_2 (in the notation of Fig. 1(d)). In the future the presented algorithm will be extended to general phylogenies. But already at this point, our analysis of three species in a star phylogeny reveals that the genomic isochores, i.e. large-scale fluctuations of the base composition along the chromosomes of humans and other vertebrates (Bernardi, 2000; Eyre-Walker and Hurst, 2001), are not static and stable in time. The three species analyzed here evolved differently after the mammalian radiation. A careful analysis of the particular regional differences of the substitution pattern which might have been evolved in the face of particular genomic or evolutionary constraints in the 3 species. This will be a very powerful tool and help us to come to a better understanding of the genomic isochore structure, its origins, and functional implications. Further, the reliable reconstruction of an ancestral genome from 90 Myr back in time will allow us to search for more ancestral repeats, which due to their continuous accumulation of mutations are indistinguishable from the genomic background in the present-day genomes.

Acknowledgment

PFA acknowledges fruitful discussions with Terence Hwa and Dmitri Petrov.

References

- Arndt, P.F., Hwa, T., 2005. Identification and measurement of neighbor dependent nucleotide substitution processes. *Bioinformatics* 21, 2322–2328.
- Arndt, P.F., Burge, C.B., Hwa, T., 2003a. DNA sequence evolution with neighbor-dependent mutation. *J Comput. Biol.* 10, 313–322.
- Arndt, P.F., Petrov, D.A., Hwa, T., 2003b. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* 20 (11), 1887–1896.
- Arndt, P.F., Hwa, T., Petrov, D.A., 2005. Substantial regional variations in the substitution rates in the human genome: importance of the GC-content, gene density and telomere-specific effects. *J. Mol. Evol.* 60, 748–763.
- Barry, D., Hartigan, J.A., 1987. Statistical analysis of hominoid molecular evolution. *Stat. Sci.* 2 (2), 191–210.
- Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241 (1), 3–17.
- Bollback, J.P., 2002. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* 19 (7), 1171–1180.
- Box, M.J., 1966. A comparison of several current optimization methods and use of transformations in constrained problems. *Comput. J.* 9 (1), 67–77.
- Chang, J.T., 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.* 137 (1), 51–73.
- Collins, T.M., Wimberger, P.H., Naylor, G.J.P., 1994. Compositional bias, character state bias and character state reconstruction using parsimony. *Syst. Biol.* 43 (4), 482–496.
- Coulondre, C., et al., 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* 274 (5673), 775–780.

- Duret, L., et al., 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* 162 (4), 1837–1847.
- Easteal, S., 1999. Molecular evidence for the early divergence of placental mammals. *Bioessays* 21 (12), 1052–1058 (discussion 1059).
- Eyre-Walker, A., 1998. Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* 47 (6), 686–690.
- Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. *Nat. Rev., Genet.* 2 (7), 549–555.
- Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17 (6), 368–376.
- Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11 (5), 725–736.
- Hedges, S.B., Parker, P.H., Sibley, C.G., Kumar, S., 1996. Continental breakup and the ordinal diversification of birds and mammals. *Nature* 381 (6579), 226–229.
- Holder, M., Lewis, P.O., 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev., Genet.* 4 (4), 275–284.
- Huelsenbeck, J.P., Bollback, J.P., 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Syst. Biol.* 50 (3), 351–366.
- Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17 (8), 754–755.
- Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294 (5550), 2310–2314.
- Hwang, D.G., Green, P., 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.* 101 (39), 13994–14001.
- Krishnan, N.M., et al., 2004. Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. *Mol. Biol. Evol.* 21 (10), 1871–1883.
- Lio, P., Goldman, N., 1998. Models of molecular evolution and phylogeny. *Genome Res.* 8, 1233–1244.
- Lunter, G., Hein, J., 2004. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* 20 (Suppl. 1), I216–I223.
- McLachlan, G.J., Krishnan, T., 1997. *The EM Algorithm and Extensions*. John Wiley & Sons, Inc, New York.
- Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3 (5), 418–426.
- Pedersen, A.-M.K., Jensen, J.L., 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* 18, 763–776.
- Perna, N., Kocher, T., 1995. Unequal base frequencies and the estimation of substitution rates. *Mol. Biol. Evol.* 12 (2), 359–361.
- Press, W.H., et al., 1992. *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press, Cambridge.
- Siepel, A., Haussler, D., 2004a. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* 11 (2–3), 413–428.
- Siepel, A., Haussler, D., 2004b. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* 21 (3), 468–488.
- Uzzell, T., Corbin, K.W., 1971. Fitting discrete probability distributions to evolutionary events. *Science* 172 (988), 1089–1096.
- Waterston, R.H., et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420 (6915), 520–562.
- Whelan, S., Goldman, N., 2004. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* 167 (4), 2027–2043.
- Yang, Z., 1994a. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39 (3), 306–314.
- Yang, Z., 1994b. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39 (1), 105–111.
- Yang, Z., Goldman, N., Friday, A., 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11 (2), 316–324.
- Zuckermandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* 8 (2), 357–366.