BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 17 2011, pages 2346–2353
doi:10.1093/bioinformatics/btr420

Genome analysis

Advance Access publication July 14, 2011

Sufficient statistics and expectation maximization algorithms in phylogenetic tree models

Hisanori Kiryu

Department of Computational Biology, Faculty of Frontier Science, The University of Tokyo, Kashiwa,
Chiba 277-8561, Japan

Associate Editor: David Posada

ABSTRACT

Motivation: Measuring evolutionary conservation is a routine step in the identification of functional elements in genome sequences. Although a number of studies have proposed methods that use the continuous time Markov models (CTMMs) to find evolutionarily constrained elements, their probabilistic structures have been less frequently investigated.

Results: In this article, we investigate a sufficient statistic for CTMMs. The statistic is composed of the fractional duration of nucleotide characters over evolutionary time, F_d, and the number of substitutions occurring in phylogenetic trees, N_s. We first derive basic properties of the sufficient statistic. Then, we derive an expectation maximization (EM) algorithm for estimating the parameters of a phylogenetic model, which iteratively computes the expectation values of the sufficient statistic. We show that the EM algorithm exhibits much faster convergence than other optimization methods that use numerical gradient descent algorithms. Finally, we investigate the genome-wide distribution of fractional duration time F_d which, unlike the number of substitutions N_s, has rarely been investigated. We show that F_d has evolutionary information that is distinct from that in N_s, which may be useful for detecting novel types of evolutionary constraints existing in the human genome.

Availability: The C++ source code of the 'Fdur' software is available at http://www.ncrna.org/software/fdur/

Contact: kiryu-h@k.u-tokyo.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 20, 2011; revised on June 17, 2011; accepted on July 10, 2011

1 INTRODUCTION

Although the number of sequenced genomes is increasing rapidly, the technologies that extract biological information from them are still in their infancy. Finding evolutionarily constrained elements in a target genome sequence is among the most successful uses of those massive genomic sequences (Asthana *et al.*, 2007; Boffelli *et al.*, 2003; Cooper *et al.*, 2005; Garber *et al.*, 2009; Mugal *et al.*, 2010; Pollard *et al.*, 2010; Prabhakar *et al.*, 2006; Siepel *et al.*, 2005). In those studies, the researchers estimated the level of the substitution rate for each genomic region and tested whether it was significantly lower or higher than an estimated rate of neutral evolution. The substitution rates were computed from a given species tree and genome-wide alignments by using variants of continuous time Markov models (CTMMs), which have been used in the maximum likelihood (ML) method of phylogenetic tree construction (Felsenstein, 1981). While the relations between the likelihoods of CTMMs and possible topologies of the phylogenetic trees have been extensively studied for many years, their probabilistic structures for a fixed tree topology, which are very important for the genome-wide studies mentioned above, have been less frequently studied.

In this article, we investigate a sufficient statistic for CTMMs that is composed of the fractional duration of each nucleotide over evolutionary time, F_d, and the number of substitutions that have occurred in a phylogenetic tree, N_s. The sufficient statistic naturally appears in expectation maximization (EM) algorithms for finding the local maximum of the likelihood function (Garber *et al.*, 2009; Hobolth and Jensen, 2005; Holmes and Rubin, 2002; Klosterman *et al.*, 2006). However, previous studies treated them only as intermediate variables for parameter optimization and did not investigate their properties or their genome-wide distributions. Siepel *et al.* (2006) developed a dynamic programming algorithm for computing the null distribution of N_s, corresponding to neutral evolution which was then used for computing P-values for constrained elements. However, no investigation of the distribution of F_d has appeared in the literature.

In the following sections, we first derive basic properties of the sufficient statistic, such as its behaviors in the limiting cases at the time where the parameter is close to zero or infinity. We derive an EM algorithm (Dempster *et al.*, 1977) for estimating the substitution matrix and the branch lengths, which iteratively computes the expectation values of the sufficient statistic, then we explain how the branch lengths used in parsimony methods are related to those in the ML method. Many previous studies of the EM algorithm optimized only some of the phylogenetic model parameters, which means that some of the expectation values required for a full parameter optimization are lacking. Holmes and Rubin (2002); Klosterman *et al.* (2006) and Garber *et al.* (2009) derived an EM algorithm for substitution matrices, but did not optimize branch lengths. Consequently, the expected durations defined by them are not appropriate expected values when the branch lengths are optimized. Hobolth and Jensen (2005) did not optimize the stationary distribution of the general time reversible model and thus missed the expectation $n^{(\zeta)}(i)$ defined in our article. Siepel and Haussler (2004) derived a partial EM algorithm for phylogenetic models but the hidden variables of their EM algorithm only included the nucleotide states on ancestral nodes, instead of considering all the possible substitution histories along tree branches, which are the random processes described by the models.

Iwasaki and Takagi (2007) presented a complete derivation of an EM algorithm for a special kind of gain-and-loss model for gene family evolution, but they explicitly used a specific parametrization for their transition matrices, so it was not possible to apply it to nucleotide substitution models. Furthermore, previous studies, except for that of (Siepel and Haussler, 2004), did not compare the run times of their EM algorithms with that of gradient descent methods of likelihood maximization. We derive a complete description of the EM algorithms for the general unrestricted (UNREST) model and the general time reversible (GTR) model (Yang, 2006). We also derive a stationary equation method for the M step in the algorithm for the GTR model. We show that the EM algorithm converges much faster than other optimization methods that use numerical gradient descent algorithms. We also show that the EM algorithm and the stationary equation method for the M step optimization have the nice property that the objective functions monotonically approach the optimal values in every iteration (Dempster *et al.*, 1977)—a property which the gradient descent methods do not possess. Finally, we investigate the genome-wide distribution of the Kullback–Leibler divergence D_F between fractional durations F_d and the stationary distribution of the CTMM, π. We show that the non-conserved promoter regions of human genes related to anatomical structure development have higher D_F values than those of other genes. We also show that the non-conserved third codon positions of human genes have higher D_F values than the background. These indicate that the fractional duration F_d contains evolutionary information that N_s does not have, which may be useful for detecting novel kinds of evolutionary constraints existing in the human genome.

2 ALGORITHMS AND IMPLEMENTATION

2.1 Sufficient statistic on a single branch

A continuous time Markov model for nucleotide evolution can be defined by a differential equation that determines the time evolution of the probability of observing each nucleotide:

$$\frac{\partial}{\partial t}P(a|b,tR)=\sum_{i\in\mathrm{Nuc}}R_{ai}P(i|b,tR),\ P(a|b,0)=\delta_{ab}\,,$$

where $P(a|b,t)$ is the probability of observing base a at time t conditioned upon base b being observed at time zero, $\mathrm{Nuc}=\{A,C,G,T\}=\{1,2,3,4\}$ is the set of nucleotides, δ_{ij} is the Kronecker delta which is 1 if $i=j$ and is 0 otherwise and $R=\{R_{ij}\}$ is the substitution rate matrix, which satisfies the conditions

$$R_{ij}>0 \text{ for } (i\neq j),\ \sum_i R_{ij}=0\,. \tag{1}$$

This differential equation can be solved by using the matrix exponential of tR:

$$P(a|b,t)=P(a|b,tR)=[\exp(tR)]_{ab}\,,$$

$$\exp(A)=I+A+\frac{A^2}{2!}+\frac{A^3}{3!}+\cdots=\sum_{k=0}^{\infty}\frac{A^k}{k!}\,.$$

Here, $[A]_{ij}$ is the (i,j) element of matrix A and I is the unit matrix. Equation (1) guarantees that the transition matrix has positive matrix elements $P(a|b,t)>0$ for $t>0$, and that it satisfies the probability conditions $P(a|b,t)\geq 0$, and $\sum_i P(i|b,tR)=1$ for all $t\geq 0$. (A proof is presented in the Supplementary Material.) Since it is difficult to handle matrix exponentials directly, we use an infinite product representation:

$$\exp(A)=\lim_{N\to\infty}\left(I+\frac{A}{N}\right)^N.$$

By applying this formula to $P(a|b,t)$, the continuous Markov model can be considered as a limit of a discrete time Markov model with transition matrix $Q=(I+tR/N)$, as follows.

$$P(a|b,tR)=\lim_{N\to\infty}P_N(a|b,tR)$$

$$P_N(a|b,tR)=\left[Q^N\right]_{ab}=\sum_{X\in\Omega_N(a,b)}\prod_{k=1}^{N}Q_{X_kX_{k-1}}$$

where $\Omega_N(a,b)$ is the set of all paths X along discrete time points $0,\dots,N$ such that $X=\{X_k\in\mathrm{Nuc}|k=0,\dots,N,X_0=a,X_N=b\}$. Q satisfies the probability conditions for a transition matrix, $Q_{ij}\geq 0$ and $\sum_i Q_{ij}=1$, if N is sufficiently large ($N\geq\max_i t|R_{ii}|$). In the following, we derive a few formulas for fixed N, then take the continuum limit $N\to\infty$. The likelihood $P(X|tR)$ of path X can be rewritten as

$$P(X|tR)=\prod_{k=1}^{N}Q_{X_kX_{k-1}}=\prod_k\left(1+\frac{tR_{kk}}{N}\right)^{N_{kk}^X}\prod_{i\neq j}\left(\frac{tR_{ij}}{N}\right)^{N_{ij}^X}$$

$$=\exp\left[\begin{array}{c}\sum_i tR_{ii}F_d(i,X)+\sum_{i\neq j}N_s(i,j,X)\log\left(tR_{ij}\right)\\ +N_{s0}(X)\log\left(\frac{1}{N}\right)+\mathcal{O}\left(\frac{1}{N}\right)\end{array}\right],$$

where N_{ij}^X is the number of $i\leftarrow j$ transitions that occurred in X, $F_d(i,X)=N_{ii}^X/N$ is the fractional duration of base i, $N_s(i,j,X)=N_{ij}^X$ is the number of substitutions $i\leftarrow j$ $(i\neq j)$, and $N_{s0}(X)=\sum_{i\neq j}N_s(i,j,X)$ is the total number of non-trivial substitutions in X. We have used the Taylor expansion $\log(1+x)=\sum_{k=1}^{\infty}(-1)^{k+1}x^k/k$ to derive the final equation. (The derivation of the last equation is described in the Supplementary Material.) The equation above indicates that the probability of each path X decreases as $\sim N^{-N_{s0}(X)}$ in the limit $N\to\infty$. On the other hand, the number of paths with the same number of non-trivial substitutions $N_{s0}(X)$ increases as $\sim N^{N_{s0}(X)}$, since the number of positioning patterns of substitution events in the time course increases roughly as the combination $\begin{pmatrix}N\\N_{s0}(X)\end{pmatrix}$. Thus, the paths with a positive number of substitutions give finite contributions in the continuum limit. Since the likelihood function depends on X only through $F_d(i,X)$ and $N_s(i,j,X)$, these form a sufficient statistic for the model (Fisher, 1922).

The expected value of $F_d(i,X)$ is

$$F_d(a,b,i)=\frac{1}{Z_N(a,b,tR)}\sum_{X\in\Omega_N(a,b)}F_d(i,X)P(X|tR)\,,$$

$$Z_N(a,b,tR)=\sum_{X\in\Omega_N(a,b)}P(X|tR)=P_N(a|b,tR)\,.$$

As shown in the Supplementary Material, the continuum limits of $F_d(i,X)$ and $N_s(i,j,X)$, which are denoted by $F_d(a,b,i,tR)$ and $N_s(a,b,i,j,tR)$, have integral representations:

$$F_d(a,b,i,tR)=\mathcal{K}(a,b,i,i,tR) \tag{2}$$

$$N_s(a,b,i,j,tR)=tR_{ij}\mathcal{K}(a,b,i,j,tR)$$

$$\mathcal{K}(a,b,i,j,tR)=\frac{1}{P(a|b,tR)}\int_0^1 ds\left[e^{(1-s)tR}\right]_{ai}\left[e^{stR}\right]_{jb}\,.$$

2.2 Sufficient statistic for phylogenetic tree models

Throughout this article, we assume that the topology of phylogenetic trees is fixed. In the following, ζ denotes the root node, $m=1,\dots,M$ denote node indexes other than the root node, in which $m=1,\dots,S$ represent the indexes of the leaf nodes, $\mathrm{Pa}(m)$ is the parent of node m, t_m is the length of branch $(m,\ \mathrm{Pa}(m))$, $T=\{t_m|m=1,\dots,M\}$ is the set of the branch lengths, and $\theta=(T,R)$. The random processes in these models correspond to the set of all the substitution histories X over the phylogenetic branches. The log likelihood of X has the following form in the finite product approximation:

$$l(X,\theta)=\sum_{m,i}t_mR_{ii}F_d^{(m)}(i,X)+\sum_{m,i\neq j}N_s^{(m)}(i,j,X)\log(t_mR_{ij})$$

$$+\sum_i n^{(\zeta)}(i,X)\log(\pi_i)+\cdots\,, \tag{3}$$

where $F_d^{(m)}(i,X)$ and $N_s^{(m)}(i,j,X)$ are the fractional duration and the number of substitutions at branch $(m,\mathrm{Pa}(m))$ computed from history X. The number $n^{(\zeta)}(i,X)$ is 1 if the base is i at the root in X, and is 0 otherwise. The ellipsis '$\cdots$' represents irrelevant terms that are independent of the model parameters T and R, or those that vanish in the continuum limit $N\to\infty$. The equation implies that the sufficient statistic of the model consists of $F_d^{(m)}(i,X)$, $N_s^{(m)}(i,j,X)$ and $n^{(\zeta)}(i,X)$. The expectation values of $F_d^{(m)}(i,X)$, $N_s^{(m)}(i,j,X)$ and $n^{(\zeta)}(i,X)$ for an alignment column C are as follows:

$$F_d^{(m)}(i,C,\theta)=\sum_{a,b}F_d(a,b,i,t_mR)P^{(m)}(a,b|C,\theta)\,,$$

$$N_s^{(m)}(i,j,C,\theta)=\sum_{a,b}N_s(a,b,i,j,t_mR)P^{(m)}(a,b|C,\theta)\,,$$

$$n^{(\zeta)}(i,C,\theta)=P^{(\zeta)}(i|C,\theta)\,.$$

Here, $P^{(m)}(a,b|C,\theta)$ is the posterior probability that the final and initial bases of branch (m, Pa(m)) are a and b, respectively. $P^{(\zeta)}(a|C,\theta)$ is the posterior probability that the base at the root node is a. These values are calculated by using inside–outside algorithms, which are described in the Supplementary Material. Finally, the fractional duration and the total number of substitutions over the entire phylogenetic tree are defined by

$$F_d(i,C,\theta)=\frac{T_d(i,C,\theta)}{\sum_i T_d(i,C,\theta)}\,,$$

$$N_{s0}(C,\theta)=\sum_m\sum_{i\neq j}N_s^{(m)}(i,j,C,\theta)\,,$$

$$T_d(i,C,\theta)=\sum_m t_mF_d^{(m)}(i,C,\theta)\,.$$

2.3 EM algorithm

The EM algorithm (Dempster *et al.*, 1977) is a method to find a local maximum of the likelihood function when there exist unobserved hidden random variables. The algorithm iteratively maximizes the likelihood function by alternately calling two distinct subroutines called the E step and the M step. In the E step, the expected values of certain statistical operators are computed by using current estimates of parameter values. In the M step, an expected log likelihood function Q_{EM} is maximized with respect to the model parameters. In the present case, the unobserved hidden variables correspond to the substitution histories X over the phylogenetic tree. We define the Q_{EM} function by

$$Q_{\mathrm{EM}}(\mathcal{A},\theta,\theta')=\frac{1}{|\mathcal{A}|}\sum_{C\in\mathcal{A}}\sum_{X\in\Omega(C)}P(X|\theta')l(X,\theta)$$

$$=\sum_{m,i}t_mR_{ii}F_{d,i}^{(m)}+\sum_{m,i\neq j}N_{s,ij}^{(m)}\log(t_mR_{ij})+\sum_i n_i^{(\zeta)}\log(\pi_i)\,,$$

where $|\mathcal{A}|$ is the number of columns in alignment $\mathcal{A}$, and $F_{d,i}^{(m)}$, $N_{s,ij}^{(m)}$ and $n_i^{(\zeta)}$ are, respectively, the means of $F_d^{(m)}(i,C,\theta')$, $N_s^{(m)}(i,j,C,\theta')$ and $n^{(\zeta)}(i,C,\theta')$ averaged over all the columns $C\in\mathcal{A}$. We have omitted the irrelevant terms in Equation (3). In the E step, we compute the expectations $F_{d,i}^{(m)}$, $N_{s,ij}^{(m)}$ and $n_i^{(\zeta)}$ for the entire alignment data. This is very compute-intensive step for genome-scale multiple alignments. In the M step, we maximize the Q_{EM} function with respect to θ with fixed expectations. Then we set θ' to the optimized values. The E and M steps are repeated until the convergence of model parameters θ. The EM algorithm has the important property that the likelihood function always increases at each iteration (Dempster *et al.*, 1977). This is not the case in the gradient descent method of parameter optimization, in which the likelihood usually undergoes several cycles of increasing and decreasing before the optimal value is reached. To maximize Q_{EM} at each M step using the gradient descent method, we must compute the gradient of Q_{EM} with respect to parameters.

We define the UNREST model, which is the most general, time-irreversible substitution model, by the parametrization $R_{ij}=\mathcal{B}_{ij}$ $(i\neq j)$, $\{\mathcal{B}_{ij}>0,\sum_{i\neq j}\mathcal{B}_{ij}=1\}$. Then, the derivatives of Q_{EM} with respect to $\mathcal{B}$ and t_m are given by

$$\frac{\partial Q_{\mathrm{EM}}}{\partial\mathcal{B}_{ij}}=\sum_m\left[-t_mF_{d,j}^{(m)}+N_{s,ij}^{(m)}\frac{1}{\mathcal{B}_{ij}}\right]+\sum_k n_k^{(\zeta)}\frac{1}{\pi_k}\frac{\partial\pi_k}{\partial\mathcal{B}_{ij}}\,,$$

$$\frac{\partial Q_{\mathrm{EM}}}{\partial t_m}=\sum_i R_{ii}F_{d,i}^{(m)}+N_{s0}^{(m)}\frac{1}{t_m}\,.$$

The derivatives of the stationary distribution are given by

$$\frac{\partial\pi_k}{\partial\mathcal{B}_{ij}}=-\sum_l(\delta_{kl}-\pi_k\cdot 1)(R_{li}^{+}-R_{lj}^{+})\pi_j\,,\tag{4}$$

where $R^{+}=UD^{+}U^{-1}$ and $D^{+}=\mathrm{diag}(0,d_2^{-1},\ldots,d_4^{-1})$ is the Moore-Penrose pseudoinverse of R (Ben-Israel and Greville, 2003). The proof of this equation is described in the Supplementary Material. In the above equations, $\mathcal{B}_{ij}$ $(i\neq j)$ are regarded as independent variables and the conditions $\mathcal{B}_{ij}>0$ and $\sum_{i\neq j}\mathcal{B}_{ij}=1$ are not taken into account. A simple method to deal with these conditions in gradient descent methods is described in the Supplementary Material.

The GTR model is defined by the parametrization $R_{ij}=\pi_iB_{ij}$ $(i\neq j)$, $\{\pi_i>0,\sum_i\pi_i=1,B_{ij}>0,B_{ji}=B_{ij}(i<j),\sum_{i<j}B_{ij}=1\}$, where π is the stationary distribution. The derivatives of Q_{EM} with respect to $\theta=(\pi,B,T)$ are given by

$$\frac{\partial Q_{\mathrm{EM}}}{\partial\pi_i}=\sum_{m,j(\neq i)}\left[-t_mB_{ij}F_{d,j}^{(m)}+N_{s,ij}^{(m)}\frac{1}{\pi_i}\right]+n_i^{(\zeta)}\frac{1}{\pi_i}\,,$$

$$\frac{\partial Q_{\mathrm{EM}}}{\partial B_{ij}}=\sum_m\left[-t_m(\pi_iF_{d,j}^{(m)}+\pi_jF_{d,i}^{(m)})+(N_{s,ij}^{(m)}+N_{s,ji}^{(m)})\frac{1}{B_{ij}}\right].$$

The derivatives $\partial Q_{\mathrm{EM}}/\partial t_m$ are the same as in the UNREST model. In the above equations, all the parameters π_i and B_{ij} $(i<j)$ are regarded as independent variables. For GTR models, there is an alternative to the gradient descent method for solving the M step problem. In this method, we maximize Q_{EM} by iteratively solving the stationary equations. The stationary equations are satisfied at any local maximum and can be obtained by setting all the derivatives to zero. The stationary equations can be formally solved in the GTR model and the solutions are given by

$$\pi_i=\frac{\sum_{j(\neq i)}N_{s,ij}+n_i^{(\zeta)}}{\sum_{j(\neq i)}B_{ij}T_{d,j}+\lambda_\pi}\,,$$

$$B_{ij}=\frac{N_{s,ij}+N_{s,ji}}{\pi_iT_{d,j}+\pi_jT_{d,i}+\lambda_B}\,,$$

$$t_m=\frac{N_{s0}^{(m)}}{\sum_i|R_{ii}|F_{d,i}^{(m)}}\,,N_{s0}^{(m)}=\sum_{i\neq j}N_{s,ij}^{(m)}\,.$$

Here, λ_π and λ_B are the Lagrange multipliers that impose the probability conditions on π and B, respectively. The Lagrange multipliers can be efficiently computed by using a Newton–Raphson method (Press *et al.*, 1992), which is described in the Supplementary Material. The above equations can be solved for each of π, B and T, but not simultaneously. Therefore, we serially update those three types of parameters until convergence. Because Q_{EM} is exactly maximized with respect to the currently updating parameter type, the method is guaranteed to converge (Drton, 2004; Lauritzen, 1996). This also indicates that the stationary method has the monotonicity property that Q_{EM} increases at every iteration (Dempster *et al.*, 1977).

Since the stationary equations are satisfied at any local maximum of the log likelihood, it is possible to derive an interesting connection between the branch lengths in ML solutions and those in parsimony methods. The stationary equations also suggest that there is a natural definition of the posterior local tree, which has the same topology as that of the global tree but has different edge lengths, for each genomic region. A more detailed description of these issues is given in the Supplementary Material.

2.4 Derivatives of the likelihood function

Using the derivative formula of matrix exponentials presented in the Supplementary Material as well as the derivative formula of the stationary distribution [Equation (4)], we can find the derivatives of the log likelihood function with respect to parameters (T, R). We present the explicit formulas in the Supplementary Material. Those formulas are used in the gradient descent methods in the following sections. We have described the comparison of the EM algorithm of Siepel and Haussler (2004) and the gradient descent method and our EM algorithm in the Supplementary Material. We have also described the relationships between the stationary equations and the approaches that optimize a part of model parameters for each alignment column to detect signals of evolutionary selection (Cooper *et al.*, 2005; Garber *et al.*, 2009).

3 DATASET AND DATA PROCESSING

3.1 Training model parameters

We trained the phylogenetic model from the Multiz 44-way dataset, which is a large set of alignments created from 44 vertebrate genome sequences including the human genome (hg18) that is provided at the UCSC genome browser site (Blanchette *et al.*, 2004; Kent *et al.*, 2002). We used a fixed tree topology that is also provided at the UCSC site and obtained the maximum likelihood estimates of the GTR model parameters and branch lengths. Because many of the alignments include only human and closely related monkey species, and partly because of limited computational resources, we used only a subset of the Multiz 44-way alignments, which was selected by the condition that fugu (fr2) is contained in each alignment. The number of alignment columns was 74×10^6 and the data size of the alignment file in MAF format was 3.1 GB. The species that is least frequently aligned is petMar1 which is included in 38% (28×10^6) of the alignment columns. Fugu was the best species for such filtering in the sense that the coverage of the species with the lowest coverage in the alignment was the maximal. For example, if we used petMar1 instead of fr2 for selecting the alignment subset, then the taeGut1 genome covers only 29% (14×10^6) of the alignment columns.

3.2 Distance measure for $F_d(i)$

To measure the changes of fractional duration $F_d(i)$, we computed the Kullback–Leibler divergence D_F between $F_d(i)$ and the stationary distribution π_i in the model:

$$D_F = \sum_i F_d(i)\log_2(F_d(i)/\pi_i),$$

which is non-negative and zero if $F_d(i) = \pi_i$. Since the trained stationary distribution is very close to the uniform distribution, the maximal value of D_F is reached when the distribution of $F_d(i)$ is close to a single-peak distribution δ_{ia} for some $a \in \mathrm{Nuc}$.

3.3 Comparison of gradient descent and EM algorithms

The run time of numerical optimization depends on many factors, and the optimal implementation can be completely different for different alignment sizes and tree topologies. This means that it is, in general, difficult to judge whether differences of running time are caused by differences between the underlying algorithms or by differences at the implementation level. However, here we are interested in optimizing parameters for genome-scale alignments. In such cases, the most time consuming step is the computation of inside–outside variables, which is required for each computation of gradient in the gradient descent methods, and is also required for each E step in the EM method. Hence, the number of times that a program passes through the alignment data until convergence is an accurate measure for comparing the convergence properties of the algorithms themselves, at least when the input contains a large number of alignments. In the following section, we used 10 000 alignment columns sampled from the Multiz 44-way alignment and plotted the relative differences of likelihood or parameter values from the optimal values. The relative difference of log likelihood l from the optimal value l_{opt} is defined by $|l - l_{\mathrm{opt}}|/|l_{\mathrm{opt}}|$). Similarly, the relative difference of parameter vector x is defined by $||x - x_{\mathrm{opt}}||/||x_{\mathrm{opt}}||$, where $||\cdot||$ represents the Euclid norm. For gradient descent, we used a BFGS (Press *et al.*, 1992) routine implemented in the PHAST package (Siepel *et al.*, 2005) and the FORTRAN code of the L-BFGS-B (Nocedal, 1980; Zhu and Byrd, 1997) routine provided by the original authors.

3.4 Gene Ontology term analyses of promoter regions

We collected the human RefSeq genes from the UCSC site and selected a representative RefSeq gene model with the longest transcript size for each Entrez gene. Then we collected a 1000 base sequence upstream of each transcription start site, and assigned the Entrez gene identity for it. To reduce the effects of possible overlaps with different gene isoforms, we masked the promoter positions that overlap with any of the putative gene models provided at the UCSC site (the UCSC genes). Then, the Entrez genes were sorted according to the conservation properties in their associated promoter regions. We considered three sorting criteria:

(i) The ratio of the number of conserved promoter sites to that of unmasked sites. The conserved sites are defined by positions with N_{s0} less than the median of the distribution $(N_{s0,\mathrm{median}} = 5.41)$.

(ii) The ratio of the number of non-conserved sites $(N_{s0,\mathrm{median}} < N_{s0})$ to that of unmasked sites.

(iii) The ratio of the number of non-conserved and highly biased F_d sites to that of non-conserved sites. The non-conserved and highly biased F_d sites are defined as follows. First, we collected the non-conserved sites by the filtering $N_{s0,\mathrm{median}} < N_{s0}$. Then, we divided the sites into 100 bins according to their N_{s0} values. For each bin, the upper 5% of the sites with high D_F values are defined as the non-conserved, highly biased sites. These sites are characterized by their stronger nucleotide preference despite the lack of conservation (see Fig 5 for more explanation).

The Entrez genes were divided into 100 bins by each sorting criterion. For each bin ID i, we performed gene ontology (GO) term enrichment analyses for the gene set with bin ID $\le i$ by computing hypergeometric P-values. The number of non-trivial GO terms was 8000 for our dataset, which indicates that the hypergeometric P-values with $-\log_{10}(P\text{-value}) > 5.2$ correspond to Bonferroni corrected $P < 0.05$. The plots in Figure 6 are inspired by the ones proposed in van Dongen *et al.* (2008) to analyze the gene repression effects of miRNAs and siRNAs.

3.5 Evolutionary characteristics of 4-fold degenerate sites

The 4-fold degenerate (4d) sites are the third codon positions where any kind of mutation does not cause a change of amino acid sequence. We collected the human 4d sites from the representative RefSeq genes. We divided the human genome into the first, second and third codon positions and the remaining positions. The third codon positions are subdivided into 4d sites and non-4d sites. For each position, we computed the pair of statistics (N_{s0}, D_F) and dropped the conserved positions which have a number of substitutions $N_{s0} < N_{s0,\mathrm{median}}$. We divided the data into 100 bins according to their N_{s0} values, and compared the D_F distributions for each bin k. We consider the D_F distribution $f_k(d)$ of bin k in the non-coding positions as the background distribution. We computed the upper quantile values $q(y) = \sum_{d=y}^{\infty} f_k(d)$ for each D_F value y of the coding sites in the same bin k. Then we plotted the distribution of quantile values q for all the non-conserved N_{s0} bins for each codon type (Fig. 8). The peaks of those distribution at small q indicate that the codon positions tend to have high D_F values compared with the background at the same conservation level.

4 RESULTS AND DISCUSSION

4.1 EM algorithm

Figure 1 shows a comparison of the convergence of log likelihood and model parameters between the gradient descent methods and the EM algorithm. The x-axes represent the number of iterations that the input alignments are traversed to compute the gradient vector or the expectation values. The y-axes are the relative distances from the optimal values. Although the differences of convergence of log likelihood values might appear small, the convergence rates of parameters are considerably different. The EM algorithm converges much faster toward the optimal value than the gradient descent methods. The figure shows the monotonic convergence of the EM algorithm (Dempster *et al.*, 1977) while the parameters of the gradient descent methods show oscillatory behavior before reaching the optimum.

Figure 2 shows the convergence of the M step maximization. The x-axes represent the number of updates of model parameters. The y-axes represent the relative distance of the parameters from the optimal values. The stationary equation method converges more rapidly than the LBFGS method. At the 10-th EM iteration (Fig. 2B), the stationary equation method makes almost no changes to the parameters since the initial parameters are already close to the optimal values. On the other hand, the gradient descent method alters them from the optimal values at least once since it needs to recognize the local landscape of the objective function.

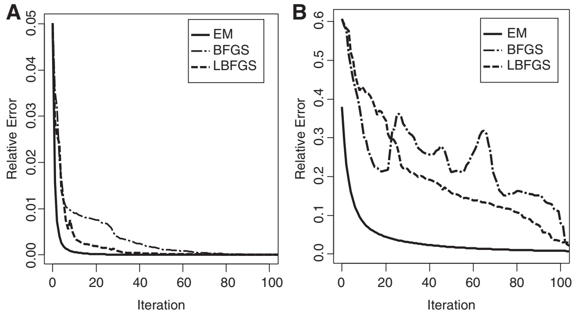

Fig. 1. Comparison of the convergence of log likelihood (**A**) and model parameters (**B**) between the gradient descent methods and the EM algorithm.

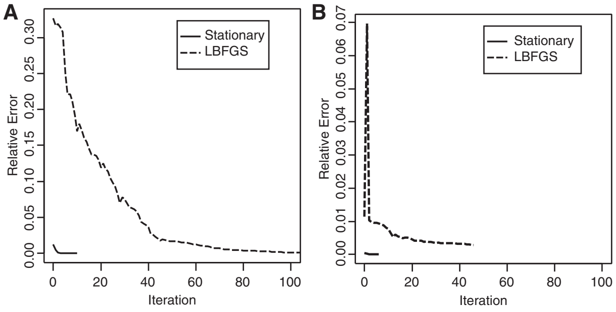

Fig. 2. Comparison of convergence in gradient descent methods and the stationary equation method. (**A**) M step maximization at the first EM iteration. (**B**) M step maximization at the 10-th EM iteration.

Table 1 shows the comparison of run time of three optimization methods. The computation was performed on a Intel Xeon E5450 3 GHz with 32 GB memory. It is noted that the BFGS method is slower than the others since the optimization code requires additional log likelihood function evaluations other than at each gradient computation.

4.2 Genome-wide distribution of F_d and N_s

Figure 3 shows the genome-wide distribution of D_F and N_{s0}. The genomic data are divided into equally sized bins according to D_F or N_{s0} values. Then we compared the proportion of each type of genomic region (Repeat, Intergenic, Intron, UTR and CDS) in each bin with that in the entire distribution. The bars represent $\log_2$(fold enrichment) of each genomic region. The figure shows that CDS is enriched and Repeat is depleted in high D_F or low N_{s0} regions. The repeat elements are depleted in both low and high N_{s0} regions, because most of them have small number of aligned species, and thus their N_{s0} values are close to the prior expectation $N_{s0} = 5.28$ for the alignment column filled with missing characters. Figure 4 shows the correlation between D_F and N_{s0} in the human genome. Although D_F and N_{s0} are roughly inversely correlated, a large variability of D_F values for the same N_{s0} values is observed.

Figure 5 illustrates how differences of D_F can occur with the same number of substitutions N_{s0}. Both (A) and (B) have a single substitution event, but (A) has a substitution only recently and (B) has a substitution in a deep ancestral branch. This difference results in an approximately single-peak $F_d(i)$ for (A) and a double-peak $F_d(i)$ distribution for (B), and hence D_F of (A) is larger than that

Table 1. The run time of three methods

Name	Iteration	Run time (s)	Run time per iteration (s)
EM	83	143	1.67
LBFGS	134	271	2.00
BFGS	107	319	2.95

The input alignment size is 10 000 columns of 44 species. The unit of time is second. Iteration represents the number of the E steps or the gradient computations that are required until the relative errors of the parameters become <0.01. Run Time shows the total time spent for these iterations. Run Time Per Iteration is the average time spent for each iteration.

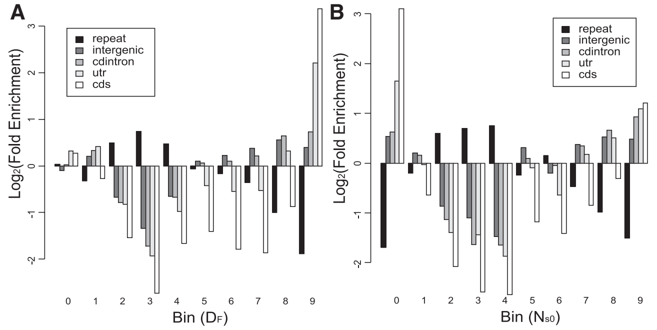

Fig. 3. Genome-wide distribution of D_F and N_{s0}. The bars represent $\log_2$(fold enrichment) of genome regions (Repeat, Intergenic, Intron, UTR and CDS) of each bin from the entire distribution. (**A**) The data points are binned by the D_F values. (**B**) The data points are binned by the N_{s0} values.

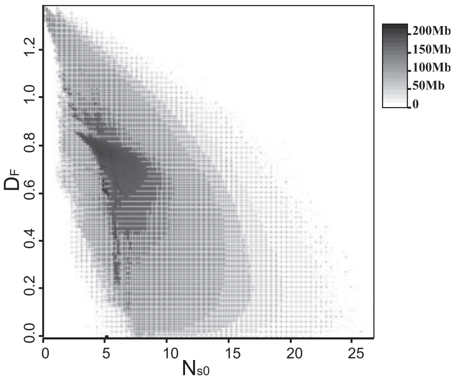

Fig. 4. Scatter plot of pair (N_{s0}, D_F).

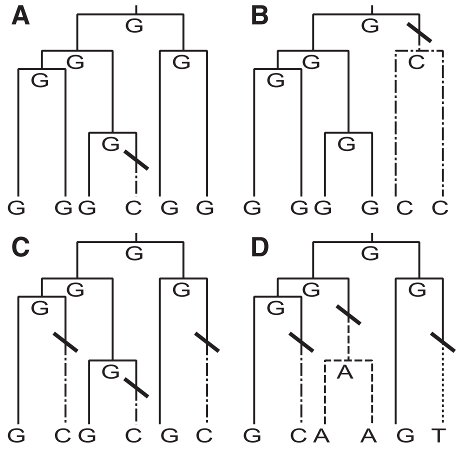

Fig. 5. Illustration of the combination of conservation patterns and D_F. The back slashes indicate substitution events. The branches that remained as base G are represented by solid lines. Similarly, The evolutionary time courses that stayed as base C, A and T are represented by dot-dash, broken and dotted lines, respectively.

of (B). Next, both (C) and (D) have three substitutions, but (C) has a strong bias which prefers base G *or* C, while (D) has no such a preference, which results in a larger D_F value for (C) than for (D).

4.3 GO term analysis of human promoters

Figure 6 shows GO term enrichment analysis of promoter regions. Positive values on the *y*-axes represent GO term enrichment and negative values represent depletion. The *x*-axes represent gene ranks (or bin identities) which are determined according to the sorting criteria: (i) high conservation (Fig. 6A); (ii) high non-conservation (Fig. 6B); and (iii) non-conserved but highly biased F_d (or high D_F) values (Fig. 6C), which correspond to the conservation pattern (C) of Figure 5. Figure 6A shows that the genes with conserved promoters are enriched with the term ‘regulation of transcription’ (GO:0045449). Figure 6B is essentially an inversion of the Figure 6A and shows that the genes with non-conserved promoters are depleted of ‘regulation of transcription’. Figure 6C shows that the genes with promoters that have highly biased F_d values despite low level of conservation, which are not reported as constrained regions by previous methods based on N_{s0} values, are enriched with the term ‘anatomical structure development’ (GO:0048856). These figures show that the fractional duration F_d captures distinct evolutionary constraints which are not contained in the expected number of substitution N_{s0}.

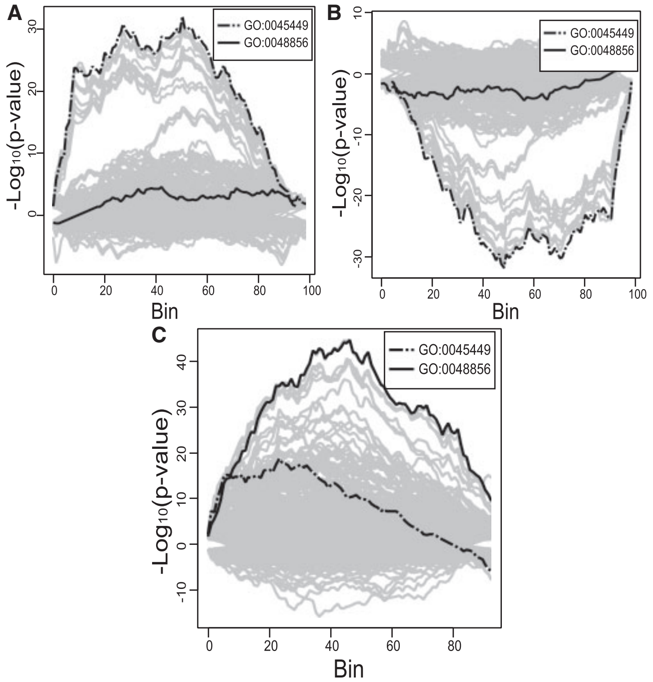

Fig. 6. GO term analysis of promoter regions. The genes are sorted in descending order according to (i) the fraction of conserved sites (**A**); (ii) the fraction of non-conserved sites (**B**); (iii) the fraction of non-conserved, highly biased sites (**C**). See Section 3 for more explanation of the sorting criteria. Solid line: GO:0048856, anatomical structure development. Dot-dash line: GO:0045449 regulation of transcription. Gray lines: other GO terms. Positive values on the *y*-axes represent GO term enrichment and negative values represent depletion of GO terms

4.4 D_F distribution in coding regions

Figure 7 shows the normalized N_{s0} distribution of each codon phase position. Large portions of the first and second codon positions are strongly conserved and have small N_{s0} values. Both the non-4d sites and the 4d sites have broader tails than the non-coding sites. The 4d sites have thick tail in the large N_{s0} values.

Figure 8 shows the D_F distribution of each codon phase position relative to the background distribution. As described in the Section 3, the data only include non-conserved sites. The sharp peaks around the zero quantile in the Figure 8A indicate that all the codon positions are subject to a strong bias that constrains F_d to be far from the stationary distribution, unlike the background distribution at the same level of conservation N_{s0}. The fact that this bias is prominent even in the 4d sites indicates that it might capture the selective pressures related to transcription and translation efficiency, such as codon bias and CpG islands. There is, however, another possible reason behind this bias for the 4d sites. Since the 4d sites are within coding regions, many species are aligned on those positions. On the other hand, non-coding regions which are used as

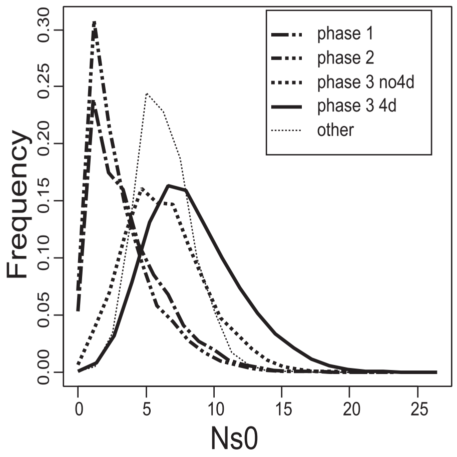

Fig. 7. Normalized N_{s0} distribution of each codon site in the human genome.

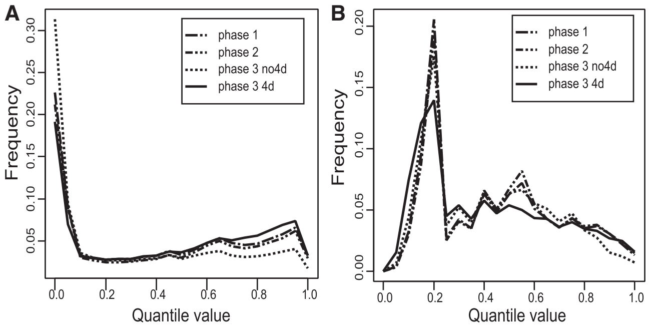

Fig. 8. D_F distribution of 4-fold degenerate sites. (**A**) D_F values are computed for the entire tree and dataset. (**B**) D_F values are computed for the subtree which has human, rhesus, mouse and dog as the tree leaves, and the data only include the sites that have no ambiguous or gap characters for these four species.

the background distribution generally have fewer aligned species, which might cause artificial differences between the distributions. To investigate this effect, we computed the fractional duration $F_d(i)$ on the subtree that has human, rhesus, mouse and dog as the tree leaves. Furthermore, we removed the data points that have any ambiguous or gap characters for these four species. Figure 8B shows the recomputed distribution. It still shows peaks at small quantile values, which indicates that the observed bias to the fractional duration F_d is not caused by the difference of the number of aligned species. These figures indicate that the codon sites including the 4d sites show distinct behaviors from the non-coding sites in terms of D_F distributions even in the extremely diverged sites.

5 CONCLUSION

We have investigated the sufficient statistic, composed of the fractional duration time F_d and the number of substitutions N_s, for continuous time Markov models. We have derived EM algorithms for the general unrestricted (UNREST) model and the general time reversible (GTR) model. In the case of the GTR model, we have derived a stationary equation method for solving the M step optimization problem. Both the EM algorithm and the stationary equation method have a property which is not shared by the gradient descent methods: the objective functions monotonically approach to their optimal values (Dempster *et al.*, 1977). We have also shown that the EM algorithm converges faster than the gradient descent methods. Although we did not describe EM algorithms for other phylogeny models with discrete gamma rate variations (Yang, 1994) and invariance sites (Gu *et al.*, 1995), the application of our method to such models is straightforward. There is a natural definition of local posterior trees which are derived from the stationary equations, and we have discussed the relationships between the branch lengths in the maximal likelihood method and those in the parsimony method. Finally, we have investigated the genome-wide distributions of the sufficient statistic. We have shown that the fractional duration F_d has evolutionary information that is not contained in the expected number of substitutions N_{s0}.

In our experiments on the F_d distributions, we have focused on only the point that the fractional duration F_d contains information different from those contained in the expected substitution count N_{s0}, which is not obvious from the outset since conserved regions ($N_{s0} \sim 0$) naturally have biased distributions (larger D_F) and thus they are strongly anti-correlated. In order to establish a novel method for finding novel constrained elements based on statistics (N_{s0}, D_F), several additional factors should be considered and investigated. First, we need to explore at least 2D space (N_{s0}, D_F) of statistics. It means that we have to determine a threshold *curve* rather than a single threshold value to determine the constrained sites. Also, more precise characterization of the correlations between (N_{s0}, D_F) and the patterns of aligned species (i.e. the 0–1 patterns representing whether each species have an aligned nucleotide or not in an alignment column). We have touched on this issue in the codon sites experiment, but discarding most of the information from the 44 species alignments and keeping only four species for computing statistics is not a satisfactory method. It requires the genome-wide distribution of 2^{44} possible patterns of alignment column types. To make any statistical statement on the level of constraint, computation of the theoretical distribution will be useful. We have only analyzed the empirical distributions of the sufficient statistic and have not analyzed the distribution determined by the phylogenetic model. There is a dynamic programming algorithm for computing the distribution of N_{s0} (Siepel *et al.*, 2006). Although it seems difficult to compute the theoretical distribution of F_d with similar methods, it can be computed by a sampling method as shown in the Supplementary Material. Furthermore, we must investigate other nuisance parameters such as the length scales of the constrained elements in order to decide if a consecutive segment is constrained or not. Lastly, we assumed that the genome-scale alignments provided at the UCSC site are completely reliable. However, there are studies that suggest that the errors contained in the Multiz alignments may be considerable (Frith *et al.*, 2008; Prakash and Tompa, 2007). It will be interesting to investigate to what extent the computed statistics are sensitive to the errors of the alignments.

ACKNOWLEDGEMENTS

The author thank their colleagues who worked on the project for discussions and comments.

Funding: 'Functional RNA Project' funded by the New Energy and Industrial Technology Development Organization (NEDO) of Japan; Grants-in-Aid for Young Scientists (21700330); Grants-in-Aid for Scientific Research (A) (22240031). Supercomputing resources were provided by the Human Genome Center, University of Tokyo (http://sc.hgc.jp/shirokane.html).

Conflict of Interest: none declared.

REFERENCES

Asthana,S. *et al.* (2007) Analysis of sequence conservation at nucleotide resolution. *PLoS Comput. Biol.*, **3**, e254.

Ben-Israel,A. and Greville,T. N. (2003) *Generalized Inverses.* Springer, New York, USA.

Blanchette,M. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.

Boffelli,D. *et al.* (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.

Cooper,G.M. *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Sect. B*, **39**, 1–38.

Drton,M. (2004) Maximum Likelihood Estimation in Gaussian AMP Chain Graph Models and Gaussian Ancestral Graph Models. PhD Thesis, University of Washington.

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

Fisher,R. (1922) On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser. A*, **222**, 309–368.

Frith,M.C. *et al.* (2008). The whole alignment and nothing but the alignment: the problem of spurious alignment flanks. *Nucleic Acids Res.*, **36**, 5863–5871.

Garber,M. *et al.* (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, 54–62.

Gu,X. *et al.* (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.*, **12**, 546–557.

Hobolth,A. and Jensen,J.L. (2005) Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article18.

Holmes,I. and Rubin,G. M. (2002) An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.*, **317**, 753–764.

Iwasaki,W. and Takagi,T. (2007). Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. *Bioinformatics*, **23**, i230–i239.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Klosterman,P.S. *et al.* (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics*, **7**, 428.

Lauritzen,S. (1996). *Graphical Models.* Clarendon Press, Oxford, UK.

Mugal,C.F. *et al.* (2010) Conservation of neutral substitution rate and substitutional asymmetries in mammalian genes. *Genome Biol. Evol.*, **2**, 19–28.

Nocedal,J. (1980) Updating Quasi-Newton Matrices with limited storage. *Math. Comput.*, **35**, 773–782.

Pollard,K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Prabhakar,S. *et al.* (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science*, **314**, 786.

Prakash,A. and Tompa,M. (2007). Measuring the accuracy of genome-size multiple alignments. *Genome Biol.*, **8**, R124.

Press,W.H. *et al.* (1992) *Numerical Recipes in C: The Art of Scientific Computing.* 2nd edn. Cambridge University Press, NY, USA.

Siepel,A. and Haussler,D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, **21**, 468–488.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Siepel,A. *et al.* (2006) New methods for detecting lineage-specific selection. In *Proceedings of 10th International Conference on Research in Computational Molecular Biology*, Springer, Berlin and Heidelbelg; GmbH & Co. KG, Berlin/DE.

van Dongen,S. *et al.* (2008) Detecting microRNA binding and siRNA off-target effects from expression data. *Nat. Methods*, **5**, 1023–1025.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.

Yang,Z. (2006) *Computational Molecular Evolution*. Oxford University Press, Oxford, UK.

Zhu,C. and Byrd,R.H. (1997) L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Trans. Math. Softw.*, **23**, 550–560.