Estimating the Frequency of Events That Cause Multiple-Nucleotide Changes

Simon Whelan*†1 and Nick Goldman*

*EMBL-European Bioinformatics Institute, Hinxton CB10 1SD, United Kingdom and †Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, United Kingdom

Manuscript received October 15, 2003
Accepted for publication April 21, 2004

ABSTRACT

Existing mathematical models of DNA sequence evolution assume that all substitutions derive from point mutations. There is, however, increasing evidence that larger-scale events, involving two or more consecutive sites, may also be important. We describe a model, denoted SDT, that allows for single-nucleotide, doublet, and triplet mutations. Applied to protein-coding DNA, the SDT model allows doublet and triplet mutations to overlap codon boundaries but still permits data to be analyzed using the simplifying assumption of independence of sites. We have implemented the SDT model for maximum-likelihood phylogenetic inference and have applied it to an alignment of mammalian globin sequences and to 258 other protein-coding sequence alignments from the Pandit database. We find the SDT model’s inclusion of doublet and triplet mutations to be overwhelmingly successful in giving statistically significant improvements in fit of model to data, indicating that larger-scale mutation events do occur. Distributions of inferred parameter values over all alignments analyzed suggest that these events are far more prevalent than previously thought. Detailed consideration of our results and the absence of any known mechanism causing three adjacent nucleotides to be substituted simultaneously, however, leads us to suggest that the actual evolutionary events occurring may include still-larger-scale events, such as gene conversion, inversion, or recombination, or a series of rapid compensatory changes.

The growing popularity of the maximum-likelihood (ML) approach to phylogenetic inference has led to increased interest in attempting to improve the modeling of the evolutionary process at the molecular level. This is due to the expectation that more realistic models will both increase our understanding of molecular evolution and lead to more accurate estimates of phylogenetic trees. For example, the models introduced by Kimura (1980) and Yang (1994a) have helped improve phylogenetic tree estimation (see, e.g., Yang et al. 1994) by the inclusion of a parameter describing a bias toward transition mutation and a parameter describing heterogeneity of evolutionary rate along a biological sequence, respectively. Models of codon evolution have provided insights into how protein-coding sequences evolve by allowing the identification of the overall type of selection occurring on specific branches of a tree (Yang 1998) and the specific sites where selection has been acting in a protein throughout its evolutionary history (Nielsen and Yang 1998). Recently, Yang and Nielsen (2002) have extended this approach by introducing a hybrid of these models that allows selection to vary both among sites and across the tree, demonstrating that an improved understanding of molecular evolution can be achieved by carefully increasing the complexity of a model through the addition of a limited number of biologically interpretable parameters. ML also provides the opportunity to compare models in a robust statistical framework (Edwards 1972; Yang et al. 1994). These techniques have been used to good effect in choosing models appropriate for phylogenetic inference (e.g., Posada and Crandall 1998; Whelan and Goldman 2001) and the detection of positive selection in phylogenies (e.g., Yang et al. 2000).

Current models applied at the nucleotide level describe the evolutionary process as point events occurring independently at individual nucleotide sites, i.e., a series of events each causing a single nucleotide to change and unaffected by surrounding sequence. Existing models describing the evolution of codons enforce the same assumption by allowing only a single nucleotide in a codon to change in any given event (Goldman and Yang 1994; Muse and Gaut 1994; Nielsen and Yang 1998; Yang 1998; Yang et al. 2000; Yang and Nielsen 2002). This implies that amino acid changes requiring two or more nucleotide changes at the codon level may occur only via a series of intermediate steps. Modeling evolution in this manner is biologically appealing because our knowledge of the molecular mechanisms causing DNA sequence to change suggests the majority of mutations that occur in the germline are due to errors made at individual sites during the replication of DNA. Additionally, the neutral theory suggests that these mutations will become fixed independently of one another
in a population at a constant rate, thus appearing over evolutionary time as independent point events (Kimura 1984).

Biological mechanisms of larger-scale events that affect contiguous nucleotides have been known for some time. If the frequency of such events is relatively high, then failure to consider them in an evolutionary model may affect estimates of phylogenetic tree topologies and other factors important in the evolutionary process, for example, selection or rate heterogeneity. The presence of large-scale events may also have an impact on current methodologies used for homology searching in databases and sequence alignment. However, such events are thought to be rare and this has led to the widely made assumption that sites evolve independently. This is also computationally appealing, having the added benefit of simplifying the mathematics of evolutionary modeling. Relaxing the independent sites assumption has proven to be difficult and there are no widely used solutions to date, although some progress has been made. For example, Pedersen and Jensen (2001) used a sophisticated approach describing the accelerated rate of C → T mutation in the context of a CpG dinucleotide. Unfortunately this approach was limited to pairwise sequence comparison and incurs a heavy computational burden.

Several sources of information have suggested that the assumption that all mutations occur as independent point events at individual nucleotide sites should be questioned. Our best estimates of protein evolution have nonzero instantaneous rates of change between amino acids whose codons differ by more that one nucleotide (for example, see Dayhoff et al. 1972, 1978; Whelan and Goldman 2001), thus suggesting single events may change multiple nucleotides on a regular basis. Furthermore, Yang et al. (1998, Table 1) compared a model where only amino acids that differ by a single nucleotide were allowed to interchange with a model that allowed all possible amino acid interchanges and found that the latter, more general, model yielded a significantly better description of the evolutionary process. Most recently, there has also been some debate over the relative importance of doublet events (an observed change in two adjacent nucleotide sites due to a single event). Averof et al. (2000) provided evidence that doublet mutation occurs at ~2% of the rate of point mutation in a globin pseudogene, an amount high enough to cause concern for those using evolutionary models. However, Smith et al. (2003) have performed similar analyses on genomic and SNP data and their results suggested that the rate may be far lower, at ~0.3%, when regional rate variation is taken into account; other authors have also contested the frequency of doublet mutations (e.g., Silva and Kondrashov 2002).

In this study we develop a new model, the singlet-doublet-triplet (SDT) model, which incorporates events that change one, two, or three adjacent nucleotides. A series of submodels, created by placing various restrictions on the SDT model, are used to investigate the relative importance of events involving single-nucleotide, doublet, and triplet changes. The models, applicable to protein-coding DNA sequences, take account of changes spanning two codons yet still permit data analysis to proceed using the simplifying assumption of independent codons. We apply these models to an aligned set of mammalian globin sequences and a substantial subset of a large database of nonredundant biological sequence alignments to estimate the relative frequency of point, doublet, and triplet events. Likelihood-ratio tests show that allowing for doublet and triplet events substantially improves the fit of the model to the data in nearly all cases. We find that the estimated frequency of doublet and triplet events prior to and after selection is surprisingly high, and that estimates of other evolutionary parameters in the model are affected by including such events in the model. Alignment errors and varying degrees of selection may explain some of this estimated abundance of doublet and triplet events, but the levels of these events are so high that alignment and selection artifacts are unlikely to be the sole source. We propose that biologically recognized large-scale events affecting long stretches of DNA occurring during the evolutionary history of the sequences (e.g., gene conversion, inversion, and recombination) or compensatory change are a likely source for these events. The model we introduce, while not completely describing such events, may help reduce the bias introduced when estimating phylogenies and evolutionary model parameters for data where large-scale events have occurred.

MATERIALS AND METHODS

In the past, models of DNA sequence evolution have assumed that only point mutations occur. To attempt to make a codon model that more accurately describes the inferred process of coding sequence evolution, we have developed a model in which evolution occurs according to three concurrent mutational processes, moderated by a common selection process.

DNA mutation processes: The first mutation process is of point (singlet) mutations, as is usual in codon models; the second is of doublet mutations, in which two consecutive nucleotides are simultaneously changed; and the third is of triplet mutations that change three consecutive nucleotides. We now describe these mutational processes and the method by which they are scaled and eventually combined within the SDT model of evolution.

On the basis of previous studies comparing models of single-nucleotide substitutions, and because of its success as the basis of modeling mutation in previous codon models, we have chosen to model single-mutation events using the Hasegawa-Kishino-Yano model (Hasegawa et al. 1985), assumed to apply independently and identically at each sequence site. This leads to the use of two parameters (totaling 4 d.f.), the transition/transversion
rate bias parameter $\kappa$ and the vector-valued parameter $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ describing the base frequencies expected at a site at equilibrium of this evolutionary process. The underlying single-nucleotide mutation process now has the following instantaneous rate of substitution of nucleotide $i$ with $j$:

$$s^*_i = \begin{cases} 
\pi_i & \text{if } i \text{ and } j \text{ differ by a transition} \\
\kappa \pi_j & \text{if } i \text{ and } j \text{ differ by a transversion} \\
0 & \text{otherwise.}
\end{cases}$$

(1)

Note that the instantaneous rate of substitution of $i$ with $j$ is defined to be zero.

Doublet and triplet mutations are defined as events that cause two and three consecutive nucleotides to change, respectively. This means, for example, that dinucleotides affected by a doublet mutation can interchange only with dinucleotides that differ in both their bases; from each of the 16 dinucleotides $j_{ij}$ there are nine permitted doublet mutations to the dinucleotides $j_{ij}$ with $i \neq j$ and $i_2 \neq j_2$. We have little prior knowledge about what form the processes of doublet or triplet mutation events might take and we expect them to be reasonably rare, so it seems best to use a very simple mathematical model to describe them. We therefore use the following underlying substitution models for doublets and triplets. The instantaneous rate of replacement of the doublet $i_{i1}i_2$ with $j_{j1}j_2$ is given by

$$d^*_i = \begin{cases} 
\pi_i \pi_j & \text{if } i_1 \neq j_1 \text{ and } i_2 \neq j_2 \\
0 & \text{otherwise.}
\end{cases}$$

(2)

Similarly, the instantaneous rate of replacement of the triplet $i_1i_2i_3$ with $j_1j_2j_3$ is given by

$$t^*_i = \begin{cases} 
\pi_i \pi_j \pi_k & \text{if } i_1 \neq j_1, i_2 \neq j_2, \text{ and } i_3 \neq j_3 \\
0 & \text{otherwise.}
\end{cases}$$

(3)

Note that we use “triplet” to mean three consecutive nucleotides irrespective of codon position in coding DNA and “codon” in the usual sense to mean an in-frame triplet that encodes an amino acid.

It is important to note that each of the singlet, doublet, and triplet mutation processes is assumed to apply to every singlet, doublet, and triplet site, respectively, of the sequences studied, which means that the doublet and triplet processes can span codons. For example, if at a given time a sequence is TGGACC, then the singlet process applies equally to the six sites ($i.e.$, to the nucleotides T, G, C, A, C, and C), the doublet process applies to each of the five doublets (TG, GC, CA, AC, and CC), and the triplet process applies to the four triplets (TGC, GCA, CAC, and ACC). Our implementation (below) combining these models ignores any “edge effects” caused by finite sequence length ($i.e.$, for a sequence of length $N$ codons there are $3N$ singlets, $3N - 1$ doublets, and $3N - 2$ triplets) and is generated as though it applied to an infinitely long sequence. We also ignore the effect of intron and exon structure in a gene, where particular codons may not be adjacent in the genomic sequence. This is justified partly by the fact that protein sequences do not evolve in isolation, but as part of larger genomic regions that will have similar stationary distributions of nucleotide content to the gene. This ignores the effect of variation in selection at the edges of genes and introns; such effects are assumed to be small and their inclusion would greatly increase the complexity of the model.

**Standardizing the relative rates of point, doublet, and triplet events:** To parameterize and estimate the relative frequencies of point, doublet, and triplet events each process needs to be scaled to have comparable units. We scale each process so that time is in units of the expected number of mutational events per codon (point, doublet, or triplet) per codon. To achieve this, we need to know the equilibrium distribution of codons in the SDT model. This is denoted $f_{i_{m_1}m_2}$ for each codon $m_1m_2m_3$. In practice, we find that $f_{i_{m_1}m_2} = \pi_{m_1}\pi_{m_2}\pi_{m_3}/(1 - \pi_{stop})$ for all sense codons, where for the universal code $\pi_{stop} = \pi_T\pi_A\pi_A + \pi_T\pi_A\pi_C + \pi_T\pi_C\pi_A$, and that for the three stop codons is $f_{\text{stop}} = f_{\text{stop}} = f_{\text{stop}} = 0$. Note that $\pi_{stop}$ would be the expected frequency of stop codons if these were not prevented from arising by the parts of the model describing the effects of selection acting on sequence evolution. However, until these values for $f_{i_{m_1}m_2}$ are verified (see below) it is convenient to retain the $f_{i_{m_1}m_2}$ notation.

Recalling that our single-nucleotide mutation events occur independently at each nucleotide site, the mean rate of single-nucleotide mutation events per codon in a sequence at equilibrium with each codon $i_{i1}i_2i_3$ occurring with frequency $f_{i_{m_1}m_2}$ is given by

$$\rho_i = \sum_{i_{m_1}m_2} f_{i_{m_1}m_2} (s^*_{i_{m_1}} + s^*_{i_{m_2}} + s^*_{i_{m_3}}),$$

(4)

where the replacement of a subscript with a + symbol indicates summation over all possible values of that subscript (so, for example, $s^*_{i_{m_1}} = s^*_{i_{m_1}A} + s^*_{i_{m_1}C} + s^*_{i_{m_1}G} + s^*_{i_{m_1}T}$). Then, to arrive at a model giving rise to one single-nucleotide mutation event per unit time, we use the following for the instantaneous rate of substitution of nucleotide $i$ with $j$ at each site of a sequence:

$$s_{ij} = \frac{s^*_i}{\rho_i}.$$  

(5)

We use a similar approach to scale time in the doublet and triplet processes to make the units the expected numbers of mutational events per codon (doublet or triplet changes). Again recalling that the mean rates must be calculated at the equilibrium state of the SDT model that combines the single, doublet, and triplet mutation processes, this is achieved by the following rescaling,

$$d^*_{i_{i1}i_2} = \frac{d^*_{i_{i1}i_2}}{\rho_i}$$

(6)

and

$$t^*_{i_{i1}i_2} = \frac{t^*_{i_{i1}i_2}}{\rho_i},$$

(7)

where
\[ p_d = \sum_{\text{codons}} f_{n_1 n_2 n_3} (d_{m_1 m_2,++} + d_{m_2 m_3,++} + \sum_{n_1 n_2 n_3} f_{n_1 n_2 n_3} d_{m_1 m_2 m_3,++}) \]

and

\[ p_t = \sum_{\text{codons}} f_{n_1 n_2 n_3} (f_{m_1 m_2,++} + \sum_{n_1 n_2 n_3} f_{n_1 n_2 n_3} (f_{m_1 m_2,++} + f_{m_1 m_3,++})). \]

Note the more complicated forms of Equations 8 and 9 relative to Equation 4. This is due to having to consider doublet and triplet events that can span two consecutive codons, for example, the pair \((m_1 m_2 m_3, n_1 n_2 n_3)\) arising consecutively with frequency \(f_{n_1 n_2 n_3} f_{n_1 n_2 n_3}\). Our approach for describing the frequency of two adjacent codons simply as the product of their individual occurrence makes the assumption that codons appear independently at neighboring sites (see below).

Finally we introduce parameters \(\sigma, \delta, \tau, \text{ and } \tau\) that describe the relative frequencies of point, doublet, and triplet mutation events, respectively. If we make the constraint that \(\sigma + \delta + \tau = 1\) then our SDT model, described by the mutation process consisting of the concurrent functioning of the singlet, doublet, and triplet processes defined by the values \(\sigma_{s,j}, \delta_{d_{i,j},j}, \text{ and } \tau_{t_{i,j}/t_{i,j}}\), also has a mean rate of one mutation event per unit time and has these different types of event occurring with frequencies in the ratio \(\sigma:\delta:\tau\).

**Formulation of the SDT model:** In existing codon models, only singlet mutation events may occur. Each event thus affects exactly one codon, and a process of evolution with independent codon sites is easily constructed. The mutation processes described above, however, allow for the possibility of single events that affect two neighboring codons. This means that a perfect implementation of these processes into a codon model would require us to analyze codon positions while making allowance for the actual neighboring codons at every point in evolutionary time. While these are known in the sequences studied (i.e., those at the tips of the phylogeny relating the sequences in an alignment), they are not known at intermediate evolutionary times. It is no longer possible to analyze the sites of a sequence alignment as though they are independent realizations of some evolutionary process on a phylogeny. The state space of the Markov process that should be used in effect becomes the space of all possible sequences, and for a codon model for a sequence of length \(N\) codons this space contains \(61^N\) possible sequences. This is clearly too large for analysis by straightforward methods, although complex statistical techniques have been used to make some progress in such cases (Pedersen and Jensen 2001; Arndt et al. 2003; Robinson et al. 2003; Siepel and Haussler 2004).

Instead, to maintain independence of sites, we choose to consider only the average effect of neighboring codons when constructing the evolutionary model from the three mutational processes described above. Some events are straightforward to deal with: all singlet events, doublet events affecting codon positions 1 and 2, or 2 and 3, and triplet events affecting codon positions 1, 2, and 3 of one codon each generate one change in a single codon. The remaining events, namely doublet and triplet events that span a codon boundary, each generate two changes, one in each of two codons. Since we wish to describe the net effect on independent codons, we cannot assume we know the precise nature of each of these two changes. In this case, we treat the first affected codon as known, so that the effect of the event on that codon is fully specified. (Exactly the same model would arise if we treated the second affected codon as known.) The affected neighboring codon will have some nucleotides specified (since the mutation event is known) and is then assumed to be drawn randomly according to expectations for individual codons over evolutionary time, conditional on the presence of the known nucleotides. With our modeling assumptions of independent codon sites and codon model equilibrium, the frequencies for each possible neighboring codon \(m_1 m_2 m_3\) are precisely proportional to the \(f_{n_1 n_2 n_3}\).

For example, consider the doublet mutation event AT → GG falling on the third position of the codon CGA and the first position of the neighboring codon. This generates two codon changes within our codon model. In the first codon, we can identify the change CGA → CGG. However, we can identify only the second codon change as TWZ → GWZ, for unknown nucleotides W and Z. Our codon model is constructed so that this doublet event generates all the changes TWZ → GWZ (i.e., all possibilities for W and Z), but with each occurring with relative frequency given by \(f_{\text{TWZ}}\). This leads to actual frequencies \(f_{\text{GWZ}}/f_{\text{TWZ}}\); these sum to one, which is appropriate to ensure that this doublet event has generated two codon changes when codons are observed independently.

More generally, the doublet change X,Y → X,Y falling on the third codon position of the triplet X,X,X and the first codon position of the neighboring triplet generates one codon change, X,X,X → X,X,X, and each of the changes Y,WZ → Y,WZ, for all nucleotides W and Z, with frequency \(f_{\text{Y,WZ}}/f_{\text{Y,WZ}}\). The exactly analogous approach is used for triplet events; so, for example, the triplet change CGG → GAT falling on the third codon position of triplet TAC and the first and second codon positions of the neighboring triplet causes one codon change TAC → TAG and causes each of the changes GGW → ATW with frequency \(f_{\text{G,G,W}}/f_{\text{G,G,W}}\).

**Selection process:** The final components of our model to be introduced are parameters to describe the effect of selection on the sequence changes described by our mutation processes. We follow an approach similar to that first used by Goldman and Yang (1994) and Muse and Gaut (1994) to introduce parameters describing the relative probability of fixation in a popu-
lation (species) of different types of change. Following those authors and the later work of Yang and co-workers (e.g., Nielsen and Yang 1998; Yang et al. 2000; Yang and Nielsen 2002), we assign different probabilities of acceptance of a mutation event depending on whether there are any consequent changes to the encoded amino acid sequence. However, whereas those authors had to consider only the possibility of single synonymous and single nonsynonymous changes, we have to consider the possibility of individual mutation events that may give rise to zero, one, or two nonsynonymous changes. (In fact, an exhaustive list of changes that can arise from our three types of mutation events is 0 nonsynonymous changes (N) + 1 synonymous change (S); 0N + 2S; 1N + 0S; 1N + 1S; 2N + 0S.) In a simple extension of earlier approaches, we use the following parameterization to describe the effect of selection on mutation events:

\[
\text{Pr (mutation events being accepted)} = \begin{cases} 
0 & \text{if any stop codons arise} \\
p_0 & \text{if all resulting substitutions are synonymous} \\
p_{N1} & \text{if one nonsynonymous substitution results} \\
p_{N2} & \text{if two nonsynonymous substitutions result.} 
\end{cases}
\]

(10)

In previous models only \(p_0\) and \(p_{N2}\) (and the probability 0 for stop codons) have been considered and are expressed as relative rates of 1 and \(\omega = p_{N1}/p_0\) for synonymous and nonsynonymous substitutions, respectively. Similarly, we work in terms of the relative rates of 1, \(\omega_1 = p_{N1}/p_0\) and \(\omega_2 = p_{N2}/p_0\). As has been widely discussed, values of \(\omega < 1\) indicate an excess of synonymous substitutions, suggesting negative selection and conservative evolution, and values of \(\omega > 1\) indicate an excess of nonsynonymous substitutions and suggest positive selection (e.g., Nielsen and Yang 1998; Yang and Bieławski 2000). The same interpretations may be placed on our selection parameters: \(\omega_0\) is directly analogous to \(\omega\) in standard models of codon evolution, and \(\omega_1\) describes the average selective pressure on two amino acids changing simultaneously. Inference under models that allow \(\omega\) values to vary between sites in the alignment has proven useful in detecting more specifically the effects of selection (for example, Yang and Bieławski 2000; Yang et al. 2000; Swanson et al. 2003), but would be computationally difficult to include in our model.

**The SDT model:** Finally, the above ideas may be combined into a Markov process model for codon evolution that permits the modeling of mutation events that affect more than one nucleotide site and that still permits the computationally convenient analysis of sequence alignments under the assumption that all codon sites evolved independently. Our codon model is embodied in a matrix \(Q\) of instantaneous replacement rates, where the element \(Q_{ij\ell;i\ell\ell}\) gives that rate of replacement of codon \(i_\ell\ell\ell\) with codon \(j_\ell\ell\ell\) when codon \(i_\ell\ell\ell\) is initially present. \(Q\) can be broken into the contributions from the three mutation processes described above, denoted \(Q^{(s)}, Q^{(d)},\) and \(Q^{(t)}\) for the singlet, doublet, and triplet processes, respectively. The full methodology for calculating these individual rate matrices is provided in the Appendix, and they are combined to generate our SDT model according to

\[
Q = Q^{(s)} + Q^{(d)} + Q^{(t)}.
\]

(11)

We are now able to verify that the SDT model has the expected equilibrium distribution \(f = (f_{ij\ell})\), as defined in terms of the elements of the parameter \(\pi\) and discussed above, by confirming that all elements of the vector \(Q\pi = 0\). We also note that the SDT model is time reversible (see, e.g., Liò and Goldman 1998), as it satisfies \(f_{ij\ell\ell}Q_{ij\ell;i\ell\ell} = f_{ij\ell\ell}Q_{ij\ell;i\ell\ell}\) for all codons \(i_\ell\ell\ell\) and \(j_\ell\ell\ell\).

So far as we are aware, no other models of codon sequence evolution have been proposed that make allowance for individual mutation events that affect more than one nucleotide. Our model is perhaps closest to the codon model denoted M0 by Yang et al. (2000) and first described by Nielsen and Yang (1998) as a simplified version of the model of Goldman and Yang (1994). Our new SDT model is almost identical to M0 when the restrictions \(\delta = \tau = 0\) are made; i.e., only singlet mutations can occur (\(\sigma = 1\)). In this case, the only difference is that M0 assumes that the rate of change to a codon \(j_\ell\ell\ell\) is proportional to the frequency with which that codon \((j_\ell\ell\ell)\) is observed, whereas the SDT model, in common with the model of Muse and Gaut (1994), assumes it is proportional to the frequency of the altered nucleotide \((j_\ell, j_\ell, or j_\ell)\). We prefer the latter assumption, since we know of no mechanism by which a mutational process might observe and be affected by the codon that a nucleotide lies in. We see this as a different issue from the fact that the occurrence of mutations will be affected by neighboring bases; this is undoubtedly the case, but would not be expected to be related in any way to a nucleotide’s position in the coding structure of protein-coding DNA.

**Likelihood implementation of the SDT model:** We implemented the SDT model for maximum-likelihood analysis in purpose-written software using standard methods for Markov process models of sequence evolution, described (for example) by Goldman and Yang (1994), Swoford et al. (1996), and Liò and Goldman (1998). Phylogenetic trees were chosen as described below, and maximum likelihood was used to estimate branch lengths and the parameters \(\kappa, \pi, \sigma, \delta, \tau, \omega_1,\) and \(\omega_2\). Gaps in the sequence alignment are considered as missing data during likelihood calculations (see, e.g., Yang 1997). Maximum-likelihood estimates under the SDT model can occasionally prove difficult to find using standard numerical optimization procedures. To avoid issues regarding bad convergence and local optima all
analyses were run from multiple starting values and results were accepted only when several starting points converged on the same maximum likelihood. Likelihood-ratio tests (LRTs; see Yang et al. 1994; Swofford et al. 1996) were used for model comparisons, for example, to test whether the SDT model gives a significantly better fit to data than a version constrained to have no doublet and triplet mutation events ($\sigma = 1$, $\delta = \tau = 0$). Precise details of the LRTs used are given below.

**Additional factors of interest:** A primary focus of this article is to see whether or not a model that allows for mutation events affecting multiple nucleotides gives an improved description of the evolution of protein-coding sequences, as assessed by LRTs of competing models that do and do not make this allowance. Assuming, as is shown below, that the new SDT model is overwhelmingly successful, it is then of interest to examine the extent to which doublet and triplet mutation events are inferred to occur. At the level of mutation, this is described in the SDT model by the parameters $\sigma$, $\delta$, and $\tau$. It is also interesting to consider the relative proportions of different mutation events occurring that are accepted by selection and the resulting numbers of different mutation events that actually contribute to observed sequence evolution (i.e., after the effects of selection are taken into account). These characteristics may all be computed from the inferred parameter estimates in the SDT model.

We denote by $M_{ij}$ the number of mutation events per codon site per unit time affecting $i$ nucleotides (i.e., $i = 1$, 2, and 3 indicating singlet, doublet, and triplet events, respectively), the first of the affected nucleotides being at codon position $j$ ($j = 1$, 2, 3). For example, $M_{32}$ is the number of triplet events affecting positions 2 and 3 of one codon and position 1 of the following codon. The events described by $M_{ij}$ are the inferred mutations that have occurred prior to selection, and we denote by $A_{ij}$ the corresponding number of events inferred to have been accepted by selection. Formulae for the computation of the $M_{ij}$ and $A_{ij}$ are given in the appendix. Note the contrasting interpretations of the $M_{ij}$ and $A_{ij}$ which relate to actual mutational events in our model, and the events described directly by the elements of the matrix $Q$, which are the observable outcomes of those events. To use an earlier example, a triplet mutation CGG $\rightarrow$ GAT occurring at position 3 of a codon TAC and positions 1 and 2 of the following codon contributes to $M_{32}$ and, according to its probability of being accepted by selection, to $A_{32}$. However, the observable outcomes are five codon changes: the singlet change TAC $\rightarrow$ TAG and the doublet changes GGW $\rightarrow$ ATW (with probability $f_{GW}/f_{ATW}$) for each of $W = A, C, G, T$, thus contributing to $Q_{TM,TAG}$ and to $Q_{GG,ATW}$ for each of $W = A, C, G, T$.

Because of the way our model is constructed $M_{i+} = \sigma$, $M_{2+} = \delta$, $M_{3+} = \tau$, and $M_{i+,+} = 1$. The total rate of accepted mutation events per site per unit time is given by $A_{i+,+}$; this will in general be different from the mean rate at equilibrium of $Q$, which indicates the rate of observable events. The relationship of these rates is explained further in the appendix.

**Sequence data and trees:** For our initial study a set of six mammalian globin sequences from human (NM_000558), rhesus monkey (J04495), cow (AJ242798), goat (J00043), Gravy’s zebra (U70191.1), and rabbit (M11113) were extracted from GenBank (accession numbers in parentheses) and aligned using ClustalX (Thompson et al. 1997) to give an alignment of 143 codons after initiation and stop codons were removed. This data set is presented as a typical example for which we can present detailed results and is ideal for studying doublet and triplet mutation because the sequences are closely related and require no insertions/deletions (indels) to produce the alignment; in our model described below, ambiguity in the alignment, which is more likely to occur in more distantly related sequences, may potentially influence estimates of the rate of doublet and triplet mutation. A phylogenetic tree for these sequences was estimated using the REV + $\Gamma$ model and using an exhaustive tree search procedure, as implemented in PAML (Yang 1994b, 1997) with the $\Gamma$-distribution discretized into four categories (Yang 1994a). The resulting topology is (((human, rhesus monkey), ((cow, goat), zebra), rabbit), which we believe agrees with the consensus opinion of the relationship between these mammals. Similar tree estimation procedures were performed using various other models of DNA evolution and no variation in the topology estimate was observed. Both the tree and sequence alignment are available from the authors on request.

In addition, a large number of coding sequence alignments were taken from the nucleotide section of the Pandit 7.6 database (Whelan et al. 2003; available at http://www.ebi.ac.uk/goldman-srv/pandit/), which is derived from the Pfam-A seed database of protein domain alignments (Bateman et al. 2004). In Pandit, nucleotide sequences for each alignment are extracted from the EMBL database (Kulikova et al. 2004) and aligned using the Pfam-A seed alignment as a template. The manual curation of these families, the quality and length of the alignments, and the levels of sequence divergence make them representative of the type of data used to perform phylogenetic inference (Whelan et al. 2003). Codon models are computationally very slow, necessitating that only a subset of Pandit was analyzed; only families with four, five, or six sequences were used. Families were constrained to have alignment lengths of >100 codons and so that the sequences all use the universal genetic code, to give a lower limit to the information present from which to estimate parameters and ensure that sequences evolve in a comparable manner, respectively. We also required that there must be no variation in tree estimate under several standard models of DNA evolution for the families used. The effect of
using an inaccurate tree topology has an unknown influence on parameter estimation under our new model, so ideally the ML trees for each codon model and data set combination should be used. Using families where only a single evolutionary tree topology is estimated under various DNA models should reduce any potential effects because tree topology has proven to be relatively stable when adding complexity to a model. To further ensure data quality we make the final constraint of ensuring that branch length estimates under the null model do not tend to extremely long values, which may occur when there are so many synonymous changes that they have become saturated. When this occurs the number of inferred synonymous changes may be severely underestimated and may potentially affect the accuracy of other parameter estimates. These constraints should ensure that the data used were of high quality and resulted in 258 families (out of a possible 4699) being used in subsequent analyses. These represent a total of 75,177 aligned codon positions, consisting of 1200 different protein-coding sequences containing in total 358,837 codons.

RESULTS

Mammalian globin sequences: Four variants of the SDT model are used to examine the evolution of the globin sequences: (a) $\sigma = 1$ and $\delta = \tau = 0$, representing a null model where no large scale events occur, (b) $\delta = 1 - \sigma$ and $\tau = 0$, a model where only point and doublet events can occur, (c) $\tau = 1 - \sigma$ and $\delta = 0$, a model where only point and triplet events can occur, and (d) the full SDT model (any $\sigma, \delta, \tau \geq 0$ satisfying $\sigma + \delta + \tau = 1$).

Relevant parameter estimates, the maximum likelihood under the tree considered, and the relative numbers of degrees of freedom of these variant models are shown in Table 1. LRTs may be performed between the models when they are nested by comparing $2\Delta$, twice the improvement in likelihood observed when progressing from the simpler (null) model to the more complex (alternative) model, to a $\chi^2$-distribution, where the $n$ degrees of freedom are the number of parameters by which the null and alternate models differ. When the parameters $\delta$ and/or $\tau$ are added, the use of the $\chi^2$-distribution is likely to be a conservative test because they can both rest on boundaries of the parameter space under null models (Self and Liang 1987). In addition, $\omega_0$ effectively does not exist under the null model (a), and any value it takes will result in exactly the same likelihood. In cases such as this $\omega_0$ is said to be unidentifiable under the null model; regularity conditions are violated and the standard theory justifying a $\chi^2$-distribution for significance testing may break down (for a pathological example see Chen and Chen 2001). Obtaining appropriate asymptotic distributions for significance testing in such cases is difficult (Davies 1977, 1987; Andrews 2001) and obtaining data-specific distributions using simulation approaches (e.g., Goldman 1993) is too computationally expensive for general use in our model. For a typical set of parameter values we used a parametric bootstrap approach (Goldman 1993) to estimate the distribution of $2\Delta$ for comparing models a and d from 200 simulations. The results (not shown) suggest that the $\chi^2$-distribution is still conservative, with the 95% points of the simulation and $\chi^2$-distribution being 3.88 and 7.82, respectively. Examination of parameter estimates from the simulation shows that both $\delta$ and $\tau$ are distributed near zero, confirming that when no doublet or triplet events occur their parameter estimates are reasonable. In the following discussion we assume the use of $\chi^2$-distributions is appropriate for significance testing.

Statistical testing: The comparison of model variants (a) and (d) represents a general test of the significance of doublet and triplet events and shows that the description of the evolutionary process provided by the SDT model is significantly better than that provided by the simpler null model that allows no large-scale events ($2\Delta = 68.26; P < 0.001$). Assessing the relative importance of doublet and triplet events on the evolution of the globin sequences is less clear because the presence of doublet and triplet events can each be tested by two model comparisons and the results are somewhat contradictory. Using a LRT to assess the importance of doublet events by comparing model variants (a) and (b), neither of which consider triplet events, yields a highly significant result ($2\Delta = 56.54; P < 0.001$). In a similar test performed in the presence of a parameter describing triplet events, i.e., the comparison of (c) and (d), the addition of doublet mutation does not significantly improve the model ($2\Delta = 0.43; P = 0.51$). A similar observation is made for triplet events. When comparing model variants without a parameter describing doublet events a much higher increase in likelihood is achieved [comparison of (a) and (c)]: $2\Delta = 67.83; P < 0.001$] than when comparing model variants that also contain a parameter describing doublet mutation [comparison of (b) and (d): $2\Delta = 11.72; P < 0.001$], although the addition of $\tau$ (triplet events) is highly significant in both these comparisons.

This behavior is probably due to the ability of the parameters describing doublet events to partially describe triplet events, and vice versa. This is illustrated by examining the test for the significance of doublet mutation given by comparing (c) and (d). The relatively small value of $2\Delta$ obtained by progressing to the SDT model is accompanied by quite a large estimate of the frequency of doublet mutation and a corresponding reduction in the estimate of the frequency of triplet mutation. It appears reasonable to assume that the statistical test comparing the null model (a) and the SDT model (d) can ably distinguish between single-nucleotide events and larger-scale events. The parameter esti-
TABLE 1
Parameter estimates from mammalian globin sequences

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value under model with</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a. ( \sigma = 1, \delta = \tau = 0 )</td>
</tr>
<tr>
<td>d.f.(^a)</td>
<td>0</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>1 (0.53/1)(^b)</td>
</tr>
<tr>
<td>( \delta )</td>
<td>0 (—/0)(^c)</td>
</tr>
<tr>
<td>( \tau )</td>
<td>0 (—/0)(^d)</td>
</tr>
<tr>
<td>( \omega_1 )</td>
<td>0.36</td>
</tr>
<tr>
<td>( \omega_2 )</td>
<td>—</td>
</tr>
<tr>
<td>( \kappa )</td>
<td>1.39</td>
</tr>
<tr>
<td>( \pi_A )</td>
<td>0.18</td>
</tr>
<tr>
<td>( \pi_C )</td>
<td>0.36</td>
</tr>
<tr>
<td>( \pi_G )</td>
<td>0.28</td>
</tr>
<tr>
<td>( \pi_T )</td>
<td>0.18</td>
</tr>
<tr>
<td>Tree length(^e)</td>
<td>2.39</td>
</tr>
<tr>
<td>( \ln L )</td>
<td>—1301.615</td>
</tr>
</tbody>
</table>

\(^a\) Degrees of freedom by which model variant varies from model a.
\(^b\) First value in parentheses is the proportion of singlet events accepted by selection, i.e., \( A_{i+}/M_{i+} = A_{i+}/\sigma \); second value is the proportion of all accepted events that are singlets, i.e., \( A_{i+}/A_{i++} \); see MATERIALS AND METHODS section and APPENDIX for details of computation.
\(^c\) First value in parentheses is the proportion of doublet events accepted by selection, i.e., \( A_{i+}/M_{i+} = A_{i+}/\delta \); second value is the proportion of all accepted events that are doublets, i.e., \( A_{i+}/A_{i++} \).
\(^d\) First value in parentheses is the proportion of triplet events accepted by selection, i.e., \( A_{i+}/M_{i+} = A_{i+}/\tau \); second value is the proportion of all accepted events that are triplets, i.e., \( A_{i+}/A_{i++} \).
\(^e\) Tree length is the sum of all branch lengths in a topology.

Parameter estimates: In the SDT model the estimates of the frequency of doublet and triplet mutation events (prior to selection; Table 1, \( \delta \) and \( \tau \), respectively) and accepted events (after selection; \( A_{i+}/A_{i++} \) and \( A_{i+}/A_{i++} \), respectively) are both surprisingly high and may indicate that both types of mutation played an important role in the evolution of the mammalian globin genes. The mutation event numbers indicate how frequently doublet and triplet mutations are estimated to occur and suggest that in the globin alignment approximately one-third of all mutations affect two or three nucleotides simultaneously (\( \delta + \tau = 0.34 \)). The numbers regarding events accepted by selection suggest that only a limited proportion of these doublet and triplet events pass through the filter of selection and impact the evolutionary history of the proteins. The relative fractions of the different types of mutation accepted by selection reflect the genetic code. More point mutations than either doublet (\( A_{i+}/\delta = 22\% \)) or triplet (\( A_{i+}/\tau = 16\% \)) mutations are accepted by selection (\( A_{i+}/\sigma = 44\% \)) because some of the point mutations at the first or third base in a codon are synonymous and occur at the same rate as neutral mutation. Fewer triplet mutations than doublet mutations survive selection because there are more synonymous doublet mutations than triplet mutations; doublet mutation may occur at the first and third sites of a codon where it is possible for synonymous mutations to occur, whereas triplet mutations always affect the second position of a codon and consequently always result in a nonsynonymous change.

It is interesting to assess the effect of the inclusion of doublet and triplet events in a model on the other parameter estimates our model contains (see Table 1). It is immediately noticeable that the inclusion of doublet and triplet events in the model has a negligible effect on the equilibrium distribution parameters contained in \( \pi \). This is reassuring, because the expected frequency of the nucleotide bases in the observed sequences is defined by \( \pi \) in a similar manner for all of the models. The inclusion of doublet and triplet mutations in the model appears to have a substantial effect on the remaining parameters. The transition/transversion ratio, \( \kappa \), appears to increase under SDT and the selection parameter, \( \omega_i \), decreases. The cause of the change in \( \kappa \) is unclear, while the decrease in \( \omega_i \) may be expected: if doublet and triplet events have occurred during the evolution of the sequence, under the null model multiple nonsynonymous point events are required to describe many of the doublet and triplet events and \( \omega_i \) needs to be artificially high to describe them; but when the larger-scale events are better accounted for (by the SDT model) the value of \( \omega_i \) is closer to its true level. For example, consider the triplet mutation that changes isoleucine (ATA) to alanine (GCC). Under the null model (a), at least two nonsynonymous events would...
be required to describe the change and ω2 would be large to make these events more likely. In the SDT model the change may be correctly described as a triplet event and ω1 is unaffected.

The value of ω2 is also of interest because no comparable parameter has previously been estimated. In the mammalian globin alignment there appears to be a preference for nonsynonymous changes to cause only a single-amino-acid residue in a protein to change rather than affecting adjacent pairs of residues, i.e., ω1 > ω2 in Table 1, particularly for the full SDT model. The final factor of interest is the effect the inclusion of doublet and triplet events has on tree length, which appears to slightly increase under the SDT model. The ratio of branch lengths between the two models also differs (i.e., differences are not a simple linear scaling of all branch lengths; results not shown), suggesting that the SDT model may affect tree estimation procedures. The cause for the increase in tree length is unclear but may be related to the increase in κ, with the SDT model inferring more synonymous transition mutations than the null model.

**Database analysis:** Application of the SDT model to a large number of alignments provides a more general assessment of the relative importance of different types of mutational event. The parameter estimates from our database analysis confirm that the conclusions of the globin analysis apply across a broader spectrum of biology, indicating that doublet and triplet events occur on a regular basis.

**Statistical testing:** The comparison between the null model (a) and the SDT model (d) is used to investigate the doublet and triplet mutation for our database analysis. The submodels (b) and (c) are not used because, as discussed above, the parameters δ and τ have a tendency to partially describe each other, making the result from these models difficult to interpret. Statistical tests show that progression from the null model to the SDT model is significant at the 95% level in 257 (99.6%) of the families examined and the mean (median) increase in likelihood was 117.79 (90.23), indicating that the SDT model provides a substantial improvement in describing evolution and is likely to provide biologically interesting information.

**Parameter estimates:** The distribution across families of the frequency of doublet and triplet events prior to selection (δ and τ, respectively) is shown in Figure 1A, with the proportion of doublet and triplet mutation events taking mean (median) values of 0.22 (0.16) and 0.28 (0.25), respectively. Figure 1B shows the distribution of the proportions of single, doublet, and triplet mutations that are accepted by selection (Aδδ/σ, Aδτ/δ, and Aττ/τ, respectively). Figure 1C indicates the proportions of all accepted events that derive from doublet and triplet mutations (Aδδ/Aδδ and Aδτ/Aδτ, respectively); the mean (median) values are 0.091 (0.041) and 0.097 (0.068), respectively. Note that the estimates of the proportion of doublet and triplet mutations from the globin analysis are consistent with the estimates made from the database. The variability observed in all of the distributions in Figure 1 should be viewed with some caution because the standard errors of the parameters that contribute to them are different in each family due to differences in the length of alignments, tree lengths, etc.

General effects on other parameters’ estimates of including doublet and triplet events in an evolutionary model are illustrated in Figure 2, A and B. Figure 2A shows the effect of including doublet and triplet events on estimates of κ, which tends to be greater under the SDT model than under the null model. Figure 2B shows the distribution of selection parameters under the null and SDT models. The estimates of ω2 tend to be much lower under the SDT model than under the null model; even cases where pervasive positive selection is implied under the null model (cases where ω1 > 1) revert to much lower levels under the SDT model. The parameter ω2 of the SDT model tends to take substantially higher values than ω2 takes in the SDT model, but slightly lower values than ω1 takes in the null model, and there is a weak correlation between ω2 and ω1 of the null model: high estimates of ω1 in the null model tend to be associated with high estimates of ω2. This unexpected general (database) finding differs from the globin result. We expected values of ω2 to be lower than ω1 (SDT), reflecting a presumed greater physicochemical disruption caused by changing two amino acids relative to a single-amino-acid change. The high values of ω2 may reflect compensatory amino acid replacements (such context effects are not considered by the SDT model) or the model describing larger-scale events than doublets and triplets (see below).

**DISCUSSION**

For the surprisingly high estimates of doublet and triplet mutation rates to be believable some support from the biological literature is required. Unlike doublet events (see Discussion in Averof et al. 2000), there is no explicit mechanism known to us that causes exactly three adjacent nucleotides to change simultaneously. It therefore seems reasonable to consider the possibility that events spanning larger stretches of DNA are the source of these high estimates. Figure 3 demonstrates how the products of gene conversion, small-scale recombination, and inversion mutations, if plausible for the data at hand, may appear as high estimates of doublet and triplet events in the SDT model. The apparent ability of a parameter describing doublet events to partially describe triplet events, and vice versa, as noted in the analysis of the globin sequences, supports this hypothesis.

We find the potential of gene conversion to be the source of our high estimates of doublet and triplet events particularly intriguing because many recent studies have suggested that gene conversion occurs more often than
Figure 1.—Distribution of parameter estimates and statistics relating to singlet, doublet, and triplet mutations on a per family basis from a database of 258 aligned protein families. (A) The proportions of doublet (shaded bars) and triplet (open bars) mutations before selection (δ and τ, respectively). (B) The proportions of singlet (solid bars), doublet (shaded bars), and triplet (open bars) mutations accepted by selection. (C) The proportions of all events accepted by selection that derive from doublet (shaded bars) and triplet (open bars) mutations. Estimates under the SDT model from the globin sequence alignment are also indicated.
Figure 2.—Effects of including parameters describing doublet and triplet events on other parameter estimates in the SDT model, on a per family basis from a database of 258 aligned protein families. (A) The change in $\kappa$ between the null and SDT models; note that 8 families with $\sigma < 0.05$ under SDT are not included because they do not contain enough information to estimate $\kappa$ accurately. (B) The distribution of $\omega_1$ under the null (solid bars) and SDT (shaded bars) models and the distribution of $\omega_2$ under the SDT model (open bars); note that two families with $\sigma > 0.95$ are not included because they do not contain enough information to estimate $\omega_2$ accurately.

Previously thought. Laboratory-based studies have found significant gene conversion on the human sex chromosomes (Skaletsky et al. 2003) and in the malaria parasite (Nielsen et al. 2005). Another study, examining human sperm, found high levels of gene conversion in meiotic crossover hot spots (Jeffreys and May 2004); between 80 and 94% of recombination events at these hot spots were observed to resolve as gene conversions and the mean length of the conversion tract was estimated to be in the range of 55–290 bp. Two computational studies have estimated the length and frequency of gene conversion events in gene families of Saccharomyces cerevisiae (Semple and Wolfe 1999) and Caenorhabditis elegans (Drouin 2002). The length and variance of the observed gene conversions are quite large but broadly agree with the lengths proposed by Jeffreys and May (2004); the shape of the distributions and the inherent difficulties in uncovering short gene conversion events suggest that relatively short segments of DNA may be altered by gene conversion (Semple and Wolfe 1999, Figure 6; Drouin 2002, Figure 2). The estimated frequency at which gene conversions occur in these studies, however, is probably too low to describe the high frequency of doublet and triplet events inferred by our model. This may be because only known, active homologs of the gene families are considered in the analyses, whereas gene conversion events have also been documented where a section of a pseudogene sequence replaces part of a functional gene. For example, in the human condition 21-hydroxylase deficiency, the most common form of adrenal hyperplasia, gene conversion is thought to account for $\sim 75\%$ of mutant alleles (Col-
Figure 3.—Possible mechanisms by which large-scale events may result in elevated estimates of doublet and triplet mutation. (A) Gene conversion. The left-hand side shows how three independent mutations may occur in a sequence; these may be correctly inferred as three point mutations (indicated by dashed boxes). The right-hand side demonstrates that a gene conversion event occurring as indicated makes these three changes appear simultaneously in the right-hand lineage, and one point and one doublet mutation will be inferred. Note that small-scale recombination will result in a similar inference except the exchange of information will be reciprocal. (B) Sequence inversion (illustrated with double-stranded DNA). The observed strand of DNA (in boldface type) appears to have evolved via a single triplet event as a consequence of this sequence inversion.

lier et al. 1993; Tusie-Luna and White 1995). Gene conversion from pseudogenes is also thought to be important in polycystic kidney disease (Watnik et al. 1998). This introduces the possibility that gene families convert not only within themselves but also with other areas of the genome with high sequence similarity. If this were occurring it may raise the frequency of gene conversion to levels that may substantially describe our high estimates of doublet and triplet mutation.

An alternative explanation for the high estimates of doublet and triplet events would be if compensatory mutations were regularly occurring. For example, if a change at the first position of a codon were accepted by selection only if other changes rapidly occurred at the second and third position of the preceding codon, this would be well described by triplet mutation in our model. A phenomenon similar to this has been modeled in RNA evolution where matching bases in regions with stem structures are described as changing simultaneously (e.g., Savill et al. 2001). If compensatory events that are accepted by selection arise quickly relative to the frequency of occurrence of the mutations that “trigger” them and of events that do not lead to any compensatory changes, then a model such as ours that treats the compensating mutations as arising simultaneously with the triggering events seems adequate: it captures the general behavior of sequence evolution, at least at the level observable with species-level data, and does so without having to model a complex pattern of heterogeneous selective pressures that vary over sequence position and over evolutionary time. We do, however, have to remember to be cautious with the interpretation of our inferred levels of instantaneous doublet and triplet events.

Another possibility is that high estimates of doublet and triplet events may be due to properties of the sequence data used. Possible problems arising from errors in tree topology or the misalignment of homologous sequences may potentially influence the estimates of the proportion of doublet and triplet events. However, results obtained by examining subsets of the data formed by removing the most divergent families, which can be expected to contain those families with the most ambiguous alignments, and by including only families with long alignments of five or more sequences do not substantially differ from the results presented above, strongly suggesting that the estimated high levels of doublet and triplet events are not artifacts of the data (results not shown). This is further supported by the detailed analysis of the globin family, where estimates typical of the results obtained from the database are
obtained from an alignment containing only closely related sequences with no ambiguity or indels in the alignment. The possibility that errors in tree estimation may be the cause of the high estimates of large-scale events cannot be discounted, but many measures have been employed to ensure that tree topology estimates are as good as possible. It is unlikely that the topology estimates for all 258 families studied are wrong, which would indicate that the inference of frequent large-scale events is correct in at least some families.

The SDT model, in common with many other models used in phylogenetics, makes many assumptions about the evolutionary process and it is possible that violations of these assumptions are the source of the high estimates of doublet and triplet mutation. The assumption of no variation in evolutionary factors (e.g., selection) along a protein may conceivably lead to overestimation of the rate of doublet and triplet mutation. If there is variation in selection in a relatively conserved protein then a site undergoing positive selection (with \( \omega > 1 \)) would have more accepted point mutations at the first and second codon position than at other codons. Consequently, in the SDT model, which takes into account only the average amount of selection occurring, this excess of changes may potentially be misinterpreted as doublet and triplet events. Another plausible contributor to the high estimates of doublet and triplet mutation would be a significant degree of rate variation in the mutational process acting on the DNA, which could be caused, for example, by the presence of mutational hot spots. However, the levels of doublet and triplet mutation events estimated from the data are very high and occur over a wide variety of biological sequences. Given the plausible biological explanation of such events, it seems unlikely that misspecifications of the model, such as those discussed above, are the sole cause of the high estimates of doublet and triplet mutation.

The misspecification argument can also be turned on its head to support the SDT model: if large-scale mutational events are occurring during evolution then estimates from studies that do not describe these events may be affected. For example, Anisimova et al. (2003) present a simulation study that demonstrates that recombination (a larger-scale event) may result in overestimation of selection parameters in standard codon models. If large-scale events are as frequent as our study suggests it may go some way to explaining the extreme level of positive selection observed by some studies; the proportion of amino acid changes driven by positive selection in primates and Drosophila has been estimated as 35% (Fay et al. 2001) and 45% (Smith and Eyre-Walker 2002), respectively. The results of this study also show that other parameter estimates may be affected by not allowing for large-scale events. The SDT model, while not completely describing events such as recombination and gene conversion, does represent a step toward being able to do so and consequently parameter estimates from it may reflect reality more accurately than those of simpler models. We hope in the future that approaches based on the formulation of the SDT model, combined with recent advances in modeling context dependence (e.g., Pedersen and Jensen 2001; Arndt et al. 2003; Siepel and Haussler 2004), may be used to more completely describe such larger-scale events in the evolutionary process.

CONCLUSION

In this study we introduce a new codon model of evolution that can estimate the frequency of singlet, doublet, and triplet mutation using an approach that allows events to span multiple codon sites while maintaining the simplifying independent-sites assumption for data analysis. The model is constructed in a manner that separates the process of mutation on the DNA sequence from the process of selection acting upon the protein. Inclusion of doublet and triplet mutations gives a statistically significant improvement in the fit of the model to data in the vast majority of 258 protein-coding sequence alignments studied, indicating that these mutation events do occur and are an important feature of genome evolution. Parameter estimates from the model suggest that doublet and triplet mutations occur far more frequently than previously thought and give some indication of the variations that may be observed over typical protein-coding DNA sequences, which reflect variations over genomic regions.

The exact cause of these high estimates, however, remains open to debate. It seems unlikely that misspecification of the model caused, for example, by high levels of variation in selection along proteins or poor data quality can be the sole cause of the high estimates of doublet and triplet mutation. The lack of a biological mechanism for triplet mutation, the more prevalent of the two types of large-scale mutation inferred, leads us to conclude that these high estimates may more likely be reflecting compensatory change and larger-scale mutation, probably in the form of gene conversion, inversion mutation, or small-scale recombination. If such events are regularly occurring during evolution, this may be a worrying departure from the independence-of-sites assumption and affect parameter estimates under current models. In these circumstances, the SDT model should provide parameter estimates that better reflect reality.

We thank Adam Siepel and David Haussler for letting us see a prepublication version of their work and two anonymous referees for comments leading to improvements to this manuscript. S.W. is supported by the Wellcome Trust. N.G. is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Science.

LITERATURE CITED


Anisimova, M., R. Nielsen and Z. Yang, 2003 Effect of recombina-
tion on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164: 1229–1236.


Davies, R. B., 1977 Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64: 247–254.

Davies, R. B., 1987 Hypothesis testing when a parameter is present only under the alternative. Biometrika 74: 33–43.


Communicating editor: Z. Yang
APPENDIX

Calculation of $Q^{(s)}$, $Q^{(d)}$, $Q^{(t)}$, and $Q$: Before calculating the constituent rate matrices of the SDT model it is convenient to define the following functions:

$$
\Delta^{(1)}(i, j) = \begin{cases} 
0 & \text{if } j \text{ is a stop codon} \\
1 & \text{if the change } i \rightarrow j \text{ is synonymous} \\
\omega_1 & \text{if the change } i \rightarrow j \text{ is nonsynonymous} 
\end{cases} 
$$

$$
\Delta^{(2)}(i, j) = \begin{cases} 
0 & \text{if either } j \text{ or } l_l \text{ are stop codons} \\
1 & \text{if changes } i \rightarrow j \text{ and } k_l \text{ represent one synonymous change} \\
\omega_2 & \text{if both } i \rightarrow j \text{ and } k_l \text{ are nonsynonymous changes} 
\end{cases} 
$$

(A1)

(A2)

to represent the probabilities of mutation events being accepted as functions of the changes between codons $i$ and $j$, and codon pairs $i, k_l$, respectively.

Of these rate matrices $Q^{(s)}$ is the simplest to describe, since all singlet events affect only one codon:

$$
Q^{(s)}_{i\rightarrow j, h} = \begin{cases} 
\Delta^{(1)}(i, j)\sigma_{n_{j=1}} & \text{if } j \neq j, i_2 = j, i_3 = j \\
\Delta^{(1)}(i, j)\sigma_{n_{j=2}} & \text{if } i_1 = j, i_2 \neq j, i_3 = j \\
\Delta^{(1)}(i, j)\sigma_{n_{j=3}} & \text{if } i_1 = j, i_2 = j, i_3 \neq j \\
0 & \text{otherwise.} 
\end{cases} 
$$

(A3)

$Q^{(d)}$ has to allow for doublet mutation events that alter only one codon (those falling on codon positions 1 and 2 or 2 and 3) and that alter two codons (falling on codon positions 3 and 1). The first two of these are straightforward to accommodate, leading to the following:

$$
Q^{(d)}_{i\rightarrow j, h} = \begin{cases} 
\Delta^{(1)}(i, j)\delta_{j, i_2, i_1} & \text{if } i_1 \neq j, i_2 \neq j, i_3 = j \\
\Delta^{(1)}(i, j)\delta_{j, i_2, i_3} & \text{if } i_1 = j, i_2 \neq j, i_3 \neq j \\
0 & \text{otherwise.} 
\end{cases} 
$$

(A4)

To complete $Q^{(d)}$, we have to consider all possible neighboring codons $i, k_l$, and all possible doublet mutation events that can alter them at codon positions 3 and 1—these are the changes $i, k_l \rightarrow j, l_l$ for all $j_1 \neq i_3$ and $l_1 \neq k_1$, giving rise to codons $i, j_1, k_1, l_1$. Recalling our assumption that codon $k_l$ follows a given instance of codon $i, j_1$ with probability $f_{i, j_1}$ and, similarly, $i, j_3$ precedes $k_1, l_1$ with probability $f_{i, j_3}$, $Q^{(d)}$ may be completed using the following iterative procedure,

for all codons $i, j, k_l$ {

for all (following) codons $k_1, l_1$ {

for all (replacement) doublets $j_1, l_1$ with $j_1 \neq i_3$ and $l_1 \neq k_1$ {

$$
Q^{(d)}_{i\rightarrow j, h} = \begin{cases} 
\Delta^{(1)}(i, j)\delta_{j, i_2, i_3} & \text{if } i_1 \neq j, i_2 \neq j, i_3 = j \\
\Delta^{(1)}(i, j)\delta_{j, i_2, i_3} & \text{if } i_1 = j, i_2 \neq j, i_3 \neq j \\
0 & \text{otherwise.} 
\end{cases} 
$$

(A5)

where the $+$ symbol indicates that in each iteration of the loops, the indicated values are cumulatively added to the values of $Q^{(d)}_{i\rightarrow j, h}$ and $Q^{(d)}_{i\rightarrow j, h}$ already computed by Equation A4.

Following the same logic, $Q^{(t)}$ can be produced by first calculating

$$
Q^{(t)}_{i\rightarrow j, h} = \begin{cases} 
\Delta^{(1)}(i, j)\tau_{j, i_2, i_1} & \text{if } i_1 \neq j, i_2 \neq j, i_3 = j \\
0 & \text{otherwise} 
\end{cases} 
$$

(A6)

and completed using the following iterative procedure:
for all codons \(i_t i_j i_k\) {

for all (following) codons \(k_t k_j k_l\) {

for all (replacement) triplets \(j_t j_j j_l\) with \(j_t \neq i_t, j_j \neq i_j, and j_l \neq i_l\) and \(l_t \neq k_t\)

\[
Q_{ij_i2j_2_j3}^{(1)} = f_{i_t k_t j_t} \Delta^{(2)}(i_t i_j i_k) k_t k_j k_l, i_t j_t j_l l_t k_t) \tau_{t_j k_t j_l k_t}\n\]

\[
Q_{ij_i2j_2_j3}^{(2)} = f_{i_t k_t j_t} \Delta^{(2)}(i_t i_j i_k) k_t k_j k_l, i_t j_t j_l l_t k_t) \tau_{t_j k_t j_l k_t}\n\]

\} \} \}.

(A7)

The matrix \(Q\) actually used in our model is now simply calculated as

\[
Q = Q^{(1)} + Q^{(2)} + Q^{(3)}
\]

(A8)

and, as is usual for Markov process models, it is convenient to define the diagonal elements of \(Q\) according to

\[
Q_{i_t i_t i_t i_t} = -\sum_{k_t k_j k_l} Q_{i_t k_t j_t k_l b}\n\]

(A9)

Calculation of the quantities \(M_{i_j}\) and \(A_{i_j}\): \(M_{i_j}\) and \(A_{i_j}\), as defined in the text, are computed using the following formulae. For singlet events,

\[
M_{i_x} = \sum_{i_t i_j i_k} \sum_{t_j i_k} \sigma s_{i_2 t} f_{i_t i_j} = \sigma \sum_{i_t i_j i_k} f_{i_t i_j} s_{i_2 t}
\]

(A10)

\[
A_{i_x} = \sum_{i_t i_j i_k} \sum_{t_j i_k} \Delta^{(1)}(i_t i_j i_k) [i_2 i_3 i_4] \sigma s_{i_2 t} f_{i_t i_j}
\]

\[
= \sigma \sum_{i_t i_j i_k} f_{i_t i_j} \sum_{i_t i_j i_k} \Delta^{(1)}(i_t i_j i_k) [i_2 i_3 i_4] s_{i_2 t}
\]

(A11)

for each \(x = 1, 2, 3\), where \([i_2 i_3 i_4]\) indicates the codon arising when codon \(i_t i_j i_k\) has nucleotide \(i_t\) at position \(x\) changed to \(j_t\) (so, for example, \([i_2 i_2 i_4]\) represents the codon \(j_t i_j i_k\) and \([i_1 i_2 i_4]\) represents \(i_t j_t i_k\)). For doublet events,

\[
M_{2_x} = \delta \sum_{i_t i_j i_k} f_{i_t i_j} = d_{2_x} + + +
\]

(A12)

\[
A_{2_x} = \delta \sum_{i_t i_j i_k} f_{i_t i_j} \sum_{i_t i_j i_k} \Delta^{(1)}(i_t i_j i_k) [i_2 i_3 i_4] s_{i_2 t} d_{2_x} + + +
\]

(A13)

for \(x = 1, 2, 3\); and

\[
M_{2_5} = \delta \sum_{i_t i_j i_k} f_{i_t i_j} \sum_{i_t i_j i_k} f_{i_t i_j} d_{2_5} + + +
\]

(A14)

\[
A_{2_5} = \delta \sum_{i_t i_j i_k} f_{i_t i_j} \sum_{i_t i_j i_k} \sum_{i_t i_j i_k} \Delta^{(2)}(i_t i_j i_k) k_t k_j k_l, i_t j_t j_l l_t k_t) d_{2_5} + + +
\]

(A15)

For triplet events,

\[
M_{3_1} = \tau \sum_{i_t i_j i_k} f_{i_t i_j} l_{i_2 i_3} + + +
\]

(A16)

\[
M_{3_2} = \tau \sum_{i_t i_j i_k} f_{i_t i_j} l_{i_2 i_3} + + +
\]

(A17)
\[ M_{3,5} = \tau \sum_{\text{codons}} f_{i,j|k,l} \sum_{\text{codons}} f_{k,l|j,i} t_{j,i|k,l} \]  
(A18)

\[ A_{3,1} = \tau \sum_{\text{codons}} f_{i,j|k,l} \sum_{j \neq l}^{\Delta^{(1)}} (i_1 i_2 i_3, j_1 j_2 j_3) t_{i,j|k,l} \] 
(A19)

\[ A_{3,2} = \tau \sum_{\text{codons}} f_{i,j|k,l} \sum_{j \neq l}^{\Delta^{(2)}} (i_1 i_2 i_3, k_1 k_2 k_3, i_4 i_5 i_6) t_{i,j|k,l} \] 
(A20)

\[ A_{3,3} = \tau \sum_{\text{codons}} f_{i,j|k,l} \sum_{j \neq l}^{\Delta^{(3)}} (i_1 i_2 i_3, k_1 k_2 k_3, i_4 i_5 i_6, l_1 l_2 l_3) t_{i,j|k,l} \] 
(A21)

As mentioned in the text, \( M_{1,4} = \sigma \), \( M_{2,4} = \delta \), \( M_{4,4} = \tau \), and \( M_{4,4} = 1 \). Although all singlet, doublet, and triplet mutation events arise independently of codon position (as seems reasonable for a model of sequence evolution), they are not independent of the affected nucleotides in each singlet, doublet, or triplet. Different singlets, doublets, and triplets have different rates of evolution (as is evident in Equations 1, 2, and 3), and the same singlet, doublet, or triplet will occur with different frequencies at different codon positions because of the effects of the selection parameters \( \omega_1 \) and \( \omega_2 \) and because stop codons are not allowed to arise. Consequently, the following relationships do not hold exactly:

\[ M_{1,1} = M_{1,2} = M_{1,3} = \sigma / 3 \]
\[ M_{2,1} = M_{2,2} = M_{2,3} = \delta / 3 \]
\[ M_{3,1} = M_{3,2} = M_{3,3} = \tau / 3 \]

However, we note that for reasonable parameter values these relationships remain approximately true.

Since each “accepted mutation” \( A_{1,1} \), \( A_{1,2} \), \( A_{1,3} \), \( A_{2,1} \), \( A_{2,2} \), and \( A_{3,1} \) leads to one observed codon change, and each of \( A_{2,3} \), \( A_{3,2} \), \( A_{3,3} \) leads to exactly two observed codon changes, we can also relate the mean rate of observable codon changes at equilibrium with the values \( A_i \) as follows:

\[- \sum_{i,j|k,l} f_{i,j|k,l} q_{i,j|k,l} = A_{1,1} + A_{1,2} + A_{1,3} + A_{2,1} + A_{2,2} + A_{3,1} + 2(A_{2,3} + A_{3,2} + A_{3,3}). \]  
(A22)