J Mol Evol (2011) 72:147–159
DOI 10.1007/s00239-010-9409-0

Regional Context in the Alignment of Biological Sequence Pairs

Raymond Sammut · Gavin Huttley

Received: 30 July 2010 / Accepted: 8 November 2010 / Published online: 24 November 2010

Abstract Sequence divergence derives from either point substitution or indel (insertion or deletion) processes. We investigated the rates of these two processes both in protein and non-protein coding DNA. We aligned sequence pairs using two pair-hidden Markov models (PHMMs) conjoined by one silent state. The two PHMMs had their own set of parameters to model rates in their respective regions. The aim was to test the hypothesis that the indel mutation rate mimics the point mutation rate. That is, indels are found less often in conserved regions (slow point substitution rate) and more often in non-conserved regions (fast point substitution rate). Both polypeptides and rRNA molecules in our data exhibited a clear distinction between slow and fast rates of the two processes. These two rates served as surrogates to conserved and non-conserved secondary structure components, respectively. With polypeptides we found both the fast indel rate and the fast replacement rate were co-located with hydrophilic residues. We also found that the average concordance, of our alignments with corresponding curated alignments, improves markedly when the model allows either of the two fast rates to colocate with hydrophilic residues. With rRNA molecules, our model did not detect colocation between the fast indel rate and the fast substitution rate. Nevertheless, coupling the indel rates with the point substitution rates across the two regions markedly increased model fit. This result suggests that rRNA pairwise alignments should be modeled after allowing for the two processes to vary simultaneously and independently in the two regions.

Keywords Indel · Polypeptide · rRNA · Pairwise alignment · PHMM · Colocation · Concordance · Hydrophilic residue

Abbreviations

rRNA	Ribosomal RNA
HMM	Hidden Markov model
PHMM	Pair-hidden Markov model
PIP	Phylogenetically independent sequence pair
PMB	Probability matrix from blocks
Replacement	The change of one amino acid in one sequence to another amino acid in the other sequence at a site of a pairwise alignment of two biological sequences
N/C-terminal	The N-terminus is the start of the polypeptide terminated by an amino acid with a free amine group (–NH_2), and the C-terminus is the end of a polypeptide terminated by an amino acid with a free carboxyl group (–COOH), by convention, a peptide sequence is written from N-terminus on the left hand side to the C-terminus on the right

Introduction

The accuracy of biological sequence alignment has important implications when making inferences about

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9409-0) contains supplementary material, which is available to authorized users.

R. Sammut · G. Huttley (✉)
Department of Genome Biology, John Curtin School of Medical Research, Building 54, The Australian National University, Canberra, ACT 0200, Australia
e-mail: gavin.huttley@anu.edu.au

R. Sammut
e-mail: mundu.sammut@gmail.com

function. This accuracy depends on our understanding of natural selection and rate heterogeneity of mutation processes. One of these processes consists of microstructural changes that take the form of insertions and deletions (indels) of nucleotides (hence amino acids) along biological sequences. When we align biological sequences, we represent indels by gaps. How precisely we position these gaps along each sequence in the alignment is one of the most important factors that lead to correct inferences.

There is evidence to suggest that, in protein coding DNA, indels are more abundant in hydrophilic regions (Pascarella and Argos 1992; Taylor et al. 2004; de la Chaux et al. 2007; Zhang et al. 2010). In non-protein coding DNA it is shown that indels are less prevalent in regions containing inverted repeats (Yamane et al. 2006). This evidence seems to suggest that indels are found less often in regions that are conserved, both in protein coding and non-protein coding DNA. Yamane et al. (2006) also found that the rate of nucleotide substitutions was relatively lower in the inverted repeats regions. We hypothesize that the rate of replacement (or substitution) is related to the rate of indel placement; that is, both rates are lower (higher) in conserved (non-conserved) regions of the DNA irrespective of coding type.

Early pairwise alignment methods assumed mutation rates to be singly uniform across all positions of the pairwise alignment (Needleman and Wunsch 1970; Gotoh 1982; Bishop and Thompson 1986; Thorne et al. 1991). It is generally agreed, however, that these methods are not biologically realistic. For example, although indels of just one position in the alignment are the most common, indels longer than one position occur biologically at relatively high percentages both in protein coding DNA (Pascarella and Argos 1992) and in non-protein coding DNA (Krawczak and Cooper 1991). Long indel models for evolutionary pairwise alignments, employing the geometric distribution to deal with random indel lengths, have been developed (Knudsen and Miyamoto 2003; Miklós et al. 2004). These models provide biologically useful information on the alignment in the form of reliability measures based on posterior probabilities. However, they do not contribute to any significant improvement in alignment accuracy since they do not exploit information obtained from putative properties of DNA. For example, clearly defined parts of secondary structure and regional replacement rates are known to be correlated in protein coding DNA (Goldman et al. 1996). A pairwise aligner designed to incorporate regional context derived from conserved and non-conserved regions of the DNA primary structure can be used to test the hypothesis of a positive relationship between the point and indel substitution rates. The regions can be imputed during the alignment procedure by exploiting putative signatures of secondary structure elements such as α-helices and β-sheets in protein coding DNA or, inverted repeats in non-protein coding DNA.

Implicit knowledge on secondary structure elements is exploited heuristically in Clustal-W (Thompson et al. 1994) to dynamically vary gap penalties during multiple alignments. For example, a patch of five contiguous hydrophilic amino acids was demonstrated by Pascarella and Argos (1992) to be indicative of a loop-region in the protein. In Clustal-W, this information triggers a reduction in the gap opening penalty during alignment of the current sequence within this region. Pascarella and Argos (1992) also found that, on average, gaps would not be longer than eight positions. On this basis, Clustal-W increases the gap opening penalty in the current sequence of the multiple alignment within eight residues (that is, columns) of existing gaps. These empirically based techniques greatly increase the sensitivity of Clustal-W which remains a commonly employed benchmark for the testing of new multiple alignment methods.

Here we evaluate the relationship between point and indel substitution rates by extending the model proposed by Knudsen and Miyamoto (2003). That is, we employ two pair-hidden Markov models (PHMMs) instead of one, and conjoin the two by a single silent state. This topology allows one PHMM to model the replacement (or substitution) rate, the indel rate, and the indel length distribution in the slow evolving region. The other PHMM models these evolutionary processes in the fast evolving region. Hence, each of the two PHMMs provides its own set of sufficient statistics independently of the other following the completion of the alignment of the two input sequences. In this setting, we aim to achieve better inference on the placement of indels when the replacement (or substitution) rate is not averaged across the entire length of the alignment. The length of indels is also modeled in the two regions under a separate geometric distribution with an average parameter estimated for each region.

Recent studies (Taylor et al. 2004; Löytynoja and Goldman 2008; Sjödin et al. 2010) have shown that rates of deletions and of insertions can have different determinants, and therefore should be modeled separately. However, our aim in this work was to investigate the association between the two processes, namely, point and indel substitutions, in two regions of DNA (conserved and non-conserved). We consider it reasonable, therefore, that the insertion and deletion rates are averaged with one parameter within each region, while differentiation is only analyzed between the two regions. Furthermore, averaging within regions also reduces computational cost.

Our modeling of amino acid replacements (or point substitutions) and of indel placements in the two regions is impelled by the notion that the slow and fast rates of these two components serve as surrogates for conserved and non-conserved regions, respectively. We use phylogenetically independent sequence pairs (PIPs) over a wide range of evolutionary distances to show that this surrogacy is generally applicable. We show further that this surrogacy also holds true both for protein coding and non-protein coding DNA by constructing separate PIP samples from the BAliBASE protein database (Thompson et al. 2005) and from the European ribosomal RNA (rRNA) database (Wuyts et al. 2004), respectively. Using these two separate samples we confirm that hydrophilic regions are associated with high rates of amino acid replacements and of indel placements coexisting spatially away from the core. We further demonstrate that processes operating on rRNA are distinctive.

Materials and Methods

The HMM–PHMM Topology

The model of Thorne et al. (1992) allowed for heterogeneity in the point substitution rate, but also imposed constraints. The latter included identical fragment size distributions, and identical indel processes, between the slow and fast regions. We eliminate both of these constraints in our modeling. Fundamental to our approach are two separate parameters to model the substitution rates in the two regions. Furthermore, to evaluate the relationship between point and indel substitution rates, we apply the same approach to the indel rates. Thus, for each of the two regions, we also employ an indel rate parameter and a corresponding parameter for the indel length probability distribution. In this setting, we would expect that rates in one category will be estimated by the model below the baseline mutation rate (slow rate region), where mutations are putatively more likely to affect function. Rates in the other category will be estimated above the baseline (fast rate region), where mutations are putatively less critical to function.

Motivated by these putative processes, we designed the two-region PHMM–HMM topology to model the pairwise alignment. This topology is two-tiered, where the lower layer consists of the two PHMMs shown in Fig. 1 and the upper layer consists of a two-state hidden Markov model (HMM) shown in Fig. 2. Each of the two PHMMs models substitutions together with indels in only one region. At each position of the pairwise alignment, the two-state HMM assigns (by convention) either PHMM$_1$ which models the slow rates in region one or PHMM$_2$ which models the fast rates in region two, to emit a symbol. By "rates", here, we mean the slow and fast substitution rates which may result, for instance, from two different intensities of natural selection. Also, by "symbol" we mean either a match (w, z), or a delete (w, -), or an insert (-, z), where $w \in \xi$ is a character in sequence S_W, $z \in \xi$ is a character in sequence S_Z, (-) is a gap, and ξ is the alphabet of the biological sequences.

The pair-hidden Markov model (PHMM) was formally introduced by Durbin et al. (1998) to address the issue of

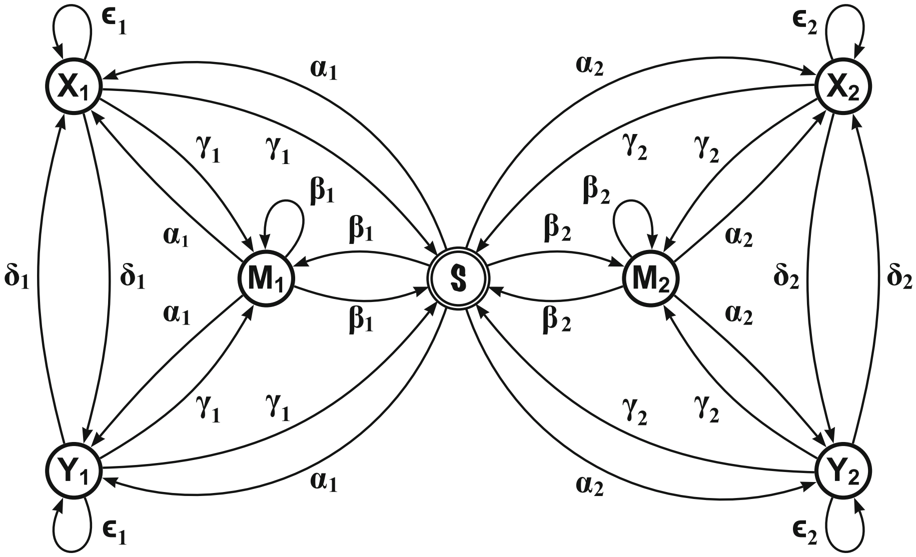
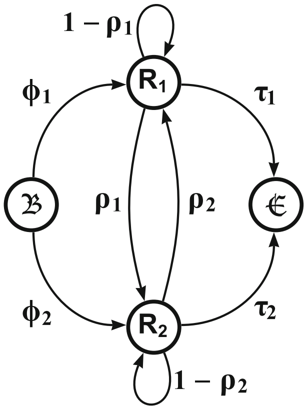


Fig. 1 Each PHMM has match state M, and has insert states X and Y for sequences S_X and S_Y, respectively. Index η, $\eta \in \{1, 2\}$, associates states $\{M_1, X_1, Y_1\}$ of PHMM$_1$ to region one and states $\{M_2, X_2, Y_2\}$ of PHMM$_2$ to region two. Silent state $\mathcal{S}$ conjoins the two PHMMs. *Arrows* and associated parameters show directional flows and transition probabilities, respectively. For example, when in state X_1, transition flow may loop back to state X_1 with probability ε_1 or reach state M_1 with probability γ_1 or reach state Y_1 with probability δ_1 or reach state $\mathcal{S}$ with probability γ_1. Where appropriate, transition probabilities are set equal to simplify the model. For example, probability of flow looping back to M_1 is set equal to probability of flows between M_1 and $\mathcal{S}$, thus economizing on number of parameters



	Transitions			Emissions	
	R_1	R_2	$\mathfrak{E}$	*PHMM*$_1$	*PHMM*$_2$
$\mathfrak{B}$	ϕ_1	ϕ_2	0	0	0
R_1	$1-\rho_1$	ρ_1	τ_1	1	0
R_2	ρ_2	$1-\rho_2$	τ_2	0	1

Fig. 2 Conceptual two-state HMM. States R_1 and R_2 emit symbols that emanate from PHMM$_1$ and PHMM$_2$, respectively. Switching parameters ρ_1 and ρ_2 determine whether R_1 or R_2 emits at each position of the alignment. $\mathcal{B}$ and $\mathcal{E}$ are *begin* and *end* states, respectively, and do not emit symbols. Stationarity parameters ϕ_1 and ϕ_2 determine initial states of R_1 and R_2. Terminating parameters τ_1 and τ_2 compensate for input sequence pair, S_X and S_Y, having finite and unequal lengths. First matrix is the state matrix and second matrix is the space matrix of the 2-state Markov process

indels in probability modeling of pairwise alignments. The formulation by these authors allows for the estimation from data of a parameter vector with five elements, and which we denote by (α, β, δ, ε, γ). These elements are shown in Fig. 1 for each PHMM in the two-region topology. Knudsen and Miyamoto (2003) derived analytically the set of equations that reduce these five elements to just three sufficient parameters, namely, the average replacement (or substitution) rate parameter t, $t > 0$, the average indel rate parameter r, $r > 0$, and the indel length probability distribution parameter a, $0 < a < 1$. We term these equations the Knudsen–Miyamoto (KM) equations, and we present them in Supplementary Material for our two-region topology where parameters in each of the two PHMMs are indexed by η, $\eta \in \{1, 2\}$; that is, η is the index of the PHMM and its corresponding region.

Figure 3 shows the conceptual matrix of transition probabilities of the lower layer of the two-tiered HMM–PHMM topology. The upper layer captures the alternating behavior of rate heterogeneity. We assume this alternating behavior to be a two-state Markov process. This process has a 2 × 2 transition matrix shown conceptually in Fig. 2 with transition probabilities ρ_1, $0 < \rho_1 < 1$, and ρ_2, $0 < \rho_2 < 1$. These *switching* probabilities determine the flow intensity in the current PHMM before they switch flow to the other PHMM of the two-region topology via the *silent* state $\mathcal{S}$ in Fig. 1. Churchill (1989) showed that under stable DNA compositional heterogeneity, a switching probability would typically be small. By extension to the substitution rate problem, a low switching probability means that we expect protein (or DNA) sections, alternately experiencing low and high rates of replacements (substitutions) along the pairwise alignment, not to be fragmented.

The *begin* state $\mathcal{B}$ in Fig. 2 plays a role only during the first step of the alignment procedure. The starting probability from $\mathcal{B}$ to the PHMMs is multiplied by stationary probabilities ϕ_η, where $\eta \in \{1, 2\}$ is the region index, to average out the initial uncertainty. Similarly, the *end* state $\mathcal{E}$ in Fig. 2 plays a role only during the last step. The last probability from each of the two PHMMs to $\mathcal{E}$ is multiplied by the sequence length parameter τ_η, $\eta \in \{1, 2\}$, in order to take into account the fact that the alignment has finite length.

The Transition Matrix

A transition from one state to another state within the same PHMM$_\eta$, $\eta \in \{1,\ 2\}$, of our two-tiered HMM–PHMM topology, takes place with the same probability as that computed by the set of KM equations belonging to that PHMM, except that we multiply this probability by $1 - \rho_\eta$ in our modeling. A transition from one state of PHMM$_\eta$ to another state of PHMM$_{3-\eta}$ takes place with a probability computed from the two sets of KM equations and the probability ρ_η. For example, a transition from state M$_1$ to state Y$_2$ would be the product of β_1, α_2, and ρ_1. Figure 4 shows all the possible probability transformations that produce our two-region transition matrix from the conceptual matrices shown in Figs. 2 and 3. The new matrix is implemented in a standard Forward algorithm (Rabiner 1989) after each row has been normalized. Note that the new matrix in Fig. 4 restores the *begin* state $\mathcal{B}$ and the *end* state $\mathcal{E}$. That is, the silent state shown in Fig. 3 was only part of the conceptual matrix, and does not need to be implemented explicitly in our modeling following the transformations.

$$\mathbf{T} = \begin{array}{c|ccccccc|}
 & \mathbf{M_1} & \mathbf{X_1} & \mathbf{Y_1} & \mathbf{M_2} & \mathbf{X_2} & \mathbf{Y_2} & \mathfrak{S} \\
\mathfrak{S} & \beta_1 & \alpha_1 & \alpha_1 & \beta_2 & \alpha_2 & \alpha_2 & 0 \\
\mathbf{M_1} & \beta_1 & \alpha_1 & \alpha_1 & 0 & 0 & 0 & \beta_1 \\
\mathbf{X_1} & \gamma_1 & \epsilon_1 & \delta_1 & 0 & 0 & 0 & \gamma_1 \\
\mathbf{Y_1} & \gamma_1 & \delta_1 & \epsilon_1 & 0 & 0 & 0 & \gamma_1 \\
\mathbf{M_2} & 0 & 0 & 0 & \beta_2 & \alpha_2 & \alpha_2 & \beta_2 \\
\mathbf{X_2} & 0 & 0 & 0 & \gamma_2 & \epsilon_2 & \delta_2 & \gamma_2 \\
\mathbf{Y_2} & 0 & 0 & 0 & \gamma_2 & \delta_2 & \epsilon_2 & \gamma_2
\end{array}$$

Fig. 3 Conceptual two-region transition matrix T of HMM–PHMM topology constructed from two 3 × 3 transition matrices of the two PHMMs. Silent state $\mathcal{S}$ acts as *begin* state of source PHMM through first row and as *end* state of sink PHMM through last column, simultaneously

The Emission Matrices

Emission probabilities, and corresponding symbols, of the HMM–PHMM topology are stored in a vector **E** (shown in Fig. 5) of matrices indexed by η, $\eta \in \{1, 2\}$. Elements in each matrix are probabilities constructed in accordance with Gonnet and Benner (1996).

$$E_{\mathrm{M}} = [P(t)\mathrm{W}]'\left[\left(\boldsymbol{q} \otimes \mathbf{1}_n'\right) \bullet I_n\right][P(t)\mathrm{Z}] \quad (1)$$

The upper-left quadrant of a matrix in **E**, denoted by E_{M}, stores the emission probabilities of all character match symbols (w, z); w, $z \in \xi$, of the PIP sequences S_{W} and S_{Z}. The distribution of these probabilities is computed using Eq. 1, where ($\otimes$) and ($\bullet$) are the Kronecker and Hadamard products, respectively. The index n is equal to the number of characters in the alphabet ξ of the biological sequences; that is, n is 20 or 4 for protein or DNA sequences, respectively. W and Z are character emission matrices of sequences S_{W} and S_{Z}, respectively, and consist of 0's and 1's. If the length of sequence S_{W} is denoted by ℓ_{W}, then W is a $n \times \ell_{\mathrm{W}}$ matrix. Also, for example, if the first character in the alphabet is A, then row one of matrix W has 1's in those positions corresponding to A's in sequence S_{W}, and 0's in all other positions; and likewise for the remaining $n - 1$ rows. The same applies to matrix Z and the corresponding sequence S_{Z} of length ℓ_{Z}. The stochastic vector $\boldsymbol{q}$ stores the n background probabilities, and is computed from sequences S_{W} and S_{Z} as described by Felsenstein (1981). $\mathbf{1}_n$ is simply a vector of n 1's. $P(t)$ is a standard $n \times n$ matrix of probabilities for substitution of amino acids (nucleotides) through time t. Its derivation is described in detail by Goldman (1993). Following multiplication, Eq. 1 produces a $\ell_{\mathrm{W}} \times \ell_{\mathrm{Z}}$ submatrix whose elements describe the probability distribution of all possible match symbols.

The upper-right quadrant of a matrix in **E**, denoted by E_{X}, stores the emission probabilities of all character-gap delete symbols (w, -); $w \in \xi$, of the PIP sequence S_{W}.

$$\mathbf{T} = \begin{array}{c|ccccccc|}
 & \mathbf{M}_1 & \mathbf{X}_1 & \mathbf{Y}_1 & \mathbf{M}_2 & \mathbf{X}_2 & \mathbf{Y}_2 & \mathfrak{E} \\
\mathfrak{B} & \beta_1\phi_1 & \alpha_1\phi_1 & \alpha_1\phi_1 & \beta_2\phi_2 & \alpha_2\phi_2 & \alpha_2\phi_2 & 0 \\
\mathbf{M}_1 & \beta_1(1-\rho_1) & \alpha_1(1-\rho_1) & \alpha_1(1-\rho_1) & \beta_1\beta_2\rho_1 & \beta_1\alpha_2\rho_1 & \beta_1\alpha_2\rho_1 & \beta_1\tau_1 \\
\mathbf{X}_1 & \gamma_1(1-\rho_1) & \epsilon_1(1-\rho_1) & \delta_1(1-\rho_1) & \gamma_1\beta_2\rho_1 & \gamma_1\alpha_2\rho_1 & \gamma_1\alpha_2\rho_1 & \gamma_1\tau_1 \\
\mathbf{Y}_1 & \gamma_1(1-\rho_1) & \delta_1(1-\rho_1) & \epsilon_1(1-\rho_1) & \gamma_1\beta_2\rho_1 & \gamma_1\alpha_2\rho_1 & \gamma_1\alpha_2\rho_1 & \gamma_1\tau_1 \\
\mathbf{M}_2 & \beta_2\beta_1\rho_2 & \beta_2\alpha_1\rho_2 & \beta_2\alpha_1\rho_2 & \beta_2(1-\rho_2) & \alpha_2(1-\rho_2) & \alpha_2(1-\rho_2) & \beta_2\tau_2 \\
\mathbf{X}_2 & \gamma_2\beta_1\rho_2 & \gamma_2\alpha_1\rho_2 & \gamma_2\alpha_1\rho_2 & \gamma_2(1-\rho_2) & \epsilon_2(1-\rho_2) & \delta_2(1-\rho_2) & \gamma_2\tau_2 \\
\mathbf{Y}_2 & \gamma_2\beta_1\rho_2 & \gamma_2\alpha_1\rho_2 & \gamma_2\alpha_1\rho_2 & \gamma_2(1-\rho_2) & \delta_2(1-\rho_2) & \epsilon_2(1-\rho_2) & \gamma_2\tau_2
\end{array}$$

Fig. 4 Implementation of two-region transition matrix T. Silent state $\mathcal{S}$ is replaced by the *begin* state $\mathcal{B}$ in first row and by *end* state $\mathcal{E}$ in last column. Each row is normalized to make T row stochastic

$$\mathbf{E} = \begin{array}{c|cccccc:c|}
[\mathrm{E_M}]_\eta & \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} & \mathtt{G} & \mathtt{T} & [\mathrm{E_X}]_\eta \\
\mathtt{C} & \mathbf{E}_{\eta,\mathtt{CA}} & \mathbf{E}_{\eta,\mathtt{CG}} & \mathbf{E}_{\eta,\mathtt{CT}} & \mathbf{E}_{\eta,\mathtt{CC}} & \mathbf{E}_{\eta,\mathtt{CG}} & \mathbf{E}_{\eta,\mathtt{CT}} & \mathbf{E}_{\eta,\mathtt{C}-} \\
\mathtt{T} & \mathbf{E}_{\eta,\mathtt{TA}} & \mathbf{E}_{\eta,\mathtt{TG}} & \mathbf{E}_{\eta,\mathtt{TT}} & \mathbf{E}_{\eta,\mathtt{TC}} & \mathbf{E}_{\eta,\mathtt{TG}} & \mathbf{E}_{\eta,\mathtt{TT}} & \mathbf{E}_{\eta,\mathtt{T}-} \\
\mathtt{C} & \mathbf{E}_{\eta,\mathtt{CA}} & \mathbf{E}_{\eta,\mathtt{CG}} & \mathbf{E}_{\eta,\mathtt{CT}} & \mathbf{E}_{\eta,\mathtt{CC}} & \mathbf{E}_{\eta,\mathtt{CG}} & \mathbf{E}_{\eta,\mathtt{CT}} & \mathbf{E}_{\eta,\mathtt{C}-} \\
\mathtt{G} & \mathbf{E}_{\eta,\mathtt{GA}} & \mathbf{E}_{\eta,\mathtt{GG}} & \mathbf{E}_{\eta,\mathtt{GT}} & \mathbf{E}_{\eta,\mathtt{GC}} & \mathbf{E}_{\eta,\mathtt{GG}} & \mathbf{E}_{\eta,\mathtt{GT}} & \mathbf{E}_{\eta,\mathtt{G}-} \\
\mathtt{A} & \mathbf{E}_{\eta,\mathtt{AA}} & \mathbf{E}_{\eta,\mathtt{AG}} & \mathbf{E}_{\eta,\mathtt{AT}} & \mathbf{E}_{\eta,\mathtt{AC}} & \mathbf{E}_{\eta,\mathtt{AG}} & \mathbf{E}_{\eta,\mathtt{AT}} & \mathbf{E}_{\eta,\mathtt{A}-} \\
\hdashline
[\mathrm{E_Y}]_\eta & \mathbf{E}_{\eta,-\mathtt{A}} & \mathbf{E}_{\eta,-\mathtt{G}} & \mathbf{E}_{\eta,-\mathtt{T}} & \mathbf{E}_{\eta,-\mathtt{C}} & \mathbf{E}_{\eta,-\mathtt{G}} & \mathbf{E}_{\eta,-\mathtt{T}} & \mathbf{0}
\end{array}$$

Fig. 5 Vector of emission matrices **E** of HMM–PHMM topology. Index η, $\eta \in \{1, 2\}$, associates matrices with regions. Upper-left quadrant shows symbols emitted by match state M in region η for specified sequences S_{W} = CTCGA and S_{Z} = AGTCGT. Similarly, upper-right quadrant and lower-left quadrant show symbols emitted by delete state X and insert state Y, respectively. Emission probabilities in each quadrant of region η sum to one. Lower-right quadrant is set arbitrarily to zero

This matrix is computed using Eq. 2, and following multiplication, Eq. 2 produces a $\ell_W \times 1$ submatrix whose elements sum to one. Similarly, the lower-left quadrant is a $1 \times \ell_Z$ submatrix, computed using Eq. 3, and whose elements describe the probability distribution for all gap-character insert symbols (-, z); $z \in \xi$, of the PIP sequence S_Z. Note that submatrices E_X and E_Y treat gaps as missing information. The lower-right quadrant is not used.

$$E_X = [P(t)\mathrm{W}]'\boldsymbol{q} \quad (2)$$

$$E_Y = \boldsymbol{q}'[P(t)\mathrm{Z}] \quad (3)$$

The Evolutionary Models

For the purpose of this study, two evolutionary models are selected with the aim of being neither too restrictive nor too general. This is because our data sets are designed to cover a wide range of evolutionary distances and are collected from across several species. At the same time, we could increase our samples to a manageable size by keeping the number of parameters to be estimated as small as possible.

We use the PMB model (Veerassamy et al. 2003) to model amino acid pairwise alignments. This model has the special feature of approximating, within an error of less than 5% on average, the entire BLOSUM series developed by Henikoff and Henikoff (1992). The PMB has also remained robust to additions of sequences to the blocks database, from which the BLOSUM series are derived (Veerassamy et al. 2003). This implies that our findings can be expected to remain valid even as more sequences continue to be added. An equally important feature of the PMB is that it can model alignment pairs over a wide range of evolutionary distances. This is because each matrix in the BLOSUM series is based on alignments that are clustered in blocks. Each block has aligned segments that share a specified percentage identity c. At the same time, the amount of information on evolution in each of these blocks also increases nearly linearly with c over a wide range of c values.

Similarly, we use the HKY model (Hasegawa et al. 1985) to model non-protein coding rRNA pairwise alignments. The HKY was chosen for this study because it allows for unequal nucleotide frequencies. In our modeling, it requires only one parameter—denoted by κ—to be estimated by numerical optimization in order to adjust for transition–transversion bias. The HKY model, therefore, also avoids excessive computational time.

The Hydrophilicity Parameter

In Pascarella and Argos (1992), the amino acid glycine most frequently flanked insertions, and amino acid isoleucine was most likely to be located away from gaps. The amino acids D, G, K, N, P, R, S, and T, which are all hydrophilic, were found more likely to appear on the flanks of indels than other amino acids. This feature was shown to be useful in heuristic modeling (Thompson et al. 1994). For our purpose, we subdivide the vector of 20 background probabilities of the set of 20 amino acids into a subvector $\boldsymbol{H}$ of 8 background probabilities of the hydrophilic subset and a subvector $\bar{\boldsymbol{H}}$ of 12 background probabilities of the nonhydrophilic subset. By introducing the hydrophilicity parameter h, $0 < h < 1$, we re-estimate from data the vector of background probabilities $\boldsymbol{q}$ using Eq. 4, where k is a suitable scalar so that the new elements in $\boldsymbol{q}$ still sum to one.

$$\boldsymbol{q} = k(h\boldsymbol{H} \cup (1-h)\bar{\boldsymbol{H}}) \quad (4)$$

The Data Sets

We used PIPs to construct our data sets. PIPs are sequence pairs that we considered to be independent and identically distributed (i.i.d.) for the purpose of this study. We deemed PIPs to be most suitable. The reason is that point and indel substitutions, resulting from the divergence of the PIP sequence pair from their common ancestor, represent events that are separate from events affecting any other PIP.

Protein PIPs were sourced from multiple alignments stored in the BAliBASE database (Thompson et al. 2005) and rRNA PIPs were sourced from multiple alignments stored in the European ribosomal RNA database (Wuyts et al. 2004). In both cases we employed the following procedure on each multiple alignment to extract PIPs. First, we obtained the set of all possible pairwise evolutionary distances, and eliminated pairs that had an extremely small distance. Second, a neighbor-joining tree (Saitou and Nei 1987) was built from the remaining pairs. Finally, we employed a post-order traversal routine to identify the most recent common ancestor for each PIP, as demonstrated in Fig. 6.

This procedure ensured that evolutionary processes that differentiated a PIP were distinct from evolutionary processes that differentiated other PIPs within each tree. The aim here was to reasonably satisfy our i.i.d. assumption among PIPs. We pooled all PIPs obtained from BAliBASE to produce the protein data set, and likewise we pooled all PIPs obtained from the European ribosomal RNA database to produce the rRNA data set. Each of these sets was then trimmed so that every taxon, and hence every PIP, in each set was unique.

The Protein Data Set

We generated 808 protein PIPs from multiple alignments stored in BBS files of BAliBASE 3.0 with references

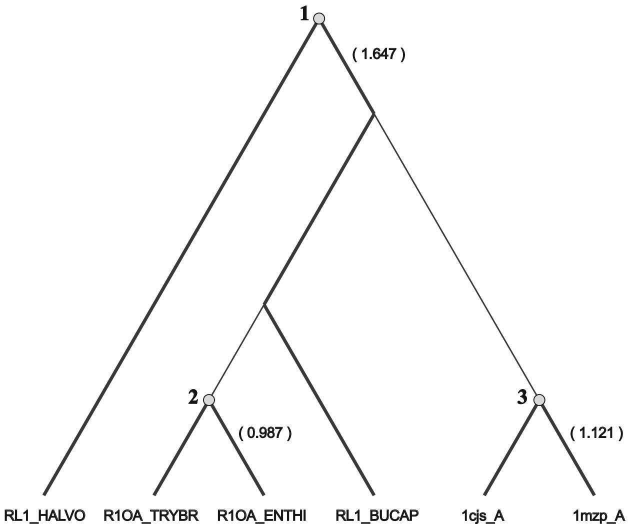


Fig. 6 Constructing PIPs using subset BBS12002 from BAliBASE 3.0. Tree is constructed as explained in text. Post-order traversal classifies six taxa into three pairs: (1) RL1_HALVO and RL1_BUCAP with evolutionary distance 1.647 and sharing ancestor number 1, (2) R10A_TRYBR and R10A_ENTHI with evolutionary distance 0.987 and sharing ancestor number 2, and (3) 1cjs_A and 1mzp_A with evolutionary distance 1.121 and sharing ancestor number 3. Distance between each pair is represented by bolded line sections. Each section is bifurcated by a unique ancestor. Each pair constitutes a PIP distinct from other two PIPs for experimental purposes

Ref1-V2, *Ref20*, *Ref30*, and *Ref50*. We did not include reference *Ref1-V1* as this consists of pairs which have less than 20% identity. This high divergence would have made it very difficult to align accurately for the purpose of this study. We also did not include reference *Ref40* because this consists of sequences with large N/C-terminal extensions. By convention, the sequence of a polypeptide (or protein) is written from the N-terminus on the left to the C-terminus on the right. Large extensions at these termini would entail large gaps at the two ends of the pairwise alignment. This would have been unsuitable for our study because one of our principal interests was indel placement with flanking residues.

Our sampling procedure produced experimental samples of 120 protein PIPs with evolutionary distances ranging from 0.25 to 1.25. One of these samples was chosen at random for this study.

The rRNA Data Set

To construct rRNA PIPs, we used two groups of rRNA sequences. Each group was randomly sampled from a large multiple alignment taken from the European ribosomal RNA database. The first group had 150 sequences and the second had 303 sequences, with non-gapped sequence lengths varying from 1088 to 1626 nucleotides. A tree was constructed from each group. The first tree yielded 74 PIPs and the second 151 PIPs, with all sequences being unique. We used a sampling procedure to produce an experimental sample of 99 rRNA PIPs from these two trees. The 99 PIPs have evolutionary distances ranging between 0.00 and 0.44.

Data and Source Code

A detailed sampling procedure for each data set is described in Supplementary Material. Data files for protein alignments are in MSF format and for rRNA alignments are in FASTA format. Source files for generating PIPs are written in Python and utilize modules available in PyCogent version 0.89.7. These files are available from the corresponding author on request.

Results

Model Validation

We validated the two-region model using simulated sets of 12 pairwise alignments. Each set was simulated following a preset regime of arbitrary parameter values. To ensure these alignments would provide enough statistical power, protein alignments were set to be 300 amino acids long and DNA alignments were set to be 900 nucleotides long. To recover parameter values, we obtained an ML estimator for one parameter at a time over two regions, while setting all other parameters at their corresponding nominal value. For each parameter, we optimized the likelihood function by varying the parameter independently in each region for each alignment. With 12 alignments and two regions, this procedure yielded 24 ML estimators for each parameter. We then used a large-sample t test to test for statistical power of each parameter in our model. Our tests showed that all estimated values were not significantly different from corresponding true values at the 5% level of significance.

We used analysis of variance to test for (i) main effects from different parameters, (ii) main effects from different nominal distances, and (iii) interaction between these two types of main effects. Interactions, and main effects from different nominal distances, were not significant at the 5% level of significance with both the protein and the rRNA samples. This was also the case with main effects from different parameters with the protein sample. With the rRNA sample, main effects from different parameters were not significant at the 1% level of significance, but were significant at the 5% level. We attribute the latter result to the κ parameter operating over a wide range of distances and with simulated sequences shorter than 1000 nucleotides. Considering that we constrained this parameter to be equal in the two regions throughout our experiments, we have no reason to suspect that this result had any adverse effect on our inferences. Full details are provided in Supplementary Material.

The Experiments

We investigated whether there exist two distinctive regions along the polypeptide (or section of non-protein coding DNA) in which rates of substitution are significantly different; for example, whether the rate of substitution in region one is significantly lower than in region two. To test whether there is a significant difference between the rate t_1 in region one and the rate t_2 in region two, of two aligned sequences S_W and S_Z in a PIP, we defined the following test:

$$H_o : t_1 = t_2 \text{ versus}$$
$$H_a : t_1 \neq t_2,$$

where all other parameters in region one were constrained to be equal to corresponding parameters in region two for both the null and the alternative hypotheses.

We maximized the likelihood function $\mathcal{L}(\boldsymbol{\theta}|S_W, S_Z)$ for each PIP using the Forward algorithm (Rabiner 1989). For the protein sample, $\mathcal{L}_o$ was maximized over the vector $\boldsymbol{\theta} = (t_1 = t_2,\ a_1 = a_2,\ r_1 = r_2,\ h_1 = h_2,\ \rho_1 = \rho_2 = c)$ under the null, giving $\hat{\boldsymbol{\theta}}_o$ following the completion of the first PIP alignment A_o. Fixed value c was some small arbitrary value as suggested in Churchill (1989), and was set to 0.001 to reduce Type II error. Likewise, under the alternative, $\mathcal{L}_a$ was maximized over the vector $\boldsymbol{\theta} = (t_1, t_2, a_1 = a_2,\ r_1 = r_2,\ h_1 = h_2,\ \rho_1, \rho_2)$, giving $\hat{\boldsymbol{\theta}}_a$ following the completion of the second PIP alignment A_a. Under this setting, we defined the χ^2 statistic shown in Eq. 5 to do an LR test for the two alignments A_o and A_a of each PIP in the data set.

$$\mathrm{LR} = 2(\mathcal{L}_a - \mathcal{L}_o) \sim \chi^2_{(\alpha=0.05,3)} \tag{5}$$

PIPs are i.i.d. for the purpose of this study, and we constructed a χ^2 statistic over the set ζ of M alignments. Each alignment in this set was required to have an LR statistic (Eq. 5) greater than zero. Because the null model was nested within the alternate model, the alternate must produce a likelihood greater than or equal to that of the null. A negative LR therefore indicates the alternate was not truly maximized. This condition can arise when the optimization method employed—simulated annealing (Goffe et al. 1994)—does not find the global optimum when maximizing the likelihood function. Accordingly, any PIPs that yielded LR ≤ 0 for any test were removed from the set ζ to ensure that they did not affect the accuracy of our results. (We note here that we used simulated annealing because of its effectiveness in dealing with large numbers of parameters (Goffe et al. 1994).)

We defined the general χ^2 statistic shown in Eq. 6 to do an LR test over all PIPs in the set ζ. In Eq. 6, n is the difference in the number of free parameters of $\mathcal{L}_a$ and $\mathcal{L}_o$ at each step of the summation, and is dependent on the test definition of ζ.

$$\varphi = \sum_{j=1}^{M} \mathrm{LR}_j \sim \chi^2_{(\alpha=0.05,\mathrm{df=nM})} \tag{6}$$

Protein Coding DNA

Table 1 shows the results from nine tests we carried out using the protein data set. Tests 1, 2, and 3 show that the replacement rate parameter t, the hydrophilicity parameter h, and the indel parameters together ℓ (=$a \cup r$), respectively, contributed to a significantly higher likelihood when allowed to vary independently in the two regions of our model, while all other parameters were constrained to be equal across the two regions.

These initial three tests led us to investigate pairwise colocation (that is, positional concomitance of two components) along polypeptides among the three components: substitution rate, hydrophilicity, and indels. We considered colocation between any two components, whose parameters were allowed to vary simultaneously and independently in the two regions under H_a, to exist if (1) H_o was rejected in favor of H_a, and (2) the levels of the corresponding estimators were both high (or both low) in the same region under H_a. We denote this colocation of two components C_1 and C_2 by $\varsigma_{(C_1,C_2)}$ for brevity.

Tests 4 and 5 in Table 1 show that on the one hand the number M of alignments that had LR > 0 decreased in each test. This is because when t and h were allowed to vary simultaneously and independently in the two regions under H_a, it was harder for the optimizer to find maximum likelihood in the presence of a weaker signal and more parameters. On the other hand, the number m of significant alignments increased from 10 to 37 when the hydrophilicity parameters h_1 and h_2 were added in the presence of the replacement rate parameters t_1 and t_2 (Tests 2 and 4). Likewise, m increased from 64 to 78 when the replacement rate parameters t_1 and t_2 were added in the presence of the hydrophilicity parameters h_1 and h_2 (Tests 1 and 5). The number m of significant alignments increased from a total of 74 in Tests 1 and 2 to a total of 115 in Tests 4 and 5, while the P value obtained from Tests 4 and 5 decreased sharply. These results strongly support the presence of colocation between substitution rate and hydrophilicity ($\varsigma_{(t,h)}$).

Table 1 Hypotheses testing with protein encoding

Test	Tests			LR tests over ζ					Concordances	
	H_o	H_a	n[a]	M[b]	m[c]	φ	df	P value	H_o, m[d]	H_a, m[e]
1	$t_1 = t_2$	$t_1 \neq t_2$	3	116	64	1304.99	348	1.30×10^{-110}	0.9425	0.9437
2	$h_1 = h_2$	$h_1 \neq h_2$	3	114	10	390.12	342	3.71×10^{-2}	0.9085	0.9081
3	$\ell_1 = \ell_2$	$\ell_1 \neq \ell_2$[f]	4	93	18	417.96	372	5.01×10^{-2}	0.8960	0.8843
4	$t_1 \neq t_2$	$h_1 \neq h_2$, $t_1 \neq t_2$	1	102	37	353.14	102	2.11×10^{-29}	0.9576	0.9658
5	$h_1 \neq h_2$	$h_1 \neq h_2$, $t_1 \neq t_2$	1	110	78	1179.88	110	1.24×10^{-178}	0.9455	0.9497
6	$t_1 \neq t_2$	$t_1 \neq t_2$, $\ell_1 \neq \ell_2$	2	73	4	136.68	146	6.98×10^{-1}	0.9544	0.9606
7	$\ell_1 \neq \ell_2$	$t_1 \neq t_2$, $\ell_1 \neq \ell_2$	1	92	54	776.08	92	8.93×10^{-109}	0.9326	0.9444
8	$h_1 \neq h_2$	$h_1 \neq h_2$, $\ell_1 \neq \ell_2$	2	99	29	408.13	198	1.09×10^{-16}	0.9187	0.9052
9	$\ell_1 \neq \ell_2$	$h_1 \neq h_2$, $\ell_1 \neq \ell_2$	1	113	40	384.62	113	2.58×10^{-31}	0.9476	0.9548

[a] n is the number of degrees of freedom for the test

[b] M is the number of PIPs in the sample which had LR > 0

[c] m is the number of PIPs from M PIPs in the previous column which were significant under H_a at the 5% level

[d] The average concordance for each test from m alignments under H_o

[e] The average concordance for each test from m alignments under H_a

[f] $\ell_1 \neq \ell_2$ means that both indel parameters are allowed to differ in the two regions; that is, ($a_1 \neq a_2$, $r_1 \neq r_2$)

To investigate further the feature $\varsigma_{(t,h)}$, we defined a new test as follows:

$$H_o : t_1 = t_2 \text{ and } h_1 = h_2 \text{ versus}$$
$$H_a : t_1 \neq t_2 \text{ and } h_1 \neq h_2,$$

where all other parameters in region one were constrained to be equal to corresponding parameters in region two for both the null and the alternative hypotheses.

The results from this test were $M = 120$, $m = 63$, and P value $= 4.39 \times 10^{-113}$ with df $= 480$. Of the 63 PIPs which expressed an LR statistic above the 5% level of significance with four degrees of freedom, 56 were found to have the feature $\varsigma_{(t,h)}$. Next we performed a large-sample test concerning the proportion P of items in a population that possess a quality of interest (Devore 1990, p. 308). In our case, the latter is the feature $\varsigma_{(t,h)}$, and we defined P as $\frac{X}{m}$, where X is the number of PIPs that had this feature and m is the number of PIPs for which LR was significant. We also assumed that X has approximately the binomial distribution. Considering that $m = 63$ is large, both $X = 56$ and $\hat{p} = \frac{X}{m}$ are also approximately normally distributed with $E(\hat{p}) = p$ and $\sigma_{\hat{p}} = \sqrt{p(1-p)/m}$. When H_o is true, $E(\hat{p}) = p_o$, $\sigma_{\hat{p}} = \sqrt{p_o(1-p_o)/m}$, and the test statistic is shown in Eq. 7 (Devore 1990, p. 308). The test showed that, conditional on the replacement rate and the hydrophilic content simultaneously present in the molecule are statistically significant, a very high percentage of protein sequences (approximately between 80 and 90% in our sample) exhibit colocation of these two components that contribute to evolutionary processes.

$$z = \frac{\hat{p} - p_o}{\sqrt{p_o(1-p_o)/m}} \sim N(0,1) \qquad (7)$$

Tests 6 and 7 in Table 1 show that both tests led to a substantial reduction in PIPs satisfying the constraint LR > 0. Adding ℓ_1 and ℓ_2 in the presence of t_1 and t_2 (Tests 3 and 6) increased the P value. Similarly, adding t_1 and t_2 in the presence of ℓ_1 and ℓ_2 (Tests 1 and 7) left the P value unchanged. In both cases, the number m of significant PIPs also decreased considerably. We therefore concluded that there was no support for the replacement rate parameter and the indel parameters varying simultaneously and independently in both regions.

This was not the case, however, with the hydrophilicity parameter. M here increased from 93 in Test 3 to 99 in Test 8, and decreased from 114 in Test 2 to 113 in Test 9. These changes are relatively small, and do not suggest any confounding between the hydrophilicity parameter and indel parameters in our two-region modeling. At the same time, in both Tests 8 and 9, the P value decreased sharply. Also, the total number m of significant PIPs increased from 28 (Tests 2 and 3) to 69 (Tests 8 and 9). These results strongly support the presence of colocation between hydrophilicity and indels ($\varsigma_{(h,\ell)}$).

To investigate further the feature $\varsigma_{(h,\ell)}$, we defined a new test as follows:

$$H_o : h_1 = h_2 \text{ and } \ell_1 = \ell_2 \text{ versus}$$
$$H_a : h_1 \neq h_2 \text{ and } \ell_1 \neq \ell_2.$$

The results from this test were $M = 114$, $m = 23$, and P value $= 1.16 \times 10^{-8}$ with df $= 570$. That is, the percentage of significant alignments was only 20%. This is much lower than that obtained from the test for the feature $\varsigma_{(t,h)}$, which was 52.5%. Considering that the percentage from the present test is lower by more than half, we do not attribute this large drop solely to statistical power. It is reasonable to say that indels seem to be less heterogeneous than replacement rates. Also, among the 23 significant alignments, 20 had the feature $\varsigma_{(h,r)}$, and an upper-tailed sign test gave a P value of 0.00024. This statistic provides further strong evidence of colocation between hydrophilicity and the indel rate ($\varsigma_{(h,r)}$) in our sample.

Non-Protein Coding DNA

Table 2 shows results from three tests using the rRNA data set. Test 1 shows once more a clear demarcation between slow and fast regions, with 91% of the alignments showing significance. The P value of these alignments was very close to zero, suggesting that the distinction between the two substitution rates in non-protein coding DNA sections is unequivocal.

Test 2 shows that there was confounding between the substitution rate parameters and the indel rate parameters when these were allowed to vary simultaneously and independently in the two regions under H_a, with M dropping from 98 to 87. Although the P value from this test shows clearly that indel rates varying independently in two regions are distinct between the two regions, this distinction is not common among alignments since m is only 7 in this case. That is, only 8% of the 87 PIPs that expressed $LR > 0$ were significant in our sample of 99 alignments, and therefore we conclude that the evidence in support of slow and fast indel rates in the two regions is weak.

Our purpose in Test 3 was to investigate potential colocation of fast substitution rates with fast indel rates. We found that the P value in this test remained very low after increasing the number of degrees of freedom. Of the 86 significant PIPs ($m = 86$), only 41 had the feature $\varsigma_{(t,r)}$. A sign test gave a P value of 0.705. This result does not provide evidence of colocation between the two rates in our non-protein coding DNA sample. We do not attribute this to the fact that we kept the indel length distribution parameter the same in the two regions. The latter was only for computational reasons. Tests (results not shown) using the protein coding data had shown that the indel length distribution parameter has no significant effect on colocation between the two rates. Nor do we attribute lack of evidence to statistical power, considering that P values obtained from Test 3 and from the sign test are very small and very large, respectively.

Concordances

To compute the concordance of our alignments with corresponding curated alignments, the column score (CS) defined in Thompson et al. (1999) was used. That is, $CS = \frac{1}{M}\sum_{i=1}^{M} C_i$, where M here is the number of sites in the test alignment, and C_i is 1 if site i, $i = 1, 2,\ldots,M$, is the same as the corresponding reference site, else C_i is 0.

For each test, in Tables 1 and 2, concordances were computed for each alignment under both H_o and H_a, but we averaged only across the m alignments that were significant at the 5% level. (This is the reason, for example, average concordances under H_o are different for Tests 1, 2, and 3 in Table 1.) This regime provided a meaningful measure of how much additional parameters varying independently in the two regions improve average concordance. For example, from Test 2 in Table 1, we can reasonably assume that allowing parameter h to vary independently in the two regions on its own is not likely to improve average concordance. However, average concordance appears to improve in the presence of t (Test 4) and in the presence of ℓ (Test 9) varying independently in the two regions simultaneously with h under both H_o and H_a in Table 1.

Discussion

Protein Coding DNA

The M values in Table 1 vary from 73 in Test 6 to 116 in Test 1. This variability shows that simulated annealing in our experiments encountered flat likelihood surfaces. As a result, the global optimum was not reached in approximately 15.6% of our optimization procedures applied to the

Table 2 Hypotheses testing with non-protein coding encoding

Test	Tests			LR tests over ζ					Concordances	
	H_o	H_a	n	M	m	φ	df	P value	H_o, m	H_a, m
1	$t_1 = t_2$	$t_1 \neq t_2$	3	98	90	5085.40	294	0.00	0.9555	0.9639
2	$t_1 \neq t_2$, $r_1 = r_2$	$t_1 \neq t_2$, $r_1 \neq r_2$	1	87	7	140.23	87	2.62×10^{-4}	0.9721	0.9704
3	$t_1 = t_2$, $r_1 = r_2$	$t_1 \neq t_2$, $r_1 \neq r_2$	4	98	86	5163.10	392	1.30×10^{-110}	0.9543	0.9630

Nomenclature same as in Table 1

protein data set. We attribute this optimization failure rate partly to the fact that we used the same initial values and the same default optimizer settings throughout our experiments. Although this regime led to missing data, it allowed us to reduce computational time and avoid subjectivity. Flat surfaces could also have been the result of the component under test being absent altogether in some of the alignments. Our results, therefore, are conditional on the optimization procedure detecting at least a small level of the component being present in the two aligned sequences. We do not have reason to suspect that this dependency has significantly affected our results.

The evidence in support of the two-region hypothesis is somewhat weak in the case of hydrophilicity (Table 1, Test 2) and in the case of indels (Table 1, Test 3). However, parameters t_1 and t_2 (Table 1, Test 1) expressed a strong demarcation between slow and fast rates, respectively, along polypeptides. This result is in accordance with intuition, but the extremely small P value compared with those obtained from Tests 2 and 3 is of note. We regard this result to be the more convincing for the fact that the PMB model contains a priori information on sequence evolution. Furthermore, our Q matrix is fixed and the same for the two regions, with only the frequency of amino acid character emissions by the PHMMs differing between the two regions. In spite of this prior information, t_1 and t_2 still came out distinctly different when allowed to vary independently in the two regions. This result is strong evidence that rates of substitution in the two broad regions of the protein coding DNA sequences in our data set were convincingly below and above the average.

Using Eq. 4 we have detected, through the h parameter, slices of the pairwise alignment that have a higher frequency of hydrophilic amino acids than the average measured across the entire alignment. That is, for an alignment whose LR test rejected the null, $h_2 > 0.5$ implied that hydrophilic amino acids were more prevalent in these slices than they were across the entire alignment. We regard these slices to be surrogates for solvent regions of the molecule. It is clear that h_1 and h_2 by themselves do not improve model fit in a convincing way (Table 1, Test 2). However, Tests 8 and 9 (Table 1) make it clear that when the indel parameters (represented jointly by ℓ) were allowed to vary simultaneously and independently in the two regions, while at the same time we also detected the solvent regions, improvement in model fit is highly significant. The results from these two tests are compatible with the heuristic applied in Clustal-W, whereby indels are assumed to occur more frequently in hydrophilic regions in order to "improve" the multiple alignment (Thompson et al. 1994).

Tests 4 and 5 show that model fit was, once more, improved significantly when we allowed the rate of substitution to vary independently in the two regions while the model was also detecting the solvent regions. This improvement can be explained by the fact that fast substitution rates are co-located with the solvent regions. That is, by allowing the background probabilities of hydrophilic amino acids to rise above average through the h_2 parameter in the fast rate region, the statistical power of the model also increases significantly as a result of the additional information in the Q matrix. The same result is obtained with indel rates in the fast rate region, except that here no colocation was found between solvent regions and indel lengths. Our colocation results for indels are compatible with what was reported in Pascarella and Argos (1992). That is, the solvent regions of protein coding DNA are more susceptible to indels.

Non-Protein Coding DNA

The two-region model was also a better fit to the rRNA data set (Table 2, Test 1). The evidence here is more convincing, where the P value was essentially zero. Of note are the M and m values, which both are much higher than they were with protein data. They are 99 and 91%, respectively, of the rRNA sample size. This high rate of success demonstrates that our optimizer performed much better with the HKY model when searching for the global optimum. This is because, unlike the PMB, the HKY has a much smaller state-space, namely, four nucleotides as opposed to 20 amino acids. The HKY, therefore, has a much higher probability of selecting the true symbols at each alignment site. Improved accuracy was achieved at a cost in terms of relatively much longer computer time due to longer sequences.

The result obtained from Test 1 (Table 2) was expected, following the result that had been obtained from the protein data set (Table 1, Test 1). This considering also, however, the nature of RNA secondary structure. For example, it has been observed that secondary structure interactions between paired nucleotides in an RNA sequence are generally stronger when compared to interactions that determine tertiary structure in the same sequence (Matthews et al. 1999). This type of interaction is essential to the stability of stem loops, and hence is highly conserved. In fact, the striking resolution between slow and fast substitution rates clearly indicates that the two rates can serve as surrogates for variations in secondary structure. That is, our method provided evidence in support of stem and non-stem regions of the RNA sequences in our sample. Since both M and m had a very high rate of success, our test suggests that this result can be extrapolated to rRNA sequences in general.

In Test 2 (Table 2) we uncoupled the indel rate from the substitution rate, and then we coupled the two rates in Test 3 (Table 2). The very small m value in Test 2, and the corresponding P values obtained from the two tests, makes it manifestly clear that coupling dramatically improves model fit. That is, meaningful modeling of rRNA data requires both parameters r and t to vary simultaneously and independently in the two regions. The large improvement in model fit would suggest that there was a strong correlation between these two components in our data set. However, we did not find statistical significance when we tested for positional colocation of high indel and high substitution rates. On the basis of our analyses, therefore, we are restricted to concluding that colocation of these two rates appears to be a property solely of protein coding DNA.

Concordances

We only measured the concordance of those alignments for which H_o was rejected in favor of H_a. We then averaged concordance measures for each test in Tables 1 and 2. This is the reason, for example, average concordances for Tests 2 and 3 in Table 2 are different in the last column, even though H_a is set in the same way for both tests. Note that $m = 7$ under Test 2, but $m = 86$ under Test 3. It should be clear that the average concordance measure under Test 3 is more reliable than that under Test 2. We also computed the corresponding average concordances of the same alignments under H_o, and these are presented in the penultimate column of the two tables.

In Table 1, the highest average concordance occurred when the substitution rate parameters and the hydrophilicity parameters were optimized simultaneously under H_a in Test 4. Here the average concordance is about 95–96%. This is a good result when compared to what is often reported in the literature. Edgar (2004), for example, reported that multiple aligners MUSCLE, MUSCLE-p, T-Coffee, and Clustal-W, all performed at about the 88% mark when benchmarked against BAliBASE alignments.

In Table 2, the highest average concordance occurred under H_o in Test 2. We emphasize that this does not mean that alignments under H_o are necessarily better than the corresponding ones under H_a. We suggest that the curated alignments in Test 2 were further away from the true alignments. This in view of the fact that the true alignment is a random variable, and hence it is also unobservable. For this reason, concordance measures should only be treated as a guide. All alignments—including curated (or *reference*) alignments—are statistics which, when constructed only tried to guess the true alignment. In general, therefore, preference should be given to the alignment that has been shown to be generated by a model that has stronger statistical support rather than to the alignment that has a better concordance, given the *reference* alignment.

Conclusions

We have evaluated the joint occurrence of indels and other attributes of biological sequences. In regard to protein coding DNA, we dealt specifically with the suggested association between indels and solvent accessibility by taking into account the hydrophilicity of amino acids in accordance with Pascarella and Argos (1992). We used a simple formulation (Eq. 4) to amplify the background probabilities of these amino acids whenever they were encountered at each site of the alignment by the Forward algorithm, irrespective of their frequency in each PIP. Our Eq. 4 is naive and mechanistic, and a future study may incorporate a more informative formulation. The fact that from Test 4 in Table 1 only about 25% of PIPs was significant (almost half as much as those from Test 1) could be because of over-simplification in Eq. 4. In spite of this simplification, Test 4 shows that our model under H_a was strongly preferred against the model under H_o. This is expected because the latter model assumes that hydrophilic residues have no effect on sequence evolution, which is incompatible with the results obtained from Tests 8 and 9 and corresponding tests for colocations.

We produced estimators using the Forward algorithm, along with posterior probabilities, in accordance with HMM theory (Rabiner 1989). Each estimator was computed by maximum likelihood across two regions using a 2-state HMM for each PIP. Our method is superior to estimating just one parameter across the entire data set. This is because our method avoided potential effects from extreme features usually present in the data set that would otherwise bias the "averaged" estimator. Another advantage is highlighted by the fact that our method can also be generalized for the study of regional heterogeneity of substitution processes. This generalization can be achieved by employing an n-state, $n \geq 2$, HMM to produce estimators across n regions.

Acknowledgments We thank Peter Maxwell for assistance with programming techniques and helpful suggestions throughout the project. We also thank Dr A. Isaev for his advice on the theory and implementation of Hidden Markov Models. Cray (Australia) are acknowledged for the provision of computational resources. We also thank the two reviewers for their comments and useful suggestions. We thank Dr J. Thorne especially for his numerous recommendations that helped us greatly to improve the manuscript. This work was supported by funding to GAH from the Australian Research Council.

References

Bishop MJ, Thompson EA (1986) Maximum likelihood alignment of DNA sequences. J Mol Biol 190:159–165

Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. Bull Math Biol 51:79–94

de la Chaux N, Messer PW, Arndt PF (2007) DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. BMC Evol Biol 7:191

Devore JL (1990) Probability and statistics for engineering and the sciences, 3rd edn. Brooks/Cole Publishing Company, Pacific Grove, California, pp 307–309

Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, pp 80–95

Edgar RC (2004) MUSCLE: low-complexity multiple sequence alignment with T-Coffee accuracy. Nucleic Acids Res 32(5): 1792–1797

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

Goffe WL, Ferrier GD, Rogers J (1994) Global optimization of statistical functions with simulated annealing. J Econom 60:65–99

Goldman N (1993) Statistical tests of models of DNA substitution. J Mol Evol 36:182–198

Goldman N, Thorne LJ, Jones TD (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J Mol Evol 263:196–208

Gonnet GH, Benner SA (1996) Probabilistic ancestral sequences and multiple alignments. In: Fifth Scandinavian Workshop on Algorithm Theory, Reykjevik

Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162:705–708

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Biol 22(2):160–174

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919

Knudsen B, Miyamoto MM (2003) Sequence alignments and pair hidden markov models using evolutionary history. J Mol Biol 333:453–460

Krawczak M, Cooper ND (1991) Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. Hum Genet 86:425–441

Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632

Matthews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911–940

Miklós I, Lunter GA, Holmes I (2004) A "Long Indel" model for evolutionary sequence alignment. Mol Biol Evol 21(3):529–540

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Pascarella S, Argos P (1992) Analysis of insertions/deletions in protein structures. J Mol Biol 224:461–471

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–285

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sjödin P, Bataillon T, Schierup MH (2010) Insertion and deletion processes in recent human history. PLoS ONE 5(1):e8650

Taylor MS, Ponting CP, Copley RR (2004) Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. Genome Res 14:555–566

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Thompson JD, Plewniak F, Poch O (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15(1):87–88

Thompson JD, Koehl P, Ripp R, Poch O (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61:127–136

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

Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol 34:3–16

Veerassamy S, Smith A, Tillier ERM (2003) A transition probability model for amino acid substitutions from blocks. J Comput Biol 10(6):997–1010

Wuyts J, Perriere G, Van de Peer Y (2004) The European ribosomal RNA database. Nucleic Acids Res 32:101–103

Yamane K, Yano K, Kawahara T (2006) Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. DNA Res 13:197–204

Zhang Z, Huang J, Wang Z, Wang L, Gao P (2010) Impact of indels on the flanking regions in structural domains. Mol Biol Evol (in press)