An Evolutionary Model for Protein-Coding Regions with Conserved RNA Structure

Jakob Skou Pedersen,[1] Roald Forsberg,*[1] Irmtraud Margret Meyer,† and Jotun Hein†*

*Bioinformatics Research Center, University of Aarhus, Aarhus, Denmark; †Genome Analysis and Bioinformatics Group, Department of Statistics, University of Oxford, Oxford, England

Here we present a model of nucleotide substitution in protein-coding regions that also encode the formation of conserved RNA structures. In such regions, apparent evolutionary context dependencies exist, both between nucleotides occupying the same codon and between nucleotides forming a base pair in the RNA structure. The overlap of these fundamental dependencies is sufficient to cause "contagious" context dependencies which cascade across many nucleotide sites. Such large-scale dependencies challenge the use of traditional phylogenetic models in evolutionary inference because they explicitly assume evolutionary independence between short nucleotide tuples. In our model we address this by replacing context dependencies within codons by annotation-specific heterogeneity in the substitution process. Through a general procedure, we fragment the alignment into sets of short nucleotide tuples based on both the protein coding and the structural annotation. These individual tuples are assumed to evolve independently, and the different tuple sets are assigned different annotation-specific substitution models shared between their members. This allows us to build a composite model of the substitution process from components of traditional phylogenetic models. We applied this to a data set of full-genome sequences from the hepatitis C virus where five RNA structures are mapped within the coding region. This allowed us to partition the effects of selection on different structural elements and to test various hypotheses concerning the relation of these effects. Of particular interest, we found evidence of a functional role of loop and bulge regions, as these were shown to evolve according to a different and more constrained selective regime than the nonpairing regions outside the RNA structures. Other potential applications of the model include comparative RNA structure prediction in coding regions and RNA virus phylogenetics.

Introduction

Some genome regions direct both the synthesis of a protein and the formation of biologically functional RNA structures. This overlap of information can be achieved because of redundancy in the genetic code and in the mapping of sequence to RNA structure, which provides a nucleotide string with considerable flexibility to optimize the composition of the encoded protein and RNA structure simultaneously. In RNA viruses several structural elements have been proposed to overlap protein-coding regions (Goodfellow, Kerrigan, and Evans 2003; Tuplin et al. 2002). One of these, the *cis*-acting replication element of the poliovirus has been shown to be involved in genome replication (Goodfellow, Kerrigan, and Evans 2003). As for cellular organisms, RNA structural elements within the protein-coding parts of the yeast ASH1 gene have been found to mediate protein localization during cell division (Chartrand et al. 1999, 2002). A recent study, however, suggests that a large fraction of protein-coding regions in bacterial and eukaryotic genomes may contain conserved local RNA secondary structure under a thermodynamic criterion (Katz and Burge 2003). The functionality of these RNA structures remains to be investigated, but in addition to protein localization several potential roles have been suggested, including an involvement in the splicing of introns, an effect on protein folding via the regulation of translation speed, and a regulation of gene expression mediated by mRNA stabilization (Katz and Burge 2003).

In this article we present a model of the nucleotide substitution process in such coding regions with conserved RNA structure (hereinafter, CORS). The model can be applied to data where there is a priori knowledge of the protein-coding and RNA structural annotation. Our focus here is to estimate parameters that contain information about the evolutionary process. Other potential applications of the model include comparative RNA secondary structure prediction in coding regions and the estimation of RNA virus phylogenies between higher taxonomical units where the double evolutionary constraints upon CORS could potentially alleviate the often incurred problems of saturation (Zanotto et al. 1996).

The functional and structural interactions of the amino acids within a protein can create a variety of evolutionary dependencies between protein-coding nucleotides. Most stochastic models of nucleotide substitution for coding regions consider only the simplest of these, namely the context dependency in the evolutionary process among nucleotides within adjacent non-overlapping three-tuples (codons) introduced by the triplet nature of the genetic code (Goldman and Yang 1994; Muse and Gaut 1994). These models ignore other interactions and assume evolutionary independence between codons, which means that the transition probability between sequences can be factored into the product of the transition probabilities between codons and calculated with relative ease.

Only certain base pairs can form the stable chemical bonds needed to maintain an RNA structure. The conservation of structure therefore introduces long-range context dependencies in the evolutionary process between base-pairing nucleotides. Existing stochastic models of nucleotide substitution for regions with RNA structure incorporate long-range correlations by considering two-tuples of base-pairing nucleotides as independent units of

[1] These authors contributed equally to this work.

Key words: RNA structure, coding region, overlapping information, context-dependent evolution, virus evolution.

E-mail: roald@birc.au.dk.

Mol. Biol. Evol. 21(10):1913–1922. 2004
doi:10.1093/molbev/msh199
Advance Access publication June 30, 2004

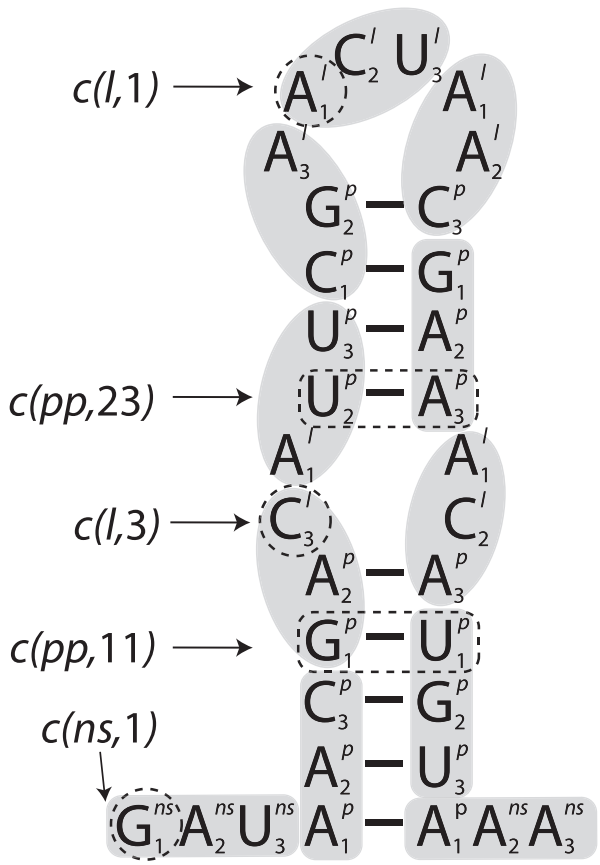

FIG. 1.—A region of coding RNA folded into an RNA secondary structure. Shaded areas indicate codon boundaries and subscripts indicate codon positions of nucleotides. The fragmentation function defines evolutionary independent tuples of nucleotides (dashed lines). Via a mapping function c, tuples are assigned a phylogenetic model based on the codon position and RNA structural annotation (ns = nonstructural, pp = base-pairing, l = loop and bulge).

evolution resulting in a similar factorization of the transition probability as described above [see Savill et al. (2001) for a description].

An evolutionary model of nucleotide substitution CORS must acknowledge that selection evaluates new mutations both in the context of the encoded protein and in the conserved RNA structure. As a consequence of the context dependencies described above, the evolutionary process of base-pairing two-tuples in stem regions can no longer be assumed to be independent of neighboring nucleotides, because these now make up the protein-coding context in which substitutions occur. The neighboring nucleotides may in turn base-pair with nucleotides from yet other codons, and thus expand the context dependency to include these (fig. 1). In this manner context dependency cascades throughout the structural regions and questions the computationally convenient assumption of evolutionary independence between short N-tuples of nucleotides.

A possible solution is to construct a model of nucleotide substitution that considers entire structures to be the unit of evolution and thus has a state space consisting of N-tuples spanning full structural regions. Here we shall refer to this type of model as *context elaborate*. These models were pioneered by Pedersen and Jensen (2001) in a study of viral genes with overlapping reading frames, and were later elaborated to model global context dependency introduced by protein-tertiary structure (Robinson et al. 2003). However, the size of the state space means that the calculation of transition probabilities in these models must rely on approximate statistical techniques such as Markov chain Monte Carlo with a computational demand that at present restricts their use to very small data sets.

An alternative solution is to reduce context dependencies to a level manageable by traditional phylogenetic models. The model presented here achieves this by replacing context dependencies within codons with codon position–specific heterogeneity in the substitution process and is inspired by the previous work of Hein and Stovlbaek (1995) and Yang (1996). Thus the input to our analysis is an alignment of DNA or RNA sequences with multiple layers of annotation, which we use to define sets of nucleotide tuples considered to evolve via independent, but annotation-specific, substitution processes. As this presents a general procedure in the construction of what we refer to as *context-reducing* models of molecular evolution, we develop a general conceptual framework for its presentation. (See Siepel and Hausler (2003) for a different approach to context reduction.) The assumption of independence between N-tuples allows for the factorization of transition probabilities and subsequent application of the model to a large data set of full-length genome sequences from the hepatitis C virus with known RNA structural elements. Using this data set, we evaluate different components of the model and estimate evolutionary parameters.

Materials and Methods

This section describes a general formalism for stating phylogenetic models for multiply annotated alignments. It begins by introducing the elements of traditional phylogenetic models. It then describes how multiply annotated alignments can be fragmented into independent N-tuples upon which a composite phylogenetic model can be defined, and how the parameter space of such composite models may be restricted. Finally, a specific model of nucleotide substitution in CORS is derived, based on the presented formalism. The notation for representing a phylogenetic model has been in part adopted in part from Siepel and Haussler (2003). A table of terms is given as Supplementary Material online.

Components of Phylogenetic Models

The data of traditional phylogenetic analysis is an alignment of n homologous sequences. Let the alignment be represented by a matrix x of dimension $n \times L$ with entries belonging to the alphabet Σ. The rows x^j ($1 \leq j \leq n$) of x correspond to the aligned sequences, and the entries of a column x_i ($1 \leq i \leq L$) correspond to homologous sequence symbols.

A standard assumption in phylogenetic analysis is that of evolutionary independence between short N-tuples

of nucleotides. In the following we consider models that split the alignment into such N-tuples.

Let a phylogenetic model for independent alignment columns be given by a five-tuple of model components $\psi = \{\Sigma, Q, \pi, \tau, \beta\}$, where Σ is the sequence alphabet of size d, Q is the instantaneous rate matrix of dimension $d \times d$, π is a vector of equilibrium frequencies of length d, τ is a rooted binary tree topology, and β is a vector of branch lengths. Phylogenetic models defined for N-tuples, instead of single columns, will have Σ replaced by Σ^N and the dimensionality of Q and π adjusted accordingly.

Σ, Q, and π define a continuous Markov process which is used to model the substitution process along the branches of the phylogenetic tree represented by τ and β. The likelihood of an N-tuple given a phylogenetic model can be calculated in time $O(n|\Sigma|^{2N})$ by the dynamic programing algorithm of Felsenstein (1981), where $|\Sigma|^N$ is the size of the N-tuple alphabet.

Let x_v be an N-tuple defined as the concatenation of the alignment columns specified by the N-long vector of indices $v(v_i \in \{1, \ldots, L\})$. The columns of x_v need not be direct neighbors in x. Let I be a set of index vectors defining a fragmentation of x into N-tuples. The likelihood of x given I is:

$$P(x \mid I, \psi) = \prod_{v \in I} P(x_v \mid \psi). \quad (1)$$

Annotated Alignments

Alignments can be annotated with information on the structure or function of different regions. This information can be given in the form of m label sequences, each drawn from a set A_k defined by annotation category k ($1 \leq k \leq m$). The annotation of x can then be represented by a matrix y of dimension $m \times L$. The rows y^j ($1 \leq j \leq m$) of y correspond to the label sequences. The complete annotation of alignment position i is contained in the column y_i ($1 \leq i \leq L$), which is a member of the combined label set $\mathcal{A} = \{u : u_k \in A_k \ \forall k\}$. The complete annotation of an N-tuple is given by the vector y_v belonging to the label set $\mathcal{A}_N = \{w : w_k \in (A_k)^N \ \forall k\}$.

The fragmentation of an alignment is determined by its annotation and the specific independence assumptions of a given model. Let $frag(x, y)$ be a mapping from an alignment and its annotation to an index set I, which then defines the fragmentation. The fragmentation of x thus partitions the sites in the alignment into nucleotide tuples which may be of varying lengths.

Defining a Composite Phylogenetic Model

Differences in the selective regime acting on the regions defined by the annotation will give rise to differences in the substitution process. It is therefore of interest to use different phylogenetic models for N-tuples with different annotations. A phylogenetic model for annotated alignments can be defined as the set $\psi = \{\psi^1, \ldots, \psi^K\}$, where the submodels $\psi^i = \{\Sigma^i, Q^i, \pi^i, \tau^i, \beta^i\}$ are traditional phylogenetic models for N-tuples as defined above.

Let c define a mapping from the annotation of an N-tuple onto $\{1, \ldots, K\}$ representing the set of phylogenetic models. The likelihood of an alignment given its annotation and a composite phylogenetic model is

$$P(x \mid y, \psi) = \prod_{v \in frag(x,y)} P(x_v \mid \psi^{c(y_v)}). \quad (2)$$

Parameterizations

The parameter space of a composite phylogenetic model is potentially large, but it can be reduced by introducing constraints on the legal parameter values. These constraints can be expressed by equations defining legal subspaces of the parameters and express assumptions about the substitution process. They can, for example, be introduced to test hypotheses or to create robust models for sparse data.

The off-diagonal entries of a rate matrix $q_{a,b}$ (with $a \neq b$, $a, b \in \Sigma$) denote the instantaneous rate of change from a to b. The diagonal entries are defined by the requirement that rows sum to zero ($q_{a,a} = -\sum_{a \neq b} q_{a,b}$). The matrix of transition probabilities P for a given time span t can be found by matrix exponentiation ($P(t) = exp(Qt) = \sum_{i=0}^{i=\infty} (Qi)^i/i!$) (Liò and Goldman 1998). It is convenient to normalize Q to one expected substitution per site per time unit by requiring the equality

$$N = \sum_a \pi_a \sum_{a \neq b} \delta(a, b) q_{a,b}, \quad (3)$$

where N is the length of a symbol and $\delta(a,b)$ count the number of positions at which a and b differ. This is a generalization of the normalization used by Siepel and Haussler (2003), and it allows direct comparison of branch length estimates between models of evolutionary units with different values of N.

The substitution process is commonly assumed to be time reversible, which can be ensured by the constraint of detailed balance: $\pi_a q_{a,b} = \pi_b q_{b,a} \ \forall a > b$. Other constraints commonly applied to nucleotide models include strand symmetry (Lobry and Lobry 1999) and a fixed ratio between transitions and transversions (Hasegawa, Kishino, and Yano 1985).

The substitution processes of the regions defined by an annotation can often be assumed to have some common properties. These properties can be intrinsic to the type of sequence being modelled—e.g., the transition bias of nucleotide sequences—or it can be due to a selective force acting across several regions. Including constraints between rate matrices allows tests of the validity of such assumptions and can considerably reduce the number of free parameters.

If there is no exchange of genetic material between the lineages of the phylogenetic tree, τ^i can be assumed to be the same between submodels. If the substitution process does not change between branches, β can also be assumed equal between submodels, in which case differences in the rate of substitution can be incorporated by defining β^i as a scaling of a general branch length vector: $\beta^i = r^i\beta$. Each submodel can then be defined as $\psi^i = (\Sigma^i, Q^i, \pi^i, r^i, \tau, \beta)$.

Likelihood Ratio Tests

When two models have nested parameter spaces, their relative fit can be evaluated by a likelihood ratio test (LRT) between the simpler (null) model ψ_i and a more complex (alternative) model ψ_j. The test statistic

$$LR = 2ln\left(\frac{P(x \mid y, \psi_j)}{P(x \mid y, \psi_i)}\right) \quad (4)$$

will be asymptotically $\chi^2_{\Delta df}$ distributed, where Δdf denotes the difference in degrees of freedom between two models (Ewens and Grant 2001).

A Composite Phylogenetic Model for Coding Regions with Conserved RNA Structures

In this section, the general framework outlined above is used to define a model of the substitution process in coding nucleotide sequences with overlapping RNA secondary structure.

Let the RNA structural annotation be given by the sequence y^S drawn from $A^S = \{ns, l, p\}$, where *ns* denotes nonstructural positions, *l* denotes loop and bulge positions, and *p* denotes RNA stem-pairing positions. Let the coding annotation be given by the sequence y^C drawn from $A^C = \{1, 2, 3\}$, where 1, 2, and 3 represent first, second, and third codon positions, respectively. (See figure 1 for examples of the annotation.)

All columns of the alignment, apart from the RNA stem-pairing ones, are assumed to evolve independently. As opposed to the standard models of coding regions, this corresponds to ignoring context dependency between nucleotides within the same codon. The fragmentation function thus maps stem-pairing positions onto index vectors for two-tuples, and everything else onto index vectors for single columns. The specific pairing of sites can be given to the *frag* function as an extra annotation sequence, which is implicit here.

Each possible labeling of the N-tuples defined by the fragmentation now defines a submodel. The possible labelings for single columns consist of the set $\mathcal{A}^{single} = \{ns, l\} \times A^C$, and for the two-tuples it consists of $\mathcal{A}^{pair} = \{p\}^2 \times (A^C)^2$, since the criteria defining the set of single columns and two-tuples were the presence or absence of the RNA-structure label *p*. The total number of submodels therefore becomes 15 ($|\mathcal{A}^{single}| + |\mathcal{A}^{pair}| = 6 + 9$). This corresponds to three different submodels for single nucleotides in the three codon positions of nonstructural regions, three different submodels for single nucleotides in the three codon positions of loop/bulge regions, and nine different submodels for the nine different codon-position combinations of a base-pairing nucleotide pair.

The enumeration of the submodels is given by the mapping $c : \mathcal{A}^{single} \cup \mathcal{A}^{pair} \rightarrow \{1, \ldots, 15\}$ (fig. 1). In the following discussion the submodels and their parameters will be referred to through their defining labels, as this is more informative than an explicit enumeration. A dot $(\cdot)$ is used to represent a label of any type. For example, the submodel for two-tuples modeling RNA stem-pairing first and third codon positions will be denoted $\psi^{c(pp,13)}$, and $\psi^{c(pp,\cdot)}$ will represent all submodels for stem-pairing two-tuples.

Rate Matrices for Single Sites

The rate matrices of submodels for single columns ($Q^{c(ns,\cdot)}$ and $Q^{c(l,\cdot)}$) are parameterized according to the HKY model (Hasegawa, Kishino, and Yano 1985). The constraints of the HKY model can be expressed by defining each off-diagonal entry in terms of four free parameters:

$$q^i_{ab} = \begin{cases} \pi^i_b & \text{if a and b differ by a transversion,} \\ \kappa^i \pi^i_b & \text{if a and b differ by a transition,} \end{cases}$$

where *i* is the index of the submodel, κ is the transition-transversion ratio (ts-tv ratio), and π is the equilibrium distribution defined by three free parameters.

Rate Matrices for Two-Tuples

The rate matrices of submodels for two-tuples ($\psi^{c(pp,\cdot)}$) are highly constrained to allow estimates from sparse data (table 2). Matrix entries are based on a pre-estimated reversible pair-symmetric two-tuple (i.e., 16×16) rate matrix (Q^{fixed}), which was estimated from a large set of stem-pairing sites from noncoding RNA-structures (Knudsen and Hein 1999). The rate matrix can be found at www.stats.ox.ac.uk/~meyer/CORSmodel. The off-diagonal entries of Q^i are defined by the equations

$$q^i_{ab} = \begin{cases} q^{fixed}_{ab} s^i_{left} & \text{if a and b differ in the left position,} \\ q^{fixed}_{ab} s^i_{right} & \text{if a and b differ in the right position,} \\ q^{fixed}_{ab} s^i_{left} s^i_{right} & \text{if a and b differ in both positions,} \end{cases}$$

where s^i_{left} and s^i_{right} represent the relative rate of substitution in the left and right site of a pair. These parameters are introduced to allow the relative rate of substitution to be dependent on the codon position involved, and they have an effect similar to the parameter that adjusts the ratio of nonsynonymous to synonymous substitutions in the codon model by Goldman and Yang (1994). The equilibrium frequencies ($\pi^{c(pp,\cdot)}$) used in the likelihood calculations for two-tuples are all fixed to the equilibrium distribution of Q^{fixed}. Because equilibrium frequencies are an implicit part of Q^{fixed}, they do not enter the parameterization of the rate matrices ($Q^{c(pp,\cdot)}$), but their values can be extracted from Q^{fixed} as this matrix fulfills detailed balance.

The Phylogenetic Tree and Tree Scales

The phylogenetic tree defined by τ and β is common to all submodels. The differences in substitution rates are modeled by the tree scales r^i, which are directly comparable because of the normalization of all rate matrices to one substitution per site per time unit (see eq. 3).

The Start Model

The most general form of the model ψ_{full} that we can construct has no shared parameters between the 15 submodels and thus contains $6 \cdot 5 + 9 \cdot 3 = 57$ parameters, excluding the phylogenetic tree.

We start, however, with a model ψ_{start} that is constrained in some dimensions. These constraints reflect the amount of available data, our hypotheses concerning the substitution process, and our desire to obtain biologically interpretable parameters.

In ψ_{start} we let the submodels for the three different codon positions in the nonstructural regions have separate parameter sets. Hence, we expect to see differences in the substitution process and rate which reflect the average effect that nucleotide substitutions in each codon position has on protein conservation. Specifically, we expect that third positions, where nearly all substitutions are synonymous, will show a higher estimated rate of substitution than first positions, where fewer substitutions are synonymous, and that these will again show higher rate estimates than second positions, where all substitutions are nonsynonymous. A similar relation is expected from the ratio of the rate of transitions to the rate of transversions (κ), again reflecting the relative number of transitions that contribute synonymous substitutions in each codon position. No scaling of the phylogenetic tree is needed for nonstructural third positions ($I^{c(ns,3)}$), because the phylogenetic tree was estimated from these. We therefore fix the rate of the third-position submodel to one ($r^{c(ns,3)} = 1$).

Although both nonstructural and loop/bulge regions are nonpairing, the latter may have a biological functionality which distinguishes them from the former. In the starting model, therefore, no ties were introduced between parameters of loop/bulge and nonstructural substitution models.

Because of the relative sparseness of data columns in each of the two-tuple annotation categories ($I^{c(pp,\cdot)}$), we have chosen to constrain their submodels considerably. Thus, we set $s_{left}^{c(pp,jk)} = s_j$, $s_{right}^{c(pp,jk)} = s_k$, $s_3 = 1\ \forall j, k \in \{1, 2, 3\}$, where s_1 and s_2 are two new free parameters shared between all relevant two-tuple models. This corresponds to assuming that the codon-position–specific effects on the relative rate of substitution are independent of the specific position combination and removes 16 free parameters compared to ψ_{full}. Because Q^{fixed} is symmetric, this specifically induces equivalence between models with symmetric codon positions, so that $\psi_{c(pp,jk)} = \psi_{c(pp,kj)}$. The normalization procedure means that s_1 and s_2 are defined relative to $s_3 = 1$. Their estimates can thus be interpreted as the effect that protein conservation has on the relative rate of substitution compared to the rate in third positions. Hence, we would expect these estimates to rank like the rates of substitution in the nonstructural regions.

We also constrain the substitution rate parameters of the two-tuple models by parameterizing the rate as $r^{c(pp,jk)} = r_p(r^{c(ns,j)} + r^{c(ns,k)}/2)\ \forall j, k \in \{1, 2, 3\}$, where r_p is a new free parameter shared between all two-tuple models. Given the previously mentioned normalization procedure, $(r^{c(ns,j)} + r^{c(ns,k)})/2$ is the expected rate of substitution for a nonstructural nucleotide pair evolving independently. This means that r_p can be interpreted as a scaling of the substitution rate in the nonstructural regions, induced by structure conservation.

Constraints on the parameters of ψ_{start} will be used to express hypotheses on the substitution process, which are then tested in a likelihood ratio framework (see *Results*).

Parameter Estimation

The maximum likelihood estimate (MLE) $argmax_{\psi} P(x \mid y, \psi)$ can be found through numerical optimization. Such optimizations are computationally intensive and prone to be caught in local optima when the dimensionality of the parameter space is large. The number of parameters subject to numerical optimization is here reduced by pre-estimating the phylogenetic tree (τ and β), and by following the normal practice of using a simple analytic estimate for the equilibrium frequencies, which is derived by counting, and thus is not based on τ and β. In the following discussion we denote the set of index vectors mapped to ψ^i by $I^i = \{v \in frag(x, y) : c(y_v) = i\}$.

The estimate of τ and β is based on third-codon positions in nonstructural regions ($I^{c(ns,3)}$). This allows the r^i estimates to be interpreted as the rate of substitution relative to sites in $I^{c(ns,3)}$ (third position). A distance matrix based on Kimura's two-parameter model (Kimura 1980) was found using DNADIST with default settings from the PHYLIP package (Felsenstein 1993). τ was estimated from this distance matrix using Weighted Neighbor Joining (Bruno, Socci, and Halpern 2000). The BASEML program from the PAML program package (Yang 2000) was used to find a MLE of β under a HKY model (which thus corresponds to $Q_0^{c(ns,3)}$), keeping τ fixed.

The estimator for π^i is the symbol frequency in the set of N-tuples defined by I^i. When the equilibrium distribution is constrained to being the same for several submodels, the estimator becomes the symbol frequency in the corresponding union of N-tuple entry sets. Recall that only $\pi^{c(ns,\cdot)}$ and $\pi^{c(l,\cdot)}$ are free parameters, because $\pi^{c(pp,\cdot)}$ are pre-estimated along with Q^{fixed}.

The MLEs of the remaining parameters of ψ (i.e., Q^i and r^i) are found using the quasi-Newton numerical optimization method, with BFGS approximation of the Hessian implemented in the OPT++ package (Meza 1994). The optimization was found to be robust to the initial parameter values. Rewriting the composite-likelihood expression (see eq. 2),

$$P(x \mid y, \psi) = \prod_{i=1}^{K} \prod_{v \in I^i} P(x_v \mid \psi^i), \tag{5}$$

makes it clear that submodels with no shared parameter constraints can be optimized independently to reduce computational time.

Approximative standard errors of the MLE found by the numerical optimization procedure were calculated from an estimate of the Fisher information matrix [e.g., Ewens and Grant (2001)]. The estimator used was a difference approximation to minus the Hessian of the log-likelihood function ($ln(P(x \mid y, \psi))$ evaluated at the MLE. This approximation of the standard errors relies on the asymptotic behavior of the MLE; the standard errors of estimates based on sparse data (i.e., parameters based on subsets of $I^{c(pp,\cdot)}$ or $I^{c(l,\cdot)}$) are therefore only indicative.

Implementation

A general framework for phylogenetic analysis has been written in C++ which allows models to be specified

Table 1
Label Distribution for Single Columns

Label	1	2	3	·
ns	2914	2915	2919	8748
l	21	21	24	66
·	2935	2936	2943	8814

NOTE.—The distribution of non-base-pairing sites in the analysed data given by the structural and coding annotation. RNA-structure labels (A^S) are given vertically and divide the sites into nonstructural sites (*ns*) i.e., sites outside RNA structural elements and loop and bulge sites (*l*) which are located inside structural elements but not involved in base-pairing. Coding labels (A^C) which correspond to codon position are given horizontally. The sum over all labels within a category is represented by a dot.

Table 2
Label Distribution for Two-Tuples

Label	1	2	3	·
1	16	10	8	34
2	10	9	15	34
3	7	13	9	29
·	33	32	32	97

NOTE.—The distribution of base-pairing nucleotide two-tuples in the analysed data set as given by the structural and coding annotation. The label of all base-pairing two-tuples consists of the codon position of the left-most nucleotide and the codon position of the right-most nucleotide ($p_{left}p_{right}$). The left part of the coding label (p_{left}) is listed vertically, and the right part of the coding label (p_{right}) is listed horizontally. The sum over all labels within a category is represented by a dot.

in XML. A Linux executable version can be downloaded from www.stats.ox.ac.uk/~meyer/CORSmodel.

The Data

Hepatitis C virus (HCV) is a flavivirus belonging to the *Flaviviridae* family with a positive-sense single-stranded RNA genome. The genome is approximately 9,500 bases long and contains a single open reading frame (ORF) which encodes one large polyprotein. On the basis of phylogeny, hepatitis C viruses are divided into genotypes 1 through 6, and these are further divided into subtypes designated by a, b, and c, in order of discovery. Two RNA structures have been described in the 5′ and 3′ untranslated regions of the HCV genome: one is involved in initiation of RNA replication (Yi and Lemon 2003) and the other functions as an internal ribosomal entry site (Reynolds et al. 1996). However, it has recently been demonstrated, via bioinformatics (Tuplin et al. 2002) and enzymatic mapping techniques (Tuplin et al. personal communication), that five RNA secondary structures exist within the 3′ part of the coding region. These structures define our structural annotation.

The alignment, which contained only 0.27% gaps, was generated manually from the coding part of 99 HCV genotype 1a and 1b genomic sequences (alignment with accession numbers available at www.stats.ox.ac.uk/~meyer/CORSmodel). The RNA structure annotation of the alignment was extrapolated from the genotype 1a sequence used in the experimental validation of the coding structures. Tables 1 and 2 provide the distribution of nucleotides between the different structural categories. All sites outside these five structures were annotated as nonstructural (*ns*). The first 50 sites of the alignment were discarded due to an RNA structure known to extend from the 5′ UTR into the beginning of the coding region (Reynolds et al. 1996).

Results

The estimated phylogenetic tree had a total branch length of 9.84 expected substitutions per site and can be downloaded from www.stats.ox.ac.uk/~meyer/CORSmodel.

Model Comparisons

The model ψ_{start} was taken as a starting point for the model comparisons. Simpler models are defined by successively adding constraints to the parameter space of ψ_{start}. This leads to a hierarchy of nested models, which is depicted in figure 2. The relative fit of successive models is evaluated by likelihood ratio tests where the simpler model represents the null hypothesis and the more general model represents the alternative hypothesis. A significant *P* value will therefore give rise to rejection of the simpler model and retention of the more general model. A nonsignificant *P* value does not lend support to rejection of the simpler model, in which case the simpler model is adopted. Table 3 defines the models in terms of their constraints. Table 4 reports the likelihoods, the test statistics, and the *P* values of each test.

The first model comparisons were made to evaluate the importance of allowing for heterogeneity in the substitution process between loop/bulge and nonstructural regions. In the model ψ_1 we allow for differing rates of substitution in the three codon positions of the single-site models, but tie these rates between nonstructural and loop/bulge regions. The comparison of ψ_{start} and ψ_1 thus tests the significance of letting codon position–specific rates of substitution in loop/bulge regions differ from those in nonstructural regions. This feature was found to be significant, but it does not illuminate whether this heterogeneity consists of a difference in the relative rates of substitution in the three codon positions or of a general rate change affecting all codon positions evenly. We therefore constructed model ψ_2, which represents the hypothesis that the relative position-specific rates of substitution are equal between loop/bulge and nonstructural regions but allows for a general scaling of all three rates in loop/bulge positions through the parameter r_l, so that $r^{c(l,k)} = r_l r^{c(ns,k)}$, $\forall k \in \{1, 2, 3\}$. This simpler model provided a fit to data not significantly worse than ψ_{start}, and it was therefore adopted as our new null model.

A further LRT showed no significant effect of letting nucleotide equilibrium frequencies differ between nonstructural and loop/bulge regions (ψ_3 vs. ψ_2), and ψ_3 therefore replaced ψ_2 as our null model. The opposite was observed when testing for difference in the transition bias between nonstructural and loop/bulge regions (ψ_4 vs. ψ_3).

Comparisons of models ψ_5 and ψ_6 versus ψ_3 showed that both transition bias and nucleotide equilibrium frequencies are significantly different between the three different codon positions in the nonstructural regions.

Next, we turned to the submodels that describe substitution of two-tuples. By comparing models ψ_7 and

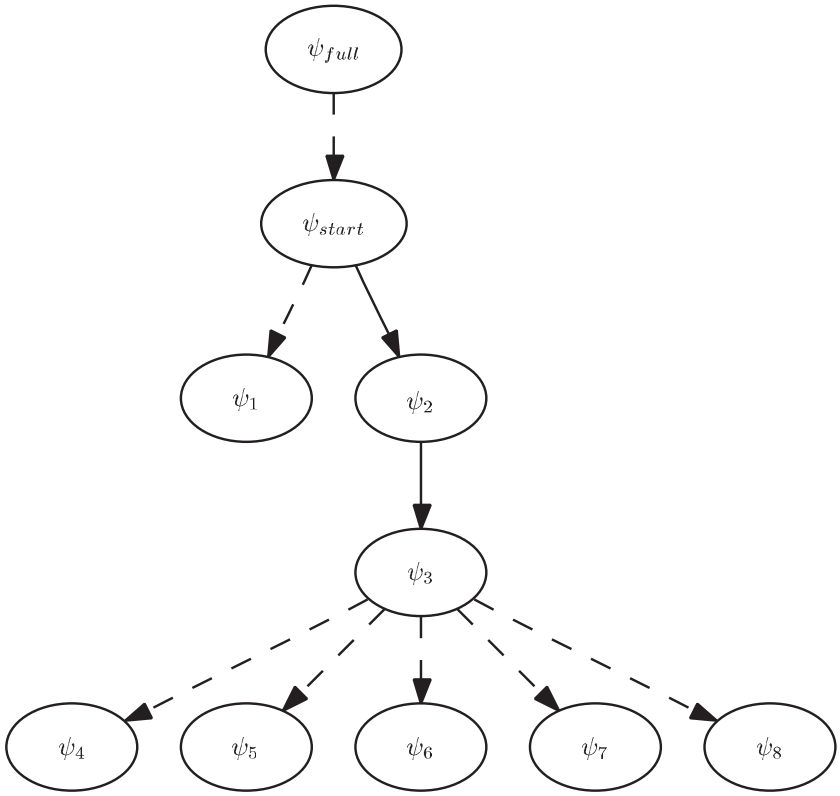

FIG. 2.—Graph of the nesting relationship between the CORS models. ψ_{full} represents the most general model. All other models are defined by introducing constraints on the free parameters of ψ_{full}. Each arrow thus corresponds to a set of parameter constraints (see table 3), and is accompanied by a likelihood ratio test (see table 4). Broken arrows represents tests which reject the constrained model as fitting the data significantly worse. Solid arrows represent tests which did not lead to rejection of the constrained model.

ψ_3, we found a significant effect of including the codon-position–specific skewing via the parameters s_1 and s_2. A similar observation was made for the parameter r_p, which scales the rate of substitution in base-pairing regions (ψ_8 vs. ψ_3).

Lastly, we tested our restricted start model ψ_{start} against the completely unrestricted full model ψ_{full}, which was found to provide a significantly better relative fit.

Parameter Estimates

Parameter estimates from the final model (ψ_3) and standard errors are listed in table 5. Values of the rate of substitution and the transition bias estimated for the nonstructural submodels followed our prediction and were ordered by codon position as follows: *third* > *first* > *second*.

This relation was not followed by the estimated parameters of transition bias in the loop/bulge regions, but the large degree of uncertainty associated with these does not permit any strong conclusion about the ordering. Estimates from third codon positions in loop/bulge regions had confidence intervals which did not overlap those of their counterparts in the nonstructural regions, showing that this position has a significant difference in the transition bias.

Considering the two-tuple submodels, we found a reduced relative rate of substitution in first and second codon positions. Contrary to expectation, the estimated effect was slightly more pronounced ($s_2 > s_1$) in first than in second positions, but these estimates are also associated with a high degree of uncertainty.

The scaling parameters for the structural regions (r_l and r_p) showed a reduction in the absolute rate of substitution to about half in loop/bulge regions and about a third in base-pairing regions.

Discussion

Here we have presented a composite phylogenetic model of nucleotide substitution in protein-coding regions with conserved RNA structures and applied it to a data set of full-length genome sequences of the hepatitis C virus.

In base-pairing regions we found that the substitution process is affected by selection to conserve both the amino acid sequence of the encoded protein and the RNA structure. The protein-coding constraint was reflected as a lowering of the relative rate of substitution in base-pairing nucleotides occupying first and second codon positions, compared to that observed in noncoding RNA structural regions, and the constraints from structural conservation caused a marked reduction of the rate of substitution in base-pairing regions compared to nonstructural regions.

We also inferred that selection imposes a significantly different filtering of mutations in loop/bulge regions compared to nonstructural regions, which results in a lowered rate of substitution and a difference in the relative number of accepted transitions and transversions. This indicates that loop/bulge regions of the investigated

Table 3
Model Descriptions

ψ_i	Preceding Model	Parameter Constraints Introduced		df
ψ_{full}				57
ψ_{start}	ψ_{full}	$r^{c(ns,3)} = 1$		32
		$s_{left}^{c(pp,jk)} = s_j$, $s_{right}^{c(pp,jk)} = s_k$, $s_3 = 1$	$\forall\, j,k \in \{1, 2, 3\}$,	
		$r^{c(pp,jk)} = r_p(\frac{r^{c(ns,j)} + r^{c(ns,k)}}{2})$ (A)	$\forall\, j,k \in \{1, 2, 3\}$	
ψ_1	ψ_{start}	$r^{c(l,j)} = r^{c(ns,j)}$(B)	$\forall\, j \in \{1, 2, 3\}$	29
ψ_2	ψ_{start}	$r^{c(l,j)} = r_l r^{c(ns,j)}$ (C)	$\forall\, j \in \{1, 2, 3\}$	30
ψ_3	ψ_2	$\pi^{c(l,j)} = \pi^{c(ns,j)}$ (D)	$\forall\, j \in \{1, 2, 3\}$	21
ψ_4	ψ_3	$\kappa^{c(l,j)} = \kappa^{c(ns,j)}$ (E)	$\forall\, j \in \{1, 2, 3\}$	18
ψ_5	ψ_3	$\pi^{c(ns,1)} = \pi^{c(ns,2)} = \pi^{c(ns,3)}$(F)		15
ψ_6	ψ_3	$\kappa^{c(ns,1)} = \kappa^{c(ns,2)} = \kappa^{c(ns,3)}$(G)		19
ψ_7	ψ_3	$s_1 = s_2 = s_3$ (H)		19
ψ_8	ψ_3	$r_p = 1$ (I)		20

NOTE.—df denotes the number of free parameters. The parameter constraints are introduced relative to the preceding model. Brief interpretation of constraints: **A**: The rate of the third position nonstructural sites set to scale with phylogenetic tree. The skewing of the two-tuple rate matrices only depends on the involved codon positions. The rate of two-tuple changes is proportional to rate of change had the codon positions been in nonstructural regions. **B**: Equal evolutionary rates between equal codon positions of loop/bulge regions and nonstructural regions. **C**: Proportional evolutionary rates between equal codon positions of loop/bulge regions and nonstructural regions. **D**: Equal equilibrium frequencies between equal codon positions between loop/bulge regions and nonstructural regions. **E**: Equal ts-tv ratio between equal codon positions between loop/bulge regions and nonstructural regions. **F**: Equal equilibrium frequencies among all single sites submodels. **G**: Equal ts-tv ratio between codon positions of nonstructural regions. **H**: No codon position–specific effect on the relative rate of changes in two-tuples. **I**: Paring regions evolve with the same evolutionary rate as nonstructural regions.

structures are not mere spacer regions but have a biological function imposing additional selective constraints. This added constraint could occur because loop/bulge regions mainly occupy slowly evolving protein regions. The amino acid composition of loop regions, however, shows no sign of a bias toward slowly evolving amino acids like proline, cysteine, and tryptophan when compared to the overall amino acid composition (results not shown). It therefore seems more likely that this functionality is related to the RNA structures where it could be mediated by the formation of pseudo knots via complementary base-pairing or the interaction with proteins as described for the loop region of the *cis*-acting replication element (CRE) of polioviruses (Goodfellow, Kerrigan, and Evans 2003).

A factor which may affect both parameter estimates and LRTs is the fidelity of the structural annotation. The experimental annotation employed in our study is based on a HCV genotype 1a strain, and our fragmentation of the alignment, and subsequent parameter estimation, is based on the assumption that its structure is conserved throughout the alignment. However, a comparison to an experimental annotation of homologous RNA structures in HCV genotype 2 shows considerable differences (Tuplin et al. in progress). Our data set does not contain any genotype 2 sequences but consists solely of sequences from the more closely related subtypes 1a and 1b. Still, we found that 10.1% of the positions annotated as pairing contain mismatching nucleotides (mismatching with respect to base-pairing—i.e., pairs other than A-T, C-G, or T-G), indicating that functional conservation within HCV genotype 1 may be achieved with some structural

Table 4
Model Comparisons and Test Statistics

ψ_i	Tested Against	Δ df	Likelihood	*LR*	*P* value
ψ_{full}	—	—	$8.34231*10^{-65371}$	—	—
ψ_{start}	ψ_{full}	25	$6.07838*10^{-65405}$	157.21	0.0
ψ_1	ψ_{start}	3	$2.47814*10^{-65419}$	66.27	0.0
ψ_2	ψ_{start}	2	$9.71299*10^{-65406}$	3.67	0.160
ψ_3	ψ_2	9	$7.47754*10^{-65407}$	5.13	0.823
ψ_4	ψ_3	3	$5.39943*10^{-65411}$	19.07	0.0003
ψ_5	ψ_3	6	$7.53933*10^{-65619}$	976.28	0.0
ψ_6	ψ_3	2	$1.04188*10^{-65843}$	2011.80	0.0
ψ_7	ψ_3	2	$4.58388*10^{-65489}$	378.60	0.0
ψ_8	ψ_3	1	$5.71316*10^{-65481}$	341.32	0.0

NOTE.—Δdf is the difference in the number of free parameters, and *LR* is the value of the likelihood ratio test statistic.

Table 5
Parameter Estimates and Standard Errors for Our Final Model ψ_3

name	value	stderr
$\kappa^{c(ns,1)}$	4.84	0.22
$\kappa^{c(ns,2)}$	4.28	0.26
$\kappa^{c(ns,3)}$	19.71	0.63
$\kappa^{c(l,1)}$	1.605	1.47
$\kappa^{c(l,2)}$	16.07	20.8
$\kappa^{c(l,3)}$	8.621	3.19
$r^{c(ns,1)}$	0.206	0.0044
$r^{c(ns,2)}$	0.115	0.0032
s_1	0.0461	0.018
s_2	0.0573	0.019
r_p	0.384	0.034
r_l	0.476	0.066

NOTE.—Estimated parameters from the substitution model include the following: (1) The ratio of the rate of transition over the rate of transversions in the single-site models (κ). These are given for each of the three codon positions and for both nonstructural sites (*ns*) and loop/bulge sites within RNA structures (*l*). (2) The rate of substitution in first and second codon positions of nonstructural regions relative to the rate of substitution at third codon-positions ($r^{c(ns,1)}$ and $r^{c(ns,2)}$). (3) The scaling of the rate of substitution of base-pairing nucleotides occupying first and second codon positions, compared to those occupying third codon positions (s_1 and s_2). (4) The scaling of the rate of substitution in the base-pairing regions (r_p) and in the loop/bulge regions (r_l), as compared to the rate of substitution in nonstructural regions. (See also text section, *The Start Model*, for an interpretation of the parameters.) stderr denotes the standard error. Approximative confidence intervals of parameters can be found as $\pm 1.96 \cdot$ stderr. (See text section, *Parameter Estimation*.)

flexibility. Some positions may thus be mis-annotated on part of the tree and could potentially affect parameter estimates. To investigate this possibility, we re-estimated parameters under model ψ_3, using a "cleaned" data set, where alignment columns containing mismatching nucleotides were treated as missing data. This showed that mis-annotation results in a small upward bias in the estimated rate of substitution in base-pairing regions ($r_p = 0.323$ vs. $r_p = 0.384$). A slight effect was also observed on the s parameters, which changed their relation from $s_1 < s_2$ to the expected relation $s_2 < s_1$.

The rate estimates of loop/bulge submodels could potentially be downward biased if regions annotated as loop/bulge are indeed base-pairing throughout a part of the tree. However, because of the strength of the constraints introduced by complementary base-pairing, we expect the effect of structural evolution to be greatest in two-tuple models.

Although the effects we observe are small, it is clear that the assumptions on which our model rests are sensitive to structural evolution. This question could be addressed through the use of models that allow the RNA structure to evolve along the tree. Such models do not exist at present but represent an exciting challenge.

There are more immediate ways in which the present approach could be improved. One would expect that the constraints of protein conservation differ between amino acids with different structural or functional roles in the protein. The resulting heterogeneity in the substitution process could be accommodated either by adding additional layers of annotation describing protein structure and function or by integrating over a distribution of, e.g., substitution rates (Yang 1993; Felsenstein and Churchill 1996).

We have stated our model via a general framework for constructing context-reducing phylogenetics models for genetic data with multiple annotations. There are several potential applications of this modeling framework, including, for example, protein sequences where both secondary and tertiary structure is taken into account, coding regions with overlying splice-sites, and coding regions annotated by genomic characteristics (e.g., isochore vs. non-isochore). The main practical limitations on the use of the presented framework will be of time usage and robustness of the parameter optimizations. Total time usage will depend on the time and number of likelihood calculations in the optimization procedure. Because the time spent in each likelihood calculation is proportional to the number of sites and squares in the alphabet size of the sites, it becomes impractical in most cases to fragment the alignment into tuples longer than three. Adding free parameters to a model will generally increase the complexity of the search space. Thus, the needed number of likelihood calculations will grow more than linearly with the number of free parameters, and the chance of finding the global optimum will decrease. This is true for a given submodel, but the relation between the total number of free parameters and the overall optimization procedure is more complex. A way of reducing the overall number of parameters is by constraining these between submodels. However, such constraints make submodels interdependent, and they increase the number of free parameters which have to be optimized simultaneously and thus the complexity of the search space. As a consequence, the effect of the total number of parameters on the speed and fidelity of the estimation procedure will depend on a complex interplay between the data and the structure of the model, and no clear guidelines are available. However, we note that the optimizations procedure employed here was both feasible and robust with more than 50 free parameters.

The comparison between the full model and our constrained starting model showed that the latter does not capture the full complexity of the evolutionary process. As more data from CORS accumulate, more elaborate models should be explored. It would also be of great interest to evaluate the goodness-of-fit that our model provides to data, and to estimate the validity of our assumption of a $\chi^2_{\Delta df}$ distributed LRT statistic. Both could be tested by parametric bootstrapping (Goldman 1993), but because of the numerical optimization procedures used, this would be extremely demanding computationally.

Furthermore, it would be interesting to evaluate how well our context-reducing model compares to context-elaborate models such as that of Pedersen and Jensen (2001), which represent a more loyal description of the true evolutionary process in structural regions. The latter approach employs a more accurate model of the evolutionary process, but it must resort to approximative computational techniques, whereas our model represents an approximation to the known context dependencies in the evolutionary process but relies on exact calculations. Thus the choice of substitution model type stands between models which treat a small amount of information with the greatest possible accuracy and approximative models which treat the greatest possible amount of information with reduced accuracy. Our objectives here were to develop a model that could be used in the estimation of evolutionary parameters and the testing of evolutionary hypotheses, in comparative RNA structure prediction in coding regions, and in RNA virus phylogenetics. At present the challenges in implementation and computation do not allow for the routine use of process-based models to solve any of these problems. This motivated our choice of a context-reducing model. As algorithms and computers improve, context-elaborate models will, however, become more attractive. The choice of model will then be determined by the type of analysis performed. Thus, context-elaborate models may become the models of choice for the estimation of evolutionary parameters and the testing of evolutionary hypotheses. However, for use in comparative RNA structure prediction and RNA virus phylogenetics, we believe that the computational demand of high throughput sequence analysis and phylogenetic algorithms will dictate the use of computationally convenient approximative models for quite some time.

Conclusion

Here we have presented a first model of nucleotide substitution in protein-coding regions with embedded, conserved RNA structure. The model is based on an approximation to the known context dependencies in the substitution process and proposes a general framework for constructing context-reducing models by fragmentation of

a multiply annotated alignment into short independent N-tuples. This framework should find use in the construction of phylogenetic models for genomic regions that are under overlapping functional constraints. We have used the model to demonstrate that the nonpairing parts of RNA structures in the hepatitis C virus (loop and bulge regions) evolve according to a selective regime different from that of nonpairing nucleotides outside the structural regions, indicating a functional role of these. The emphasis here has been on exploring the importance of different model features and the estimation of evolutionary parameters. Other immediate applications of the model are in RNA virus phylogenetics and in comparative RNA structure prediction in coding regions. We are at present pursuing the latter by incorporating the presented model into a stochastic context-free grammar which is capable of generating alignments of coding regions with RNA structures.

Acknowledgments

This study acknowledges support from the following grants: The Danish Natural Science Research Council grants 21-02-0206 and 51-00-0392 (R.F.), grant 51-00-0283 (R.F., J.S.P., J.H.); EPSRC grant HAMJW and MRC grant HAMKA (R.F., I.M. and J.H.); The National Institutes of Health, USA grant 1-R01-GM60729-01 (R.F.). The authors are grateful to Peter Simmonds and Andrew Tuplin for providing information on HCV structures and comments on the manuscript, to Kirsten Saabye for suggestions which greatly improved the readability of the manuscript, and to Mark Springer and four anonymous reviewers for their comments which improved the content and clarity of this manuscript.

Literature Cited

Bruno, W. J., Socci, N. D., and Halpern, A. L. 2000. Weighted Neighbor Joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. **17**:189–197.

Chartrand, P., Meng, X. H., Huttelmaier, S., Donato, D., and Singer, R. H. 2002. Asymmetric sorting of ashlp in yeast results from inhibition of translation by localization elements in the mRNA. Mol. Cell **10**:1319–1330.

Chartrand, P., Meng, X. H., Singer, R. H., and Long, R. M. 1999. Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. Current Biol. **9**:333–336.

Ewens, W. J., and Grant, G. R. 2001. Statistical methods in bioinformatics. Springer–Verlag, New York.

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

———. 1993. *PHYLIP*, Phylogeny Inference Package, 3.5c edition. University of Washington, Seattle, Wash.

Felsenstein, J., and Churchill, G. A. 1996. A hidden markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

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

Goldman, N., and Yang, Z. 1994. A codon-based model of nucleotide substitution of protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–735.

Goodfellow, I. G., Kerrigan, D., and Evans, D. J. 2003. Structure and function analysis of the poliovirus cis-acting replication element (CRE). RNA **9**:124–137.

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

Hein, J., and Stovlbaek, J. 1995. A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. J. Mol. Evol. **40**:181–189.

Katz, L., and Burge, C. B. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. **13**:2042–2051.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol Evol. **16**:111–120.

Knudsen, B., and Hein, J. 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics **15**:446–454.

Liò, P., and Goldman, N. 1998. Models of molecular evolution and phylogeny. Genome Res. **8**:1233–1244.

Lobry, J., and Lobry, C. 1999. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. Mol. Biol. Evol. **16**:719–723.

Meza, J. C. 1994. OPT++: an object-oriented class library for nonlinear optimization. Technical Report SAND94-8225, Sandia National Laboratories, Livermore, Calif.

Muse, S. V., and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol. Evol. **11**:715–724.

Pedersen, A. M., and Jensen, J. L. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. **18**:763–776.

Reynolds, J. E., Kaminski, A., Carroll, A. R., Clarke, B. E., Rowlands, D. J., and Jackson, R. J. 1996. Internal initiation of translation of hepatitis C virus RNA: the ribosome entry site is at the authentic initiation codon. RNA **2**:867–878.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. **20**:1692–1704.

Savill, N. J., Hoyle, D. C., and Higgs, P. G. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics **157**:399–411.

Siepel, A., and Haussler, D. 2003. Phylogenetic estimation of context dependent substitution rates by maximum likelihood. Mol. Biol. Evol. **23**:468–88.

Tuplin, A., Wood, J., Evans, D. J., Patel, A. H., and Simmonds, P. 2002. Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. RNA **8**:824–841.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

———. 1996. Maximum-Likelihood Models for Combined Analyses of Multiple Sequences Data. J. Mol. Evol. **42**: 587–596.

———. 2000. Phylogenetic Analysis by Maximum Likelihood (PAML), 3.0 edition. University College London.

Yi, M., and Lemon, S. M. 2003. 3′ nontranslated RNA signals required for replication of hepatitis C virus RNA. J. Virol. **77**:3557–3568.

Zanotto, P. M., Gibbs, M. J., Gould, E. A., and Holmes, E. C. 1996. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. J. Virol. **70**:6083–6096.

Mark Springer, Associate Editor

Accepted June 22, 2004