View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository

Compensatory Evolution in RNA Secondary Structures Increases Substitution Rate Variation among Sites

Jennifer L. Knies,†[2,3] Kristen K. Dang,‡[3] Todd J. Vision,* Noah G. Hoffman,§[1] Ronald Swanstrom,|| and Christina L. Burch**

*Department of Biology, University of North Carolina, Chapel Hill; †Curriculum in Genetics and Molecular Biology, University of North Carolina, Chapel Hill; ‡Department of Biomedical Engineering, University of North Carolina, Chapel Hill; §Department of Microbiology and Immunology, University of North Carolina, Chapel Hill; ||Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill; and ¶The UNC Center for AIDS Research, University of North Carolina at Chapel Hill

There is growing evidence that interactions between biological molecules (e.g., RNA–RNA, protein–protein, RNA–protein) place limits on the rate and trajectory of molecular evolution. Here, by extending Kimura's model of compensatory evolution at interacting sites, we show that the ratio of transition to transversion substitutions (κ) at interacting sites should be equal to the square of the ratio at independent sites. Because transition mutations generally occur at a higher rate than transversions, the model predicts that κ should be higher at interacting sites than at independent sites. We tested this prediction in 10 RNA secondary structures by comparing phylogenetically derived estimates of κ in paired sites within stems (κ_p) and unpaired sites within loops (κ_u). Eight of the 10 structures showed an excellent match to the quantitative predictions of the model, and 9 of the 10 structures matched the qualitative prediction $\kappa_p > \kappa_u$. Only the Rev response element from the human immunovirus (HIV) genome showed the reverse pattern, with $\kappa_p < \kappa_u$. Although a variety of evolutionary forces could produce quantitative deviations from the model predictions, the reversal in magnitude of κ_p and κ_u could be achieved only by violating the model assumption that the underlying transition (or transversion) mutation rates were identical in paired and unpaired regions of the molecule. We explore the ability of the APOBEC3 enzymes, host defense mechanisms against retroviruses, which induce transition mutations preferentially in single-stranded regions of the HIV genome, to explain this exception to the rule. Taken as a whole, our findings suggest that κ may have utility as a simple diagnostic to evaluate proposed secondary structures.

Introduction

Compensatory mutations, or mutations that are individually deleterious but neutral or beneficial in combination, permit deleterious mutations to be fixed in populations without causing a net fitness loss (Poon and Otto 2000). Experimental evidence from laboratory populations shows that most deleterious mutations can be compensated by numerous mutations at alternative sites (Burch and Chao 1999; Poon and Chao 2005) and that fitness recovery following the fixation of a deleterious mutation most often occurs via compensatory rather than back mutation (Schrag et al. 1997; Maisnier-Patin et al. 2002; Hoffman et al. 2005).

Kimura developed a population genetics model in which deleterious and compensatory mutations can arise within a single genome and occasionally drift to fixation as a pair (Kimura 1985). Because genomes containing single deleterious mutations are not required to become fixed in this process, pairs of compensatory mutations can be fixed at an appreciable rate even if their individual deleterious effects are large and even in large populations. However, because the waiting time for both mutations to arise in the same genome is longer than the waiting time for an independent neutral mutation to arise, the rate of molecular evolution is predicted to be lower at compensatory sites than at independently evolving neutral sites.

The contribution of compensatory mutations to molecular evolution in natural populations has been most thoroughly investigated in regions of RNA secondary structure. RNA secondary structure offers a convenient model for investigating compensatory evolution because individual sites are readily identified as independently evolving (unpaired sites) or involved in a compensatory interaction (stems or paired sites). Compensatory evolution clearly plays a role in the evolution of these regions because mutations in stems are generally accompanied by compensating mutations that maintain base pairing in the stem (Kirby et al. 1995; Wilke et al. 2003; Kern and Kondrashov 2004). As predicted by Kimura's compensatory neutral model, the rates of substitution at compensatory sites are decreased relative to those at independent sites in RNA secondary structures (Stephan 1996; Innan and Stephan 2001), with few exceptions (Wang and Hickey 2002). In fact, this difference in substitution rates at compensatory and independent sites has been used to predict secondary structure (Muse 1995; Pedersen et al. 2004).

In addition to the slowdown in molecular evolution predicted at compensatory sites, inherent differences in rates are also expected to be exaggerated at paired sites in RNA. Specifically, we expect the ratio of transitions to transversions to be exaggerated in the pairing regions. This expectation is derived from the biochemical constraints of base pairing, where a transition mutation can only be compensated by another transition and likewise for transversions. Because transitions occur more frequently than transversions (reviewed in Wakeley 1996), transitions will be compensated more quickly than transversions, resulting in an elevated transition:transversion rate ratio (κ) at paired sites.

Here, by extending Kimura's model specifically to molecular evolution in RNA secondary structures, we make the prediction that the transition to transversion substitution rate ratio (κ) at paired sites should be the square of that for

[1] Present address: Department of Laboratory Medicine, University of Washington.

[2] Present address: Department of Ecology and Evolutionary Biology, Brown University.

[3] J.L.K. and K.K.D. contributed equally to this work.

Key words: molecular evolution, RNA secondary structure, compensatory evolution, transition–transversion ratio.

E-mail: Jennifer_Knies@brown.edu.

Mol. Biol. Evol. 25(8):1778–1787. 2008
doi:10.1093/molbev/msn130
Advance Access publication June 4, 2008

© The Author 2008. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

Table 1
RNA Secondary Structures Analyzed

Structure	C/N	Organism	Source of Sequences
RRE	C	HIV	Los Alamos HIV Database (http://hiv-web.lanl.gov/content/index)[a]
IRES	N	Pestivirus	Viral RNA Structure Database (http://rna.tbi.univie.ac.at/cgi-bin/virusdb.cgi)
CRE[b]	C	Hepatitis C	Los Alamos HCV database (http://hcv.lanl.gov/content/hcv-db/index)
5S rRNA	N	Firmicutes bacteria	5S ribosomal RNA database (http://www.man.poznan.pl/5SData/)
16S rRNA	N	Bacteria	Comparative RNA Web site http://www.rna.ccbb.utexas.edu
23S rRNA	N	Bacteria	Comparative RNA Web site http://www.rna.ccbb.utexas.edu
tRNA[c]	N	Amphibian mitochondria	Organellar Genome Retrieval system (http://drake.physics.mcmaster.ca/ogre/)
tRNA[c]	N	Mammalian mitochondria	Organellar Genome Retrieval system (http://drake.physics.mcmaster.ca/ogre/)
12S rRNA	N	Mammalian mitochondria	GenBank (AB074968, AY172335, U33494–UU3948)
RnaseP	N	Mammals	RNaseP database (http://www.mbio.ncsu.edu/RNaseP/home.html)

NOTE.—C, coding; N, noncoding.

[a] Only one sequence per patient was used.

[b] The sequences used were nonrecombinant, nonrelated type 1 HCV NS5B sequences (positions 7394–9170 of reference sequence M62321).

[c] The tRNAs analyzed were a concatenated alignment of tRNA^ Ala, tRNA^ Cys, tRNA^ Glu, tRNA^ Asn, tRNA^ Gln, and tRNA^ Tyr.

unpaired sites, all other factors being equal. We tested this prediction in 10 functionally and taxonomically diverse RNA molecules and found common, but not universal, quantitative agreement with the model. The prediction that we test may be useful in increasing the accuracy of methods of RNA secondary structure prediction.

Kimura's Model of Compensatory Neutral Evolution

Following Kimura (1985), we consider the substitution process at a pair of loci involved in a compensatory interaction in a diploid population of size N (equivalent results are obtained for a haploid population of size $2N$). Let μ represent the rate of mutation from the wild type to the mutated allele at both loci and ignore back mutation by assuming that selection is sufficiently strong to keep both mutations at a low enough frequency that back mutations are improbable. Selection is assumed to act equivalently on mutations at both sites, so that the fitness of genomes containing either of the 2 mutations is $1 - s$. Because the 2 mutations are involved in a compensatory interaction, however, the fitness of genomes that contain both mutations is equivalent to the wild type (i.e., fitness = 1). Finally, we assume that the loci are sufficiently close together that recombination between them can be ignored. Recombination is minimal in RNA secondary structures of small to moderate size (Parsch et al. 2000). Moreover, although recombination is known to slow down the rate of compensatory evolution (Higgs 1998), it does not measurably affect the predictions we lay out below (data not shown: Using data from table 1 in Innan and Stephan [2001], we constructed rate ratios of time for compensatory evolution at 2 different classes of nucleotide sites that had different mutation rates but experienced equal amounts of recombination and found that recombination had a minimal affect on this ratio). In Kimura's model, the deleterious mutations are assumed to be present initially at an equilibrium frequency determined by the balance between mutation and selection. At this equilibrium, the expected number of alleles carrying either of the 2 deleterious mutations is $4N\mu/s$. Compensating mutations are assumed to arise in genomes that already carry an initial deleterious mutation at rate μ, and the probability that the newly arisen linked pair of mutations drifts to fixation in the population is $1/2N$. Combining these effects, the rate at which compensatory substitutions are introduced at a pair of sites is

$$d_C = \left(\frac{4N\mu}{s}\right) \cdot \mu \cdot \left(\frac{1}{2N}\right) = \frac{2\mu^2}{s}. \qquad (1)$$

Equation (1) is equivalent to equation (8b) of the bidirectional, symmetric model of Stephan (1996). We can compare this compensatory substitution rate to the expectation at independently evolving neutral sites (Kimura 1985):

$$d_I = 2N\mu \cdot \left(\frac{1}{2N}\right) = \mu. \qquad (2)$$

Now we consider evolution in 2 classes of sites—one class in which both sites participating in the compensatory interaction mutate at a faster rate μ_1 and one class in which both sites mutate at a slower rate μ_2. We find that the ratio of the substitution rate between the 2 classes of sites:

$$\frac{d_{C,1}}{d_{C,2}} = (\frac{2\mu_1^2/s}{2\mu_2^2/s}) = \left(\frac{\mu_1}{\mu_2}\right)^2 \qquad (3)$$

is the square of the ratio at independently evolving neutral sites

$$\frac{d_{I,1}}{d_{I,2}} = \frac{\mu_1}{\mu_2}. \qquad (4)$$

To adapt this scenario to the evolution of RNA secondary structures, we make use of the widespread observation that transition mutations (purine-to-purine or pyrimidine-to-pyrimidine) are generally more common than transversion mutations. Furthermore, we note that a transition mutation in one side of an RNA stem structure can only be compensated by another transition mutation and likewise for transversions. Thus, we expect 2 rates of compensatory evolution, one for transitions ($d_{C,\mathrm{Ti}}=2\mu_{\mathrm{Ti}}^2/s$) and another for transversions ($d_{C,\mathrm{Tv}}=2\mu_{\mathrm{Tv}}^2/s$). By assuming that selection against both types of deleterious intermediates acts with the same strength, we predict that the rate ratio of transition to transversion substitutions (κ) in paired regions of RNA secondary structure (stems) should be:

$$\kappa_{\mathrm{p}} = \frac{d_{C,\mathrm{Ti}}}{d_{C,\mathrm{Tv}}} = \left(\frac{\mu_{\mathrm{Ti}}}{\mu_{\mathrm{Tv}}}\right)^2, \tag{5}$$

which is again the square of the rate ratio in unpaired regions (loops):

$$\kappa_{\mathrm{u}} = \frac{d_{I,\mathrm{Ti}}}{d_{I,\mathrm{Tv}}} = \frac{\mu_{\mathrm{Ti}}}{\mu_{\mathrm{Tv}}}. \tag{6}$$

Methods

To test the predictions of the model, we selected 10 RNA molecules with well-documented secondary structures for which a large number of diverse sequences are available. The structures used here were predicted from comparative sequence analyses (12S rRNA, 5S rRNA, and tRNAs), experimental evidence (Rev response element [RRE]), or both (RNase P, internal ribosome entry site [IRES], *cis*-acting replication element [CRE], 16S, and 23S rRNA). For each set of sequences, an alignment and phylogeny were inferred. The value of κ was then estimated for paired versus unpaired sites, and a test was performed to determine if these 2 values differed significantly.

Sources of Sequence Data and Secondary Structures

Sequences were obtained from a variety of sources, as listed in table 1. For each molecule, sequence positions were classified as either paired or unpaired. In most cases, we used structures reported in the literature: RRE (Phuphuakrat and Auewarakul 2003), CRE (Tuplin et al. 2002, 2004), and 12S rRNA (Springer et al. 1995). Structures for IRES, 5S rRNA (Fox and Woese 1975), and RNase P (Haas et al. 1991) were obtained, respectively, from the Viral RNA Structure Database (Thurner et al. 2004), the 5S Ribosomal RNA Database (Szymanski et al. 2002), and the RNase P database (http://www.mbio.ncsu.edu/RNaseP/home.html) (Brown 1999; Harris et al. 2001). The 16S and 23S rRNA structures were obtained from the comparative RNA Web site (http://www.rna.ccbb.utexas.edu), and sites were assigned secondary structure positions according to the reference sequence—*Escherichia coli* (J01695) (Cannone et al. 2002). Finally, tRNA structures were obtained using Mfold v3.0 (http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html) (Zuker 2003) with manual adjustments to fit the canonical model described in Sprinzl et al. (1998).

Alignment and Phylogenetic Inference

The sequences were either obtained having already been aligned or aligned de novo in ClustalW (Chenna et al. 2003) or MAFFT (Katoh et al. 2005). We constructed phylogenies in MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001). In order to constrain parameter values as little as possible, we used the General Time Reversible (GTR) + gamma + invariant model of nucleotide substitution and set the remaining parameters at their default value. For IRES, 5S rRNA, and RNase P alignments 100,000 generations with a 10% (10,000 generations) burn-in was sufficient for convergence. For the 16S and 23S rRNA alignments, convergence was achieved in 250,000 generations with a 50% burn-in. For the 12S rRNA alignment and the mitochondrial alignments, convergence was achieved in 500,000 generations with a 10% and 20% burn-in, respectively. For the CRE and RRE alignments, convergence was achieved in 1 million generations with a 20% and 50% burn-in, respectively. Three independent runs were conducted for each alignment, and the log likelihood values of these runs were compared to confirm that the chains converged on the same posterior distribution. The consensus tree was obtained by majority rule. There was no evidence of saturation as the distance from root to tip was <1 in all phylogenies.

In 4 cases, we were unable to use all the available sequences because either the phylogenetic method did not converge (RRE), the resulting phylogenetic tree was poorly supported (5S rRNA), or the time required for the phylogenetic method was excessive (16S and 23S rRNA). In the case of RRE, we used an arbitrary subset of 199 of the available sequences. In the case of 5S rRNA, we chose sequences from 2 genera (*Bacillus* and *Clostridium*) because the phylogeny built from these sequences achieved convergence. For 16S and 23S rRNA, we trimmed the sequence alignment to 94 and 100 sequences, respectively, by choosing every 10th sequence from the highly refined seed alignment downloaded from the comparative RNA Web site.

The alignments and trees for each molecule have been uploaded to Dryad (http://hdl.handle.net/10255/dryad.162).

Estimation of Substitution Rate Parameters

We used the program HyPhy (Kosakovsky Pond et al. 2005) to estimate κ separately for paired and unpaired regions of each molecule. We incorporated rate heterogeneity with a discretized gamma distribution of mutation rates (4 rate classes) and used the HKY85 model of nucleotide substitution (Hasegawa et al. 1985), which allows for unequal base frequencies, one substitution rate for all transitions, and one for all transversions. We chose this model in order to obtain a single estimate for the transition–transversion rate ratio (κ) with reasonably small confidence intervals (CIs), even though it is not necessarily the best fit for each alignment (see below). For each alignment, we report κ and the approximate 95% CIs derived from the Fisher information matrix. Parameter estimates of κ are expected to be robust to minor errors in tree topology (Hillis 1999).

We also investigated whether the HKY85 model, which allows a single rate for all transitions and a single rate for all transversions, gives a reasonable approximation of the observed substitution patterns. For each alignment, we found the best nucleotide substitution model using the model-testing procedure implemented in HyPhy (Kosakovsky Pond et al. 2005), which is based on that of ModelTest (Posada and Crandall 2001). We fit models with discrete gamma distributed rate variation and a fraction of invariant sites.

Table 2
Description of Variation in Sequence Alignments

		No. Sites[a]			Diversity[b]	
Structure	No. Taxa	Overall	Stems	Loops	Stems	Loops
RRE	199	231 (140)	161 (91)	70 (49)	0.01[c]	0.04[c]
IRES	11	347 (162)	198 (92)	149 (70)	0.27	0.27
CRE	98	255 (119)	189 (72)	66 (45)	0.05	0.08
5S rRNA	15	117 (86)	73 (61)	44 (25)	0.27	0.27
16S rRNA[d]	94	1542 (972)	936 (688)	606 (304)	0.23[c]	0.11[c]
23S rRNA[d]	100	2904 (2102)	1684 (1377)	1220 (752)	0.29[c]	0.14[c]
A tRNA	40	414 (344)	266 (224)	148 (120)	0.20	0.19
M tRNA	40	419 (187)	243 (98)	176 (89)	0.05[c]	0.15[c]
12S rRNA[e]	7	930 (406)	457 (173)	473 (233)	0.29[c]	0.43[c]
RNase P[f]	10	229 (106)	129 (72)	100 (34)	0.20	0.30

[a] Numbers in parentheses indicate the number of variable positions.

[b] Diversity was calculated as the median across all positions of the fraction of taxa different from the consensus nucleotide. Majority nucleotide is designated as consensus. Gaps do not change the consensus. In the case of a tie, an ambiguity character is reported as consensus and the number of differences from consensus is counted as the total number minus the number of one of the types of the majority character.

[c] Significantly different by a Wilcoxon rank sum test (P value < 0.01). The null hypothesis for this test is that diversity at stem positions is not significantly different from diversity at loop positions.

[d] Sites not present in the reference sequence (*Escherichia coli*: J01695) were removed from the alignment.

[e] Excluding the region between stem 38 and its complement 38′, which is highly variable and difficult to align.

[f] Excluding loops P3, P9, P10, P15.1, and P18, which are highly variable in length and are difficult to align.

Tests of Predictions

Likelihood ratio tests were used to decide whether the data support estimation of a separate κ for paired and unpaired regions. Formally,

$$\begin{aligned} &H_O\colon \kappa = \kappa_p = \kappa_u \text{ and} \\ &H_A\colon \kappa_p \neq \kappa_u, \end{aligned} \tag{6}$$

where κ is the transition–transversion ratio estimated from the entire molecule, κ_p is the transition–transversion ratio estimated from an analysis of paired positions only, and κ_u is the transition–transversion ratio estimated from an analysis of unpaired positions only. We calculated a test statistic, λ, for each alignment in the following way:

$$\lambda = -2\left(\ln L\kappa - \left(\ln L\kappa_p + \ln L\kappa_u\right)\right). \tag{7}$$

The statistical significance of λ was evaluated assuming a χ^2 distribution with one degree of freedom. For each molecule (or alignment), the likelihood values were calculated by holding constant the phylogenetic tree and the rate variation parameter α (estimated from the complete molecule). Only the parameter values of the nucleotide substitution model (HKY85) were allowed to vary.

Results

We tested the prediction that compensatory evolution in regions of RNA secondary structure should result in the transition to transversion rate ratio (κ) at paired sites being the square of that at unpaired sites by examining 10 different RNA secondary structures. These molecules were selected from a diverse group of organisms: viruses, bacteria, amphibians, and mammals (table 1). In order to accurately estimate κ, we used alignments that had at least 20 variable paired and 20 variable unpaired sites and for which the secondary structure was known to be conserved and functionally important. The alignments had varying amounts of sequence diversity (table 2) and complexity of secondary structures (fig. 1) ranging from relatively simple tRNAs with 3 stem-loops to the 12S rRNA with approximately 20 stem-loops. In addition, 2 of the alignments (16S and 23S rRNA) showed significantly more diversity among paired sites (stems) than unpaired sites (loops), consistent with previous studies that reported lineage-specific elevations of substitution rates in paired as compared with unpaired regions among archaea and bacteria RNA secondary structures (Smit et al. 2007).

Testing Kimura's Model of Compensatory Evolution

Our approach assumes that transition substitution rates (G ↔ A, C ↔ T) are more similar to each other than they are to transversion substitution rates, and vice versa, and that the 2 rates differ substantially. To examine these assumptions, we used a model-testing procedure that allowed any combination of the 6 time-reversible substitution rates (see Methods) to find the best-fitting nucleotide substitution model for each alignment. The best-fit models are shown in figure 2. Although the HKY85 model, which specifies one rate for transitions and one rate for transversions, was not statistically the best-fit model for any alignment (as determined by Akaike Information Criterion scores), the best-fit models did show differences between transition and transversion rates, and transition rates were almost universally more similar to each other than to transversion rates. Only in the case of the 5S rRNA, alignment was one of the transversion rates (A ↔ T) similar to the transition rates. Thus, it is justifiable to use the HKY85 model as a reasonable approximation for the purposes of comparing estimates of κ.

We estimated κ in 2 ways. First, we considered the molecule as a whole and estimated a single κ to describe all sites (paired and unpaired). Second, we divided the molecule into paired and unpaired sites and estimated κ_p from all paired sites and κ_u from all unpaired sites. Nine of the 10 structures showed $\kappa_p > \kappa_u$ (fig. 3), qualitatively supporting

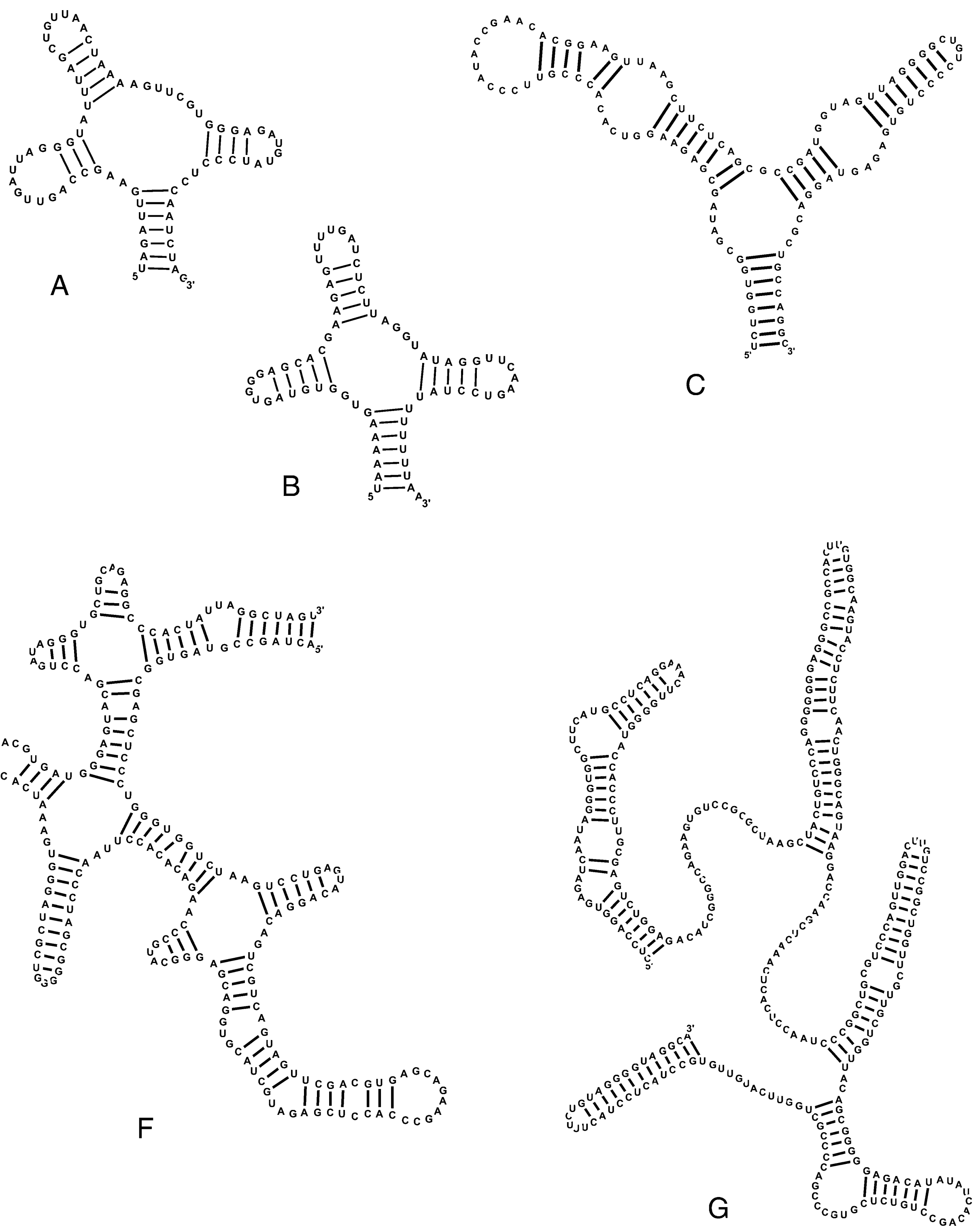

FIG. 1.—RNA secondary structures. Representative (*A*) mitochondrial tRNA for asparagine (from mammalian); (*B*) mitochondrial tRNA for glutamine (from amphibian); (*C*) 5s rRNA; (*D*) RNaseP (shown is the sequence of *Bacillus brevis*)—black bars represent a pseudoknot; (*E*) RRE from HIV-1; (*F*) IRES; (*G*) CRE; (*H*) 12s rRNA (shown is the sequence of *Bos taurus*). The secondary structures of 16S and 23S rRNA can be found in Cannone et al. (2002).

the prediction of Kimura's model. Only the RRE structure showed $\kappa_p < \kappa_u$. Likelihood ratio tests (see Methods for details) led to rejection of H_O: $\kappa = \kappa_p = \kappa_u$ in favor of H_A: $\kappa_p \neq \kappa_u$ for all 10 molecules at $P < 0.0001$ (table 3).

We also examined the quantitative fit of the estimates to the model predictions. A visual inspection of figure 3 confirms the close match of most structures to the expectation that $\kappa_p = \kappa_u^2$. For 8 structures (12S

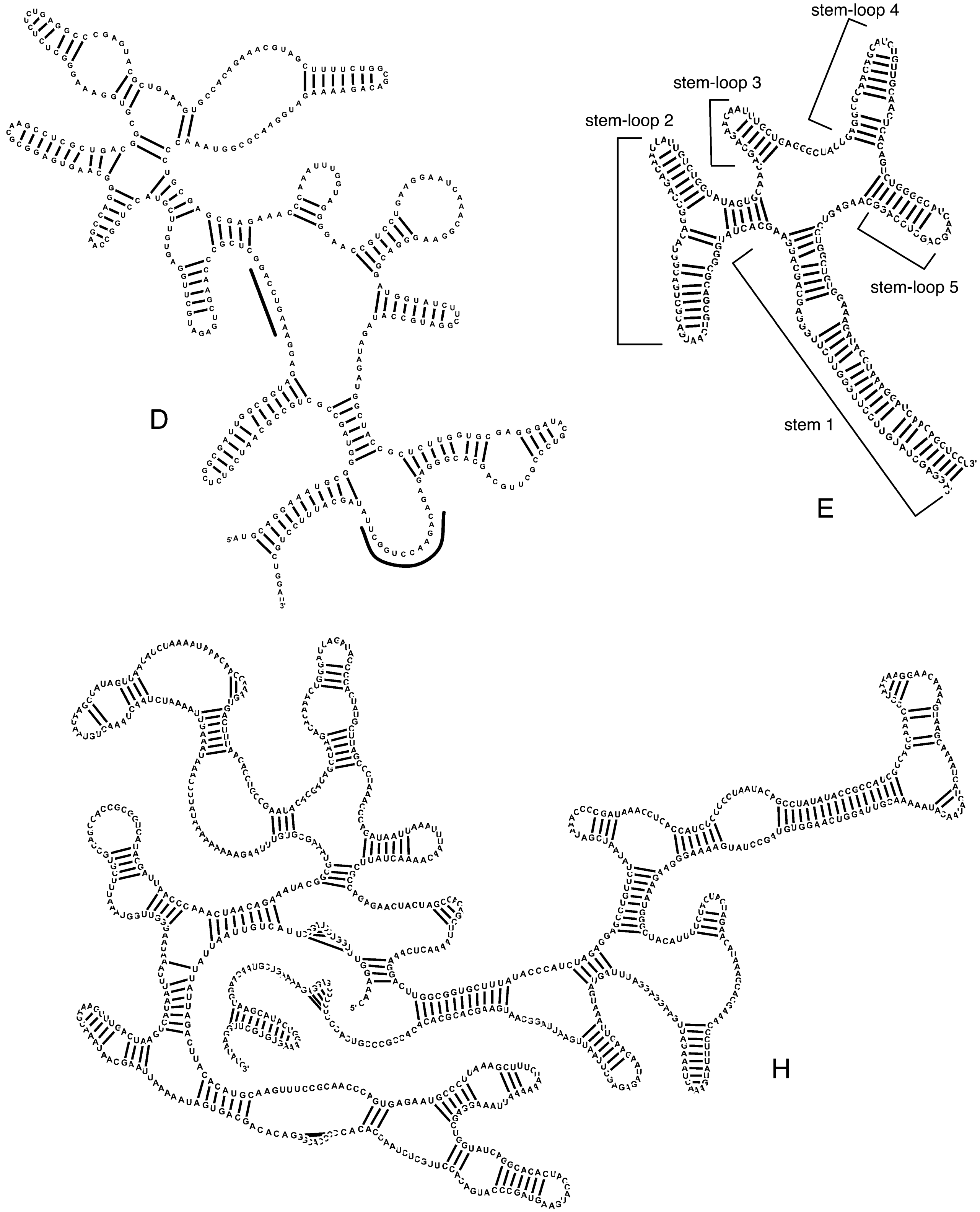

FIG. 1.—(Continued)

rRNA, 5S rRNA, 16S rRNA, 23S rRNA, amphibian tRNAs, RNase P, IRES, and CRE), estimates of κ_p cannot be statistically distinguished from κ_u^2, that is, 95% CIs estimated from the Fisher Information matrix overlap the $\kappa_p = \kappa_u^2$ line. Only the human immunovirus (HIV) RRE structure, and to a lesser extent the mammalian tRNAs, deviate significantly from the model prediction.

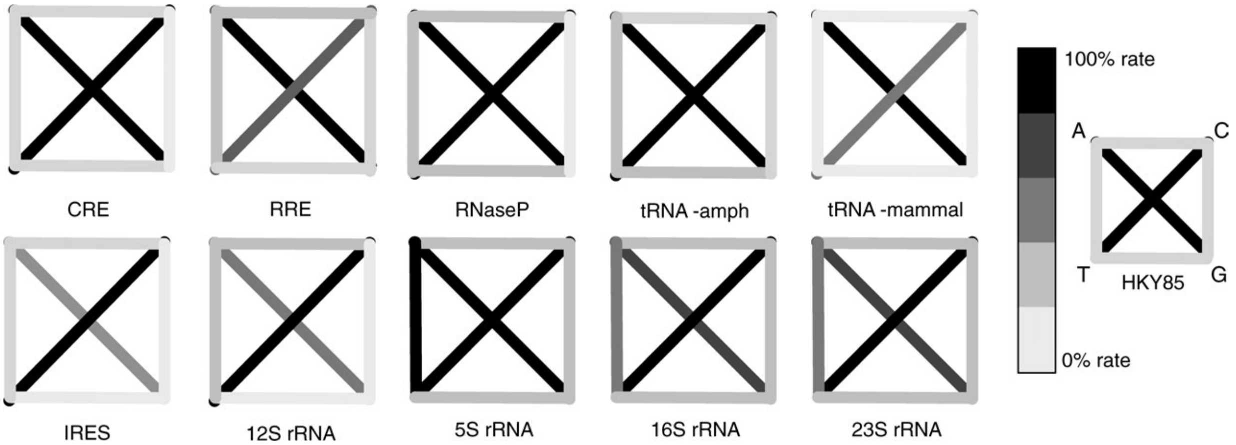

FIG. 2.—Best-fit nucleotide substitution models for each alignment. Shown is a cartoon illustration of the rate categories of the best-fit nucleotide substitution models for each molecule. Within a molecule, rates were scaled to the maximum rate (black). Diagonal lines depict transitions; the edges of the square depict transversions. The HKY85 model, which was used for the rate ratios reported throughout this article, is shown for comparison on the right.

Substitution Patterns in RRE

We examined 3 possible explanations for the surprising result that $\kappa_p < \kappa_u$ in RRE. First, because both the RRE and CRE secondary structures occur within coding regions, we examined the possibility that the difference between κ_p and κ_u is diminished by selection on the protein sequence. We recalculated κ_p and κ_u for both molecules using only data from 4-fold degenerate sites in paired and unpaired regions. In CRE, the presence of codons affects the estimates in the predicted direction (4-fold degenerate sites: $\kappa_p = \kappa_u^{2.89}$; all sites: $\kappa_p = \kappa_u^{1.45}$), though the 4-fold sites overshoot the predicted pattern. We had less power to compare 4-fold degenerate sites at the paired and unpaired sites of RRE because there were too few 4-fold degenerate unpaired sites, and there was insufficient sequence variability at these sites. However, the 4-fold degenerate paired sites did show a higher κ_p ($\kappa_p = 7.61$ with 95% CI [4.79–18.48]) than the paired sites as a whole ($\kappa_p = 4.21$ with 95% CI [3.51–5.28]). This suggests that the presence of protein-coding constraints does impede compensatory evolution at paired sites in RNA secondary structures, although it does not explain why κ_u would be "greater" than κ_p in RRE.

Second, we examined the possibility that we had used a nonrepresentative sample of RRE sequences. To confirm that the observed substitution patterns in RRE were not specific to the particular set of HIV sequences we examined (which were all derived from subtype B), we estimated κ_p and κ_u from 2 additional RRE alignments of sequences drawn from higher taxonomic levels: sequences from different subtypes (1 sequence each from A, B, C, F, G, H, J, and K) and sequences from different groups (1–2 sequences each from M, N, and O) of HIV. In both these alignments, the results were qualitatively similar to those for subtype B: κ_u was significantly higher than κ_p (table 4).

Third, we considered whether the RRE estimates were disproportionately influenced by a portion of the molecule that experiences a type of selection that differs from the molecule as a whole. We systematically removed each stem-loop of RRE and reestimated κ_p and κ_u for the resulting partial structures. The κ_p and κ_u estimates were qualitatively similar for all these partial structures (table 5).

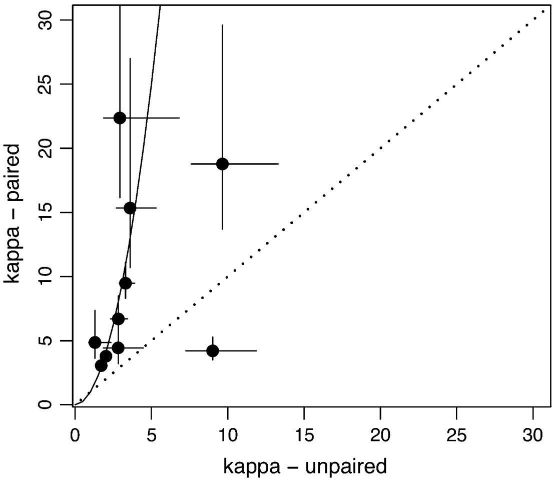

FIG. 3.—Transition–transversion rate ratios (κ) for each alignment. The dotted line represents a 1:1 relationship between κ_p and κ_u. The solid line represents the predicted relationship $\kappa_p = \kappa_u^2$. Note that the CRE data point is from the analysis of 4-fold degenerate sites in paired and unpaired regions.

Table 3
Transition–Transversion Rate Ratios (κ_p)

Structure	κ	κ_p	κ_u	λ
RRE	5.19	4.21	9.01	546.05[a]
IRES	6.50	15.34	3.60	73.46[a]
CRE	12.52	22.36	2.93	177.32[a]
5S rRNA	3.70	4.44	2.82	35.05[a]
16S rRNA	3.24	3.79	2.02	665.64[a]
23S rRNA	2.57	3.06	1.71	1281.71[a]
A tRNA	6.04	9.48	3.30	204.73[a]
M tRNA	11.98	18.78	9.65	122.24[a]
12S rRNA	3.90	6.69	2.83	131.93[a]
RNase P	2.98	4.86	1.30	59.21[a]

[a] LRT value significant at $P < 0.0001$

Table 4
Transition–Transversion Rate Ratios (κ) for RRE alignments at Different Taxonomic Levels

RRE alignment	κ_p[a]	κ_p[a]
Subtype B	4.21 (3.51–5.28)	9.01 (7.25–11.89)
Subtypes	2.10 (1.45–3.09)	7.26 (3.88–15.09)
Groups	5.56 (3.21–10.46)	15.96 (7.00–46.29)

[a] 95% CIs in parentheses.

Table 5
Transition–Transversion Rate Ratios (κ) for Partial RRE Structures

Removal of RRE stem-loop[a]	κ_p[b]	κ_p[b]
None[c]	4.21 (3.51–5.28)	9.01 (7.25–11.89)
I	4.70 (3.67–6.54)	8.03 (6.38–10.77)
II	5.43 (4.37–7.19)	9.73 (7.40–14.17)
III	3.36 (2.78–4.24)	8.18 (6.55–10.92)
IV	4.47 (3.70–5.66)	11.21 (8.77–15.51)
V	4.02 (3.34–5.05)	8.35 (6.70–11.07)

[a] See figure 1.
[b] 95% CIs in parentheses.
[c] See table 3.

Discussion

We have extended Kimura's population genetics model of compensatory evolution to make the prediction that substitution rate variation due to underlying mutation rate differences is exaggerated by compensatory interactions. We made use of the fact that transition rates generally exceed transversion rates to test this prediction in regions of RNA secondary structure. Specifically, we predict that κ, the ratio of transition to transversion substitutions, should be higher in paired than in unpaired regions. Nine of the 10 RNA secondary structures we examined confirmed this qualitative prediction. Moreover, 8 of the 10 structures (RNase P, 5S rRNA, 16S rRNA, 20S rRNA, IRES, tRNA A, 12S rRNA, and CRE) closely matched the quantitative prediction that the ratio in paired regions should be the square of the ratio in unpaired regions (i.e., $\kappa_p = \kappa_u^2$).

Remarkably, we observed a close quantitative match even in sequence alignments where we had an a priori reason to believe that model assumptions were violated. Alignments that showed more diversity among paired sites (stems) than unpaired sites (loops), in which selection may have been acting on the RNA primary sequence to constrain evolution in loops (Smit et al. 2007), nonetheless showed a close match to the model prediction. Finally, the model prediction appears robust to the amount of signal in the sequence data, in that sequence alignments with both low and high amounts of sequence diversity performed equally well in our analysis.

Implications for RNA Secondary Structure Prediction

The close match to the $\kappa_p = \kappa_u^2$ expectation in most structures confirms the role of compensatory interactions in the molecular evolution of RNA secondary structures and suggests the use of the transition to transversion rate ratio, κ, as a simple diagnostic to evaluate proposed secondary structures. In addition, the observed large deviations from the neutral expectation $\kappa_p = \kappa_u$ should make substitution rate variation a useful tool for structure prediction. Indeed, recently developed methods for predicting RNA secondary structure from sequence data capitalize on the expectation that substitution rates should be lower at paired than unpaired sites (Muse 1995; Pedersen et al. 2004). These methods have been shown to be as strong or stronger in validating known structures and predicting new ones than previous methods (Muse 1995; Parsch et al. 2000; Pedersen et al. 2004).

Our results can be used to refine these methods for secondary structure prediction by providing an exact expectation for the ratio of different classes of substitution at unpaired and paired sites. In particular, we find that $\kappa_p \approx \kappa_u^2$ because transversions are suppressed in stems to a greater extent than transitions (i.e., $d_{C,\mathrm{Tv}}/d_{I,\mathrm{Tv}} < d_{C,\mathrm{Ti}}/d_{I,\mathrm{Ti}}$). By contrast, the secondary structure prediction method of Muse (1995) assumes that transition and transversion rates are equally (multiplicatively) reduced in stems. The secondary structure prediction method of Pedersen et al. (2004) uses empirically measured substitution patterns in known stems to estimate transition and transversion rates; that is, it allows $\kappa_p \neq \kappa_u$, but the measure of κ_p is specified identically for all structures in all organisms.

Structures not Explained by the Model

Although most structures showed a good fit to the model predictions, 2 structures deviated either quantitatively (mammalian tRNAs) or qualitatively (RRE). We were particularly puzzled by the observation that the relative magnitude of the rate ratios in RRE was reversed from the model prediction, so that $\kappa_p < \kappa_u$. The only feature that the 2 molecules share is a high κ_u value compared with the others. However, a high ratio of the underlying mutation rates μ_{Ti} to μ_{Tv} does not, in itself, violate the model's assumptions. Here, we consider how the assumptions of the model could be violated in a way that preferentially elevates κ_u, with a particular focus on mechanisms that could explain the observation in RRE that $\kappa_p < \kappa_u$.

The model assumes that 1) the only factor that differs among sites is their paired or unpaired status within the RNA secondary structure, 2) recombination is infrequent, and 3) the underlying mutation rate μ_{Ti} is the same in paired and unpaired regions and likewise for μ_{Tv}. Violations of the first 2 assumptions appear to have had only minor effects on our estimates of κ_u and κ_p. Coding regions cause additional selective differences among sites, and the presence of codons was shown to affect the estimates of κ for CRE and RRE, but the size of the effect was not sufficient to explain a reversal of the relative magnitudes of κ_u and κ_p in RRE. Recombination between HIV genomes of the same or very closely related genotypes will minimally affect κ_u and κ_p. Recombination between divergent genotypes may have a greater effect on κ_u and κ_p, but the frequency of this process is controversial and is believed to be low (Smith et al. 2005).

In contrast, violation of the third assumption that the underlying mutation rate μ_{Ti} is the same in paired and unpaired regions (and likewise for μ_{Tv}) may provide a plausible explanation even for the observation that $\kappa_p < \kappa_u$. In RRE, we identified a potential mechanism for raising the transition mutation rate in unpaired regions of the genome—the host cytosine deaminating enzyme family APOBEC3. After entering a host cell, the HIV RNA genome is reverse transcribed into single-stranded DNA prior to incorporation into the host's genome. During this process, the unpaired regions in the DNA genome (including RRE) are vulnerable to the action of APOBEC3G/F, which preferentially deaminates C to T (a transition) within specific motifs in unpaired regions of retroviral DNA. These mutations are observed as G to A transitions on the resulting plus-strand genome of HIV (Harris and Liddament 2004). The unpaired motifs are, thus, predicted to experience an elevated μ_{Ti}, which would explain the reversal of κ_u and κ_p in RRE. Our strongest test of this hypothesis comes from an examination of the sequence alignments in which adenine is overrepresented in unpaired regions (~40%) of RRE compared with paired regions (~19%), consistent with the preferential action of APOBEC3 on unpaired sites. In contrast, we did not find evidence that the APOBEC3 F and G target motifs GA and GG were underrepresented in unpaired regions or that the G ↔ A substitution rate was elevated in unpaired regions compared with the C ↔ T rate (data not shown). However, the power of the latter tests was limited by the small statistical power associated with assessing dinucleotide frequencies and the inability to specify the G → A substitution rate separately from the A → G rate in the GTR substitution model used here. Together, we take these results to provide at least some support for the hypothesis that the action of the APOBEC enzymes on the minus single-stranded DNA containing some secondary structure caused an elevation of μ_{Ti} in unpaired sites, explaining the observation in RRE that $\kappa_p < \kappa_u$. However, a rigorous test of this hypothesis would require a more detailed analysis of molecular evolution across the HIV genome.

Implications for Secondary Structure Evolution

Extending Kimura's model of compensatory evolution allowed the successful quantitative prediction of substitution rates (Stephan 1996; Innan and Stephan 2001) and rate variation (our study) in RNA secondary structures. The success of these predictions confirms the existence of strong constraints on both nucleotide substitutions and structural evolution in these molecules. If nucleotide substitutions in RNA secondary structures were not constrained by compensatory interactions, but often proceeded by a neutral process, we would expect the difference between estimates of κ_u and κ_p to be smaller than the model prediction $\kappa_p = \kappa_u^2$ and to approach the neutral expectation $\kappa_p = \kappa_u$. Similarly, structural evolution would cause some sites that are paired in the reference sequence to be unpaired in nonreference sequences, and vice versa, resulting in a smaller difference between estimates of κ_u and κ_p than predicted by the model. Thus, the close quantitative agreement with the model for most molecules confirms that substitution rates in secondary structures have been governed by a history of compensatory evolution and suggests that there has been little structural evolution in these molecules even over long evolutionary time periods.

Funding

National Institutes of Health (NIH) (grants GM067940 to C.L.B. and AI44667 to R.S.); National Science Foundation (grant DBI-0227314 to T.J.V.); NIH Training Grant (T32-AI07001 to J.L.K. and T32-AI007419 to K.K.D.).

Acknowledgments

The authors thank Drs Stefanie Hartmann, Sergei Kosakovsky Pond, Derrick Zwickl, and Daniel Weinreich for advice and technical assistance.

Literature Cited

Brown J. 1999. The Ribonuclease P Database. Nucleic Acids Res. 27:314.

Burch C, Chao L. 1999. Evolution by small steps and rugged landscapes in the RNA virus phi6. Genetics. 151:921–927.

Cannone JJ, Subramanian S, Schnare MN, Pande N, Shang Z, Yu N, Gutell RR. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structural information for ribosomal, intron, and other RNAs. BMC Bioinformatics. 3:2.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson T, Higgins D, Thompson J. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31:3497–3500.

Fox GE, Woese CR. 1975. 5S-Rna secondary structure. Nature 256:505–507.

Haas ES, Morse DP, Brown JW, Schmidt FJ, Pace NR. 1991. Long-range structure in ribonuclease-P Rna. Science. 254: 853–856.

Harris J, Haas E, Williams D, Frank D, Brown J. 2001. New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. RNA. 7:220–232.

Harris RS, Liddament MT. 2004. Retroviral restriction by APOBEC proteins. Nature Rev Immunol. 4:868–877.

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

Higgs P. 1998. Compensatory neutral mutations and the evolution of RNA. Genetics. 102/103:91–101.

Hillis DM. 1999. Phylogenetics and the study of HIV. In: Crandall KA, editor. The evolution of HIV. Baltimore (MD): Johns Hopkins University Press. p. 504.

Hoffman NG, Schiffer CA, Swanstrom R. 2005. Covariation of amino acid positions in HIV-1 protease. Virology. 331: 206–207.

Huelsenbeck J, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Innan H, Stephan W. 2001. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics. 159:389–399.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Kern AD, Kondrashov FA. 2004. Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. Nat Genetics. 36:1207–1212.

Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. J Genet. 64:7–19.

Kirby DA, Muse SV, Stephan W. 1995. Maintenance of pre-messenger-Rna secondary structure by epistatic selection. Proc Natl Acad Sci USA. 92:9047–9051.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21: 676–679.

Maisnier-Patin S, Berg OG, Liljas L, Andersson DI. 2002. Compensatory adaptation to the deleterious effect of antibiotic resistance in Salmonella typhimurium. Mol Microbiol. 46: 355–366.

Muse SV. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics. 139:1429–1439.

Parsch J, Braverman J, Stephan W. 2000. Comparative sequence analysis and patterns of covariation in RNA secondary structure. Genetics. 154:909–921.

Pedersen J, Meyer I, Forsberg R, Simmonds P, Hein J. 2004. A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic Acids Res. 32:4925–4936.

Phuphuakrat A, Auewarakul P. 2003. Heterogeneity of HIV-1 Rev response element. Aids Res Hum Retroviruses. 19:569–574.

Poon A, Chao L. 2005. The rate of compensatory mutation in the DNA bacteriophage {phi}X174. Genetics. 170:989–999.

Poon A, Otto S. 2000. Compensating for our load of mutations: freezing the meltdown of small populations. Evol Int J Org Evol. 54:1467–1479.

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. Syst Biol. 50:580–601.

Schrag SJ, Perrot V, Levin BR. 1997. Adaptation to the fitness costs of antibiotic resistance in Escherichia coli. Proc R Soc Lond B Biol Sci. 264:1287–1291.

Smit S, Widmann J, Knight R. 2007. Evolutionary rates vary among rRNA structural elements. Nucleic Acids Res. 1–16.

Smith DM, Richman DD, Little SJ. 2005. HIV superinfection. J Infect Dis. 192:438.

Springer M, Hollar L, Burk A. 1995. Compensatory substitutions and the evolution of the mitochondrial 12S rRNA gene in mammals. Mol Biol Evol. 12:1138–1150.

Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. 26:148–153.

Stephan W. 1996. The rate of compensatory evolution. Genetics. 144:419–426.

Szymanski M, Barciszewska M, Erdmann V, Barciszewski J. 2002. 5S Ribosomal RNA Database. Nucleic Acids Res. 30:176–178.

Thurner C, Witwer C, Hofacker IL, Stadler PF. 2004. Conserved RNA secondary structures in Flaviviridae genomes. J Gen Virol. 85:1113–1124.

Tuplin A, Evans D, Simmonds P. 2004. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. J Gen Virol. 85: 3037–3047.

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

Wakely J. 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends in Ecology & Evolution. 11:158–162.

Wang H, Hickey D. 2002. Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. Nucleic Acids Res. 30:2501–2507.

Wilke C, Lenski R, Adami C. 2003. Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding. BMC Evol Biol. 3:3.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406–3415.

Jeffrey Thorne, Associate Editor

Accepted May 30, 2008